ENCYCLOPEDIA OF MODERN OPTICS



EDITED BY
B.D. GUENTHER
A. MILLER
L. BAYVEL
J.E. MIDWINTER

VOLUME ONE



EDITOR-IN-CHIEF

Robert D. Guenther
Duke University
Durham, NC, USA

EDITORS

Duncan G. Steel
University of Michigan
Ann Arbor, MI, USA

Leopold Bayvel
London, UK

CONSULTING EDITORS

John E. Midwinter
University College London, London, UK

Alan Miller
University of St Andrews, Fife, UK

EDITORIAL ADVISORY BOARD

R Alferness
Lucent Technologies Bell Laboratories
Holmdel, NJ, USA

D J Brady
Duke University
Durham, NC, USA

J Caulfield
Diversified Research Corporation
Cornersville, TN, USA

R L Donofrio
Display Device Consultants LLC
Ann Arbor, MI, USA

J Dudley
Université de Franche-Comté
Besançon, France

J G Eden
University of Illinois
Urbana, IL, USA

H O Everitt
U.S. Army Research Office
Research Triangle Park, NC, USA

E L Hecht
Adelphi University
Garden City, NY, USA

H S Hinton
Utah State University
Logan, UT, USA

T N Hurst
Hewlett-Packard Laboratories
Tucson, AZ, USA

D Killinger
University of South Florida
Tampa, FL, USA

T A King
University of Manchester
UK

E Leith
University of Michigan
Ann Arbor, MI, USA

A Macleod
Thin Film Center, Inc.
Tucson, AZ, USA

P F McManamon
Air Force Research Laboratory
Wright-Paterson AFB, OH, USA

J N Mait
U.S. Army Research Laboratory
Adelphi, MD, USA

D Phillips
Imperial College
London, UK

C R Pidgeon
Heriot-Watt University
Edinburgh, UK

R Pike
King's College London
UK

E Van Stryland
University of Central Florida
Orlando, FL, USA

Preface

We live in a world powered by light, but much of the understanding of light was developed around the time of the French Revolution. Before the 1950s, optics technology was viewed as well established and there was little expectation of growth either in scientific understanding or in technological exploitation. The end of the Second World War brought about huge growth in scientific exploration, and the field of optics benefited from that growth. The key event was the discovery of methods for producing a source of coherent radiation, with the key milestone being the demonstration of the first laser by Ted Maiman in 1960. Other lasers, nonlinear optical phenomena, and technologies such as holography and optical signal processing, followed in the early 1960s. In the 1970s the foundations of fiber optical communications were laid, with the development of low-loss glass fibers and sources that operated in the wavelength region of low loss. The 1980s saw the most significant technological accomplishment: the development of efficient optical systems and resulting useful devices. Now, some forty years after the demonstration of the first coherent light source, we find that optics has become the enabling technology in areas such as:

- information technology and telecommunications;
- health care and the life sciences;
- sensing, lighting, and energy;
- manufacturing;
- national defense;
- manufacturing of precision optical components and systems; and
- optics research and education.

We find ourselves depending on CDs for data storage, on digital cameras and printers to produce our family photographs, on high speed internet connections based on optical fibers, on optical based DNA sequencing systems; our physicians are making use of new therapies and diagnostic techniques founded on optics.

To contribute to such a wide range of applications requires a truly multidisciplinary effort drawing together knowledge spanning many of the traditional academic boundaries. To exploit the accomplishments of the past forty years and to enable a revolution in world fiber-optic communications, new modalities in the practice of medicine, a more effective national defense, exploration of the frontiers of science, and much more, a resource to provide access to the foundations of this field is needed. The purpose of this Encyclopedia is to provide a resource for introducing optical fundamentals and technologies to the general technical audience for whom optics is a key capability in exploring their field of interest.

Some 25 internationally recognized scientists and engineers served as editors. They helped in selecting the topical coverage and choosing the over 260 authors who prepared the individual articles. The authors form an international group who are expert in their discipline and come from every part of the technological community spanning academia, government and industry. The editors and authors of this Encyclopedia hope that the reader finds in these pages the information needed to provide guidance in exploring and utilizing optics.

As Editor-in-Chief I would like to thank all of the topical editors, authors and the staff of Elsevier for each of their contributions. Special thanks should go Dr Martin Ruck of Elsevier who provided not only organizational skills but also technological knowledge which allowed all of the numerous loose ends to be tied.

B D Guenther
Editor-in-Chief

A

ALL-OPTICAL SIGNAL REGENERATION

O Leclerc, Alcatel Research & Innovation, Marcoussis, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The breakthrough of optical amplification, combined with the techniques of wavelength division multiplexing (WDM) and dispersion management, have made it possible to exploit a sizeable fraction of the optical fiber bandwidth (several terahertz). Systems based on 10 Gbit/s per channel bit-rate and showing capacities of several terabit/s, with transmission capabilities of hundreds or even thousands of kilometers, have reached the commercial area.

While greater capacities and spectral efficiencies are likely to be reached with current technologies, there is potential economic interest in reducing the number of wavelength channels by increasing the channel rate (e.g., 40 Gbit/s). However, such fourfold increase in the channel bit-rate clearly results in a significant increase in propagation impairments, stemming from the combined effects of noise accumulation, fiber dispersion, fiber nonlinearities, and inter-channel interactions and contributing to two main forms of signal degradation. The first one is related to the amplitude domain; power levels of marks and spaces can suffer from random deviations arising from interaction between signal and amplified spontaneous emission (ASE) noise or with signals from other channels through cross-phase modulation (XPM) from distortions induced by chromatic dispersion. The second type of signal degradations occurs in the time domain; time position of pulses can also suffer from random deviations arising from interactions between signal and ASE noise through fiber dispersion. Preservation of high power contrast between '1' and '0', and of both amplitude fluctuations and timing jitter below some acceptable levels, are mandatory for high transmission quality, evaluated through bit-error-rate (BER) measurements or estimated by Q-factors. Moreover, in future optical networks, it appears mandatory to ensure similar but high optical signal quality at the output of whatever nodes in the networks, as to enable successful transmission of the data over arbitrary distance.

Among possible solutions to overcome such systems limitations is the implementation of Optical Signal Regeneration, either in-line for long-haul transmission applications or at the output of network nodes. Such Signal Regeneration performs, or should be able to perform, three basic signal-processing functions that are Re-amplifying, Re-shaping, and Re-timing, hence the generic acronym '3R' (**Figure 1**). When Re-timing is absent, one usually refers to the regenerator as '2R' device, which has only re-amplifying and re-shaping capabilities. Thus, full 3R regeneration with retiming capability requires clock extraction.

Given system impairments after some transmission distance, two solutions remain for extending the actual reach of an optical transmission system or the scalability of an optical network. The first consists in segmenting the system into independent trunks, with full electronic repeater/transceivers at interfaces (we shall refer to this as 'opto-electronic regeneration' or O/E Regeneration forthwith). The second solution, i.e., all-optical Regeneration, is not the optical version of the first which would have higher bandwidth capability but still performs the same signal-restoring functions with far reduced complexity. At this point, it should be noted that Optical 3R techniques are not necessarily void of any electronic functions (e.g., when using electronic clock recovery and O/E modulation), but the main feature is that these electronic functions are narrowband (as opposed to broadband in the case of electronic regeneration).

Some key issues have to be considered when comparing such Signal Regeneration approaches. The first is that today's and future optical transmission systems or/and networks are WDM networks.

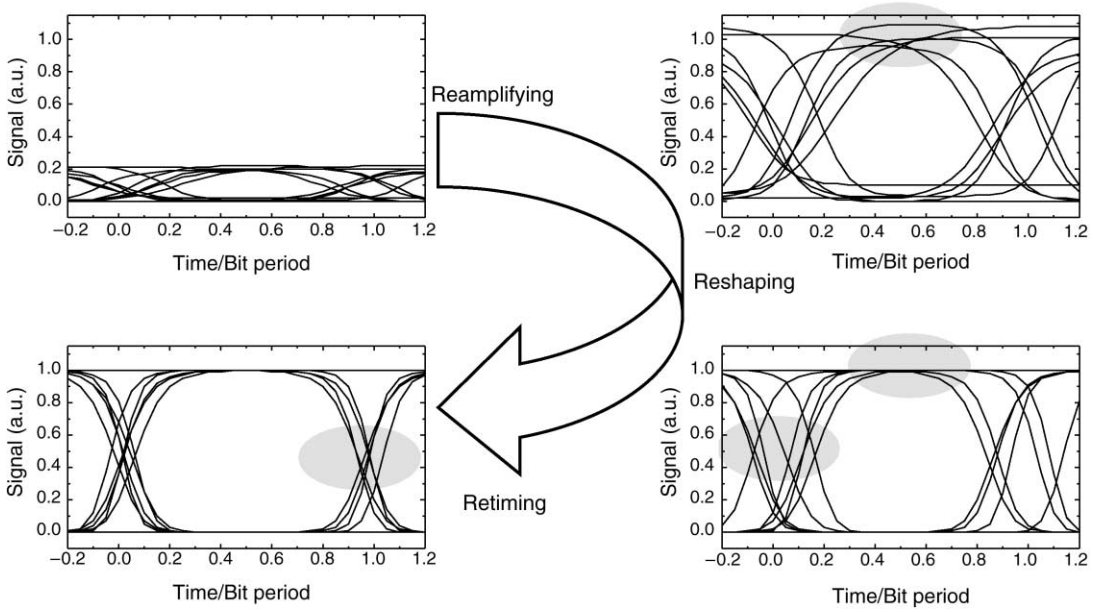


Figure 1 Principle of 3R regeneration, as applied to NRZ signals. (a) Re-Amplifying; (b) Re-Shaping; (c) Re-Timing. NB: Such eye diagrams can be either optical or electrical eye diagrams.

Under this condition, the WDM compatibility – or the fact that any Regeneration solution can simultaneously process several WDM channels – represents a key advantage. The maturity of the technology – either purely optical or opto-electronic – also plays an important role in the potential (pre-)development of such solutions. But the main parameter that will decide the actual technology (and also technique) relies on the tradeoff between actual performance of the regeneration solutions and their costs (device and implementation), depending on the targeted applications (long-haul system, medium haul transport, wide area optical network, etc.).

In this article, we review the current alternatives for all-optical Signal Regeneration, considering both theoretical and experimental performance and practical implementation issues. Key advantages and possible drawbacks of each solutions are discussed, to sketch the picture in this field. However, first we must focus on some generalities about Signal Regeneration and the way to define (and qualify) such regenerator performance. In a second part, we will detail the currently-investigated optical solutions for Signal Regeneration with a specific highlight for semiconductor-based solutions using either semiconductor optical amplifiers (SOA) technology or newly-developed saturable absorbers. Optical Regeneration techniques based on synchronous modulation will also be discussed in a third section. The conclusion will summarize the key features of each solution, so as to underline the demanding challenge optical components are facing in this application.

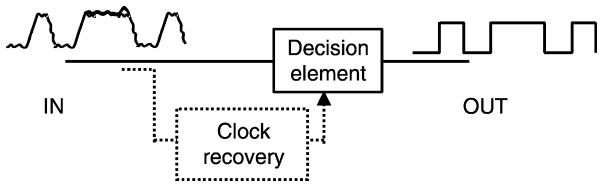


Figure 2 Generic structure of Signal 2R/3R Regenerator based on Decision Element (2R) and Decision Element and Clock Recovery (3R).

Generalities on Signal Regeneration

Principles

In the general case, Signal Regeneration is performed using a decision element exhibiting a nonlinear transfer function. Provided with a threshold level and when associated with an amplifier, such an element then performs the actual Re-shaping of the incoming data (either in the electrical or optical domain) and completes a 2R Signal Regenerator. Figure 2 shows the generic structure of such a Signal Regenerator in the general case as applied to non return to zero (NRZ) data. A clock recovery block can be added (dotted lines) to provide the decision element with time reference and hence perform the third R (Re-timing) of full Signal 3R Regeneration. At this point, it should be mentioned that the decision

element can operate either on electrical signals (standard electrical DFF) provided that optical → electrical and electrical → optical signal conversion stages are added or directed onto optical signals using the different techniques described below. The clock signal can be of an electrical nature, as for electrical decision element in O/E regenerator – or either an electrical or a purely optical signal in all-optical regenerators.

Prior to reviewing and describing the various current technology alternatives for such Optical Signal Regeneration, the issue of the actual characterization of regenerator performance needs to be explained and clarified. As previously mentioned, the core element of any Signal Regenerator is the decision element showing a nonlinear transfer function that can be of varying steepness. As will be seen in Figure 3, the actual regenerative performance of the regenerator will indeed depend upon the degree of nonlinearity of the decision element transfer function.

Figure 3 shows the principle of operation of a regenerator incorporating a decision element with two steepnesses of the nonlinear transfer function. In any case, the '1' and '0' symbols amplitude probability densities (PD) are squeezed after passing through the decision element. However, depending upon the addition of a clock reference for triggering the decision element, the symbol arrival time PD will be also squeezed (clocked decision = 3R regeneration) or enlarged (no clock reference = 2R regeneration) resulting in conversion of amplitude fluctuations to time position fluctuations.

As for system performance – expressed through BER – regenerative capabilities of any regenerator simultaneously depend upon both the output amplitude and arrival time PD of the '1' and '0' symbols. In the unusual case of 2R regeneration (no clocked decision), a tradeoff has then to be derived, considering both the reduction of amplitude PD and the enlarged arrival time PD induced by the regenerator, to ensure sufficient signal improvement. In Figure 3a, we consider a step function for the transfer function of the decision element. In this case, amplitude PD are squeezed to Dirac PD after the decision element, and depending upon addition or not of a clock reference, arrival time PD is reduced (3R) or dramatically enlarged (2R). In Figure 3b, the decision element exhibits a moderately nonlinear transfer function. This results in an asymmetric and less-pronounced squeezing of the amplitude PD compared to the previous case, but in turn results in a significantly less enlarged arrival time PD when no clock reference is added (2R regeneration). Comparison of these two decision element of different nonlinear transfer function indicates that for 3R regeneration applications, the more nonlinear the transfer function of the decision element the better performance, the ideal case being the step function. In the case of 2R regeneration applications, a tradeoff between the actual reduction of the amplitude PD and enlargement of timing PD is to be derived and clearly depends upon the actual shape of the nonlinear transfer function of the decision element.

Qualification of Signal Regenerator Performance

To further illustrate the impact of the actual shape of the nonlinear transfer function of the decision

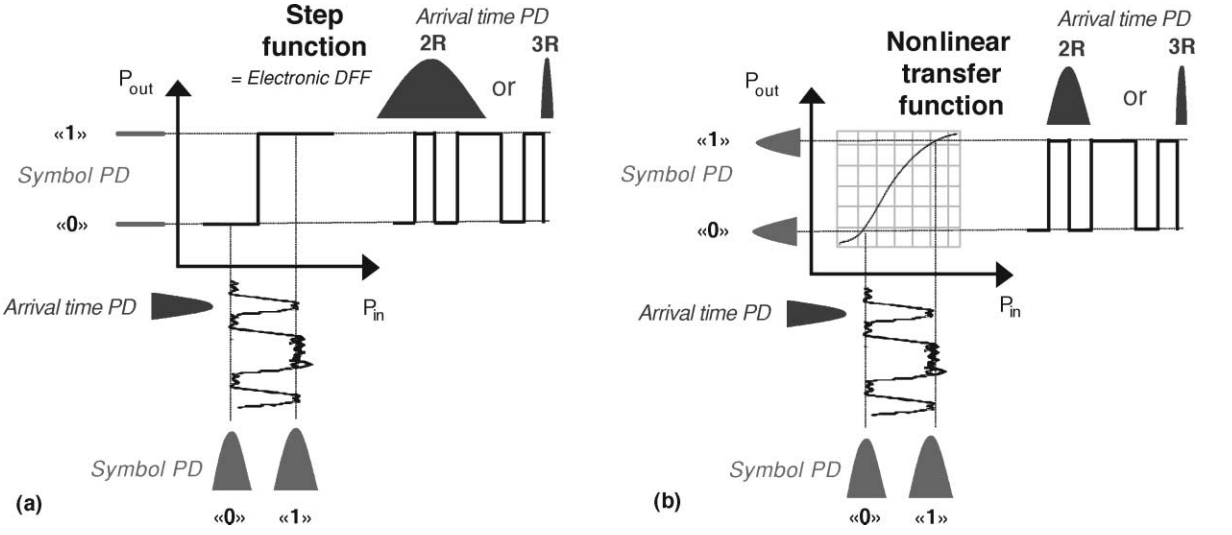


Figure 3 Signal Regeneration process using Nonlinear Gates. (a) Step transfer function (= Electronic DFF); (b) 'moderately' nonlinear transfer function. As an illustration of the Regenerator operation '1' and '0' symbols amplitude probability density (PD) and arrival time probability density (PD) are shown in light gray and dark gray, respectively.

element in 3R application, the theoretical evolution of BER with number of concatenated regenerators have been plotted for regenerators having different nonlinear responses. Figure 4 shows the numerically calculated evolution of the BER of a 10 Gbit/s NRZ signal with fixed optical signal-to-noise ratio (OSNR) at the input of the 3R regenerator, as a function of the cascaded regenerator incorporating nonlinear gates with nonlinear transfer function of different depths. From Figure 4, can be seen different behaviors depending on the nonlinear function shape. As previously stated, the best regeneration performance is obtained with an ideal step function (case a), which is actually the case for O/E regenerator using electronic decision flip-flop (DFF). In that case, BER linearly increase (i.e., more errors) in the cascade. Conversely, when nonlinearity is reduced (cases (b) and (c)), both BER and noise accumulate, until the concatenation of nonlinear functions reach some steady-state pattern, from which BER linearly increases. Concatenation of nonlinear devices thus magnifies shape differences in their nonlinear response, and hence their regenerative capabilities.

Moreover, as can be seen in Figure 4, all curves standing for different regeneration efficiencies pass through a common point defined after the first device. This clearly indicates that it is not possible to qualify the regenerative capability of any regenerator when considering the output signal after only one regenerator. Indeed, the BER is the same for either a 3R regenerator or a mere amplifier if only measured after a single element. This originates from the initial overlap between noise distributions associated with marks and spaces, that cannot be suppressed but only minimized by a single decision element through threshold adjustment.

As a general result, the actual characterization of the regenerative performance of any regenerator should *in fine* be conducted considering a cascade of regenerators. In practice this can easily be done with the experimental implementation of the regenerator under study in a recirculating loop. Moreover, such an investigation tool will also enable access to the regenerator performance with respect to the transmission capabilities of the regenerated signal, which should not be overlooked.

Let us now consider the physical implementation of such all-optical Signal Regenerators, along with the key features offered by the different alternatives.

All-Optical 2R/3R Regeneration Using Optical Nonlinear Gates

Prior to describing the different solutions for all-optical in-line Optical Signal Regeneration, it should be mentioned that since the polarization states of the optical data signals cannot be preserved during propagation, it is required that the regenerator exhibits an extremely low polarization sensitivity. This clearly translates to a careful optimization of all the different optical components making up the 2R/3R regenerator. It should be noted that this also applies to the O/E solutions but is of limited impact, since only the photodiode has to be polarization insensitive.

Figure 5 illustrates the generic principle of operation of an all-optical 3R Regenerator using optical nonlinear gates. Contrary to what occurs in O/E, regenerator where the extracted clock signal drives the electrical decision element, the incoming and distorted optical data signal triggers the nonlinear gate, hence generating a switching window which is

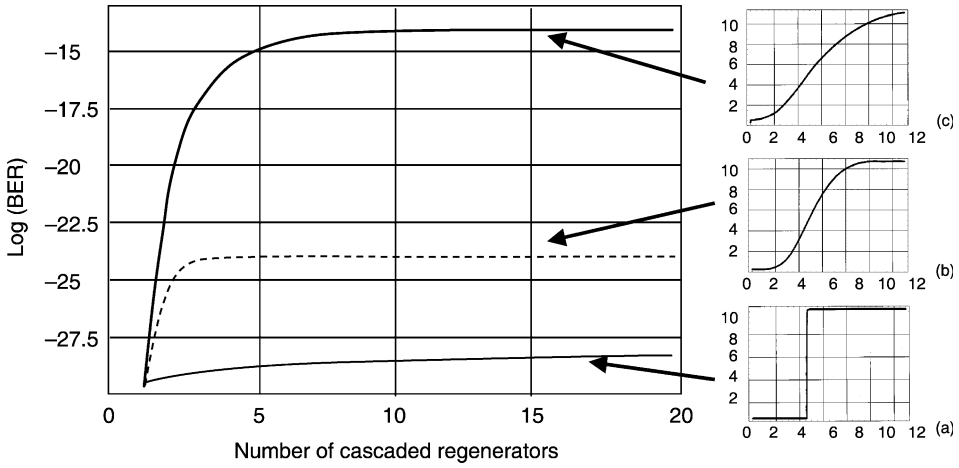


Figure 4 Evolution of the BER with concatenated regenerators for nonlinear gates with nonlinear transfer function of decreasing depths from case (a)–(c) (step function).

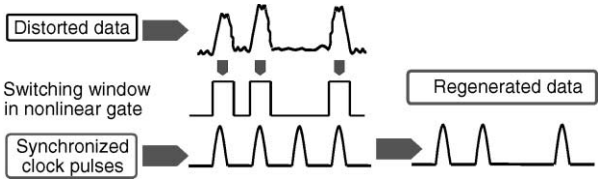


Figure 5 Principle of operation of an all-Optical Signal 3R Regenerator using nonlinear gates.

applied to a newly generated optical clock signal so as to reproduce the initial data stream on the new optical carrier.

In the case of 2R Optical Signal Regeneration, a continuous wave (CW) signal is substituted for the synchronized optical clock pulses. As previously mentioned, the actual regeneration performance of the 2R/3R devices will mainly depend upon the nonlinearity of the transfer function of the decision element but in 3R applications the quality of the optical clock pulses has also to be considered. In the following, we describe current solutions to explain the two main building blocks of all-optical Signal Regenerators: the decision element (i.e., nonlinear gate) and the clock recovery (CR) elements.

Optical Decision Element

In the physical domain, optical decision elements with ideal step response – as for electrical DFF – do not exist. Different nonlinear optical transfer functions, approaching more or less the ideal case, can be realized in various media such as fiber, SOA, electro-absorption modulators (EAM), and lasers. Generally, as described below, the actual response (hence the regenerative properties of the device) of such optical gates directly depends upon the incoming signal instantaneous power. Under these conditions, it appears essential to add an adaptation stage so as to reduce intensity fluctuations (as caused by propagation or crossing routing/switching node) and provide the decision element with fixed power conditions. In practice, this results in the addition of a control circuitry (either optical or electrical) in the Re-amplification block, whose complexity directly depends on actual system environment (ultra-fast power equalization for packet-switching applications and compensation of slow power fluctuations in transmission applications).

As previously described the decision gate performs Re-shaping (and Re-timing when clock pulses are added) of the incoming distorted optical signal, and represent the regenerator's core element. Ideally, it should also act as a transmitter with characteristics ensuring the actual propagation of the regenerated data stream. In that respect, the chirp possibly induced by the optical decision gate onto the regenerated signal – and the initial quality of the optical clock pulses in 3R applications – should be carefully considered (ideally by means of loop transmission) as to adequately match line transmission requirements.

Different solutions for the actual realization of the optical decision element have been proposed and extensively investigated using, for example, cross gain modulation in semiconductor optical amplifier (SOA) devices but the most promising and flexible devices probably are interferometers, for which, descriptions of the generic principle of operation follows. Consider a CW signal (probe) at λ_2 wavelength injected into an optical interferometer, in which one arm incorporates a nonlinear medium in which an input signal carried by λ_1 wavelength (command) is, in turn, injected. Such a signal, at λ_1 wavelength, induces a phase shift through cross-phase modulation (XPM) in this arm of the interferometer, the amount depending upon power level $P_{\text{in},\lambda 1}$. In turn, such phase modulation (PM) induces amplitude modulation (AM) on the signal at λ_2 wavelength when recombined at the output of the interferometer and translates the information carried by wavelength λ_1 onto λ_2. Under these conditions, such optical gates clearly act as wavelength converters (it should be mentioned that Wavelength Conversion is not necessarily equivalent to Regeneration; i.e., a linear transfer function performs suitable Wavelength Conversion but by no means Signal Regeneration).

Optical interferometers can be classified according to the nature of the nonlinearity exploited to achieve a π phase shift. In the case of fiber-based devices, such as the nonlinear optical loop mirror (NOLM), the phase shift is induced through the Kerr effect in an optical fiber. The key advantage of fiber-based devices such as NOLM lies in the near-instantaneous (fs) response of the Kerr nonlinearity, making them very attractive for ultra-high bit-rate operation ($\geq$160 Gbit/s). Polarization-insensitive NOLMs have been realized, although with the same drawbacks concerning integrability. With recent developments in highly nonlinear (HNL) fibers, however, the required NOLM fiber length could be significantly reduced, hence dramatically reducing environmental instability.

A second type of device is the integrated SOA-based Mach-Zehnder interferometers (MZI). In MZIs, the phase shift is due to the effect of photo-induced carrier depletion in the gain saturation regime of one of the SOAs. The control and probe can be launched in counter- or co-directional ways. In the first case, no optical filter is required at the output

of the device for rejecting the signal at λ_1 wavelength but operation of the MZI is limited by its speed. At this point, one should mention that the photo-induced modulation effects in SOAs are intrinsically limited in speed by the gain recovery time, which is a function of the carrier lifetime and the injection current. An approach referred to as differential operation mode (DOM) and illustrated on Figure 6, which takes advantage of the MZI's interferometric properties, makes it possible to artificially increase the operation speed of such 'slow' devices up to 40 Gbit/s.

As discussed earlier, the nonlinear response is a key parameter for regeneration efficiency. Combining two interferometers is a straightforward means to improve the nonlinearity of the decision element transfer function, and hence regeneration efficiency. This approach was validated at 40 Gbit/s using a cascade of two SOA-MZI, (see Figure 7 (**left**)). Such a scheme offers the advantage of restoring data polarity and wavelength, hence making the regenerator inherently transparent. Finally, the second conversion stage can be used as an adaptation interface to the transmission link achieved through chirp tuning in this second device.

Such an Optical 3R Regenerator was upgraded to 40 Gbit/s, using DOM in both SOA-MZIs with validation in a 40 Gbit/s loop RZ transmission. The 40 Gbit/s eye diagram monitored at the regenerator output after 1, 10, and 100 circulations are shown in Figure 7 (**right**) and remain unaltered by distance. With this all-optical regenerator structure, the minimum OSNR tolerated by the regenerator (1 dB sensitivity penalty at 10^{-10} BER) was found to be as low as 25 dB/0.1 nm. Such results clearly illustrate the high performance of this SOA-based regenerator structure for 40 Gbit/s optical data signals.

Such a complex mode of operation for addressing 40 Git/s bit-rates will probably be discarded when we consider the recent demonstration of standard-mode wavelength conversion at 40 Gbit/s, which uses a newly-designed active-passive SOA-MZI incorporating evanescent-coupling SOAs. The device architecture is flexible in the number of SOAs, thus enabling easier operation optimization and reduced power consumption, leading to simplified architectures and operation for 40 Gbit/s optical 3R regeneration.

Based on the same concept of wavelength conversion for Optical 3R Regeneration, it should be noted many devices have been proposed and experimentally validated as wavelength converters at rates up to 84 Gbit/s, but with cascadability issues still to be demonstrated to assess their actual regenerative properties.

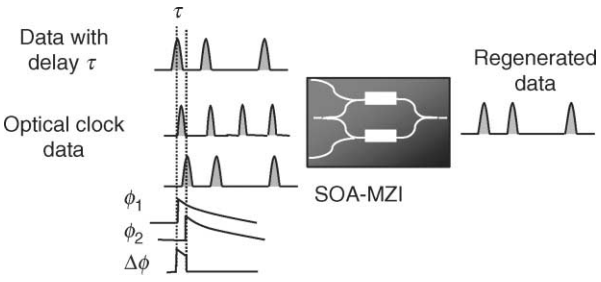


Figure 6 Schematic and principle of operation of SOA-MZI in differential mode.

Optical Clock Recovery (CR)

Next to the decision, the CR is a second key function in 3R regenerators. One possible approach for CR uses electronics while another only uses optics. The former goes with OE conversion by means of a photodiode and subsequent EO conversion through a modulator. This conversion becomes more complex and power-hungry as the data-rate increases. It is clear that the maturity of electronics gives a current advantage to this approach. But considering the pros and cons of electronic CR for cost-effective implementation, the all-optical approach seems more promising, since full regenerator integration is potentially possible with reduced power consumption. In this view, we shall focus here on the optical approach and more specifically on the self-pulsating effect in three-sections distributed feedback (DFB) lasers or more recently in distributed Bragg reflector

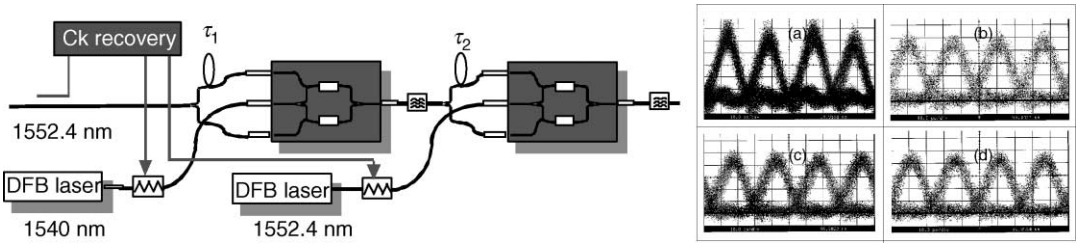


Figure 7 Optimized structure of a 40 Gbit/s SOA-based 3R regenerator. 40 Gbit/s eye diagram evolution: (a) B-to-B; (b) 1 lap; (c) 10 laps; (d) 100 laps.

(DBR) lasers. Recent experimental results illustrate the potentials of such devices for high bit rates (up to 160 Gbit/s), broad dynamic range, broad frequency tuning, polarization insensitivity, and relatively short locking times (1 ns). This last feature makes these devices good candidates for operation in asynchronous-packet regimes.

Optical Regeneration by Saturable Absorbers

We next consider saturable absorbers (SA) as nonlinear elements for optical regeneration. Figure 8 (**left**) shows a typical SA transfer function and illustrates the principle of operation. When illuminated with an optical signal with peak power below some threshold (P_{sat}), the photonic absorption of the SA is high and the device is opaque to the signal (low transmittance). Above P_{sat}, the SA transmittance rapidly increases and asymptotically saturates to transparency (passive loss being overlooked). Such a nonlinear transfer function only applies to 2R optical regeneration.

Different technologies for implementing SAs are available, but the most promising approach uses semiconductors. In this case, SA relies upon the control of carrier dynamics through the material's recombination centers. Parameters such as on–off contrast (ratio of transmittance at high and low incident powers), recovery time (1/e) and saturation energy, are key to device optimization. In the following, we consider a newly-developed ion-irradiated MQW-based device incorporated in a micro-cavity and shown on Figure 8 (**right**). The device operates as a reflection-mode vertical cavity, providing both a high on/off extinction ratio by canceling reflection at low intensity and a low saturation energy of 2 pJ. It is also intrinsically polarization-insensitive. Heavy ion-irradiation of the SA ensures recovery times (at 1/*e*) shorter than 5 ps (hence compatible with bit-rate above 40 Gbit/s), while maintaining a dynamic contrast in excess of 2.5 dB at 40 GHz repetition rate.

The regenerative properties of SA make it possible to reduce cumulated amplified spontaneous emission (ASE) in the '0' bits, resulting in a higher contrast between mark and space, hence increasing system performance. Yet SAs do not suppress intensity noise in the marks, which makes the regenerator incomplete. A solution for this noise suppression is optical filtering with nonlinear (soliton) pulses. The principle is as follows. In absence of chirp, the soliton temporal width scales in the same way as the reciprocal of its spectral width (Fourier-transform limit) times its intensity (fundamental soliton relation). Thus, an increase in pulse intensity corresponds to both time narrowing and spectral broadening. Conversely, a decrease in pulse intensity corresponds to time broadening and spectral narrowing. Thus, the filter causes higher loss when intensity increases, and lower loss when intensity decreases. The filter thus acts as an automatic power control (APC) in feed-forward mode, which causes power stabilization. The resulting 2R regenerator (composed by the SA and the optical filter) is fully passive, which is of high interest for submarine systems where the power consumption must be minimal, but it does not include any control in the time domain (no Re-timing).

System demonstrations of such passive SA-based Optical Regeneration have been reported with a 20 Gbit/s single-channel loop experiment. Implementation of the SA-based 2R Regenerator with 160 km-loop periodicity made it possible to double the error-free distance (Q = 15.6 dB or 10^{-9} BER) of a 20 Gbit/s RZ signal. So as to extend the capability of passive 2R regeneration to 40 Gbit/s systems, an improved configuration was derived from numerical optimization and experimentally demonstrated in a 40 Gbit/s WDM-like, dispersion-managed loop transmission, showing more than a fourfold increase in the WDM transmission distance at 10^{-4} BER (1650 km without the SA-based regenerator and 7600 km when implementing the 2R regenerator with 240 km periodicity).

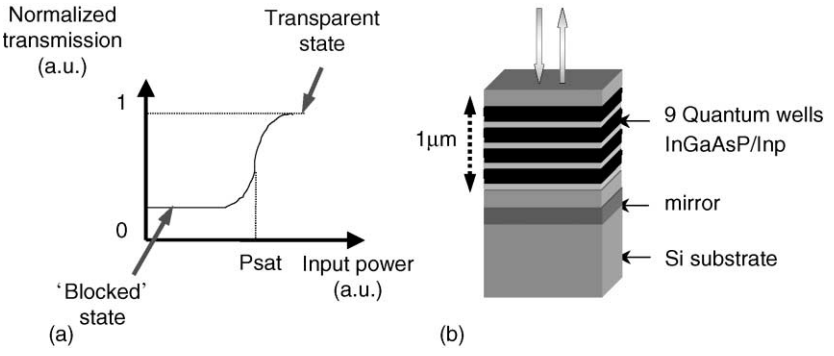


Figure 8 (a) Saturable Absorber (SA) ideal transfer function. (b) Structure of Multi-Quantum Well SA.

Such a result illustrates the potential high interest of such passive optical 2R regeneration in long-haul transmission (typically in noise-limited systems) since the implementation of SA increases the system's robustness to OSNR degradation without any extra power consumption. Reducing both saturation energy and insertion loss along with increasing dynamic contrast represent key future device improvements. Regeneration of WDM signals from the same device, such as one SA chip with multiple fibers implemented between Mux/DMux stages, should also be thoroughly investigated. In this respect, SA wavelength selectivity in quantum dots could possibly be advantageously exploited.

Synchronous Modulation Technique

All-optical 3R regeneration can be also achieved through in-line synchronous modulation (SM) associated with narrowband filtering (NF). Figure 9 shows the basic layout of such an Optical 3R Regenerator. It is composed of an optical filter followed by an Intensity and Phase Modulator (IM/PM) driven by a recovered clock. Periodic insertion of SM-based modulators along the transmission link provides efficient jitter reduction and asymptotically controls ASE noise level, resulting in virtually unlimited transmission distances. Re-shaping and Re-timing provided by IM/PM intrinsically requires nonlinear (soliton) propagation in the trunk fiber following the SM block. Therefore, one can refer to the approach as distributed optical regeneration. This contrasts with lumped regeneration, where 3R is completed within the regenerator (see above with Optical Regeneration using nonlinear gates), and is independent of line transmission characteristics.

However, when using a new approach referred to 'black box' optical regeneration (BBOR), it is possible to make the SM regeneration function and transmission work independently in such a way that any type of RZ signals (soliton or non-soliton) can be transmitted through the system. The BBOR technique includes an adaptation stage for incoming RZ pulses in the SM-based regenerator, which ensures high regeneration efficiency regardless of RZ signal format (linear RZ, DM-soliton, C-RZ, etc.). This is achieved using a local and periodic soliton conversion of RZ pulses by means of launching an adequate power into some length of fiber with anomalous dispersion. The actual experimental demonstration of the BBOR approach and its superiority over the 'classical' SM-based scheme for DM transmission was experimentally investigated in 40 Gbit/s DM loop transmission. Under these conditions, one can then independently exploit dispersion management (DM) techniques for increasing spectral efficiency in long-haul transmission, while ensuring high transmission quality through BBOR.

One of the key properties of the SM-based all-optical Regeneration technique relies on its WDM compatibility. The first (Figure 10, **left**) and straightforward solution to apply Signal Regeneration to WDM channels amounts to allocating a regenerator to each WDM channel. The second consists in sharing a single modulator, thus processing the WDM channels at once in serial fashion. This approach requires WDM synchronicity, meaning that all bits be synchronous with the modulation, that can be achieved either by use of appropriate time-delay lines located within a DMux/Mux apparatus (Figure 10, **upper right**), or by making the WDM channels inherently time-coincident at specific regenerator locations (Figure 10, **bottom right**). Clearly, the serial regeneration scheme is far simpler and cost-effective than the parallel version; however, optimized re-synchronization schemes still remain to be

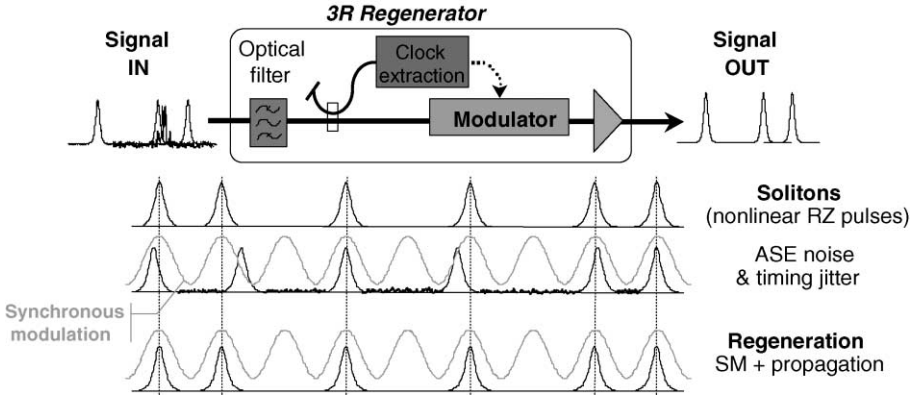


Figure 9 Basic layout of the all-optical Regenerator by Synchronous Modulation and Narrowband Filtering and illustration of the principle of operation.

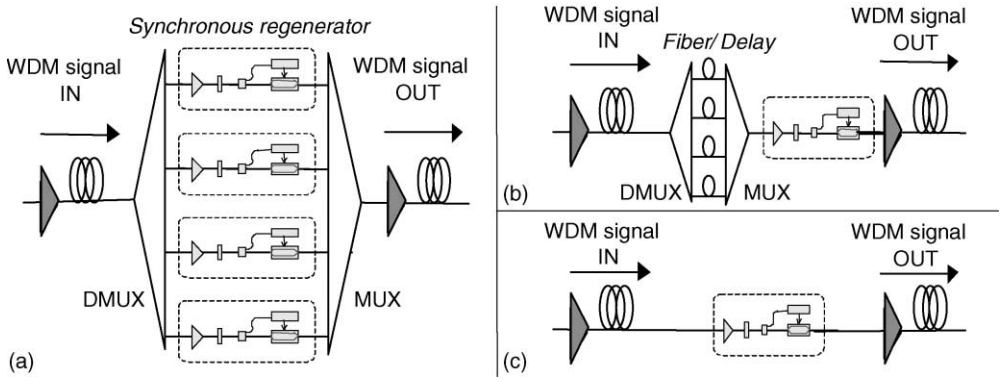


Figure 10 Basic implementation schemes for WDM all-Optical Regeneration. (a) parallel asynchronous; (b) serial re-synchronized; (c) serial self-synchronized.

developed for realistic applications. Experimental demonstration of this concept was assessed by means of a 4×40 Gbit/s dispersion-managed transmission over 10 000 km (BER $< 5.10^{-8}$) in which a single modulator was used for the simultaneous regeneration of the 4 WDM channels.

Considering next all-optical regeneration schemes with ultra-high speed potential, a compact and loss-free 40 Gbit/s Synchronous Modulator, based on optically-controlled SOA-MZI, was proposed and loop-demonstrated at 40 Gbit/s with an error-free transmission distance in excess of 10 000 km. Moreover, potential ultra-high operation of this improved BBOR scheme was recently experimentally demonstrated by means of a 80 GHz clock conversion with appropriate characteristics through the SOA-MZI. One should finally mention all fiber-based devices such as NOLM and NALM for addressing ultra-high speed SM-based Optical Regeneration, although no successful experimental demonstrations have been reported so far in this field.

Conclusion

In summary, optical solutions for Signal Regeneration present many key advantages. These are the only advantages to date to possibly ensure WDM compatibility of the regeneration function (mostly 2R related). Such optical devices clearly exhibits the best 2R regeneration performance (wrt to O/E solutions) as a result of the moderately nonlinear transfer function (which in turn can be considered as a drawback in 3R applications), but the optimum configuration is still to be clearly derived and identified depending upon the system application. Optics also allow to foresee and possibly target ultrafast applications above 40G, for signal regeneration if needed. Among the current drawbacks, one should mention the relative lack of wavelength/formats flexibility of these solutions (compared to O/E solutions). It is complex or difficult to restore the input wavelength or address any C-band wavelength at the output of the device or to successfully regenerate modulation formats other than RZ. In that respect, investigations should be conducted to derive new optical solutions capable of processing more advanced modulation formats at 40G. Finally, the fact that the nonlinear transfer function of the optical is in general triggered by the input signal instantaneous power also turns out to be a drawback since it requires control circuitry. The issue of cost (footprint, power consumption, etc.) of these solutions, compared to O/E ones, is still open. In this respect, purely optical solutions incorporating all-optical clock recovery, the performance of which is still to be technically assessed, are of high interest for reducing costs. Complete integration of an all-optical 2R/3R regenerator or such parallel regenerators onto a single semiconductor chip should also contribute to make all-optical solutions cost-attractive even though acceptable performance of such fully integrated devices is still to be demonstrated.

From today's status concerning the two alternative approaches for in-line regeneration (O/E or all-optical), it is safe to say that the choice between either solution will be primarily dictated by both engineering and economical considerations. It will result from a tradeoff between overall system performance, system complexity and reliability, availability, time-to-market, and rapid returns from the technology investment.

See also

Interferometry: Overview. **Optical Amplifiers:** Semiconductor Optical Amplifiers. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Bigo S, Leclerc O and Desurvire E (1997) All-optical fiber signal processing for solitons communications. *IEEE Journal of Selected Topics on Quantum Electronics* 3(5): 1208–1223.

Dagens B, Labrousse A, Fabre S, *et al.* (2002) New modular SOA-based active–passive integrated Mach–Zehnder interferometer and first standard mode 40 Gb/s all-optical wavelength conversion on the C-band. paper PD 3.1, Proceedings of ECOCOL 02, Copenhague.

Durhuus T, Mikkelsen B, Joergensen C, Danielsen SL and Stubkjaer KE (1996) All-optical wavelength conversion by semiconductor optical amplifiers. *Journal of Lightwave Technology* 14(6): 942–954.

Lavigne B, Guerber P, Brindel P, Balmefrezol E and Dagens B (2001) Cascade of 100 optical 3R regenerators at 40 Gbit/s based on all-active Mach–Zehnder interferometers. paper We.F.2.6, Proceedings of ECOC 2001, Amsterdam.

Leclerc O, Lavigne B, Chiaroni D and Desurvire E (2002) All-optical regeneration: Principles and WDM implementation. In: Kaminow IP and Li T (eds) *Optical Fiber Technologies IVA*, chap. 15, pp. 732–784. San Diego: Academic Press.

Leuthold J, Raybon G, Su Y, *et al.* (2002) 40 Gbit/s transmission and cascaded all-optical wavelength conversion over 1,000,000 km. *Electronic Letters* 38(16): 890–892.

Öhlén P and Berglind E (1997) Noise accumulation and BER estimates in concatenated nonlinear optoelectronic repeaters. *IEEE Photon Technol. Letters* 9(7): 1011–1013.

Otani T, Suzuki M and Yamamoto S (2001) 40 Gbit/s optical 3R regenerator for all-optical networks. We.F.2.1, Proceedings of ECOC 2001, Amsterdam.

Oudar J-L, Aubin G, Mangeney J, *et al.* (2004) Ultra-fast quantum-well saturable absorber devices and their application to all-optical regeneration of telecommunication optical signals. *Annales des Télécommunications* 58(11–12): 1667–1707.

Sartorius B, Mohrle M, Reichenbacher S, *et al.* (1997) Dispersive self-Q-switching in self-pulsating DFB lasers. *IEEE Journal of Quantum Electronics* 33(2): 211–218.

Ueno Y, Nakamura S, Sasaki J, *et al.* (2001) Ultrahigh-speed all-optical data regeneration and wavelength conversion for OTDM systems. Th.F.2.1, Proceedings of ECOC 2001, Amsterdam.

B

BABINET'S PRINCIPLE

B D Guenther, Duke University, Durham, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

In the development of the theory of diffraction, the diffraction field is due to a surface integral, the Fresnel integral, but no restrictions are imposed on the choice of the surface over which the integration must be performed. This fact leads to a very useful property called Babinet's principle, first stated by Jacques Babinet (1794–1872) in 1837 for scalar waves. We will discuss only the scalar Babinet's principle; discussion of the rigorous vector formulation can be found in a number of books on electromagnetic theory.

To introduce Babinet's principle, we label a plane, separating the source, S, and observation point, P, as Σ. If no obstructions are present, a surface integration over Σ yields the light distribution at P. If we place an opaque obstruction in this plane with a clear aperture, Σ_1, then the field at P is given by integrating over only Σ_1; contributions from Σ, outside of Σ_1, are zero since the obstruction is opaque.

We may define an aperture, Σ_2, as complementary to Σ_1 if the obstruction is constructed by replacing the transparent regions of Σ, i.e., Σ_1, by opaque surfaces and the opaque regions of Σ by clear apertures. Figure 1 shows two complementary obstructions, where the shaded region indicates an opaque surface.

The surface integral over Σ generates the field, E, in the absence of any obstruction. If obstruction Σ_1 is present then the diffraction field is E_1, obtained by integrating over Σ_1. According to Babinet's principle, the diffraction field due to obstruction Σ_2, must be

$$\tilde{E}_2 = \tilde{E} - \tilde{E}_1$$

We will look at examples of the application of the principle for both Fraunhofer and Fresnel diffraction.

Fraunhofer Diffraction

The electric field due to Fraunhofer diffraction is given by

$$E_{\mathrm{P}}(\omega_x, \omega_y) = \frac{\mathrm{i}\alpha \mathrm{e}^{-\mathrm{i}kR_0}}{\lambda R_0} \iint_{\Sigma} f(x, y) \mathrm{e}^{-\mathrm{i}(\omega_x x + \omega_y y)} \, \mathrm{d}x \, \mathrm{d}y$$

where α is the amplitude of the incident plane wave and R_0 is the distance from the obstruction to the observation point. We have defined the spatial frequencies, ω_x and ω_y by the equations

$$\omega_x = kL = -\frac{2\pi\xi}{\lambda R_0} \qquad \omega_y = kM = -\frac{2\pi\eta}{\lambda R_0}$$

This surface integral is an approximation of the Hygens–Fresnel integral and can be identified as a Fourier transform of the aperture transmission function, $f(x, y)$ in two dimensions. For our discussion we will consider only one dimension and ignore the parameters in front of the integral.

Assume that $a(x)$ is the amplitude transmission of an aperture in an obstruction, $b(x)$ is the amplitude transmission function of the complementary obstruction and g is the amplitude of the wave when no

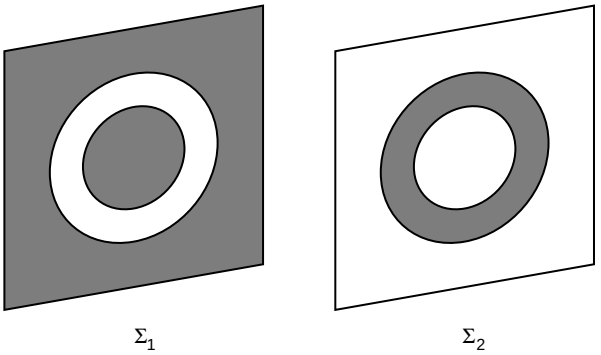


Figure 1 An aperture Σ_1 shown by the unshaded region and its complement Σ_2. Reproduced with permission from Guenther R (1990) *Modern Optics*. New York: Wiley. Copyright © 1990 John Wiley & Sons.

obstruction is present. The Fourier transforms of the one-dimensional amplitude transmission functions of the two apertures are equal to the Fraunhofer diffraction patterns that will be generated by the two apertures

$$A(k) = \int_{-\infty}^{\infty} a(x)\mathrm{e}^{-ikx}\,\mathrm{d}x$$

$$B(k) = \int_{-\infty}^{\infty} b(x)\mathrm{e}^{-ikx}\,\mathrm{d}x$$

With no aperture present, the far field amplitude is

$$G(k) = g\int_{-\infty}^{\infty} \mathrm{e}^{-ikx}\,\mathrm{d}x$$

Babinet's principle states that

$$B(x) = G - A(x)$$

must be the Fraunhofer diffraction field of the complementary aperture. We may rewrite this equation for the diffraction field as

$$B(k) = G\delta(k) + A(k)\mathrm{e}^{i\pi} \qquad [1]$$

The first term on the right of the Fraunhofer diffraction field for the complementary obstruction [1] is located at the origin of the observation plane and is proportional to the amplitude of the unobstructed wave. The second term in the equation for the Fraunhofer diffraction pattern [1] is identical to the Fraunhofer diffraction pattern of the original aperture, except for a constant phase. Thus the Fraunhofer diffraction from the two complementary apertures are equal, except for a constant phase and the bias term at the origin. Physically this means that the diffraction intensity distributions of complementary apertures will be identical but their brightness will differ!

Fresnel Diffraction

We can calculate the Fresnel diffraction from an opaque disk by applying Babinet's principle to the diffraction pattern calculated for a circular aperture.

We assume that a circular aperture of radius a is illuminated by a point source a distance Z' from the aperture. We observe the transmitted wave at the point P, located a distance Z from the aperture (see Figure 2).

The Fresnel diffraction integral is

$$\tilde{E}_{\mathrm{P}} = \frac{i\alpha}{\lambda\rho D}\mathrm{e}^{-ikD} \times \int\int f(x,y)\mathrm{e}^{-\frac{ik}{2\rho}\left[(x-x_0)^2+(y-y_0)^2\right]}\,\mathrm{d}x\,\mathrm{d}y \qquad [2]$$

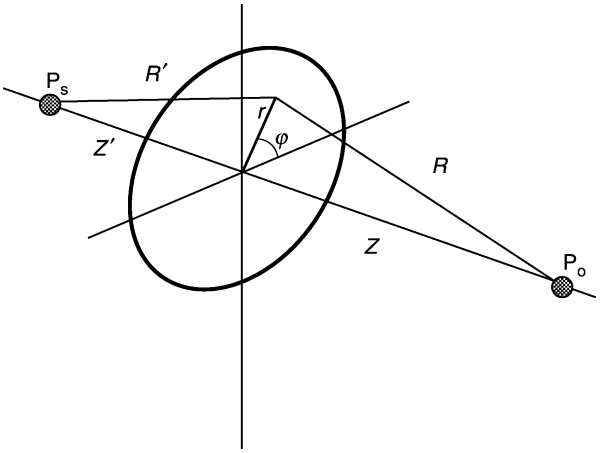


Figure 2 Geometry for the analysis of Fresnel diffraction of a circular aperture. Reproduced with permission from Guenther R (1990) *Modern Optics*. New York: Wiley. Copyright © 1990 John Wiley & Sons.

where the distance between the source and observation point is

$$D = Z + Z' + \frac{(\xi - x_s)^2 + (\eta - y_s)^2}{2(Z+Z')} \qquad \frac{1}{\rho} = \frac{1}{Z} + \frac{1}{Z'}$$

and the coordinates of the stationary point are

$$x_0 = \frac{Z'\xi + Zx_s}{Z+Z'} \qquad y_0 = \frac{Z'\eta + Zy_s}{Z+Z'}$$

The parameters of the integral are simplified by the geometry selected for Figure 2. Both the observation point and source are located on the axis of symmetry of the aperture which lies along the z-axis of a cylindrical coordinate system thus

$$x_0 = y_0 = \xi = \eta = 0 \qquad \text{and} \qquad D = Z' + Z$$

The approximation for the distance from the source and observation point to the aperture yields

$$R + R' \approx Z + Z' + \frac{r^2}{2\rho} = D + \frac{r^2}{2\rho}$$

where $r^2 = x^2 + y^2$· The cylindrical coordinate version of [2] is then

$$\tilde{E}_{\mathrm{P}} = \frac{i\alpha}{\lambda\rho D}\mathrm{e}^{-ikD}\int_0^a\int_0^{2\pi} f(r,\varphi)\mathrm{e}^{-i\frac{kr^2}{2\rho}}\,r\,\mathrm{d}r\,\mathrm{d}\phi \qquad [3]$$

We assume that the transmission function of the aperture is a constant, $f(r,\phi) = 1$, to make the integral as simple as possible. After performing the integration over the angle ϕ, [3] can be

rewritten as

$$\tilde{E}_{\mathrm{Pa}} = \frac{\mathrm{i}\alpha\pi}{D}\mathrm{e}^{-\mathrm{i}kD}\int_0^{\frac{a^2}{\rho\lambda}} \mathrm{e}^{-\mathrm{i}\pi r^2/\rho\lambda}\,\mathrm{d}\left(\frac{r^2}{\rho\lambda}\right) \quad [4]$$

Performing the integration of [4], results in the Fresnel diffraction amplitude

$$\tilde{E}_{\mathrm{Pa}} = \frac{\alpha}{D}\mathrm{e}^{-\mathrm{i}kD}\left[1 - \mathrm{e}^{-\mathrm{i}\pi a^2/\rho\lambda}\right] \quad [5]$$

The intensity of the diffraction field is

$$\begin{aligned} I_{\mathrm{Pa}} &= 2I_0\left(1 - \cos\frac{\pi a^2}{\rho\lambda}\right) \\ &= 4I_0\sin^2\frac{\pi a^2}{2\rho\lambda} \end{aligned} \quad [6]$$

[6] predicts a sinusoidal variation in intensity as the observation point moves toward the aperture.

Babinet's principle can now be used to evaluate Fresnel diffraction by an opaque disk, the same size as the circular aperture, i.e., of radius a. The field for the disk is obtained by subtracting the field, diffracted by a circular aperture, from the field of an unobstructed spherical wave, E_{P}:

$$\begin{aligned} \tilde{E}_{\mathrm{Pd}} &= \tilde{E}_{\mathrm{P}} - \tilde{E}_{\mathrm{Pa}} \\ \tilde{E}_{\mathrm{Pd}} &= \frac{\alpha}{D}\mathrm{e}^{-\mathrm{i}kD} - \tilde{E}_{\mathrm{Pa}} \\ &= \frac{\alpha}{D}\mathrm{e}^{-\mathrm{i}kD}\mathrm{e}^{-\mathrm{i}\pi a^2/\rho\lambda} \end{aligned} \quad [7]$$

The intensity is

$$I_{\mathrm{Pd}} = I_0$$

The result of the application of Babinet's principle is the conclusion that, at the center of the geometrical shadow cast by the disk, there is a bright spot with the same intensity as would exist if no disk were present. This is Poisson's spot. The intensity of this spot is independent of the choice of the observation point along the z-axis.

See also

Diffraction: Fraunhofer Diffraction; Fresnel Diffraction.

Further Reading

Abramowitz M and Stegun IA (1964) *Handbook of Mathematical Functions*. NBS Applied Mathematics Series-55 Washington, DC: National Bureau of Standards.

Born M and Wolf E (1959) *Principles of Optics*. New York: Pergamon Press.

Guenther R (1990) *Modern Optics*. New York: Wiley.

Haus HA (1984) *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall.

Hecht E (1998) *Optics*, 3rd edn. Reading, MA: Addison-Wesley.

Klein MV (1970) *Optics*. New York: Wiley.

Rossi B (1957) *Optics*. Reading, MA: Addison-Wesley.

Sommerfeld A (1964) *Optics*. New York: Academic Press.

C

CHAOS IN NONLINEAR OPTICS

R G Harrison and W Lu, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Instabilities in laser emission, notably in the form of spontaneous coherent pulsations, have been observed almost since the first demonstration of laser action. Indeed, the first laser operated in 1960 by Maiman produced noisy spiked output even under conditions of quasi-steady excitation and provided the early impetus for studies of such effects. Subsequent theoretical efforts towards understanding these phenomena occurred up to the 1980s at a modest level, due in part to the wide scope of alternative areas of fertile investigation provided by lasers during this period. However, since the 1990s, there has been a major resurgence of interest in this area. This has been due to profound mathematical discoveries over this period which have revolutionized our understanding of classical dynamical systems. It is now clear that many systems containing some form of nonlinearity are dynamically unstable and even chaotic and that such behavior is deterministic. Further, the discovery that chaos evolves through particular routes with well-defined scenarios and that such routes are universal, has stimulated experimentalists to search for physical systems that exhibit these properties. These phenomena have now been observed in areas as diverse as fluid flow, chemical reactions, population ecology, and superconductor devices. The laser, perhaps the most recent newcomer to the field, is a paradigm for such investigation, owing to its simplicity both in construction and in the mathematics that describe it, and already a wide range of these phenomena have been observed. Along with a proliferation of new observations, many of the instabilities known to occur in these systems are now being reinterpreted in lieu of our new insight in this area. From this, fascinating new concepts of control and synchronization of chaos have emerged and spawned new fields of applications, not least in secure communication. In this article we introduce the general principles of dynamical instabilities, chaos, and control in lasers. For in-depth study, the reader is referred to texts and review articles cited in the Further Reading section at the end of this article.

Laser Physics

Our conventional understanding of laser physics is concerned with how cooperative order evolves from randomness. This transition is explained by first considering the ensemble of lasing atoms in thermal equilibrium. Every atom executes a small motion, giving rise to an oscillating electric dipole, described by linear dynamics. For larger motions, the atomic dipoles interfere with each other's motion and beyond a particular threshold, the motion becomes 'co-operative' or ordered over a long range. The role of the electromagnetic field in the laser cavity in ordering the induced atomic dipoles, is collectively described as giving rise to a macroscopic polarization, the magnitude of which depends on the number of dipoles (excited atoms). The dynamical interplay between the cavity field amplitude (E) as one variable and the atomic material variables of polarization (P) and population (D) of excited atoms, all of which are in general complex quantities, provide a full description of lasing action. For the simplest laser, a single mode two-level system with a homogeneously broadened gain lasing at resonance, these reduce to just three real variables, the Maxwell–Bloch equations:

$$\dot{E} = -\kappa E + \kappa P \qquad [1]$$

$$\dot{P} = \gamma_{\perp} ED - \gamma_{\perp} P \qquad [2]$$

$$\dot{D} = \gamma_{II}(\lambda + 1) - \gamma_{II} D - \gamma_{II} \lambda EP \qquad [3]$$

where κ is the cavity decay rate, $\gamma_\perp$ is the decay rate of atomic polarization, γ_{II} is the decay rate of population inversion, and λ the pumping parameter.

Described in this way the laser acts as a single oscillator in much the same way as a mechanical nonlinear oscillator, where laser output represents damping of the oscillator and excitation of the atoms is the driving mechanism to sustain lasing (oscillation). For ordered or coherent emission, it is necessary for one or both of the material variables (which are responsible for further lasing emission) to respond sufficiently fast to ensure a phase correlation with the existing cavity field. This is readily obtained in many typical lasers with output mirrors of relatively high reflectivity, since the field amplitude within the laser cavity will then vary slowly compared to the fast material variables which may then be considered through their equilibrium values. This situation, commonly referred to as adiabatic elimination of fast variables, reduces the dynamics of lasing action to that of one (or two) variables, the field (and population), all others being forced to adapt constantly to the slowly varying field variable. Our familiar understanding of DC laser action presupposes such conditions to hold. However when, for example, the level of excitation of the atoms is increased to beyond a certain value, i.e., the second laser threshold, all three dynamical variables may have to be considered, which satisfies the minimum requirement of three degrees of freedom (variables) for a system to display chaos. So this simple laser is capable of chaotic motion, for which the emission is aperiodic in time and has a broadband spectrum. Prediction of such behavior was initially made by Haken in 1975 through establishing the mathematical equivalence of the Maxwell–Bloch equations to those derived earlier by Lorentz describing chaotic motion in fluids.

Nonlinear Dynamics and Chaos in Lasers

In general there is no general analytical approach for solving nonlinear systems such as the Maxwell–Bloch equations. Instead solutions are obtained by numerical means and analyzed through geometric methods originally developed by Poincaré as early as 1892. This involves the study of topological structures of the dynamical trajectories in phase space of the variables describing the system. If an initial condition of a dissipative nonlinear dynamical system, such as a laser, is allowed to evolve for a long time, the system, after all the transients have died out, will eventually approach a restricted region of the phase space called an attractor. A dynamical system can have more than one attractor, in which case different initial conditions lead to different types of long-time behavior. The simplest attractor in phase space is a fixed point which is a solution with just one set of values for its dynamical variables; the nonlinear system is attracted towards this point and stays there, giving a DC solution. For other control conditions the system may end up making a periodic motion; the attractor of this motion is called a limit cycle. However, when the operating conditions exceed a certain critical value, the periodic motion of the system breaks down into a more complex dynamical trajectory, which never repeats. This motion represents a third kind of attractor in phase space called a chaotic or strange attractor.

Figure 1 shows a sequence of trajectories of the Lorentz equations in the phase space of (x, y, z) on increasing the value of one of the control parameters; this corresponds to (E, P, D) in the Maxwell–Block equations on increasing the laser pump. For a control setting near zero, all trajectories approach stable equilibrium at the origin, the topological structure of the basin of attraction being hyperbolic about zero (Figure 1a). For the laser equations, this corresponds to operation below lasing threshold for which the magnitude of the control parameter (laser gain) is insufficient to produce lasing. As the control parameter is increased, the basin lifts at its zero point to create an unstable fixed point here but with now two additional fixed points located in the newly formed troughs on either side of the zero point, which is now a saddle point. This is illustrated in Figure 1b where the system eventually settles to one or other of the two new and symmetrical fixed points, depending on the initial conditions. This corresponds to DC or constant lasing as defined by the parameter values of one or other of these points which are indistinguishable. Pictorially, think of a conventional saddle shape comprising a hyperbolic and inverted hyperbolic form in mutually perpendicular planes and connected tangentially at the origin. With the curve of the inverted hyperbola turned up at its extremity and filling in the volume with similar profiles which allow the two hyperbolic curves to merge into a topological volume, one sees that a ball placed at the origin is constrained to move most readily down either side of the inverted hyperbolas into one or other of the troughs formed by this volume. For the Lorentz system, chaotic behavior occurs at larger values of the control parameter when all three fixed points become saddles. Since none of the equilibrium points are now attracting, the behavior of the system cannot be a steady motion. Although perhaps difficult to visualize topologically, it is then possible to find a region in this surface enclosing all three points and

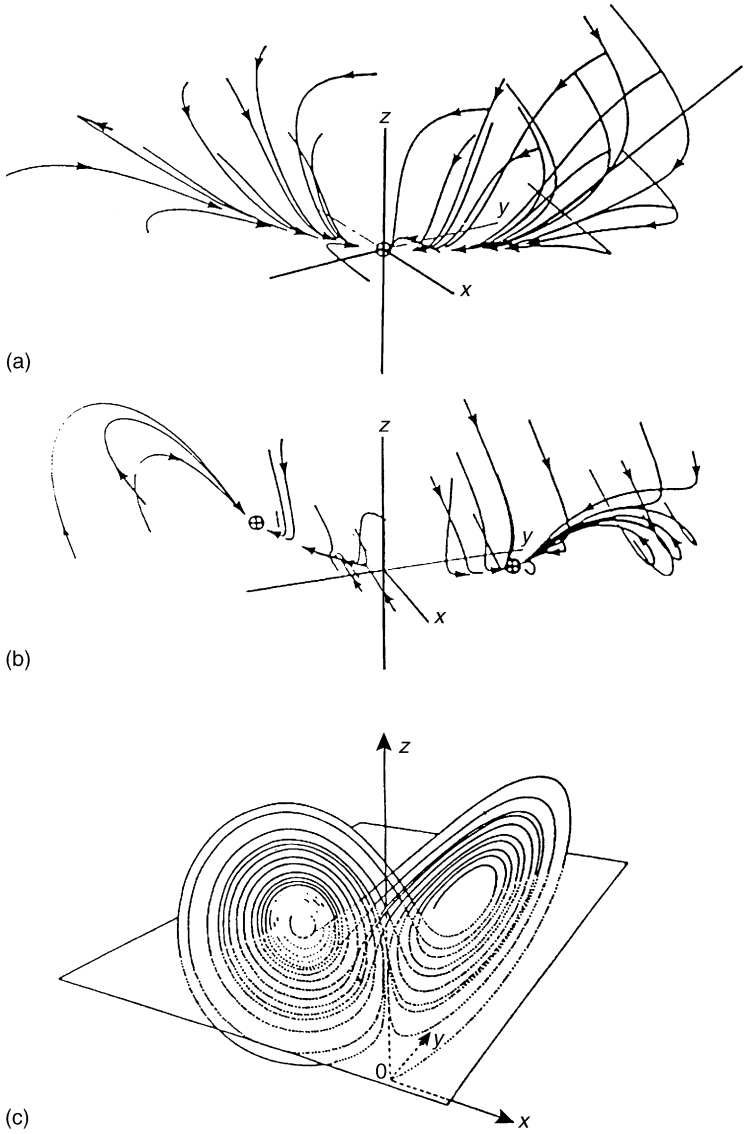


Figure 1 Trajectories in the three-dimensional phase space of the Lorentz attractor on increasing one of the control parameters. (Reproduced with permission from Thompson JMT and Stewart HB (1987) *Nonlinear Dynamics and Chaos*. New York: John Wiley.)

large enough so that no trajectories leave the region. Thus, all initial conditions outside the region evolve into the region and remain inside from then on. A corresponding chaotic trajectory is shown in Figure 1c. A point outwardly spirals from the proximity of one of the new saddle points until the motion brings it under the influence of the symmetrically placed saddle, the trajectory then being towards the center of this region from where outward spiraling again occurs. The spiraling out and switching over continues forever though the trajectory never intersects. In this case, arbitrarily close initial conditions lead to trajectories, which, after a sufficiently long time, diverge widely. Since a truly exact assignment of the initial condition is never possible, even numerically, a solution comprising several such trajectories therefore evolves and, as a consequence, long-term predictability is impossible. This is in marked contrast to that of the fixed point and limit cycle attractors, which settle down to the same solutions. A recording of one of these variables in time, say for a laser the output signal amplitude (in practice recorded as signal intensity, proportional to the square of the field amplitude), then gives oscillatory emission of increasing intensity with sudden discontinuities (resulting from flipping from one saddle to the other in Figure 1c) as expected.

While the Lorentz/Haken model is attractive for its relative simplicity, many practical lasers cannot be reduced to this description. Nevertheless, the



Figure 2 Lorentz-type chaos in the NH_3 laser (emitting at 81 μm optically pumped by a N_2O laser.

underlying topology of the trajectory of solutions in phase space for these systems is often found to be relatively simple and quite similar, as for three-level models descriptive of optically pumped lasers. An experimental example of such behavior for a single mode far infared molecular laser is shown in Figure 2. In other lasers, for which there is adiabatic elimination of one or other of the fast variables, chaotic behavior is precluded if, as a consequence, the number of variables of the system is reduced to less than three. Indeed this is the case for many practical laser systems. For these systems the addition of an independent external control parameter to the system, such as cavity length or loss modulation, have been extensively used as a means to provide the extra degrees of freedom. Examples include gas lasers with saturable absorbers, solid state lasers with loss modulation, and semiconductor lasers with external cavities, to name but a few. In contrast, for multimode rather than single mode lasers, intrinsic modulation of inversion (or photon flux) by multimode parametric interaction ensures the additional degrees of freedom. Furthermore, when the field is detuned from gain center, the dynamical variables E, P, D are complex, providing five equations for single mode systems, which is more than sufficient to yield deterministic chaos for suitable parameter values. Also, of significance is the remarkably low threshold found for the generation of instabilities and chaos in single-mode laser systems in which the gain is inhomogeneously broadened, an example being the familiar He–Ne laser.

Erratic and aperiodic temporal behavior of any of the system's variables implies a corresponding continuous spectrum for its Fourier transform, which is thus a further signature of chaotic motion. Time series, power spectra, and routes to chaos collectively provide evidence of deterministic behavior. Of the wide range of possible routes to chaos, three have emerged as particularly common and are frequently observed in lasers. These are period doubling, intermittency, and two-frequency scenarios. In the first, a solution, which is initially stable is found to oscillate, the period of which successively doubles at distinct values of the control parameter. This continues until the number of fixed points becomes infinite at a finite parameter value, where the variation in time of the solution becomes irregular. For the intermittency route, a signal that behaves regularly in time becomes interrupted by statistically distributed periods of irregular motion, the average number of which increases with the external control parameter until the condition becomes chaotic. The two-frequency route is more readily identified with early concepts of turbulence, considered to be the limit of an infinite sequence of instabilities (Hopf bifurcation) evolving from an initial stable solution each of which creates a new basic frequency. It is now known that only two or three instabilities (frequencies) are sufficient for the subsequent generation of chaotic motion.

Applications of Chaos in Lasers

It has been accepted as axiomatic since the discovery of chaos that chaotic motion is in general neither predictable nor controllable. It is unpredictable because a small disturbance will produce exponentially growing perturbation of the motion. It is uncontrollable because small disturbances lead to other forms of chaotic motion and not to any other stable and predictable alternative. It is, however, this very sensitivity to small perturbations that has been more recently used to stabilize and control chaos, essentially using chaos to control itself. Among the many methods proposed in the late 1980s, a feedback control approach was proposed by Ott, Grebogi, and Yorke (OGY) in which tiny feedback was used to stabilize unstable periodic orbits or fixed points of chaotic attractors. This control strategy can be best understood by a schematic of the OGY control algorithm for stabilizing a saddle point P^*, as shown in Figure 3. Curved trajectories follow a stable (unstable) manifold towards (away from) the saddle point. Without perturbation of a parameter, the starting state, s_1, would evolve to the state s_2.

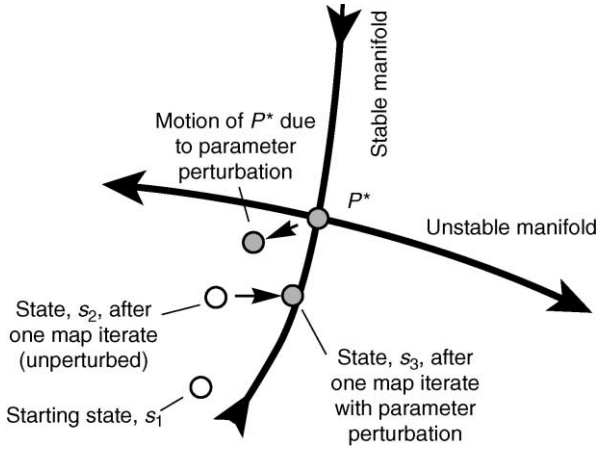


Figure 3 The OGY control strategy for stabilizing an orbit in a chaotic attractor.

The effect of changing a parameter of the system is depicted as shifting states near P^* along the solid black arrows, whereas the combination of the unperturbed trajectory and the effect of the perturbation is to bring the state to a point, s_3, on the stable manifold. Once on the stable manifold, the trajectory naturally tends towards the desired point. This algorithm has been successfully applied in numerous experimental systems without *a priori* modeling of these systems, examples being in cardiology, electronics, and lasers. Based on the concept of feedback control, various other approaches have been developed where emphasis has been given to algorithms which are more readily implemented in practical systems, in particular that utilizing occasional proportional feedback (OPF). These pioneering studies have since inspired prolific activities, both in theory and experiment, of control of chaos across many disciplines, opening up possibilities of utilizing chaos in many diverse systems. In optics, Roy *et al*. first demonstrated dynamical control of an autonomously chaotic and high-dimensional laser system on microsecond time-scales. The laser, a diode pumped solid-state Nd:YAIG system with an intracavity KTP doubling crystal, exhibited chaotic behavior from coupling of the longitudinal modes through nonlinear sum-frequency generation. To control the system, the total laser output was sampled within a window of selected offset and width. A signal proportional the deviation of the sampled intensity from the center of the window was generated and applied to perturb the driving current of the diode laser. The sampling frequency is related to the relaxation oscillation frequency of the system. This control signal repeatedly attempts to bring the system closer to a periodic unstable orbit that is embedded in the chaotic attractor, resulting in a realization of the periodic orbit. By adjusting the frequency of the feedback signal they observed the basic (relaxation oscillation) and many higher-order periodic waveforms of the laser intensity. In a subsequent experiment, Gills *et al*. showed that an unstable steady state of this laser could also be stabilized by the OPF control technique and tracked with excellent stability to higher output power with increase of the pump excitation.

Another area to receive considerable attention in the last few years is that of synchronization of chaos. It may be expected that chaotic systems will defy synchronization because two such identical systems started at nearly the same initial conditions have dynamical trajectories that quickly becomes uncorrelated. However, it has now been widely demonstrated that when these systems are linked, their chaotic trajectories converge to be the same and remain in step with each other. Further, such synchronization is found to be structurally stable and does not require the systems to be precisely identical. Not surprisingly, these findings have attracted interest from the telecommunication community; the natural masking of information by chaotic fluctuations offering a means to a certain degree of security. In this new approach, a chaotic carrier of information can be considered as a generalization of the more traditional sinusoidal carrier. In communication systems that use chaotic waveforms, information can be recovered from the carrier using a receiver synchronized, or tuned, to the dynamics of the transmitter. Optical systems are particularly attractive since they display fast dynamics, offering the possibility of communication at bandwidths of a hundred megahertz or higher. Van Wiggeren and Roy first demonstrated data transmission rates of 10 Mbites per second using erbium-doped fiber ring lasers. These lasers are particularly well-suited for communication purposes since their lasing wavelength is close to the minimum-loss wavelength in optical fiber. Figure 3 shows a schematic of this system. The tiny message, in order of one thousandth of the carrier intensity, is decoded in the transmitter. The receiver tends to synchronize its behavior to the chaotic part of the transmitted wave (but not the message). Subtracting the waveform created in the receiver from the transmitted signal recovers the tiny message. In principle, it is possible to communicate information at ultrahigh data rate with the use of this scheme because the chaotic dynamics in the ring laser has a very large spectral width. In a later experiment they showed that a receiver can recover information at 126 Mbits/sec from the chaotic carrier (Figure 4).

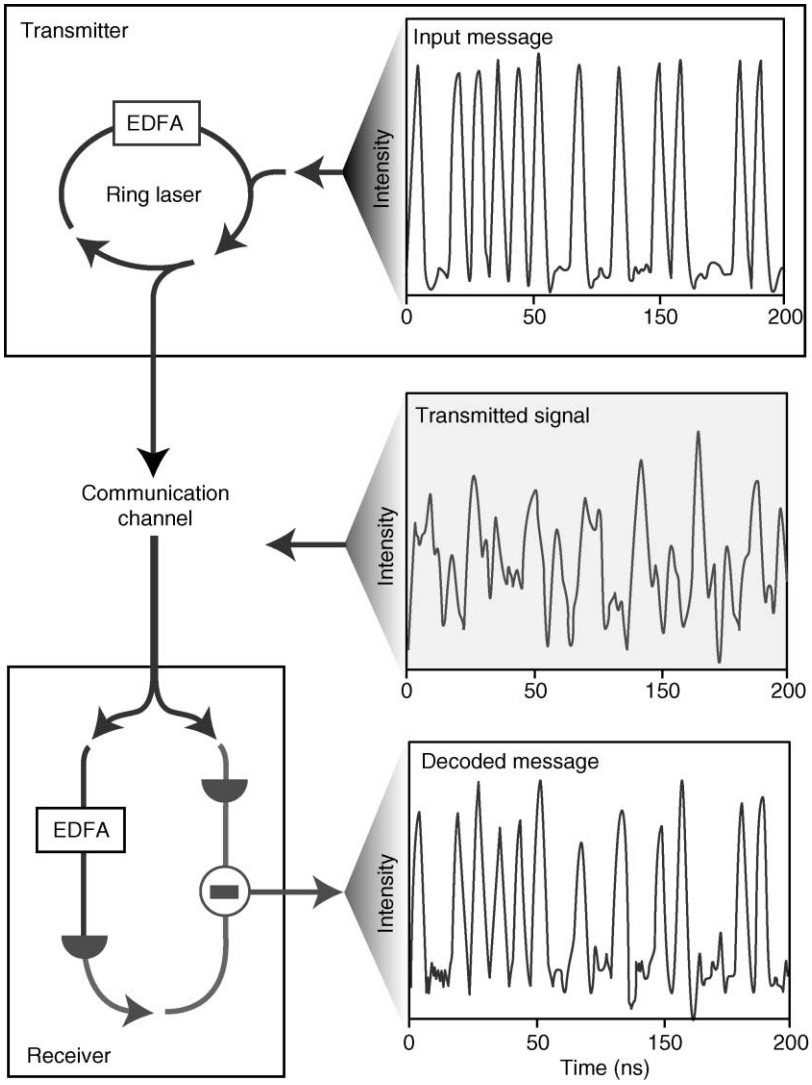


Figure 4 Schematic of communication with chaotic erbium-doped fiber amplifiers (EDFAs). Injecting a message into the transmitter laser folds the data into the chaotic frequency fluctuations. The receiver reverses this process, thereby recovering a high-fidelity copy of the message. (Reproduced with permission from Gauthier DJ (1998) Chaos has come again. *Nature* 279: 1156–1157.)

See also

Fourier Optics. **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems. **Polarization:** Introduction.

Further Reading

Abraham NB, Lugiato LA and Narducci LM (eds) (1985) Instabilities in active optical media. *Journal of the Optical Society of America B* 2: 1–272.

Arecchi FT and Harrison RG (1987) *Instabilities and Chaos in Quantum Optics*. Synergetics Series 34, 1–253. Berlin: Springer Verlag.

Arecchi FT and Harrison RG (eds) (1994) *Selected Papers on Optical Chaos*. SPIE Milestone Series, vol. MS 75, SPIE Optical Engineering Press.

Gauthier DJ (1998) Chaos has come again. *Nature* 279: 1156–1157.

Gills Z, Iwata C and Roy R (1992) Tracking unstable steady states: extending the stability regime of a multimode laser system. *Physical Review Letters* 69: 3169–3172.

Haken H (1985) Light. In: *Laser Light Dynamics*, vol. 2. Amsterdam: North-Holland.

Harrison RG (1988) Dynamical instabilities and chaos in lasers. *Contemporary Physics* 29: 341–371.

Harrison RG and Biswas DJ (1985) Pulsating instabilities and chaos in lasers. *Progress in Quantum Electronics* 10: 147–228.

Hunt ER (1991) Stabilizing high-period orbits in a chaotic system – the diode resonator. *Physical Review Letters* 67: 1953–1955.

Ott E, Grebogi C and Yorke JA (1990) Controlling chaos. *Physical Review Letters* 64: 1196–1199.

Pecore LM and Carrol TL (1990) Synchronization of chaotic systems. *Physical Review Letters* 64: 821–824.
Roy R, *et al.* (1992) Dynamical control of a chaotic laser: experimental stabilization of a globally coupled system. *Physical Review Letters* 68: 1259–1262.
VanWiggeren GD and Roy R (1998) Communication with chaotic lasers. *Science* 279: 1198–1200.
VanWiggeren GD and Roy R (1998) Optical communication with chaotic waveforms. *Physical Review Letters* 279: 1198–1200.

CHEMICAL APPLICATIONS OF LASERS

Contents

Detection of Single Molecules in Liquids
Diffuse-Reflectance Laser Flash Photolysis
Laser Manipulation in Polymer Science
Nonlinear Spectroscopies
Photodynamic Therapy of Cancer
Pump and Probe Studies of Femtosecond Kinetics
Time-Correlated Single-Photon Counting
Transient Holographic Grating Techniques in Chemical Dynamics

Detection of Single Molecules in Liquids

A J de Mello, J B Edel and E K Hill, Imperial College of Science, Technology and Medicine, London, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A significant challenge facing experimentalists in the physical and biological sciences is the detection and identification of single molecules. The ability to perform such sensitive and selective measurements is extremely valuable in applications such as DNA analysis, immunoassays, environmental monitoring, and forensics, where small sample volumes and low analyte concentrations are the norm. More generally, most experimental observations of physical systems provide a measurement of ensemble averages, and yield information only on mean properties. In contrast, single molecule measurements permit observation of the interactions and behavior of a heterogeneous population in real time.

Over the past few years a number of techniques with sufficient sensitivity have been developed to detect single molecules. Scanning probe microscopies (most notably scanning tunneling microscopy and atomic force microscopy) have been used to great effect in the analysis of surface bound species, but for the detection of single molecules in liquids, optical methods incorporating the measurement of absorption or emission processes, have proved most successful.

The Absorption–Emission Cycle

The key concept underlying most emissive approaches to single molecule detection is that a single molecule can be cycled repeatedly between its ground state and an excited electronic state to yield multiple photons. The process can be understood by reference to Figure 1. Fluorescence emission in the condensed phase can be described using a four-step cycle. Excitation from a ground electronic state to an excited state is followed by rapid (internal) vibrational relaxation. Subsequently, radiative decay to the ground state is observed as fluorescence emission and is governed by the excited state lifetime. The final stage is internal relaxation back to the original ground state. Under saturating illumination, the rate-limiting step for this cycle is governed by the fluorescence lifetime (τ_f), which is typically of the order of a few nanoseconds. If a single molecule diffuses through an illuminated zone (e.g., the focus of a laser beam) it may reside in that

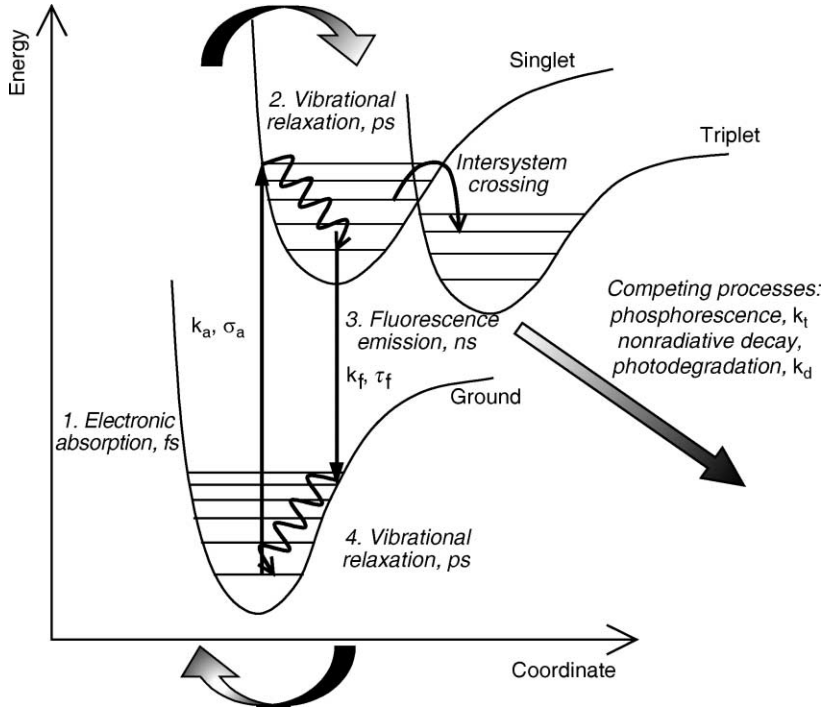


Figure 1 Schematic illustration of the molecular absorption–emission cycle and timescales for the component processes. Competing processes that may reduce the ultimate photon yield are also shown.

region for several milliseconds. The rapid photon absorption-emission cycle may therefore be repeated many times during the residence period, resulting in a burst of fluorescence photons as the molecule transits the beam. The burst size is limited theoretically by the ratio of the beam transit time (τ_r) and the fluorescence lifetime:

$$N_{\text{photons}} = \frac{\tau_r}{\tau_f} \qquad [1]$$

For typical values of τ_r (5 ms) and τ_f (5 ns) up to one million photons may be emitted by the single molecule. In practice, photobleaching and photodegradation processes limit this yield to about 10^5 photons. Furthermore, advances in optical collection and detection technologies enable registration of about 1–5% of all photons emitted. This results in a fluorescence burst signature of up to a few thousand photons or photoelectrons.

Successful single molecule detection (SMD) depends critically on the optimization of the fluorescence burst size and the reduction of background interference from the bulk solvent and impurities. Specifically a molecule is well suited for SMD if it is efficiently excited by an optical source (i.e., possesses a large molecular absorption cross-section at the wavelength of interest), has a high fluorescence quantum efficiency (favoring radiative deactivation of the excited state), has a short fluorescence lifetime, and is exceptionally photostable.

Ionic dyes are often well suited to SMD as fluorescence quantum efficiencies can be close to unity and fluorescence lifetimes below 10 nanoseconds. For example, xanthene dyes such as Rhodamine 6G and tetramethyl-rhodamine isothiocyanate are commonly used in SMD studies. However, other highly fluorescent dyes such as fluorescein are unsuitable for such applications due to unacceptably high photodegradation rate coefficients. Furthermore, some solvent systems may enhance nonradiative processes, such as intersystem crossing, and yield significant reduction in the photon output. Structures of three common dyes suitable for SMD are shown in Figure 2.

Signal vs Background

The primary challenge in SMD is to ensure sufficient reduction in background levels to enable discrimination between signal and noise. As an example, in a 1 nM aqueous dye solution each solute molecule occupies a volume of approximately 1 fL. However, this same volume also contains in excess of 10^{10} solvent molecules. Despite the relatively small scattering cross-section for an individual water molecule ($\sim 10^{-28}$ cm^2 at 488 nm), the cumulative scattering

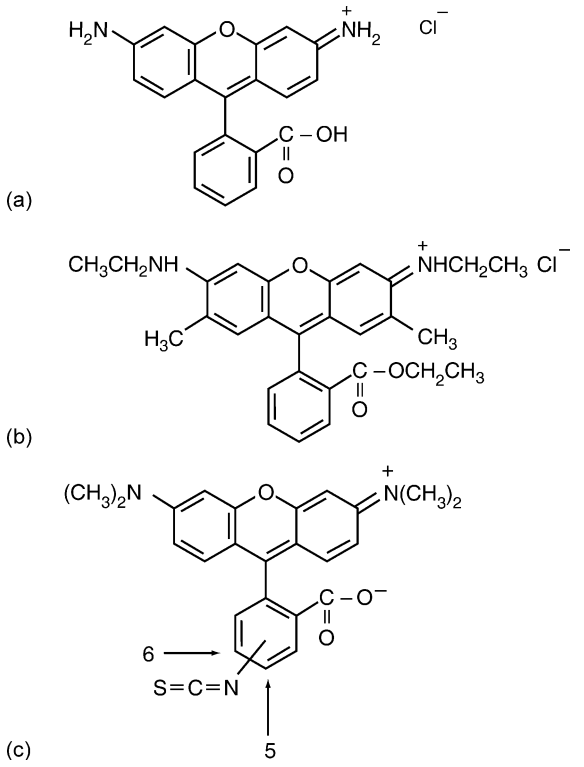


Figure 2 Structures of common dye molecules suitable for SMD in solution: (a) 3,6-diamino-9-(2-carboxyphenyl)-chloride (rhodamine 110); (b) 9-(2-(ethoxycarbonyl)phenyl)-3,6-bis (ethylamino)-2,7-dimethyl chloride (rhodamine 6G); (c) 9-(2-carboxy-isothiocyanatophenyl)-3,6-bis(dimethylamino)-inner salt (tetra methylrhodamine-5-(and-6)-isothiocyanate).

signal from the solvent may swamp the desired fluorescence signal. The principal method of reducing the solvent background is to minimize the optical detection volume: the signal from a single molecule is independent of probe volume dimensions, but the background scales proportionally with the size of the detection region. Although there are several experimental approaches to SMD in solution, several factors hold common:

1. Tiny detection volumes (10^{-12}–10^{-15} L) are used to reduce background signals.
2. A low analyte concentration combined with the small observation volume, ensures that less than one analyte molecule is present in the probe volume on average.
3. High-efficiency photon collection (optics) and detection maximize the proportion of the isotropic fluorescence burst that is registered.
4. Background reduction methods are employed to improve signal-to-noise ratios. These include: optical rejection of Raman and Rayleigh scatter, time-gated discrimination between prompt scatter and delayed emission, and photobleaching of the solvent immediately before detection.

The minute volumes within which single molecules are detected can be generated in a variety of ways. Picoliter volumes can be defined by mutually orthogonal excitation and detection optics focused in a flowing stream. Much smaller, femtoliter probe volumes are generated using confocal microscopes. At this level, background emission is significantly reduced and high signal-to-noise ratios can be achieved. Confocal detection techniques are versatile and have been widely adopted for SMD in freely diffusing systems. Consequently, confocal methods will be discussed in detail in this article. The other general approach to performing SMD in solution involves the physical restriction of single molecules within defined volumes. Of particular note are techniques where single molecules are confined within a stream of levitated microdroplets. Droplet volumes are typically less than 1 fL and imaging of the entire microdroplet enables single molecule fluorescence to be contrasted against droplet 'blanks' with good signal-to-noise. Furthermore, since molecules are confined within discrete volumes, the technique can be utilized for high-efficiency molecular counting applications. More recently, spatial confinement of molecules in capillaries and microfabricated channels (with submicron dimensions) has been used to create probe volumes between 1 fL and 1 pL, and immobilized molecules on surfaces have been individually probed using wide-field microscopy with epi-illumination or evanescent wave excitation.

Single Molecule Detection using Confocal Microscopy

As previously stated, the confocal fluorescence microscope is an adaptable and versatile tool for SMD. In its simplest form, a confocal microscope is one in which a point light source, a point focus in the object plane, and a pinhole detector are all confocal with each other. This optical superposition generates superior imaging properties and permits definition of ultra-small probe volumes. The concepts behind a basic confocal microscope and its use in SMD are schematically illustrated in Figure 3. Coherent light (typically from a laser and tuned to an optical transition of the molecule under investigation) behaves as a point light source and is focused into a sample chamber using a high-numerical aperture objective lens. As a single molecule traverses the laser beam it is continuously cycled between the ground and an excited electronic state, emitting a burst of fluorescence photons. Fluorescence emission is isotropic (spontaneous emission), so photons are emitted in all directions (4π steradians).

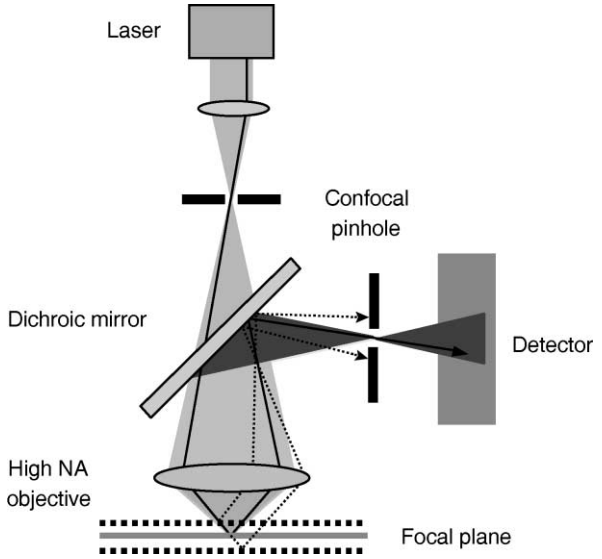


Figure 3 Principle of confocal detection. A confocal pinhole only selects light that emanates from the focal region. Dashed lines indicate paths of light sampled above and below the focal plane that are rejected by the pinhole. The solid ray derives from the focal point, and is transmitted through the pinhole to a detector.

Consequently, the high numerical aperture is used to collect as large a fraction of photons emitted from the focal plane as possible. Designing the objective to be used with an immersion medium, such as oil, glycerin, or water, can dramatically increase the objective numerical aperture, and thus the number of collected photons. Light collected by the objective is then transmitted towards a dichroic beam splitter. In the example shown, fluorescence photons (of lower energy) are reflected towards the confocal detector pinhole, whereas scattered radiation (of higher energy) is transmitted through the dichroic towards the light source. Creation of a precise optical probe volume is effected through the definition of the confocal pinhole. The detector is positioned such that only photons that pass through the pinhole are detected. Consequently, light emanating from the focal plane in the sample is transmitted through the pinhole and detected, whereas light not deriving from the focal plane is rejected by the aperture, and therefore not detected (Figure 3).

To ensure that the maximum number of photons are detected by the system, high efficiency detectors must be used. Photomultiplier tubes (the most common detectors for light-sensing applications) are robust and versatile but have poor detection efficiencies (approximately 5% of all photons that fall on the photocathode yield an electrical signal). Consequently, the most useful detectors for SMD (or low light level) applications are single photon-counting avalanche photodiodes (SPADs). A SPAD is essentially a p–n junction reverse biased above the breakdown voltage, that sustains an avalanche current when triggered by a photon-generated carrier. Detection efficiencies for typical SPADs are normally between 60–70% and are thus ideal for SMD in solution. An approximation of the overall detection efficiency of a confocal system for SMD can be made using eqn [2], which incorporates an estimation of photon losses at all stages of the collection/detection process. Typical transmission efficiencies for each step are also shown.

overall detection efficiency	≈	objective collection efficiency	×	dichroic transmission coefficient	×	additional optical losses	×	detector efficiency
0.06		0.24		0.9		0.5		0.6

[2]

Optical Probe Volumes

The precise nature of the probe volume is determined by the image of the pinhole in the sample and the spherical aberration of the microscope objective. Routinely, confocal probe volumes are approximated as resembling a cylinder with a radius defined by the diffraction-limited waist of a Gaussian beam. This approximation is useful when the incident beam is narrow or not tightly focused. However, when the radius of the incident beam is large, the corresponding diffraction limited focus is narrowed, and the probe volume more closely resembles a pair of truncated cones. Figure 4a illustrates the dependence of the curvature of the $1/e^2$ intensity contour on the collimated beam radius.

Consequently, it is clear that a simple cylindrical approximation for the probe volume breaks down for wide, tightly focused beams. If a broad incident beam (diameter >1.5 mm) is used, a large noncylindrical contribution to the probe volume is anticipated and a more appropriate model is required. An alternative and more accurate model for the confocal probe volume considers the Gaussian profile of the focused beam. The $1/e^2$ intensity contour radius of a Gaussian waveform with wavelength λ, at some distance z from the beam waist radius w_0, is given by eqn [3]:

$$w(z) = w_0\sqrt{1 + \left(\frac{\lambda z}{\pi w_0^2}\right)^2} \qquad [3]$$

In this case, the probe volume V is given by the volume of rotation of $w(z)$ around the z-axis between Z' and $-Z'$. The volume of rotation can therefore be

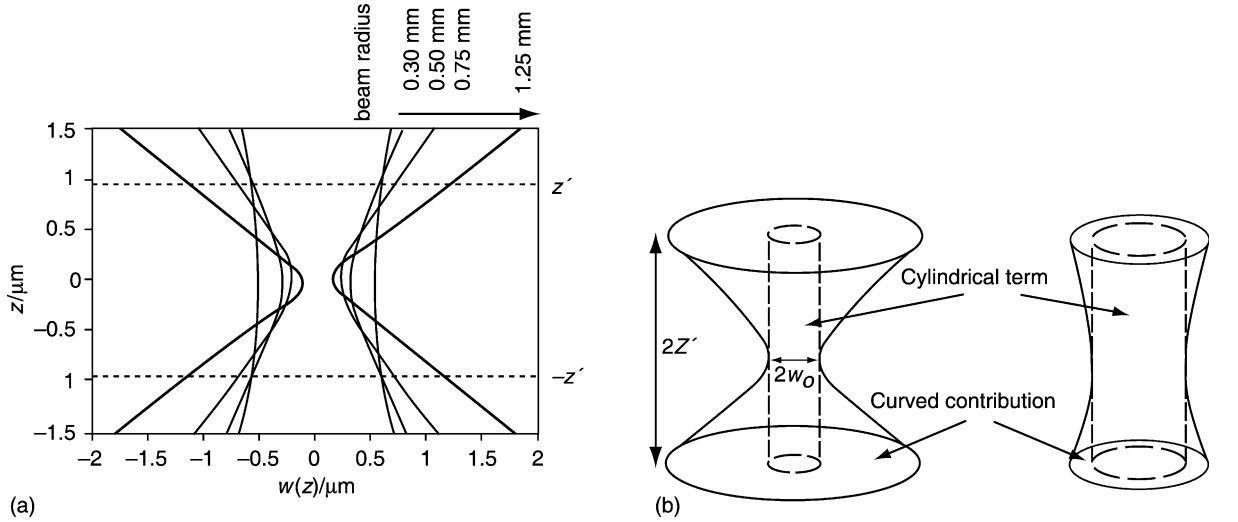


Figure 4 (a) $1/e^2$ Gaussian intensity contours plotted for a series of laser beam radii ($\lambda = 488$ nm, $f = 1.6$ mm, $n = 1.52$). (b) Cylindrical and curved components of the Gaussian probe volume. The curved contribution is more significant for larger beam radii and correspondingly tight beam waists.

simply defined according to,

$$V = \int_{-Z'}^{Z'} \pi w(z)^2 dz \qquad [4]$$

Solution of eqn [4] yields

$$V = 2\pi w_0^2 Z' + \frac{2\lambda^2}{3\pi w_0^2} Z'^3 \qquad [5]$$

The Gaussian volume expression contains two terms. The first term, $2\pi w_0^2 Z'$, corresponds to a central cylindrical volume; the second term has a more complex form that describes the extra curved volume (Figure 4b). The diffraction-limited beam waist radius w_0 can be defined in terms of the focusing objective focal length f, the refractive index n, and the collimated beam radius R according to

$$w_0 = \frac{\lambda f}{n\pi R} \qquad [6]$$

Substitution in eqn [5] yields,

$$V = 2\pi\left(\frac{\lambda f}{n\pi R}\right)^2 Z' + \frac{2\lambda^2}{3\pi}\left(\frac{n\pi R}{\lambda f}\right)^2 Z'^3$$

$$= \frac{2\lambda^2 f^2}{\pi n^2 R^2} Z' + \frac{2\pi n^2 R^2}{3f^2} Z'^3 \qquad [7]$$

The volume is now expressed in terms of identifiable experimental variables and constants. Once again, the first term may be correlated with the cylindrical contribution to the volume, and the second term is the additional volume due to the curved contour. It is clear from Figure 4a that, for a given focal length, the noncylindrical contribution to the probe volume increases with incident beam diameter, when the diffraction limited focus is correspondingly sharp and narrow. Furthermore, it can also be seen that the second term in eqn [7] is inversely proportional to f^2, and thus the extent to which the probe volume is underestimated by the cylindrical approximation increases with decreasing focal length. This fact is significant when performing confocal measurements, since high numerical aperture objectives with short focal lengths are typical. Some realistic experimental parameters give an indication of typical dimensions for the probe volume in confocal SMD systems. For example, if $\lambda = 488$ nm, $f = 1.6$ mm, $Z' = 1.0$ μm and $n = 1.52$, a minimum optical probe volume of 1.1 fL is achievable with a collimated beam diameter of 1.1 mm.

Intensity Fluctuations: Photon Burst Statistics

When sampling a small volume within a system that may freely exchange particles with a large surrounding analyte bath, a Poisson distribution of particles is predicted. A Poisson distribution is a discrete series that is defined by a single parameter μ equating to the mean and variance of the distribution:

$$P(n = x) = \frac{\mu^x e^{-\mu}}{x!} \qquad [8]$$

Common Poisson processes include radioactive disintegration, random walks and Brownian motion. Although particle number fluctuations in the excitation volume are Poissonian in nature, the corresponding fluorescence intensity modulation induces a stronger correlation between photon counts. For a single molecular species the model is described by two parameters: an intrinsic molecular brightness and the average occupancy of the observation volume. A super-Poissonian distribution has a width or variance that is greater than its mean; in a Poisson distribution the mean value and the variance are equal. The fractional deviation Q is defined as the scaled difference between the variance and the expectation value of the photon counts, and gives a measure of the broadening of the photon counting histogram (PCH). Q is directly proportional to the molecular brightness factor ε and the shape factor γ of the optical point spread function. (γ is constant for a given experimental geometry.)

$$Q = \frac{\langle \Delta n^2 \rangle - \langle n \rangle}{\langle n \rangle} = \gamma \varepsilon \qquad [9]$$

A pure Poisson distribution has $Q = 0$; for super-Poissonian statistics $Q > 0$. Deviation from the Poisson function is maximized at low number density and high molecular brightness.

In a typical SMD experiment raw data are generally collected with a multichannel scalar and photons are registered in binned intervals. Figure 5 illustrates typical photon burst scans demonstrating the detection of single molecules (R-phycoerythrin) in solution. Fluorescence photon bursts, due to single molecule events, are clearly distinguished above a low background baseline (top panel) of less than 5 counts per channel in the raw data. It is noticeable that bursts vary in both height and size. This is in part due to the range of possible molecular trajectories through the probe volume, photobleaching kinetics, and the nonuniform illumination intensity in the probe region. In addition, it can be seen that the burst frequency decreases as bulk solution concentration is reduced. This effect is expected since the properties of any given single-molecule event are determined by molecular parameters alone (e.g., photophysical and diffusion constants) and concentration merely controls the frequency/number of events.

Although many fluorescence bursts are clearly distinguishable from the background, it is necessary to set a count threshold for peak discrimination in order to correctly identify fluorescence bursts above the background. A photocount distribution can be used as the starting point for determining an appropriate threshold for a given data set. The overlap between signal and background photocount distributions affects the efficiency of molecular detection. Figure 6 shows typical signal and background photocount probability distributions, with a threshold set at approximately 2 photocounts per channel. The probability of spurious (or false)

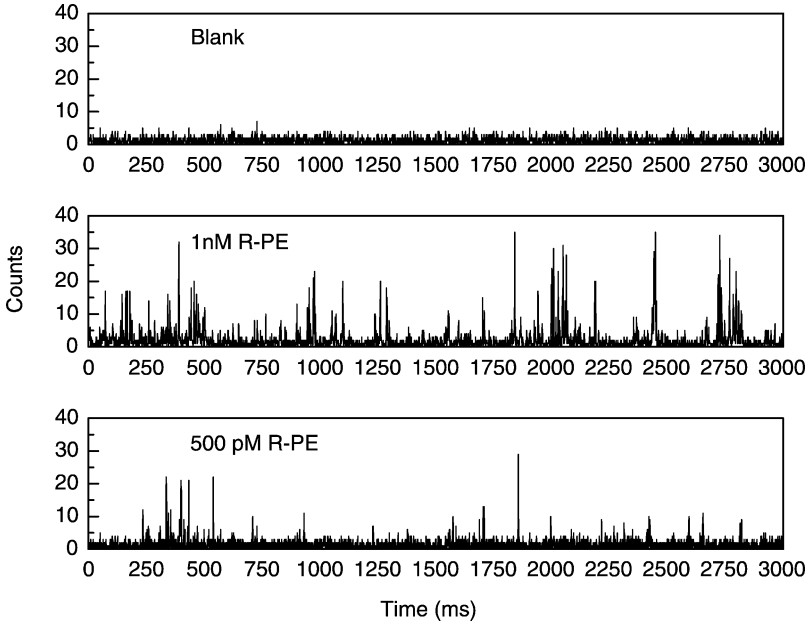


Figure 5 Photon burst scans originating from 1 nM and 500 pM R-phycoerythrin buffered solutions. Sample is contained within a 50 μm square fused silica capillary. Laser illumination = 5 μW, channel width = 1 ms. The top panel shows a similar burst scan originating from a deionized water sample measured under identical conditions.

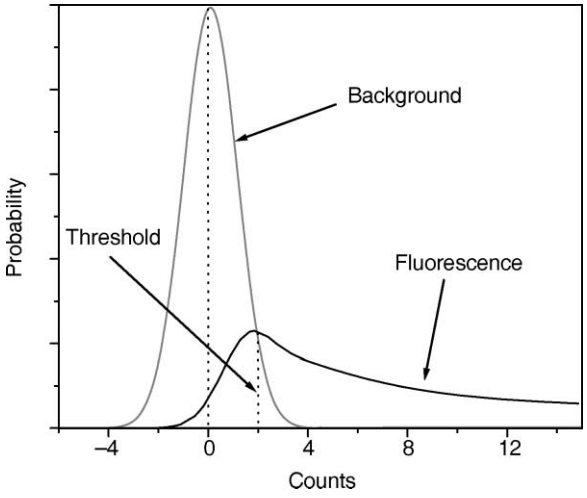


Figure 6 Simulated fluorescence and background photocount probability distributions. The vertical dashed line at 2 counts represents an arbitrarily defined threshold value for peak determination.

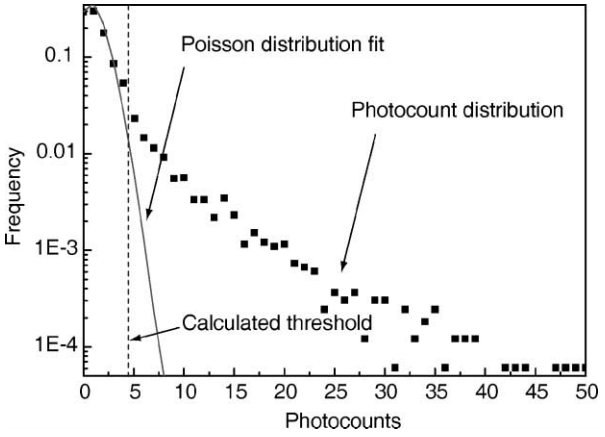


Figure 7 A photon counting histogram generated from a 16 second photon burst scan originating from a 10 μg/mL solution of 1000 nm fluorescent microbeads. The dotted curve shows a least-squares fit of early channels to a Poisson distribution, and the dashed vertical line marks the peak threshold (defined as $\mu + 3\sqrt{\mu} = 4.47$ counts).

detection resulting from statistical fluctuations in the background can be quantified by the area under the 'background' curve at photocount values above the threshold. Similarly, the probability that 'true' single molecule events are neglected can be estimated by the area under the 'fluorescence' curve at photocount values below the threshold. Choice of a high threshold value will ensure a negligible probability of calling a false positive, but will also exclude a number of true single molecule events that lie below the threshold value. Conversely, a low threshold value will generate an unacceptably high number of false positives. Consequently, choice of an appropriate threshold is key in efficient SMD.

Since the background shot noise is expected to exhibit Poisson statistics, the early part of the photocount distribution (i.e., the portion that is dominated by low, background counts) can be modeled with a Poisson distribution, to set a statistical limit for the threshold. Photon counting events above this threshold can be defined as photon bursts associated with the presence of single molecules. In an analogy with Gaussian systems the selected peak discrimination threshold can be defined as three standard deviations from the mean count rate:

$$n_{\text{threshold}} = \mu + 3\sqrt{\mu} \qquad [10]$$

Adoption of a threshold that lies 3 standard deviations above the mean yields a confidence limit that is typically greater than 99%. Figure 7 illustrates a sample photocount distribution, a least-squares fit to an appropriate Poisson distribution, and the calculated threshold that results. Once the threshold has been calculated, its value is subtracted from all channel counts and a peak search utility used to identify burst peaks in the resulting data set.

Data Filtering

As stated, the primary challenge in detecting single molecules in solution is not the maximization of the detected signal, but the maximization of the signal-to-noise ratio (or the reduction of background interferences). Improving the signal-to-noise ratio in such experiments is important, as background levels can often be extremely high.

Several approaches have been used to smooth SMD data with a view to improving signal-to-noise ratios. However, the efficacy of these methods is highly dependent on the quality of the raw data obtained in experiment. As examples, three common methods are briefly discussed. The first method involves the use of a weighted quadratic sum (WQS) smoothing filter. The WQS function creates a weighted sum of adjacent terms according to

$$\tilde{n}_{k,WQS} = \sum_{j=0}^{m-1} w_j (n_{k+j})^2 \qquad [11]$$

The range of summation m is the same order as the burst width, and the weighting factors w_j are chosen

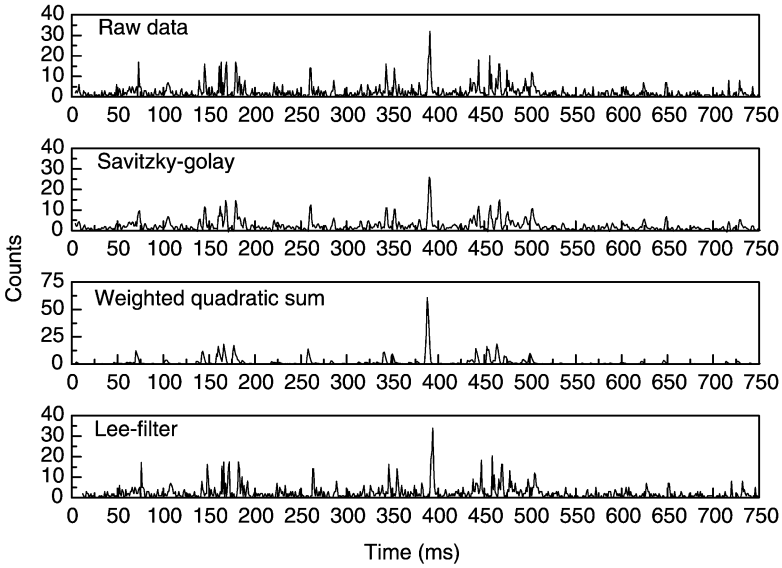


Figure 8 Effect of various smoothing filters on a 750 ms photon burst scan originating from a 1 nM R-phycoerythrin buffered solution. Raw data are shown in the top panel.

to best discriminate between single-molecule signals and random background fluctuations. This method proves most useful for noisy systems, in which the raw signal is weak. There is a practical drawback, in that peak positions are shifted by the smoothing function, and subsequent burst analysis is therefore hampered.

Another popular smoothing filter is the Lee-filtering algorithm. The Lee filter preferentially smoothes background photon shot noise and is defined according to

$$\tilde{n}_k = \bar{n}_k + (n_k - \bar{n}_k)\frac{\sigma_k^2}{\sigma_k^2 + \sigma_0^2} \qquad [12]$$

where the running mean (n_k) and running variance (σ_k^2) are defined by

$$\bar{n}_k = \frac{1}{(2m+1)}\sum_{j=-m}^{m} n_{k+j} \quad m < k \leq N - m \qquad [13]$$

$$\sigma_k^2 = \frac{1}{(2m+1)}\sum_{j=-m}^{m} (n_{k+j} - \bar{n}_{k+j})^2 \quad 2m < k \leq N - 2m \qquad [14]$$

for a filter $(2m+1)$ channels wide. Here, n_k is the number of detected photons stored in a channel k, σ_0 is a constant filter parameter, and N is the total number of data points.

A final smoothing technique worth mentioning is the Savitsky-Golay smoothing filter. This filter uses a least-squares method to fit an underlying polynomial function (typically a quadratic or quartic function) within a moving window. This approach works well for smooth line profiles of a similar width to the filter window and tends to preserve features such as peak height, width and position, which may be lost by simple adjacent averaging techniques. Figure 8 shows the effects of using each approach to improve signal-to-noise for raw burst data.

Photon Burst Statistics

A valuable quantitative analysis method for analysis of fluorescence bursts utilizes the analysis of Poisson statistics. Burst interval distributions are predicted to follow a Poissonian model, in which peak separation frequencies adopt an exponential form. The probability of a single molecule (or particle) event occurring after an interval Δt is given by eqn [15]:

$$N(\Delta t) = \lambda \exp(-\beta t) \qquad [15]$$

where λ is a proportionality constant and β is a characteristic frequency at which single molecule events occur. The recurrence time τ_R can then be simply defined as

$$\tau_R = \frac{1}{\beta} \qquad [16]$$

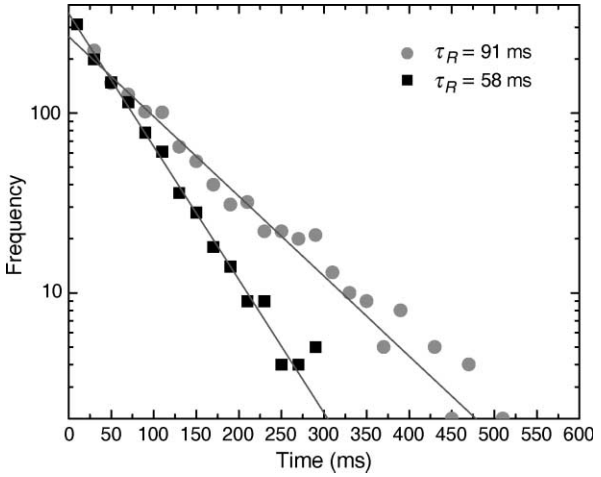


Figure 9 Burst interval distribution analysis of photon burst scans. Data originate from 1 μm fluorescent beads moving through 150 μm wide microchannels at flow rates of 200 nL min^{-1} (circles) and 1000 nL min^{-1} (squares). Least squares fits to a single exponential function are shown by the solid lines.

Equation [6] simply states that longer intervals between photon bursts are less probable than shorter intervals. Furthermore, the recurrence time reflects a combination of factors that control mobility, probe volume occupancy, or other parameters in the single molecule regime. Consequently, it is expected that τ_R should be inversely proportional to concentration, flow rate or solvent viscosity in a range of systems. Figure 9 shows an example of frequency $N(\Delta t)$ versus time plots for two identical particle systems moving at different velocities through the probe volume. A least-squares fit to a single exponential function yields values of $\tau_R = 91$ ms for a volumetric flow rate of 200 nL/min and $\tau_R = 58$ ms for a volumetric flow rate of 1000 nL/min.

Temporal Fluctuations: Autocorrelation Analysis

Autocorrelation analysis is an extremely sensitive method for detecting the presence of fluorescence bursts in single molecule experiments. This approach essentially measures the average of a fluctuating signal as opposed to the mean spectral intensity. As previously discussed, the number of molecules contained within a probe volume at any given time is governed by Poisson statistics. Consequently, the root mean square fluctuation can be defined according to eqn [17],

$$\frac{\sqrt{\langle(\delta N)^2\rangle}}{\langle N\rangle} = \frac{\sqrt{\langle(N-\langle N\rangle)^2\rangle}}{\langle N\rangle} = \frac{1}{\sqrt{\langle N\rangle}} \quad [17]$$

where N is the number of molecules. It is observed that the relative fluctuation diminishes as the number of particles measured is increased. Hence, it is important to minimize the number of molecules present in the probe volume. It should be noted, however, that if there are too few molecules in the solution there will be long dark periods were no single molecule bursts are observed. If the probe volume is bathed in radiation of constant intensity, fluctuation of the resulting fluorescence signal can simply be defined as deviations from the temporal signal average:

$$\langle F(t)\rangle = \frac{1}{T}\int_0^T F(t)\mathrm{d}t \quad [18]$$

Here, t is defined as the total measurement time, $F(t)$ is the fluorescence signal at time t, and $\langle F(t)\rangle$ is the temporal signal average. Fluctuations in the fluorescence intensity, $\delta F(t)(\delta F(t) = F(t) - \langle F(t)\rangle)$, with time t, about an equilibrium value $\langle F\rangle$, can be statistically investigated by calculating the normalized autocorrelation function, $G(\tau)$, where

$$G(\tau) = \frac{\langle F(t+\tau)F(t)\rangle}{\langle F\rangle^2} = \frac{\langle \delta F(t+\tau)\delta F(t)\rangle}{\langle F\rangle^2} + 1 \quad [19]$$

In dedicated fluorescence correlation spectroscopy experiments, the autocorrelation curve is usually generated in real time in a high-speed digital correlator. Post data acquisition calculation is also possible using the following expression

$$G(\tau) = \sum_{t=0}^{N-1} g(t)g(t+\tau) \quad [20]$$

Here $g(t)$ is the total number of counts during the time interval $(t, t+\Delta t)$, $g(t+\tau)$ is the number of counts detected in an interval of Δt at a later time $t+\tau$, and N is the total number of time intervals in the dataset. In a diffusion controlled system with a single fluorescent molecule that is irradiated with a three dimensional Gaussian intensity profile, the autocorrelation curve is governed by the mean probe volume occupancy N and the characteristic diffusion time (τ_D). The laser beam waist radius ω and the probe depth $2z$ describe the Gaussian profile:

$$G(\tau) = 1 + \frac{1}{N}\left(1+\frac{\tau}{\tau_d}\right)^{-1}\left(1+\left(\frac{\omega}{z}\right)^2\frac{\tau}{\tau_d}\right)^{-1/2} \quad [21]$$

The diffusion time is a characteristic molecular residence time in the probe volume and inversely

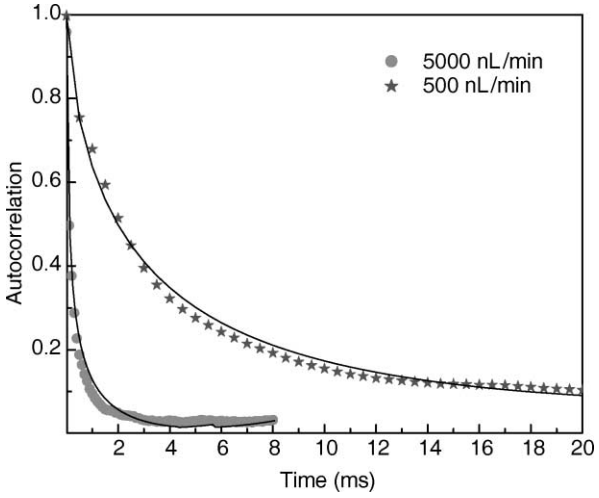


Figure 10 Autocorrelation analysis of photon burst scans of 1 μm fluorescent beads moving through 150 μm wide microchannels at flow rates of 500 nL min^{-1} (stars) and 1000 nL min^{-1} (circles). Solid lines represent fits to the data according to eqn [23].

related to the translational diffusion coefficient for the molecule:

$$D = \frac{w^2}{4\tau_D} \qquad [22]$$

In a flowing system, the autocorrelation function depends on the average flow time through the probe volume τ_{flow}. A theoretical fit to the function can be described according to

$$G(\tau) = 1 + \frac{1}{N} A \exp\left\{\left(\frac{\tau}{\tau_{flow}}\right)^2 A\right\}$$
$$A = \left(1 + \frac{\tau}{\tau_d}\right)^{-1}\left(1 + \left(\frac{\omega}{z}\right)^2 \frac{\tau}{\tau_d}\right) \qquad [23]$$

where N is the mean probe volume occupancy; the flow velocity v can then be extracted from the characteristic flow time according to

$$v = \frac{w}{\tau_{flow}} \qquad [24]$$

It should be noted that in the case that directed flow is negligible or defined to be zero, the autocorrelation function simplifies to eqn [21].

Figure 10 illustrates experimentally determined autocorrelation curves for two identical particle systems moving at different velocities through the probe volume. As particle flow velocity is increased, the width of the autocorrelation curves is seen to narrow as a result of the reduced residence time in the probe volume. A plot of the reciprocal of the full width half maximum of the autocorrelation curve as a function of volumetric flow rate is linear, and provides a simple way of calculating particle/molecule velocities within flowing systems.

Applications

The basic tools and methods outlined in this chapter have been used to perform SMD in a variety of chemically and biologically relevant systems, and indeed there is a large body of work describing the motion, conformational dynamics and interactions of individual molecules (see Further Reading). A primary application area has been in the field of DNA analysis, where SMD methods have been used in DNA fragment sizing, single-molecule DNA sequencing, high-throughput DNA screening, single-molecule immunoassays, and DNA sequence analysis. SMD methods have also proved highly useful in studying protein structure, protein folding, protein-molecule interactions, and enzyme activity.

More generally, SMD methods may prove to be highly important as a diagnostic tool in systems where an abundance of similar molecules masks the presence of distinct molecular anomalies that are markers in the early stages of disease or cancer.

See also

Microscopy: Confocal Microscopy.

Further Reading

Ambrose WP, Goodwin PM, Jett JH, Van Orden A, Werner JH and Keller RA (1999) Single molecule fluorescence spectroscopy at ambient temperature. *Chem. Rev.* 99: 2929–2956.

Barnes MD, Whitten WB and Ramsey JM (1995) Detecting single molecules in liquids. *Anal. Chem.* 67: 418A–423A.

Basche T, Orrit M and Rigler R (2002) *Single Molecule Spectroscopy in Physics, Chemistry, and Biology.* New York: Springer-Verlag.

Zander CJ, Enderlein J and Keller RA (2002) *Single Molecule Detection in Solution: Methods and Applications*. New York: John Wiley & Sons.

Diffuse-Reflectance Laser Flash Photolysis

D R Worrall and S L Williams, Loughborough University, Loughborough, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Flash photolysis was developed as a technique to study short-lived intermediates in photoinduced reactions by George Porter (later Lord Porter of Luddenham) and Ronald Norrish in 1949, drawing on Porter's wartime experience with radar techniques. Such was the impact of this development that it earned the Nobel prize jointly for Porter, Norrish, and Eigen in 1967. In this article we will describe the application of flash photolysis to opaque, scattering samples, detailing how light propagation in such samples can be treated theoretically, and will discuss methods by which the data obtained from diffuse reflectance flash photolysis experiments may be analyzed.

The technique of flash photolysis was originally based on using an intense flash of light (the photolysis flash) from a xenon tube to excite the sample, followed a certain time delay later by a spectroscopic flash of lower intensity from a second flash tube, the light from the latter being detected using a photographic plate. The photolysis flash is of sufficient intensity to produce a large population of intermediates (radicals, ions, excited states, isomers) in the sample, which then absorb light from the spectroscopic flash depending on the concentration of intermediates according to the Beer–Lambert law:

$$A = \varepsilon c l \qquad [1]$$

where A is the sample absorbance, ε the molar absorption coefficient, c the concentration and l the pathlength. The absorbance is related to the incident and transmitted intensities as

$$A = \log_{10} I_0/I \qquad [2]$$

With I_0 the incident and I the transmitted intensities, there is therefore an exponential fall-off of intensity with pathlength for a homogeneous absorber. Hence, by monitoring the evolution of the absorption spectra, the changes in concentration of the photoproduced intermediates and hence the kinetics of the processes in which they are involved are elucidated.

Flash photolysis has evolved subsequently to make use of laser sources and sophisticated electronic detection apparatus to push the limits of time resolution to the femtosecond regime. Indeed, recently the Nobel prize for chemistry was awarded to Ahmed Zewail for his work with ultrafast pump-probe techniques. However, in its conventional geometry, flash photolysis is limited to transparent samples, since it is necessary to be able to probe the excited species by monitoring the absorption spectra.

Many biological systems and industrially important samples are opaque or highly scattering, and hence the attenuation of light through the sample is no longer described by the Beer–Lambert law. In 1981, Frank Wilkinson and Rudolph Kessler had the idea of using diffusely reflected light to interrogate the changes in concentration within a scattering sample subjected to a high-intensity excitation pulse. When photons enter a sample, they may be absorbed or scattered. Those which are scattered may re-emerge from the irradiated surface as diffusely reflected light. The intensity of diffusely reflected light emerging from the surface at a particular wavelength is a unique function of the ratio of scattering to absorption. The more scattering events occurring before absorption, the more likely the photon is to escape from the sample as diffusely reflected light. Hence the probability of escape decreases as the absorption probability increases, and the diffusely reflected light is deficient in those wavelengths where the absorption is strongest, i.e., the ratio of the incident to absorbed light intensity at a given wavelength is related to the absorption of the sample at that wavelength.

Kubelka–Munk Theory of Reflectance

The theory which describes the relationship between incident and scattered light intensity, absorption and scatter, and concentration which is widely applied in this context, is the Kubelka–Munk theory of reflectance. The theory was originally developed to describe the reflectance characteristics of paint films but it works quite well for many samples containing a homogeneous distribution of scatterers and absorbers. The limiting assumption in this theory is that the scatterers from which the scattering layer is composed are very much smaller than the total layer thickness. Additionally, the layer should be optically thick such that all of the light entering the layer should be either absorbed or reflected, with a negligible fraction transmitted.

For a layer of thickness X diffusely irradiated with monochromatic light, the diagram shown in Figure 1

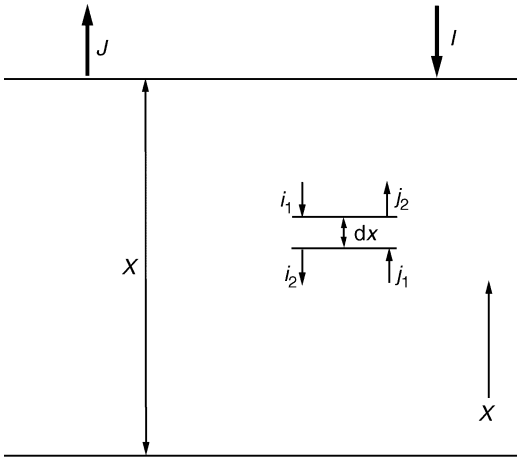


Figure 1 Schematic of counterpropagating fluxes in a diffusely irradiated scattering medium.

can be constructed, with I the incident flux and J the diffusely reflected flux, and i and j the fluxes traveling upwards and downwards through an infinitesimally small thickness element dx. Two further parameters may be defined which are characteristic of the medium.

K The absorption coefficient. Expresses the attenuation of light due to absorption per unit thickness.
S The scattering coefficient. Expresses the attenuation of light due to scattering per unit thickness.

Both of these parameters can be thought of as arising due to the particle (or chromophore) acting as a sphere of characteristic size, which casts a shadow either due to the prevention of on-axis transmission or due to absorption. The scattering or absorption coefficient then depends on the effective size of this sphere, and the number density of spheres in the medium. In each case, the probability P of a photon being transmitted through a particular thickness X of a medium is related exponentially to the absorption or scattering coefficient:

$$P = \exp(-KX) \qquad [3]$$

$$P = \exp(-SX) \qquad [4]$$

The scattering coefficient S depends on the refractive index difference between the particle and the dispersion medium. The scattering coefficient is also dependent upon the particle size, and shows an inverse correlation, i.e., the scattering coefficient increases as the particle size decreases. This effect is a function of the size of the particle relative to the wavelength of light impinging on it, with the scattering coefficient increasing as the wavelength decreases. However, this change with wavelength is small provided the particle size is large relative to the wavelength of the light.

The effect of the material in the element dx on the counterpropagating fluxes i and j depends on the absorption and scattering coefficients. Both i and j will be attenuated by both absorption and scattering:

$$i_2 = i_1 - i_1(S + K)\mathrm{d}x \qquad [5]$$

$$j_2 = j_1 - j_1(S + K)\mathrm{d}x \qquad [6]$$

Both i and j are reinforced by backscattering from the other flux:

$$i_2 = i_1 + j_1 S\mathrm{d}x \qquad [7]$$

$$j_2 = j_1 + i_1 S\mathrm{d}x \qquad [8]$$

The net effect of this in the attenuation of the flux propagating into the sample (i) and the flux backscattered from the sample (j) can be expressed as the following differential equations:

$$\mathrm{d}i = (S + K)i_1\mathrm{d}x - j_1 S\mathrm{d}x \qquad [9]$$

$$\mathrm{d}j = -(S + K)j_1\mathrm{d}x + i_1 S\mathrm{d}x \qquad [10]$$

For a sample of infinite thickness, these equations can be solved to give an analytical solution for the observed reflectance of the sample in terms of the absorption and scattering coefficients:

$$\frac{K}{S} = \frac{(1 - R_\infty)^2}{2R_\infty} \qquad [11]$$

where R_∞ is the reflectance from a sample of such thickness that an increase in sample thickness has no effect on the observed reflectance. The absorption coefficient K is dependent upon the concentration of absorbers in the sample through

$$K = 2\varepsilon c \qquad [12]$$

with ε the naperian absorption coefficient, c the concentration of absorbers and the factor 2 which is the geometrical factor for an isotropic scatterer. For multiple absorbers, K is simply the linear sum of absorption coefficients and concentrations. Hence for a diffusely scattering medium, an expression analogous to the Beer–Lambert law can be derived to relate concentration to a physically measurable parameter, in this case the sample reflectance. The ratio K/S is usually referred to as the Kubelka–Munk remission function, and is the parameter usually quoted in this context. It is important to note that the relationship

with concentration is only valid for a homogeneous distribution of absorbers within a sample.

Transient Concentration Profiles

As has been discussed in the previous section, a light flux propagating through a scattering medium is attenuated by both scattering and absorption events, whilst in a nonturbid medium attenuation is by absorption only. Hence even in a sample with a very low concentration of absorbers, the light flux is rapidly attenuated as it penetrates into the sample by scattering events. This leads to a significant flux gradient through the sample, resulting in a reduction in the concentration of photo-induced species as penetration depth into the sample increases. There are three distinct concentration profiles which may be identified within a scattering sample which are dependent on the scattering and absorption coefficients. These concentration profiles are interrogated by a beam of analyzing light, and hence an understanding of the effect of these differing profiles on the diffusely reflected intensity is vital in interpreting transient absorption data. The transient depth profiles are illustrated in Figure 2.

Kubelka–Munk Plug

This occurs when the photolysis flash, i.e., laser fluence is high and the absorption coefficient is low at the laser wavelength. If we assume a simple photophysical model involving simply the ground state S_0, the first excited singlet state S_1 and first excited triplet state T_1, and we assume that either the quantum yield of triplet state production is high or the S_1 lifetime is very much shorter than the laser pulse allowing re-population of the ground state, then it is possible at high enough fluence to completely convert all of the S_0 states to T_1 states within the sample. Provided the fluence is high enough, this complete conversion will penetrate some way into the sample (Figure 2). Under circumstances where the T_1 state absorbs strongly at some wavelength other than the photolysis wavelength, probing at this wavelength will result in the probe beam being attenuated significantly within a short distance of the surface of the sample, and thus will only interrogate regions where there is a homogeneous excited state concentration. Under these circumstances the reflectance of the sample as a function of the concentration of excited states is described by the Kubelka–Munk equation, and the change in remission function can be used to probe the change in excited state concentration.

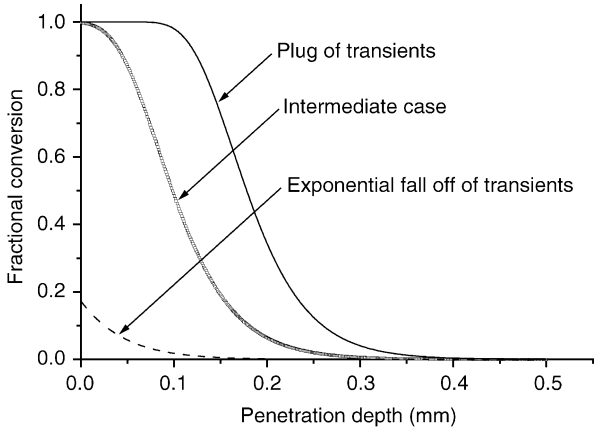


Figure 2 Transient concentration profiles following pulsed laser excitation of a scattering and absorbing sample.

Exponential Fall-Off of Concentration

This occurs when the laser fluence is low and the absorption coefficient at the laser wavelength high. Again considering the simple photophysical model, most of the laser flux will be absorbed by the ground state in the first few layers of sample, and little will penetrate deeply into the sample. In the limiting case this results in an exponential fall-off of transient concentration with penetration depth into the sample (Figure 2). Here the distribution of absorbers is not random and the limiting Kubelka–Munk equation is no longer applicable since the mean absorption coefficient varies with sample penetration depth. Lin and Kan solved eqs [9] and [10] with the absorption coefficient K varying exponentially with penetration depth and showed that the series solution converges for changes in reflectance of less than 10% such that the reflectance change is a linear function of the number of absorbing species.

Intermediate Case

Between the two extremes described above is a case where significant transient conversion takes place at the front surface, but with little penetration into the sample. This can occur, for example, with high laser fluences and large ground state absorption coefficients at the laser wavelength. Under these circumstances, illustrated in Figure 2, the analyzing light interrogates not only the transient concentration profile, but also a significant amount of the analyzing light may penetrate through the transient layer into the unconverted sample behind, if the transient absorption at the analyzing wavelength is low. This creates a more complex problem for analysis since effectively the sample is irradiated from both front and back faces, with consequent effects on the diffusely reflected intensity. It is possible to numerically model the reflectance change as a function of

transient concentration under these circumstances but a precise knowledge of the absorption and scattering coefficients is required. Under most circumstances, this case is avoided and diffuse reflectance flash photolysis experiments are arranged such that one of the two limiting cases above, generally the exponential fall-off (usually achieved by attenuation of the laser beam), prevails in a particular experiment.

Sample Geometry

In the case of conventional nanosecond laser flash photolysis, it is generally the case that right-angled geometry is employed for the photolysis and probe beams (Figure 3a). This geometry has a number of advantages over alternatives. The photolysis beam and analyzing beam are spatially well separated, such that the analyzing beam intensity is largely unaffected by scattered photolysis light. Also, the fluorescence generated by the sample will be emitted in all directions, whilst the analyzing beam is usually collimated allowing spatial separation of fluorescence and analyzing light; sometimes an iris is used to aid this spatial discrimination. This geometry is appropriate for quite large beam diameters and fluences; where smaller beams are used collinear geometry may be more appropriate in order to achieve long interaction pathlengths.

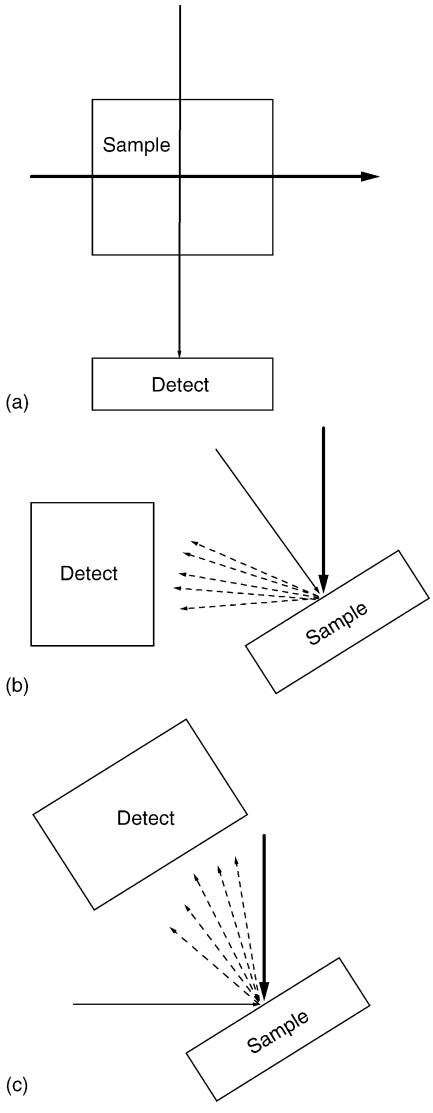


Figure 3 Sample geometries for (a) conventional and (b,c) diffuse reflectance flash photolysis. ⟶, photolysis beam; ⟶, analyzing beam.

In the case of nanosecond diffuse reflectance flash photolysis, the geometry required is quite different (Figure 3b,c). Here, the photolysis beam and analyzing beam must be incident on the same sample surface, and the diffusely reflected analyzing light is collected from the same surface. The geometry is often as shown in Figure 3b, where the analyzing light is incident almost perpendicularly on the sample surface, with the photolysis beam incident at an angle such that the specular reflection of the photolysis beam passes between detector and analyzing beam (not shown). Alternatively, the geometry shown in Figure 3c may be employed, where diffusely reflected light is detected emerging perpendicular to the sample surface. In both cases, the geometry is chosen such that specular (mirror) reflection of either exciting or analyzing light from the sample is not detected, since specular reflection is light which has not penetrated the sample and therefore contains no information regarding the concentrations of species present. A requirement, as in conventional flash photolysis, is that the analyzing beam probes only those areas which are excited by the photolysis beam, requiring the latter to be larger than the former. The nature of the scattering described previously means that the required geometry does not give spatial discrimination at the detector between photolysis and analyzing light, and fluorescence. This is since both analyzing and photolysis light undergo scattering and absorption processes (although with wavelength-dependent absorption and scattering coefficients) and emerge with the same spatial profiles. Fluorescence, principally that stimulated by the photolysis beam since this is of greatest intensity, originates from within the sample but again undergoes absorption and scattering and emerges with the same spatial distribution as the exciting light. In diffuse reflectance flash photolysis, separation of the analyzing light from the excitation or fluorescence must be achieved using spectral (filters and/or monochromators) or temporal, rather than spatial discrimination. Time-gated charge coupled devices

(CCD) or photodiode array detectors can be effectively used in nanosecond laser flash photolysis to exclude excitation light and fluorescence, since these occur on time-scales usually much shorter than the transient absorption of species of interest. The usual analyzing light source used in nanosecond diffuse reflectance flash photolysis is a xenon arc lamp, due to its good spectral coverage and high intensity. High probe intensity is particularly important in diffuse reflectance flash photolysis since the scattered intensities are often low, and the scattered light emerges in a large solid angle, only part of which can be effectively collected and detected. Also, high intensities allow for the light from the analyzing source to dominate over fluorescence if the latter is relatively weak.

In femtosecond diffuse reflectance laser flash photolysis, sample geometry considerations are also important. Such experiments are performed using the pump-probe technique, with the probe often being a femtosecond white-light continuum. The sample geometry employed is illustrated in Figure 4.

Here the pump and probe are almost colinear, and are incident on the same sample area; again, the requirement is for the pumped area to be larger than the probed area. Diffusely reflected light is then collected and analyzed, time resolution being achieved by varying the delay between pump and probe beams. It should be noted that the temporal resolution is worse than in conventional pump-probe techniques. In conventional femtosecond flash photolysis, the time resolution is generally limited by the widths of pump and probe pulses; in diffuse reflectance mode, the pulses undergo numerous refractions, reflections, and diffractions such that the pulse is temporally broadened during its transit through the material. The extent of this broadening is a sensitive function of the optical properties of the individual sample.

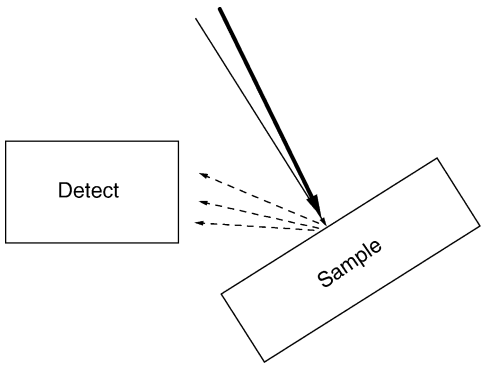


Figure 4 Sample geometry for femtosecond diffuse reflectance flash photolysis. ⟶, photolysis beam; ⟶, analyzing beam.

Kinetic Analysis

Kinetic analysis, and of course time-resolved spectroscopic analysis, require a quantitative treatment of the concentration changes within a sample following an excitation pulse as a function of time. When studying transient absorption phenomena in opaque samples, it is usual to define the reflectance change in terms of the sample reflectance before and after excitation, such that in spectrally resolved data a transient difference spectrum rather than an absolute reflectance spectrum of the transient species is obtained. The latter can, however, be reconstructed from a knowledge of the absorption coefficients and concentration of the species involved. It is possible to define the reflectance change as

$$\frac{\Delta J_t}{J_0} = 1 - \frac{R_t}{R_0} \qquad [13]$$

where R_0 and R_t represent the intensity of probe light diffusely reflected from the sample before excitation and at a time t after excitation, respectively. Frequently reflectance change is expressed as '% absorption', which is defined in eqn [14].

$$\%\ \text{absorption} = 100 \times \left(1 - \frac{R_t}{R_0}\right) \qquad [14]$$

These parameters are often used as being proportional to transient concentration, subject to satisfying the criteria for an exponential fall-off of transient concentration with penetration depth as discussed previously, and are used to replace transient concentration in kinetic equations used in data analysis.

It is generally the case that the samples studied using diffuse reflectance laser flash photolysis have either some degree of heterogeneity, for example paper, microcrystalline cellulose, silica gel or alumina, or have well-defined porous structures such as zeolites. Molecules adsorbed to these supports may be present on the surface, within micro- or mesopores, or intimately included within the structure. Hence each molecule may experience its own unique environment, and this will obviously influence its observed photophysics. It is therefore the case that in these systems, even very simple photo-induced reactions such as unimolecular photoisomerizations do not follow first-order kinetics; rather, a distribution of rates is observed which reflect the differing environments experienced by the molecules, and hence the molecules act as probes of this heterogeneity.

There are a number of approaches to kinetic analysis in opaque, often heterogeneous systems, as described below.

Rate Constant Distributions

Here the sample heterogeneity is treated as producing a series of micro-environments, and the reaction studied will have its own unique rate constant in each of these environments. The width of the observed rate constant distribution therefore describes the number of these possible environments, with the distribution amplitude at any particular rate constant, reflecting the probability of the molecule existing in the corresponding environment.

Exponential series lifetime distribution analysis is the simplest approach to distributed lifetime analysis. This analysis has the inherent advantage that there are no presuppositions regarding the kinetic model describing the data. Rather, a large number of exponentials with fixed rate constants and with amplitudes as adjustable parameters are applied to the data, and a least-squares procedure used to optimize the amplitude of each exponential. Generally there is a relationship between the rate constants, they being equally spaced on either a linear or logarithmic scale. Physically this can be interpreted as a large number of first-order or pseudo-first-order (in the case of bimolecular reactions) rates arising from the sample heterogeneity. Hence the rate constant distribution emerges naturally from the fitting procedure, with no preimposed constraints. Since there are a large number of adjustable parameters, these models are applied most successfully to data with good signal-to-noise.

Alternative approaches involve imposing a distribution shape onto a set of exponentials, and optimizing this distribution to the data. This approach has the advantage that by assuming a rate constant distribution, the relative amplitudes of the exponentials are fixed and hence the number of fitting parameters greatly reduced. One of the more successful models applied in this context is that developed by Albery *et al*. In the development of their model it is assumed that, for the reaction in question, the free energy of activation, $\Delta G^{\neq}$, is distributed normally around a mean value $\Delta \bar{G}^{\neq}$ according to eqn [15]:

$$\Delta G^{\neq} = \Delta \bar{G}^{\neq} - \gamma x RT \qquad [15]$$

with γ being the width of the distribution for $\infty \geq x \geq -\infty$. This assumed normal distribution of the free energy of activation leads to a log-normal distribution of the decay rate constants distributed around some average rate constant $\bar{k}$. The dispersion in the first-order rate constant is then

$$\ln(k) = \ln(\bar{k}) + \gamma x \qquad [16]$$

The equation used to describe the data is given as:

$$\frac{c}{c_0} = \frac{\int_{-\infty}^{\infty} \exp(-t^2) \exp\left[- \bar{k}t \exp(\gamma t)\right] \mathrm{d}t}{\int_{-\infty}^{\infty} \exp(-t^2) \mathrm{d}t} \qquad [17]$$

with c and c_0 being concentration at time $t = t$ and $t = 0$ relative to the excitation flash, respectively. Where reflectance change is proportional to transient concentration, reflectances can be used directly in place of concentrations. This equation can be solved by making the following substitutions:

$$\int_{-\infty}^{\infty} \exp(-t^2) \mathrm{d}t = \pi^{1/2} \qquad [18]$$

$$\int_{-\infty}^{\infty} \exp(-t^2) \exp\left[-\bar{k}t \exp(\gamma t)\right] \mathrm{d}t = \int_0^1 \lambda^{-1} \exp\Big(-(\ln \lambda^2)\Big)\Big(\exp(-\bar{k}t\lambda^{\gamma}) + \exp(-\bar{k}t\lambda^{-\gamma})\Big) \mathrm{d}\lambda \qquad [19]$$

Hence here there are only two fitting parameters, namely the width of the distribution γ and the distribution center, $\bar{k}$. Note that for $\gamma = 0$, eqn [17] reduces to a single exponential function.

A further model which constrains the rate constant distribution, which has been successfully applied to describe the rates of photo-induced processes on silica gel surfaces, is a Lévy stable distribution of rate constants, described as:

$$\mathrm{P}(\bar{k}, \alpha, \gamma) = \frac{1}{\pi} \int_0^{\infty} \exp(- \gamma q^{\alpha}) \cos(\bar{k}\mathrm{q}) \mathrm{dq} \qquad [20]$$

where α is the characteristic power law exponent ($0 < \alpha \leq 2$) and γ is the distribution width ($\gamma > 0$). Special cases of the Lévy stable distribution occur for $\alpha = 1$ and $\alpha = 2$, where the distribution shape becomes Lorentzian and Gaussian, respectively. The Lévy distribution gives an increasing weighting to the tails of the distribution as the distribution width decreases, and can be described as a random walk consisting of long jumps followed by several short walks. It has been shown that this type of foraging behavior is more efficient at seeking out randomly distributed targets than a simple random walk. The Lévy stable distribution has three adjustable parameters which allows greater flexibility in the distribution of rate constants than does the Gaussian model, but still constrains the distribution to be symmetrical about some mean value.

Physical Models

A number of these have been applied to very specific data sets, where parameters of the systems are accurately known. These include random walk models in zeolites, and time-dependent fractal-dimensional rate constants to describe kinetics on silica gel surfaces. The advantage of these methods over rate constant distributions is that since they are based on the physical description of the system, they yield physically meaningful parameters such as diffusion coefficients from analysis of kinetic data. However, for many systems the parameters are known in insufficient detail for accurate models to be developed.

Examples

An example of bimolecular quenching data is shown in Figure 5. Here, anthracene at a concentration of 1 μmol g^{-1} is co-adsorbed to silica gel from acetonitrile solution together with azulene at a concentration of 0.8 μmol g^{-1}. Laser excitation at 355 nm from an Nd:YAG laser produces the excited triplet state of the anthracene, which undergoes energy transfer to the co-adsorbed azulene molecule as a result of the rapid diffusion of the latter.

The data shown in Figure 5 are recorded monitoring at 420 nm, and the laser energy (approximately 5 mJ per pulse) is such that an exponential fall-off of transient concentration with penetration depth is expected such that reflectance change is proportional to transient concentration (see section on Transient Concentration Profiles above). The data are shown on a logarithmic time axis for clarity. The fitted line is obtained by applying the model of Albery *et al.*

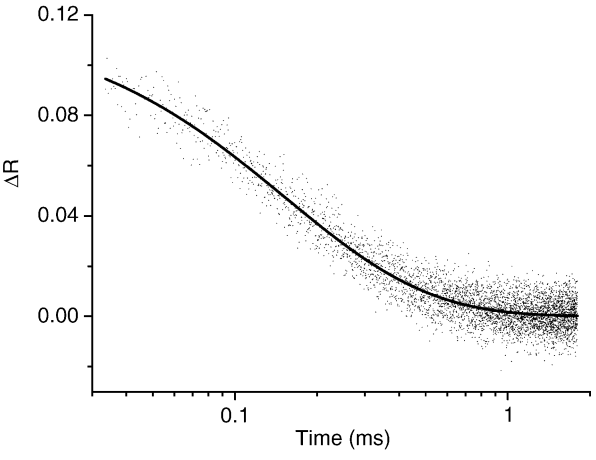


Figure 5 Transient absorption decay for anthracene (1 μmol g^{-1}) co-adsorbed with azulene (0.8 μmol g^{-1}) to silica gel monitored at 420 nm. Fitted using the model of Albery *et al.*

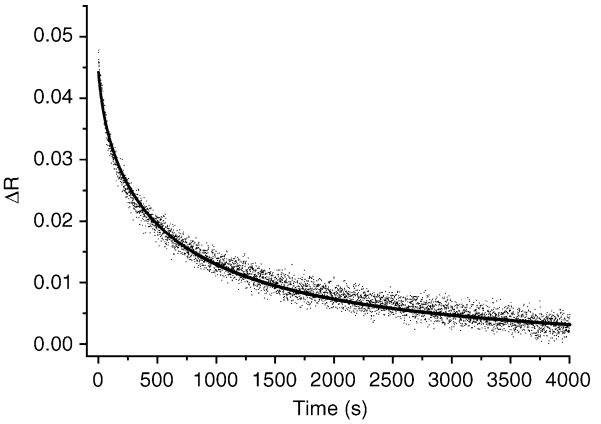


Figure 6 Transient absorption decay for the naphthalene (1 μmol g^{-1}) radical cation monitored at 680 nm on silica gel. Fitted using a Lévy stable distribution.

(see section on Rate Constant Distributions above) with fitting parameters $\bar{k} = 1.01 \times 10^4$ s^{-1} and $\gamma = 0.73$.

Where ion–electron recombination is concerned, the Albery model often fails to adequately describe the data obtained since it does not allow sufficient small rate constants relative to the value of $\bar{k}$, given the constraints of the Gaussian distribution. The Lévy stable distribution is ideal in this application since it allows greater flexibility in the shape of the distribution.

Figure 6 shows example data for naphthalene adsorbed to silica gel (1 μmol g^{-1}). The laser pulse energy at 266 nm (approximately 40 mJ per pulse) is such that photo-ionization of the naphthalene occurs producing the radical cation. The subsequent decay of the radical cation via ion–electron recombination can be monitored at 680 nm. Note that decay is observed on a time-scale of several thousand seconds. The fitted line is according to a Lévy stable distribution with parameters $\bar{k} = 8.2 \times 10^{-4}$ s^{-1}, $\gamma = 0.5$ and $\alpha = 1.7$ (see section on Rate Constant Distributions above). This model allows the shape of the distribution to deviate from a Gaussian, and can be more successful than the model of Albery *et al.* in modeling the complex kinetics which arise on surfaces such as silica gel. Note that where the model of Albery *et al.* can successfully model the data, the data can also be described by a Lévy stable distribution with $\alpha = 2$.

List of Units and Nomenclature

A	Sample absorbance
c	Concentration
I	Incident flux
J	Diffusely reflected flux

k	Rate constant
$\bar{k}$	Mean rate constant
K	Absorption coefficient (absorption per unit thickness) [cm^{-1}]
l	Pathlength [cm]
P	Probability of transmission of a photon through a sample of defined thickness
R_∞	Reflectance of an infinitely thick sample
S	Scattering coefficient (scattering per unit thickness) [cm^{-1}]
t	Time
X	Thickness of sample layer [cm]
α	Characteristic Lévy power law exponent
γ	Width of rate constant distribution
$\Delta G^{\neq}$	Free energy of activation [kJ mol^{-1}]
$\Delta \bar{G}^{\neq}$	Mean free energy of activation [kJ mol^{-1}]
ε	Molar decadic or naperian absorption coefficient

See also

Chemical Applications of Lasers: Pump and Probe Studies of Femtosecond Kinetics. **Optical Materials**: Measurement of Optical Properties of Solids. **Scattering**: Scattering from Surfaces and Thin Films; Scattering Theory.

Further Reading

Albery WJ, Bartlett PN, Wilde CP and Darwent JR (1985) A general-model for dispersed kinetics in heterogeneous systems. *Journal of the American Chemical Society* 107: 1854–1858.

Anpo M (ed.) (1996) Surface photochemistry. *Wiley Series in Photoscience and Photoengineering*, vol. 1. Chichester, UK: John Wiley and Sons.

Anpo M and Matsuura T (eds) (1989) Photochemistry on solid surfaces. *Elsevier Series in Studies in Surface Science and Catalysis*, vol. 47. Amsterdam, The Netherlands: Elsevier Science Publishers B.V.

Asahi T, Furube A, Fukumura H, Ichikawa M and Masuhara H (1998) Development of a femtosecond diffuse reflectance spectroscopic system, evaluation of its temporal resolution, and applications to organic powder systems. *Review of Scientific Instruments* 69: 361–371.

Bertoin J (1998) *Lévy Processes*. Cambridge, UK: Cambridge University Press.

Hapke B (1993) *Introduction to the Theory of Reflectance and Emittance Spectroscopy*. Cambridge, UK: Cambridge University Press.

Kamat PV (1993) Photochemistry on nonreactive and reactive (semiconductor) surfaces. *Chemical Reviews* 93: 287–300.

Kan HKA and Lin TP (1970) Calculation of reflectance of a light diffuser with non-uniform absorption. *Journal of the Optical Society of America* 60: 1252–1256.

Kessler RW and Wilkinson F (1981) Diffuse reflectance triplet–triplet absorption spectroscopy of aromatic hydrocarbons chemisorbed on γ-alumina. *Journal of the Chemical Society – Faraday Transactions I* 77: 309–320.

Kortum G (1969) *Reflectance Spectroscopy*. Berlin: Springer-Verlag.

Kubelka P and Munk F (1931) Ein beitrag zur optik der farbanstriche. *zeitschrift fuer Technische Physik* 12: 593–601.

Ramamurthy V (ed.) (1991) *Photochemistry in Organised & Constrained Media*. New York: VCH Publishers.

Thomas JK (1993) Physical aspects of photochemistry and radiation chemistry of molecules adsorbed on SiO_2, γ-Al_2O_3, zeolites and clays. *Chemical Reviews* 93: 301–320.

Thomas JK and Ellison EH (2001) Various aspects of the constraints imposed on the photochemistry of systems in porous silica. *Advances in Colloid and Interface Science* 89–90: 195–238.

Wendlandt WW and Hecht HG (1966) Reflectance Spectroscopy. *Chemical Analysis Series*, vol. 21. New York: Interscience Publishers.

Wilkinson F and Kelly G (1989) Diffuse reflectance flash photolysis. In: Scaiano JC (ed.) *Handbook of Organic Photochemistry*, vol. 1, pp. 293–314. Florida: CRC Press.

Laser Manipulation in Polymer Science

S Ito, Y Hosokawa and H Masuhara, Osaka University, Osaka, Japan

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Laser manipulation is a method for manipulating single particles with a size of less than a few tens of microns in which optical pressure of a focused laser beam is used to trap and control the particles without any mechanical contact. As infrared lasers, as a trapping light source, have become more user-friendly, even operators who are not familiar with lasers and microscopes can perform laser manipulation. Those laser manipulators already on sale have attracted significant attention as a new tool. It is especially interesting to combine this method with nanotechnology and biotechnology, which have progressed rapidly during the 1990s. In this article,

the principle and method of laser manipulation is described and then its applications and possibilities in nanotechnology and biotechnology are summarized.

Principle and Method

When light is reflected or refracted at the interface between two media with different refractive indices, the momentum of photons is changed, which leads to the generation of photon force as a reaction to the momentum change. For example, when a laser beam is reflected by a mirror, the momentum of photons is changed by Δp (Figure 1a). As a result, a photon force, F_{phot}, acts on the mirror to deflect it vertically away from the reflected beam. In refraction, the momentum of photons is also changed, so that the photon force acts on the interface, as shown in Figure 1b. Thus light does not only give its energy to materials via absorption but also applies a mechanical force to them. However, when we are exposed to light, such as from a halogen lamp, we are never aware of the photon force because its magnitude is less than an order of pN. However, the photon force of the pN order acquired from a focused laser beam is sufficient to manipulate nm–μm-sized particles in solution, under an optical microscope.

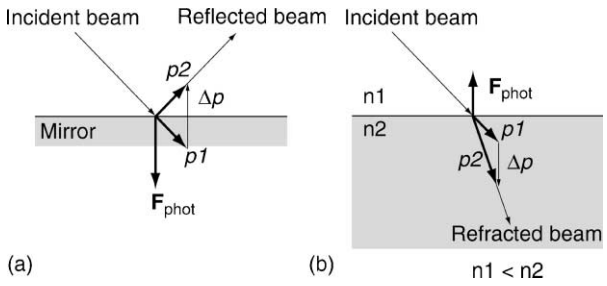


Figure 1 Optical pressure originating from the momentum change of photons.

Principle of Laser Trapping

If the size of a particle is larger than the wavelength of a trapping laser beam (μm-sized particles), the principle of single-beam gradient force optical trapping can be explained in terms of geometrical optics. When a tightly focused laser beam is irradiated onto a transparent dielectric particle, an incident laser beam is refracted at the interface between the particle and medium, as represented by Figure 2a. The propagation direction of the beam is changed, i.e., the momentum of the photon is changed, and consequently, a photon force is generated. As the laser beam leaves the particle and enters the surrounding medium, this refraction causes a photon force to be exerted again on that interface. Summing up the force contributions of all rays, if the refractive index of the particle is higher than that of the medium, a resultant force exerted on the particle is directed toward the focal point as an attractive force. However, reflection at the surface of the particle is negligible if that particle is transparent and the refractive index ratio of the particle to medium is close to unity. Therefore, the incident beam is reflected at two surfaces by a small amount, and as a result, the particle is directed slightly to the propagation direction of the incident light. Where the particle absorbs the incident beam, the photon force is also generated to push it in the propagation direction. The negligible effects of the

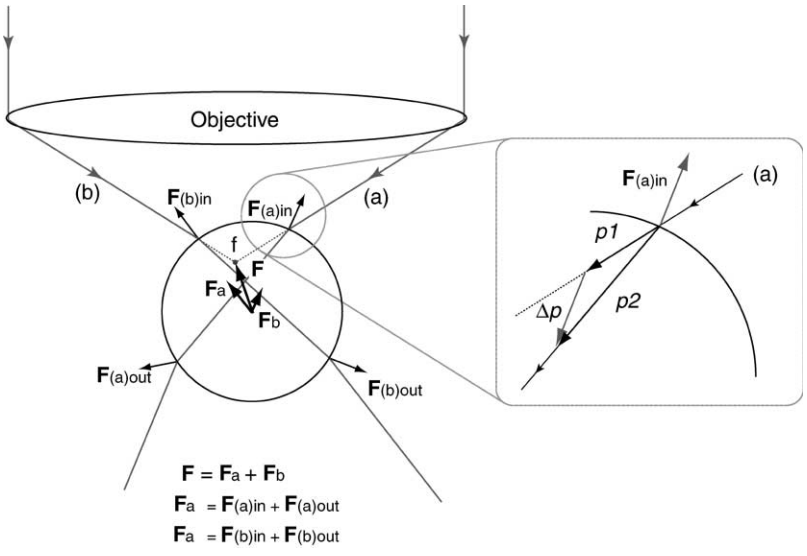


Figure 2 Principle of single-beam gradient force optical trapping explained by ray optics.

reflection and absorption occur in trapping experiment of transparent particles, such as polymer particles, silica microspheres, etc. However, particles with high reflectance and absorption coefficients such as metallic particles, exert a more dominant force at the surface where absorption and reflection occurs. This force is a repulsive force away from the focal point; consequently, metallic microparticles cannot be optically trapped.

If a dielectric particle is much smaller than the wavelength of a trapping light (nm-sized), it can be regarded as a point dipole (Rayleigh approximation) and the photon force ($\mathbf{F}_{\text{phot}}$) acted on it is given by

$$\mathbf{F}_{\text{phot}} = \mathbf{F}_{\text{grad}} + \mathbf{F}_{\text{scat}} \qquad [1]$$

Here, $\mathbf{F}_{\text{grad}}$ and $\mathbf{F}_{\text{scat}}$ are called the gradient force and scattering force, respectively. The scattering force is caused by the scattering of light, and it pushes the particle toward the direction of light propagation. On the other hand, the gradient force is generated when a particle is placed in a heterogeneous electric field of light. If the dielectric constant of the particle is higher than that of the surrounding medium, the gradient force acts on the particle to push it toward the higher intensity region of the beam. In the case of laser trapping of dielectric nanoparticles, the magnitude of gradient force is much larger than the scattering force. Consequently, the particle is trapped at the focal point of the trapping laser beam, where the beam intensity (electric field intensity) is maximum. The photon force can be expressed as follows:

$$\mathbf{F}_{\text{phot}} \cong \mathbf{F}_{\text{grad}} = \frac{1}{2}\varepsilon_{\text{m}}\alpha\nabla|\mathbf{E}|^2 \qquad [2]$$

$$\alpha = 3V\frac{\varepsilon_{\text{p}} - \varepsilon_{\text{m}}}{\varepsilon_{\text{p}} + 2\varepsilon_{\text{m}}} \qquad [3]$$

where, $\mathbf{E}$ is electric field of the light, V is volume of the particle, and ε_{p} and ε_{m} are dielectric constants of the particle and medium. In trapping nanoparticles that absorb the laser beam, such as gold nanoparticles, the complex dielectric constant and attenuation of the electric field in the particle need to be taken into consideration. Although the forces due to light absorption and scattering are both propelling the particle toward the direction of light propagation, the magnitude of gradient force is much larger than these forces. This is in contrast to the μm-sized metallic particle, which cannot be optically trapped.

Laser Manipulation System

An example of a laser micromanipulation system, with a pulsed laser to induce photoreactions, is shown in Figure 3. A linearly polarized laser beam from a CW Nd^{3+}:YAG laser is modulated to the circularly polarized light by a λ/4 plate and then split into horizontally and vertically polarized beams by a

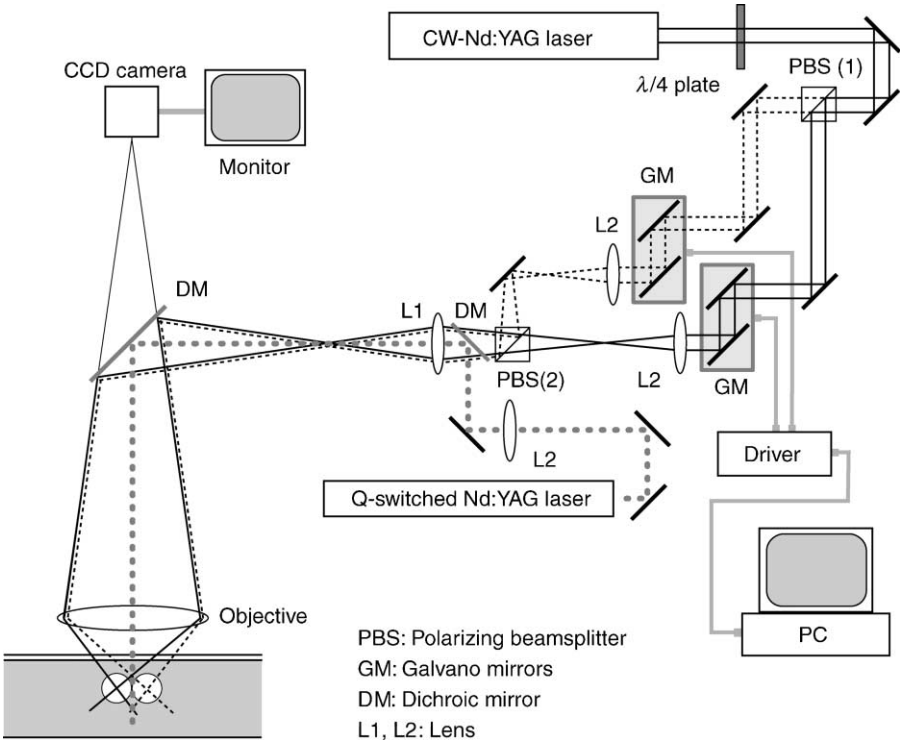


Figure 3 A block diagram of a dual beam laser manipulation-reaction system.

polarizing beamsplitter (PBS1). Both laser beams are combined by another (PBS2), introduced coaxially into an optical microscope, and then focused onto a sample through a microscope objective. The focal spots of both beams in the sample solution can be scanned independently by two sets of computer-controlled galvano mirrors. Even if the beams are spatially overlapping, interference does not take place because of their orthogonal polarization relation. Pulsed lasers, such as a Q-switched Nd^{3+}:YAG laser, are used to induce photoreactions, such as photopolymerization, photothermal reaction, laser ablation, etc.

Laser Manipulation and Patterning of Nanoparticles

Laser manipulation techniques enable us to capture and mobilize fine particles in solution. Most studies using this technique have been conducted on μm-sized objects such as polymer particles, microcrystals, living cells, etc. Because it is difficult to identify individual nanoparticles with an optical microscope, laser manipulation techniques have rarely been applied to nanotechnology and nanoscience. However, the laser trapping technique can be a powerful tool for manipulation of nanoparticles in solution where individual particles can be observed. This is achieved by detection of fluorescence emission from labeled dyes or scattered light. Now, even single molecule fluorescence spectroscopy has been achieved by the use of a highly sensitive photodetector, and a single metallic nanoparticle can be examined by detecting the light scattered from it. There have been several reports on the application of laser manipulation techniques for patterning of nm-sized particles. Here, fixation methods of nanoparticles, using the laser manipulation technique and local photoreactions, are introduced.

Patterning of Polymer Nanoparticles

Patterning of individual polymer nanoparticles onto a substrate can be achieved by using local photopolymerization. The following example shows the strength of this method. Polystyrene nanoparticles of 220 nm with fluorescent dye were dispersed in ethylene glycol solution containing polymerizable vinyl monomer (acrylamide, 31 wt%), crosslinker (*N*,*N*′-methylenebisacrylamide, 2.2 wt%), and radical photoinitiator (Irgacure2959, Ciba Specialty Chemicals, 1.1 wt%). When the sample was irradiated by blue light from a high-pressure mercury lamp, green fluorescence from dye molecules within each nanoparticle was observed. A nanoparticle that entered the region, irradiated by a near-infrared laser beam (1064 nm), was trapped at the focal point, and moved onto the surface of the glass substrate by handling the 3D stage of the microscope. An additional fixation laser beam (355 nm, 0.03 μJ, pulse duration ~6 ns, repetition rate ~5 Hz) was then focused to the same point for ~10 s, which led to the generation of acrylamide gel around the trapped nanoparticle. By repeating the procedure, patterning of single polymer nanoparticles on a glass substrate was achieved, and a fluorescence image of single nanoparticles as a letter 'H', is shown in Figure 4. A magnified atomic force microscope (AFM) image of one of the fixed nanoparticles is also shown, which confirms that only one polymer nanoparticle was contained in the polymerized gel.

Multiple polymer nanoparticles can also be gathered, patterned, and fixed on a glass substrate by scanning both trapping and fixation laser beams with the use of two pairs of galvano mirrors. The optical transmission and fluorescence images of the 'H' patterned nanoparticles on a glass substrate are shown in Figure 5a and b, respectively. The letter 'H' consists of three straight lines of patterned and fixed nanoparticles. The trapping laser beam (180 mW) was scanned at 30 Hz along each line with a length of ~10 μm on a glass substrate for 300 s. Nanoparticles were gathered and patterned along the locus of the focal point on the substrate. Then the fixation laser beam (0.097 μJ) was scanned for another 35 s. As a result, the straight lines of

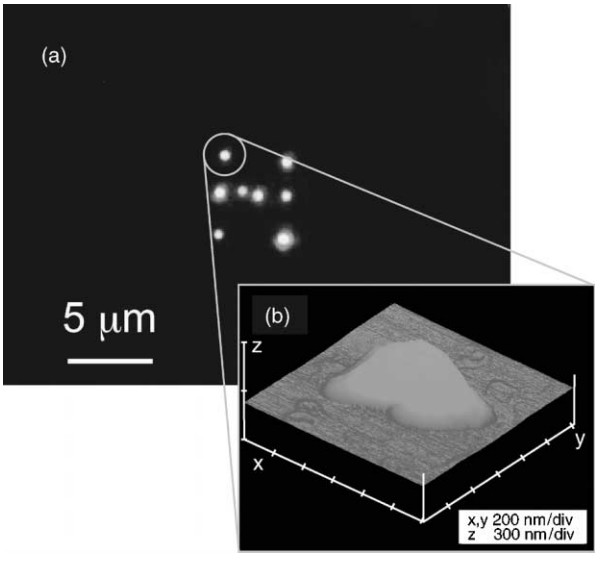


Figure 4 (a) A fluorescence image of spatially patterned individual polymer nanoparticles as the letter 'H' in distilled water. (b) A magnified AFM image of one of the produced acrylamide gels containing only one polymer nanoparticle on a glass substrate in the air.

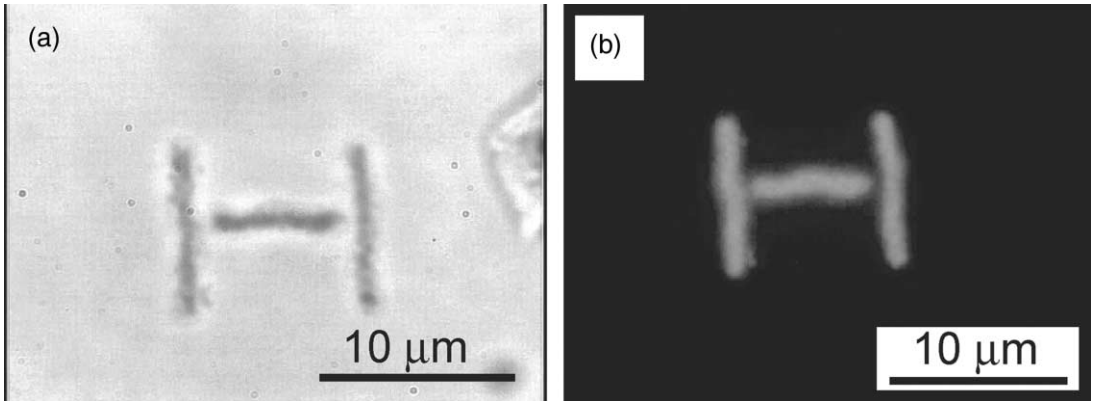


Figure 5 Optical transmission (a) and fluorescence (b) images of patterned and fixed polymer nanoparticles on the glass substrate, as the letter 'H'.

patterned nanoparticles were fixed in generated acrylamide gel on the substrate. Combining several simple fixed patterns, a more complex arrangement can be created with use of the present manipulation and fixation techniques.

Fixation of Individual Gold Nanoparticles

Manipulation and fixation of single metallic nanoparticles in solution has been achieved by means of photo-induced transient melting of nanoparticles. As an example, the following method describes this technique. Gold nanoparticles (diameter ~ 80 nm) were dispersed in ethylene glycol. In order to identify a single gold nanoparticle, extinction of light from the halogen lamp of the optical microscope by the particle was utilized. Thus, the trapped particle was observed as a black spot in transmission image. The optically trapped single gold nanoparticle was then transferred to a precise position on the surface of glass substrate in the sample solution. The focused 355 nm pulse was additionally irradiated onto the pressed nanoparticle, which led to a transient temperature elevation of the gold nanoparticle to enable its fixation. It was confirmed by AFM observation that, at suitable laser fluence ($32-64$ mJ cm^{-2}), a gold nanoparticle was fixed on the glass substrate without fragmentation. Repeating the same manipulation and fixation procedure, single gold nanoparticles could be patterned on a glass substrate. Figure 6 shows an AFM image of successive spatial patterning of single 80 nm gold particles.

The significance of the laser manipulation-fixation technique is that we can trap, manipulate, and fix single and/or many nanoparticles in solution at room temperature. We have already demonstrated assembling entangled polymer chains of $10-20$ nm mean radius by laser trapping, and formation of a single μm-sized particle. Thus, these various kinds of

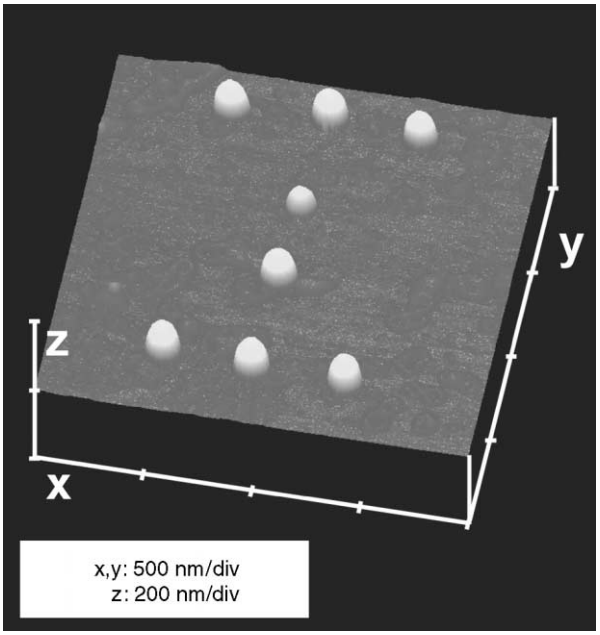


Figure 6 An AFM image of fixed and patterned gold nanoparticles, as the letter I.

nanoscopic materials can be well manipulated using these techniques and it is believed that the present nanomanipulation-fixation method will be useful for future nanoscience and nanotechnology.

Application to Biotechnology

Recent progress in biotechnology, using single cell manipulation by laser and microscope, has been attracting significant attention. In conventional methods, cell manipulation has been performed by mechanical manipulation using microneedles and micropipettes. However, laser manipulation can be applied as microtweezers. In comparison with the former, this form of laser manipulation has the

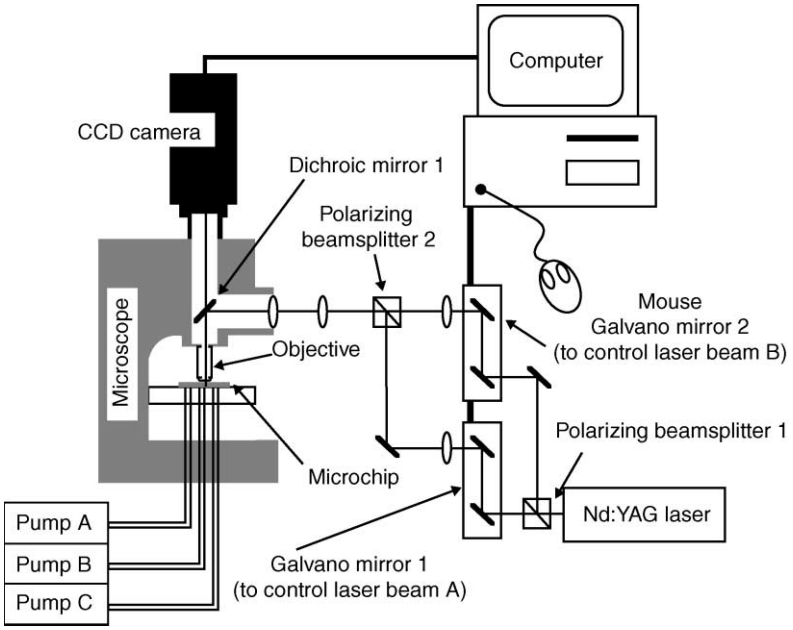


Figure 7 An experimental setup of dual-beam laser manipulation and microfluidic devices.

advantage of no contact with cells and can perform some complex and characteristic movement, for example, rotation, separation, accumulation, etc. Furthermore, by combining it with a microcutter using laser ablation, cell fusion and extraction/injection of organs from/into the cells can be achieved. In this section, single-cell manipulation achieved by using the laser manipulation, is introduced.

Noncontact Rotation of Cells Using Dual Beam Laser Manipulation

As a special feature of laser-cell manipulation, it is noticeable that multiple laser tweezers can be operated independently without any contact. Here, noncontact rotation of the fission yeast, *Schizosaccharomyces pombe* (h^-), demonstrates this method. The yeast has the shape of an ellipse of length 8 and 3 μm for major and minor axes, and is shown set on a dual-beam laser manipulation system in Figure 7. One trapping beam A was focused at the end of the cell to anchor it, and another beam, B, was at the other end and scanned circularly around beam A by controlling the Galvano mirrors. The rotation of the yeast cell was realized at a frequency of less than 2 Hz, as shown in Figure 8. Such a cell manipulation is impossible by mechanical manipulation and indicates the superior performance of laser manipulation.

Transfer of Cells in Microchannel

A flow cytometry to physically separate and identify specific types of cells from heterogeneous populations

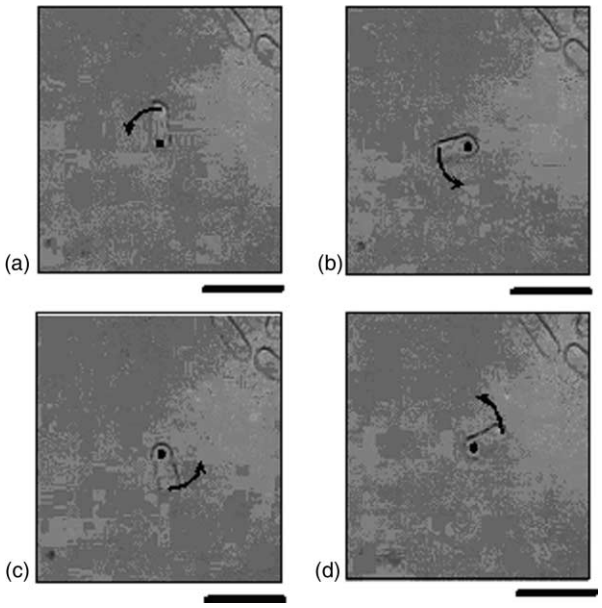


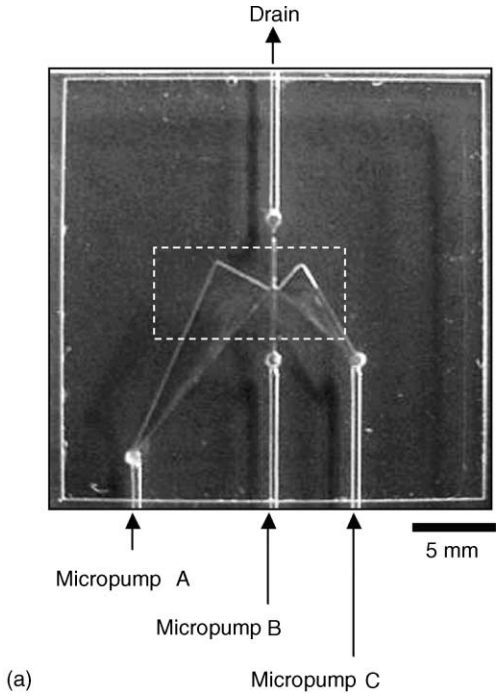
Figure 8 Rotation of fission yeast cells using dual-beam laser manipulation. The bars are 10 μm.

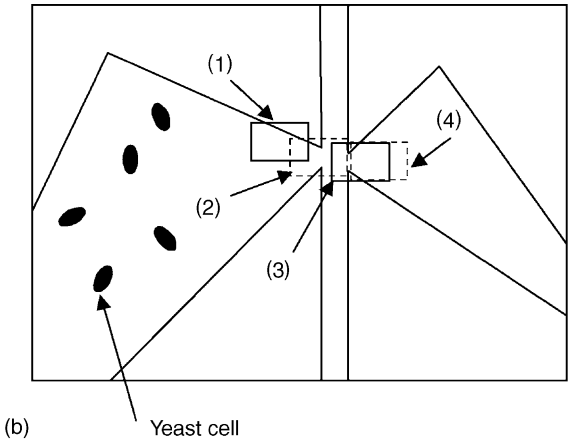
by fluorescence, which is called fluorescence-activated cell sorter (FACS), has attracted significant attention as an important technique in biotechnology. In the separation process, charged single droplets containing single, fluorescence labeled cells are selected by passing them between two high-voltage deflection plates. Since it is a sequential process, which does not use microscopy, flexible and high-purity cell separation is limited. To overcome this problem, a selective cell separation system,

combining the laser trapping with the microchannel, is used. The method using microscopy is useful and flexible, because cells are not simply detected by fluorescence but also identified by their shape, size, surface morphology, absorption, etc. However, cell sorting under a microscope is not processed efficiently by laser trapping. A more effective method is described by the following.

The microchannel was prepared with laser microfabrication and photopolymerization, whose top view is shown in Figure 9a, and was then set on the microscope stage. Three syringes were equipped with homemade pumps A, B, and C, and connected with a microchip. There were two wing-like chambers in the microchip, that were connected with pumps A and C. Between chambers, there was a microchannel 100 μm wide, which joined these chambers and crossed a long microchannel connected with the pump B, thus forming a drain. Thickness of these chambers was 100 μm. By controlling these pumps, the left and right chambers were filled by the culture medium including the yeast cells and the neat culture medium, respectively, as shown in Figure 9b.

Individual cells were transferred from left to right chambers by using a single laser beam and handling an electrically movable microscope stage. The trapping laser irradiated a yeast cell in the left chamber and the trapped cell was transferred to the right chamber. A representative demonstration is given in Figure 9c. The position of the trapped yeast cell was controlled by the microscope stage with the velocity of 20 μm/s. The single yeast cells were successfully transferred from left to right chambers at the rate of 26 s. By combining laser trapping with the microchannel, single separation of cells was achieved under a microscope.





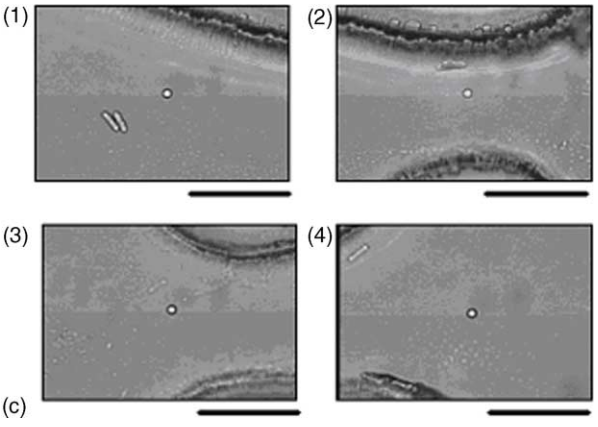


Figure 9 (a) A microchip fabricated by laser polymerization technique and (b) its schematic representation corresponding to dashed area in the microchip. (c) Cell transfer in microchip using laser trapping. The bars are 100 μm. A trapping laser beam is fixed at the center of each picture, by which a particle is trapped and transferred from left to right chambers. The position of each picture is shown in (b).

Collection and Alignment of Cells

When cell sorting using laser trapping is applied to biotechnology, the time to transfer cells should be short compared to the above-mentioned times. One method is to transfer multiple cells simultaneously by developing dual-beam laser manipulation, which is shown schematically in Figure 10. One trapping beam A was scanned on a line given as an arrow in Figure 10a at a rate faster than motions of cells.

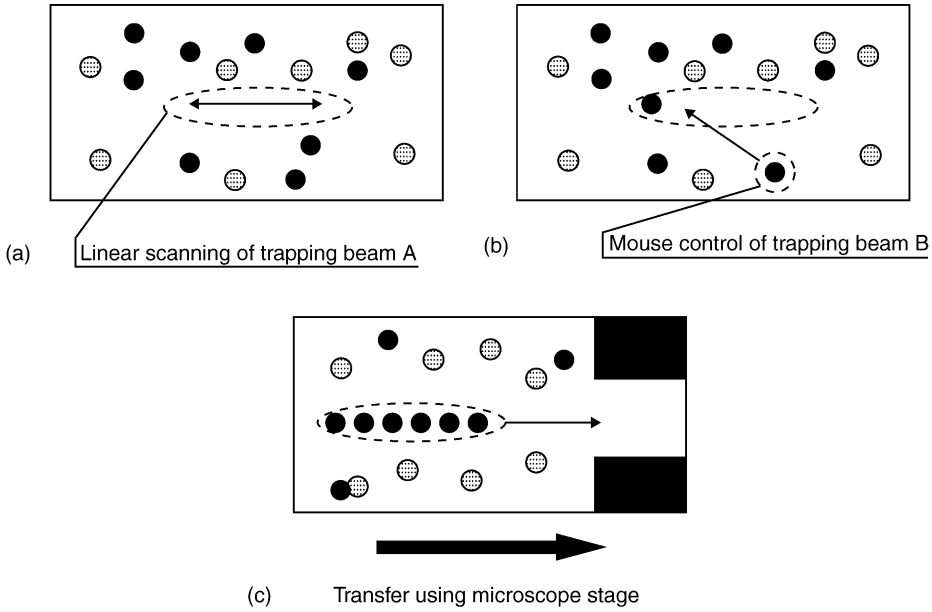


Figure 10 A schematic illustration of efficient collection and transfer of cells based by dual-beam laser manipulation.

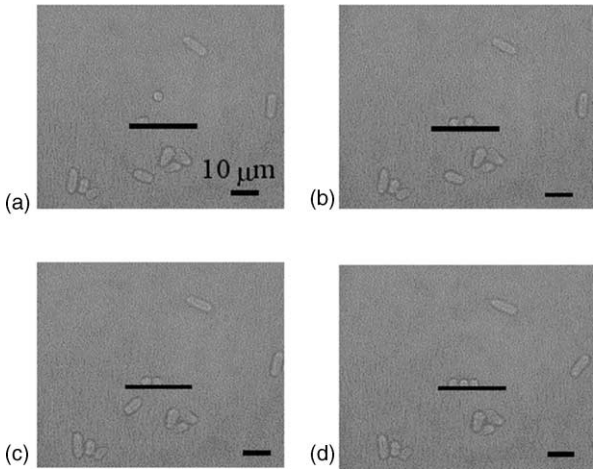


Figure 11 Collection and alignment of fission yeast cells using dual-beam laser manipulation.

The other trapping beam B was used to move individual cells to the line sequentially (Figure 10b). Finally, by moving the microscope stage, many particles trapped at the scanning line can be transferred to another chamber (Figure 10c).

An example is shown in Figure 11. The trapping beam A was scanned on a line with the rate of 15 Hz whose length was 30 μm. The trapping beam B was used to trap a yeast cell and move it to the line, which was controlled by a computer mouse pointer. The cells stored on the line were successfully isolated and with this method three cells can be collected within 15 s. Furthermore, the cells could be transferred with the velocity of 20 μm/s by driving the microscope stage. By improving the presented cell-sorting system, it is expected to realize flexible and high-purity cell separation.

List of Units and Nomenclature

Laser fluence (light energy density per pulse)	[mJ cm^{-2}]

See also

Time-Resolved Fluorescence: Measurements in Polymer Science.

Further Reading

Ashkin A, Dziedzic MJ, Bjorkholm JE and Chu S (1986) Observation of a single-beam gradient force optical trap for dielectric particles. *Optical Letters* 11: 288–290.

Ashkin A, Dziedzic MJ and Yamane T (1987) Optical trapping and manipulation of single cells using infrared laser beams. *Nature* 330: 769–771.

Berns WM, Aist J, Edwards J, *et al.* (1981) Laser microsurgery in cell and developmental biology. *Science* 213: 505–513.

Hoffmann F (1996) Laser microbeams for the manipulation of plant cells and subcellular structures. *Plant Science* 113: 1–11.

Hosokawa Y, Masuhara H, Matsumoto Y and Sato S (2002) Dual-beam laser micromanipulation for sorting

biological cells and its device application. *Proceedings of SPIE* 4622: 138–142.

Ito S, Yoshikawa H and Masuhara H (2001) Optical patterning and photochemical fixation of polymer nanoparticles on glass substrates. *Applied Physics Letters* 78: 2566–2568.

Ito S, Yoshikawa H and Masuhara H (2002) Laser manipulation and fixation of single gold nanoparticles in solution at room temperature. *Applied Physics Letters* 80: 482–484.

Kamentsky AL, Melamed RM and Derman H (1965) Spectrophotometer: new instrument for ultrarapid cell analysis. *Science* 150: 630–631.

Kim BH, Kogi O and Kitamura N (1999) Single-microparticle measurements: laser trapping-absorption microscopy under solution-flow conditions. *Analytical Chemistry* 71: 4338–4343.

Masuhara H, De Schryver FC, Kitamura N and Tamai N (eds) (1994) *Microchemistry: Spectroscopy and Chemistry in Small Domains*. Amsterdam: Elsevier-Science.

Sasaki K, Koshioka M, Misawa H, Kitamura N and Masuhara H (1991) Laser-scanning micromanipulation and spatial patterning of fine particles. *Japanese Journal of Applied Physics* 30: L907–L909.

Schütze K, Clement-Segelwald A and Ashkin A (1994) Zona drilling and sperm insertion with combined laser microbeam and optical tweezers. *Fertility and Sterility* 61: 783–786.

Shikano S, Horio K, Ohtsuka Y and Eto Y (1999) Separation of a single cell by red-laser manipulation. *Applied Physics Letters* 75: 2671–2673.

Svoboda K and Block SM (1994) Optical trapping of metallic Rayleigh particles. *Optical Letters* 19: 930–932.

Nonlinear Spectroscopies

S R Meech, University of East Anglia, Norwich, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Chemistry is concerned with the induction and observation of changes in matter, where the changes are to be understood at the molecular level. Spectroscopy is the principal experimental tool for connecting the macroscopic world of matter with the microscopic world of the molecule, and is, therefore, of central importance in chemistry. Since its invention, the laser has greatly expanded the capabilities of the spectroscopist. In linear spectroscopy, the monochromaticity, coherence, high intensity, and high degree of polarization of laser radiation are ideally suited to high-resolution spectroscopic investigations of even the weakest transitions. The same properties allowed, for the first time, the investigation of the nonlinear optical response of a medium to intense radiation. Shortly after the foundations of nonlinear optics were laid, it became apparent that these nonlinear optical signals could be exploited in molecular spectroscopy, and since then a considerable number of nonlinear optical spectroscopies have been developed. This short article is not a comprehensive review of all these methods. Rather, it is a discussion of some of the key areas in the development of the subject, and indicates some current directions in this increasingly diverse area.

Nonlinear Optics for Spectroscopy

The foundations of nonlinear optics are described in detail elsewhere in the encyclopedia, and in some of the texts listed in the Further Reading section at the end of this article. The starting point is usually the nonlinearity in the polarization, P_i, induced in the sample when the applied electric field, E, is large:

$$P_i = \varepsilon_0\left[\chi_{ij}^{(1)}E_j + \frac{1}{2}\chi_{ijk}^{(2)}E_jE_k + \frac{1}{4}\chi_{ijkl}^{(3)}E_jE_kE_l + \cdots\right] \quad [1]$$

where ε_0 is the vacuum permittivity, $\chi^{(n)}$ is the nth order nonlinear susceptibility, the indices represent directions in space, and the implied summation over repeated indices convention is used. The signal field, resulting from the nonlinear polarization, is calculated by substituting it as the source polarization in Maxwell's equations and converting the resulting field to the observable, which is the optical intensity.

The nonlinear polarization itself arises from the anharmonic motion of electrons under the influence of the oscillating electric field of the radiation. Thus, there is a microscopic analog of eqn [1] for the induced molecular dipole moment, μ_i:

$$\mu_i = [\varepsilon_0\alpha_{ij}d_j + \varepsilon_0^{-2}\beta_{ijk}d_jd_k + \varepsilon_0^{-3}\gamma_{ijkl}d_jd_kd_l + \cdots] \quad [2]$$

in which α is the polarizability, β the first molecular hyperpolarizability, γ the second, etc. The power series is expressed in terms of the displacement field $\mathbf{d}$ rather than $\mathbf{E}$ to account for local field effects. This somewhat complicates the relationship between the molecular hyperpolarizabilities, for example, γ_{ijkl} and

the corresponding macroscopic susceptibility, $\chi^{(3)}_{ijkl}$, but it is nevertheless generally true that a molecule exhibiting a high value of γ_{ijkl} will also yield a large third-order nonlinear polarization. This relationship between the macroscopic and microscopic para meters is the basis for an area of chemistry which has influenced nonlinear optics (rather than the inverse). The synthesis of molecules which have giant molecular hyperpolarizabilities has been an active area, because of their potential applications in lasers and electro-optics technology. Such molecules commonly exhibit a number of properties, including an extended linear π electron conjugation and a large dipole moment, which changes between the ground and excited electronic states. These properties are in accord with the predictions of theory and of quantum chemical calculations of molecular hyperpolarizabilities.

Nonlinear optical spectroscopy in the frequency domain is carried out by measuring the nonlinear signal intensity as a function of the frequency (or frequencies) of the incident radiation. Spectroscopic information is accessible because the molecular hyperpolarizability, and therefore the nonlinear susceptibility, exhibits resonances: the signal is enhanced when one or more of the incident or generated frequencies are resonant with the frequency of a molecular transition. The rich array of nonlinear optical spectroscopies arises in part from the fact that with more input fields there are more accessible resonances than is the case with linear spectroscopy. As an example we consider the third-order susceptibility, $\chi^{(3)}_{ijkl}E_jE_kE_l$, in the practically important case of two incident fields at the same frequency, ω_1, and a third at ω_2. The difference between the two frequencies is close to the frequency of a Raman active vibrational mode, Ω_{rg} (Figure 1). The resulting susceptibility can be calculated to have the form:

$$\chi^{(3)}_{ijkl}(-2\omega_1+\omega_2;\omega_1,\omega_1,-\omega_2) = \frac{N\Delta\rho_{rg}\Omega_{rg}(\alpha^{\mathrm{R}}_{ij}\alpha^{\mathrm{R}}_{kl}+\alpha^{\mathrm{R}}_{ik}\alpha^{\mathrm{R}}_{jl})}{12\hbar[\Omega^2_{rg}-(\omega_1-\omega_2)^2+\Gamma^2-2\mathrm{i}(\omega_1-\omega_2)\Gamma]} \quad [3]$$

in which the α^{R} are elements of the Raman susceptibility tensor, Γ is the homogeneous linewidth of the Raman transition, $\Delta\rho_{rg}$ a population difference, and N the number density. A diagram illustrating this process is shown in Figure 1, where the signal field is at the anti-Stokes Raman frequency ($2\omega_1-\omega_2$). The spectroscopic method which employs this scheme is called coherent anti-Stokes Raman spectroscopy (CARS) and is one of the most widely applied nonlinear optical spectroscopies (see below). Clearly, from eqn [3] we can see that the signal will be enhanced when the difference frequency is resonant with the Raman transition frequency.

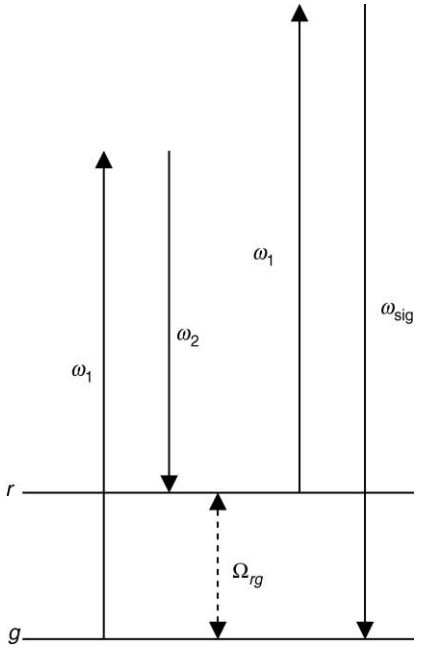


Figure 1 An illustration of the resonance enhancement of the CARS signal, $\omega_{sig} = 2\omega_1 - \omega_2$ when $\omega_1 - \omega_2 = \Omega_{rg}$.

With essentially the same experimental geometry there will also be nonlinear signals generated at both the Stokes frequency ($2\omega_2 - \omega_1$) and at the lower of the incident frequencies, ω_2. These signals have distinct phase matching conditions (see below), so they can easily be discriminated from one another, both spatially and energetically. Additional resonance enhancements are possible if either of the individual frequencies is resonant with an electronic transition of the molecule, in which case information on Raman active modes in the excited state is also accessible.

It is worthwhile noting here that there is an equivalent representation of the nth order nonlinear susceptibility tensor $\boldsymbol{\chi}^{(n)}(\omega_{\mathrm{sig}}:\omega_1,\ldots,\omega_n)$ as a time domain response function, $\mathbf{R}^{(n)}(\tau_1,\ldots,\tau_n)$. While it is possible to freely transform between them, the frequency domain representation is the more commonly used. However, the response function approach is increasingly applied in time domain nonlinear optical spectroscopy when optical pulses shorter than the homogeneous relaxation time are used. In that case, the time ordering of the incident fields, as well as their frequencies, is of importance in defining an experiment.

An important property of many nonlinear optical spectroscopies is the directional nature of the signal, illustrated in Figure 2. The directional nonlinear

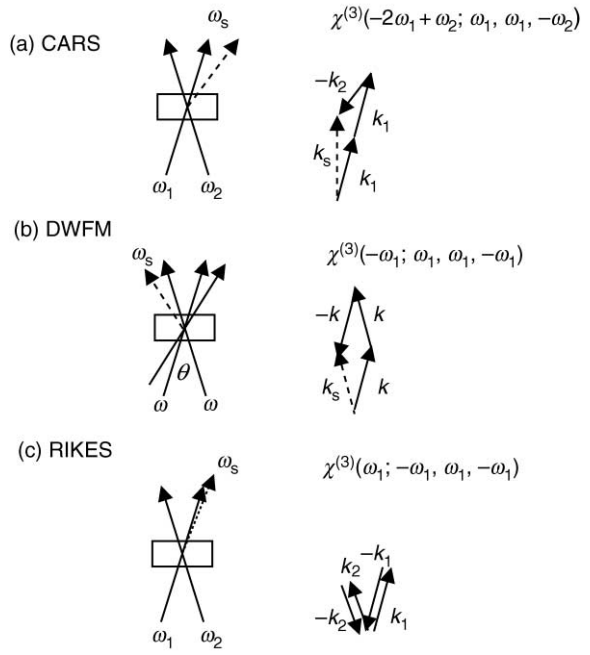


Figure 2 Experimental geometries and corresponding phase matching diagrams for (a) CARS (b) DFWM (c) RIKES. Note that only the RIKES geometry is fully phase matched.

signal arises from the coherent oscillation of induced dipoles in the sample. Constructive interference can lead to a large enhancement of the signal strength. For this to occur the individual induced dipole moments must oscillate in phase – the signal must be phase matched. This requires the incident and the generated frequencies to travel in the sample with the same phase velocity, $k_\omega/\omega = c/n_\omega$ where n_ω is the index of refraction and k_ω is the wave propagation constant at frequency ω. For the simplest case of second-harmonic generation (SHG), in which the incident field oscillates at ω and the generated field at 2ω, phase matching requires $2k_\omega = k_{2\omega}$. The phase-matching condition for the most efficient generation of the second harmonic is when the phase mismatch, $\Delta k = 2k_\omega - k_{2\omega} = 0$, but this is not generally fulfilled due to the dispersion of the medium, $n_\omega < n_{2\omega}$. In the case of SHG, a coherence length L can be defined as the distance traveled before the two waves are 180° out of phase, $L = |\pi/\Delta k|$. For cases in which the input frequencies also have different directions, as is often the case when laser beams at two frequencies are combined (e.g., CARS), the phase-matching condition must be expressed as a vectorial relationship, hence, for CARS, $\Delta\mathbf{k} = \mathbf{k}_{AS} - 2\mathbf{k}_1 + \mathbf{k}_2 \approx 0$, where $\mathbf{k}_{AS}$ is the wavevector of the signal at the anti-Stokes frequency (Figure 2). Thus for known input wavevectors one can easily calculate the expected direction of the output signal, $\mathbf{k}_s$. This is illustrated for a number of important cases in Figure 2.

Nonlinear Optical Spectroscopy in Chemistry

As already noted, there is a vast array of nonlinear optical spectroscopies, so it is clear some means of classification will be required. For coarse graining the order of the nonlinear process is very helpful, and that is the scheme we will follow here.

Second-Order Spectroscopies

Inspection of eqn [1] shows that in gases, isotropic liquids, and solids where the symmetry point group contains a center of inversion, $\boldsymbol{\chi}^{(2)}$ is necessarily zero. This is required to satisfy the symmetry requirement that polarization must change sign when the direction of the field is inverted, yet for a quadratic, or any even order dependence on the field strength, it must remain positive. Thus second-order nonlinearities might not appear very promising for spectroscopy. However, there are two cases in which second-order nonlinear optical phenomena are of very great significance in molecular spectroscopy, harmonic generation and the spectroscopy of interfaces.

Harmonic conversion

Almost every laser spectroscopist will have made use of second-harmonic or sum frequency generation for frequency conversion of laser radiation. Insertion of an oscillating field of frequency ω into eqn [1] yields a second-order polarization oscillating of 2ω. If two different frequencies are input, the second-order polarization oscillates at their sum and difference frequencies. In either case, the second-order polarization acts as a source for the second-harmonic (or sum, or difference frequency) emission, provided $\boldsymbol{\chi}^{(2)}$ is nonzero. The latter can be arranged by selecting a noncentrosymmetric medium for the interaction, the growth of such media being an important area of materials science. Optically robust and transparent materials with large values of $\boldsymbol{\chi}^{(2)}$ are available for the generation of wavelengths shorter than 200 nm to longer than 5 μm. Since such media are birefringent by design, a judicious choice of angle and orientation of the crystal with respect to the input beams allows a degree of control over the refractive indices experienced by each beam. Under the correct phase matching conditions $n_\omega \approx n_{2\omega}$, and very long interaction lengths result, so the efficiency of signal generation is high.

Higher-order harmonic generation in gases is an area of growing importance for spectroscopy in the

deep UV and X-ray region. The generation of such short wavelengths depends on the ability of amplified ultrafast solid state lasers to generate extremely high instantaneous intensities. The mechanism is somewhat different to the one outlined above. The intense pulse is injected into a capillary containing an appropriate gas. The high electric field of the laser causes ionization of atoms. The electrons generated begin to oscillate in the applied laser field. The driven recombination of the electrons with the atoms results in the generation of the high harmonic emission. Although the mechanism differs from the SHG case, questions of phase matching are still important. By containing the gas in a corrugated waveguide phase matching is achieved, considerably enhancing the intensity of the high harmonic. Photon energies of hundreds of electronvolts are possible using this technique. The generation of such high energies is not yet routine, but a number of potential applications are already apparent. A powerful coherent source of X-ray and vacuum UV pulses will certainly aid surface analysis techniques such as UV photoemission and X-ray photoelectron spectroscopy. Much excitement is currently being generated by the possibility of using intense ultrashort X-ray pulses to record structural dynamics on an ultrafast timescale.

Surface second-order spectroscopy

At an interface inversion symmetry is absent by definition, so the second-order nonlinear susceptibility is finite. If the two bulk phase media are themselves isotropic, then even a weak second-order signal necessarily arises from the interface. This surface-specific all-optical signal is unique, because it can be used to probe the interface between two condensed phases. This represents a great advantage over every other form of surface spectroscopy. In linear optical spectroscopy, the signal due to species at the interface are usually swamped by contributions from the bulk phase. Other surface-specific signals do exist, but they rely on the scattering of heavy particles (electrons, atoms) and so can only be applied to the solid vacuum interface. For this reason the techniques of surface SHG and SFG are widely applied in interface spectroscopy.

The most widely used method is sum frequency generation (SFG) between temporally overlapped tuneable infrared and fixed frequency visible lasers, to yield a sum frequency signal in the visible region of the spectrum. The principle of the method is shown in Figure 3. The surface nonlinear susceptibility exhibits resonances at vibrational frequencies, which are detected as enhancements in the visible SFG intensity. Although the signal is weak, it is directional and background free, so relatively easily measured by photon counting techniques. Thus, SFG is used to measure the vibrational spectra of essentially any optically accessible interface. There are, however, some limitations on the method. The surface must exhibit a degree of order – if the distribution of adsorbate orientation is isotropic the signal again becomes zero by symmetry. Second, a significant enhancement at the vibrational frequency requires the transition to be both IR and Raman allowed, as suggested by the energy level diagram (Figure 3).

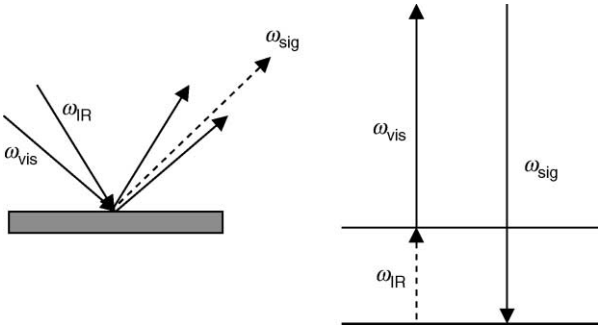


Figure 3 The experimental geometry for SFG, and an illustration of the resonance enhancement of $\omega_{sig} = \omega_{IR} + \omega_{vis}$ at a Raman and IR allowed vibrational transition.

The SHG signal can also be measured as a function of the frequency of the incident laser, to recover the electronic spectrum of the interface. This method has been used, particularly in the study of semiconductor surfaces, but generally the electronic spectra of adsorbates contain less information than their vibrational spectra. However, by measuring the SHG intensity as a function of time, information on adsorbate kinetics is obtained, provided some assumptions connecting the surface susceptibility to the molecular hyperpolarizability are made. Finally, using similar assumptions, it is possible to extract the orientational distribution of the adsorbate, by measuring the SHG intensity as a function of polarization of the input and output beams. For these reasons, SHG has been widely applied to analyze the structure and dynamics of interfaces.

Third-Order Spectroscopies

The third-order coherent Raman spectroscopies were introduced above. One great advantage of these methods over conventional Raman is that the signal is generated in a coherent beam, according to the appropriate phase matching relationship (Figure 2). Thus, the coherent Raman signal can easily be distinguished from background radiation by spatial filtering. This has led to CARS finding widespread application in measuring spectra in (experimentally)

hostile environments. CARS spectroscopy has been widely applied to record the vibrational spectra of flames. Such measurements would obviously be very challenging for linear Raman or IR, due to the strong emission from the flame itself. The directional CARS signal in contrast can be spatially filtered, minimizing this problem. CARS has been used both to identify unstable species formed in flames, and to probe the temperature of the flame (e.g., from measured population differences eqn [3]). A second advantage of the technique is that the signal is only generated when the two input beams overlap in space. Thus, small volumes of the sample can be probed. By moving the overlap position around in the sample, spatially resolved information is recovered. Thus, it is possible to map the population of a particular transient species in a flame.

CARS is probably the most widely used of the coherent Raman methods, but it does have some disadvantages, particularly in solution phase studies. In that case the resonant $\boldsymbol{\chi}^{(3)}$ signal (eqn [3]) is accompanied by a nonresonant third-order background. The interference between these two components may result in unusual and difficult to interpret lineshapes. In this case, some other coherent Raman methods are more useful. The phase matching scheme for Raman-induced Kerr effect spectroscopy (RIKES) was shown in Figure 2. The RIKES signal is always phase matched, which leads to a long interaction length. However, the signal is at ω_2 and in the direction of ω_2, which would appear to be a severe disadvantage in terms of detection. Fortunately, if the input polarizations are correctly chosen the signal can be isolated by its polarization. In an important geometry, the signal (ω_2) is isolated by transmission through a polarizer oriented at 45° to a linearly polarized pump (ω_1). The pump is overlapped in the sample with the probe (ω_2) linearly polarized at $-45°$. Thus the probe is blocked by the polarizer, but the signal is transmitted. This geometry may be viewed as pump-induced polarization of the isotropic medium to render it birefringent, thus inducing ellipticity in the transmitted probe, such that the signal leaks through the analyzing polarizer (hence the alternative name optical Kerr effect).

In this geometry, it is possible to greatly enhance signal-to-noise ratios by exploiting interference between the probe beam and the signal. Placing a quarterwave plate in the probe beam with its fast axis aligned with the probe polarization, and reorienting it slightly ($<1°$) yields a slightly elliptically polarized probe before the sample. A fraction of the probe beam, the local oscillator (LO), then also leaks through the analyzing polarizer, temporally and spatially overlapped with the signal. Thus, the signal and LO fields are seen by the detector, which measures the intensity as:

$$I(t) = \frac{nc}{8\pi}|\mathbf{E}_{\mathrm{LO}}(t) + \mathbf{E}_{\mathrm{S}}(t)|^2$$

$$= I_{\mathrm{LO}}(t) + I_{\mathrm{S}}(t) + \frac{nc}{4\pi}\mathrm{Re}[\mathbf{E}^*_{\mathrm{S}}(t)\cdot\mathbf{E}_{\mathrm{LO}}(t)] \qquad [4]$$

where the final term may be very much larger than the original signal, and is linear in $\boldsymbol{\chi}^{(3)}$. This term is usually isolated from the strong I_{LO} by lock-in detection. This method is called optical heterodyne detection (OHD), and generally leads to excellent signal to noise. It can be employed with other coherent signals by artificially adding the LO to the signal, provided great care is taken to ensure a fixed phase relationship between LO and signal. In the RIKES experiment, however, the phase relationship is automatic. The arrangement described yields an out-of-phase LO, and measures the real part of $\boldsymbol{\chi}^{(3)}$, the birefringence. Alternatively, the quarterwave plate is omitted, and the analyzing polarizer is slightly reoriented, to introduce an in-phase LO, which measures the imaginary part of $\boldsymbol{\chi}^{(3)}$, the dichroism. This is particularly useful for absorbing media. The OHD-RIKES method has been applied to measure the spectroscopy of the condensed phase, and has found particularly widespread application in transient studies (below).

Degenerate four wave mixing (DFWM) spectroscopy is a simple and widely used third-order spectroscopic method. As the name implies, only a single frequency is required. The laser beam is split into three, and recombined in the sample, in the geometry shown in Figure 2. The technique is also known as laser-induced grating scattering. The first two beams can be thought of as interfering in the sample to write a spatial grating, with fringe spacing dependent on the angle between them. The signal is then scattered from the third beam in the direction expected for diffraction from that grating. The DFWM experiment has been used to measure electronic spectra in hostile environments, by exploiting resonances with electronic transitions. It has also been popular in the determination of the numerical value of $\boldsymbol{\chi}^{(3)}$, partly because it is an economical technique, requiring only a single laser, but also because different polarization combinations make it possible to access different elements of $\boldsymbol{\chi}^{(3)}$. The technique has also been used in time resolved experiments, where the decay of the grating is monitored by diffraction intensity from a time delayed third pulse.

Two-photon or, more generally, multiphoton excitation has applications in both fundamental spectroscopy and analytical chemistry. Two relevant level

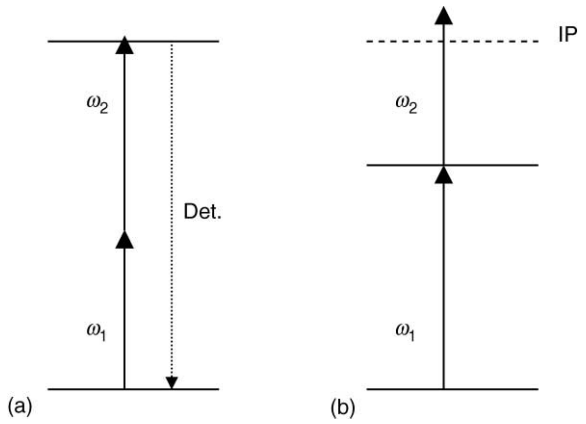


Figure 4 Two cases of resonant two photon absorption. In (a) the excited state is two-photon resonant, and the process is detected by the emission of a photon. In (b) the intermediate state is resonant, and the final energy is above the ionization potential (IP) so that photocurrent or mass detection can be used.

schemes are shown in Figure 4. Some property associated with the final state permits detection of the multiphoton absorption, for example, fluorescence in (a) and photocurrent in (b).

Excitation of two-photon transitions, as in Figure 4a, is useful in spectroscopy because the selection rules are different to those for the corresponding one-photon transition. For example the change in angular momentum quantum number, ΔL, in a two-photon transition is 0, ± 2, so, for example, an atomic S to D transition can be observed. High spectroscopic resolution may be attained using Doppler free two-photon absorption spectroscopy. In this method, the excitation beams are arranged to be counter-propagating, so that the Doppler broadening is cancelled out in transitions where the two excitation photons arise from beams with opposing wavevectors. In this case, the spectroscopic linewidth is governed only by the homogeneous dephasing time.

The level scheme in Figure 4b is also widely used in spectroscopy, but in this case the spectrum of the intermediate state is obtained by monitoring the photocurrent as a function of ω_1. The general technique is known as resonance enhanced multiphoton ionization (REMPI) and yields high-quality spectra of intermediate states which are not detectable by standard methods, such as fluorescence. The sensitivity of the method is high, and it is the basis of a number of analytical applications, often in combination with mass spectrometry.

Ultrafast Time Resolved Spectroscopy

The frequency and linewidth of a Raman transition may be extracted from the CARS measurement, typically by combining two narrow bandwidth pulsed lasers, and tuning one through the resonance while measuring the nonlinear signal intensity. The time resolved analogue requires two pulses, typically of a few picoseconds duration (and therefore a few wavenumbers bandwidth) at ω_1 and ω_2 to be incident on the sample. This pair coherently excites the Raman mode. A third pulse at ω_1 is incident on the sample a time t later, and stimulates the CARS signal at $2\omega_1 - \omega_2$ in the phase-matched direction. The decay rate of the vibrational coherence is measured from the CARS intensity as a function of the delay time. Thus, the frequency of the mode is measured in the frequency domain, but the linewidth is measured in the time domain. If very short pulses are used (such that the pulsewidth is shorter than the inverse frequency of the Raman active mode) the Raman transition is said to be impulsively excited, and the CARS signal scattered by the time delayed pulse reveals an oscillatory response at the frequency of the Raman active mode, superimposed on its decay. Thus, in this case, spectroscopic information is measured exclusively in the time domain. In the case of nonlinear Raman spectroscopy, similar information is available from the frequency and the time domain measurements, and the choice between them is essentially one of experimental convenience. For example, time domain CARS, RIKES, and DFWM spectroscopy have turned out to be particularly powerful routes to extracting low-frequency vibrational and orientational modes of liquids and solutions, thus providing detailed insights into molecular interactions and reaction dynamics in the condensed phase.

Other time domain experiments contain information that is not accessible in the frequency domain. This is particularly true of photon echo methods. The name suggests a close analogy with nuclear magnetic resonance (NMR) spectroscopy, and the (optical) Bloch vector approach may be used to describe both measurements, although the transition frequencies and time-scales involved differ by many orders of magnitude. In the photon echo experiment, two or three ultrafast pulses with carefully controlled interpulse delay times are resonant with an electronic transition of the solute. In the two-pulse echo, the echo signal is emitted in the phase match direction at twice the interpulse delay, and its intensity as a function of time yields the homogeneous dephasing time associated with the transition. In the three-pulse experiment the pulses are separated by two time delays. By measuring the intensity of the stimulated echo as a function of both delay times it is possible to separately determine the dephasing time and the population relaxation time associated with the resonant transition. Such information is not accessible from linear spectroscopy, and can be extracted

only with difficulty in the frequency domain. The understanding of photon echo spectroscopy has expanded well beyond the simple description given here, and it now provides unprecedented insights into optical dynamics in solution, and thus informs greatly our understanding of chemistry in the condensed phase. The methods have recently been extended to the infra red, to study vibrational transitions.

Higher-Order and Multidimensional Spectroscopies

The characteristic feature of this family of spectroscopies is the excitation of multiple resonances, which may or may not require measurements at $\chi^{(n)}$ with $n > 3$. Such experiments require multiple frequencies, and may yield weak signals, so they only became experimentally viable upon the availability of stable and reliable solid state lasers and optical parametric generators. Measurements are made in either the time or the frequency domain, but in either case benefit from heterodyne detection.

One of the earliest examples was two-dimensional Raman spectroscopy, where multiple Raman active modes are successively excited by temporally delayed pulse pairs, to yield a fifth-order nonlinear signal. The signal intensity measured as a function of both delay times (corresponding to the two dimensions) allows separation of homogeneous and inhomogeneous contributions to the line shape. This prodigiously difficult $\chi^{(5)}$ experiment has been completed in a few cases, but is plagued by interference from third-order signals.

More widely applicable are multidimensional spectroscopies using infrared pulses or combinations of them with visible pulses. The level scheme for one such experiment is shown in Figure 5 (which is one of many possibilities). From the scheme, one can see that the nonlinear signal in the visible depends on two resonances, so both can be detected. This can be regarded as a multiply resonant nondegenerate four-wave mixing (FWM) experiment. In addition, if the two resonant transitions are coupled, optical excitation of one affects the other. Thus, by measuring the signal as a function of both frequencies, the couplings between transitions are observed. These appear as cross peaks when the intensity is plotted in the two frequency dimensions, very much as with 2D NMR. This technique is already providing novel information on molecular structure and structural dynamics in liquids, solutions, and proteins.

Spatially Resolved Spectroscopy

A recent innovation is nonlinear optical microscopy. The nonlinear dependence of signal strength on

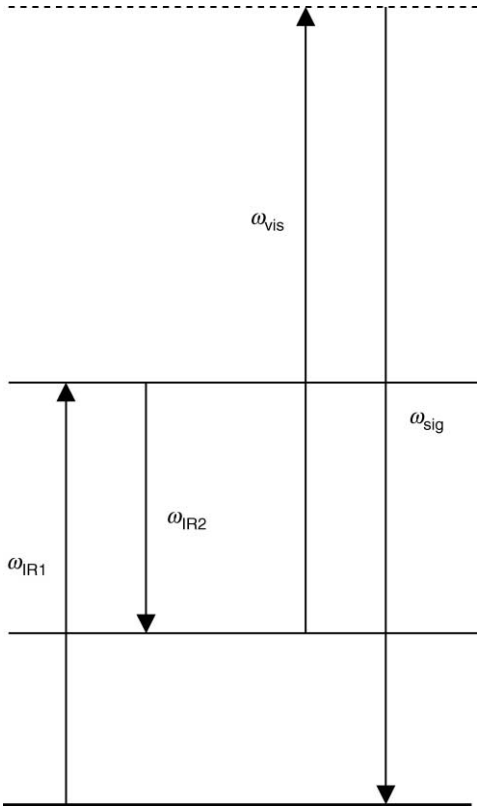


Figure 5 Illustration of multiple resonance enhancements in a FWM geometry, from which 2D spectra may be generated.

intensity means that nonlinear processes are localized at the focal point of a lens. When focusing is strong, such as in a microscope objective, spatial localization of the nonlinear signal can be dramatic. This is the basis of the two-photon fluorescence microscopy method, where a high repetition rate source of low-energy ultrafast pulses is focused by a microscope objective into a sample labeled with a fluorescent molecule, which has absorption at half the wavelength of the incident photons (Figure 4). The fluorescence is necessarily localized at the focal point because of its dependence on the square of the incident intensity. By measuring intensity while scanning the position of the focal point in space, a 3D image of the distribution of the fluorophore is constructed. This technique turns out to have a number of advantages over one photon fluorescence microscopy, most notably in terms of ease of implementation, minimization of sample damage, and depth resolution. The technique is widely employed in cell biology.

Stimulated by the success of two-photon microscopy, further nonlinear microscopies have been developed, all relying on the spatial localization of the signal. CARS microscopy has been

demonstrated to yield 3D images of the distribution of vibrations in living cells. It would be difficult to recover such data by linear optical microscopy. SHG has been applied in microscopy. In this case, by virtue of the symmetry selection rule referred to above, a 3D image of orientational order is recovered. Both these and other nonlinear signals provide important new information on complex heterogeneous samples, most especially living cells.

See also

Spectroscopy: Nonlinear Laser Spectroscopy; Raman Spectroscopy.

Further Reading

Andrews DL and Allcock P (2002) *Optical Harmonics in Molecular Systems*. Weinheim, Germany: Wiley-VCH.

Andrews DL and Demidov AA (eds) (2002) *An Introduction to Laser Spectroscopy*. New York: Kluwer Academic.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Cheng JX and Xie XS (2004) Coherent anti-Stokes Raman scattering microscopy: Instrumentation, theory, and applications. *Journal of Physical Chemistry B* 108: 827.

de Boeij WP, Pshenichnikov MS and Wiersma DA (1998) Ultrafast solvation dynamics explored by femtosecond photon echo spectroscopies. *Annual Review of Physical Chemistry* 49: 99.

Eisenthal KB (1992) Equilibrium and dynamic processes at interfaces by 2nd harmonic and sum frequency generation. *Annual Review of Physical Chemistry* 43: 627.

Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. Oxford, UK: Oxford University Press.

Fleming GR and Cho M (1996) Chromophore-solvent dynamics. *Annual Review of Physical Chemistry* 47: 109.

Fourkas JT (2002) Higher-order optical correlation spectroscopy in liquids. *Annual Review of Physical Chemistry* 53: 17.

Hall G and Whitaker BJ (1994) Laser-induced grating spectroscopy. *Journal of Chemistry Society Faraday Transactions* 90: 1.

Heinz TF (1991) Second order nonlinear optical effects at surfaces and interfaces. In: Ponath H-E and Stegeman GI (eds) *Nonlinear Surface Electromagnetic Phenomena*, pp. 353.

Hesselink WH and Wiersma DA (1983) Theory and experimental aspects of photon echoes in molecular solids. In: Hochstrasser RM and Agranovich VM (eds) *Spectroscopy and Excitation Dynamics of Condensed Molecular Systems*, chap. 6. Amsterdam: North Holland.

Levenson MD and Kano S (1988) *Introduction to Nonlinear Laser Spectroscopy*. San Diego, CA: Academic Press.

Meech SR (1993) Kinetic application of surface nonlinear optical signals. In: Lin SH, Fujimura Y, and Villaeys A (eds) *Advances in Multiphoton Processes and Spectroscopy*, vol. 8, pp. 281. Singapore: World Scientific.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. Oxford: Oxford University Press.

Rector KD and Fayer MD (1998) Vibrational echoes: a new approach to condensed-matter vibrational spectroscopy. *International Review of Physical Chemistry* 17: 261.

Richmond GL (2001) Structure and bonding of molecules at aqueous surfaces. *Annual Review of Physical Chemistry* 52: 357.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: Wiley.

Smith NA and Meech SR (2002) Optically heterodyne detected optical Kerr effect – Applications in condensed phase dynamics. *International Review of Physical Chemistry* 21: 75.

Tolles WM, Nibler JW, McDonald JR and Harvey AB (1977) Review of theory and application of coherent anti-Stokes Raman spectroscopy (CARS). *Applied Spectroscopy* 31: 253.

Wright JC (2002) Coherent multidimensional vibrational spectroscopy. *International Review of Physical Chemistry* 21: 185.

Zipfel WR, Williams RM and Webb WW (2003) Nonlinear magic multiphoton microscopy in the biosciences. *Nature Biotechnology* 21: 1368.

Zyss J (ed.) (1994) *Molecular Nonlinear Optics*. San Diego: Academic Press.

Photodynamic Therapy of Cancer

A J MacRobert and T Theodossiou, University College London, London, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

One of the most active areas of photomedical research, in recent years, has been the exploration of the use of light-activated drugs known as photosensitizers. These compounds may be activated using light, usually provided by a laser via an optical fiber which is placed at the site of the target lesion. This treatment is known as photodynamic therapy (PDT) and is being applied to the local destruction of malignant tumors and certain nonmalignancies. Activation of the photosensitizer results in the

generation of reactive oxidizing intermediates which are toxic to cells, and this process ultimately leads to tumor destruction. PDT is a relatively low-power, nonthermal, photochemical technique that uses fluence rates not exceeding 200 mW/cm^2 and total light doses or fluences of typically 100 J/cm^2. Generally red or near-infrared light is used, since tissue is relatively transparent at these wavelengths. It is a promising alternative approach to the local destruction of tumors for several reasons. First, it is a minimally invasive treatment since laser light can be delivered with great accuracy to almost any site in the body via thin flexible optical fibers with minimal damage to overlying normal tissues. Second, the nature of PDT damage to tissues is such that healing is safer and more complete than after most other forms of local tissue destruction (e.g., radiotherapy). PDT is also capable of some degree of selectivity for tumors when the sensitizer levels, light doses, and irradiation geometry are carefully controlled. This selectivity is based on the higher sensitizer retention in tumors after administration relative to the adjacent normal tissues in which the tumor arose (generally 3:1 for extracranial tumors but up to 50:1 for brain tumors).

The photosensitizer is administered intravenously to the patient and time allowed (3–96 hours depending on the sensitizer) for it to equilibrate in the body before the light treatment (Figure 1). This time is called the drug–light interval. PDT may also be useful for treating certain nonmalignant conditions, in particular, psoriasis and dysfunctional menorrhagia (a disorder of the uterus), and local treatment of infections; such as genital papillomas, infections in the mouth and the upper gastrointestinal tract. In certain cases, the photosensitizer may be applied directly to the lesions, particularly for treatment of skin tumors, as discussed later. The main side-effect of PDT is skin photosensitivity, owing to retention of the drug in the skin, so patients must avoid exposure to sunlight in particular for a short period following treatment. Retreatment is then possible once the photosensitizer has cleared from the skin since these drugs have little intrinsic toxicity, unlike many conventional chemotherapy agents.

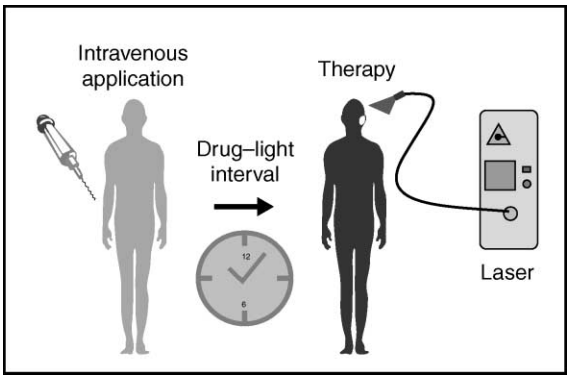


Figure 1 Photodynamic therapy: from sensitization to treatment.

Photoproperties of Photosensitizers

By definition, there are three fundamental requirements for obtaining a photodynamic effect: (a) light of the appropriate wavelength matched to the photosensitizer absorption; (b) a photosensitizer; and (c) molecular oxygen. The ideal photochemical and biological properties of a photosensitizer may be easily summarized, although the assessment of a sensitizer in these terms is not as straightforward as might be supposed because the heterogeneous nature of biological systems can sometimes profoundly affect these properties. Ideally though, a sensitizer should possess the following attributes: (a) red or near infrared light absorption; (b) nontoxic, and with low skin photosensitizing potency; (c) selective retention in tumors relative to normal adjacent tissue; (d) an efficient generator of cytotoxic species, usually singlet oxygen; (e) fluorescence, for visualization; (f) a defined chemical composition, and (g) preferably water soluble. A list of several photosensitizers possessing the majority of the above-mentioned attributes is given in Table 1.

The reasons for these requirements are partly self-evident, but worth amplifying. Strong absorption is desirable in the red and near-infrared spectral region, where tissue transmittance is optimum enabling penetration of the therapeutic light within the tumor (Figure 2). To minimize skin photosensitization by solar radiation, the sensitizer absorption spectrum should ideally consist of a narrow red wavelength band, with little absorption at other wavelengths down to 400 nm, below which solar irradiation falls off steeply. Another advantage of red wavelength irradiation is that the potential mutagenic effects encountered with UV-excited sensitisers (e.g., psoralens) are avoided. Since the object of the treatment is the selective destruction of tumor tissue, leaving surrounding normal tissue undamaged, some degree of selective retention of the dye in tumor tissue is desirable. Unfortunately, the significance of this aspect has been exaggerated in the literature and in many cases treatment selectivity owes more to careful light irradiation geometry. Nevertheless, many normal tissues have the capacity to heal safely following PDT damage.

The key photochemical property of photosensitizers is to mediate production of some active

Table 1 Properties of several photosensitizers

Compound	*λ/nm (ε/M^{-1} cm^{-1})*	*Drug dose/mg kg^{-1}*	*Light dose/J cm^{-2}*	*Diseases treated*
Hematoporphyrin (HpD–Photofrin–Photosan)	628 (3.0×10^3)	1.5–5	75–250	Early stage esophagus, bladder, lung, cervix, stomach, and mouth cancers. Palliative in later stages
ALA (converted to protoporphyrin IX)	635 (5×10^3)	60	50–150	Skin, stomach, colon, bladder, mouth cancers. Esophageal dysplasia. Various nonmalignant conditions
Benzoporphyrin derivative (BpD)	690 (3.5×10^4)	4	150	Age-related macular degeneration (AMD)
Tin etiopurpurin (SnET2–Purlytin)	665 (3.0×10^4)	1.2	150–200	Breast and skin cancers, AMD
Monoaspartyl chlorin e_6 (MACE)	660 (4.0×10^4)	1.0	25–200	Skin cancers
Lutetium texaphyrin (Lu–Tex)	732 (4.2×10^4)	1.0	150	Metastatic brain tumors, breast cancers, atherosclerotic plaques
Aluminum disulfonated phthalocyanine	675 (2.0×10^5)	1.0	50–200	Brain, colon, bladder, and pancreatic cancers. Head and neck cancers in animal studies only
Metatetrahydroxychlorin (mTHPC–temoporphin–Foscan)	652 (3.5×10^4)	0.15	5–20	Head, neck, prostate, pancreas, lung, brain, biliary tract, and mouth cancers. Superior to HpD and ALA in mouth cancers
Palladium pheophorbide (Tookad)	763 (8.6×10^4)	2.0	–	Prostate cancer
Hypericin	590 (3×10^4)	–	–	Psoriasis

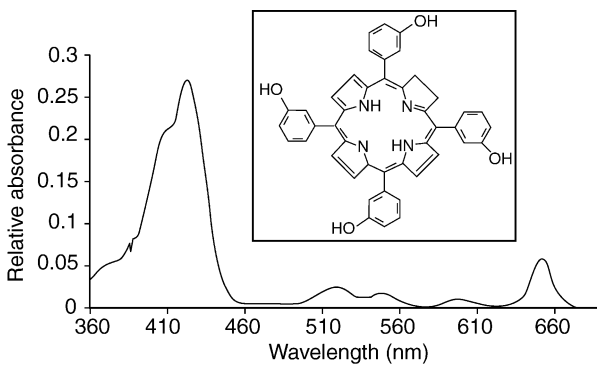


Figure 2 The absorption spectrum of m-THPC. The structure of m-THPC is shown in the inset.

molecule which is cytotoxic, that is, will destroy cells. The first electronically excited state of molecular oxygen, so-called singlet oxygen, fulfills this role very well and may be produced via the interaction of an excited electronic state of the sensitizer with oxygen present in the tissue. Thus, in summary, to achieve effective photosensitization, a sensitizer should exhibit appreciable absorption at red to near-infrared wavelengths and generate cytotoxic species via oxygen-dependent photochemical reactions. The first clinical photosensitizer was hematoporphyrin derivative (HpD), which is derived synthetically from hematoporphyrin by reaction with acetic and sulfuric acids to give a complex mixture of porphyrins. A purified fraction of these (Photofrin) is available commercially, and this has been used most widely in clinical applications to date. Second-generation (photosensitizers are now becoming available, including phthalocyanine and chlorin compounds as shown in Table 1.

A new approach to PDT has recently emerged involving the administration of a natural porphyrin precursor, 5-aminolaevulinic acid (ALA), which is metabolized within cells to produce protoporphyrin IX (see Table 1). This porphyrin is known to be a powerful photosensitizing agent but suffers from the drawback of being a poor tumor localizer when used directly. In contrast, administration of ALA induces protoporphyrin biosynthesis, particularly in rapidly proliferating cells, which may then be destroyed using irradiation at 630 nm. Therefore, this new therapy may offer enhanced treatment selectivity with little risk of skin photosensitivity owing to the rapid clearance after 24 h of the protoporphyrin. Investigation of this new approach has already proved successful in clinical treatment of

skin tumors using topical application of ALA in a thick emulsion.

Considerable effort is also being expended on exploiting the retention of sensitizers in tumors for diagnostic purposes, although the results are rather mixed to date. The prospects with ALA-induced protoporphyrin IX are, however, more promising, as illustrated in Figure 3 where the fluorescence is selectively confined to the skin tumor (basal cell carcinoma).

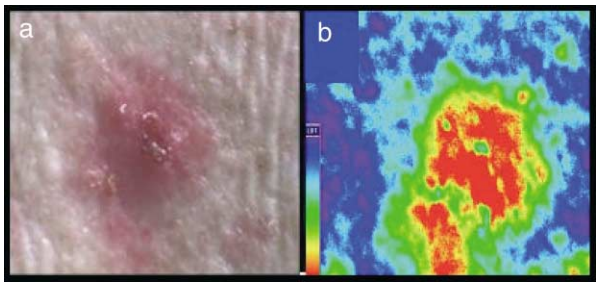


Figure 3 (a) Image of a basal cell carcinoma; (b) fluorescence imaging of the same lesion after ALA sensitization and 405 nm light excitation.

Mechanisms of Photodynamic Therapy

The main principles of the photosensitization mechanism are now well established with the initial step being excitation of the sensitizer from its electronic ground state to the short-lived fluorescent singlet state. The lifetime of the singlet state is generally only a few nanoseconds and the main role of this state in the photosensitization mechanism is to act as a precursor of the metastable triplet state through intersystem crossing. Efficient formation of this metastable state is required because it is the interaction of the triplet state with tissue components that generates cytotoxic species such as singlet oxygen. Thus the triplet state quantum yield (i.e., probability of triplet state formation per photon absorbed) of photosensitizers should ideally approach unity. Interaction of the metastable triplet state (which in de-aerated solutions has a lifetime extending to the millisecond range) with tissue components may proceed via either a type I or II mechanism or a combination (see Figure 4). A type I process can involve hydrogen abstraction from the sensitizer to

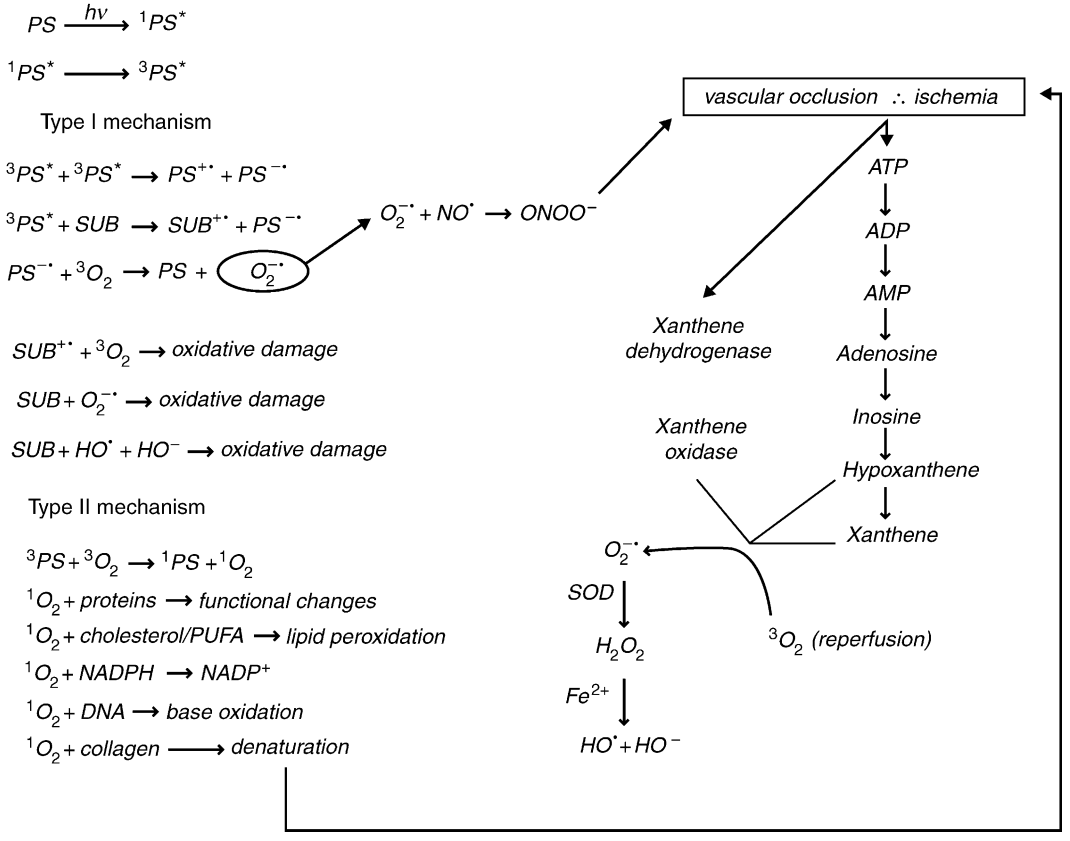


Figure 4 The photophysical and photochemical mechanisms involved in PDT. PS denotes the photosensitizer, SUB denotes the substrate, either biological, a solvent or another photosensitizer; * denotes the excited state and • denotes a radical.

produce free radicals or electron transfer, resulting in ion formation. The type II mechanism, in contrast, exclusively involves interaction between molecular oxygen and the triplet state to form singlet oxygen which is highly reactive in biological systems. The near-resonant energy transfer from the triplet state to O_2 can be highly efficient and the singlet oxygen yield can approach the triplet state yield provided that the triplet state energy exceeds 94 kJ/mol, the singlet oxygen excitation energy. It is widely accepted that the type II mechanism underlies the oxygen-dependent phototoxicity of sensitizers used for photodynamic therapy. Both proteins and lipids (the main constituents of membranes) are susceptible to photooxidative damage induced via a type II mechanism which generally results in the formation of unstable peroxide species. For example, unsaturated lipids may be oxidized by the 'ene' reaction where the singlet oxygen reacts with a double bond. Other targets include such important biomolecules as cholesterol, certain amino acid residues, collagen, and the coenzyme, NADPH. Another synergistic mode of action involves the occlusion of the microcirculation to tumors and reperfusion injury. Reperfusion injury relates to the formation of xanthene oxidase from xanthene dehydrogenase in anoxic conditions which reacts with xanthene or hypoxanthene products of ATP dephosphorylation to convert the restored oxygen into superoxide anion which directly or via the Fenton reaction causes cellular damage.

An important feature of the type II mechanism, which is sometimes overlooked, is that when the sensitizer transfers electronic energy to O_2 it returns to its ground state. Thus the cytotoxic singlet oxygen species is generated without chemical transformation of the sensitizer which may then absorb another photon and repeat the cycle. Effectively, a singlet photosensitizer molecule is capable of generating many times its own concentration of singlet oxygen, which is clearly a very efficient means of photosensitization provided the oxygen supply is adequate.

Lasers in PDT

Lasers are the most popular light source in PDT since they have several key characteristics that differentiate them from conventional light sources, namely coherence, monochromaticity, and collimated output. The two main practical features that make them so useful in PDT are their monochromaticity and their combination with fiber-optic delivery. The monochromaticity is important since the laser can be tuned to a specific absorption peak of a photosensitizer, thus ensuring that all the energy delivered is utilized for the excitation and photodynamic activation of the photosensitizer. This is not true for a conventional light source (as for example a tungsten lamp) where the output power is divided over several hundred nanometers throughout the UV, visible, and IR regions and only a fraction of the power lies within the absorption band of the photosensitizer. To numerically illustrate this, let us simplify things by representing the lamp output as a square profile and also consider a photosensitizer absorption band about 30 nm full width half maximum. If a laser with a frequency span of about 2 nm and a lamp with the same power but with a frequency span of about 600 nm are used, then the useful portion of it is about 0.05 and if we consider a Gaussian profile for the sensitizer absorption band, the above fraction drops even lower, perhaps even to 0.01. That means that the lamp would achieve 100 times less the excitation rate of the laser, or in other words, we would require a lamp with an output power of 100 times that of the laser to achieve the same rate of excitation and consequently the same time of treatment, provided we use a low enough laser power to remain within the linear regime of the photosensitizer excitation. The other major disadvantage of a lamp source is the lack of collimation which results in low efficiency for fiber-optic delivery.

We now review the different lasers that have found application in PDT. Laser technology has significantly advanced in recent years and there are now a range of options in PDT laser sources for closely matching the absorption profiles of the various photosensitizers. Moreover, since PDT is becoming more widely used clinically, practical considerations, such as portability and turn-key operation are increasingly important. In Figure 5 the optical spectrum is shown from about 380 nm (violet–blue) to 800 nm (near-infrared, NIR). We have superimposed on this spectrum, the light penetration depth of tissue to roughly illustrate how deeply light can penetrate (and consequently activate photosensitizer molecules) into lower-lying tissue layers. The term 'penetration depth' is a measure of the light attenuation by the tissue so the figure of 2 mm corresponds to $1/e$ attenuation of the incident intensity at a tissue depth of 2 mm. It is obvious that light at the blue end of the spectrum has only a superficial effect whereas the more we shift further into the red and NIR regions the penetration depth increases. This is due to two main reasons: absorption and scattering of light by various tissue component molecules. For example, in the blue/green region for the absorption of melanin and hemoglobin is relatively high. But there is a noticeable increase in penetration depth beyond about 600 nm leading to an optimal wavelength region, or

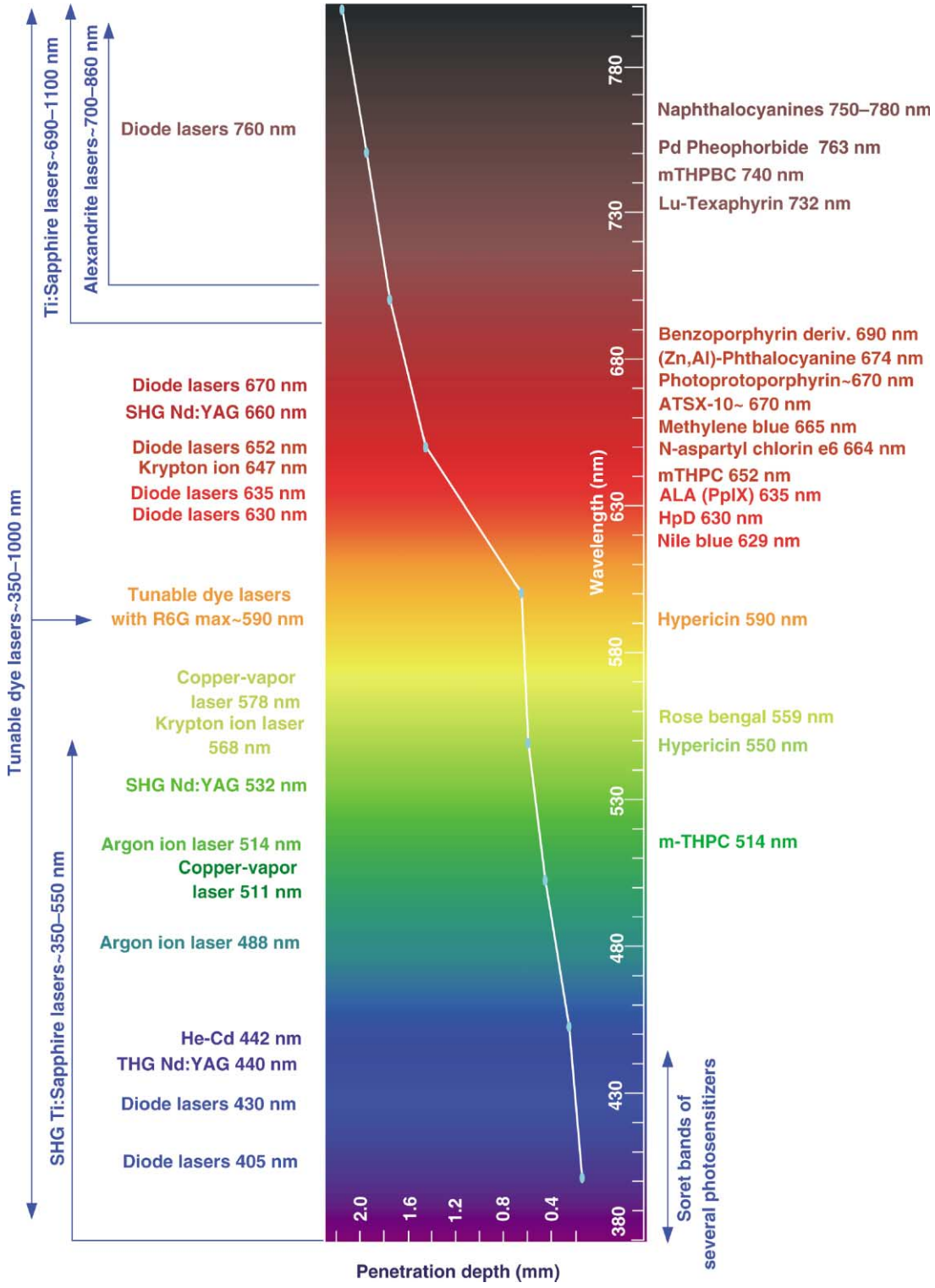


Figure 5 Photosensitizers and laser light sources available in the visible and near-infrared spectral regions.

'therapeutic window' for laser therapy of around 600–1100 nm (Figure 5).

The fundamental wavelength of the Nd:YAG laser lies within this therapeutic window at 1064 nm and this laser is now widely used in thermal laser therapy. The Nd:YAG lasers can be operated either in cw (output power ~ 200 W multimode), long-pulse (~500 W average power at 50 Hz), or Q-switched

(50 MW peak power at around 10 ns pulse duration) modes. Nd:YAG is a solid-state laser with a yttrium aluminum garnet crystal doped with about 1% of trivalent Nd ions as the active medium of the laser; and with another transition of the Nd ion, this laser (with the choice of different optics) can also operate at 1320 nm. Although dyes are available which absorb from 800–1100 nm, the generation of singlet oxygen via the type II mechanism is energetically unfavorable because the triplet state energies of these dyes are too low. However, the fundamental frequency of the laser can be doubled (second-harmonic generation, SHG) or tripled (third-harmonic generation, THG) with the use of nonlinear crystals, to upconvert the output radiation to the visible range, thus rendering this laser suitable for PDT: i.e., frequency doubling to 532 nm and tripling to 355 nm. Note that the 532 nm output is suitable for activation of the absorption band of hypericin with a maximum at 550 nm even though it is not tuned to the maximum of this absorption.

In the early days of PDT and other laser therapies, ion lasers were widely used. The argon ion laser uses ionized argon plasma as gain medium, and produces two main wavelengths at 488 nm and 514 nm. The second of the two lies exactly at the maximum of the 514 nm m-THPC absorption. Argon ion lasers are operated in cw mode and usually have output powers in the region of 5–10 W at 514 nm (the most powerful argon ion line). Krypton ion lasers, which emit at 568 or 647 nm, are similar to their argon ion counterparts; however, they utilize ionized krypton plasma as gain medium. The 568 nm output can be used to activate the Rose Bengal peak at 559 nm, whereas the 647 nm most powerful line of krypton ion lasers, has been used for activation of m-THPC (652 nm).

Ideally, for optimal excitation, it is best to exactly tune the laser wavelength to the maxima of photosensitizer absorption bands. For this reason tunable organic dye lasers have been widely used for PDT. In these lasers the gain medium is an organic dye with a high quantum yield of fluorescence. Due to the broad nature of their fluorescence gain profile, tuning elements within the laser cavity can be used for selection of the lasing frequency within the gain band. Tunable dye lasers with the currently available dyes can quite easily cover the range from about 350–1000 nm. However, their operation is not of a 'turn key' nature since they require frequent replacement of their active material either for tuning to a different spectral region or for a better performance when the dye degrades. Tunable dye lasers also need to be pumped by some other light source like an argon ion laser, an excimer laser, a solid state laser (e.g., Nd:YAG 532 or 355 nm), a copper vapor laser, or lamps. Dye lasers have been used for the excitation of HpD at 630 nm, protoporphyrin IX at 635 nm, photoprotoporphyrin at 670 nm, and phthalocyanines around 675 nm. A further example is hypericin which, apart from the band at 550 nm, has a second absorption band at 590 nm. This is the optimum operation wavelength of dye lasers with rhodamine 590, better known as rhodamine 6G.

Dye lasers can operate in either pulsed or cw mode. However, there is a potential disadvantage in using a pulsed laser for PDT which becomes apparent when using lower repetition rates and higher pulse energies (>1 mJ per pulse) laser excitation. High pulse energies can induce saturation or transient bleaching of the sensitizer during the laser pulse and consequently much of the energy supplied is not efficiently absorbed by the sensitizer. It has been suggested that this effect accounts for the lack of photosensitized damage in tissue sensitized with a phthalocyanine and irradiated by a 5 Hz, 25 mJ-per-pulse flash lamp pumped dye laser. However, using a low pulse energy, high repetition rate copper vapor pumped dye laser (see below), the results were indistinguishable from cw irradiation with an Argon ion dye laser.

Another laser that has found clinical application in PDT is the copper vapor laser. In this laser the active medium is copper vapor at high temperature maintained in the tube by a repetitively pulsed discharge current. The copper vapor lasers are pulsed with typical pulse duration of about 50 ns and repetition rates reaching 20 kHz. Copper vapor lasers have been operated at quite high average powers up to about 40 W. The output radiation is produced at two distinct wavelengths, namely 511 and 578 nm both of which have found clinical use. Copper vapor pumped dye lasers have also been widely used for activating red-absorbing photosensitizers, and the analogous gold vapor lasers operating at 628 nm have been used to activate HpD.

A relatively new class of tunable lasers is the solid state Ti:Sapphire laser. The gain medium in this laser is a $\sim$1%Ti doped sapphire (Al_2O_3) crystal and its output may be tuned throughout the 690–1100 nm region, covering many photosensitizer absorption peaks. For example, most of the bacteriochlorin type photosensitizers have their absorption bands in that spectral region, e.g., m-THPBC with an absorption maximum at 740 nm. Also, Texaphyrin sensitizers absorb highly in this spectral region: lutetium texaphyrin has an absorption maximum at 732 nm.

Although tunable dye or solid-state lasers are widely used in laboratory studies for PDT, fixed-wavelength semiconductor diode lasers are gradually supplanting tunable lasers, owing to their practical convenience. It is now generally the case that when

a photosensitizer enters clinical trials or enters standard clinical use, laser manufacturers provide dedicated diode lasers with outputs matched to the chosen sensitizer. In this context there are diode lasers available at 630 nm for use with HpD, at 635 nm for use with ALA-induced protoporphyrin IX, 670 nm for ATSX-10 or phthalocyanines, or even in the region of 760 nm, for use with bacteriochlorins. The advantage of these diode lasers is that they are tailormade for use with a particular photosensitizer, they are highly portable, and they have a relatively easy turn-key operation.

So far we have concentrated on red wavelengths but most of the porphyrin and chlorin family of photosensitizers exhibit quite strong absorption bands in the blue region, known as 'Soret' bands. Despite the fact that tissue penetration in this spectral region is minimal, these bands are far more intense than the corresponding red absorption bands of the sensitizer. In this respect it is possible to activate these sensitizers at blue wavelengths, especially in the case of treatment of superficial (e.g., skin) malignant lesions. There are now blue diode lasers (and light-emitting diodes) for the activation of these bands, in particular at 405 nm where the Soret band of protoporphyrin IX lies or at 430 nm for selective excitation of the photoprotoporphyrin species. For very superficial lesions (e.g., in the retina) it may even be possible to use multiphoton excitation provided by femtosecond diode lasers operating at about 800 nm.

Clinical Applications of Lasers in PDT

In the previous section we reviewed the various types of lasers being used for PDT but their combination with fiber-optic delivery is also of key importance for their clinical application. The laser light may be delivered to the target lesion either externally using surface irradiation, or internally within the lesion which is denoted as interstitial irradiation, as depicted in **Figure 6**. For example, in case of superficial cutaneous lesions the laser irradiation is applied externally (**Figure 6a**). The area of the lesion is marked prior to treatment to determine the diameter of the laser spot. For a given fluence rate (100–200 W/cm^2) and the area of the spot, the required laser output power is then calculated. A multimode optical fiber, terminated with a microlens to ensure uniform irradiation, is then positioned at a distance from the lesion yielding the desired beam waste, and the surrounding normal tissue is shielded with dark material. However, if the lesion is a deeper-lying solid tumor, interstitial PDT is employed

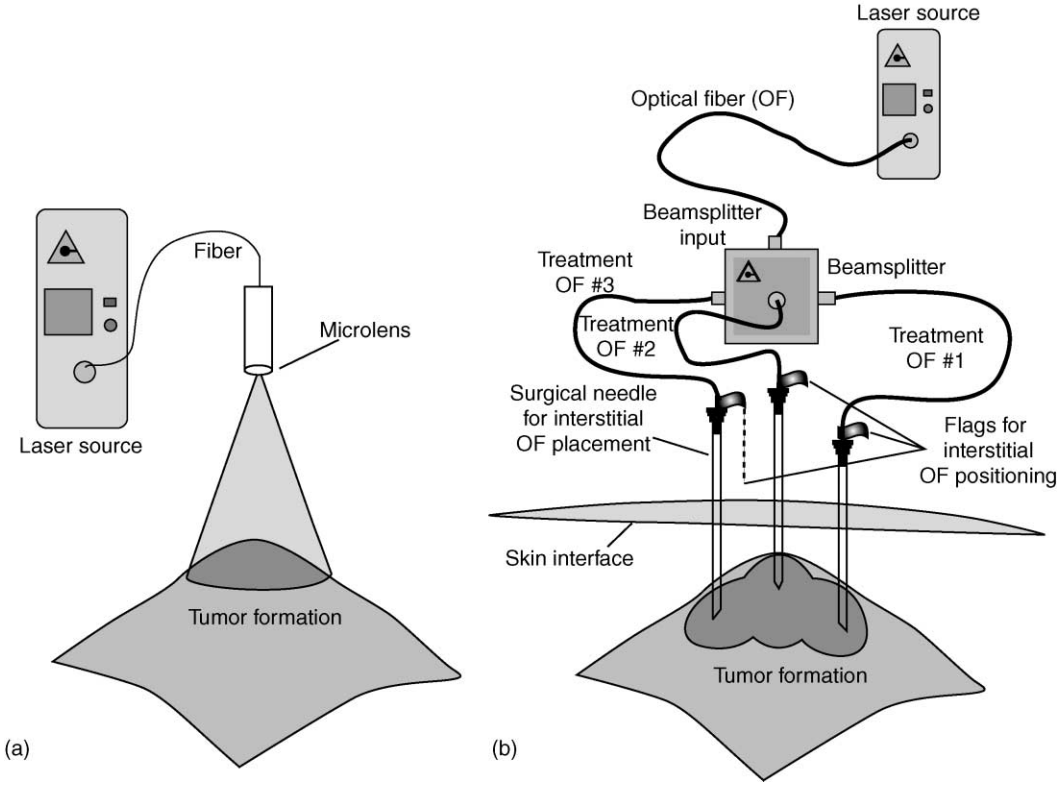


Figure 6 Clinical application of PDT. (a) Surface treatment. (b) Interstitial treatment.

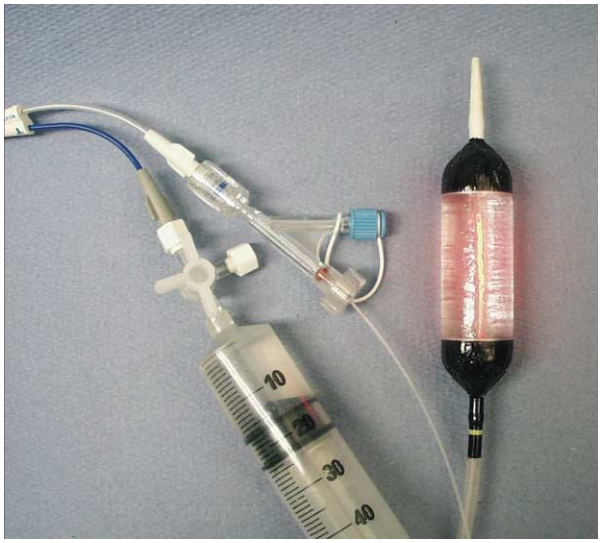


Figure 7 Balloon applicator and diffuser fiber used for PDT in Barrett's esophagus.

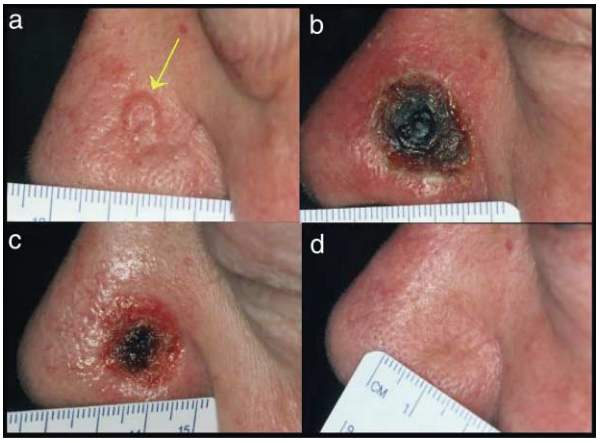


Figure 8 PDT treatment of a nasal basal cell carcinoma with m-THPC (Foscan®): (a) prior to PDT; (b) tissue necrosis a week after treatment; (c) four weeks after treatment; gradual recession of the necrosis and inflammation; (d) healing two months after treatment. Courtesy of Dr Alex Kuebler.

(Figure 6b). Surgical needles are inserted in an equidistant parallel pattern within the tumor, either freehand or under image (MRI/CT) guidance. Bare tip optical fibers are guided through the surgical needles to the lesion and flagged for position. A beamsplitter is used to divide the laser output into 2–4 components so that two to four optical fibers can be simultaneously employed. The laser power is adjusted so that the output fluence of each of the optical fibers is 100–200 mW. Even light distribution in the form of overlapping spheres approximately 1 cm diameter ensures the treatment of an equivalent volume of tissue around each fiber tip. Once each 'station' is likewise treated the fibers are repositioned accordingly, employing a pull-back technique in order for another station to be treated. Once the volume of the tumor is scanned in this manner the treatment is complete. Interstitial light delivery allows PDT to be used in the treatment of large, buried tumors and is particularly suitable for those in which surgery would involve extensive resection.

The treatment of tumors of internal hollow organs is also possible with PDT. The most representative example of this is the treatment of precancerous lesions in the esophagus, known medically as 'Barrett's esophagus'. In this case, a balloon applicator is used to house a special fiber with a light-diffusing tip or 'diffuser' (Figure 7). These diffusers are variable in length and transmit light uniformly along the whole of their length in order to facilitate treatment of circumferential lesions. The balloon applicator is endoscopically inserted into the patient. The balloon is inflated to stretch the esophageal walls and the flagged diffuser fiber is placed in position, central to the balloon optical window. In this way the whole of the lesion may be treated circumferentially at the same time.

Finally, the clinical response to PDT treatment is shown in Figure 8. The treated area becomes necrotic and inflamed within two to four days following PDT. This necrotic tissue usually sloughs off or is mechanically removed. The area eventually heals with minimal scarring (Figure 8).

Future Prospects

A major factor in the future development of PDT will be the application of relatively cheap and portable semiconductor diode lasers. High-power systems consisting of visible diode arrays coupled into multimode fibers with an output of several Watts, have recently become available which will greatly simplify the technical difficulties which have held back PDT. The use of new photosensitizers, in the treatment of certain tumors by PDT, is making steady progress, and although the toxicology testing and clinical evaluation are lengthy and costly processes, the expectation is that these compounds will become more widely available for patient use during the next five years. Regarding the current clinical status of PDT, Photofrin has recently been approved for treatment of lung and esophageal cancers in Europe, USA, and Japan. Clinical trials are in progress for several of the second-generation photosensitizers which offer significant improvements over Photofrin in terms of chemical purity, photoproperties, and skin

clearance. With the increasing clinical application of compact visible diode lasers the prospects for photodynamic therapy are therefore encouraging.

See also

Lasers: Dye Lasers; Free Electron Lasers; Metal Vapor Lasers. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Bonnet R (2000) *Chemical Aspects of Photodynamic Therapy*. Amsterdam: Gordon and Breach Science Publishers.

Dolmans DEJGJ, Fukumura D and Jain RK (2003) Photodynamic therapy for cancer. *Nature Reviews (Cancer)* 3: 380–387.

Hopper C (2000) Photodynamic therapy: a clinical reality in the treatment of cancer. *Lancet Oncology* 1: 212–219.

Milgrom L and MacRobert AJ (1998) Light years ahead. *Chemistry in Britain* 34: 45–50.

Rosen GM, Britigan BE, Halpern HJ and Pou S (1999) *Free Radicals Biology and Detection by Spin Trapping*. New York: Oxford University Press.

Svelto O (1989) *Principles of Lasers*, 3rd edn. New York: Plenum Press.

Vo-Dinh T (2003) *Biomedical Photonics Handbook*. Boca Raton, FL: CRC Press.

Pump and Probe Studies of Femtosecond Kinetics

G D Scholes, University of Toronto, Toronto, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Pump-probe spectroscopy is a ubiquitous time-resolved optical spectroscopy that has found application to the study of all manner of ultrafast chemical processes that involve excited states. The principle of the measurement is that a 'pump pulse' – usually an intense, short laser pulse – impulsively excites a sample, thus defining the start time for the ensuing photophysical dynamics. A probe pulse interrogates the system at later times in order to capture 'snap shots' of the state of the system. Many kinds of pump-probe experiments have been devised. The present article will mainly focus on transient absorption measurements.

Femtosecond pump-probe experiments have been employed to reveal the kinetics of excited state chemical processes such as solvation, isomerization reactions, proton transfer, electron transfer, energy transfer, or photochromism. It is becoming more common to use pump-probe techniques to study systems of increasing complexity, such as photosynthetic proteins, photoreceptor proteins, molecular aggregates, nanostructured materials, conjugated polymers, semiconductors, or assemblies for artificial light harvesting. However, the underlying chemical kinetics are not always easily revealed by pump-probe signals.

The principle of the pump-probe measurement is related to the seminal flash-photolysis experiment devised by Norrish and Porter. However, in order to achieve ultrafast time resolution – beyond that attainable by electronic detection – the probe pulse is temporally short and is controlled to arrive at the sample at variable time delays after the pump pulse has excited the sample. The change in intensity of the probe pulse, as it is transmitted through the sample, is monitored at each pump-probe time delay using a 'slow' detector that integrates over the probe pulse duration. One possible xprocess that can diminish the probe transmission is transient absorption. After formation of an excited state S_1 by the pump pulse, the probe pulse can monitor a resonance with a higher excited state $S_1 \rightarrow S_n$ absorption. Figure 1 shows transient absorption spectra, corresponding to excited state absorption at various time delays, that has been probed by a white light continuum probe and monitored by a multichannel CCD detector. The transient spectrum is seen to blue-shift with

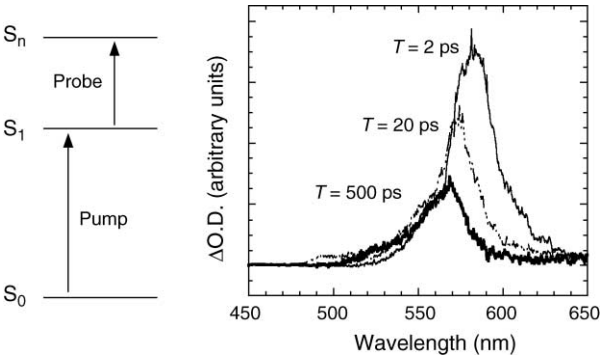


Figure 1 Excited state absorption spectra of 4′-*n*-pentyl-4-cyanoterphenyl in octanol solvent. The pump wavelength is 277 nm (100 nJ, 150 fs, 40 kHz) and the probe is a white light continuum (500 to 700 nm) generated in a sapphire crystal. The pump-probe delay T for each spectrum is indicated on the plot.

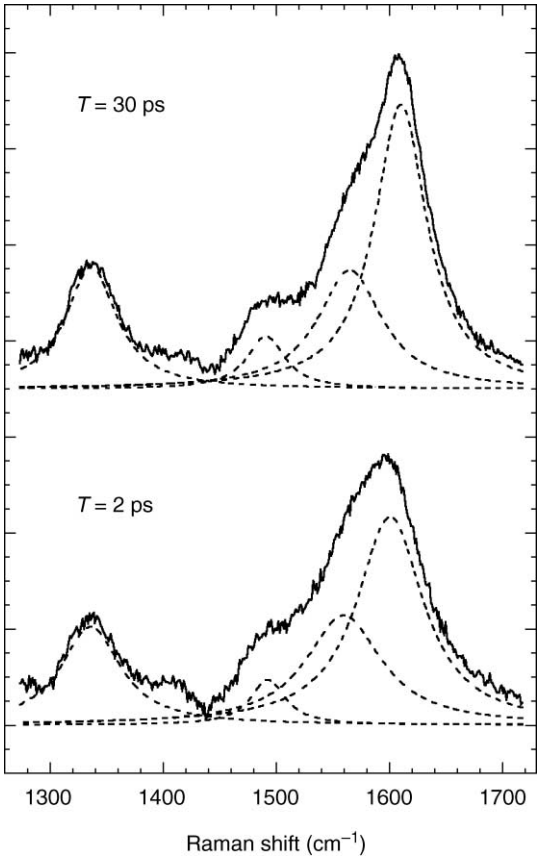


Figure 2 TR^3 spectra of 4′-*n*-pentyl-4-cyanoterphenyl in octanol solvent, recorded using the same set-up as summarized in Figure 1. The dashed lines are deconvolutions of the data to show the Raman bands.

pump-probe delay owing to solvation of the S_1 state of the molecule – the dynamic Stokes' shift. The dynamic Stokes' shift is the shift to lower energy of the emission spectrum as the solvent relaxes in response to the change in dipole moment of the excited electronic state compared to the ground state. The overall intensity of the induced absorption decays according to the lifetime of the excited state.

Although they will not be discussed in detail in the present article, various other pump-probe experiments are possible. For example, when in resonance with this transient absorption, the probe pulse can induce resonance Raman scattering, which reveals the evolution of the vibrational spectrum of the excited state. In this kind of pump-probe experiment we do not monitor the change in probe transmission through the sample. Instead, we monitor the difference between probe-induced resonance Raman scatter from the excited and ground states of the system. The resonance Raman (TR^3) scatter, time-resolved as a function of pump-probe delay, can be detected provided that it is not overwhelmed by fluorescence emission. Typical TR^3 data for different pump-probe delays are shown in Figure 2. The dashed lines show a deconvolution of the spectra to reveal the underlying vibrational bands. These bands are rather broad owing to the spectral width of the 1 ps laser pulses used in this experiment. It is clear that the time-resolution of a TR^3 experiment is limited to the picosecond regime, otherwise most of the vibrational band information is lost. TR^3 and complementary pump-probe infrared transient absorption experiments, utilizing an infrared probe beam, have been useful for observing the evolution of structural information, such as geometry changes, subsequent to photoexcitation.

In the present article we will describe the foundation of experiment and theory necessary to understand resonant ultrafast pump-probe spectroscopy applied to chemical processes in the condensed phase. By resonant, it is meant that the pump and probe pulses have frequencies (energies) resonant with electronic transitions of the molecules being studied. The implications of condensed phase are that the experiment interrogates an ensemble of molecules and the electronic transitions of these molecules are coupled to the random motions and environments of a bath, for example the solvent.

Experimental Measurement of Pump-Probe Data

The experimental setup for a one-color pump-probe experiment (i.e. pump and probe of the same color) is shown in Figure 3. The setup is easily adapted for a two-color experiment. Most of the laser intensity (70%) is transmitted through the first beam splitter so that the pump is more intense than the probe. A second beam splitter is used to split off a small amount of probe light for the reference beam. Note that the arrangement of these beam splitters results in both the pump and probe passing through the same amount of dispersive material on their way to the sample. The pump and probe beams each propagate through half-wave plates in order to control polarization. Usually the probe polarization is set to the magic angle (54.7°) relative to the pump in order to remove polarization bias. The beams are sent towards the sample by way of retroreflectors mounted on $x-y-z$ translation stages. The pump retroreflector is mounted on a precision computer-controlled translation stage such that it can be scanned in the x-direction to arrive at variable delays before and after the probe. The pump and probe beams are aligned using the $y-z$ controls on the retroreflector translation stages together with the pick-off mirrors

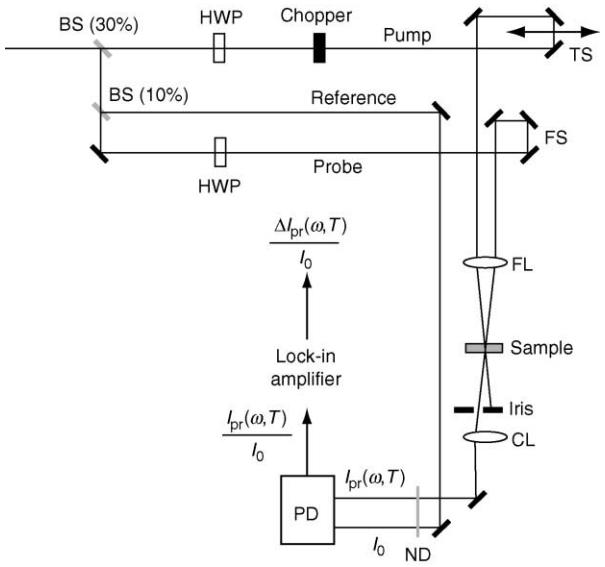


Figure 3 An experimental layout for femtosecond pump-probe spectroscopy. The reference beam ideally passes through the sample (not shown as such in aid of clarity). See text for a description. HWP, half-wave plate; BS, beamsplitter (% reflection); FL, focusing lens; CL, collimating lens; ND, neutral density filter; PD, photodetectors.

such that they travel in parallel towards the sample. The overall setup of the beams is in a box geometry so that pump and probe arms have the same path length.

The pump and probe beams are focused into the sample using a transmissive (i.e. lens) or reflective optic FL. The focal length is typically 20 to 30 cm, providing a small crossing angle to improve phase matching. The pump beam should have a larger spot size at its focal point in the sample than the probe beam, in order to avoid artifacts arising from the wandering of the pump and probe spatial overlap as a function of delay time – a potential problem, particularly for long delays. Good spatial overlap of the beams in the sample is important. It can be checked using a 50 μm pinhole. Fine adjustment of the pump-probe delay, to establish temporal overlap of the pump and probe pulses, can be dictated by autocorrelating the beams in a second harmonic generating crystal mounted at the sample position. At the same time the pulse compression is tweaked to ensure that the pulse dispersion in the experiment has been pre-compensated.

The sample is usually flowed or mounted in a spinning cell if it is liquid, in order to avoid thermal effects or photochemical bleaching. The path length of the sample is typically ≤ 1 mm. The transmitted probe beam is spatially isolated from the pump beam using an iris, is collimated, and then directed onto the photodetector.

The reference beam, which provides the relative reference intensity I_0 should preferably pass through the sample (away from the pump spot). This provides the probe-only transmission – that is, the reference beam intensity is attenuated according to sample ground state optical density. The photodetector outputs are used to ratio the pump-probe arm intensity transmitted through the sample with the reference intensity, $I_{pr}(\omega, T)/I_0$. Since the pump beam is being chopped at the lock-in reference frequency, the lock-in amplifier outputs the pump-induced fractional change in transmission intensity, $\Delta I_{pr}(\omega, T)/I_0$, usually simply written as $\Delta T/T$. When the pump-induced signal I_{pr} is small compared to the reference intensity I_0, then the detected $\Delta I_{pr}(\omega, T)/I_0$ signal is approximately equal to the change in optical density, ΔO.D.:

$$\Delta\text{O.D.} = \log\left(\frac{I_0 - \Delta I_{pr}}{I_0}\right) \approx \frac{-\Delta I_{pr}}{I_0} \equiv \frac{-\Delta T}{T} \quad [1]$$

Here we have used a Taylor expansion $\log(1 + x) \approx x$ for small x.

The power dependence of the signal in the $\chi^{(3)}$ limit should be linear in both the pump and probe intensities. It is also possible to pump the sample using two-photon absorption, then probe an excited state absorption. This is a $\chi^{(5)}$ experiment, so that the signal depends quadratically on the pump intensity.

What is Measured in a Pump-Probe Experiment?

Origin of the Signal

The pump-probe measurement is a third-order nonlinear spectroscopy. A complete calculation of the signal requires computation of the third-order polarization induced by the interaction of the pump and probe with the material sample, together with an account for the kinetic evolution of populations (e.g., excited state reactant and product states, etc.).

The pump pulse, wavevector $\mathbf{k}_1$, interacts twice with the sample, thereby creating excited state population density $|e\rangle\langle e|$ and a hole-in-the-ground state population density $|g\rangle\langle g|$. These population densities propagate until the probe pulse, with wavevector $\mathbf{k}_2$, interacts with the system to induce a polarization that depends on the state of the system at time T after the pump. This induced polarization $P_s(t)$ is radiated in the $\mathbf{k}_s = \mathbf{k}_1 - \mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_2$ direction. Because the signal is radiated in the probe direction, the probe pulse acts as a local oscillator to heterodyne the signal. The change in probe-pulse spectral intensity after transmission through an

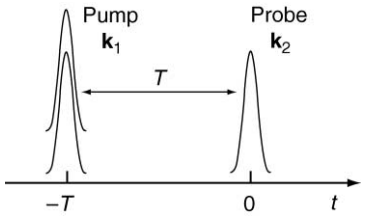


Figure 4 Labeling and time variable for the description of a pump-probe experiment. T defines the pump-probe delay.

optically thin sample of path length l, with concentration of absorbers c and refractive index n, is given by eqn [2]

$$\Delta I_{\mathrm{pr}}(\omega,T) \propto \frac{cl\omega_{\mathrm{pr}}}{n}\mathrm{Im}\left[\int dt E^{*}_{\mathrm{pr}}(t)P^{(3)}(0,T,t)e^{i\omega t}\right] \quad [2]$$

In eqn [2] $E_{\mathrm{pr}}(t)$ is the probe pulse electric field, ω is the center frequency of the probe, and * indicates complex conjugate. The time variables correspond to those indicated in Figure 4, and the signal is resolved as a function of frequency by Fourier transformation of $t \rightarrow \omega$. Experimentally this is achieved by dispersing the transmitted probe using a spectrograph. The first-time variable, the time interval between the two pump pulse interactions, is set to zero because we are interested here in the limit where the pulse duration is much shorter than T. In this regime – impulsive pump-probe – the measurement is sensitive to formation and decay of excited state species.

The induced polarization is given by eqn [3]

$$P^{(3)}(0,T,t) = \int_0^\infty \mathrm{d}t \int_0^\infty \mathrm{d}T[R^{\mathrm{SE}} - R^{\mathrm{ESA}} + R^{\mathrm{GSR}}] \times E_1^{*}(t-T)E_1(t-T)E_2(t) \quad [3]$$

where the response functions that contain information about the lineshape functions and kinetics are given by eqn [4]:

$$R^{\mathrm{SE}}(0,T,t) = |\mu_{eg}|^2|\mu_{eg}|^2 \exp(-i\omega_{eg}t) \times \exp[-g^{*}(t) + 2ig''(T+t) - 2ig''(T)] \times K^{\mathrm{SE}}(T) \quad [4a]$$

$$R^{\mathrm{ESA}}(0,T,t) = |\mu_{eg}|^2|\mu_{fe}|^2 \exp(-i\omega_{fe}t) \times \exp[-g^{*}(t) - 2ig''(T+t) + 2ig''(T)] \times K^{\mathrm{ESA}}(T) \quad [4b]$$

$$R^{\mathrm{GSR}}(0,T,t) = |\mu_{eg}|^2|\mu_{eg}|^2 \exp(-i\omega_{eg}t)\exp[-g(t)] \times K^{\mathrm{GSR}}(T) \quad [4c]$$

The overall intensity of each contribution to the signal is scaled by the transition dipole moment that connects the ground and excited state μ_{eg} or excited state and a higher excited state μ_{fe}. The intensity is further influenced by resonance of the probe-pulse spectrum with the transition frequencies ω_{eg} and ω_{fe}. These transition energies have a time-dependence owing to the relaxation of the excited state density in the condensed phase environment – the dynamic Stokes' shift. The details of the dynamic Stokes' shift depend on the bath, and are therefore contained in the imaginary part of the lineshape function $g(t) = g'(t) + ig''(t)$. The evolution of the Stokes' shift with pump-probe delay T is clearly seen in the excited state absorption spectrum shown in Figure 1. The lineshape function contains all the details of the timescales of fluctuations of the bath and the coupling of the electronic transition of the molecule to these fluctuations. The lineshape function can be cast in the simple form given by eqn [5] using the Brownian oscillator model in the high temperature, high friction limit:

$$g(t) \approx (2\lambda k_B T_B/\hbar\Lambda)[\Lambda t - 1 + \exp(-\Lambda t)] + i(\lambda/\Lambda)[1 - \exp(-\Lambda t)] \quad [5]$$

The Brownian oscillator model attributes fluctuations of the transition frequencies to coupling between electronic states and an ensemble of low-frequency bath motions. In eqn [5]; λ is the solvent reorganization energy (half the Stokes' shift) and Λ is the modulation frequency of the solvent oscillator (τ_{L}^{-1} for a Debye solvent, where τ_{L} is the longitudinal dielectric relaxation time); k_{B} is the Boltzmann constant, and T_{B} is the temperature.

Information about the kinetics of the evolution of the system initiated by the pump pulse is contained in the terms K^{SE}, K^{ESA}, and K^{GSR}. These terms will be discussed in the next section.

The evolution of the excited state density is mapped onto the response functions R^{SE} and R^{ESA}. R^{SE} is the contribution arising from the probe pulse stimulating emission from the excited state e in the probe direction, which increases the probe intensity. Thus the spectrum of the signal arising from R^{SE} resembles the fluorescence emission of the sample, as shown in Figure 5. The excited state absorption contribution to the signal R^{ESA} depletes the probe intensity, and corresponds to an e $\rightarrow$ f electronic resonance, such as that shown in Figure 1. Note that the ESA signal contributes to the overall $\Delta I_{\mathrm{pr}}(\omega, T)$ with an opposite sign to the SE and GSR contributions. The ground-state recovery term R^{GSR} arises from depletion of the ground-state population by the pump pulse, thereby increasing the transparency of the sample over the

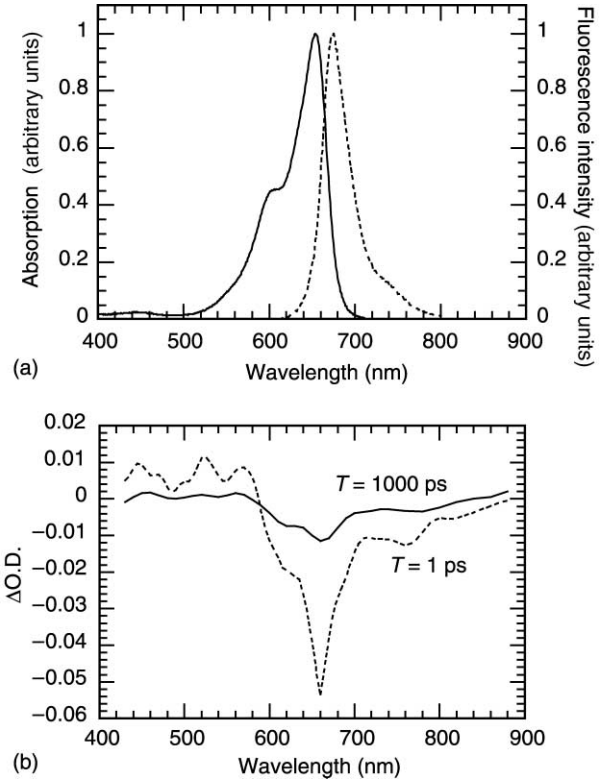


Figure 5 (a) shows the absorption and emission spectrum of a dilute solution of oxazine-1 in water. (b) shows the transient absorption spectrum of this system at various delays after pumping at 600 nm using picosecond pulses. The transient spectrum, ground state recovery and stimulated emission resemble the sum of the absorption and emission spectra.

spectrum of the ground-state absorption (see Figure 5). Thus the transmitted probe intensity is increased relative to that without the pump pulse. The GSR contribution to $\Delta I_{pr}(\omega, T)$ decreases with T according to the excited state lifetime of the photo-excited species.

The Coherent Spike

To calculate the pump-probe signal over delay times that are of the order of the pulse duration, one must calculate all possible ways that a signal can be generated with respect to all possible time orderings of the two pump interactions and the probe interactions (Liouville space pathways). Such a calculation reveals coherent as well as sequential contributions to the response of the system owing to the entanglement of time orderings that arise from the overlap of pulses of finite duration. The nonclassical, coherent, contributions dominate the signal for the time period during which pump and probe pulses overlap, leading to the coherent spike (or coherent 'artifact') (Figure 6).

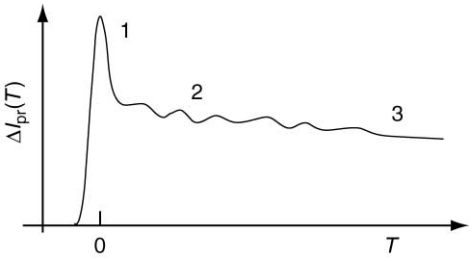


Figure 6 A schematic depiction of a pump-probe signal. The coherent spike is labeled 1. In region 2 we can see quantum beats as a sum of sinusoidal modulations of the signal. The label 3 denotes the long time decay of the signal, as determined by the population dynamics, K^{SE}, K^{GSR}, and K^{ESA}.

Nuclear Wavepackets

The electronic absorption spectrum represents a sum over all possible vibronic transitions, each weighted by a corresponding nuclear overlap factor according to the Franck–Condon principle. It was recognized by Heller and co-workers that in the time-domain representation of the electronic absorption spectrum, the sum over vibrational overlaps is replaced by the propagation of an initially excited nuclear wavepacket on the excited state according to $|i(t)\rangle = \exp(-Ht/\hbar)\,|i(0)\rangle$. Thus the absorption spectrum is written as

$$\sigma_i(\omega) = \frac{2\pi\omega}{3c}\int_{-\infty}^{\infty} \mathrm{d}t\, \exp[-i(\omega - \omega_{eg}^{i})t - g(t)] \times \langle i(0)\,|i(t)\rangle \qquad [6]$$

for mode i. The Hamiltonian H is determined by the displacement of mode i, Δ_i. A large displacement results in a strong coupling between the vibration and the electronic transition from state g to e. In the frequency domain this is seen as an intense vibronic progression. In the time domain this translates to an oscillation in the time-dependent overlap $\langle i(0)\,|i(t)\rangle$, with frequency ω_i, thereby modulating the intensity of the linear response as a function of t.

It follows that an impulsively excited coherent superposition of vibrational modes, created by an optical pulse that is much shorter than the vibrational period of each mode, propagates on the excited state as a wavepacket. The oscillations of this wavepacket then modulate the intensity of the transmitted probe pulse as a function of T to introduce features into the pump-probe signal known as quantum beats (see Figure 6). In fact, nuclear wavepackets propagate on both the ground and excited states in a pump-probe experiment. The short probe pulse interrogates the evolution of the nuclear wavepackets by projecting them onto the manifold of vibrational wavefunctions in an excited or ground state.

Femtosecond Kinetics

According to eqn [4] the kinetics of the evolution of the system initiated by the pump pulse are governed by the terms K^{SE}, K^{ESA}, and K^{GSR}. Each of these terms describes the kinetics of decay, production, or recovery of excited state and ground state populations. In the simplest conceivable case, each might be a single exponential function. The stimulated emission term K^{SE} contains information about the depopulation of the initially excited state, which for example, may occur through a chemical reaction involving the excited state species or energy transfer from that state. The excited state absorption term K^{ESA} contains information on the formation and decay of all excited state populations that give a transient absorption signal. The ground-state recovery term K^{GSR} contains information on the time-scales for return of all population to the ground state, which will occur by radiative and nonradiative processes. Usually the kinetics are assumed to follow a multi-exponential law such that

$$K^m(T, \lambda_{exc}, \omega) = \sum_{j=1}^{N} A_j(\lambda_{exc}, \omega) \exp(-t/\tau_j) \qquad [7]$$

where the amplitude A_j, but not the decay time coefficient τ_j, of each contribution in the coupled N-component system depends on the excitation wavelength λ_{exc} and the signal frequency ω. The index m denotes SE, ESA, or GSR.

In order to extract meaningful physical parameters from analysis of pump-probe data, a kinetic scheme in the form of eqn [7] needs to be constructed, based on a physical model to connect the decay of initially excited state population to formation of a product excited state and subsequent recovery of the ground-state population. An example of such a model is depicted in Figure 7. Here the stimulated emission contribution to the signal is attenuated according to the rate that state 1 goes to 2. The transient absorption appears at that same rate and decays according to the radiationless process 2 to 3. The rate of ground state recovery depends on 1 to 2 to 3 to 5 and possibly pathways involving the isomer 4.

The kinetic scheme is not easily extracted from pump-probe data because contributions from each of K^{SE}, K^{ESA}, and K^{GSR} contribute additively at each detection wavelength ω, as shown in eqns [2]–[4]. Moreover, additional, wavelength-dependent, decay or rise components may appear in the data on ultrafast time-scales owing to the time-dependent Stokes' shift (the $g''(t)$ terms indicated in eqns [4]). A spectral shift can resemble either a decay or rise component. In addition, the very fastest dynamics may be hidden beneath the coherent spike. For these reasons, simple fitting of pump-probe data usually cannot reveal the underlying kinetic model. Instead the data should be simulated according to eqns [3] and [4].

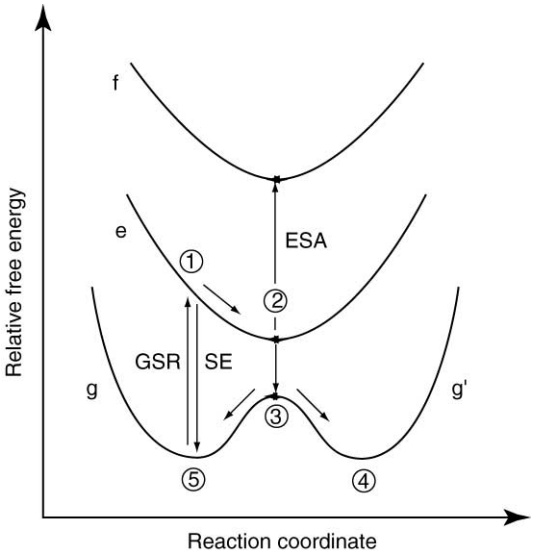


Figure 7 Free energy curves corresponding to a model for an excited state isomerization reaction, with various contributions to a pump-probe measurement indicated. See text for a description. (Courtesy of Jordanides XJ and Fleming GR.)

Alternative strategies involve global analysis of wavelength-resolved data to extract the kinetic evolution of spectral features, known as 'species associated decay spectra'. Global analysis of data effectively reduces the number of adjustable parameters in the fitting procedure and allows one to obtain physically meaningful information by associating species with their spectra. These methods are described in detail by Holzwarth.

List of Units and Nomenclature

ESA	excited state absorption
GSR	ground state recovery
SE	stimulated emission
S_n	nth singlet excited state
TR^3	time-resolved resonance Raman
$\chi^{(3)}$	nth-order nonlinear optical susceptibility

See also

Coherent Lightwave Systems. **Coherent Transients:** Ultrafast Studies of Semiconductors. **Nonlinear Optics, Applications:** Raman Lasers. **Optical Amplifiers:** Raman, Brillouin and Parametric Amplifiers. **Scattering:** Raman Scattering. **Spectroscopy:** Raman Spectroscopy.

Further Reading

Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. New York: Oxford University Press.

Heller EJ (1981) The semi-classical way to molecular spectroscopy. *Accounts of Chemical Research* 14: 368–375.

Holzwarth AR (1996) Data analysis of time-resolved measurements. In: Amesz J and Hoff AJ (eds) *Biophysical Techniques in Photosynthesis*, pp. 75–92. Dordrecht, Germany: Kluwer.

Ippen EP, Shank CV, Wiensenfeld JM and Migus A (1980) Subpicosecond pulse techniques. *Philosophical Transactions of the Royal Society of London A* 298: 225–232.

Jonas DM and Fleming GR (1995) Vibrationally abrupt pulses in pump-probe spectroscopy. In: El-Sayed MA, Tanaka I and Molin Y (eds) *Ultrafast Processes in Chemistry and Photobiology*, pp. 225–256. Oxford, UK: Blackwell Scientific.

Klimov VI and McBranch DW (1998) Femtosecond high-sensitivity, chirp-free transient absorption spectroscopy using kilohertz lasers. *Optics Letters* 23: 277–279.

Lee SY (1995) Wave-packet model of dynamic dispersed and integrated pump-probe signals in femtosecond transition-state spectroscopy. In: Manz J and Wöste L (eds) *Femtosecond Chemistry*, vol. 1, pp. 273–298. Weinheim, Germany: VCH.

Lin SH, Alden R, Islampour R, Ma H and Villaeys AA (1991) *Density Matrix Method and Femtosecond Processes*. Singapore: World Scientific.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. New York: Oxford University Press.

Norrish RGW and Porter G (1949) Chemical reactions produced by very high light intensities. *Nature* 164: 658.

Towrie M, Parker AW, Shaikh W and Matousek P (1998) Tunable picosecond optical parametric generator-amplifier system for time resolved Raman spectroscopy. *Measurement Science and Technology* 9: 816–823.

Yan YJ and Mukamel S (1990) Femtosecond pump-probe spectroscopy of polyatomic molecules in condensed phases. *Physical Reviews A* 41: 6485–6504.

Time-Correlated Single-Photon Counting

A Beeby, University of Durham, Durham, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Time-correlated single-photon counting, TCSPC, is a well-established technique for the determination of fluorescence lifetimes and related time-resolved fluorescence properties. Since the 1980s, there have been dramatic changes in the laser sources used for this work, which in turn have opened up the technique to a wide user base.

TCSPC: The Basics

In the TCSPC experiment, the sample is repeatedly excited by a high-repetition rate, low-intensity light source. Fluorescence from the sample is collected and may be passed though a cut-off filter or monochromator to remove scattered excitation light and to select the emission wavelength. The fluorescence is then focused onto a detector, typically a photomultiplier (PMT) or single-photon avalanche diode (SPAD) which detects single emitted photons. The 'raw' signal from the detector fluctuates considerably from event to event. In order to remove any timing error that could be induced by this, the signal is processed first by a constant fraction discriminator (CFD) in order to provide a stable and consistent timing pulse. This signal is used as an input for the time-to-amplitude converter (TAC) which determines the time interval between the excitation of the sample and the emission of the photon. The output from the TAC is a relatively slow pulse whose intensity is proportional to the time interval between the start and stop signals. A pulse height analyzer (PHA) is used to process the output from the TAC and increments a channel in its memory corresponding to a time window during which the photon was detected. As the measurement is repeated many times over, a histogram is constructed in the memory of the PHA, showing the number of photons detected as a function of the time interval. This histogram corresponds to the fluorescence decay.

As outlined below, it is desirable to have a light source with a moderately high repetition rate, typically of the order of MHz. Because it is not practical to start the TAC each time the light source 'fires', it is more usual to operate the TAC in the so-called 'reversed mode', whereby the start signal is derived from the detected photon and the stop is derived from the light source. The electronic circuitry is illustrated in Figure 1. In practice, many researchers still use modular components that are wired together manually, although there is increasingly a tendency towards the use of single PC-based cards which contain all the necessary discriminators, timing circuits, and data acquisition electronics. There are a number of commercial PC

cards that contain all the timing electronics, etc., required for TCSPC, for example, Becker and Hickl (Germany).

This approach to obtaining the fluorescence decay profile can only work if the detected photons are 'randomly selected' from the total emission from the sample. In practice, this means that the rate of detection of photons should be of the order of 1% of the excitation rate: that is, only one in a hundred excitation pulses gives rise to a detected photon. This histogram mirrors the fluorescence decay of the sample convolved with the instrument response function (IRF). The IRF is itself a convolution of the temporal response of the optical light pulse, the optical path in the collection optics, and the response of the detector and ancillary electronics. Experimentally, the IRF is usually recorded by placing a scattering material in the spectrometer.

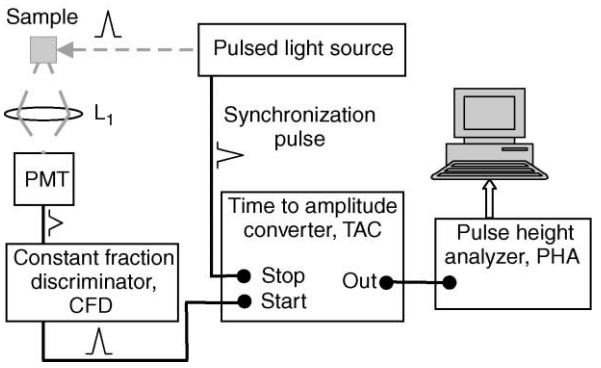


Figure 1 Schematic diagram of a TSCPC spectrometer. Fluorescence from the sample is collected by L_1, which may also include a monochromator/wavelength selection filter. A synchronization pulse is obtained either from the laser driver electronics directly, or via a photodiode and second constant fraction discriminator. The TAC is shown operating in a 'reversed mode' as is normal with a high repetition rate laser source.

In principle, the decay function of the sample can be extracted by deconvolution of the IRF from the decay function, although in practice it is more common to use the method of iterative re-convolution using a model decay function(s). Thus, the experimentally determined IRF is convolved with a model decay function and parameters within the function are systematically varied to obtain the best fit for a specific model. The most commonly used model assumes the decaying species to follow one or more exponential functions (sum of exponentials) or a distribution of exponentials. Further details regarding the mechanics of data analysis and fitting procedures can be found in the literature. A typical fit is illustrated in Figure 2.

The main aim of this review is to illustrate how recent advances in laser technology have changed the face of TCSPC and to discuss the types of laser light sources that are in use. The easiest way of approaching this topic is to define the criteria for an ideal light source for this experiment, and then to discuss the ways in which new technologies are meeting these criteria.

High Repetition Rate

A typical decay data set may contain something of the order of 0.1–5 million 'counts': it is desirable to

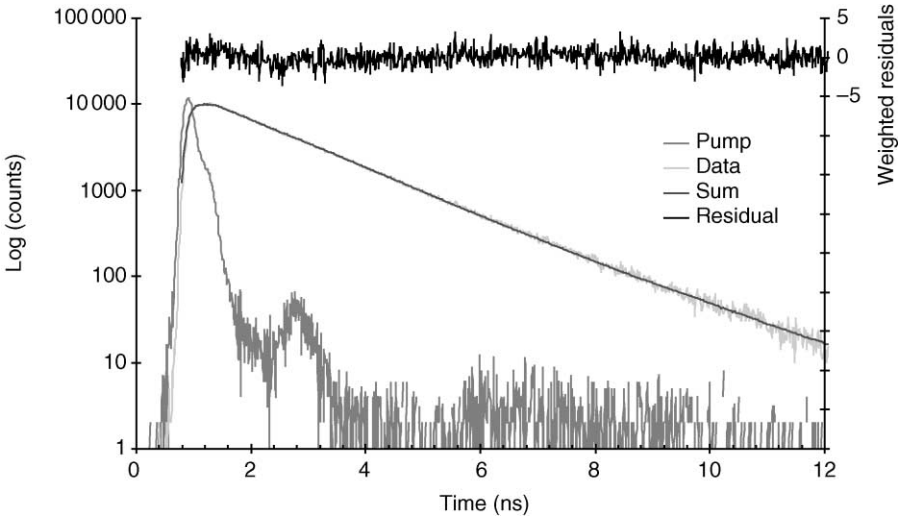


Figure 2 The fluorescence decay of Rhodamine B in water. The sample was excited using a 397 nm pulsed laser diode (1 MHz, $<$200 ps FWHM), and the emission was observed at 575 nm. Deconvolution of the IRF (dark gray trace) and observed decay (light gray trace) with a single exponential function gave a lifetime of 1.50 ns, with a $\chi^2 R = 1.07$, Durbin–Watson parameter = 1.78. The fitted curve is shown overlaid in gray, and the weighted residuals are shown offset in black.

acquire such a number of events in order that we can reliably employ statistical parameters to judge the quality of the fitted decay function(s). As stated above, the rate of photon acquisition should be of the order of 1% of the excitation rate in order to obtain reliable, good-quality data. It is clear then that a high repetition rate source is desirable in order to reduce the overall data acquisition time. However, it is important that the repetition rate of the source is not too high, otherwise the data show 'overlap' effects, caused by the fluorescence intensity from one excitation pulse not completely decaying before the next laser pulse arrives at the sample. Under these conditions the observed fluorescence decay is perturbed and extraction of complex kinetic behavior is more difficult. As a general rule of thumb, it is recommended that the inter-pulse separation is greater than five times the lifetime of the longest component of the fluorescence decay. Fluorescence lifetimes can be as long as 100 ns, indicating a maximum repetition rate of 2 MHz, although when investigating fluorophores with a shorter lifetime correspondingly higher frequencies can be selected. Clearly the ability to vary the repetition rate according to the type of sample under investigation is an advantage.

Short Pulse Duration

Typically the minimum lifetime that can be determined by a TCSPC spectrometer is limited by the instrument response function: as a general rule it is possible to extract fluorescence lifetimes of about one tenth of the full width at half maximum (FWHM) of the IRF by reconvolution methods. As discussed above, the IRF is the product of a number of factors, including the optical pulse duration, differences in the optical path traversed by the fluorescence, the response time of the detector, and the temporal response of the associated timing electronics. Ideally the excitation source should be as short as possible, and with modern laser-based TCSPC systems, it is usually the case that the time-response of the detector is the limiting factor. The introduction of high-speed microchannel plate (MCP) photomultipliers have lead to very short IRFs, which can be <50 ps FWHM. When using these detectors is it essential to keep the optical transit-time spread as low as possible: that is, the path taken by light through a monochromator should be independent of wavelength and trajectory taken through the optics. Many research groups advocate the use of a subtractive dispersive double monochromator for this purpose.

Tunability

Early TCSPC experiments were carried out using hydrogen- or nitrogen-filled flashlamps which produced broadband ns-duration pulses with low repetition rates (<100 kHz). Variable wavelength excitation is possible using these sources, but their low intensity and low repetition rates mean that data acquisition periods can be very long. The early 1980s saw the introduction of cavity-dumped ion lasers and synchronously pumped mode-locked cavity-dumped dye lasers, which provided significantly higher repetition rates, typically up to 4 MHz, and greater intensities. However, the TCSPC experiments that could be carried out using these laser-based sources were somewhat limited by the wavelengths that could be generated by the argon ion and synchronously pumped dye lasers, and their second harmonics. The dye lasers were restricted to the use of Rhodamine 6G as the gain medium, which has a limited tuning range ($\sim$580–620 nm). It is often desirable to vary the excitation wavelength as part of a photophysical study, either as a systematic study of the effects of excitation energy upon the excited state behavior, or in order to optimize excitation of a particular chromophore in a system. To some extent in the past, the wavelengths available from the laser systems dictated the science that they could be used to address. The mid- to late-1980s saw the introduction of mode-locked continuous wave (CW) Nd:YAG lasers as pump lasers and increased number of dyes that could be synchronously pumped, providing a broader range of excitation wavelengths as well as increased stability. One drawback of these complex and often fickle laser systems was the level of expertise and attention they required to keep their performance at its peak.

Since the 1990s, the introduction of two important new types of laser have had a significant impact on the use of TCSPC as an investigative method. The first is based upon the mode-locked titanium sapphire laser which offers tuneable radiation from the deep UV to the NIR, whilst the second series of laser are based upon (relatively) low-cost, turn-key solid state diode lasers. The developments and merits of these new laser sources, which now dominate the market for TCSPC sources, are discussed below.

Ti-Sapphire

The discovery of the titanium doped sapphire laser has led to a revolution in ultrafast laser sources. Coupled with efficient diode-pumped solid state

Nd:YAG pump lasers, this medium currently dominates the entire ultrafast laser market, from high repetition, low pulse energy systems through to systems generating very high pulse energies at kHz repetition rates.

Ti-sapphire is, in many respects, an ideal laser gain medium: it has a high energy storage capacity, it has a very broad-gain bandwidth, and it has good thermal properties. Following the first demonstration of laser action in a Ti-sapphire crystal, the CW Ti-sapphire laser was quickly commercialized. It proved to be a remarkable new laser medium, providing greater efficiency, with outputs of the order of 1 W. The laser also provided an unprecedented broad tuning range, spanning from just below 700 nm to 1000 nm. The most important breakthrough for time-resolved users came in 1991, when self-mode-locking in the medium was demonstrated. This self-mode-locking, sometimes referred to as Kerr-lens mode-locking, arises from the nonlinear refractive index of the material and is implemented by simply placing an aperture within the cavity at an appropriate point. The broad-gain bandwidth of the medium results in very short pulses: pulse durations of the order of tens of femtoseconds can be routinely achieved with commercial laser systems.

The mode-locked Ti-sapphire laser was quickly seized upon and commercialized. The first generation of systems used multiwatt CW argon ion lasers to pump them, making the systems unwieldy and expensive to run, but in recent years these have been replaced by diode-pumped Nd:YAG lasers, providing an all solid-state, ultrafast laser source. Equally significantly, the current generation of lasers are very efficient and do not require significant electrical power or cooling capacity. There are several commercial suppliers of pulsed Ti-sapphire lasers, including Coherent and Spectra Physics.

A typical mode-locked Ti-sapphire laser provides an 80 MHz train of pulses with a FWHM of ~100 fs, with an average power of up to 1.5 W and are tuneable from 700 nm to almost 1000 nm. The relatively high peak powers associated with these pulses means that second- or even third-harmonic generation is relatively efficient, providing radiation in the ranges 350–480 nm and 240–320 nm. The output powers of the second and third harmonics far exceeds that required for TCSPC experiments, making them an attractive source for the study of fluorescence lifetime measurements. The simple mode-locked Ti-sapphire laser has some drawbacks: the very short pulse duration provides a broad-bandwidth excitation, the repetition rate is often too high with an inter-pulse separation of ~12.5 ns, and the laser does not provide continuous tuning from the deep-UV though to the visible. Incrementally, all of these points have been addressed by the laser manufacturers.

As well as operating as a source of femtosecond pulses, with their associated bandwidths of many tens of nm, some modern commercial Ti-sapphire lasers can, by adjustment of the optical path and components, be operated in a mode that provides picosecond pulses whilst retaining their broad tuning curve. Under these conditions, the bandwidth of the output is greatly reduced, making the laser resemble a synchronously pumped mode-locked dye laser.

As stated above, a very high repetition rate can be undesirable if the system under study has a long fluorescence lifetime due to the fluorescence decay from an earlier pulse not decaying completely before the next excitation pulse. Variation of the repetition of output frequency can be addressed by two means. The first is to use a pulse-picker. This is an opto-accoustic device which is placed outside the laser cavity. An electrical signal is applied to a device, synchronized to the frequency of the mode-locked laser to provide pulses at frequencies from ~4 MHz down to single-shot. The operation of a pulse picker is somewhat inefficient, with less than half of the original pulse energy being obtained in the output, although output power is not usually an issue when using these sources. More importantly, when using pulse-pickers, it is essential to ensure that there is no pre- or post-pulse which can lead to artifacts in the observed fluorescence decays. The second means of reducing the pulse duration is cavity dumping. This is achieved by extending the laser cavity and placing an opto-acoustic device, known as a Bragg cell, at a beam-waist within the cavity. When a signal is applied to the intracavity Bragg cell, synchronized with the passage of a mode-locked pulse though the device, a pulse of light is extracted from the cavity. The great advantage of cavity-dumping is that it builds up energy within the optical cavity and when pulses are extracted these are of a higher pulse energy than the mode-locked laser alone. This in turn leads to more efficient harmonic generation. This addition of the cavity dumper system extends the laser cavity and hence reduces the overall frequency of the mode-locked laser to ~50 MHz, and can provide pulses at 9 MHz down to single shot.

The extension of the tuning range of the Ti-sapphire laser has been facilitated by the introduction of the optical parametric oscillator (OPO). In the OPO, a short wavelength pump beam passes through a nonlinear optical crystal and generates two tuneable

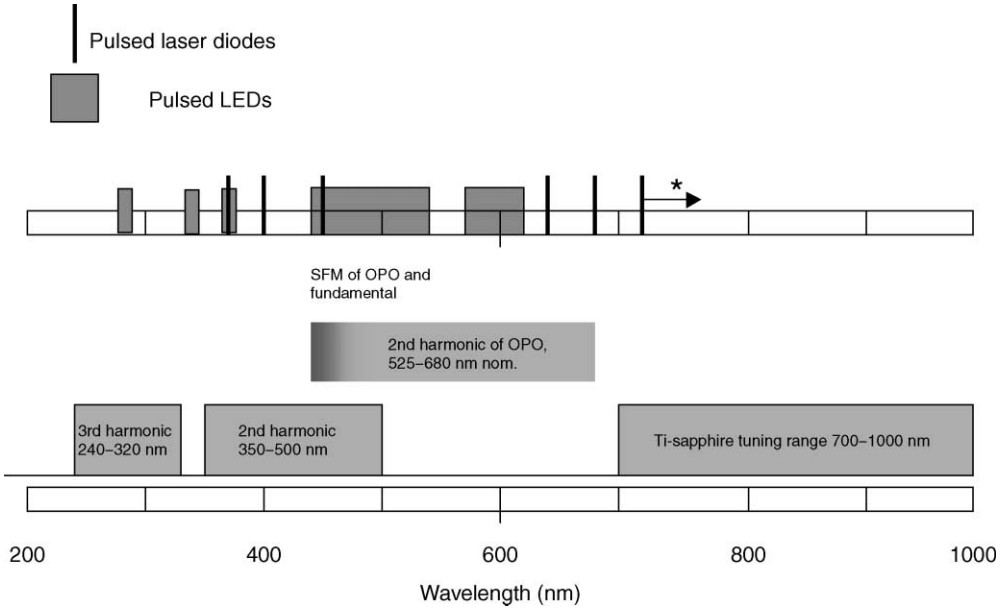


Figure 3 An illustration of the spectral coverage provided by currently available pulsed semiconductor lasers and LEDs, and a modern Ti-sapphire laser and OPO. Note that the OPO's output is frequency doubled, and the wavelengths in the range 525– ~ 440 nm may be obtained by sum frequency mixing of the Ti-sapphire's fundamental output with the output of the OPO. * denotes that there are many long wavelength pulsed laser diodes falling in the range $\lambda > 720$ nm.

beams of longer wavelengths. Synchronous pumping of the OPOs with either a fs or ps Ti-sapphire laser ensures reasonably high efficiencies. The outputs from the OPO may be frequency doubled, with the option of intracavity doubling enhancing the efficiency still further. Using an 800 nm pump wavelength, an OPO provides typical signal and idler tuning ranges of 1050– >1350 nm and 2.1–3.0 μm, respectively. The signal can be doubled to give 525–680 nm, effectively filling the visible spectrum. The average powers obtainable from Ti-sapphire pumped OPOs makes them eminently suitable for TCSPC experiments. A summary of the spectral coverage of the Ti-sapphire laser and OPO, and low-cost solid-state sources are illustrated in Figure 3.

An additional advantage of the Ti-sapphire laser is the possibility of multiphoton excitation. The combination of high average power and very short pulse duration of a mode-locked Ti-sapphire laser means that the peak-powers of the pulses can be very high. Focusing the near-IR output of the laser gives rise to exceedingly high photon densities within a sample which can lead to the simultaneous absorption of two or more photons. This is of particular significance when recording fluorescence lifetimes in conjunction with a microscope, particularly when the microscope is operated in a confocal configuration. Such systems are now being used for a derivative of TCSPC, known as fluorescence lifetime imaging (FLM).

In summary, the Ti-sapphire laser is a very versatile and flexible source that is well suited to the TCSPC experiment. Off-the-shelf systems provide broad spectral coverage and variable repetition rates. Undoubtedly this medium will be the mainstay of TCSPC for some time to come.

Diode Lasers

Low-power diode lasers, with outputs in the range 630–1000 nm, have been commercially available for some time. These devices, which are normally operated as CW sources, can be operated in a pulsed mode, providing a high frequency train of sub-ns pulses. However, the relatively long output wavelengths produced by this first generation of diode lasers severely restricted the range of chromophores that could be excited and this proved to be a limiting factor for their use as TCSPC sources. Attempts to generate shorter wavelengths from these NIR laser diodes by second harmonic generation (SHG), were never realized: the SHG process is extremely inefficient due to the very low peak output power of these lasers. Shorter wavelengths could be obtained by pulsing conventional light-emitting diodes (LEDs), which were commercially available with emission wavelengths covering the visible spectrum down to ~450 nm. However, these had the disadvantage of exhibiting longer pulse durations than those achievable from laser diodes and have

much broader emission bandwidths than a laser source.

The introduction of violet and blue diode lasers and UV-LEDs by the Nichia Corporation (Japan) in the late 1990s resulted in a renewed interest in sources utilizing these small, solid-state devices. At least three manufacturers produce ranges of commercial TCSPC light sources based upon pulsed LEDs and laser diodes and can now boast a number of pulsed laser sources in their ranges including 375, 400, 450, 473, 635, 670, and 830 nm. Currently commercial pulsed laser diodes and LEDs are manufactured by Jobin-Yvon IBH (UK), Picoquant (Germany), and Hamamatso (Japan). These small, relatively low-cost devices offer pulse durations of <100 ps and excitation repetition rates of up to tens of MHz making them ideally suited to 'routine' lifetime measurements. Although these diode laser sources are not tunable and do not span the full spectrum, they are complemented by pulsed LEDs which fill the gap in the visible spectrum. Pulsed LEDs generally have a longer pulse duration and broader, lower intensity spectral output, but are useful for some applications. At the time of writing there are no diode laser sources providing wavelengths of <350 nm, although Jobin-Yvon IBH (UK) have recently announced pulsed LEDs operating at 280 nm and 340 nm. As LEDs, and possibly diode lasers, emitting at these shorter wavelengths become more widely available, the prospects of low running costs and ease of use makes these semiconductor devices an extremely attractive option for many TCSPC experiments. Commercial TCSPC spectrometers based around these systems offer a comparatively low-cost, turnkey approach to fluorescence lifetime based experiments, opening up what was a highly specialized field to a much broader range of users.

See also

Lasers: Noble Gas Ion Lasers. **Light Emitting Diodes**. **Ultrafast Laser Techniques:** Pulse Characterization Techniques.

Further Reading

Lakowicz JR (ed.) (2001) *Principles of Fluorescence Spectroscopy*. Plenum Press.

O'Connor DV and Phillips D (1984) *Time-Correlated Single Photon Counting*.

Spence DE, Kean PN and Sibbett W (1991) *Optics Letters* 16: 42–44.

Transient Holographic Grating Techniques in Chemical Dynamics

E Vauthey, University of Geneva, Geneva, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Over the past two decades, holographic techniques have proved to be valuable tools for investigating the dynamics of chemical processes. The aim of this chapter is to give an overview of the main applications of these techniques for chemical dynamics. The basic principle underlying the formation and the detection of elementary transient holograms, also called transient gratings, is first presented. This is followed by a brief description of a typical experimental setup. The main applications of these techniques to solve chemical problems are then discussed.

Basic Principle

The basic principle of the transient holographic technique is illustrated in Figure 1. The sample material is excited by two laser pulses at the same wavelength and crossed at an angle $2\theta_{pu}$. If the two pump pulses have the same intensity, I_{pu}, the intensity distribution at the interference region, assuming plane waves, is

$$I(x) = 2I_{pu}\left[1 + \cos\left(\frac{2\pi x}{\Lambda}\right)\right] \qquad [1]$$

where $\Lambda = \lambda_{pu}/(2\sin\theta_{pu})$ is the fringe spacing, λ_{pu} is the pump wavelength, and I_{pu} is the intensity of one pump pulse.

As discussed below, there can be many types of light–matter interactions that lead to a change in the optical properties of the material. For a dielectric material, they result in a spatial modulation of the optical susceptibility and thus of the complex refractive index, $\tilde{n}$. The latter distribution can be described as a Fourier cosine series:

$$\tilde{n}(x) = \tilde{n}_0 + \sum_{m=1}^{\infty} \tilde{n}_m \cos\left(\frac{m2\pi x}{\Lambda}\right) \qquad [2]$$

where $\tilde{n}_0$ is the average value of $\tilde{n}$. In the absence of saturation effects, the spatial modulation of $\tilde{n}$ is harmonic and the Fourier coefficients with $m > 1$ vanish. In this case, the peak to null variation of the complex refractive index, $\Delta\tilde{n}$, is equal to the Fourier coefficient $\tilde{n}_1$. The complex refractive index can be split into its real and imaginary components:

$$\tilde{n} = n + iK \qquad [3]$$

where n is the refractive index and K is the attenuation constant.

The hologram created by the interaction of the crossed pump pulses consists in periodic, one-dimensional spatial modulations of n and K. Such distributions are nothing but phase and amplitude gratings, respectively. A third laser beam at the probe wavelength, λ_{pr}, striking these gratings at Bragg angle, $\theta_B = \arcsin(\lambda_{pr}/2\Lambda)$, will thus be partially diffracted (Figure 1b). The diffraction efficiency, η, depends on the modulation amplitude of the optical properties. In the limit of small diffraction efficiency ($\eta < 0.01$), this relationship is given by

$$\eta = \frac{I_{dif}}{I_{pr}} \cong \left[\left(\frac{\ln 10 \Delta A}{4\cos\theta_B}\right)^2 + \left(\frac{\pi d \Delta n}{\lambda_{pr} \cos\theta_B}\right)^2\right] \times \exp\left(-\frac{\ln 10 A}{\cos\theta_B}\right) \qquad [4]$$

where I_{dif} and I_{pr} are the diffracted and the probe intensity respectively, d is the sample thickness, and $A = 4\pi dK/(\lambda \ln 10)$ is the average absorbance.

The first and the second terms in the square bracket describe the contribution of the amplitude and phase gratings respectively, and the exponential term accounts for the reabsorption of the diffracted beam by the sample.

The main processes responsible for the variation of the optical properties of an isotropic dielectric material are summarized in Figure 2.

The modulation of the absorbance, ΔA, is essentially due to the photoinduced concentration change,

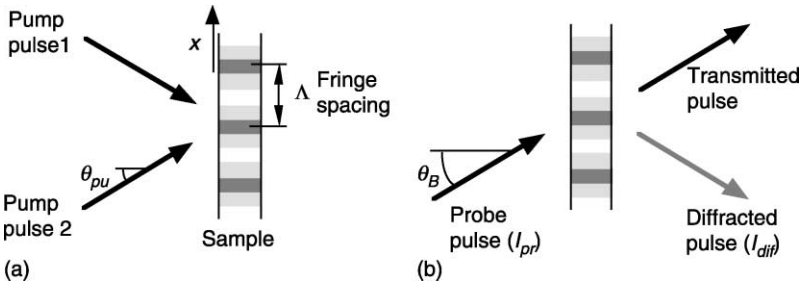


Figure 1 Principle of the transient grating technique: (a) grating formation (pumping), (b) grating detection (probing).

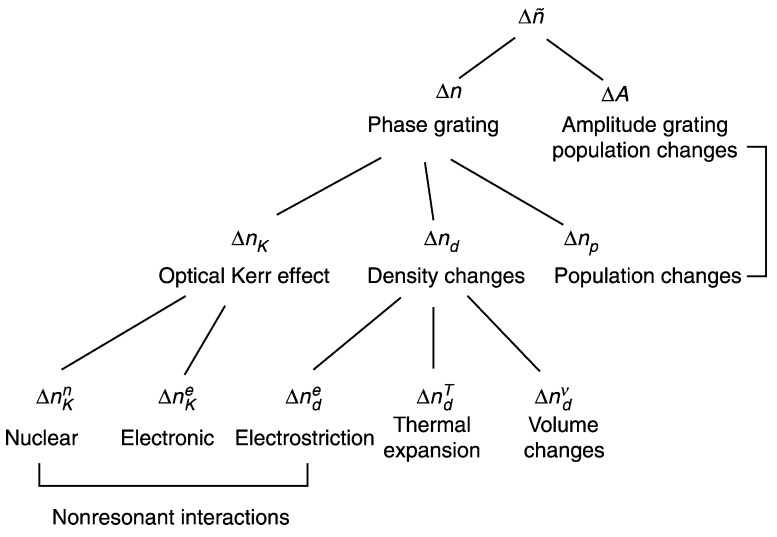


Figure 2 Classification of the possible contributions to a transient grating signal.

ΔC, of the different chemical species i (excited state, photoproduct, …):

$$\Delta A(\lambda_{pr}) = \sum_i \varepsilon_i(\lambda_{pr})\Delta C_i \qquad [5]$$

where ε_i is the absorption coefficient of the species i.

The variation of the refractive index, Δn, has several origins and can be expressed as

$$\Delta n = \Delta n_K + \Delta n_p + \Delta n_d \qquad [6]$$

Δn_K is the variation of refractive index due to the optical Kerr effect (OKE). This nonresonant interaction results in an electronic polarization (electronic OKE) and/or in a nuclear reorientation of the molecules (nuclear OKE) along the direction of the electric field associated with the pump pulses. As a consequence, a transient birefringence is created in the material. This effect is usually discussed within the framework of nonlinear optics in terms of intensity dependent refractive index or third order nonlinear susceptibility. Electronic OKE occurs in any dielectric material under sufficiently high light intensity. On the other hand, nuclear OKE is mostly observed in liquids and gases and depends strongly on the molecular shape.

Δn_p is the change of refractive index related to population changes. Its magnitude and wavelength dependence can be obtained by Kramers–Kronig transformation of $\Delta A(\lambda)$ or $\Delta K(\lambda)$:

$$\Delta n_p(\lambda) = \frac{1}{2\pi^2}\int_0^\infty \frac{\Delta K(\lambda')}{1-(\lambda'/\lambda)^2}\,\mathrm{d}\lambda' \qquad [7]$$

Δn_d is the change of refractive index associated with density changes. Density phase gratings can have essentially three origins:

$$\Delta n_d = \Delta n_d^t + \Delta n_d^v + \Delta n_d^e \qquad [8]$$

Δn_d^t is related to the temperature-induced change of density. If a fraction of the excitation energy is converted into heat, through a nonradiative transition or an exothermic process, the temperature becomes spatially modulated. This results in a variation of density, hence to a modulation of refractive index with amplitude Δn_d^t. Most of the temperature dependence of n originates from the density. The temperature-induced variation of n at constant density is much smaller than Δn_d^t.

Δn_d^v is related to the variation of volume upon population changes. This volume comprises not only the reactant and product molecules but also their environment. For example, in the case of a photodissociation, the volume of the product is larger than that of the reactant and a positive volume change can be expected. This will lead to a decrease of the density and to a negative Δn_d^v.

Finally, Δn_d^e is related to electrostriction in the sample by the electric field of the pump pulses. Like OKE, this is a nonresonant process that also contributes to the intensity dependent refractive index. Electrostriction leads to material compression in the regions of high electric field strength. The periodic compression is accompanied by the generation of two counterpropagating acoustic waves with wave vectors, $\vec{k}_{ac} = \pm(2\pi/\Lambda)\vec{i}$, where $\vec{i}$ is the unit vector along the modulation axis. The interference of these acoustic waves leads to a temporal modulation of Δn_d^e at the acoustic frequency υ_{ac}, with $2\pi\upsilon_{ac} = k_{ac}\upsilon_s$, υ_s being the speed of sound. As Δn_d^e oscillates between negative and positive values, the diffracted intensity, which is proportional to $(\Delta n_d^e)^2$, shows at temporal oscillation at twice the acoustic frequency. In most cases, Δn_d^e is weak and can be neglected if the pump pulses are within an absorption band of the sample.

The modulation amplitudes of absorbance and refractive index are not constant in time and their temporal behavior depends on various dynamic processes in the sample. The whole point of the transient grating techniques is precisely the measurement of the diffracted intensity as a function of time after excitation to deduce dynamic information on the system.

In the following, we will show that the various processes shown in Figure 2, that give rise to a diffracted signal, can in principle be separated by choosing appropriate experimental parameters, such as the timescale, the probe wavelength, the polarization of the four beams, or the crossing angle.

Experimental Setup

A typical experimental arrangement for pump–probe transient grating measurements is shown in Figure 3. The laser output pulses are split in three parts. Two parts of equal intensity are used as

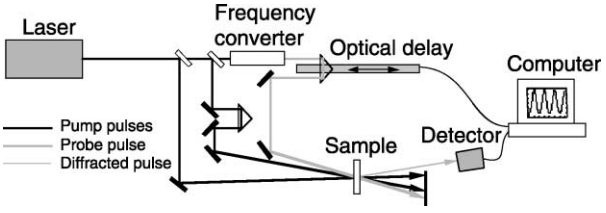


Figure 3 Schematic of a transient grating setup with pump–probe detection.

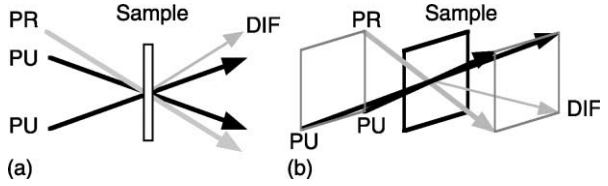


Figure 4 Beam geometry for transient grating: (a) in plane; (b) boxcars.

pump pulses and are crossed in the sample. In order to ensure time coincidence, one pump pulse travels along an adjustable optical delay line. The third part, which is used for probing, can be frequency converted using a dye laser, a nonlinear crystal, a Raman shifter, or white light continuum generation. The probe pulse is sent along a motorized optical delay line before striking the sample at the Bragg angle. There are several possible beam configurations for transient grating and the two most used are illustrated in Figure 4. When the probe and pump pulses are at different wavelengths, they can be in the same plane of incidence as shown in Figure 4a. However, if the pump and probe wavelengths are the same, the folded boxcars geometry shown in Figure 4b has to be used. The transient grating technique is background free and the diffracted signal propagates in a well-defined direction. In a pump–probe experiment, the diffracted signal intensity is measured as a function of the time delay between the pump and probe pulses. Such a setup can be used to probe dynamic processes occurring in timescales going from a few fs to a few ns, the time resolution depending essentially on the duration of the pump and probe pulses. For slower processes, the grating dynamics can be probed in real time with a *cw* laser beam and a fast photodetector.

Applications

The Transient Density Phase Grating Technique

If the grating is probed at a wavelength far from any absorption band, the variation of absorbance, $\Delta A(\lambda_{pr})$, is zero and the corresponding change of refractive index, $\Delta n_p(\lambda_{pr})$, is negligibly small. In this case, eqn [4] simplifies to

$$\eta \cong \left(\frac{\pi d}{\lambda_{pr} \cos \theta_B}\right)^2 \times \Delta n_d^2 \qquad [9]$$

In principle, the diffracted signal may also contain contributions from the optical Kerr effect, Δn_K, but we will assume here that this non-resonant and ultrafast response is negligibly small. The density change can originate from both heat releasing processes and volume differences between the products and reactants, the former contribution usually being much larger than the latter. Even if the heat releasing process is instantaneous, the risetime of the density grating is limited by thermal expansion. This expansion is accompanied by the generation of two counterpropagating acoustic waves with wave vectors, $\vec{k}_{ac} = \pm(2\pi/\Lambda)\vec{i}$. One can distinguish two density gratings: 1) a diffusive density grating, which reproduces the spatial distribution of temperature and which decays by thermal diffusion; and 2) an acoustic density grating originating from the standing acoustic wave and whose amplitude oscillates at the acoustic frequency υ_{ac}.

The amplitudes of these two gratings are equal but of opposite sign. Consequently, the time dependence of the modulation amplitude of the density phase grating is given by

$$\Delta n_d(t) = \left(\frac{\beta Q}{\rho C_v} + \Delta V\right)\rho\left(\frac{\partial n}{\partial \rho}\right)R(t) \qquad [10a]$$

with

$$R(t) = 1 - \cos(2\pi\upsilon_{ac}t) \times \exp(-\alpha_{ac}\upsilon_s t) \qquad [10b]$$

where ρ, β, C_v, and α_{ac} are the density, the volume expansion coefficient, the heat capacity, and the acoustic attenuation constant of the medium, respectively, Q is the amount of heat deposited during the photoinduced process, and ΔV is the corresponding volume change. As the standing acoustic wave oscillates, its corresponding density grating interferes with the diffusive density grating. Therefore, the total modulation amplitude of the density and thus Δn_d exhibits the oscillation at υ_{ac}. Figure 5 shows the time profile of the diffracted

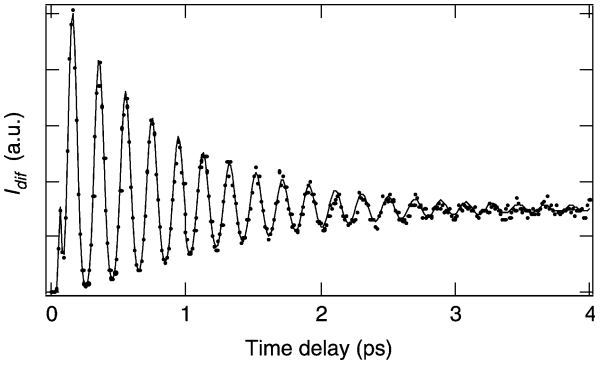


Figure 5 Time profile of the diffracted intensity measured with a solution of malachite green after excitation with two pulses with close to counterpropagating geometry.

intensity measured with a solution of malachite green. After excitation to the S_1 state, this dye relaxes nonradiatively to the ground state in a few ps. For this measurement, the sample solution was excited with two 30 ps laser pulses at 532 nm crossed with an angle close to 180°. The continuous line is the best fit of eqns [9] and [10]. The damping of the oscillation is due to acoustic attenuation. After complete damping, the remaining diffracted signal is due to the diffusive density grating only.

$R(t)$ can be considered as the response function of the sample to a prompt heat release and/or volume change. If these processes are not instantaneous compared to an acoustic period ($\tau_{ac} = \upsilon_{ac}^{-1}$), the acoustic waves are not created impulsively and the time dependence of Δn_d is

$$\Delta n_d(t) = \left(\frac{\beta Q}{\rho C_v} + \Delta V\right)\rho\left(\frac{\partial n}{\partial \rho}\right)F(t) \qquad [11a]$$

with

$$F(t) = \int_{-\infty}^{t} R(t - t') \cdot f(t')\mathrm{d}t' \qquad [11b]$$

where $f(t)$ is a normalized function describing the time evolution of the temperature and/or volume change. In many cases, $f(t) = \exp(-k_r t)$, k_r being the rate constant of the process responsible for the change. Figure 6 shows the time profile of the diffracted intensity calculated with eqns [9] and [11] for different values of k_r.

If several processes take place, the total change of refractive index is the sum of the changes due to the individual processes. In this case, Δn_d should be expressed as

$$\Delta n_d(t) = \sum_i \left(\frac{\beta Q_i}{\rho C_v} + \Delta V_i\right)\rho\left(\frac{\partial n}{\partial \rho}\right)F_i(t) \qquad [12]$$

The separation of the thermal and volume contributions to the diffracted signal is problematic. Several approaches have been proposed, the most used being the measurement in a series of solvents with different expansion coefficient β. This method requires that the other solvent properties like refractive index, dielectric constant, or viscosity, are constant or that the energetics of the system investigated does not depend on them. This separation is easier when working with water because its β value vanishes at 4 °C. At this temperature, the density variations are due to the volume changes only.

The above equations describe the growth of the density phase grating. However, this grating is not permanent and decays through diffusive processes. The phase grating originating from thermal expansion decays via thermal diffusion with a rate constant k_{th} given by

$$k_{th} = D_{th}\left(\frac{2\pi}{\Lambda}\right)^2 \qquad [13]$$

where D_{th} is the thermal diffusivity. Table 1 shows k_{th} values in acetonitrile for different crossing angles.

The decay of the phase grating originating from volume changes depends on the dynamics of the population responsible for ΔV (vide infra).

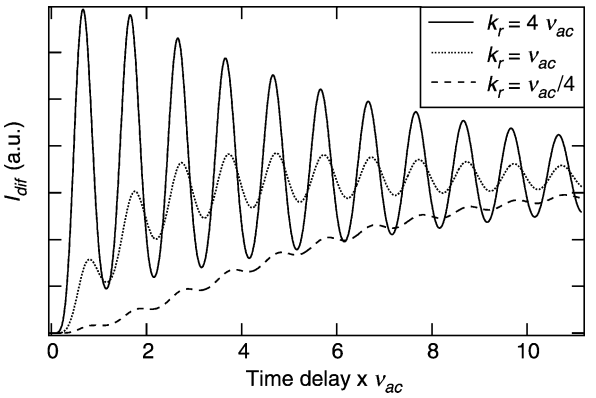


Figure 6 Time profiles of the diffracted intensity calculated using eqns [9] and [11] with various k_r values.

Time resolved optical calorimetry

A major application of the density phase grating in chemistry is the investigation of the energetics of photo-induced processes. The great advantage of this technique over other optical calorimetric methods, like the thermal lens or the photoacoustic spectroscopy, is its superior time resolution. The time constant of the fastest heat releasing process that can be time resolved with this technique is of the order of the acoustic period. The shortest acoustic period is achieved when forming the grating with two counterpropagating pump pulses ($2\theta_{pu} = 180°$). In this case the fringe spacing is

Table 1 Fringe spacing, Λ, acoustic frequency, ν_{ac}, thermal diffusion rate constant, k_{tr}, for various crossing angles of the pump pulses, 2 θ_{pu}, at 355 nm and in acetonitrile

$2\,\theta_{pu}$	Λ (μm)	$\upsilon_{ac}(s^{-1})$	k_{th} (s^{-1})
0.5°	40.7	3.2×10^7	4.7×10^3
50°	0.42	3.1×10^9	4.5×10^7
180°	0.13	9.7×10^9	4.7×10^8

$\Lambda = \lambda_{pu}/2n$ and, with UV pump pulses, an acoustic period of the order of 100 ps can be obtained in a typical organic solvent.

For example, Figure 7a shows the energy diagram for a photoinduced electron transfer (ET) reaction between benzophenone (BP) and an electron donor (D) in a polar solvent. Upon excitation at 355 nm, $^1BP^*$ undergoes intersystem-crossing (ISC) to $^3BP^*$ with a time constant of about 10 ps.

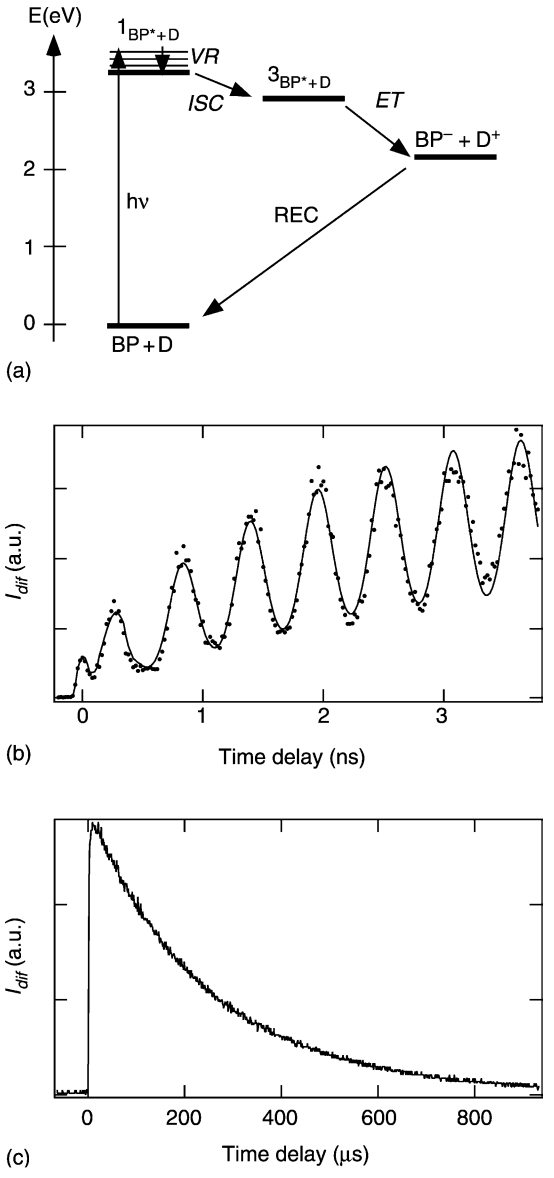


Figure 7 (a) Energy diagram of the states involved in the photoinduced electron transfer reaction between benzophenone (BP) and an electron donor (D) (VR: vibrational relaxation, ISC: intersystem crossing; ET: electron transfer; REC: charge recombination). (b) Time profile of the diffracted intensity measured at 590 nm after excitation at 355 nm of a solution of BP and 0.05 M D. (c) Same as (b) but measured at 1064 nm with a *cw* laser.

After diffusional encounter with the electron donor, ET takes place and a pair of ions is generated. With this system, the whole energy of the 355 nm photon ($E = 3.49$ eV) is converted into heat. The different heat releasing processes can be differentiated according to their timescale. The vibrational relaxation to $^1BP^*$ and the ensuing ISC to $^3BP^*$ induces an ultrafast release of 0.49 eV as heat. With donor concentrations of the order of 0.1 M, the heat deposition process due to the electron transfer is typically in the ns range. Finally, the recombination of the ions produces a heat release in the microsecond timescale. Figure 7b shows the time profile of the diffracted intensity measured after excitation at 355 nm of BP with 0.05 M D in acetonitrile. The 30 ps pump pulses were crossed at 27° ($\tau_{ac} = 565$ ps) and the 10 ps probe pulses were at 590 nm. The oscillatory behavior is due to the ultrafast heat released upon formation of $^3BP^*$, while the slow dc rise is caused by the heat dissipated upon ET. As the amount of energy released in the ultrafast process is known, the energy released upon ET can be determined by comparing the amplitudes of the fast and slow components of the time profile. The energetics as well as the dynamics of the photoinduced ET process can thus be determined. Figure 7c shows the decay of the diffracted intensity due to the washing out of the grating by thermal diffusion. As both the acoustic frequency and the thermal diffusion depends on the grating wavevector, the experimental time window in which a heat releasing process can be measured, depends mainly on the crossing angle of the pump pulses as shown in Table 1.

Determination of material properties

Another important application of the transient density phase grating technique is the investigation of material properties. Acoustics waves of various frequencies, depending on the pump wavelength and crossing angle, can be generated without physical contact with the sample. Therefore, acoustic properties, such as the speed of sound and the acoustic attenuation of the material, can be easily obtained from the frequency and the damping of the oscillation of the transient grating signal.

Similarly, the optoelastic constant of a material, $\rho\partial n/\partial\rho$, can be determined from the amplitude of the signal (see eqn [11]). This is done by comparing the signal amplitude of the material under investigation with that obtained with a known standard.

Finally, the thermal diffusivity, D_{th}, can be easily obtained from the decay of the thermal density

phase grating, such as that shown in Figure 7c. This technique can be used with a large variety of bulk materials as well as films, surfaces and interfaces.

Investigation of Population Dynamics

The dynamic properties of a photogenerated species can be investigated by using a probe wavelength within its absorption or dispersion spectrum. As shown by eqn [4], either ΔA or Δn_p has to be different from zero. For this application, Δn_K and Δn_d should ideally be equal to zero. Unless working with ultrashort pulses (< 1 ps) and a weakly absorbing sample, Δn_K can be neglected. Practically, it is almost impossible to find a sample system for which some fraction of the energy absorbed as light is not released as heat. However, by using a sufficiently small crossing angle of the pump pulses, the formation of the density phase grating, which depends on the acoustic period, can take as much as 30 to 40 ns. In this case, Δn_d is negligible during the first few ns after excitation and the diffracted intensity is due to the population grating only:

$$I_{dif}(t) \propto \sum_i \Delta C_i^2(t) \qquad [14]$$

where ΔC_i is the modulation amplitude of the concentration of every species i whose absorption and/or dispersion spectrum overlaps with the probe wavelength.

The temporal variation of ΔC_i can be due either to the dynamics of the species i in the illuminated grating fringes or to processes taking place between the fringes, such as translational diffusion, excitation transport, and charge diffusion. In liquids, the decay of the population grating by translational diffusion is slow and occurs in the microsecond to millisecond timescale, depending on the fringe spacing. As thermal diffusion is typically hundred times faster, eqn [14] is again valid in this long timescale. Therefore if the population dynamics is very slow, the translational diffusion coefficient of a chemical species can be obtained by measuring the decay of the diffracted intensity as a function of the fringe spacing. This procedure has also been used to determine the temperature of flames. In this case however, the decay of the population grating by translational diffusion occurs typically in the sub-ns timescale.

In the condensed phase, these interfringe processes are of minor importance when measuring the diffracted intensity in the short timescale, i.e., before the formation of the density phase grating. In this case, the transient grating technique is similar to transient absorption, and it thus allows the measurement of population dynamics. However, because holographic detection is background free, it is at least a hundred times more sensitive than transient absorption.

The population gratings are usually probed with monochromatic laser pulses. This procedure is well suited for simple photoinduced processes, such as the decay of an excited state population to the ground state. For example, Figure 8 shows the decay of the diffracted intensity at 532 nm measured after excitation of a cyanine dye at the same wavelength. These dynamics correspond to the ground state recovery of the dye by non-radiative deactivation of the first singlet excited state. Because the diffracted intensity is proportional to the square of the concentration changes (see eqn [14]), its decay time is twice as short as the ground state recovery time. If the reaction is more complex, or if several intermediates are involved, the population grating has to be probed at several wavelengths. Instead of repeating many single wavelength experiments, it is preferable to perform multiplex transient grating. In this case, the grating is probed by white light pulses generated by focusing high intensity fs or ps laser pulses in a dispersive medium. If the crossing angle of the pump pulses is small enough ($< 1°$), the Bragg angle for probing is almost independent on the wavelength. A transient grating spectrum is obtained by dispersing the diffracted signal in a spectrograph. This spectrum consists in the sum of the square of the transient absorption and transient dispersion spectra. Practically, it is very similar to a transient absorption spectrum, but with a much superior signal to noise ratio. Figure 9 shows the transient grating spectra measured at various time delays after excitation of a solution of chloranil (CA) and

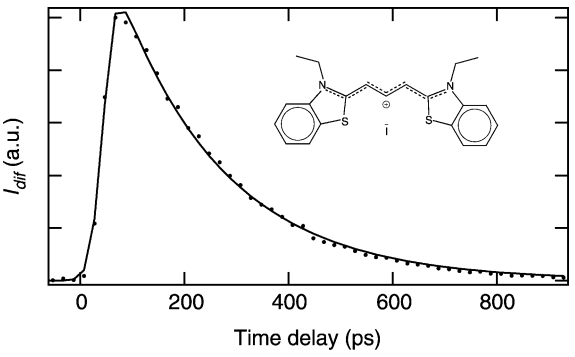


Figure 8 Time profile of the diffracted intensity at 532 nm measured after excitation at the same wavelength of a cyanine dye (inset) in solution and best fit assuming exponential ground state recovery.

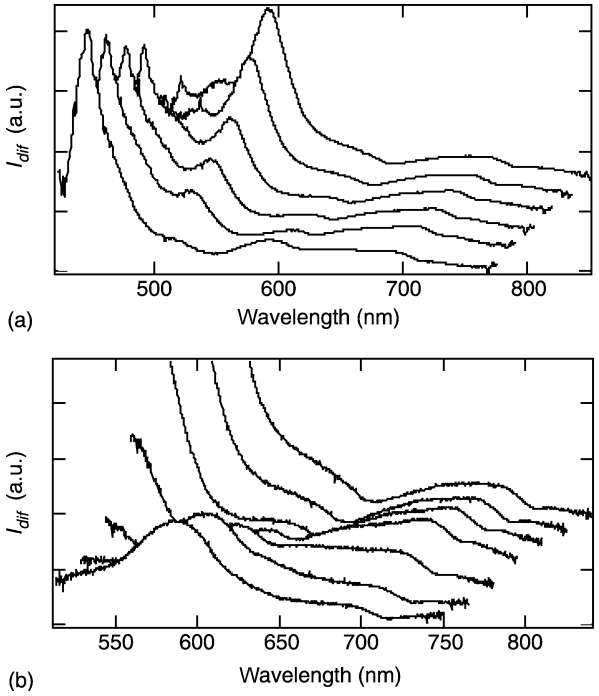


Figure 9 Transient grating spectra obtained at various time delays after excitation at 355 nm of a solution of chloranil and 0.25 M methylnaphthalene (a): from top to bottom: 60, 100, 180, 260, 330 and 600 ps; (b): from top to bottom: 100, 400, 750, 1100, 1500 ps).

methylnaphthalene (MNA) in acetonitrile. The reaction that takes place is

$$^3CA^* \cdot MNA \xrightarrow{\text{diffusion}} (^3CA^* \cdot MNA) \xrightarrow{ET}$$
$$^3(CA^- \cdot MNA^+) \xrightarrow{MNA} {}^3(CA^- \cdot MNA^+)$$

After its formation upon ultrafast ISC, $^3CA^*$, which absorbs around 510 nm, decays upon ET with MNA to generate CA^-(450 nm) and MNA^+ (690 nm). The ion pair reacts further with a second MNA molecule to form the dimer cation (580 nm).

The time profile of the diffracted intensity reflects the population dynamics as long as these populations follow first- or pseudo-first order kinetics. Higher order kinetics leads to inharmonic gratings and in this case the time dependence of the diffracted intensity is no longer a direct measure of the population dynamics.

Polarization Selective Transient Grating

In the above applications, the selection between the different contributions to the diffracted signal was essentially made by choosing the probe wavelength and the crossing angle of the pump pulses. However, this approach is not always sufficient. Another important parameter is the polarization of the four waves involved in a transient grating experiment. For example, when measuring population dynamics, the polarization of the probe beam has to be at magic angle (54.7°) relatively to that of the pump pulses. This ensures that the observed time profiles are not distorted by the decay of the orientational anisotropy of the species created by the polarized pump pulses.

The magnitude of ΔA and Δn depends on the pump pulses intensity, and therefore the diffracted intensity can be expressed by using the formalism of nonlinear optics:

$$I_{dif}(t) = C \int_{-\infty}^{+\infty} \mathrm{d}t\, I_{pr}(t - t'') \times \left[\int_{-\infty}^{t} \mathrm{d}t' |R_{ijkl}(t'' - t')| I_{pu}(t') \right]^2 \quad [15]$$

where C is a constant and R_{ijkl} is an element of the fourth rank tensor $\mathbf{R}$ describing the nonlinear response of the material to the applied optical fields. In isotropic media, this tensor has only 21 nonzero elements, which are related as follows:

$$R_{1111} = R_{1122} + R_{1212} + R_{1221} \quad [16]$$

where the subscripts are the Cartesian coordinates. Going from right to left, they design the direction of polarization of the pump, probe, and signal pulses. The remaining elements can be obtained by permutation of these indices ($R_{1122} = R_{1133} = R_{2211} = \ldots$). In a conventional transient grating experiment, the two pump pulses are at the same frequency and are time coincident and therefore their indices can be interchanged. In this case, $R_{1212} = R_{1221}$ and the number of independent tensor elements is further reduced:

$$R_{1111} = R_{1122} + 2R_{1212} \quad [17]$$

The tensor $\mathbf{R}$ can be decomposed into four tensors according to the origin of the sample response: population and density changes, electronic and nuclear optical Kerr effects:

$$\mathbf{R} = \mathbf{R}(p) + \mathbf{R}(d) + \mathbf{R}(K, e) + \mathbf{R}(K, n) \quad [18]$$

As they describe different phenomena, these tensors do not have the same symmetry properties. Therefore, the various contributions to the diffracted intensity can, in some cases, be measured selectively by choosing the appropriate polarization of the four waves.

Table 2 shows the relative amplitude of the most important elements of these four tensors.

The tensor $\mathbf{R}(p)$ depends on the polarization anisotropy of the sample, r, created by excitation with polarized pump pulses:

$$r(t) = \frac{N_{\parallel}(t) - N_{\perp}(t)}{N_{\parallel}(t) + 2N_{\perp}(t)} = \frac{2}{5}\mathrm{P}_2[\cos(\gamma)]\exp(-k_{re}t) \qquad [19]$$

where $N_{\parallel}$ and $N_{\perp}$ are the number of molecules with the transition dipole oriented parallel and perpendicular to the polarization of the probe pulse, respectively, γ is the angle between the transition dipoles involved in the pump and probe processes, P_2 is the second Legendre polynomial, and k_{re} is the rate constant for the reorientation of the transition dipole, for example, by rotational diffusion of the molecule or by energy hopping.

The table shows that $R_{1212}(d) = 0$, i.e., the contribution of the density grating can be eliminated with the set of polarization $(0°, 90°, 0°, 90°)$, the so-called crossed grating geometry. In this geometry, the two pump pulses have orthogonal polarization, the polarization of the probe pulse is parallel to that of one pump pulse and the polarization component of the signal that is orthogonal to the polarization of the probe pulse is measured. In this geometry, $R_{1212}(p)$ is nonzero as long as there is some polarization anisotropy, ($r \neq 0$). In this case, the diffracted intensity is

$$I_{dif}(t) \propto |R_{1212}(p)|^2 \propto [\Delta C(t) \times r(t)]^2 \qquad [20]$$

The crossed grating technique can thus be used to investigate the reorientational dynamics of molecules, through $r(t)$, especially when the dynamics of r is faster than that of ΔC. For example, Figure 10 shows the time profile of the diffracted intensity measured in the crossed grating geometry with rhodamine 6G in ethanol. The decay is due to the reorientation of the molecule by rotational diffusion, the excited lifetime of rhodamine being about 4 ns.

On the other hand, if the decay of r is slower than that of ΔC, the crossed grating geometry can be used to measure the population dynamics without any interference from the density phase grating. For example, Figure 11 shows the time profile of the diffracted intensity after excitation of a suspension of TiO_2 particles in water. The upper profile was measured with the set of polarization $(0°, 0°, 0°, 0°)$ and thus reflects the time dependence of $R_{1111}(p)$ and $R_{1111}(d)$. $R_{1111}(p)$ is due to the trapped electron population, which decays by charge recombination, and $R_{1111}(d)$ is due to the heat dissipated upon both charge separation and recombination. The lower time profile was measured in the crossed grating geometry and thus reflects the time dependence of R_{1212} and in particular that of $R_{1212}(p)$.

Each contribution to the signal can be selectively eliminated by using the set of geometry $(\zeta, 45°, 0°, 0°)$ where the value of the 'magic angle' ζ for each

Table 2 Relative value of the most important elements of the response tensors $\mathbf{R}(i)$ and polarization angle ζ of the signal beam, where the contribution of the corresponding process vanishes for the set of polarization $(\zeta, 45°, 0°, 0°)$.

Process	R_{1111}	R_{1122}	R_{1212}	ζ
Electronic OKE	1	1/3	1/3	−71.6°
Nuclear OKE	1	−1/2	3/4	63.4°
Density	1	1	0	−45°
Population:				
$\gamma = 0°$	1	1/3	1/3	−71.6°
$\gamma = 90°$	1	2	−1/2	−26.6°
No correlation: $r = 0$	1	1	0	−45°

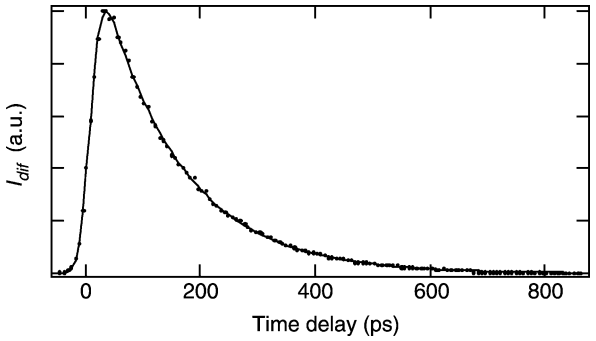


Figure 10 Time profile of the diffracted intensity measured with a solution of rhodamine 6G with crossed grating geometry.

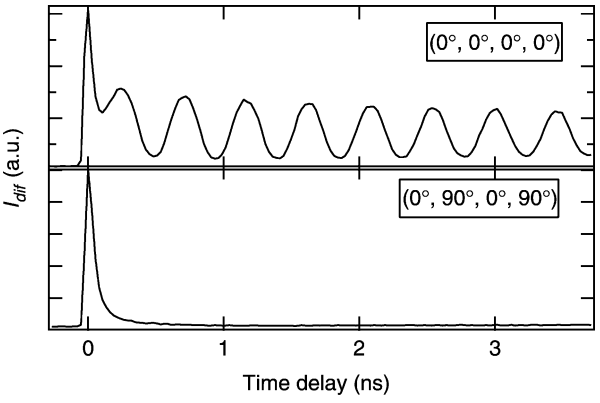


Figure 11 Time profiles of the diffracted intensity after excitation at 355 nm of a suspension of TiO_2 particles in water and using different set of polarization of the four beams.

contribution is listed in Table 2. This approach allows for example the nuclear and electronic contributions to the optical Kerr effect to be measured separately.

Concluding Remarks

The transient grating techniques offer a large variety of applications for investigating the dynamics of chemical processes. We have only discussed the cases where the pump pulses are time coincident and at the same wavelength. Excitation with pump pulses at different wavelengths results to a moving grating. The well-known CARS spectroscopy is such a moving grating technique. Finally, the three pulse photon echo can be considered as a special case of transient grating where the sample is excited by two pump pulses, which are at the same wavelength but are not time coincident.

List of Units and Nomenclature

C	concentration [mol L^{-1}]
C_v	heat capacity [J K^{-1} kg^{-1}]
d	sample thickness [m]
D_{th}	thermal diffusivity [m^2 s^{-1}]
I_{dif}	diffracted intensity [W m^{-2}]
I_{pr}	intensity of the probe pulse [W m^{-2}]
I_{pu}	intensity of a pump pulse [W m^{-2}]
k_{ac}	acoustic wavevector [m^{-1}]
k_r	rate constant of a heat releasing process [s^{-1}]
k_{re}	rate constant of reorientation [s^{-1}]
k_{th}	rate constant of thermal diffusion [s^{-1}]
K	attenuation constant
n	refractive index
$\tilde{n}$	complex refractive index
N	number of molecule per unit volume [m^{-3}]
r	polarization anisotropy
$\mathbf{R}$	fourth rank response tensor [m^2 V^{-2}]
v_s	speed of sound [m s^{-1}]
V	volume [m^3]
α_{ac}	acoustic attenuation constant [m^{-1}]
β	cubic volume expansion coefficient [K^{-1}]
γ	angle between transition dipoles
Λ	fringe spacing [m]
Δn_d	variation of refractive index due to density changes
Δn_d^e	Δn_d due to electrostriction
Δn_K	variation of refractive index due to optical Kerr effect
Δn_d^p	Δn_d due to volume changes
Δn_p	variation of refractive index due to population changes
Δn_d^t	Δn_d due to temperature changes
Δx	peak to null variation of the parameter x
ε	molar decadic absorption coefficient [cm L mol^{-1}]
ζ	angle of polarization of the diffracted signal
η	diffraction efficiency
θ_B	Bragg angle (angle of incidence of the probe pulse)
θ_{pu}	angle of incidence of a pump pulse
λ_{pr}	probe wavelength [m]
λ_{pu}	pump wavelength [m]
ρ	density [kg m^{-3}]
τ_{ac}	acoustic period [s]
υ_{ac}	acoustic frequency [s^{-1}]

See also

Chemical Applications of Lasers: Nonlinear Spectroscopies; Pump and Probe Studies of Femtosecond Kinetics. **Holography, Applications:** Holographic Recording Materials and their Processing. **Holography, Techniques:** Holographic Interferometry. **Materials Characterization Techniques:** $\chi^{(3)}$. **Nonlinear Optics Basics:** Four-Wave Mixing. **Ultrafast Technology:** Femtosecond Chemical Dynamics, Gas Phase; Femtosecond Condensed Phase Spectroscopy – Structural Dynamics.

Further Reading

Bräuchle C and Burland DM (1983) Holographic methods for the investigation of photochemical and photophysical properties of molecules. *Angewandte Chemie International Edition Engl* 22: 582–598.

Eichler HJ, Günter P and Pohl DW (1986) *Laser-Induced Dynamics Gratings*. Berlin: Springer.

Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. Oxford: Oxford University Press.

Fourkas JT and Fayer MD (1992) The transient grating: a holographic window to dynamic processes. *Accounts of Chemical Research* 25: 227–233.

Hall G and Whitaker BJ (1994) Laser-induced grating spectroscopy. *Journal of the Chemistry Society Faraday Trans* 90: 1–16.

Levenson MD and Kano SS (1987) *Introduction to Nonlinear Laser Spectroscopy*, revised edn. Boston: Academic Press.

Mukamel S (1995) *Nonlinear Optical Spectroscopy*. Oxford: Oxford University Press.

Rullière C (ed.) (1998) *Femtosecond Laser Pulses*. Berlin: Springer.

Terazima M (1998) Photothermal studies of photophysical and photochemical processes by the transient grating method. *Advances in Photochemistry* 24: 255–338.

cross-sectional area 10^4 times and dispersed the beam across it. The CPA technique makes it possible to use conventional laser amplifiers and to stay below the onset of nonlinear effects.

See also

Diffraction: Diffraction Gratings. **Ultrafast Laser Techniques:** Pulse Characterization Techniques.

COHERENCE

Contents

Overview
Coherence and Imaging
Speckle and Coherence

Overview

A Sharma and A K Ghatak, Indian Institute of Technology, New Delhi, India

H C Kandpal, National Physical Laboratory, New Delhi, India

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Coherence refers to the characteristic of a wave that indicates whether different parts of the wave oscillate in tandem, or in a definite phase relationship. In other words, it refers to the degree of confidence by which one can predict the amplitude and phase of a wave at a point, from the knowledge of these at another point at the same or a different instant of time. Emission from a thermal source, such as a light bulb, is a highly disordered process and the emitted light is incoherent. A well-stabilized laser source, on the other hand, generates light in a highly ordered manner and the emitted light is highly coherent. Incoherent and coherent light represent two extreme cases. While describing the phenomena of physical optics and diffraction theory, light is assumed to be perfectly coherent in both spatial as well as temporal senses, whereas in radiometry it is generally assumed to be incoherent. However, practical light sources and the fields generated by them are in between the two extremes and are termed as partially coherent sources and fields. The degree of order that exists in an optical field produced by a source of any kind may be described in terms of various correlation functions. These correlation functions are the basic theoretical tools for the analyses of statistical properties of partially coherent light fields.

Light fields generated from real physical sources fluctuate randomly to some extent. On a microscopic level quantum mechanical fluctuations produce randomness and on macroscopic level the randomness occurs as a consequence of these microscopic fluctuations, even in free space. In real physical sources, spontaneous emission causes random fluctuations and even in the case of lasers, spontaneous emission cannot be suppressed completely. In addition to spontaneous emission, there are many other processes that give rise to random fluctuations of light fields. Optical coherence theory was developed to describe the random nature of light and it deals with the statistical similarity between light fluctuations at two (or more) space–time points.

As mentioned earlier, in developing the theory of interference or diffraction, light is assumed to be perfectly coherent, or, in other words, it is taken to be monochromatic and sinusoidal for all times. This is, however, an idealization since the wave is obviously generated at some point of time by an atomic or molecular transition. Furthermore, a wavetrain generated by such a transition is of a finite duration, which is related to the finite lifetime of the atomic or molecular levels involved in the transition. Thus, any wave emanating from a source is an ensemble of a large number of such wavetrains of finite duration, say τ_c. A simplified visualization of such an ensemble is shown in Figure 1 where a wave is shown as a series of wavetrains of duration τ_c. It is evident from the figure that the fields at time t and $t + \Delta t$ will have a definite phase relationship if $\Delta t \ll \tau_c$ and will have no or negligible phase relationship when $\Delta t \gg \tau_c$. The time τ_c is known as the coherence time of the

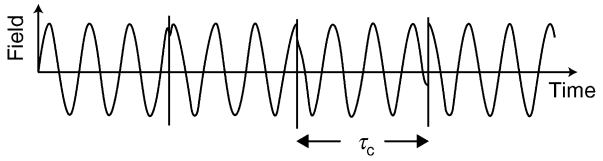


Figure 1 Typical variation of the radiation field with time. The coherence time $\sim \tau_c$.

radiation and the field is said to remain coherent for time $\sim \tau_c$. This property of waves is referred to as the time coherence or the temporal coherence and is related to the spectral purity of the radiation. If one obtains the spectrum of the wave shown in Figure 1 by taking the Fourier transform of the time variation, it would have a width of $\Delta\nu$ around ν_0 which is the frequency of the sinusoidal variation of individual wavetrains. The spectral width $\Delta\nu$ is related to the coherence time as

$$\Delta\nu \sim 1/\tau_c \qquad [1]$$

For thermal sources such as a sodium lamp, $\tau_c \sim 100$ ps, whereas for a laser beam it could be as large as a few milliseconds. A related quantity is the coherence length l_c, which is the distance covered by the wave in time τ_c,

$$l_c = c\tau_c \sim \frac{c}{\Delta\nu} = \frac{\lambda_0^2}{\Delta\lambda} \qquad [2]$$

where λ_0 is the central wavelength ($\lambda_0 = c/\nu_0$) and $\Delta\lambda$ is the wavelength spread corresponding to the spectral width $\Delta\nu$. In a two-beam interference experiment (e.g., Michelson interferometer, Figure 2), the interfering beam derived from the same source will produce good interference fringes if the path difference between the two interfering beams is less than the coherence length of the radiation given out by the source.

It must be added here that for the real fields, generated by innumerable atoms or molecules, the individual wavetrains have different lengths around an average value τ_c. Furthermore, several wavetrains in general are propagating simultaneously, overlapping in space and time, to produce an ensemble whose properties are best understood in a statistical sense as we shall see in later sections.

Another type of coherence associated with the fields is the space coherence or the spatial coherence, which is related to the size of the source of radiation. It is evident that when the source is an ideal point source, the field at any two points (within the coherence length) would have definite phase relationship. However, the field from a thermal source of finite area S can be thought as resultant of the fields

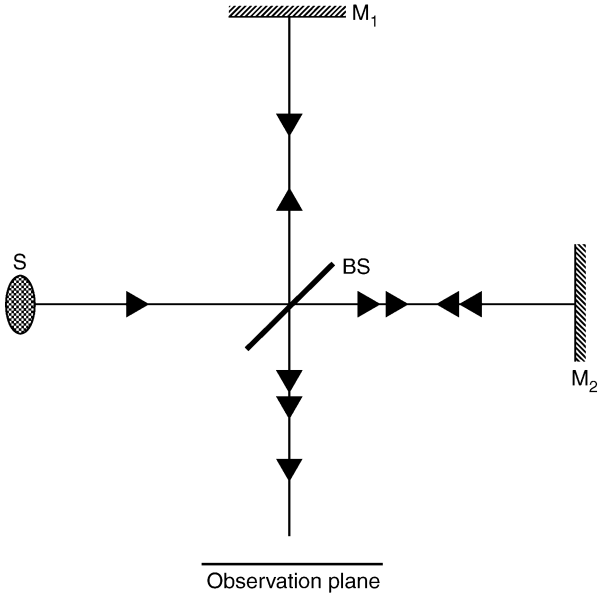


Figure 2 Michelson interferometer to study the temporal coherence of radiation from source S; M_1 and M_2 are mirrors and BS is a 50%–50% beamsplitter.

from each point on the source. Since each point source would usually be independent of the other, the phase relationship between the fields at any two points would depend on their position and size of the source. It can be shown that two points will have a strong phase relationship if they lie within the solid angle $\Delta\Omega$ from the source (Figure 3) such that

$$\Delta\Omega \sim \frac{\lambda_0^2}{S} \qquad [3]$$

Thus, on a plane R distance away from the source, one can define an area $A_c = R^2\Delta\Omega$ over which the field remains spatially coherent. This area A_c is called the coherence area of the radiation and its square root is sometimes referred to as the transverse coherence length. It is trivial to show that

$$A_c \sim \frac{\lambda_0^2}{\Delta\Omega_S} \qquad [4]$$

where $\Delta\Omega_S$ is the solid angle subtended by the source on the plane at which the coherence area is defined. In Young's double-hole experiment, if the two pinholes lie within the coherence area of the radiation from the primary source, good contrast in fringes would be observed.

Combining the concepts of coherence length and the coherence area, one can define a coherence volume as $V_c = A_c l_c$. For the wavefield from a thermal source, this volume represents that portion of the space in which the field is coherent and any

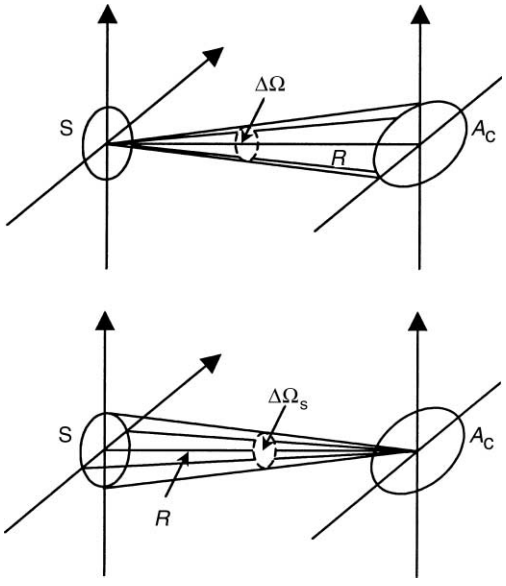


Figure 3 Spatial coherence and coherence area.

interference produced using the radiation from points within this volume will produce fringes of good contrast.

Mathematical Description of Coherence

Analytical Field Representation

Coherence properties associated with fields are best analyzed in terms of complex representation for optical fields. Let the real function $V^{(r)}(\mathbf{r},t)$ represent the scalar field, which could be one of the transverse Cartesian components of the electric field associated with the electromagnetic wave. One can then define a complex analytical signal $V(\mathbf{r},t)$ such that

$$V^{(r)}(\mathbf{r},t) = \mathrm{Re}[V(\mathbf{r},t)],$$
$$V(\mathbf{r},t) = \int_0^\infty \tilde{V}^{(r)}(\mathbf{r},\nu)\mathrm{e}^{-2\pi i\nu t}\,\mathrm{d}\nu \qquad [5]$$

where the spectrum $\tilde{V}^{(r)}(\mathbf{r},\nu)$ is the Fourier transform of the scalar field $V^{(r)}(\mathbf{r},t)$ and the spectrum for negative frequencies has been suppressed as it does not give any new information since $\tilde{V}^{(r)^*}(\mathbf{r},\nu) = \tilde{V}^{(r)}(\mathbf{r},-\nu)$.

In general, the radiation from a quasi-monochromatic thermal source fluctuates randomly as it is made of a large number of mutually independent contributions from individual atoms or molecules in the source. The field from such a source can be regarded as an ensemble of a large number of randomly different analytical signals such as $V(\mathbf{r},t)$. In other words, $V(\mathbf{r},t)$ is a typical member of an ensemble, which is the result of a random process representing the radiation from a quasi-monochromatic source. This process is assumed to be stationary and ergodic so that the ensemble average is equal to the time average of a typical member of the ensemble and that the origin of time is unimportant. Thus, the quantities of interest in the theory of coherence are defined as time averages:

$$\langle f(t)\rangle = \underset{T\to\infty}{\mathrm{Lim}}\,\frac{1}{2T}\int_{-T}^{T} f(t)\mathrm{d}t \qquad [6]$$

Mutual Coherence

In order to define the mutual coherence, the key concept in the theory of coherence, we consider Young's interference experiment, as shown in Figure 4, where the radiation from a broad source of size S is illuminating a screen with two pinholes P_1 and P_2. The light emerging from the two pinholes produces interference, which is observed on another screen at a distance R from the first screen. The field at point P, due to the pinholes, would be $K_1V(\mathbf{r}_1,t-t_1)$ and $K_2V(\mathbf{r}_2,t-t_2)$ respectively, where $\mathbf{r}_1$ and $\mathbf{r}_2$ define the positions of P_1 and P_2, t_1 and t_2 are the times taken by the light to travel to P from P_1 and P_2, and K_1 and K_2 are imaginary constants that depend on the geometry and size of the respective pinhole and its distance from the point P. Thus, the resultant field at point P would be given by

$$V(\mathbf{r},t) = K_1V(\mathbf{r}_1,t-t_1) + K_2V(\mathbf{r}_2,t-t_2) \qquad [7]$$

Since the optical periods are extremely small as compared to the response time of a detector, the detector placed at point P will record only the time-averaged intensity:

$$I(P) = \langle V^*(\mathbf{r},t)V(\mathbf{r},t)\rangle \qquad [8]$$

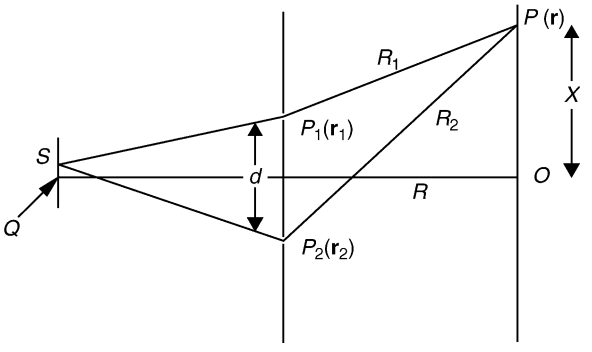


Figure 4 Schematic of Young's double-hole experiment.

where some constants have been ignored. With eqn [7], the intensity at point P would be given by

$$I(P) = I_1 + I_2 + 2\,\mathrm{Re}\{K_1 K_2^* \Gamma(\mathbf{r}_1, \mathbf{r}_2, t - t_1, t - t_2)\} \quad [9]$$

where I_1 and I_2 are the intensities at point P, respectively, due to radiations from pinholes P_1 and P_2 independently (defined as $I_i = \langle |K_i V(\mathbf{r}_i, t)|^2 \rangle$) and

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, t - t_1, t - t_2) \equiv \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t + \tau) \rangle \quad [10]$$

is the mutual coherence function of the fields at P_1 and P_2, and it depends on the time difference $\tau = t_2 - t_1$, since the random process is assumed to be stationary. This function is also sometimes denoted by $\Gamma_{12}(\tau)$. The function $\Gamma_{ii}(\tau) \equiv \Gamma(\mathbf{r}_i, \mathbf{r}_i, \tau)$ defines the self-coherence of the light from the pinhole at P_i, and $|K_i|^2 \Gamma_{ii}(0)$ defines the intensity I_i at point P, due to the light from pinhole P_i.

Degree of Coherence and the Visibility of Interference Fringes

One can define a normalized form of the mutual coherence function, namely:

$$\gamma_{12}(\tau) \equiv \gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \frac{\Gamma_{12}(\tau)}{\sqrt{\Gamma_{11}(0)}\sqrt{\Gamma_{22}(0)}} = \frac{\langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t + \tau) \rangle}{\left[\langle |V(\mathbf{r}_1, t)|^2 \rangle \langle |V(\mathbf{r}_2, t)|^2 \rangle \right]^{1/2}} \quad [11]$$

which is called the complex degree of coherence. It can be shown that $0 \leq |\gamma_{12}(\tau)| \leq 1$. The intensity at point P, given by eqn [9], can now be written as

$$I(P) = I_1 + I_2 + 2\sqrt{I_1 I_2}\,\mathrm{Re}[\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)] \quad [12]$$

Expressing $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ as

$$\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = |\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| \exp[i\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau) - 2\pi i \nu_0 \tau] \quad [13]$$

where $\alpha(\vec{r}_1, \vec{r}_2, \tau) = \arg[\gamma(\vec{r}_1, \vec{r}_2, \tau)] + 2\pi\nu_0\tau$, and ν_0 is the mean frequency of the light, the intensity in eqn [12] can be written as

$$I(P) = I_1 + I_2 + 2\sqrt{I_1 I_2}|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| \times \cos[\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau) - 2\pi\nu_0\tau] \quad [14]$$

Now, if we assume that the source is quasi-monochromatic, i.e., its spectral width $\Delta\nu \ll \nu_0$, the quantities $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ and $\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau)$ vary slowly on the observation screen, and the interference fringes are mainly obtained due to the cosine term. Thus, defining the visibility of fringes as $v = (I_{\max} - I_{\min})/(I_{\max} + I_{\min})$, we obtain

$$v = \frac{2(I_1 I_2)^{1/2}}{I_1 + I_2} |\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| \quad [15]$$

which shows that for maximum visibility, the two interfering fields must be completely coherent ($|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| = 1$). On the other hand, if the fields are completely incoherent ($|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| = 0$), no fringes are observed ($I_{\max} = I_{\min}$). The fields are said to be partially coherent when $0 < |\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| < 1$. When $I_1 = I_2$, the visibility is the same as $|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)|$. The relation in eqn [15] shows that in an interference experiment, one can obtain the modulus of the complex degree of coherence by measuring I_1, I_2, and the visibility. Similarly, eqn [14] shows that from the measurement the positions of maxima, one can obtain the phase of the complex degree of coherence, $\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau)$.

Temporal and Spatial Coherence

If the source illuminating the pinholes is a point source of finite spectral width situated at point Q, the fields at point P_1 and P_2 (Figure 4) at any given instant are the same and the mutual coherence function becomes

$$\Gamma_{11}(\tau) = \Gamma(\mathbf{r}_1, \mathbf{r}_1, \tau) = \langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_1, t + \tau) \rangle = \langle V^*(\mathbf{r}_2, t) V(\mathbf{r}_2, t + \tau) \rangle = \Gamma_{22}(\tau) \quad [16]$$

The self-coherence, $\Gamma_{11}(\tau)$, of the light from pinhole P_1, is a direct measure of the temporal coherence of the source. On the other hand, if the source is of finite size and we observe the interference at point O which corresponds to $\tau = 0$, the mutual coherence function would be

$$\Gamma_{12}(0) = \Gamma(\mathbf{r}_1, \mathbf{r}_2, 0) = \langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t) \rangle \equiv J_{12} \quad [17]$$

which is called the mutual intensity and is a direct measure of the spatial coherence of the source. In general, however, the function $\Gamma_{12}(\tau)$ measures, for a source of finite size and spectral width, a combination of the temporal and spatial coherence, and in only some limiting cases, the two types of coherence can be separated.

Spectral Representation of Mutual Coherence

One can also analyze the correlation between two fields in the spectral domain. In particular, one can define the cross-spectral density function $W(\mathbf{r}_1, \mathbf{r}_2, \nu)$ which defines the correlation between the amplitudes of the spectral components of frequency ν of the light

at the points P_1 and P_2. Thus:

$$W(\mathbf{r}_1, \mathbf{r}_2, \nu)\delta(\nu - \nu') = \left\langle V^*(\mathbf{r}_1, \nu)V(\mathbf{r}_2, \nu')\right\rangle \qquad [18]$$

Using the generalized Wiener–Khintchine theorem, the cross-spectral density function can be shown to be the Fourier transform of the mutual coherence function:

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \int_0^\infty W(\mathbf{r}_1, \mathbf{r}_2, \nu)\mathrm{e}^{-2\pi\mathrm{i}\nu t}\,\mathrm{d}\nu \qquad [19]$$

$$W(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_{-\infty}^\infty \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)\mathrm{e}^{2\pi\mathrm{i}\nu t}\,\mathrm{d}\tau \qquad [20]$$

If the two points P_1 and P_2 coincide (i.e., there is only one pinhole), the cross-spectral density function reduces to the spectral density function of the light, which we denote by $S(\mathbf{r}, \nu)[\equiv W(\mathbf{r}, \mathbf{r}, \nu)]$. Thus, it follows from eqn [20] that the spectral density of light is the inverse Fourier transform of the self-coherence function. This leads to the Fourier transform spectroscopy, as we shall see later. One can also define spectral degree of coherence at frequency ν as

$$\mu(\mathbf{r}_1, \mathbf{r}_2, \nu) = \frac{W(\mathbf{r}_1, \mathbf{r}_2, \nu)}{\sqrt{W(\mathbf{r}_1, \mathbf{r}_1, \nu)W(\mathbf{r}_2, \mathbf{r}_2, \nu)}} = \frac{W(\mathbf{r}_1, \mathbf{r}_2, \nu)}{\sqrt{S(\mathbf{r}_1, \nu)S(\mathbf{r}_2, \nu)}} \qquad [21]$$

It is easy to see that $0 \le |\mu(\mathbf{r}_1, \mathbf{r}_2, \nu)| \le 1$. It is also sometimes referred to as the complex degree of coherence at frequency ν. It may be noted here that in the literature the notation $G^{(1,1)}(\nu) \equiv G(\mathbf{r}_1, \mathbf{r}_2, \gamma)$ has also been used for $W(\mathbf{r}_1, \mathbf{r}_2, \nu)$.

Propagation of Coherence

The van Cittert–Zernike Theorem

Perfectly coherent waves propagate through diffraction formulae, which have been discussed in this volume elsewhere. However, incoherent and partially coherent waves would evidently propagate somewhat differently. The earliest treatment of propagation noncoherent light is due to van Cittert, which was later generalized by Zernike to obtain what is now the van Cittert–Zernike theorem. The theorem deals with the correlations developed between fields at two points, which have been generated by a quasi-monochromatic and spatially incoherent planar source. Thus, as shown in Figure 5, we consider a quasi-monochromatic ($\Delta\nu \ll \nu_0$) planar source σ, which has an intensity distribution $I(\mathbf{r}')$ on its plane and is spatially incoherent, i.e., there is no correlation between the fields at any two points on the source.

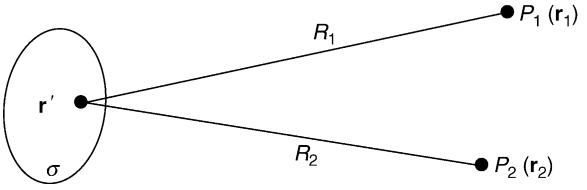


Figure 5 Geometry for the van Cittert–Zernike theorem.

The field, due to this source, would develop finite correlations after propagation, and the theorem states that

$$\gamma(\mathbf{r}_1, \mathbf{r}_2, 0) = \frac{\Gamma_{12}(0)}{\sqrt{\Gamma_{11}(0)}\sqrt{\Gamma_{12}(0)}} = \frac{1}{\sqrt{I(\mathbf{r}_1)I(\mathbf{r}_2)}}\int\int_\sigma I(\mathbf{r}')\frac{\mathrm{e}^{2\pi\mathrm{i}\nu_0(R_2-R_1)/c}}{R_1R_2}\mathrm{d}^2r' \qquad [22]$$

where $R_i = |\mathbf{r}_i - \mathbf{r}'|$ and $I(\mathbf{r}_i)$ is the intensity at point P_i. This relation is similar to the diffraction pattern produced at point P_1 due to a wave, with a spherical wavefront converging towards P_2 and with an amplitude distribution $I(\mathbf{r}')$ when it is diffracted by an aperture of the same shape, size, and the intensity distribution as those of the incoherent source σ. Thus, the theorem shows that a radiation, which was incoherent at the source, becomes partially coherent as it propagates.

Generalized Propagation

The expression $\mathrm{e}^{-2\pi\mathrm{i}\nu_0R_1/c}/R_1$ can be interpreted as the field obtained at P_1 due to a point source located at the point $\mathbf{r}'$ on the planar source. Thus, this expression is simply the point spread function of the homogeneous space between the source and the observation plane containing the points P_1 and P_2. Hopkins generalized this to include any linear optical system characterized by a point spread function $h(\mathbf{r}_i - \mathbf{r}')$ and obtained the formula for the complex degree of coherence of a radiation emerging from an incoherent, quasi-monochromatic planar source after it has propagated through such a linear optical system:

$$\gamma(\mathbf{r}_1,\mathbf{r}_2,0)=\frac{1}{\sqrt{I(\mathbf{r}_1)I(\mathbf{r}_2)}}\int\int_\sigma I(\mathbf{r}')h(\mathbf{r}_1-\mathbf{r}')h^*(\mathbf{r}_2-\mathbf{r}')\mathrm{d}^2r' \qquad [23]$$

It would thus seem that the correlations propagate in much the same way, as does the field. Indeed, Wolf showed that the mutual correlation function

$\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)$ satisfies the wave equations:

$$\nabla_1^2\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\frac{1}{c^2}\frac{\partial^2}{\partial\tau^2}\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)$$
$$\nabla_2^2\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\frac{1}{c^2}\frac{\partial^2}{\partial\tau^2}\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau) \qquad [24]$$

where ∇_i^2 is the Laplacian with respect to the point $\mathbf{r}_i$. Here the first of eqns [24], for instance, gives the variation of the mutual coherence function with respect to $\mathbf{r}_1$ and τ for fixed $\mathbf{r}_2$. Further, the variable τ is the time difference defined through the path difference, since $\tau=(R_2-R_1)/c$, and the actual time does not affect the mutual coherence function (as the fields are assumed to be stationary).

Using the relation in eqn [20], one can also obtain from eqn [24], the propagation equations for the cross-spectral density function $W(\mathbf{r}_1,\mathbf{r}_2,\nu)$:

$$\nabla_1^2 W(\mathbf{r}_1,\mathbf{r}_2,\nu)+k^2W(\mathbf{r}_1,\mathbf{r}_2,\nu)=0$$
$$\nabla_2^2 W(\mathbf{r}_1,\mathbf{r}_2,\nu)+k^2W(\mathbf{r}_1,\mathbf{r}_2,\nu)=0 \qquad [25]$$

where $k=2\pi\nu/c$.

Thompson and Wolf Experiment

One of the most elegant experiments for studying various aspects of coherence theory was carried out by Thompson and Wolf by making slight modifications in the Young's double-hole experiment. The experimental set-up shown in Figure 6 consists of a quasi-monochromatic broad incoherent source S of diameter $2a$. This was obtained by focusing filtered narrow band light from a mercury lamp (not shown in the figure) on to a hole of size $2a$ in an opaque screen A. A mask consisting of two pinholes, each of diameter $2b$ with their axes separated by a distance d, was placed symmetrically about the optical axis of the experimental setup at plane B between two lenses L_1 and L_2, each having focal length f. The source was at the front focal plane of the lens L_1 and the observations were made at the plane C at the back focus of the lens L_2. The separation d was varied to study the spatial coherence function on the mask plane by measuring the visibility and the phase of the fringes formed in plane C.

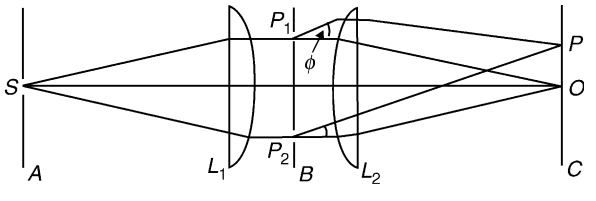


Figure 6 Schematic of the Thompson and Wolf experiment.

Using the van Cittert–Zernike theorem, the complex degree of coherence at the pinholes P_1 and P_2 on the mask after the lens L_1 is

$$\gamma_{12}=|\gamma_{12}|e^{i\beta_{12}}=\frac{2J_1(v)}{v} \text{ with } v=\left(\frac{2\pi vad}{cf}\right) \qquad [26]$$

for symmetrically placed pinholes about the optical axis. Here β_{12} is the phase of the complex degree of coherence and in this special case where the two holes are equidistant from the axis, β_{12} is either zero or π, respectively, for positive or negative values of $2J_1(v)/v$. Let us assume that the intensities at two pinholes P_1 and P_2 are equal. The interference pattern observed at the back focal plane of lens L_2 is due to the superposition of the light diffracted from the pinholes. The beams are partially coherent with degree of coherence given by eqn [26]. Since the pinholes are symmetrically placed, the intensity due to either of the pinholes at a point P at the back focal plane of the lens L_2 is the same and is given by the Fraunhofer formula for diffraction from circular apertures, i.e.:

$$I_1(P)=I_2(P)=\left|\frac{2J_1(u)}{u}\right|^2, \text{ with } u=\frac{2\pi\nu}{c}b\sin\phi \qquad [27]$$

and ϕ is the angle that the diffracted beam makes from normal to the plane of the pinholes. The intensity of the interference pattern produced at the back focal plane of the lens L_2 is:

$$I(P)=2I_1(P)\left[1+\left|\frac{2J_1(v)}{v}\right|^2\cos(\delta+\beta_{12})\right] \qquad [28]$$

where $\delta=\mathrm{d}\sin\phi$ is the phase difference between the two beams reaching P from P_1 and P_2. For the on-axis point O, the quantity δ is zero.

The maximum and minimum values of $I(P)$ are given by

$$I_{\max}(P)=2I_1(P)\left[1+|2J_1(v)/v|^2\right] \qquad [29(a)]$$

$$I_{\max}(P)=2I_1(P)\left[1-|2J_1(v)/v|^2\right] \qquad [29(b)]$$

Figure 7 shows an example of the observed fringe patterns obtained by Thompson (1958) in a subsequent experiment.

Types of Fields

As mentioned above, $\gamma_{12}(\tau)$ is a measure of the correlation of the complex field at any two points P_1

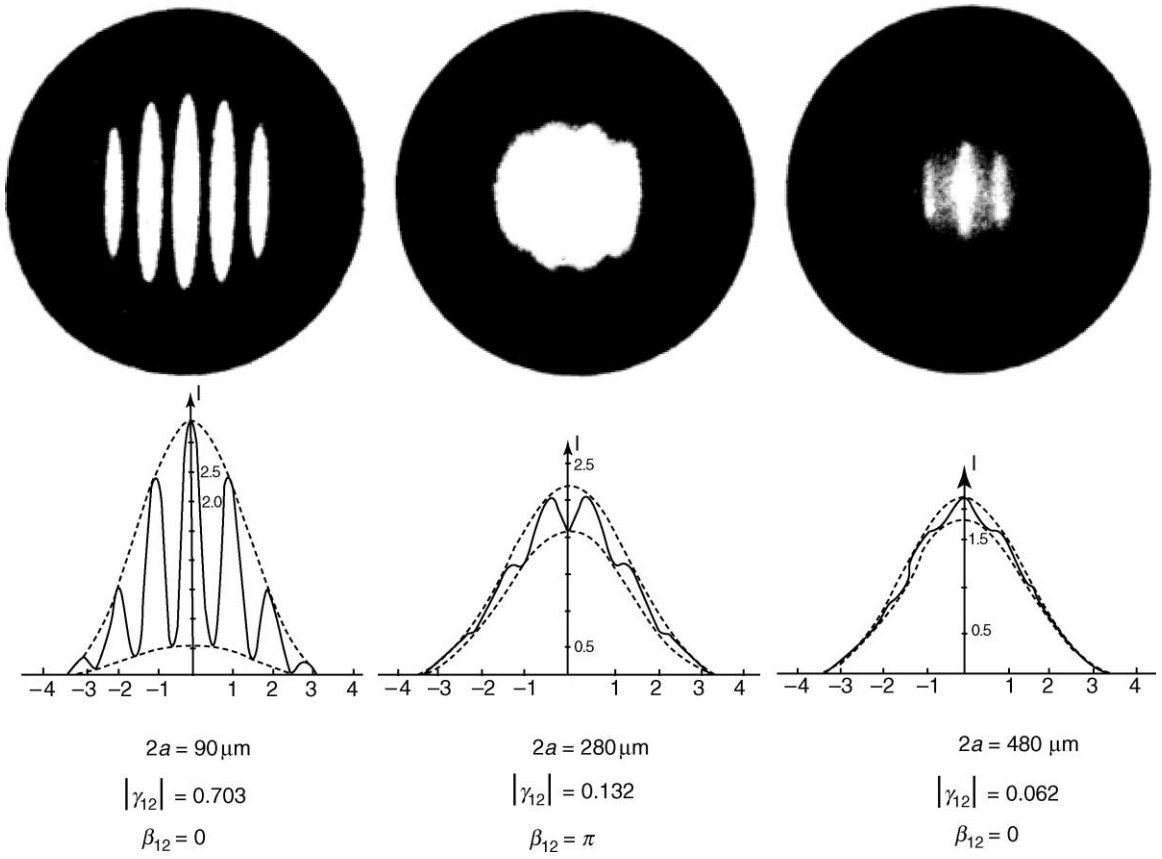


Figure 7 Two beam interference patterns obtained by using partially coherent light. The wavelength of light used was 0.579 μm and the separation *d* of the pinholes in the screen at B was 0.5 cm. The figure shows the observed fringe pattern and the calculated intensity variation for three sizes of the secondary source. The dashed lines show the maximum and minimum intensity. The values of the diameter, 2*a*, of the source and the corresponding values of the magnitude, $|\gamma_{12}|$, and the phase, β_{12}, are also shown in each case. Reproduced with permission from Thompson BJ (1958) Illustration of phase change in two-beam interference with partially coherent light. *Journal of the Optical Society of America* 48: 95–97.

and P_2 at specific time delay τ. Extending this definition, an optical field may be coherent or incoherent, if $|\gamma_{12}(\tau)| = 1$ or $|\gamma_{12}(\tau)| = 0$, respectively, for all pairs of points in the field and for all time delays. In the following, we consider some specific cases of fields and their properties.

Perfectly Coherent Fields

A field would be termed as perfectly coherent or self-coherent at a fixed point, if it has the property that $|\gamma(\mathbf{R}, \mathbf{R}, \tau)| = 1$ at some specific point $\mathbf{R}$ in the domain of the field for all values of time delay τ. It can also be shown that $\gamma(\mathbf{R}, \mathbf{R}, \tau)$ is periodic in time, i.e.:

$$\gamma(\mathbf{R}, \mathbf{R}, \tau) = \mathrm{e}^{-2\pi \mathrm{i} \nu_0 \tau} \quad \text{for} \quad \nu_0 > 0 \qquad [30]$$

For any other point $\mathbf{r}$ in the domain of the field, $\gamma(\mathbf{R}, \mathbf{r}, \tau)$ and $\gamma(\mathbf{r}, \mathbf{R}, \tau)$ are also unimodular and also periodic in time.

A perfectly coherent optical field at two fixed points has the property that $|\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau)| = 1$ for any two fixed points $\mathbf{R}_1$ and $\mathbf{R}_2$ in the domain of the field for all values of time delay τ. For such a field, it can be shown that $\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau) = \exp[\mathrm{i}(\beta - 2\pi\nu_0\tau)]$ with ν_0 (> 0) and β are real constants. Further, as a corollary, $|\gamma(\mathbf{R}_1, \mathbf{R}_1, \tau)| = 1$ and $|\gamma(\mathbf{R}_2, \mathbf{R}_2, \tau)| = 1$ for all τ, i.e., the field is self-coherent at each of the field points $\mathbf{R}_1$ and $\mathbf{R}_2$, and

$$\gamma(\mathbf{R}_1, \mathbf{R}_1, \tau) = \gamma(\mathbf{R}_2, \mathbf{R}_2, \tau) = \exp(-2\pi \mathrm{i} \nu_0 \tau) \text{ for } \nu_0 > 0$$

It can be shown that for any other point $\mathbf{r}'$ within the field

$$\begin{aligned} &\gamma(\mathbf{r}, \mathbf{R}_2, \tau) = \gamma(\mathbf{R}_1, \mathbf{R}_2, 0)\gamma(\mathbf{r}, \mathbf{R}_1, \tau) = \gamma(\mathbf{r}, \mathbf{R}_1, \tau)\mathrm{e}^{\mathrm{i}\beta} \\ &\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau) = \gamma(\mathbf{R}_1, \mathbf{R}_2, 0) \exp[-2\pi \mathrm{i} \nu_0 \tau] \end{aligned} \qquad [31]$$

A perfectly coherent optical field for all points in the domain of the field has the property that $|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| = 1$ for all pairs of points $\mathbf{r}_1$ and $\mathbf{r}_2$ in the domain of the field, for all values of time delay τ.

It can then be shown that $\gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\exp\{i[\alpha(\mathbf{r}_1)-\alpha(\mathbf{r}_2)]-2\pi i\nu_0\tau\}$, where $\alpha(\mathbf{r}')$ is real function of a single position $\mathbf{r}'$ in the domain of the field. Further, the mutual coherence function $\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)$ for such a field has a factorized periodic form as

$$\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=U(\mathbf{r}_1)U^*(\mathbf{r}_2)\ \exp(-2\pi i\nu_0\tau) \qquad [32]$$

which means the field is monochromatic with frequency ν_0 and $U(\mathbf{r})$ is any solution of the Helmholtz equation:

$$\nabla^2U(\mathbf{r})+k_0^2U(\mathbf{r})=0, \quad k_0=2\pi\nu_0/c \qquad [33]$$

The spectral and cross-spectral densities of such fields are represented by Dirac δ-functions with their singularities at some positive frequency ν_0. Such fields, however, can never occur in nature.

Quasi-Monochromatic Fields

Optical fields are found in practice for which spectral spread $\Delta\nu$ is much smaller than the mean frequency $\bar{\nu}$. These are termed as quasi-monochromatic fields. The cross-spectral density $W(\mathbf{r}_1,\mathbf{r}_2,\nu)$ of the quasi-monochromatic fields attains an appreciable value only in the small region $\Delta\nu$, and falls to zero outside this region.

$$W(\mathbf{r}_1,\mathbf{r}_2,\nu)=0, \quad |\nu-\bar{\nu}|>\Delta\nu \ \text{ and } \ \Delta\nu\ll\bar{\nu} \qquad [34]$$

The mutual coherence function $\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)$ can be written as the Fourier transform of cross-spectral density function as

$$\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=e^{-2\pi i\bar{\nu}\tau}\int_0^\infty W(\mathbf{r}_1,\mathbf{r}_2,\nu)e^{-2\pi i(\nu-\bar{\nu})\tau}\,d\nu \qquad [35]$$

If we limit our attention to small τ such that $\Delta\nu|\tau|\ll1$ holds, the exponential factor inside the integral is approximately equal to unity and eqn [35] reduces to

$$\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\exp(-2\pi i\bar{\nu}\tau)\int_0^\infty[W(\mathbf{r}_1,\mathbf{r}_2,\nu)]\,d\nu \qquad [36]$$

which gives

$$\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\Gamma(\mathbf{r}_1,\mathbf{r}_2,0)\exp(-2\pi i\bar{\nu}\tau) \qquad [37]$$

Equation [37] describes the behavior of $\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)$ for a limited range of τ values for quasi-monochromatic fields and in this range it behaves as a monochromatic field of frequency $\bar{\nu}$. However, due to the factor $\Gamma(\mathbf{r}_1,\mathbf{r}_2,0)$, the quasi-monochromatic field may be coherent, partially coherent, or even incoherent.

Cross-Spectrally Pure Fields

The complex degree of coherence, if it can be factored into a product of a component dependent on spatial coordinates and a component dependent on time delay, is called reducible. In the case of perfectly coherent light, the complex degree of coherence is reducible, as we have seen above, e.g., in eqn [32], and in the case of quasi-monochromatic fields, this is reducible approximately as shown in eqn [37]. There also exists a very special case of a cross-spectrally pure field for which the complex degree of coherence is reducible. A field is called a cross-spectrally pure field if the normalized spectrum of the superimposed light is equal to the normalized spectrum of the component beams, a concept introduced by Mandel. In the space-frequency domain, the intensity interference law is expressed as the so-called spectral interference law:

$$S(\mathbf{r},\nu)=S^{(1)}(\mathbf{r},\nu)+S^{(2)}(\mathbf{r},\nu)+2\left[\sqrt{S^{(1)}(\mathbf{r},\nu)}\sqrt{S^{(2)}(\mathbf{r},\nu)}\right]\mathrm{Re}[\mu(\mathbf{r}_1,\mathbf{r}_2,\nu)e^{-2\pi i\nu\tau}] \qquad [38]$$

where $\mu(\mathbf{r}_1,\mathbf{r}_2,\nu)$ is the spectral degree of coherence, defined in eqn [21] and τ is the relative time delay that is needed by the light from the pinholes to reach any point on the screen; $S^{(1)}(\mathbf{r},\nu)$ and $S^{(2)}(\mathbf{r},\nu)$ are the spectral densities of the light reaching P from the pinholes P_1 and P_2, respectively (see **Figure 4**) and it is assumed that the spectral densities of the field at pinholes P_1 and P_2 are the same [$S(\mathbf{r}_1,\nu)=CS(\mathbf{r}_2,\nu)$]. Now, if we consider a point for which the time delay is τ_0, then it can be seen that the last term in eqn [38] would be independent of frequency, provided that

$$\mu(\mathbf{r}_1,\mathbf{r}_2,\nu)\ \exp(-2\pi i\nu\tau_0)=f(\mathbf{r}_1,\mathbf{r}_2,\tau_0) \qquad [39]$$

where $f(\mathbf{r}_1,\mathbf{r}_2,\tau_0)$ is a function of $\mathbf{r}_1$, $\mathbf{r}_2$ and τ_0 only and the light at this point would have the same spectral density as that at the pinholes. If a region exists around the specified point on the observation plane, such that the spectral distribution of the light in this region is of the same form as the spectral distribution of the light at the pinholes, the light at the pinholes is cross-spectrally pure light.

In terms of the spectral distribution of the light $S(\mathbf{r}_1,\nu)$ at pinhole P_1 and $S(\mathbf{r}_2,\nu)[=CS(\mathbf{r}_1,\nu)]$ at pinhole P_2, the mutual coherence function at the pinholes can be written as

$$\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\sqrt{C}\int S(\mathbf{r}_1,\nu)\mu(\mathbf{r}_1,\mathbf{r}_2,\nu)\ \exp(-2\pi i\nu\tau)d\nu \qquad [40]$$

and using eqn [39], we get the very important condition for the field to be cross-spectrally pure, i.e.:

$$\gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\gamma(\mathbf{r}_1,\mathbf{r}_2,\tau_0)\gamma(\mathbf{r}_1,\mathbf{r}_1,\tau-\tau_0) \qquad [41]$$

The complex degree of coherence $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ is thus expressible as the product of two factors: one factor characterizes the spatial coherence at the two pinholes at time delay τ_0 and the other characterizes the temporal coherence at one of the pinholes. Equation [41] is known as the reduction formula for cross-spectrally pure light. It can further be shown that

$$\mu(\mathbf{r}_1, \mathbf{r}_2, \nu) = \gamma(\mathbf{r}_1, \mathbf{r}_2, \tau_0)\ \exp(2\pi i \nu \tau_0) \qquad [42]$$

Thus, the absolute value of the spectral degree of coherence is the same for all frequencies and is equal to the absolute value of the degree of coherence for the point specified by τ_0. It has been shown that cross-spectrally pure light can be generated, for example, by linear filtering of light that emerges from the two pinholes in Young's interference experiments.

Types of Sources

Primary Sources

A primary source is a set of radiating atoms or molecules. In a primary source the randomness comes from true source fluctuations, i.e., from the spatial distributions of fluctuating charges and currents. Such a source gives rise to a fluctuating field. Let $Q(\mathbf{r}, t)$ represent the fluctuating source variable at any point $\mathbf{r}$ at time t, then the field generated by the source is represented by fluctuating field variable $V(\mathbf{r}, t)$. The source is assumed to be localized in some finite domain such that $Q(\mathbf{r}, t) = 0$ at any time $t > 0$ outside the domain. Assuming that field variable $V(\mathbf{r}, t)$ and the source variable $Q(\mathbf{r}, t)$ are scalar quantities, they are related by an inhomogeneous equation as

$$\nabla^2 V(\mathbf{r}, t) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} V(\mathbf{r}, t) = -4\pi Q(\mathbf{r}, t) \qquad [43]$$

The mutual coherence functions of the source $\Gamma_Q(\mathbf{r}_1, \mathbf{r}_2, \tau) = \langle Q^*(\mathbf{r}_1, t) Q(\mathbf{r}_2, t + \tau)\rangle$ and of the field $\Gamma_V(\mathbf{r}_1, \mathbf{r}_2, \tau) = \langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t + \tau)\rangle$ characterize the statistical similarity of the fluctuating quantities at the points $\mathbf{r}_1$ and $\mathbf{r}_2$. Following eqn [20], one can define, respectively, the cross-spectral density function of the source and the field as

$$W_Q(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_{-\infty}^{\infty} \Gamma_Q(\mathbf{r}_1, \mathbf{r}_2, \tau) e^{-2\pi i \nu t}\, d\tau \qquad [44]$$

$$W_V(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_{-\infty}^{\infty} \Gamma_V(\mathbf{r}_1, \mathbf{r}_2, \tau) e^{-2\pi i \nu t}\, d\tau$$

The cross-spectral density functions of the source and of the field are related as

$$(\nabla_2^2 + k^2)(\nabla_1^2 + k^2) W_V(\mathbf{r}_1, \mathbf{r}_2, \nu) = 4\pi^2 W_Q(\mathbf{r}_1, \mathbf{r}_2, \nu) \qquad [45]$$

The solution of eqn [45] is represented as

$$W_V(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_S \int_S W_Q(\mathbf{r}'_1, \mathbf{r}'_2, \nu) \frac{e^{ik(R_2 - R_1)}}{R_1 R_2} d^3 r'_1\, d^3 r'_2 \qquad [46]$$

where $R_1 = |\mathbf{r}_1 - \mathbf{r}'_1|$ and $R_2 = |\mathbf{r}_2 - \mathbf{r}'_2|$ (see Figure 8). Using eqn [46], one can then obtain an expression for the spectrum at a point ($\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r} = r\mathbf{u}$) in the far field ($r \gg r'_1, r'_2$) of a source as

$$S^{\infty}(r\mathbf{u}, \nu) = \frac{1}{r^2} \int_S \int_S W_Q(\mathbf{r}'_1, \mathbf{r}'_2, \nu) e^{-ik\mathbf{u}\cdot(\mathbf{r}'_1 - \mathbf{r}'_2)} d^3 r'_1\, d^3 r'_2 \qquad [47]$$

where $\mathbf{u}$ is the unit vector along $\mathbf{r}$. The integral in eqn [47], i.e., the quantity defined by $r^2 S^{\infty}(r\mathbf{u}, \nu)$, is also defined as the radiant intensity, which represents the rate of energy radiated at the frequency ν from the source per unit solid angle in the direction of $\mathbf{u}$.

Secondary Sources

Sources used in a laboratory are usually secondary planar sources. A secondary source is a field, which arises from the primary source in the region outside the domain of the primary sources. This kind of source is an aperture on an opaque screen illuminated either directly or via an optical system by primary sources. Let $V(\boldsymbol{\rho}, t)$ represent the fluctuating field in a secondary source plane σ at $z = 0$ (Figure 9) and $W_0(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \nu)$ represent its cross-spectral density (the subscript 0 refers to $z = 0$). One can then solve eqn [25] to obtain the propagation of the cross-spectral density from this planar source. For two points P_1 and P_2 located at distances which are large compared to wavelength, the cross-spectral density is

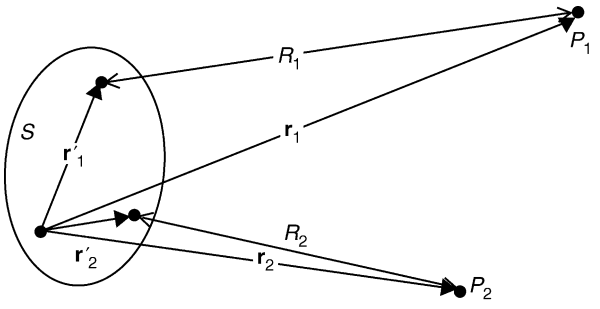


Figure 8 Geometry of a 3-dimensional primary source S and the radiation from it.

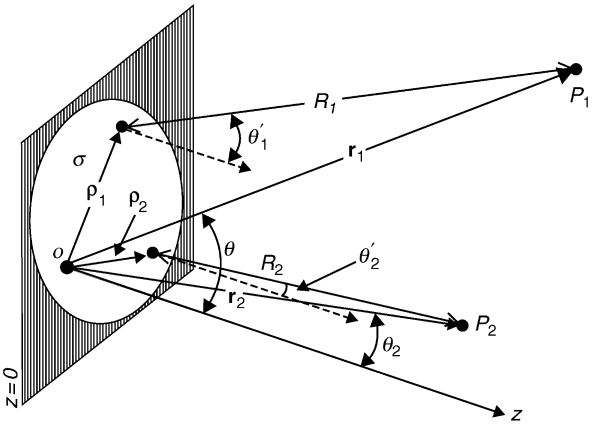


Figure 9 Geometry of a planar source σ and the radiation from it.

then given by

$$W(\mathbf{r}_1, \mathbf{r}_2, \nu) = \left(\frac{k}{2\pi}\right)^2 \int_\sigma \int_\sigma W_0(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \nu) \times \frac{e^{ik(R_2 - R_1)}}{R_1 R_2} \cos\theta'_1 \cos\theta'_2 \, d^2\rho_1 \, d^2\rho_2 \quad [48]$$

where $R_j = |\mathbf{r}_j - \boldsymbol{\rho}_j|$, $(j = 1, 2)$, θ'_1 and θ'_2 are the angles that R_1 and R_2 directions make with the z-axis (Figure 9). Using eqn [48], one can then obtain an expression for the spectrum at a point $(\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r} = r\mathbf{u})$ in the far field $(r \gg \rho_1, \rho_2)$ of a planar source as

$$S^\infty(r\mathbf{u}, \nu) = \frac{k^2}{2\pi^2} \frac{\cos^2\theta}{r^2} \int_\sigma \int_\sigma W_0(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \nu) \times e^{-ik\mathbf{u}_\perp \cdot (\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2)} d^2\rho_1 \, d^2\rho_2 \quad [49]$$

where $\mathbf{u}_\perp$ is the projection of the unit vector $\mathbf{u}$ on the plane σ of the source.

Schell-Model Sources

In the framework of coherence theory in space–time domain, two-dimensional planar model sources of this kind were first discussed by Schell. Later, the model was adopted for formulation of coherence theory in space–frequency domain. Schell-model sources are the sources whose degree of spectral coherence $\mu_A(\mathbf{r}_1, \mathbf{r}_2, \nu)$ (for either primary or secondary source) is stationary in space. It means that $\mu_A(\mathbf{r}_1, \mathbf{r}_2, \nu)$ depends on $\mathbf{r}_1$ and $\mathbf{r}_2$ only through the difference $\mathbf{r}_2 - \mathbf{r}_1$, i.e., of the form:

$$\mu_A(\mathbf{r}_1, \mathbf{r}_2, \nu) \equiv \mu_A(\mathbf{r}_2 - \mathbf{r}_1, \nu) \quad [50]$$

for each frequency ν present in the source spectrum. Here A stands for field variables V, in the case of a Schell-model secondary source and for source variables Q, in the case of a Schell-model primary source. The cross-spectral density function of a Schell-model source is of the form:

$$W_A(\mathbf{r}_1, \mathbf{r}_2, \nu) = [S_A(\mathbf{r}_1, \nu)]^{1/2} [S_A(\mathbf{r}_2, \nu)]^{1/2} \mu_A(\mathbf{r}_2 - \mathbf{r}_1, \nu) \quad [51]$$

where $S_A(\mathbf{r}, \nu)$ is the spectral density of the light at a typical point in the primary source or on the plane of a secondary source. Schell model sources do not assume low coherence, and therefore, can be applied to spatially stationary light fields of any state of coherence. The Schell-model of the form given in eqn [51] has been used to represent both three-dimensional primary sources and two-dimensional secondary sources.

Quasi-Homogeneous Sources

Useful models of partially coherent sources that are frequently encountered in nature or in the laboratory are the so-called quasi-homogeneous sources. These are an important sub-class of Schell-model sources. A Schell-model source is called quasi-homogeneous if the intensity of a Schell model source is essentially constant over any coherence area. Under these approximations, the cross-spectral density function for a quasi-homogeneous source is given by

$$W_A(\mathbf{r}_1, \mathbf{r}_2, \nu) = S_A\left[\tfrac{1}{2}(\mathbf{r}_1 + \mathbf{r}_2), \nu\right] \mu_A(\mathbf{r}_2 - \mathbf{r}_1, \nu) = S_A[\mathbf{r}, \nu] \mu_A(\mathbf{r}', \nu) \quad [52]$$

where $\mathbf{r} = (\mathbf{r}_1 + \mathbf{r}_2)/2$, and $\mathbf{r}' = \mathbf{r}_2 - \mathbf{r}_1$. The subscript A stands for either V or Q for the field variable or a source variable, respectively. It is clear that for a quasi-homogeneous source the spectral density $S_A(\mathbf{r}, \nu)$ varies so slowly with position that it is approximately constant over distances across the sources that are of the order of the correlation length Λ, which is a measure of the effective width of $|\mu_A(\mathbf{r}', \nu)|$. Therefore, $S_A(\mathbf{r}, \nu)$ is a slowly varying function of $\mathbf{r}$ (Figure 10b) and $|\mu_A(\mathbf{r}', \nu)|$ is a fast varying function of $\mathbf{r}'$ (Figure 10a). In addition, the linear dimensions of the source are large compared with the wavelength of light and with the correlation length Δ (Figure 10c).

Quasi-homogeneous sources are always spatially incoherent in the 'global' sense, because their linear dimensions are large compared with the correlation length. This model is very good for representing two-dimensional secondary sources with sufficiently low coherence such that the intensity does not vary over the coherence area on the input plane. It has also been applied to three-dimensional primary sources,

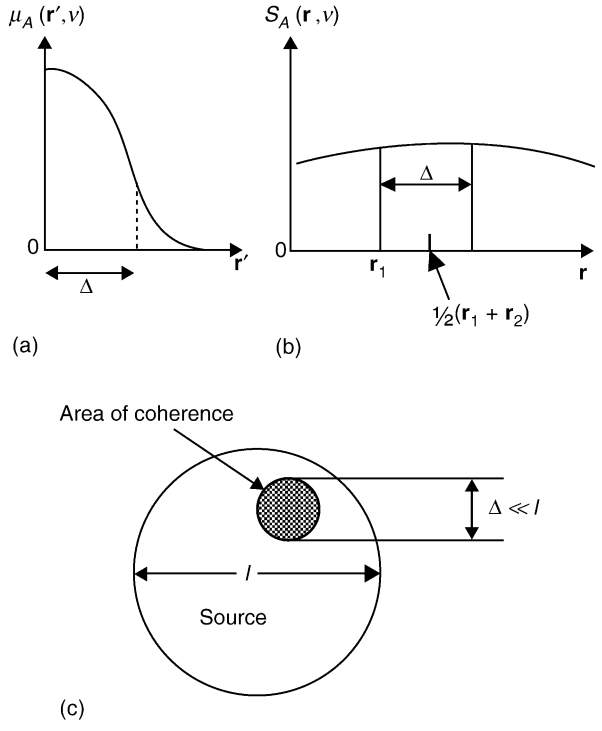


Figure 10 Concept of quasi-homogeneous sources.

three-dimensional scattering potentials, and two-dimensional primary and secondary sources.

Equivalence Theorems

The study of partially coherent sources led to the formulations of a number of equivalence theorems, which show that sources of any state of coherence can produce the same distribution of the radiant intensity as a fully spatially coherent laser source. These theorems provide conditions under which sources of different spatial distribution of spectral density and of different state of coherence will generate fields, which have the same radiant intensity. It has been shown, by taking examples of Gaussian–Schell model sources, that sources of completely different coherence properties and different spectral distributions across the source generate identical distribution of radiant intensity. Experimental verifications of the results of these theorems have also been carried out. For further details on this subject, the reader is referred to Mandel and Wolf (1995).

Correlation-Induced Spectral Changes

It was assumed that spectrum is an intrinsic property of light that does not change as the radiation propagates in free-space, until studies on partially coherent sources and radiations from them in 1980s, revealed that this was true only for specific type of sources. It was discovered on general grounds that the spectrum of light, which originates in an extended source, either a primary source or a secondary source, depends not only on the source spectrum but also on the spatial coherence properties of the source. It was also predicted theoretically by Wolf that the spectrum of light would, in general, be different from the spectrum of the source, and be different at different points in space on propagation in free space.

For a quasi-homogeneous planar secondary source defined by eqn [52], whose normalized spectrum is the same at each source point, one can write the spectral density as

$$S_0(\boldsymbol{\rho}, \nu) = I_0(\boldsymbol{\rho}) g_0(\nu) \quad \text{with} \quad \int_0^\infty g_0(\nu) \mathrm{d}\nu = 1 \qquad [53]$$

where $I_0(\boldsymbol{\rho})$ is the intensity of light at point $\boldsymbol{\rho}$ on the plane of the source, $g_0(\nu)$ is the normalized spectrum of the source and the subscript 0 refers to the quantities of the source plane. Using eqn [49], one can obtain an expression for the far-field spectrum due to this source as

$$S^{\infty}(r\mathbf{u}, \nu) = \frac{k^2}{2\pi^2} \frac{\cos^2 \theta}{r^2} \int_\sigma \int_\sigma I_0(\boldsymbol{\rho}) g_0(\nu) \mu_0(\boldsymbol{\rho}', \nu) \times \mathrm{e}^{-\mathrm{i}k\mathbf{u}_\perp \cdot (\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2)} \, \mathrm{d}^2\rho_1 \, \mathrm{d}^2\rho_2 \qquad [54]$$

Noting that $\boldsymbol{\rho} = (\boldsymbol{\rho}_1 + \boldsymbol{\rho}_2)/2$ and $\boldsymbol{\rho}' = \boldsymbol{\rho}_2 - \boldsymbol{\rho}_1$, one can transform the variables of the integration and obtain after some manipulation:

$$S^{(\infty)}(r\mathbf{u}, \nu) = \frac{k^2 \cos^2 \theta}{(2\pi r)^2} \tilde{I}_0 \tilde{\mu}_0(k\mathbf{u}_\perp, \nu) g_0(\nu) \qquad [55]$$

where

$$\tilde{\mu}_0(k\mathbf{u}_\perp, \nu) = \int_\sigma \mu_0(\boldsymbol{\rho}', \nu) \mathrm{e}^{-\mathrm{i}k\mathbf{u}_\perp \cdot \boldsymbol{\rho}'} \, \mathrm{d}^2\rho' \qquad [56]$$

and

$$\tilde{I}_0 = \int_\sigma I_0(\boldsymbol{\rho}) \mathrm{d}^2\rho \qquad [57]$$

Equation [55] shows that the spectrum of the field in the far-zone depends on the coherence properties of the source through its spectral degree of coherence $\mu_0(\boldsymbol{\rho}', \nu)$ and on the normalized source spectrum $g_0(\nu)$.

Scaling Law

The reason why coherence-induced spectral changes were not observed until recently is that the usual thermal sources employed in laboratories or

commonly encountered in nature have special coherence properties and the spectral degree of coherence has the function form:

$$\mu^{(0)}(\boldsymbol{\rho}_2-\boldsymbol{\rho}_1,\nu)=f[k(\boldsymbol{\rho}_2-\boldsymbol{\rho}_1)] \quad \text{with} \quad k=\frac{2\pi\nu}{c} \qquad [58]$$

which shows that the spectral degree of coherence depends only on the product of the frequency and space coordinates. This formula expresses the so-called scaling law, which was enunciated by Wolf. Commonly used sources satisfy this property. For example, the spectral degree of coherence of Lambertian sources and black-body sources can be shown to be

$$\mu_0(\boldsymbol{\rho}_2-\boldsymbol{\rho}_1,\nu)=\frac{\sin(k|\boldsymbol{\rho}_2-\boldsymbol{\rho}_1|)}{k|\boldsymbol{\rho}_2-\boldsymbol{\rho}_1|} \qquad [59]$$

This expression evidently satisfies the scaling law. If the spectral degree of coherence does not satisfy the scaling law, the normalized far-field spectrum will, in general, vary in different directions in the far-zone and will differ from the source spectrum.

Spectral Changes in Young's Interference Experiment

Spectral changes in Young's interference experiment with broadband light are not as well understood as in experiments with quasi-monochromatic, probably because in such experiments no interference fringes are formed. However, if one were to analyze the spectrum of the light in the region of superposition, one would observe changes in the spectrum of light in the region of superposition in the form of a shift in the spectrum for narrowband spectrum and spectral modulation for broadband light. One can readily derive an expression for the spectrum of light in the region of superposition. Let $S^{(1)}(P,\nu)$ be the spectral density of the light at P which would be obtained if the small aperture at P_1 alone was open; $S^{(2)}(P,\nu)$ has a similar meaning if only the aperture at P_2 was open (Figure 4).

Let us assume, as is usually the case, that $S^{(2)}(P,\nu)\approx S^{(1)}(P,\nu)$ and let d be the distance between the two pinholes. Consider the spectral density at the point P, at distance x from the axis of symmetry in an observation plane located at distance of R from the plane containing the pinholes. Assuming that $x/R \ll 1$, one can make the approximation $R_2 - R_1 \approx xd/R$. The spectral interference law (eqn [38]) can then be written as

$$S(P,\nu)\approx 2S^{(1)}(P,\nu)\{1+|\mu(P_1,P_2,\nu)| \times\cos[\beta(P_1,P_2,\nu)+2\pi\nu xd/cR]\} \qquad [60]$$

where $\beta(P_1,P_2,\nu)$ denotes the phase of the spectral degree of coherence. Equation [60] implies the two results:

(i) at any fixed frequency ν, the spectral density varies sinusoidally with the distance x of the point from the axis, with the amplitude and the phase of the variation depending on the (generally complex) spectral degree of coherence $\mu(P_1,P_2,\nu)$; and
(ii) at any fixed point P in the observation plane the spectrum $S(P,\nu)$ will, in general, differ from the spectrum $S^{(1)}(P,\nu)$, the change also depending on the spectral degree of coherence $\mu(P_1,P_2,\nu)$ of the light at the two pinholes.

Experimental Confirmations

Experimental tests of the theoretical prediction of spectral invariance and noninvariance due to correlation of fluctuations across the source were performed just after the theoretical predictions. Figure 11 shows results of one such experiment in which spectrum changes in the Young's experiment were studied. Several other experiments also reported confirmation of the source correlation-dependent spectral changes. One of the important applications of these observations has been to explain the discrepancies in the maintenance of the spectroradiometric scales by national laboratories in different countries. These studies also have potential application in determining experimentally the spectral degree of coherence of partially coherent fields. The knowledge of spectral degree of coherence is often important in remote sensing, e.g., for determining angular diameters of stars.

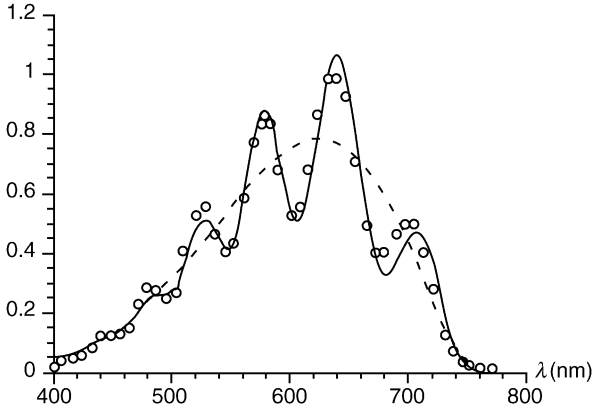


Figure 11 Correlation-induced changes in spectrum in Young's interference. Dashed line represents the spectrum when only one of the slits is open, the continuous curve shows the spectrum when both the slits are open and the circles are the measured values in the latter case. Reproduced with permission from Santarsiero M and Gori F (1992) Spectral changes in Young interference pattern. *Physics Letters* 167: 123–128.

Applications of Optical Coherence

Stellar Interferometry

The Michelson stellar interferometer, named after Albert Michelson, was used to determine the angular diameters of stars and also the intensity distribution across the star. The method was devised by Michelson without using any concept of coherence, although subsequently the full theory of the method was developed on the basis of propagation of correlations. A schematic of the experiment is shown in Figure 12. The interferometer is mounted in front of a telescope, a reflecting telescope in this case. The light from a star is reflected from mirrors M_1 and M_2 and is directed towards the primary mirror (or the objective lens) of the telescope. The two beams thus collected superpose in the focal plane F of telescope where an image crossed with fringes is formed. The outer mirrors M_1 and M_2 can be moved along the axis defined as $M_1M_3M_4M_2$ while the inner mirrors M_3 and M_4 remain fixed. The fringe spacing depends on the position of mirrors M_3 and M_4 and hence is fixed, while the visibility of the fringes depends on the separation of the mirrors M_1 and M_2 and hence, can be varied. Michelson showed that from the measurement of the variation of the visibility with the separation of the two mirrors, one could obtain information about the intensity distribution of the stars, which are rotationally symmetric. He also showed that if the stellar disk is circular and uniform, the visibility curve as a function of the separation d of the mirrors M_1 and M_2 will have zeros for certain values of d, and that the smallest of these d values for which zero occurs is given by $d_0 = 0.61\lambda_a/\alpha$, where α is the semi-angular diameter of the star and λ_a is the mean wavelength of the filtered quasi-monochromatic light from the star. Angular diameters of several stars down to 0.02 second of an arc were determined.

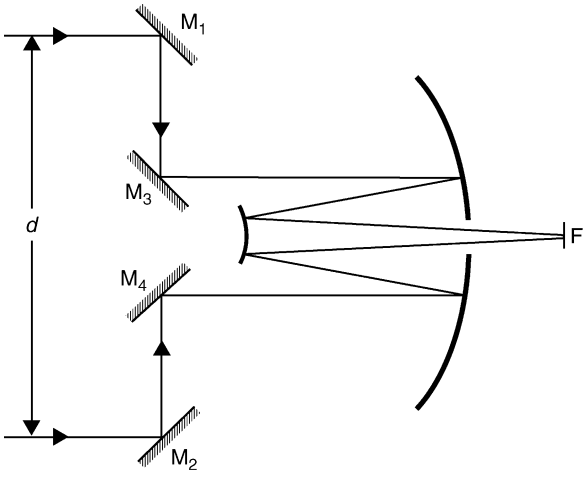


Figure 12 Schematic of the Michelson stellar interferometer.

From the standpoint of second-order coherence theory the principles of the method can be readily understood. The star is considered an incoherent source and according to the van Cittert–Zernike theorem, the light reaching the outer mirrors M_1 and M_2 of the interferometer will be partially coherent. This coherence would depend on the size of and the intensity distribution across the star. Let (x_1, y_1) and (x_2, y_2) be the coordinates of the positions of the mirrors M_1 and M_2, respectively, and (ξ, η) the coordinates of a point on the surface plane of the star which is assumed to be at a very large (astronomical) distance R from the mirrors. The complex degree of coherence at the mirrors would then be given by eqn [22] which can now be written as

$$\gamma(\Delta x, \Delta y, 0) = \frac{\int_\sigma I(u,v)\mathrm{e}^{-\mathrm{i}k_a(u\Delta x + v\Delta y)}\,\mathrm{d}u\,\mathrm{d}v}{\int_\sigma I(u,v)\mathrm{d}u\,\mathrm{d}v} \qquad [61]$$

where $I(u, v)$ is the intensity distribution across the star disk σ as a function of the angular coordinates $u = \xi/R$, $v = \eta/R$, $\Delta x = x_1 - x_2$, $\Delta y = y_1 - y_2$, and $k_a = 2\pi/\lambda_a$, λ_a being the mean wavelength of the light from the star. Equation [61] shows that the equal-time ($\tau = 0$) complex degree of coherence of the light incident at the outer mirrors of the interferometer is the normalized Fourier-transform of the intensity distribution across the stellar disk. Further, eqn [15] shows that the visibility of the interference fringes is the absolute value of γ, if the intensity of the two interfering beams is equal, as in the present case. The phase of γ can be determined by the position of the intensity maxima of the fringe pattern (eqn [14]). If one is interested in determining only the angular size of the star and the star is assumed to be a circularly symmetric disk of angular diameter 2α and of uniform intensity [$I(u, v)$ is constant across the disk], then eqn [61] reduces to

$$\gamma(\Delta x,\Delta y)=\frac{2J_1(v)}{v}, \quad v=\frac{2\pi\alpha}{\lambda_a}d, \quad d=\sqrt{(\Delta x)^2+(\Delta y)^2} \qquad [62]$$

The smallest separation of the mirrors for which the visibility γ vanishes corresponds to $v = 3.832$, i.e., $d_0 = 0.61\lambda_a/\alpha$, which is in agreement with Michelson's result.

Interference Spectroscopy

Another contribution of Michelson, which was subsequently identified as an application of the

coherence theory, was the use of his interferometer (Figure 2) to determine the energy distribution in the spectral lines. The method he developed is capable to resolving spectral lines that are too narrow to be analyzed by the spectrometer. The visibility of the interference fringes depends on the energy distribution in the spectrum of the light and its measurement can give information about the spectral lines. In particular, if the energy distribution in a spectral line is symmetric about some frequency ν_0, its profile is simply the Fourier transform of the visibility variation as a function of the path difference between the two interfering beams. This method is the basis of interference spectroscopy or the Fourier transform spectroscopy.

Within the framework of second-order coherence theory, if the mean intensity of the two beams in a Michelson's interferometer is the same, then the visibility of the interference fringes in the observation plane is related to the complex degree of coherence of light at a point at the beamsplitter where the two beams superimpose (eqn [15]). The two quantities are related as $v(\tau) = |\gamma(\tau)|$ where $\gamma(\tau) \equiv \gamma(\mathbf{r}_1, \mathbf{r}_1, \tau)$ is the complex degree of self-coherence at the point $\mathbf{r}_1$ on the beamsplitter. Following eqn [19], $\gamma(\tau)$ can be represented by a Fourier integral as

$$\gamma(\tau) = \int_0^\infty s(\nu)\exp(-\mathrm{i}2\pi\nu\tau)\mathrm{d}\nu \qquad [63]$$

where $s(\nu)$ is the normalized spectral density of the light defined as $s(\nu) = S(\nu)/\int_0^\infty S(\nu)\mathrm{d}\nu$ and $S(\nu) \equiv S(\mathbf{r}_1, \nu) = W(\mathbf{r}_1, \mathbf{r}_1, \nu)$ is the spectral density of the beam at the point $\mathbf{r}_1$. The method is usually applied to very narrow spectral lines for which one can assume that the peak that occurs at ν_0, and $\gamma(\tau)$ can be represented as

$$\begin{aligned} \gamma(\tau) &= \bar{\gamma}(\tau)\exp(-2\pi\mathrm{i}\nu_0\tau) \quad \text{with} \\ \bar{\gamma}(\tau) &= \int_{-\infty}^\infty \bar{s}(\mu)\exp(-\mathrm{i}2\pi\mu\tau)\mathrm{d}\mu \end{aligned} \qquad [64]$$

where $\bar{s}(\mu)$ is the shifted spectrum such that

$$\begin{aligned} \bar{s}(\mu) &= s(\nu_0 + \mu) \quad \mu \geq -\nu_0 \\ &= 0 \qquad\qquad \mu < -\nu_0 \end{aligned}$$

From the above one readily gets

$$v(\tau) = |\gamma(\tau)| = \left| \int_{-\infty}^\infty \bar{s}(\mu)\,\exp(-\mathrm{i}2\pi\mu\tau)\mathrm{d}\mu \right| \qquad [65]$$

If the spectrum is symmetric about ν_0, then $v(\tau)$ would be an even function of τ and the Fourier inversion would give:

$$\bar{s}(\mu) = s(\nu_0 + \mu) = 2\int_0^\infty v(\tau)\cos(2\pi\mu\tau)\mathrm{d}\tau \qquad [66]$$

which can be used to calculate the spectral energy distribution for a symmetric spectrum about ν_0 from the visibility curve. However, for an asymmetric spectral distribution, the visibility and the phase of the complex degree of coherence must be determined as the Fourier transform of the shifted spectrum is no longer real everywhere.

Higher-Order Coherence

So far we have considered correlations of the fluctuating field variables at two space–time $(\mathbf{r}, t)$ points, as defined in eqn [10]. These are termed as second-order correlations. One can extend the concept of correlations to more than two space–time points, which will involve higher-order correlations. For example, one can define the space–time cross correlation function of order (M, N) of the random field $V(\mathbf{r}, t)$, represented by $\Gamma^{(M,N)}$, as an ensemble average of the product of the field $V(\mathbf{r}, t)$ values at N space–time points and $V^*(\mathbf{r}, t)$ at other M points. In this notation, the mutual coherence function as defined in eqn [10] would now be $\Gamma^{(1,1)}$. Among higher-order correlations, the one with $M = N = 2$, is of practical significance and is called the fourth-order correlation function, $\Gamma^{(2,2)}(\mathbf{r}_1, t_1, \mathbf{r}_2, t_2; \mathbf{r}_3, t_3, \mathbf{r}_4, t_4)$. The theory of Gaussian random variables tells us that any higher correlation can be written in terms of second-order correlations over all permutations of pairs of points. In addition, if we assume that $(\mathbf{r}_3, t_3) = (\mathbf{r}_1, t_1)$ and $(\mathbf{r}_4, t_4) = (\mathbf{r}_2, t_2)$, and that the field is stationary, then $\Gamma^{(2,2)}$ is called the intensity–intensity correlation and is given as

$$\begin{aligned} &\Gamma^{(2,2)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1) \\ &= \left\langle V(\mathbf{r}_1, t_1)V(\mathbf{r}_2, t_2)V^*(\mathbf{r}_1, t_1)V^*(\mathbf{r}_2, t_2)\right\rangle \\ &= \left\langle I(\mathbf{r}_1, t_1)I(\mathbf{r}_2, t_2)\right\rangle \\ &= \left\langle I(\mathbf{r}_1, t_1)\right\rangle\left\langle I(\mathbf{r}_2, t_2)\right\rangle\left(1 + |\gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1)|^2\right) \end{aligned} \qquad [67]$$

where

$$\gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1) = \frac{\Gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1)}{[\langle I(\mathbf{r}_1, t_1)\rangle]^{1/2}[\langle I(\mathbf{r}_2, t_2)\rangle]^{1/2}} \qquad [68]$$

We now define fluctuations in intensity at $(\mathbf{r}_j, t_j)$ as

$$\Delta I_j = I(\mathbf{r}_j, t_j) - \langle I(\mathbf{r}_j, t_j)\rangle$$

and then the correlation of intensity fluctuations becomes

$$\begin{aligned}\langle \Delta I_1 \Delta I_2 \rangle &= \langle I(\mathbf{r}_1,t_1) I(\mathbf{r}_2,t_2)\rangle - \langle I(\mathbf{r}_1,t_1)\rangle\langle I(\mathbf{r}_2,t_2)\rangle \\ &= \langle I(\mathbf{r}_1,t_1)\rangle\langle I(\mathbf{r}_2,t_2)\rangle |\gamma^{(1,1)}(\mathbf{r}_1,\mathbf{r}_2,t_2-t_1)|^2\end{aligned} \quad [69]$$

where we have used eqn [67]. Equation [69] forms the basis for intensity interferometry.

Hanbury-Brown and Twiss Experiment

In this landmark experiment conducted both on the laboratory scale and astronomical scale, Hanbury-Brown and Twiss demonstrated the existence of intensity–intensity correlations in terms of the correlations in the photocurrents in the two detectors and thus measured the squared modulus of complex degree of coherence. In the laboratory experiment, the arc of a mercury lamp was focused onto a circular hole to produce a secondary source. The light from this source was then divided equally into two parts through a beamsplitter. Each part was, respectively, detected by two photomultiplier tubes, which had identical square apertures in front. One of the tubes could be translated normally to the direction of propagation of light and was so positioned that its image through the splitter could be made to coincide with the other tube. Thus, by suitable translation a measured separation d between the two square apertures could be introduced. The output currents from the photomultiplier tubes were taken by cables of equal length to a correlator. In the path of each cable a high-pass filter was inserted, so that only the current fluctuations could be transmitted to the correlator. Thus, the normalized correlations between the two current fluctuations:

$$C(d) = \frac{\langle \Delta J_1(t) \Delta J_2(t)\rangle}{\langle [\Delta J_1(t)]^2\rangle^{1/2}\langle [\Delta J_2(t)]^2\rangle^{1/2}} \quad [70]$$

as a function of detector separation d could be measured. Now when the detector response time is much larger than the time-scale of the fluctuations in intensity, then it can be shown that the correlations in the fluctuations of the photocurrent are proportional to the correlations of the fluctuations in intensity of the light being detected. Thus, we would have

$$C(d) \approx \delta |\gamma^{(1,1)}(\mathbf{r}_1,\mathbf{r}_2,0)|^2 \quad [71]$$

where δ is the average number of photocounts of the light of one polarization during the time-scale of the fluctuations (for general thermal sources, this is much less than one). Equation [71] represents the Hanbury-Brown–Twiss effect.

Stellar Intensity Interferometry

Michelson stellar interferometry can resolve stars which have angular sizes of the order of $0.01''$, since for smaller stars, the separation between the primary mirrors runs into several meters and maintaining stability of mirrors such that the optical paths do not change, even by fraction of a wavelength, is extremely difficult. The atmospheric turbulence further adds to this problem and obtaining stable fringe pattern becomes next to impossible for very small stars. Hanbury-Brown and Twiss applied the intensity interferometry based on their photoelectric correlation technique for determining the angular sizes of such stars. Two separate parabolic mirrors collected light from the star and the output of the photodetectors placed at the focus of each mirror was sent to a correlator. The cable lengths were made unequal so as to compensate for the time difference of the light arrival at the two mirrors. The normalized correlation of the fluctuations of the photocurrents was determined as described above. This would give the variation of the modulus-squared degree of coherence as a function of the mirror separation d from which the angular size of the stars can be estimated. The advantage of the stellar intensity interferometer over the stellar (amplitude) interferometer is that the light need not interfere as in the latter, since the photodetectors are mounted directly at the focus of the primary mirrors of the telescope. Thus, the constraint on the large path difference between the two beams is removed and large values of d can now be used. Moreover, the atmosphere turbulence and the mirror movements have very small effect. Stellar angular diameters as small as $0.0004''$ of arc with resolution of $0.00003''$ could be measured by such interferometers.

See also

Coherence: Coherence and Imaging. **Coherent Lightwave Systems**. **Coherent Transients:** Coherent Transient Spectroscopy in Atomic and Molecular Vapours. **Coherent Control:** Applications in Semiconductors; Experimental; Theory. **Interferometry:** Overview. **Information Processing:** Coherent Analogue Optical Processors. **Terahertz Technology:** Coherent Terahertz Sources.

Further Reading

Beran MJ and Parrent GB (1964) *Theory of Partial Coherence*. Englewood Cliffs, NJ: Prentice-Hall.

Born M and Wolf E (1999) *Principles of Optics*. New York: Pergamon Press.
Carter WH (1996) Coherence theory. In: Bass M (ed.) *Hand Book of Optics*. New York: McGraw-Hill.
Davenport WB and Root WL (1960) *An Introduction to the Theory of Random Signals and Noise*. New York: McGraw-Hill.
Goodman JW (1985) *Statistical Optics*. Chichester: Wiley.
Hanbury-Brown R and Twiss RQ (1957) Interferometry of intensity fluctuations in light: I basic theory: the correlation between photons in coherent beams of radiation. *Proceedings of the Royal Society* 242: 300–324.
Hanbury-Brown R and Twiss RQ (1957) Interferometry of intensity fluctuations in light: II an experimental test of the theory for partially coherent light. *Proceedings of the Royal Society* 243: 291–319.
Kandpal HC, Vaishya JS and Joshi KC (1994) Correlation-induced spectral shifts in optical measurements. *Optical Engineering* 33: 1996–2012.
Mandel L and Wolf E (1965) Coherence properties of optical fields. *Review of Modern Physics* 37: 231–287.
Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge, UK: Cambridge University Press.
Marathay AS (1982) *Elements of Optical Coherence*. New York: John Wiley.
Perina J (1985) *Coherence of Light*. Boston, MA: M.A. Reidel.
Santarsiero M and Gori F (1992) Spectral changes in Young interference pattern. *Physics Letters* 167: 123–128.
Schell AC (1967) A technique for the determination of the radiation pattern of a partially coherent aperture. *IEEE Transactions on Antennas and Propagation* AP-15: 187.
Thompson BJ (1958) Illustration of phase change in two-beam interference with partially coherent light. *Journal of the Optical Society of America* 48: 95–97.
Thompson BJ and Wolf E (1957) Two beam interference with partially coherent light. *Journal of the Optical Society of America* 47: 895–902.
Wolf E and James DFV (1996) Correlation-induced spectral changes. *Report on Progress in Physics* 59: 771–818.

Coherence and Imaging

J van der Gracht, HoloSpex, Inc., Columbia, MD, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical imaging systems are strongly affected by the coherence of the light that illuminates the object of interest. In many cases, the light is approximately coherent or incoherent. These approximations lead to simple mathematical models for the image formation process and allow straightforward analysis and design of such systems. When the light is partially coherent, the mathematical models are more complicated and system analysis and design is more difficult. Partially coherent illumination is often used in microscopy, machine vision, and optical lithography. The intent of this article is to provide the reader with a basic understanding of the effects of coherence on imaging. The information should enable the reader to recognize when coherence effects are present in an imaging system and give insight into when coherence can be modified to improve imaging performance. The material relies mostly on concepts drawn from the Fourier optics perspective of imaging and a rigorous coherence theory treatment is avoided. We encourage the reader to consult the Further Reading list at the end of this article, for more complete definitions of terms for coherence theory. A number of experimental results highlight the effects of spatial coherence of the illumination on image formation. Discussions of the role of coherence in such key applications are interspersed throughout the article.

Image Formation – Ideal and Optimal

An image is typically defined as the reproduction or likeness of the form of an object. An image that is indistinguishable from the original object is generally considered to be ideal. In a general context, the sound of a voice coming from a loudspeaker can be thought of as the image of the sound coming directly from the original speaker's mouth. In optical imaging, the ideal image replicates the light emanating from the object. Taken to the extreme, the ideal image replicates the light leaving the object in terms of intensity, wavelength, polarization, and even coherence. When the final image is viewed by the eye, the ideal image only needs to replicate the spatial distribution of the light leaving the object in terms of color and relative intensity at each point on the object. (In this article, intensity is defined as optical power per unit area (watts per meter squared).)

A general diagram of a direct view image formation system is shown in Figure 1. The condenser optics gathers light from a primary source and illuminates a transmissive object having a complex wave amplitude

transmission of $O(x, y)$. The imaging optics produce an optical image that is viewed directly by the viewer. A growing number of image formation systems now include a solid-state image detector as shown in Figure 2. The raw intensity optical image is converted to an electronic signal that can be digitally processed and then displayed on a monitor. In this system, the spatial intensity distribution emanating from the monitor is the final image. The task of the optical system to the left of the detector is to gather spatial information about the light properties of the object. In fact, the information gathered by the detector includes information about the object, the illumination system, and the image formation optics. If the observer is only interested in the light transmission properties of the object, the effects of the illumination and image formation optics must be well understood. When the light illuminating the object is known to be coherent or incoherent, reasonably simple models for the overall image formation process can be used. More general, partially coherent illumination can produce optically formed images that differ greatly from the intensity pattern leaving the object. Seen from this perspective, partially coherent illumination is an undesirable property that creates nonideal images and complicates the image analysis.

The ideal image as described above is not necessarily the optimum image for a particular task. Consider the case of object recognition when the object has low optical contrast. Using the above definition, an ideal image would mimic the low contrast and render recognition difficult. An image formation system that alters the contrast to improve the recognition task would be better than the so-called ideal image. The image formation system that maximized the appropriate performance metric for a particular recognition task would be considered optimal. In optimizing the indirect imaging system of Figure 2, the designer can adjust the illumination, the imaging optics, and the post-detection processing. In fact, research microscopes often include an adjustment that modifies illumination coherence and often alters the image contrast. Darkfield and phase contrast imaging microscopes usually employ partially coherent illumination combined with pupil modification to view otherwise invisible objects. Seen from this perspective, partially coherent illumination is a desirable property that provides more degrees of freedom to the imaging system designer. Quite often, partially coherent imaging systems provide a compromise between the performance of coherent and incoherent systems.

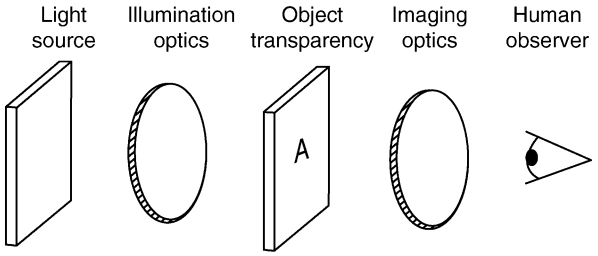


Figure 1 Direct view optical system. Imaging optics conveys the light from an illuminated object directly to the human viewer.

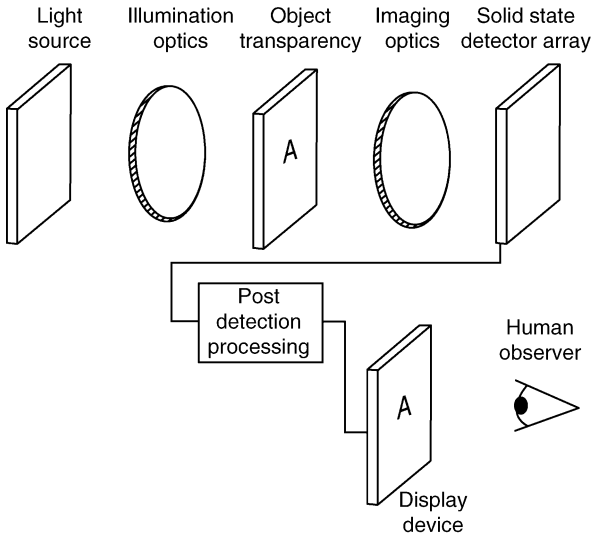


Figure 2 Indirect view optical system. The raw intensity optical image is detected electronically, processed and a final image is presented to the human observer on a display device.

Elementary Coherence Concepts

Most readers are somewhat familiar with the concept of temporal coherence. In Figure 3, a Michelson interferometer splits light from a point source into two paths and recombines the beams to form interference fringes. The presence of interference fringes indicates that the wave amplitude fluctuations of the two beams are highly correlated so the light adds in wave amplitude. If the optical path difference between the two paths can be made large without reducing the fringe contrast, the light is said to be

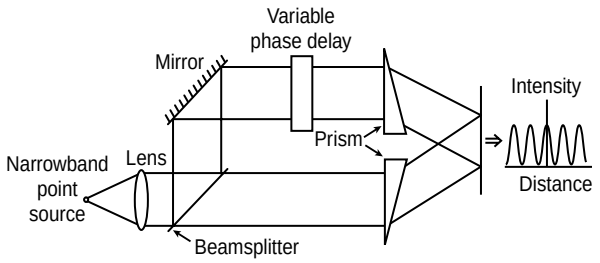


Figure 3 The presence of high contrast fringes in a Michelson interferometer indicates high temporal coherence.

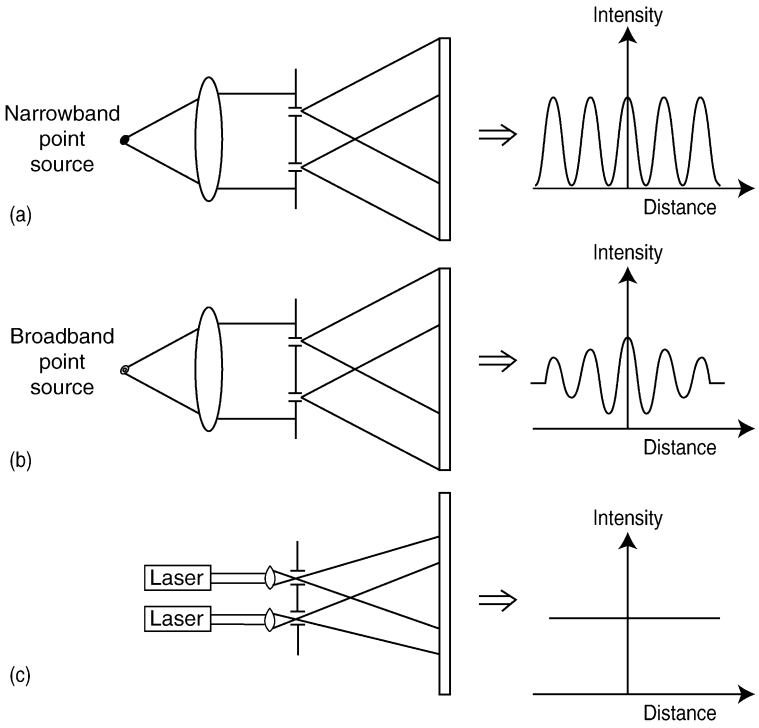


Figure 4 A Young's two pinhole interferometer produces: (a) a uniform high contrast fringe pattern for light that is highly spatially and temporally coherent; (b) high contrast fringes only near the axis for low temporal coherence and high spatial coherence light; and (c) no fringe pattern for high temporal coherence, low spatial coherence light.

highly temporally coherent. Light from very narrowband lasers can maintain coherence over very large optical path differences. Light from broadband light sources requires very small optical path differences to add in wave amplitude.

Spatial coherence is a measure of the ability of two separate points in a field to interfere. The Young's two pinhole experiment in Figure 4 measures the coherence between two points sampled by the pinhole mask. Only a one-dimensional pinhole mask is shown for simplicity. Figure 4a shows an expanded laser beam illuminating two spatially separated pinholes and recombining to form an intensity fringe pattern. The high contrast of the fringe indicates that the wave amplitude of the light from the two pinholes is highly correlated. Figure 4b shows a broadband point source expanded in a similar fashion. The fringe contrast is high near the axis because the optical path difference between the two beams is zero on the axis and relatively low near the axis. For points far from the axes, the fringe pattern disappears because the low temporal coherence from the broadband source results in a loss in correlation of the wave amplitudes.

A final Young's experiment example in Figure 4c shows that highly temporally coherent light can be spatially incoherent. In the figure, the light illuminating the two pinholes comes from two separate and highly temporally coherent lasers that are designed to have the same center frequency. Since the light from the lasers is not synchronized in phase, any fringes that might form for an instant will move rapidly and average to a uniform intensity pattern over the integration time of a typical detector. Since the fringe contrast is zero over a practical integration time, the light at the two pinholes is effectively spatially incoherent.

Two-point Imaging

In a typical partially coherent imaging experiment we need to know how light from two pinholes adds at the optical image plane, as shown in Figure 5. Diffraction and system aberrations cause the image of a single point to spread so that the images of two spatially separated object points overlap in the image plane. The wave amplitude image of a single pinhole is called the coherent point spread function (CPSF). Since the CPSF is compact, two CPSFs will only overlap when the corresponding object points are closely spaced. When the two points are sufficiently close, the relevant optical path differences will be small so full temporal coherence can be assumed. Spatial coherence will be the critical factor in determining how to add the responses to pairs of

point images. We assume full temporal coherence in the subsequent analyses and the light is said to be *quasimonochromatic*.

Coherent Two-Point Imaging

Consider imaging a two-point object illuminated by a spatially coherent plane wave produced from a point source, as depicted in Figure 6a. In the figure, the imaging system is assumed to be space-invariant. This means that the spatial distribution of the CPSF is the same regardless of position of the input pinhole. The CPSF in the figure is broad enough such that the image plane point responses overlap for this particular pinhole spacing. Since the wave amplitudes from the two pinholes are correlated, the point responses add in wave amplitude, resulting in a two-point image intensity given by

$$I_{2\text{coh}}(x) = I_0|h(x - x_1)|^2 + I_0|h(x - x_2)|^2 \qquad [1]$$

where $h(x)$ is the normalized CPSF and I_0 is a scaling factor that determines the absolute image intensity value. Since the CPSF has units of wave amplitude, the magnitude squaring operation accounts for the square law response of the image plane detector.

More generally, the image intensity for an arbitrary object distribution can be found by breaking the object into a number of points and adding the CPSFs due to each point on the object. The resultant image plane intensity is the spatial convolution of the object amplitude transmittance with the CPSF and is given by

$$I_{\text{coh}}(x) = I_0\left|\int O(\xi)h(\xi - x)\mathrm{d}\xi\right|^2 \qquad [2]$$

where $O(x)$ is the object amplitude transmittance.

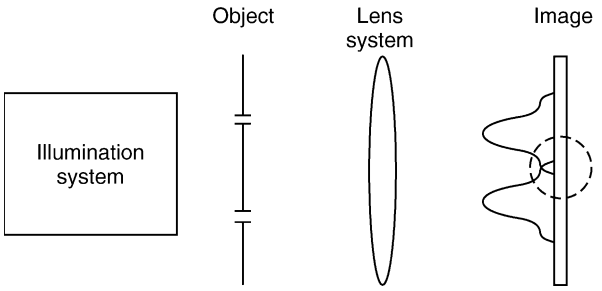


Figure 5 Imaging of two points. The object plane illumination coherence determines how the light adds in the region of overlap of the two image plane points.

Incoherent Two-Point Imaging

Two pinhole imaging with spatially incoherent light is shown in Figure 6b, where a spatially extended blackbody radiator is placed directly behind the pinholes. Once again the imaging optics are assumed to be space-invariant and have the same CPSF as the system in Figure 6a. Since the radiation is originating from two completely different physical points on the source, no correlation is expected between the wave amplitude of the light leaving the pinholes and the light is said to be spatially incoherent. The resultant image intensity corresponding to the two pinhole object with equal amplitude transmittance values is calculated by adding the individual intensity responses to give

$$I_{2\text{inc}}(x) = I_0|h(x - x_1)|^2 + I_0|h(x - x_2)|^2 \qquad [3]$$

For more general object distributions, the image intensity is the convolution of the object intensity transmittance with the incoherent PSF, and is given by

$$I_{\text{inc}}(x) = I_0\int |o(\xi)|^2|h(x - \xi)|^2\mathrm{d}\xi \qquad [4]$$

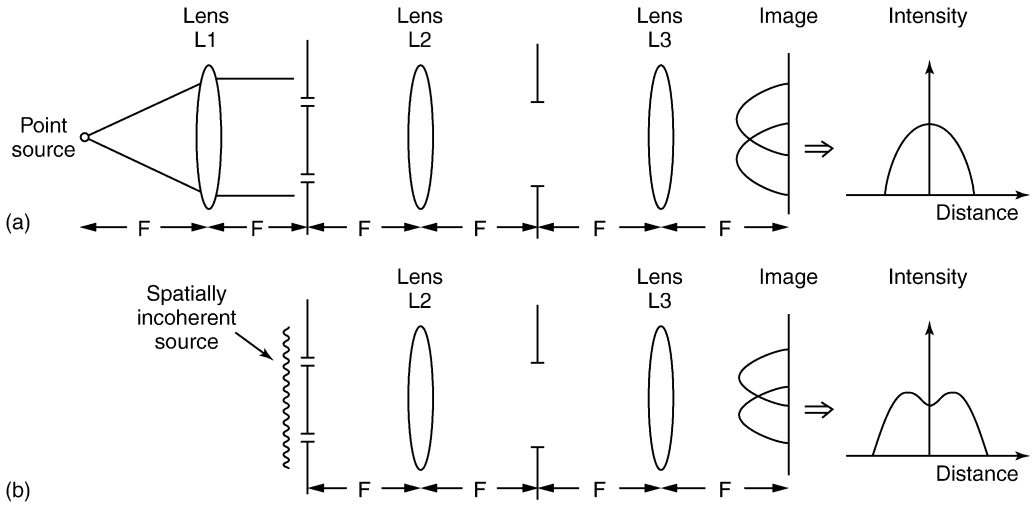


Figure 6 Imaging of two points for (a) spatially coherent illumination and (b) spatially incoherent imaging.

where the intensity transmittance is the squared magnitude of the amplitude transmittance and the incoherent PSF is the squared magnitude of the CPSF.

Partially Coherent Two-Point Imaging

Finally we consider two pinhole imaging illuminated by partially spatially coherent light. Now the two-point responses do not add simply in amplitude or intensity. Rather, the image intensity is given by the more general equation:

$$I_{2pc}(x) = I_0[|h(x-x_1)^2 + |h(x-x_2)|^2 + 2\mathrm{Re}\{\mu(x_1;x_2)h(x-x_1)h^*(x-x_2)\}] \quad [5]$$

where the $*$ denotes complex conjugations and $\mu(x_1;x_2)$ is the normalized form of the mutual intensity of the object illumination evaluated at the object points in question. The mutual intensity function is often denoted as $J_0(x_1;x_2)$ and is a measure of the cross correlation of the wave amplitude distributions leaving the two pinholes. Rather than providing a rigorous definition, we note that the magnitude of $J_0(x_1;x_2)$ corresponds to the fringe contrast that would be produced if the two object pinholes were placed at the input of a Young's interference experiment. The phase is related to the relative spatial shift of the fringe pattern. When the light is uncorrelated, $J_0(x_1;x_2) = 0$ and eqn [5] collapses to the incoherent limit of eqn [4]. When the light is coherent, $J_0(x_1;x_2) = 1$ and eqn [5] reduces to the coherent form of eqn [2].

Image formation for a general object distribution is given by the bilinear equation:

$$I_{pc}(x) = I_0 \int o(\xi_1)o^*(\xi_2)h(x-\xi_1)h^*(x-\xi_2) \times J_o(\xi_1-\xi_2)\mathrm{d}\xi_1\,\mathrm{d}\xi_2 \quad [6]$$

Note that, in general, $J_0(x_1;x_2)$ must be evaluated for all pairs of object points. Close examination of this equation reveals that just as a linear system can be evaluated by considering all possible single point responses, a bilinear system requires consideration of all possible pairs of points. This behavior is much more complicated and does not allow the application of the well-developed linear system theory.

Source Distribution and Object Illumination Coherence

According to eqn [6], the mutual intensity of all pairs of points at the object plane must be known, to calculate a partially coherent image. Consider the telecentric Kohler illumination imaging system shown in Figure 7. The primary source is considered to be spatially incoherent and illuminates the object after passing through lens L1 located one focal distance away from the source and one focal distance away from the object. Even though the primary source is spatially incoherent, the illumination at the object plane is partially coherent. The explicit mutual intensity function corresponding to the object plane illumination is given by applying the van Cittert Zernike theorem:

$$J_0(\Delta x) = \int S(\xi)\exp[\mathrm{j}2\pi\Delta x\xi/\lambda F]\mathrm{d}\xi \quad [7]$$

where $S(x)$ is the intensity distribution of the spatially incoherent primary source, F is the focal length of the lens, λ is the illumination wavelength, and $\Delta x = x_1 - x_2$. The van Cittert Zernike theorem reflects a Fourier transform relationship between the source image intensity distribution and the mutual intensity at the object plane. When the source plane is effectively spatially incoherent, the object plane mutual intensity is only a function of separation distance. For a two-dimensional object, the mutual intensity needs to be characterized for all pairs of unique spacings in x and y. In Figure 7, the lens arrangement ensures that the object plane is located in the far field of the primary source plane. In fact, the van Cittert Zernike theorem applies more generally, even in the Fresnel propagation regime, as long as the standard paraxial approximation for optics is valid.

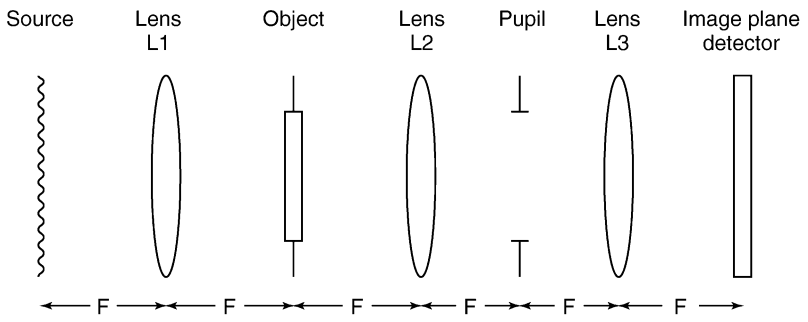


Figure 7 A telecentric Kohler illumination system with a spatially incoherent primary source imaged onto the pupil.

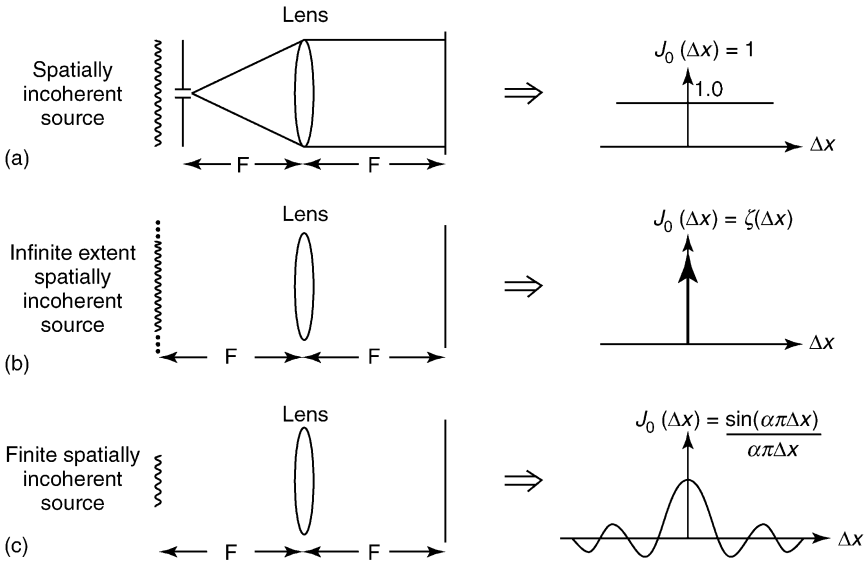


Figure 8 Light propages from the source plane to the object plane and produces (a) coherent illumination from a single source point; (b) incoherent illumination from an infinite extent source; and (c) partially coherent illumination from a finite extent primary source.

Figure 8 shows some simple examples of primary source distributions and the corresponding object plane mutual intensity function. Figure 8a assumes an infinitesimal point source and the corresponding mutual intensity is 1.0 for all possible pairs of object points. Figure 8b assumes an infinite extent primary source and results in a dirac delta function for the mutual intensity function. This means that there is no correlation between any two object points having a separation greater than zero so the object plane illumination is spatially incoherent. In Figure 5c, a finite extent uniform source gives a mutual intensity function of the form $\sin(\alpha\pi\Delta x)/(\alpha\Delta x)$. The finite-sized source corresponds to partially coherent imaging and shows that the response to pairs of points is affected by the spatial distribution of the source in a complicated way. Note that a large primary source corresponds to a large range of angular illumination at the object plane.

Varying the size of the source in the imaging system of Figure 8 will affect the spatial coherence of the illumination and hence the optical image intensity. Many textbook treatments discuss the imaging of two points separated by the Rayleigh resolution criterion which corresponds to the case where the first minimum of one point image coincides with the maximum of the adjacent point image. With a large source that provides effectively incoherent light, the two point image has a modest dip in intensity between the two points, as shown in Figure 9a. Fully coherent illumination of two points separated by the Rayleigh resolution produces a single large spot with no dip in intensity, as shown in Figure 9b. Varying the source size and hence the illumination spatial coherence produces a dip that is less than the incoherent intensity dip. This result is often used to suggest that coherent imaging gives poorer resolution than incoherent imaging. In fact, generalizations about two-point resolution can be misleading.

Recent developments in optical lithography have shown that coherence can be used to effectively increase two-point resolution beyond traditional diffraction limits. In Figure 9c, one of the pinholes in a coherent two-point imaging experiment has been modified with a phase shift corresponding to one half of a wavelength of the illuminating light. The two images add in wave amplitude and the phase shift creates a distinct null at the image plane and effectively enhances the two-point resolution. This approach is termed phase screen lithography and has been exploited to produce finer features in lithography by purposely introducing small phase shift masks at the object mask. In practice, the temporal and spatial coherence of the light is engineered to give sufficient coherence to take advantage of the two-point enhancement while maintaining sufficient incoherence to avoid speckle-like artifacts associated with coherent light.

Spatial Frequency Modeling of Imaging

Spatial frequency models of image formation are also useful in understanding how coherence affects image formation. A spatially coherent imaging system has a particularly simple spatial frequency model. In the

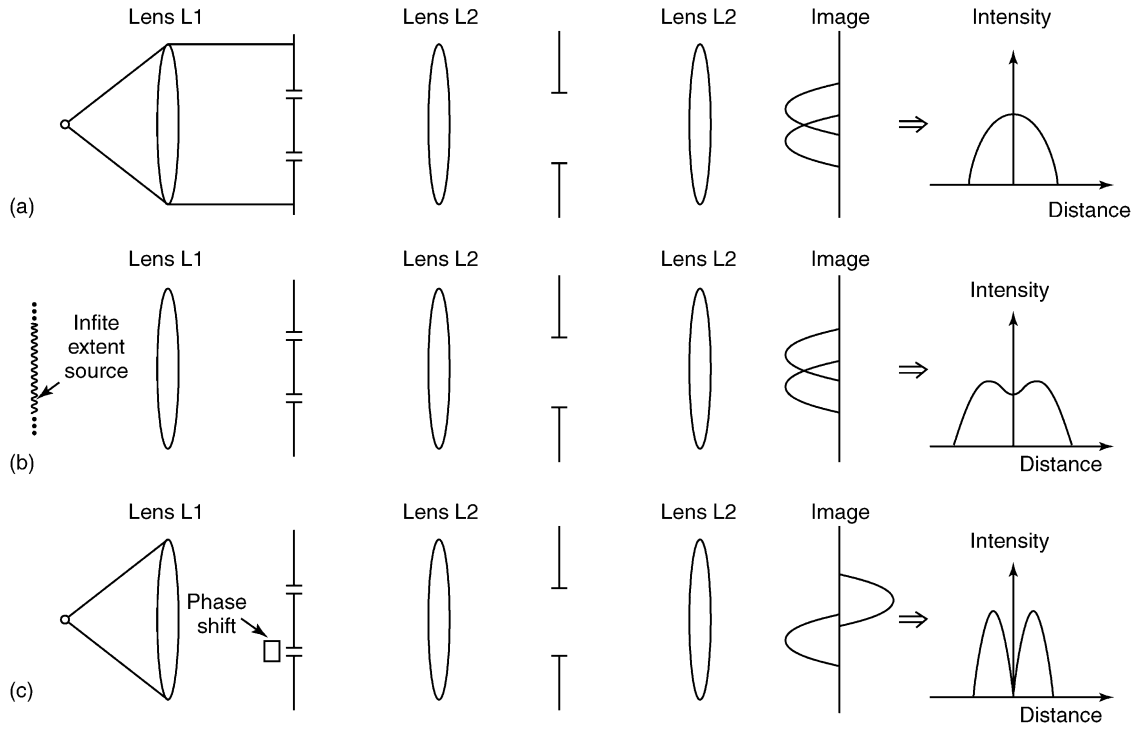


Figure 9 Comparison of coherent and two point imaging of two pinholes separated by the Rayleigh distance. (a) Coherent image cannot resolve the two points. (b) Incoherent image barely resolves the two points. (c) Coherent illumination with a phase shifting plate at one pinhole produces a null between the two image points.

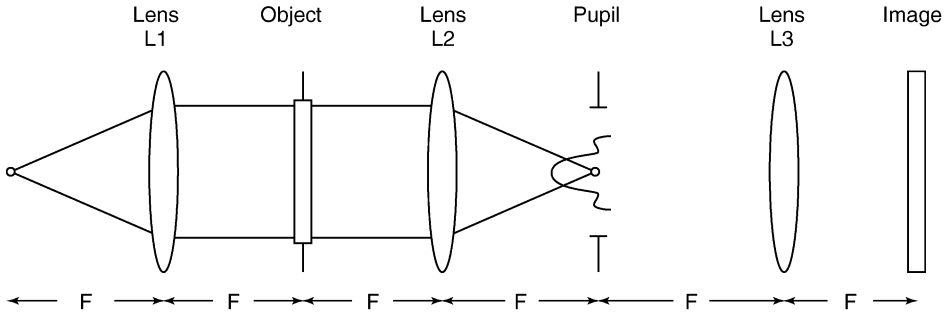


Figure 10 The object Fourier transform is filtered by the pupil function in a coherent imaging system.

spatially coherent imaging system shown in Figure 10, the on-axis plane wave illumination projects the spatial Fourier transform of the object, or Fraunhofer pattern, onto the pupil plane of the imaging system. The pupil acts as a frequency domain filter that can be modified to perform spatial frequency filtering. The complex transmittance of the pupil is the wave amplitude spatial frequency transfer function of the imaging system. It follows that the image plane wave amplitude frequency spectrum, $\tilde{U}_c(f_x)$, is given by

$$\tilde{U}_c(f_x) = \tilde{O}(f_x)H(f_x) \qquad [8]$$

where $\tilde{O}(f_x)$ is the spatial Fourier transform of the object amplitude transmittance function and $H(f_x)$ is proportional to the pupil plane amplitude transmittance. Equation [8] is the frequency domain version of the convolution representation given by eqn [2], but does not account for the squared magnitude response of the detector.

In the previous section, we learned that an infinite extent source is necessary to achieve fully incoherent imaging for the imaging system of Figure 7. As a thought experiment, one can start with a single point source on the axis and keep adding mutually incoherent source points to build up from a coherent imaging system to an incoherent imaging system. Since the individual source points are assumed to be incoherent with each other, the images from each point source can be added in intensity. In Figure 11,

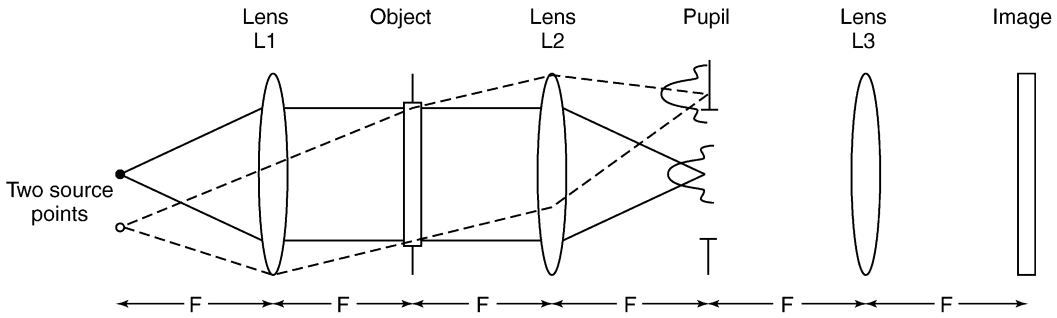


Figure 11 A second source point projects a second object Fourier transform onto the pupil plane. The images produced by the two source points add in intensity at the image plane.

only two point sources are shown. Each point source projects an object Fraunhofer pattern onto the pupil plane. The centered point source will result in the familiar coherent image. The off-axis point source projects a displaced object Fraunhofer pattern that is also filtered by the pupil plane before forming an image of the object. The image formed from this second point source can be modeled with the same coherent imaging model with a shifted pupil plane filter. Since the light from the two source points is uncorrelated, the final image is calculated by adding the intensities of the two coherently formed images.

The two source point model of Figure 11 can be generalized to an arbitrary number of source points. The final image is an intensity superposition of a number of coherent images. This suggests that partially coherent imaging systems behave as a number of redundant coherent imaging systems, each having a slightly different amplitude spatial frequency transfer function due to the relative shift of the pupil filter with respect to the object Fraunhofer pattern. As the number of point sources is increased to infinity, the primary source approaches an infinite extent spatially incoherent source. In practice, the source need not be infinite. When the source is large enough to effectively produce linear-in-intensity imaging, the imaging system is effectively spatially incoherent. The corresponding image intensity in the spatial frequency domain, $\tilde{I}_{\text{inc}}(f_x)$, is given by

$$\tilde{I}_{\text{inc}}(f_x) = \tilde{I}_{\text{obj}}(f_x)\text{OTF}(f_x) \qquad [9]$$

where $\tilde{I}_{\text{obj}}(f_x)$ is the spatial Fourier transform of the object intensity transmittance and $\text{OTF}(f_x)$ is the incoherent optical transfer function which is proportional to the spatial autocorrelation of the pupil function.

Many texts include detailed discussions comparing the coherent transfer function and the OTF and attempt to make comparisons about the relative performance of coherent and incoherent systems. These comparisons are often misleading since one transfer function describes the wave amplitude spatial frequency transfer function and the other describes the intensity spatial frequency transfer function. Here we note that incoherent imaging systems do indeed allow higher object amplitude spatial frequencies to participate in the image formation process. This argument is often used to support the claim that incoherent systems have higher resolution. However, both systems have the same image intensity spatial frequency cutoff. Furthermore, the nature of the coherent transfer function tends to produce high contrast images that are typically interpreted as higher-resolution images than their incoherent counterparts. Perhaps the real conclusion is that the term resolution is not well defined and direct comparisons between coherent and incoherent imaging must be treated carefully.

The frequency domain treatment for partially coherent imaging of two-dimensional objects involves a four-dimensional spatial frequency transfer function that is sometimes called the bilinear transfer function or the transmission cross-coefficient model. This model describes how constituent object wave amplitude spatial frequencies interact to form image plane intensity frequencies. The utility of this approach to analysis is limited for someone new to the field, but is often used in numerical simulations of partially coherent imaging systems used in optical lithography.

Experimental Examples of Important Coherence Imaging Phenomena

Perhaps the best way to gain an understanding of coherence phenomena in imaging is to examine experimental results. In the following section we use experimental data to see how coherence affects edge response, noise immunity, and depth of field. Several experimental configurations were used to collect

the image data, but all of the systems can be represented generically by the Kohler illumination system shown in Figure 12. Kohler illumination is often employed in many projection illumination systems. In order to obtain Kohler illumination, the primary spatially incoherent source is imaged onto the pupil plane of the imaging portion of the system and the object is placed at the pupil plane of the condenser optics.

Primary Source Generation

Figure 13 shows a highly coherent illumination system produced by focusing a laser beam to a point and imaging the focused spot onto the pupil of the imaging system to produce spatially coherent illumination. The use of a laser produces highly temporally coherent light.

Two methods were used to create an extended spatially incoherent primary source with control over the spatial intensity distribution. Figure 14a shows a collimated laser beam with a 633 nm center wavelength illuminating a rotating diffuser. A photographically produced mask defines the spatial shape of the primary source. The laser provides highly temporally coherent light and the diffuser destroys the spatial coherence of the light. Consider the thought experiment of two pinholes placed immediately to the right of the moving diffuser. Without the diffuser, the two wave amplitudes would be highly correlated. Assuming that the diffuser can be modeled as a spatially random phase plate, a fixed diffuser would only introduce a fixed phase difference between the amplitudes leaving the two pinholes and would not destroy the coherence. When the diffuser is rotated, the light from each pinhole encounters a different phase modulation that is changing over time. This random modulation destroys the effective correlation between the wave amplitude of the light leaving the two pinholes provided that the rotation speed is sufficiently fast.

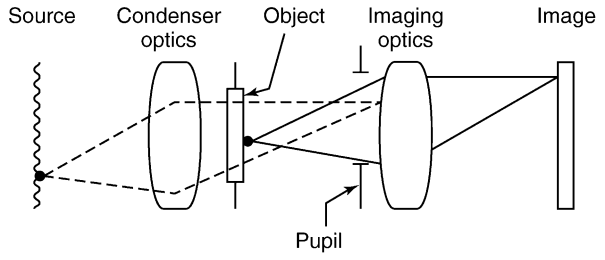


Figure 12 General representation of the Kohler illumination imaging systems used in the experimental result section. The primary incoherent source is imaged onto the imaging system pupil plane and the object resides in the pupil of the condenser optics.

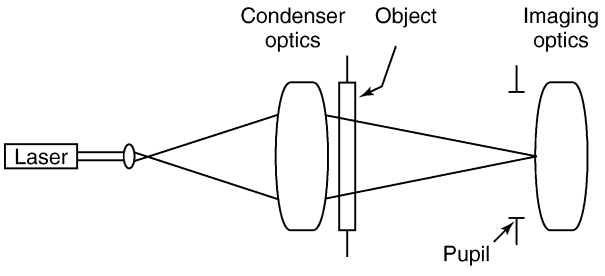


Figure 13 Highly spatially and temporally coherent illumination produced by imaging a focused laser beam into the imaging system pupil.

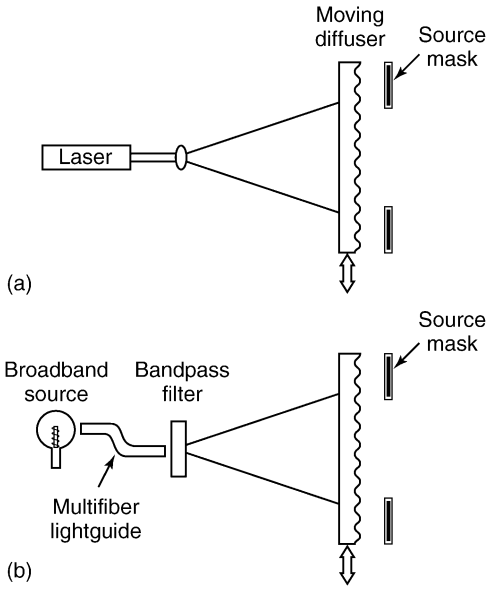


Figure 14 Two methods for generating a spatially incoherent primary source. (a) An expanded laser beam passes through a moving diffuser followed by a mask to define the extent and shape of the source. (b) A broadband source exits a multifiber lightguide and propagates to a moving diffuser followed by a source mask.

The moving diffuser method is light inefficient but is a practical way of exploring coherence in imaging in a laboratory environment. The choice of the diffuser is critical. The diffuser should spread light out uniformly over an angular subtense that overfills the object of interest. Many commercially available diffusers tend to pass too much light in the straight through direction. Engineered diffusers can be purchased to produce an optimally diffused beam. When quick and inexpensive results are required, thin plastic sheets used in day-to-day packaging often serve as excellent diffusers. When the diffusion angles are not high enough, a number of these plastic sheets can be layered on top of each other to achieve the appropriate angular spread.

The second method for producing a spatially incoherent source is shown in Figure 14b. Broadband

light is delivered through a large multifiber lightguide and illuminates a moving diffuser. The main purpose of the diffuser is to ensure that the entire object is illuminated uniformly. The low temporal coherence of the source and the long propagation distances are usually enough to destroy any spatial coherence at the plane of the diffuser. The rotation further ensures that no residual spatial coherence exists at the primary source plane. A chromatic filter centered around 600 nm, with a spectral width of approximately 100 nm, is shown in the light path. The filter was used to minimize the effects of chromatic aberrations in the system. The wide spectral width certainly qualifies as a broadband source in relation to a laser. High-contrast photographically defined apertures determine the spatial intensity distribution at the plane of the primary source. In the following results, the primary source distributions were restricted to circular sources of varying sizes as well as annular sources. These shapes are representative of most of the sources employed in microscopy, machine vision, and optical lithography.

Noise Immunity

Coherent imaging systems are notorious for introducing speckle-like noise artifacts at the image. Dust and optically rough surfaces within an optical system result in a complicated textured image plane intensity distribution that is often called speckle noise. We refer to such effects as coherent artifact noise. **Figure 15a** shows the image of a standard binary target as imaged by a benchtop imaging system, with the highly coherent illumination method shown in **Figure 13**. Some care was taken to clean individual lenses and optical surfaces within the system and no dust was purposely introduced. The image is corrupted by a complicated texture that is due in

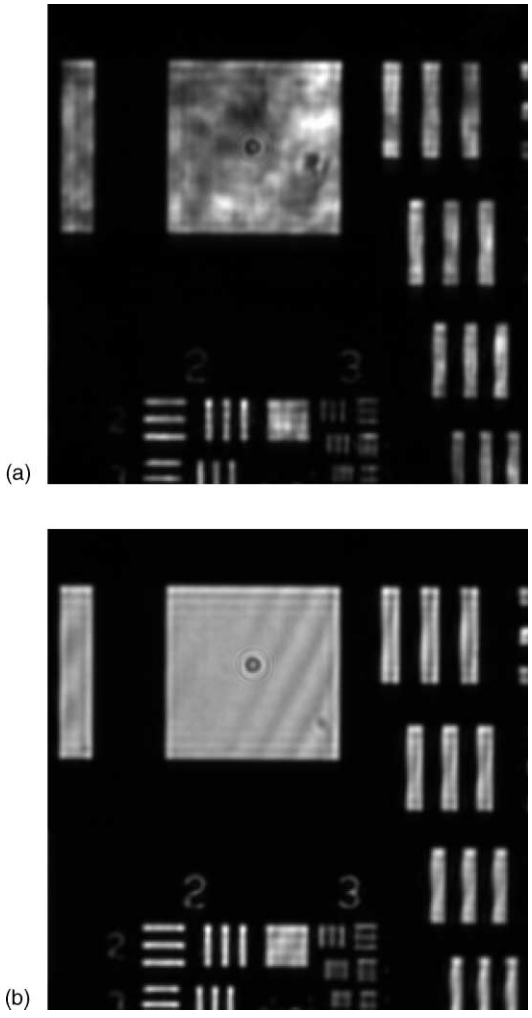


Figure 15 High temporal coherence imaging with disk sources corresponding to (a) extremely high spatial coherence ($K = 0.05$) and (b) slightly reduced but high spatial coherence ($K = 0.1$).

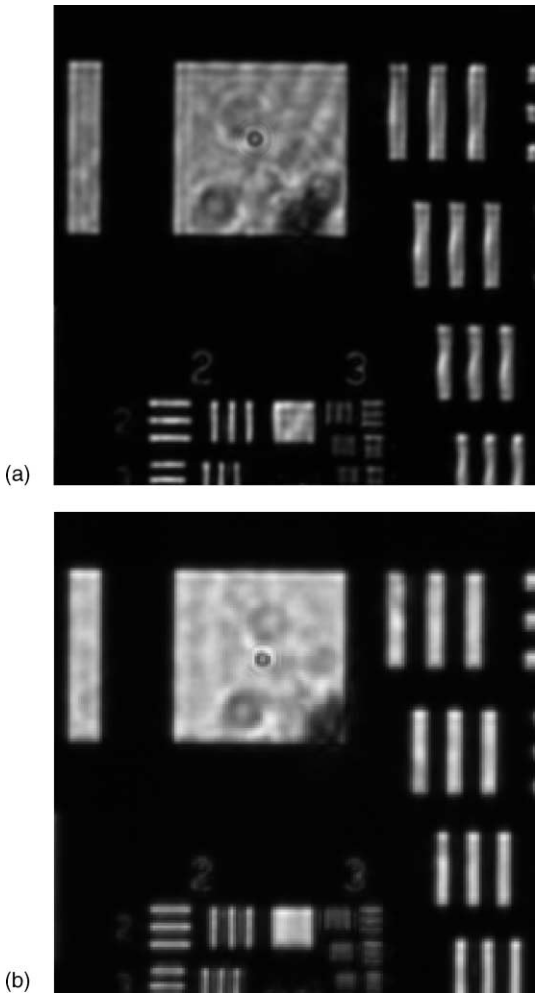


Figure 16 High spatial coherence disk illumination ($K = 0.1$) imaging through a dust covered surface with (a) narrowband laser illumination and (b) with broadband illumination.

part to imperfections in the laser beam itself and in part to unwanted dust and optically rough surfaces. Certainly, better imaging performance can be obtained with more attention to surface cleanliness and laser beam filtering, but the result shows that it can be difficult to produce high-quality coherent images in a laboratory setting.

Lower temporal coherence and lower spatial coherence will reduce the effect of speckle noise. The image of Figure 15b was obtained with laser illumination according to Figure 14a with a source mask corresponding to a source to pupil diameter ratio of $K = 0.1$. The object plane illumination is still highly temporally coherent and the spatial coherence has been reduced but is still very high. The artifact noise has been nearly eliminated by the modest reduction of spatial coherence. The presence of diagonal fringes in some regions is a result of multiple reflections produced by the cover glass in front of the CCD detector. The reflections produce a weak secondary image that is slightly displaced from the main image and the high temporal coherence allows the two images to interfere.

The noise performance was intentionally perturbed in the image of Figure 16a by inserting a slide with a modest sprinkling of dust at an optical surface in between the object and the pupil plane. The illumination conditions are the same as for the image of Figure 15b. The introduction of the dust has further degraded the image. The image of Figure 16b maintains the same spatial coherence ($K = 0.1$) but employs the broadband source. The lower temporal coherence eliminates the unwanted diagonal fringes but the speckle noise produced by the dust pattern does not improve significantly relative to the laser illumination $K = 0.1$ system. Figures 17a–c show that increasing the source size (and hence the range of angular illumination) reduces

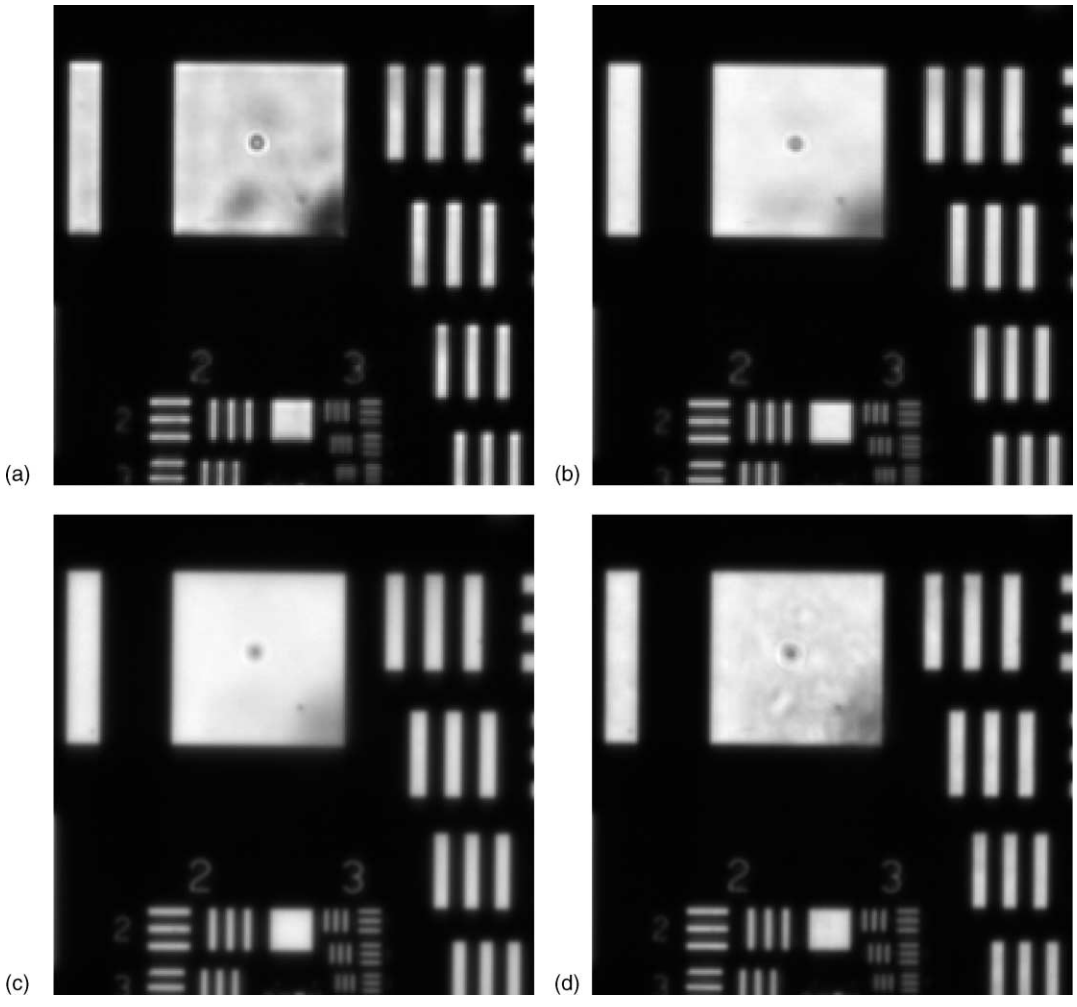


Figure 17 Low temporal coherence imaging through a dust-covered surface with various source sizes and shapes: (a) disk source with $K = 0.3$; (b) disk source with $K = 0.7$; (c) disk source with $K = 2.0$; and (d) a thin annular source with an outer diameter corresponding to $K = 0.5$ and inner diameter corresponding to $K = 0.45$.

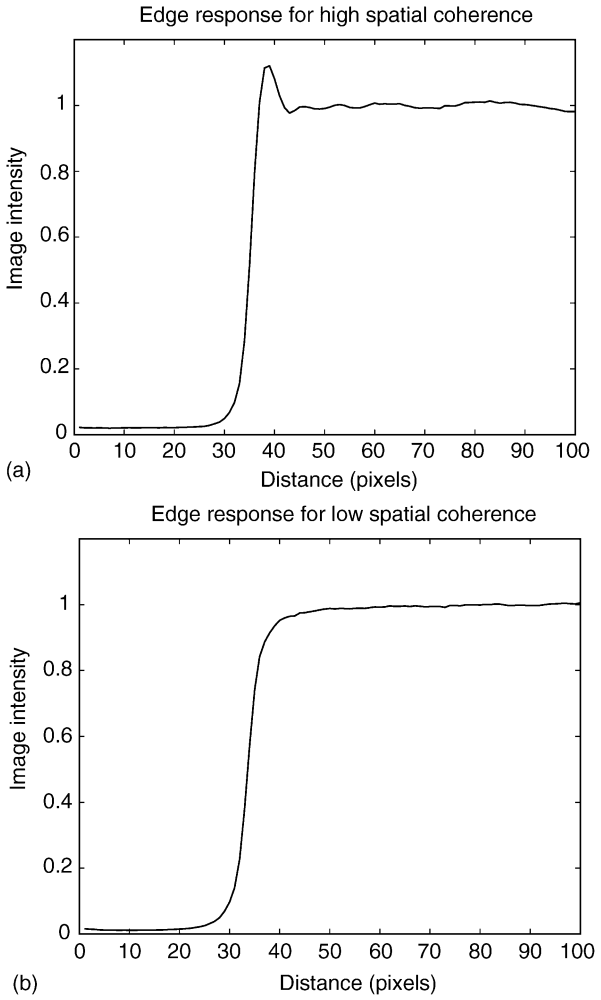


Figure 18 Slices of edge intensity images of high contrast edges taken by a laboratory system with low temporal coherence and (a) high spatial coherence (disk source with $K = 0.1$) and (b) effectively incoherent (disk source with $K = 2$).

the spatial coherence and virtually eliminates the unwanted speckle pattern. Finally, Figure 17d shows that an annular source can also provide some noise immunity. The amount of noise immunity is related to the total area of the source rather than the outer diameter of annular source. Seen from the spatial frequency model, each source point produces a noise-corrupted coherent image. Since the effect of the noise is different for each coherent image, the image plane noise averages out as the images add incoherently. This perspective suggests that an extended source provides redundancy in the transfer of object information to the image plane.

Edge Response

Spatial coherence has a strong impact on the images of edges. The sharp spatial frequency cutoff of spatially coherent imaging systems creates oscillations at the images of an edge. Figure 18 shows slices of experimentally gathered edge imagery as a function of the ratio of source diameter to pupil diameter. Higher coherence systems produce sharper edges, but tend to have overshoots. The sharper edges contribute to the sense that high-coherence systems produce higher-resolution images. One advantage to an incoherent imaging system is that the exact location of the edge corresponds to the 50% intensity location. The exact location of an edge in a partially coherent imaging system is not as easily determined. The presence or lack of edge ringing can be used to assess whether a given imaging system can be modeled as spatially incoherent or partially coherent.

Depth of Field

Coherent imaging systems exhibit an apparent increase in depth-of-field compared to spatially incoherent systems. Figure 19 shows spatially incoherent imagery for four different focus positions. Figure 20 shows imagery with the same amount of defocus produced with highly spatially coherent light. Finally, Figure 21 gives an example of how defocused imagery depends on the illumination coherence. The images of a spoke target were all gathered with a fixed amount of defocus and the source size was varied to control the illumination coherence. The images of Figure 21 differ greatly, even though the CPSF was the same for all the cases.

Digital Post-detection Processing and Partial Coherence

The model of Figure 2 suggests that post-detection image processing can be considered as part of the image formation system. Such a general view can result in imaging that might be otherwise unobtainable by classical means. In fact, microscopists routinely use complicated deblurring methods to reconstruct out-of-focus imagery and build up three-dimensional images from multiple image slices.

The coherence of the illumination should be considered when undertaking such image restoration. Rigorous restoration of partially coherent imagery is computationally intensive and requires precise knowledge of the coherence. In practice, a linear-in-intensity model for the image formation is almost always used in developing image restoration algorithms. Even nonlinear restoration algorithms have built-in linear assumptions about the image formation models which imply spatially incoherent illumination.

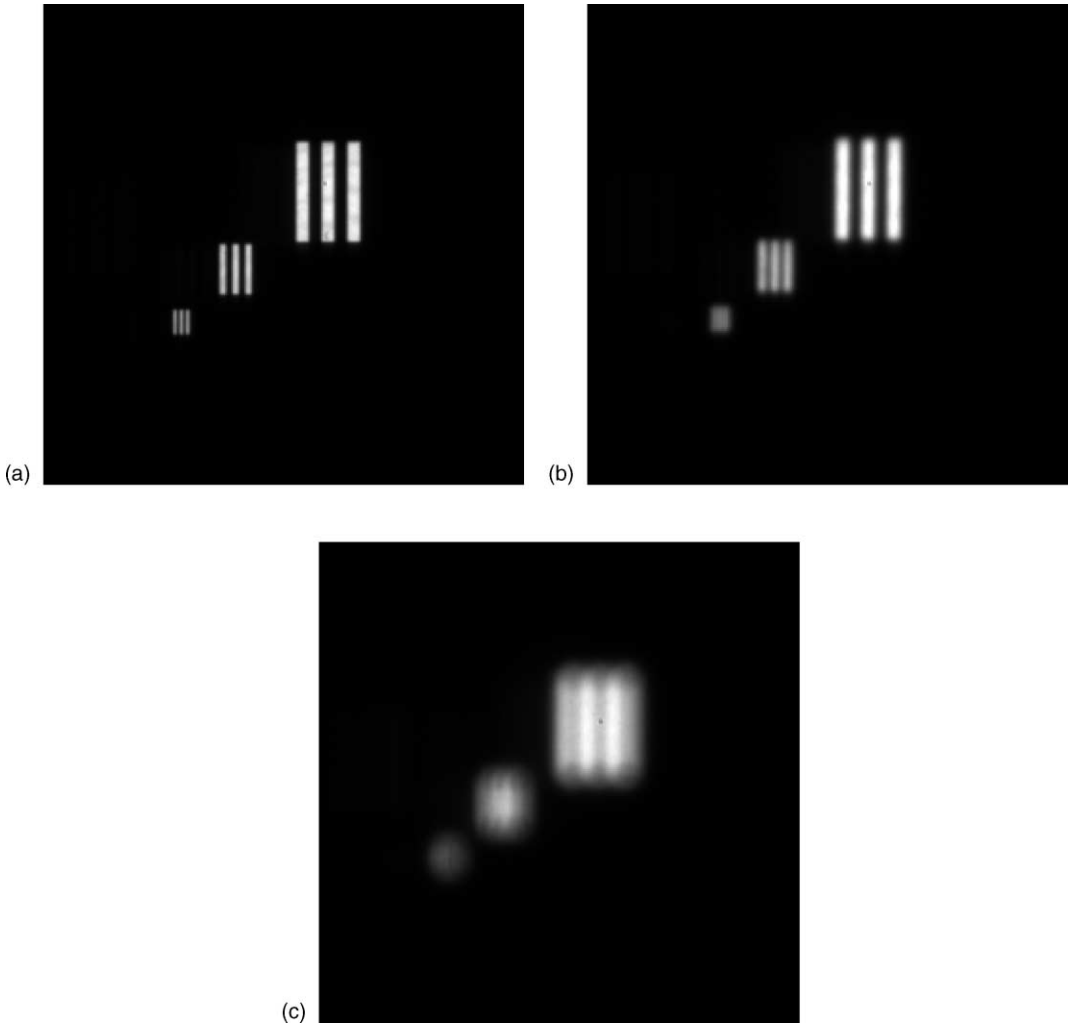


Figure 19 Spatially incoherent imaging for (a) best focus, (b) moderate misfocus, and (c) high misfocus.

The images of Figures 22a and b are digitally restored versions of the images of Figure 19b and Figure 20b. The point spread function was directly measured and used to create a linear restoration filter that presumed spatially incoherent illumination. The restored image in Figure 22a is faithful since the assumption of spatially incoherent light was reasonable. The restored image of Figure 22b suffers from a loss in fidelity since the actual illumination coherence was relatively high. The image is visually pleasing and correctly conveys the presence of three bar targets. However, the width of the bars is not faithful to the original target which had spacings equal to the widths of the bars.

As discussed earlier, the ideal image is not always the optimal image for a given task. In general, restoration of partially coherent imaging with an implicit spatially incoherent imaging model will produce visually pleasing images that are not necessarily faithful to the object plane intensity. While they are not faithful, they often preserve and even enhance edge information and the overall image may appear sharper than the incoherent image. When end task performance is improved, the image may be considered to be more optimal than the ideal image. It is important to keep in mind that some of the spatial information may be misleading and a more complete understanding of the coherence may be needed for precise image plane measurements. This warning is relevant in microscopy where the user is often encouraged to increase image contrast by reducing illumination spatial coherence. Annular sources can produce highly complicated spatial coherence functions that will strongly impact the restoration of such images.

Summary and Discussion

The coherence of the illumination at the object plane is important in understanding image formation.

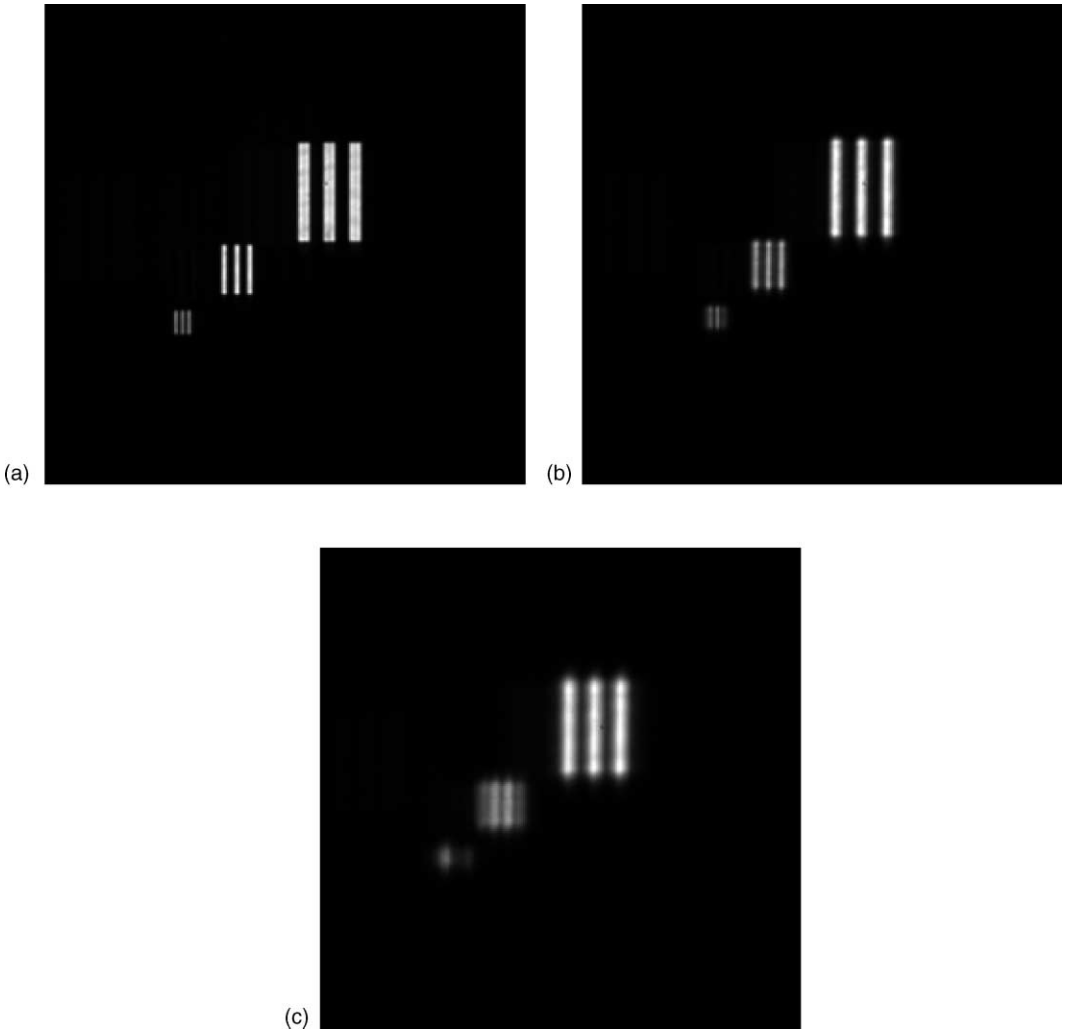


Figure 20 High spatial coherence disk source ($K = 0.2$) imaging for (a) best focus, (b) moderate misfocus, and (c) high misfocus.

Highly coherent imaging systems produce high-contrast images with high depth of field and provide the opportunity for sophisticated manipulation of the image with frequency plane filtering. Darkfield imaging and phase contrast imaging are examples of frequency plane filtering. Unfortunately, coherent systems are sensitive to optical noise and are generally avoided in practical system design. Relaxing the temporal coherence of the illumination can provide some improvement, but reducing the spatial coherence is more powerful in combating noise artifacts. Research microscopes, machine vision systems, and optical lithography systems are the most prominent examples of partially coherent imaging systems. These systems typically employ adjustable spatially incoherent extended sources in the shapes of disks or annuli. The exact shape and size of the primary source, shapes the angular extent of the illumination at the object plane and determines the spatial coherence at the object plane. Spatial coherence effects can be significant, even for broadband light. Control over the object plane spatial coherence allows the designer to find a tradeoff between the various strengths and weaknesses of coherent and incoherent imaging systems.

As more imaging systems employ post-detection processing, there is an opportunity to design fundamentally different systems that effectively split the image formation process into a physical portion and a post-detection portion. The simple example of image deblurring cited in this article shows that object plane coherence can affect the nature of the digitally restored image. The final image can be best understood when the illumination coherence effects are well understood. The spatially incoherent case results in the most straightforward model for image interpretation, but is not always the best choice since the coherence can often be manipulated to increase the contrast and hence the amount of useful information in the raw image. A completely general approach to

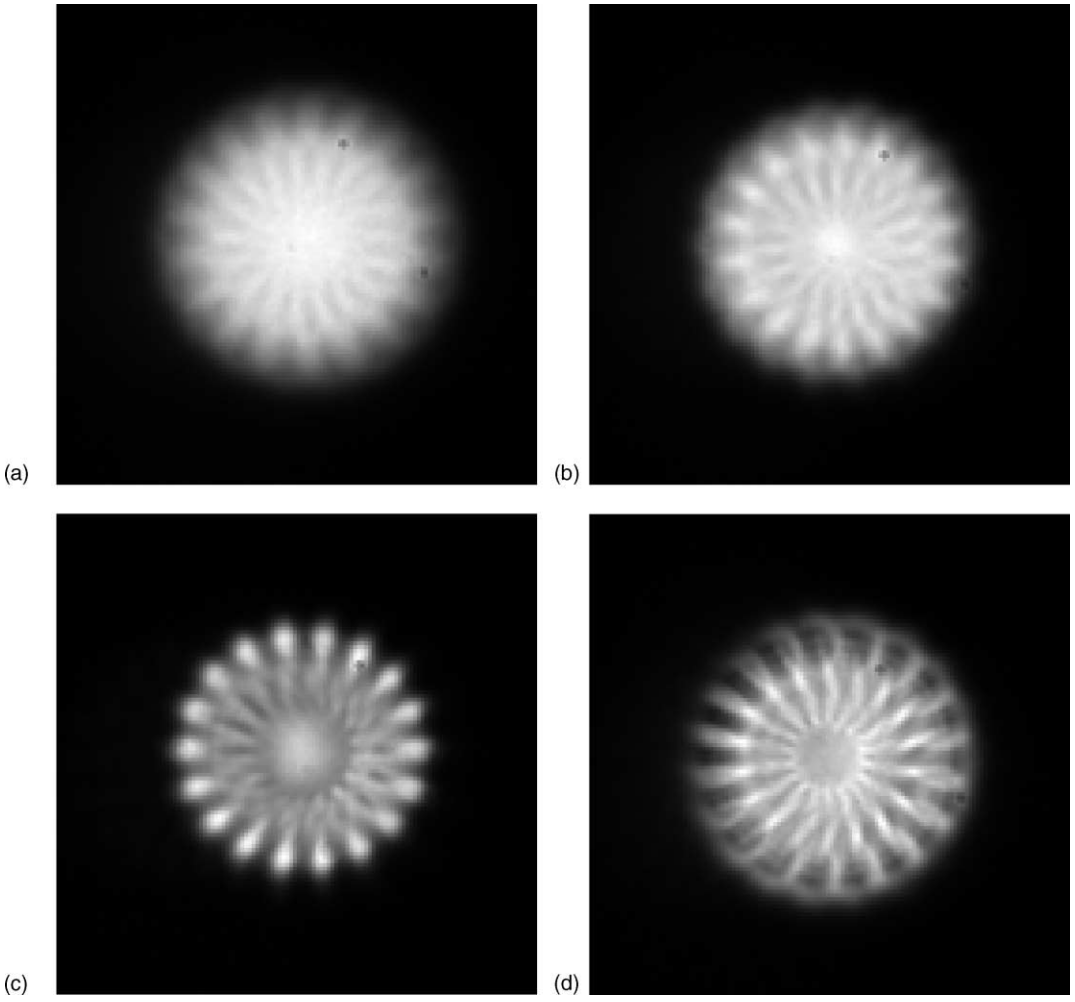


Figure 21 Imaging with a fixed amount of misfocus and varying object spatial coherence produced by: (a) highly incoherent disk source illumination ($K = 2$); (b) moderate coherence disk source ($K = 0.5$); (c) high coherence disk source ($K = 0.2$); and (d) annular source with outer diameter corresponding to $K = 0.5$ and inner diameter corresponding to $K = 0.45$.

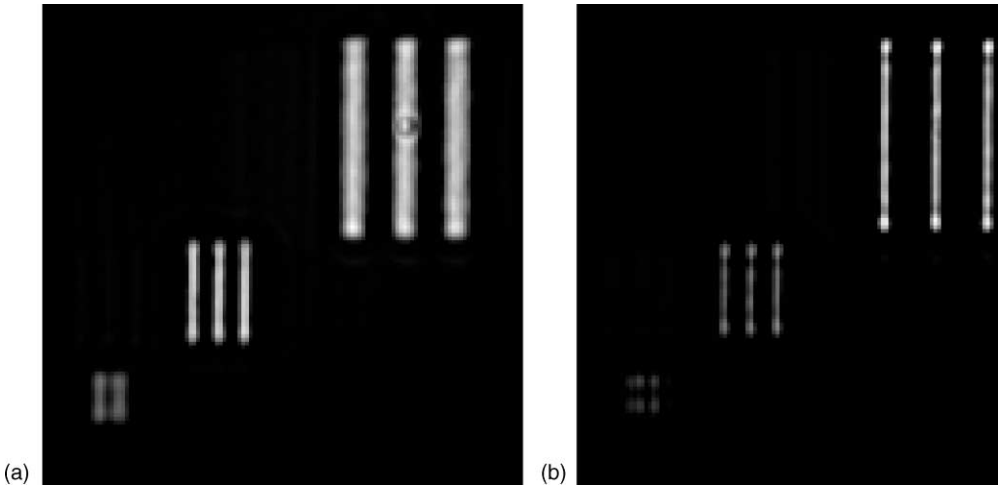


Figure 22 Digital restoration of blurred imagery with an inherent assumption of linear-in-intensity imaging for (a) highly spatially incoherent imaging ($K = 2$) and (b) high spatial coherence ($K = 0.2$).

imaging would treat the coherence of the source, the imaging optics, and the post-detection restoration all as free variables that can be manipulated to produce the optimal imaging system for a given task.

See also

Coherent Lightwave Systems. **Information Processing:** Coherent Analogue Optical Processors. **Terahertz Technology:** Coherent Terahertz Sources.

Further Reading

Bartelt H, Case SK and Hauck R (1982) Incoherent Optical Processing. In: Stark H (ed.) *Applications of Optical Fourier Transforms*, pp. 499–535. London: Academic Press.

Born M and Wolf E (1985) *Principles of Optics*, 6th edn. (corrected), pp. 491–554. Oxford, UK: Pergamon Press.

Bracewell RN (1978) *The Fourier Transform and its Applications*. New York: McGraw Hill.

Gonzales RC and Woods EW (1992) *Digital Image Processing*, 3rd edn., pp. 253–304. Reading, MA: Addison-Wesley.

Goodman JW (1985) *Statistical Optics*, 1st edn. New York: John Wiley & Sons.

Goodman JW (1996) *Introduction to Fourier Optics*, 2nd edn. New York: McGraw Hill.

Hecht E and Zajac A (1979) *Optics*, 1st edn., pp. 397–435, Reading, MA: Addison-Wesley.

Inoue S and Spring KR (1997) *Video Microscopy*, 2nd edn. New York: Plenum.

Reynolds GO, DeVelis JB, Parrent GB and Thompson JB (1989) *Physical Optics Notebook: Tutorials in Fourier Optics*, 1st edn. New York: SPIE, Bellingham and AIP.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*, pp. 343–381. New York: John Wiley & Sons.

Speckle and Coherence

G Häusler, University of Erlangen-Nürnberg, Erlangen, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Opticists are aware that the amount of coherence plays a significant role in imaging systems: laser speckles are known to add significant noise to the image, as well as parasitic interferences from dusty lenses. Optical systems are often called coherent, if a laser is used (right), and incoherent if other sources come into play (wrong). Many users of optical systems are unaware that it is not the high temporal coherence but the spatial coherence that commonly afflicts the image quality, and that this parasitic spatial coherence is ubiquitous, even though not obvious. Coherent artifacts can occur without the use of lasers, although speckle noise is more prominent with lasers. Even opticists sometimes underestimate the damage that residual coherent noise can cause, and as laser oriented sensor funding programs are 'en vogue', nonexperts are disappointed if some metrology device does not include a laser.

This encyclopedia addresses the many uses of lasers. In this article, we will discuss the costs of coherence. The commonly pretended incoherent approach of everyday optics may lead to significant quantitative measuring errors of illumination or reflectivity, 3d shape, distance or size. Spatial coherence is the dominant source of noise. We will give some rules of thumb to estimate these errors and a few tricks to reduce coherent noise. These rules will help to minimize coherent noise, however, it turns out that as spatial coherence is ubiquitous, there are only limited options available to clear it. One of the options to build good sensors that measure shape, reflectivity, etc. is avoiding the use of lasers!

To become familiar with some basics of the theory of coherence we refer the reader the Further Reading section at the end of this article.

Coherence can be boon or disaster for the opticist, as is explained in other articles of this encyclopedia about interferometry, diffraction, and holography. A specific topic is information acquisition from coherently scattered light. An enlightening example is where speckles in white light interferometry at rough surfaces and in speckle interferometry are exploited. We will briefly discuss white light interferometry at rough surfaces in the section on speckles as carriers of information below.

Practical Coherence Theory

A major issue of this chapter will be corrupting properties of coherence in the daily life of an optical metrologist. We will demonstrate that 'speckle' noise is ever present, and essentially unavoidable, in the images of (diffusively reflecting) objects. Its influence on the quality of optical measurements leads to a lower limit of the physical measuring uncertainty.

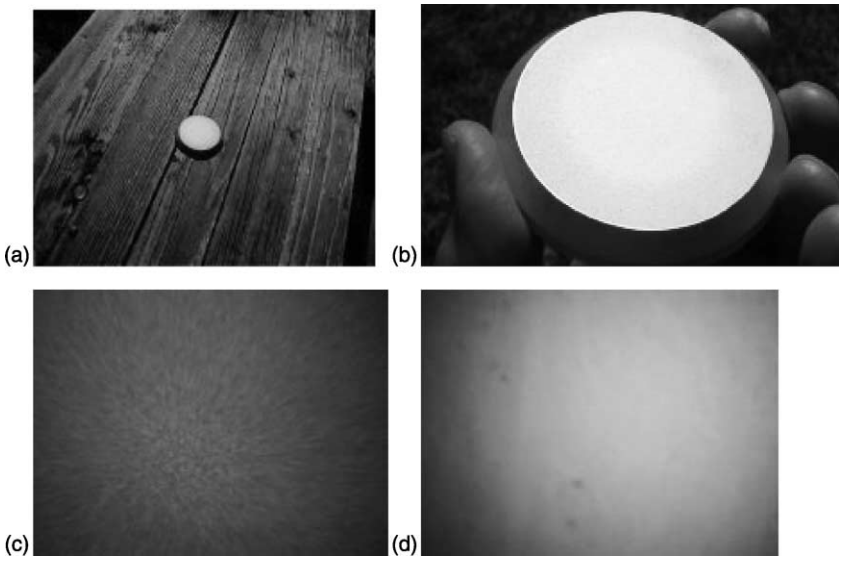


Figure 1 Ground glass in sunlight: (a) the image of a ground glass, illuminated by the sun, observed with small aperture (from large distance), displays no visible speckles; (b) a medium observation aperture (smaller distance) displays weak visible speckles; (c) observation with very high aperture displays strong contrast speckles; and (d) the image of the ground glass on a cloudy day does not display speckles, due to the very big aperture of illumination. Reproduced with permission from Häusler G (1999) Optical sensors and algorithms for reverse engineering. In: Donah S (ed.) Proceedings of the 3rd Topical Meeting on Optoelectronic Distance/Displacement Measurements and Applications. IEEE-LEOS.

This 'coherent noise' limit is, not surprisingly, identical to Heisenberg's limit. We will start by summarizing the important results, and will give some simple explanations later.

Some Basic Observations

(i) The major source of noise in optical measurements is spatial coherence, the temporal coherence status, generally, does not play a significant role.
(ii) An optically rough surface displays subjective speckles with contrast $C = 1$, if the observation aperture $\sin u_o$ is larger than the illumination aperture $\sin u_q$.
(iii) The speckle contrast is $C \sim \sin u_q / \sin u_o$, if the observation aperture is smaller than the illumination aperture.

Figure 1 illustrates the situation by a simple experiment.

The results of Figure 1 can be explained simply by summarizing some 80 pages of coherence theory in a nutshell, by useful simplifications and approximations (Figure 2).

If the source is at an infinite distance, the coherence function no longer depends on the two variables but just on the slit distance $d = |x_1 - x_2|$. In this case, the coherence function $\Gamma(d)$ can be easily calculated as the Fourier transform of the spatial intensity

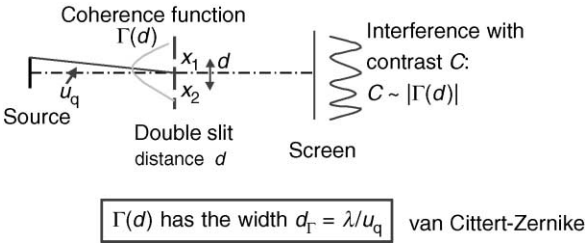


Figure 2 Basic experiment for spatial coherence. An extended source, such as the sun or some incandescent lamp illuminates a double slit, from a large distance. On a screen behind the double slit we can observe 'double slit interference', if the waves coming from x_1 and x_2 display coherence. The contrast of the interference fringes is given by the magnitude of the coherence function Γ (neglecting some normalization factor). Γ is a function of the two slit locations x_1 and x_2.

distribution of the source (van Cittert-Zernike theorem). Let us assume a one dimensional source with an angular size of $2u_q$, where u_q is commonly called the aperture of the illumination. In this case, the coherence function will be:

$$\Gamma(d) \sim \sin c(2u_q d/\lambda) \qquad [1]$$

The width of the coherence function which gives the size of the coherently illuminated area (coherence area) can be approximated from eqn [1]:

$$d_\Gamma = \lambda/u_q \qquad [2]$$

Equation [1] includes some approximations: for circular sources such as the sun, we get an additional

factor 1.22 for the width of the coherence function – the 'sinc-function' has to be replaced by the Bessel function $J_1(r)/(r)$. There is another approximation: we should replace the aperture angle u_q by $\sin u_q$, for larger angles. We choose the wavelength of the maximum sensitivity of the eye, at $\lambda = 0.55\ \mu m$, which is the wavelength of the maximum emission of the sun also.

In conclusion, if two points at the object are closer than d_Γ these points will be illuminated with a high degree of coherence. The light waves, scattered from these object-locations can display interference contrast. Specifically, they may cancel each other out, causing a low signal-to-noise ratio of only 1:1. This does not happen if the points have a distance larger than d_Γ.

From Figure 3 we learn that the width of the spatial coherence function from sunlight illumination at the earth's surface is ~110 μm. (Stars more distant than the sun appear smaller, and hence have a wider coherence function, at the earth's surface. Michelson was able to measure the angular size of some close stars, by measuring the width of this coherence function, which was about $d_\Gamma \sim 10$ m.)

Figure 4 again displays an easy to perform experiment. We can see speckles at sunlight illumination. We can observe speckles as well in shops because they often use small halogen spot illumination.

So far we have only discussed the first stage (from the source to the object). We still have to discuss the way from the object to the retina (or CCD-chip of a video camera). This is illustrated by Figure 5.

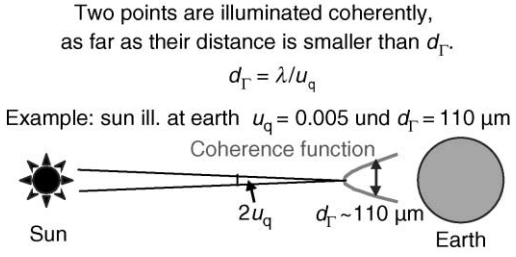


Figure 3 Coherence from sunlight illumination. With the van Cittert–Zernike theorem and an approximate illumination aperture of the sun $u_q^{sun} \sim 0.005$ (0.25°), we get a width of the coherence function at the Earth's surface of about $d_\Gamma \sim 110\ \mu m$.



Figure 4 Finger nail in sunlight, with speckles.

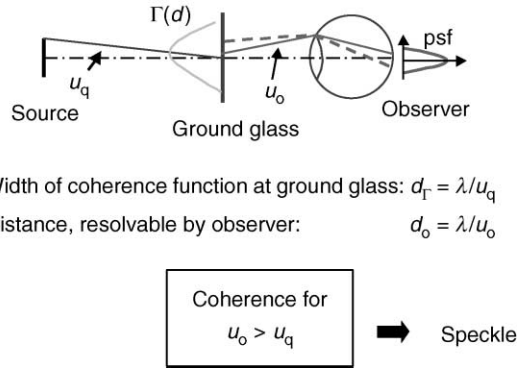


Figure 5 Coherence in a nutshell: spatial coherence and observer.

As an object, we choose some diffusely reflecting or transmitting surface (such as a ground glass). The coherence function at the ground glass has again a width of d_Γ. What happens if we image this ground glass onto our retina? Let us assume our eye to be diffraction limited (which, in fact, it is at sunlight conditions, where the eye pupil has only a diameter $\Phi_{pupil} = 2.5$ mm diameter or less). Then a point x_1 at the ground glass will cause a diffraction spot at the retina, with a size of:

$$d'_{diffr} = \lambda/u'_o = 8\ \mu m \qquad [3]$$

The image sided aperture of observation u'_0 of the eye is calculated from Φ_{pupil} and from its focal length $f_{eye} = 18$ mm as $u'_o = \Phi_{pupil}/2f_{eye} = 0.07$. If we project the diffraction spot at the retina back onto the object, we get its size d from:

$$d_{diffr} = \lambda/u_o = \lambda/(\Phi_{pupil}/2z_o) \qquad [4]$$

with z_o = observation distance of the object from the eye. We call $u_o = \Phi_{pupil}/2z_o$ the object sided observation aperture. Let us calculate the laterally resolvable distance d_{diffr} at the object, with a distance $z_o = 250$ mm, which is called the 'comfortable viewing distance':

$$d_{diffr}(z_o = 250\ mm) = 110\ \mu m \qquad [5]$$

After these preparations we come to the crucial issue: how does the image at the retina appear if we cannot resolve distances at an object smaller than d_Γ, and how does it appear if the resolution of the eye is sufficient to resolve distances smaller than d_Γ?

Let us start with the first assumption: $d_{diffr} \gg d_\Gamma$. Now the images of points at the object, over an area of the diffraction spot are more or less incoherently averaged at the retina, so we will see little interference contrast or 'speckles'. From eqn [4] we see that at

this incoherent averaging starts for distances larger than 250 mm, which is the comfortable viewing distance. Adults cannot see objects from much closer distances. We generally do not see speckles, not just at the limit of maximum accommodation. For larger distances, if we take $z_o = 2500$ mm, the resolvable distance at the object will be 1.1 mm, which is 10 times larger than the diameter of the coherence area. Averaging over such a large area will drop the interference contrast by a factor of 10. Note that such a small interference contrast might not be visible, but it is not at zero!

Coming to the second assumption we can laterally resolve distances smaller than d_Γ. In order to understand this, we first have to learn what is an 'optically rough' object surface. Figure 6 illustrates the problem.

A surface commonly is called 'rough' if the local height variations are larger than λ. However, a surface appears only rough, if the height variation within the distance that can be resolved by the observer, is larger than λ (reflected light will travel the distance twice, hence a roughness of $\lambda/2$ will be sufficient). Then the scattered different phasors, or 'Huygens elementary waves':

$$u_s \sim \exp[i2kz(x, y)] \qquad [6]$$

scattered from the object at x, y may have big phase differences so that destructive interferences (and speckle) can occur in the area of the diffraction image. So the attribute 'rough' depends on the object as well as on the observation aperture. With a high observation aperture (microscope), the diffraction image is small and within that area the phase differences might be small as well. So a ground glass may then look locally like a mirror, while with a small observation aperture it appears 'rough'.

With the 'rough' observation mode assumption, we can summarize what we have assumed: with resolving distances smaller than the coherence width d_Γ we will see high interference contrast; in fact we see a speckle contrast of $C = 1$, if the roughness of the object is smaller than the coherence length of the source. This is the case for most metal surfaces, for ground glasses, or for worked plastic surfaces. The assumption is not true for 'translucent' surfaces, such as skin, paper, or wood. This will be discussed below.

The observation that coherent imaging is achieved, if we can resolve distances smaller than the coherence width d_Γ, is identical to the simple rule that fully coherent imaging occurs if the observation aperture u_o is larger than the illumination aperture u_q.

As mentioned, we will incoherently average in the image plane, over some object area, determined by the size of the diffraction spot. According to the rules of speckle-averaging, the speckle contrast C decreases with the inverse square root of the number N of incoherently averaged speckle patterns. This number N is equal to the ratio of the area A_{diff} of the diffraction spot divided by the coherence area $A_\Gamma = d_\Gamma^2$. So we obtain for the speckle contrast C:

$$C = 1 \quad \text{for } u_q < u_o \qquad [7a]$$

$$C = 1/\sqrt{(N)} = u_o/u_q, \quad \text{for } u_q >= u_o \qquad [7b]$$

We summarize the results in Figure 7.

Equation [7b] has an interesting consequence: we never get rid of speckle noise, even for large illumination apertures and small observation apertures. In many practical instruments, such as a slide projector, the illumination 'aperture stop' cannot be greater than the imaging lens. Fortunately, the observer's eye commonly has a pupil smaller than that of the projector, and/or looks from a distance at the projection screen. Laser projection devices however cause strong and disturbing speckle effects for the user and significant effort is invested to cope with this effect.

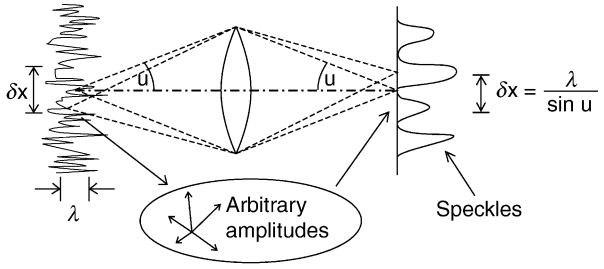


Figure 6 What is a rough surface?

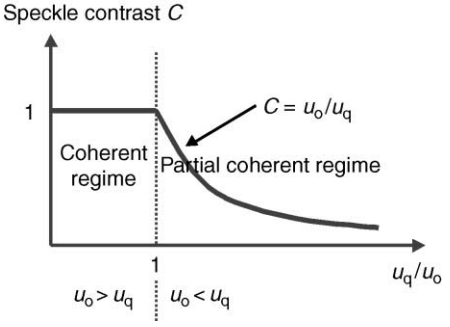


Figure 7 Coherence in a nutshell: what are the conditions for incoherent and coherent imaging?

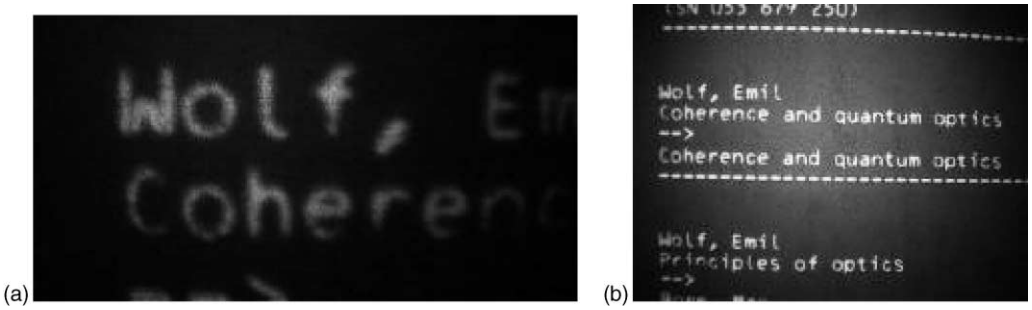


Figure 8 Microfiche projection. For high observation aperture (a) strong speckles occur, and for small observation aperture (b), the speckle contrast is low.

In Figure 8, an image is depicted from a microfiche reading projector. Figure 8a displays the close-up look (high observation aperture) and Figure 8b depicts a more distant view (small observation aperture). The amount of noise is much less in the second case, as it should be according to the rules of eqn [7].

Let us finish this section with some speculation. How would a fly see the sunny world, with a human type of eye? With high observation aperture and at short distance, the world is full of speckle noise. Fortunately, nature invented the facet-eye, for insects, as depicted in Figure 9.

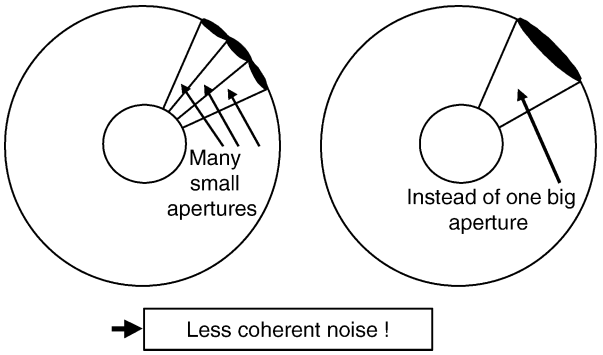


Figure 9 Facet eye and 'human' eye. The facet eye consists of many low aperture lenses, in contrast to the human eye.

Speckle Limits of Metrology

The consequences of the effects discussed above are often underestimated. We should be suspicious, if some effect in nature that disturbs us – like our coherent noise – is ubiquitous, that there might be some deep underlying principle that does not allow us to know everything about the object under observation. Indeed, it turns out that Heisenberg's uncertainty principle is strongly connected with coherent noise (Figure 10).

We can see this from the following experiment: a laser spot is projected onto a ground glass, and imaged with high magnification by a video camera, with an aperture sin u_o. The ground glass is macroscopically planar. When the ground glass is laterally shifted, we find that the observed spot is 'dancing' at the video target. Its observed position is not constant, although its projected position is. It turns out that the standard deviation of the observed position is equal to the uncertainty calculated from the aperture by Heisenberg's principle.

We can calculate the limit for the distance measuring uncertainty, from speckle theory, as well from Heisenberg's principle (within some factor of order 1):

$$\delta z \cdot \delta p_z \gg h/4\pi \qquad [8]$$

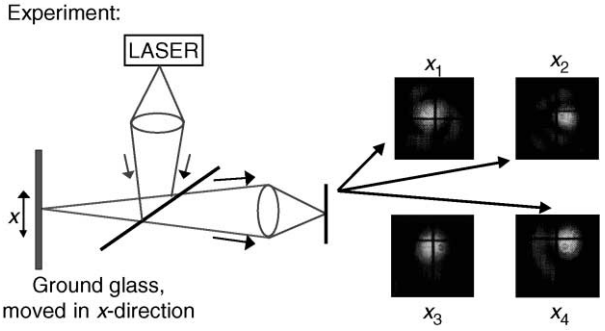


Figure 10 Coherence makes it impossible to localize objects with very high accuracy. Laser spots projected onto a ground glass cannot be imaged without some uncertainty of the lateral position. The four pictures of the spot images, taken with different lateral positions of the ground glass, display the cross bar (true position) and the apparent position of the spot images. The position uncertainty is equal to the uncertainty calculated from Heisenberg's uncertainty principle.

where δp_z is the uncertainty of the photon impulse h/λ in the z-direction (along the optical axis of the measuring system) and h is Planck's constant. For a small measurement uncertainty of the distance, we should allow a big uncertainty δp_z. This can be achieved by a large aperture of the observation system, giving the photons a wide range of possible directions to the lens. We can also allow different

wavelengths, and come to white light interferometry, see section on speckles as carriers of information below. The result is – not surprising – the same as Rayleigh scattering found for the depth of field:

$$\delta z = \lambda / \sin^2 u_o \qquad [9]$$

Coherent noise is the source of the fundamental limit of the distance measuring uncertainty δz of triangulation based sensors.

$$\delta z = C \cdot \lambda / (\sin u_o \sin \theta) \qquad [10]$$

where C is the speckle contrast, λ is the wavelength, $\sin u_o$ is the aperture of observation, and θ is the angle of triangulation (between the direction of the projection and the direction of observation). For a commercially available laser triangulation sensor, (see Figure 11), with $\sin u_o \sim 0.01, C = 1$ and $\theta = 30°$, the measuring uncertainty δz will be larger than 100 μm. We may add, that for $\sin u_o = \sin \theta$, which is valid for auto-focus sensors (such as the confocal scanning microscope), eqn [10] degenerates to the well known Rayleigh depth of field (eqn [9]).

The above results are remarkable in that we cannot know the accurate position of the projected spot or about the position of an intrinsic reflectivity feature; we also cannot know the accurate local reflectivity of a coherently illuminated object. A further consequence is that we do not know the accurate shape of an object in 3D-space. We can calculate this 'physical measuring uncertainty' from the considerations above.

These consequences hold for optically rough surfaces, and for measuring principles that exploit the local intensity of some image, such as with all triangulation type of sensors. The consequences of these considerations are depressing: triangulation with a strong laser is not better than triangulation with only one single photon. The deep reason is, that all coherent photons stem from the same quantum mechanical phase cell and are indistinguishable.

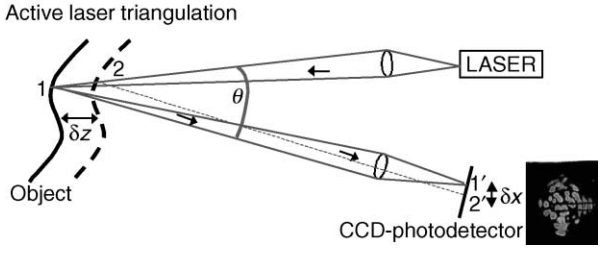


Figure 11 Principle of laser triangulation. The distance of the projected spot is calculated from the location of the speckled spot image and from the angle of triangulation. The speckle noise introduces an ultimate limit of the achievable measuring uncertainty. Reproduced with permission from *Physikalische Plätter* (May 1997): 419.

Hence, many photons do not supply more information than one single photon.

Why Are We Not Aware of Coherence

Generally, coherent noise is not well visible, even in sophisticated technical systems the visibility might be low. There are two main reasons; first, the observation aperture might be much smaller than the aperture of illumination. This is often true, for large distance observation, even with small apertures of illumination. The second reason holds for technical systems, if the observation is implemented via pixelized video targets. If the pixel size is much larger than a single speckle, which is commonly so, then, by averaging over many speckles, the noise is greatly reduced. However, we have to take into account that we pay for this by loss of lateral resolution $1/\delta x$. We can formulate another uncertainty relation:

$$\delta x > \lambda / (C \cdot \sin u_o) \qquad [11]$$

which says that if we want to reduce the speckle contrast C (noise), by lateral averaging, we can do this but lose lateral resolution $1/\delta x$.

Can We Overcome Coherence Limits?

Since the daily interaction of light with matter is coherent scattering, we can overcome the limit of eqn [10] only by looking for measuring principles that are not based on the exploitation of local reflectivity, i.e., on 'conventional imaging'. Concerning optical 3D-measurements, the principle of triangulation uses the position of some image detail, for the calculation of the shape of objects, and has to cope with coherent noise.

Are there different mechanisms of photon–matter interaction, with noncoherent scattering? Fluorescence and thermal excitation are incoherent mechanisms. It can be shown (see Figure 12) that triangulation utilizing fluorescent light, displays much less noise than given by eqn [10]. This is exploited in fluorescent confocal microscopy.

Thermally excited matter emits perfectly incoherent radiation as well. We use this incoherence to measure the material wear in laser processing (see Figure 13), with, again, much better accuracy than given by eqns [9,10].

The sensor is based on triangulation. Nevertheless, by its virtually zero coherent noise, it allows a measuring uncertainty which is limited only by camera noise and other technical imperfections. The uncertainty of the depth measurement through the narrow aperture of the laser nozzle is only a few

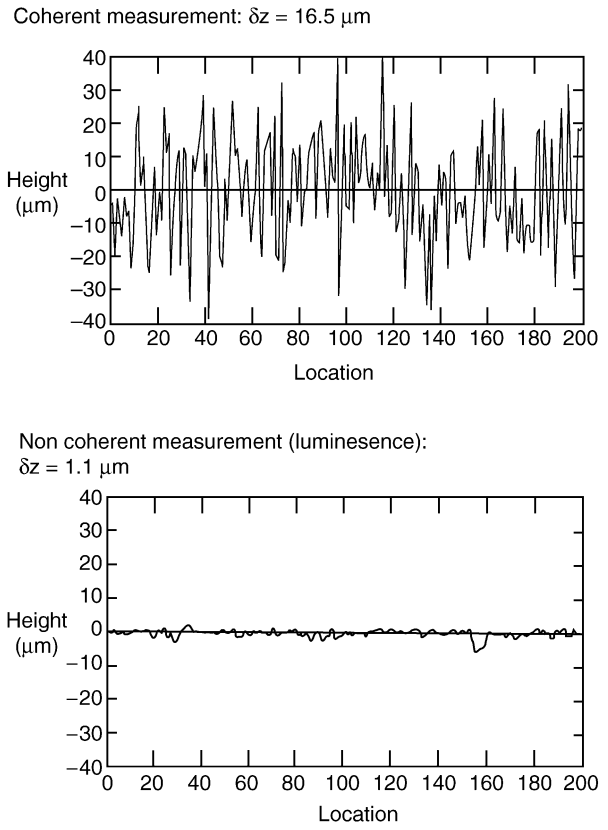


Figure 12 Reduction of speckle noise by triangulation with fluorescent light. A flat metal surface is measured by laser triangulation. Top: surface measured by laser triangulation with full speckle contrast. Bottom: the surface is covered with a very thin fluorescent layer, and illuminated by the same laser, however, the triangulation is done after suppressing the scattered laser light, and utilizing the (incoherent!) fluorescent light. This experiment proves that it is the coherent noise that causes the measuring uncertainty.

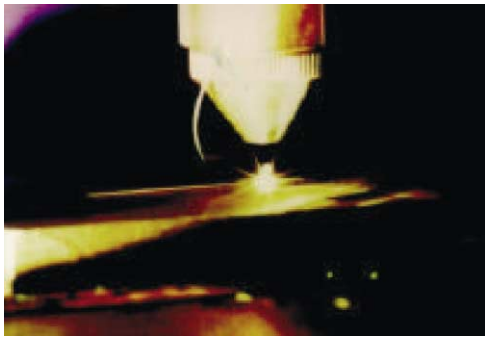


Figure 13 A 100 W laser generates a plasma that emits perfectly incoherent radiation at the object surface. The emitted light can be used to measure on-line the distance of the object surface. A triangulation sensor, exploiting the incoherent plasma emission, controls surface ablation with an accuracy of only a few microns. Reproduced with permission from F & M, *Feinwerktechnik Mikrotechnik Masstechnik* (1995) Issue 9.

microns, beating the limitation of eqn [10] by a factor of about 5, even in the presence of a turbulent plasma spot with a temperature of 3000 K.

Broadband Illumination

Speckle noise can hardly be avoided by broadband illumination. The speckle contrast C is related to the surface roughness σ_z and the coherence length l_c by eqn [12]:

$$C^2 \sim l_c/2\sigma_z \quad \text{for} \quad \sigma_z \gg l_c \qquad [12]$$

So, only for very rough surfaces, and white light illumination, C can be reduced. However, the majority of technical objects is smooth, with a roughness of a few micrometers, hence speckles can be observed even with white light illumination ($l_c \sim 3\ \mu m$). This situation is different for 'translucent' objects such as skin, paper, or some plastic material.

Speckles as a Carrier of Information

So far we discussed the problems of speckles when measuring rough surfaces. We can take advantage of speckles, if we do not stick to triangulation as the basic mechanism of distance measurement. We may utilize the fact that although the phase within each speckle has a random value, this phase is constant over some area. This observation can be used to build an interferometer that works even at rough surfaces. So, with proper choice of the illumination aperture and observation aperture, and with a surface roughness less than the coherence length, we can measure distances by the localization of the temporal coherence function. This function can easily be measured by moving the object under test along the optical axis and measuring the interference intensity (correlogram) at each pixel of the camera. The signal generating mechanism of this 'white light interferometer' distance measurement is not triangulation but 'time of flight'. Here we do not suffer from the limitations of eqn [10]. The limits of white light rough surface interferometry are discussed in the Further Reading section at the end of this article.

The 'coherence radar' and the correlograms of different speckles are depicted in Figures 14 and 15.

The physical mechanism of the signal formation in white light interferometry at rough surfaces is different from that of white light interferometry at smooth surfaces. That is why the method is sometimes referred as 'coherence radar'. We will briefly summarize some advantageous, and – probably unexpected – features of the coherence radar.

We call the real shape of the object, or the surface profile, $z(x, y)$, where z is the local distance at the position (x, y). We give the measured data an index 'm^2':

(i) The physical measuring uncertainty of the measured data is independent from the observation aperture. This is quite a remarkable property, as it enables us to take accurate measurements at the bottom of deep boreholes.

(ii) The standard deviation σ_z of the object can be calculated from the measured data z_m: $\sigma_z = <|z_m|>$. According to German standards, σ_z corresponds to the roughness measure R_q. This measure R_q was calculated from measurements at different roughness gauges, with different observation apertures. Two measurements are depicted in Figure 16.

The experiments shown in Figure 16 display the correct roughness measure, even if the higher frequencies of the object spectrum $Z(\nu, \mu)$ (Figure 17), are not optically resolved. The solution of this paradox might be explained as follows.

From the coherently illuminated object each point scatters a spherical wave. The waves scattered from different object points have, in general, a different phase. At the image plane, we see the laterally

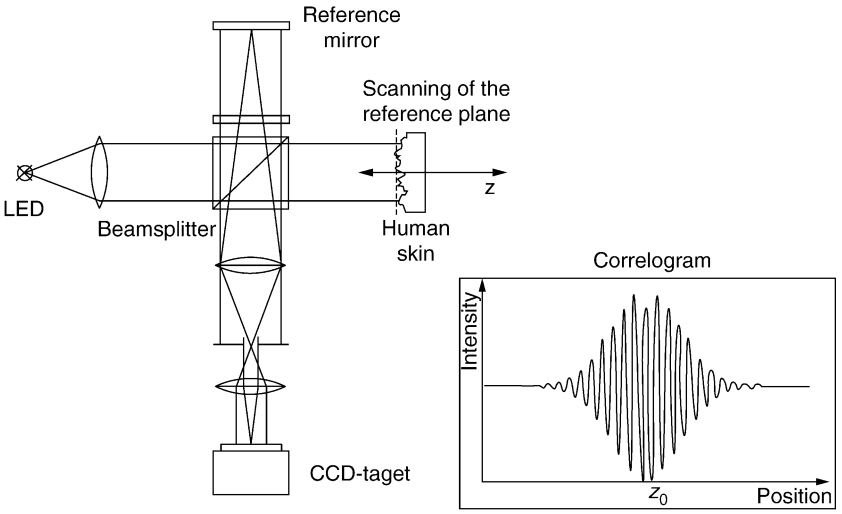


Figure 14 'Coherence radar', white light interferometry at rough surfaces. The surface under test is illuminated with high spatial coherence and low temporal coherence. We acquire the temporal coherence function (correlogram) within each speckle, by scanning the object under test in depth.

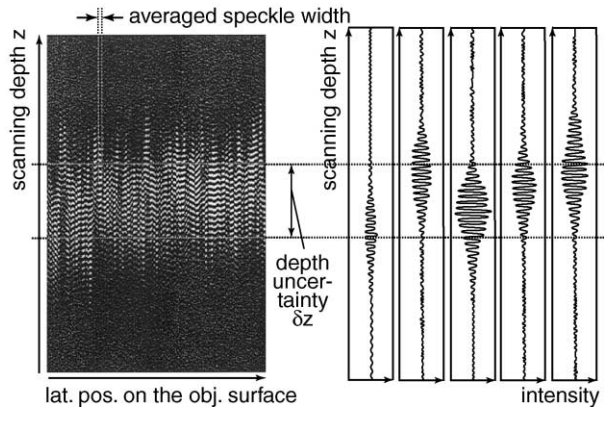


Figure 15 White light speckles in the $x-z$ plane. The left side displays the acquired correlograms, for a couple of speckles, the right side shows the graphs of some of the correlograms. We see that the correlograms are not located at a constant distance, but display some 'distance uncertainty'. This uncertainty does not originate from the instrument, but from the roughness of the object.

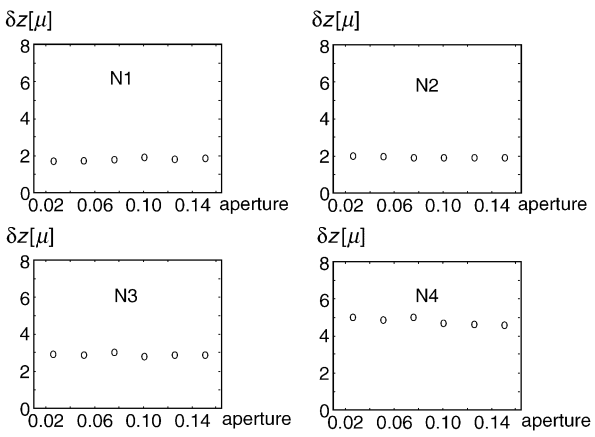


Figure 16 Optical measurement of the roughness beyond the Abbe resolution limit. The roughness is measured for 4 roughness standards N1–N4, with different observation apertures. The roughness is measured correctly, although the microtopology is not resolved by the observing optics.

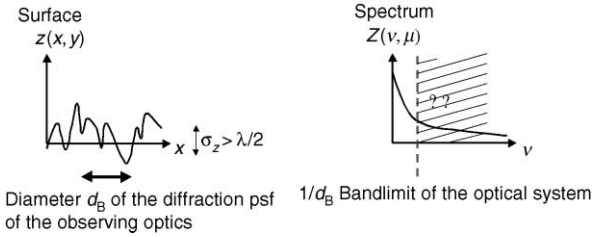


Figure 17 Signal generation beyond the Abbe limit. The microtopology of the object $z(x, y)$ is not resolved, due to the large diffraction image with diameter d_B. Yet we get information from beyond the bandlimit $1/d_B$ of the observing optics.

averaged complex amplitude $<u>$, according to eqn [6]. The averaging is due to the diffraction limited resolution.

Equation [6] reveals that the (spatially) averaged field amplitude $<u>$ is a nonmonotonic, nonlinear function of the surface profile $z(x, y)$, if the surface is rough (i.e., $z \gg \lambda$). We do not average over the profile z but over the complex amplitude. As a consequence of this nonlinearity, the limited observation aperture does not only collect the spatial frequencies ν, μ of the spatial Fourier spectrum $Z(\nu, \mu)$ within the bandlimit $1/d_B$, but also acquires information beyond this bandlimit. The reason is the 'down conversion' of higher spatial frequencies by nonlinear mixing. Simulations and experiments confirm the consequences.

Thus we can evaluate the surface roughness, even without laterally resolving the microtopology. However, there is a principal uncertainty on the measured data. We do not see the true surface but a surface with a 'noise' equivalent to the roughness of the real surface.

Nonlinear nonmonotonic nonlinearities in a sequence of operations may cause 'chaotic' behavior, in other words, small parameter changes such as vibrations, humidity on the surface, etc. may cause large differences of the outcome. In fact, we observe a significant variation of the measured data $z_m(x, y)$ within a sequence of repeated measurements. Our conjecture is that the complex signal formation may be involved in this irregular behavior. The hypothesis may be supported by the observation that much better repeatability can be achieved at specular surfaces, where eqn [6] degenerates to a linear averaging over the surface profile:

$$<u> \sim 1 + ik <z(x, y)>, \quad \text{for } z \ll \lambda \qquad [13]$$

Summary

Spatial coherence is ubiquitous and unavoidable. Spatial coherence disturbs most optical measurements, where optically rough surfaces exist. Then, coherence gives the ultimate limit of the achievable measuring uncertainty. There are simple rules to estimate the measuring uncertainty, by calculating the width of the coherence function and the resolution of the observation system. Spatial coherence and Heisenberg's uncertainty principle lead to the same results of measuring uncertainty.

On the other hand – the formation of signals from coherent light that is scattered at rough surfaces is quite complex and a strongly nonlinear process, which sometimes might encode information which is not otherwise available. One example appears to be the measurement of the roughness with white light interferometry – which is possible, even if the microtopology of the surface is not optically resolved.

See also

Coherence: Overview; Coherence and Imaging.

Further Reading

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge: Cambridge University Press.

Brooker G (2003) *Modern Classical Optics*. Oxford: Oxford University Press.

Dainty JC (ed.) (1984) *Laser Speckle and Related Phenomena*. Berlin, Heidelberg, New York: Springer Verlag.

Dorsch R, Herrmann J and Häusler G (1994) Laser triangulation: fundamental uncertainty in distance measurement. *Applied Optics* 33: 1306–1314.

Dresel T, Venzke H and Häusler G (1992) 3D-sensing of rough surfaces by coherence radar. *Applied Optics* 31: 919–925.

Ettl P (2002) *Über die Signalenstehung bei Weisslichtinterferometrie*. PhD Dissertation, University of Erlangen-Nürnberg.

Häusler G, Ettl P, Schenk M, Bohn G and Laszlo I (1999) Limits of optical range sensors and how exploit them. In: Asakura T (ed.) *Trends in Optics and Phototonics*. Ico IV, Springer Series in Optical Sciences, vol. 74, pp. 328–342. Berlin, Heidelberg, New York: Springer Verlag.

Häusler G and Herrmann J (1992) Physical limits of 3D-sensing. Proceedings of the SPIE Conf. 1822. In: Svetkov D (ed.) *Optics, Illumination, and Image Sensing for Machine Vision VII*, pp. 150–158. Boston, MA: SPIE.

Klein M and Furtak TE (1986) *Optics*. New York: John Wiley & Sons.

Klinger P, Spellenberg B, Herrmann JM and Häusler G (2001) *In Process 3D-Sensing for Laser Material Processing*. Proceedings of Third Int. Conf. on 3-D Digital Imaging and Modeling, Quebec City, Canada. Los Alamitos: IEEE Computer Society, 38–41.

Pedrotti F and Pedrotti L (1993) *Introduction to Optics*, 2nd edn. New Jersey: Prentice Hall.
Pérez J-Ph (1996) *Optik, Spektrum Akademischer*. Heidelberg: Verlag.
Sharma A, Kandpal HC and Ghatak AK (2004) *Optical Coherence*. See this Encyclopedia.
van de Gracht (2004) *Coherence and Imaging*. See this Encyclopedia.
Young M (1993) *Optics and Lasers, Including Fibers and Optical Wave Guides*, 4th revised edn. Berlin, Heidelberg: Springer Verlag.

COHERENT CONTROL

Contents

Theory
Experimental
Applications in Semiconductors

Theory

H Rabitz, Princeton University, Princeton, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since the inception of quantum mechanics almost a century ago, a prime activity has been the observation of quantum phenomena in virtually all areas of chemistry and physics. However, the natural evolution of science leads to the desire to go beyond passive observation to active manipulation of quantum mechanical processes. Achieving control over quantum phenomena could be viewed as engineering at the atomic scale guided by the principles of quantum mechanics, for the alteration of system properties or dynamic behavior. From this perspective, the construction of quantum mechanically operating solid-state devices through selective material growth would fall into this category. The focus of this article is principally on the manipulation of quantum phenomena through tailored laser pulses. The suggestion of using coherent radiation for the active alteration of microworld processes may be traced to the early 1960s, almost immediately after the discovery of lasers. Since then, the subject has grown enormously to encompass the manipulation of (1) chemical reactions, (2) quantum electron transport in semiconductors, (3) excitons in solids, (4) quantum information systems, (5) atom lasers, and (6) high harmonic generation, amongst other topics. Perhaps the most significant use of these techniques may be their provision of refined tools to ultimately better understand the basic physical interactions operative at the atomic scale.

Regardless of the particular application of laser control over quantum phenomena, there is one basic operating principle involved: active manipulation of constructive and destructive quantum wave interferences. This process is depicted in Figure 1, showing the evolution of a quantum system from the initial state $|\psi_i\rangle$ to the desired final state $|\psi_f\rangle$ along three of possibly many interfering pathways. In general, there may be many possible final accessible states $|\psi_f\rangle$, and often the goal is to achieve a high amplitude in one of these states and low amplitude in all the others. The target state might actually be a superposition of states, and an analogous picture to

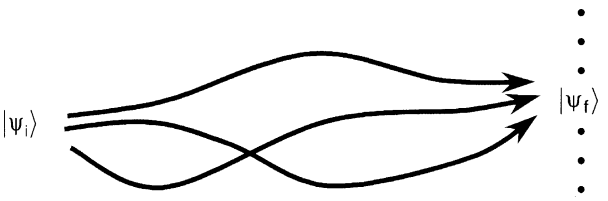


Figure 1 The evolution of a quantum system under laser control from the initial state $|\psi_i\rangle$ to the final state $|\psi_f\rangle$. The latter state is chosen to have desirable physical properties, and three of possibly many pathways between the states are depicted. Successful control of the process $|\psi_i\rangle \rightarrow |\psi_f\rangle$ by a tailored laser pulse generally requires creating constructive quantum wave interferences in the state $|\psi_f\rangle$ from many pathways, and destructive interferences in all other accessible final states $|\psi_{f'}\rangle \neq |\psi_f\rangle$.

that in Figure 1 would also apply to steering about the density matrix. The process depicted in Figure 1 may be thought of as a microscale analog of the classic double-slit experiment for light waves. As the goal is often high-finesse focusing into a particular target quantum state, success may call for the manipulation of many quantum pathways (i.e., the notion of a 'many-slit' experiment). Most applications are concerned with achieving a desirable outcome for the expectation value $\langle\psi_f|O|\psi_f\rangle$ associated with some observable Hermitian operator O.

The practical realization of the task above becomes a control problem when the system is expressed through its Hamiltonian $H = H_0 + V_c$, where H_0 is the free Hamiltonian describing the dynamics without explicit control; it is assumed that the free evolutionary dynamics under H_0 will not satisfactorily achieve the physical objective. Thus, a laboratory-accessible control term V_c is introduced in the Hamiltonian to achieve the desired manipulation. Even just considering radiative interactions, the form of V_c could be quite diverse depending on the nature of the system (e.g., nuclear spins, electrons, atoms, molecules, etc.) and the intensity of the radiation field. Many problems may be treated through an electric dipole interaction $V_c = -\boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)$ where $\boldsymbol{\mu}$ is a system dipole moment and $\boldsymbol{\varepsilon}(t)$ is the laser electric field. Where appropriate, this article will consider the interaction in this form, but other suitable radiative control interactions may be equally well treated. Thus, the laser field $\boldsymbol{\varepsilon}(t)$ as a function of time (or frequency) is at our disposal for attempted manipulation of quantum systems.

Before considering any practical issues associated with the identification of shaped laser control fields, a fundamental question concerns whether it is, in principle, possible to steer about any particular quantum system from an arbitrary initial state $|\psi_i\rangle$ to an arbitrary final state $|\psi_f\rangle$. Questions of this sort are addressed by a controllability analysis of Schrödinger's equation

$$i\hbar\frac{\partial|\psi(t)\rangle}{\partial t} = [H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)]|\psi(t)\rangle, \quad |\psi(0)\rangle = |\psi_i\rangle \qquad [1]$$

Controllability concerns whether, in principle, some field $\boldsymbol{\varepsilon}(t)$ exists, such that the quantum system described by $H_0 = H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)$ permits arbitrary degrees of control. For finite-dimensional quantum systems (i.e., those described by evolution amongst a discrete set of quantum states), the formal tools for such a controllability analysis exist both for evaluating the controllability of the wave function, as well as the more general time evolution operator $U(t)$, which satisfies

$$i\hbar\frac{\partial U}{\partial t} = [H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)]U, \quad U(0) = 1 \qquad [2]$$

Analyses of this type can be quite insightful, but they require detailed knowledge about H_0 and $\boldsymbol{\mu}$. Cases involving quantum information science applications are perhaps the most demanding with regard to achieving total control. Most other physical applications would likely accept much more modest levels of control and still be categorized as excellent achievements.

Theoretical tools and concepts have a number of roles in considering the control of quantum systems, including (1) an exploration of physical/chemical phenomena under active control, (2) the design of viable control fields, (3) the development of algorithms to actively guide laboratory control experiments towards achieving their dynamical objectives, and (4) the introduction of special algorithms to reveal the physical mechanisms operative in the control of quantum phenomena. Activities (1)–(3) have thus far been the primary focus of theoretical studies in this area, and it is anticipated that item (4) will grow in importance in the future.

A few comments on the history of laser control over quantum systems are relevant, as they speak to the special nature of the currently employed successful closed-loop quantum control experiments. Starting in the 1960s and spanning roughly 20 years, it was thought that the design of lasers to manipulate molecular motion could be achieved by the application of simple physical logic and intuition. In particular, the thinking at the time focused on using cw laser fields resonant with one or more local modes of the molecule, as state or energy localization was believed to be the key to successful control. Since quantum dynamics phenomena typically occur on ultrafast time-scales, possibly involving spatially dispersed wave packets, expecting to achieve high quality control with a light source operating at one or two resonant frequencies is generally wishful thinking. Quantum dynamics phenomena occur in a multifrequency domain, and controls with a few frequencies will not suffice. Over approximately the last decade, it became clear that successful control often calls for manipulating multiple interfering quantum pathways (cf., Figure 1). In turn, this recognition led to the need for broad-bandwidth laser sources (i.e., tailored laser pulses). Fortunately, the necessary laser pulse-shaping technologies have become available, and these sources continue to expand into new frequency ranges with enhanced bandwidth capabilities. Many chemical and physical

applications of control over quantum phenomena involve performing a significant amount of work (i.e., quantum mechanical action), and sufficient laser field intensity is required. A commonly quoted adage is 'no field, no yield', and at the other extreme, there was much speculation that operating nonlinearly at high field intensities would lead to a loss of control because the quantum system would effectively act as an amplifier of even weak laser field noise. Fortunately, the latter outcome has not occurred, as is evident from a number of successful high field laser control experiments.

Quantum control may be expressed as an inverse problem with a prescribed chemical or physical objective, and the task being the discovery of a suitable control field $\boldsymbol{\varepsilon}(t)$ to meet the objective. Normally, Schrödinger's equation [1] is viewed as linear, and this perspective is valid if the Hamiltonian is known a priori, leaving the equation to be solved for the wave function. However, in the case of quantum control, neither the wave function nor the control field is known a priori, and the two enter bilinearly on the right-hand side of eqn [1]. Thus, quantum control is mathematically a nonlinear inverse problem. As such, one can anticipate possibly diverse behavior in the process of seeking successful controls, as well as in the ensuing quantum dynamical control behavior. The control must take into account the evolving system dynamics, thereby resulting in the control field being a function(al) of the current state of the system $\boldsymbol{\varepsilon}(\psi)$ in the case of quantum control design, or dependent on the system observations $\boldsymbol{\varepsilon}(\langle O\rangle)$ in the case of laboratory closed-loop field discovery efforts where $\langle O\rangle = \langle\psi|O|\psi\rangle$. Furthermore, the control field at the current time t will depend in some implicit fashion on the future state of the evolving system at the final target time T. It is evident that control field design and laboratory discovery present highly nontrivial tasks.

Although the emphasis in this review is on the theoretical aspects of quantum control theory, the subject exists for its laboratory realizations, and those realizations, in turn, intimately depend on the capabilities of theoretical and algorithmic techniques for their successful implementations. Accordingly, theoretical laser control field design techniques will be discussed, along with algorithmic aspects of current laboratory practice. This material respectively will focus on optimal control theory (OCT) and optimal control experiments (OCEs), as seeking optimality provides the best means to achieve any posed objective. Finally, the last part of the article will present some general conclusions on the state of the quantum control field.

Quantum Optimal Control Theory for Designing Laser Fields

When considering control in any domain of application, a reasonable approach is to computationally design the control for subsequent implementation in the laboratory. Quantum mechanical laser field design has taken on a number of realizations. At one limit is the application of simple intuition for design purposes, and in some special cases (e.g., the weak field perturbation theory regime), this approach may be applicable. Physical insights will always play a central role in laser field design, but to be especially useful, they need to be channeled into the proper mathematical framework. Many of the interesting applications operate in the strong-field nonperturbative regime. In this domain, serious questions arise regarding whether sufficient information is available about the Hamiltonian to execute reliable designs. Regardless of whether the Hamiltonian is known accurately or is an acknowledged model, the theoretical study of quantum control can provide physical insight into the phenomena involved, as well as possibly yielding trial laser pulses for further refinement in the laboratory.

Achieving control over quantum mechanical phenomena often involves a balance of competing dynamical processes. For example, in the case of aiming to create a particular excitation in a molecule or material, there will always be the concomitant need to minimize other unwanted excitations. Often, there are also limitations on the form, intensity, or other characteristics of the laser controls that must be adhered to. Another goal is for the control outcome to be as robust as possible to the laser field fluctuations and Hamiltonian uncertainties. Overriding all of these issues is merely the desire to achieve the best possible physical result for the posed quantum mechanical objective. In summary, all of these desired extrema conditions translate over to posing the design of control laser fields as an optimization problem. Thus, optimal control theory forms a basic foundation for quantum control field design.

Control field design starts by specification of the Hamiltonian components H_0 and $\boldsymbol{\mu}$ in eqn [1], with the goal of finding the best control electric field $\boldsymbol{\varepsilon}(t)$ to balance the competing objectives. Schrödinger's equation must be solved as part of this process, but this effort is not merely a forward propagation task, as the control field is not known a priori. As an optimal design is the goal, a cost functional $J = J(\text{objectives, penalties, } \boldsymbol{\varepsilon}(t))$ is prescribed, which contains the information on the physical objectives, competing penalties, and any costs or constraints associated with the structure of the laser field. The functional J could

contain many terms if the competing physical goals are highly complex. As a specific simple illustration, consider the common goal of steering the system to achieve an expectation value $\langle\psi(T)|O|\psi(T)\rangle$ as close as possible to the target value O_{target} associated with the observable operator O at time T. An additional cost is imposed to minimize the laser fluence. These criteria may be embodied in a cost functional of the form

$$J = \left(\langle\psi(T)|O|\psi(T)\rangle - O_{\text{target}}\right)^2 + \omega\int_0^T \varepsilon^2(t)\,\mathrm{d}t - 2\Im\int_0^T \mathrm{d}t\left\langle\lambda(t)|\mathrm{i}\hbar\frac{\partial}{\partial t} - H_0 + \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)|\psi(t)\right\rangle \quad [3]$$

Here, the parameter $\omega \geq 0$ is a weight to balance the significance of the fluence term relative to achieving the target expectation value, and $|\lambda(t)\rangle$ serves as a Lagrange multiplier, assuring that Schrödinger's equation is satisfied during the variational minimization of J with respect to the control field. Carrying out the latter minimization will lead to eqn [1], along with the additional relations

$$\mathrm{i}\hbar\frac{\partial}{\partial t}|\lambda(t)\rangle = [H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)]|\lambda(t)\rangle, \quad |\lambda(T)\rangle = 2\sigma O|\lambda(T)\rangle \quad [4]$$

$$\boldsymbol{\varepsilon}(t) = \frac{1}{\omega}\Im\langle\lambda(t)|\boldsymbol{\mu}|\psi(t)\rangle \quad [5]$$

$$\sigma = \langle\psi(T)|O|\psi(T)\rangle - O_{\text{target}} \quad [6]$$

Equations [1] and [4]–[6] embody the OCT design equations that must be solved to yield the optimal field $\boldsymbol{\varepsilon}(t)$ given in eqn [5]. These design equations have an unusual mathematical structure within quantum mechanics. First, insertion of the field expression in eqn [5] into eqns [1] and [4] produces two cubically nonlinear coupled Schrödinger-like equations. These equations are in the same family as the standard nonlinear Schrödinger equation, and as such, one may expect that unusual dynamical behavior could arise. Although eqn [4] for $|\lambda(t)\rangle$ is identical in form to the Schrödinger equation [1], only $|\psi(t)\rangle$ is the true wavefunction, with $|\lambda(t)\rangle$ serving to guide the controlled quantum evolution towards the physical objective. Importantly, eqn [1] is an initial value problem, while eqn [4] is a final value problem. Thus, the two equations together form a two-point boundary value problem in time, which is an inherent feature of temporal engineering optimal control as well. The boundary value nature of these equations typically leads to the existence of multiple solutions, where σ in eqn [6] plays the role of a discrete eigenvalue, specifying the quality of the particular achieved control field design. The fact that there are generally multiple solutions to the laser design equations can be attractive from a physical perspective, as the designs may be sorted through to identify those being most attractive for laboratory implementation.

The overall mathematical structure of eqns [1] and [4]–[6] can be melded together into a single design equation

$$\mathrm{i}\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = [H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(\psi, \sigma, t)]|\psi(t)\rangle, \quad |\psi(0)\rangle = |\psi_{\text{i}}\rangle \quad [7]$$

The structure of this equation embodies the comments in the introduction that control field design is inherently a nonlinear process. The main source of complexity arising in eqn [7] is through the control field $\boldsymbol{\varepsilon}(\psi, \sigma, t)$ depending not only on the wave function at the present time t, but also on the future value of the wavefunction at the target time T contained in σ. This structure again reflects the two-point boundary value nature of the control equations.

Given the typical multiplicity of solutions to eqns [1] and [4]–[6], or equivalently, eqn [7], it is attractive to consider approximations to these equations (except perhaps for the inviolate Schrödinger equation [1], and much work continues to be done along these lines. One case involves what is referred to as tracking control, whereby a path is specified for $\langle\psi(t)|O|\psi(t)\rangle$ evolving from $t = 0$ out to $t = T$. Tracking control eliminates σ to produce a field of the form $\boldsymbol{\varepsilon}(\psi,t)$, thereby permitting the design equations to be explicitly integrated as a forward-marching problem toward the target; the tracking equations are still nonlinear with respect to the evolving state. Many variations on these concepts can be envisioned, and other approximations may also be introduced to deal with special circumstances. One technique appropriate for at least few-level systems is stimulated Raman adiabatic passage (STIRAP), which seeks robust adiabatic passage from an initial state to a particular final state, typically with nearly total destructive interference occurring in the intermediate states. This design technique can have special attractive robustness characteristics with respect to field errors. STIRAP, tracking, and various perturbation theory-based control design techniques can be expressed as special cases of OCT, with suitable cost functionals and constraints. Further approximation methods will surely be developed, with the rigorous OCT concepts forming the general foundation for control field design.

As the OCT equations are inherently nonlinear, their solution typically requires numerical iteration,

and a variety of procedures may be utilized for this purpose. Almost all of the methods employ local search techniques (e.g., gradient methods), and some also show monotonic convergence with respect to each new iteration step. Local techniques typically evolve to the nearest solution in the space of possible control fields. Such optimizations can be numerically efficient, although the quality of the attained result can depend on the initial trial for the control field and the details of the algorithm involved. In contrast, global search techniques, such as simulated annealing or genetic algorithms, can search more broadly for the best solution in the control field space. These more expansive searches are attractive, but at the price of typically requiring more intensive computations. Effort has also gone into seeking global or semiglobal input $\rightarrow$ output maps relating the electric field $\boldsymbol{\varepsilon}(t)$ structure to the observable $O[\boldsymbol{\varepsilon}(t)]$. Such maps may be learned, hopefully, from a modest number of computations with a selected set of fields, and then utilized as high-speed interpolators over the control field space to permit the efficient use of global search algorithms to attain the best control solution possible. At the foundation of all OCT design procedures is the need to solve the Schrödinger equation. Thus, computational technology to improve this basic task is of fundamental importance for designing laser fields.

Many computations have been carried out with OCT to attain control field designs for manipulating a host of phenomena, including rotational, vibrational, electronic, and reactive dynamics of molecules, as well as electron motion in semiconductors. Every system has its own rich details, which in turn, are compounded by the fact that multiple control designs will typically exist in many applications producing comparable physical outcomes. Collectively, these control design studies confirm the manipulation of constructive and destructive quantum wave interferences as the general mechanism for achieving successful control over quantum phenomena. This conclusion may be expressed as the following principle:

> **Control field–system cooperativity principle:** Successful quantum control requires that the field must have the proper structure to take full advantage of all of the dynamical opportunities offered by the system, to best satisfy the physical objective.

This simple statement of system–field cooperativity embodies the richness, as well as the complexity, of seeking control over quantum phenomena, and also speaks to why simple intuition alone has not proved to be a generally viable design technique. Quantum systems can often exhibit highly complex dynamical behavior, including broad dispersion of the wave packet over spatial domains or multitudes of quantum states. Handling such complexity can require fields with subtle structure to interact with the quantum system in a global fashion to manage all of the motions involved. Thus, we may expect that successful control pulses will often have broad bandwidth, including amplitude and phase modulation. Until relatively recently, laser sources with these characteristics were not available, but the technology is now in hand and rapidly evolving (see the discussion later).

The cooperativity principle above is of fundamental importance in the control of all quantum phenomena, and a simple illustration of this principle is shown in Figure 2 for the control of wave packet motion on an excited state of the NO molecule. The target for the control is a narrow wave packet located over the well of the excited B state. The optimal control field consists of two coordinated pulses, at early and late times, with both features having internal structure. The figure indicates that the ensuing excited state wave packet has two components, one of which actually passes through the target region during the initial evolution, only to return again and meet the second component at just the right place and time, to achieve the target objective as best as possible. Similar cooperativity interpretations can be found in virtually all implementations with OCT.

The main reason for performing quantum field design is to ultimately implement the designs in the laboratory. The design procedures must face laboratory realities, which include the fact that most Hamiltonians are not known to high accuracy (especially for polyatomic molecules and complex solid-state structures), and secondly, a variety of laboratory field imperfections may unwittingly be present. Notwithstanding these comments, OCT has been fundamental to the development of quantum control, including laying out the logic for how to perform the analogous optimal control experiments. Design implementations and further OCT development will continue to play a basic role in the quantum control field. At present, perhaps the most important contribution of OCT has been to (1) highlight the basic control cooperativity principle above, and (2) provide the basis for developing algorithms to successfully guide optimal control experiments, as discussed below.

Algorithms for Implementing Optimal Control Experiments

The ultimate purpose of considering control theory within quantum mechanics is to take the matter into

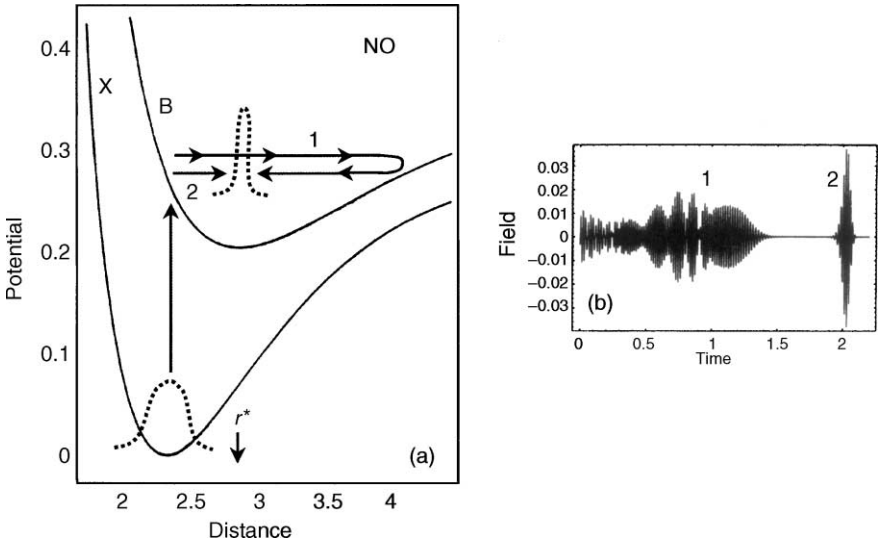


Figure 2 Control of wave packet evolution on the electronically excited B state of NO, with the goal of creating a narrow final packet over the excited state well. (a) indicates that the ground-state packet is brought up in two pieces using the coordinated dual pulse in (b). The piece of the packet from the first pulse actually passes through the target region, then bounces off the right side of the potential, to finally meet the second piece of the packet at just the right time over the target location r^* for successful control. This behavior is an illustration of the control field–system cooperativity principle stated in the text.

the laboratory for implementation and exploitation of its capabilities. Ideally, theoretical control field designs may be attained using the techniques discussed above, followed by the achievement of successful control upon execution of the designs in the laboratory. This appealing approach is burdened with three difficulties: (1) Hamiltonians are often imprecisely known, (2) accurately solving the design equations can be a significant task, and (3) realization of any given design will likely be imperfect due to laboratory noise or other unaccounted-for systematic errors. Perhaps the most serious of these difficulties is point (1), especially considering that the best quality control will be achieved by maximally drawing on subtle constructive and destructive quantum wave interference effects. Exploiting such subtleties will generally require high-quality control designs that, in turn, depend on having reliable Hamiltonians. Although various designs have been carried out seeking robustness with respect to Hamiltonian uncertainties, the issue in point (1) should remain of significance in the foreseeable future, especially for the most complex (and often, the most interesting!) chemical/physical applications. Mitigating this serious problem is the ability to create shaped laser pulse controls and apply them to a quantum system, followed by a probe of their effects at an unprecedented rate of thousands or more independent trials per minute. This unique capability led to the suggestion of partially, if not totally, sidestepping the design process by performing closed-loop experiments to let the quantum system teach the laser how to achieve its control in the laboratory. Figure 3 schematically shows this closed loop process, drawing on the following logic:

1. **The molecular view.** Although there may be theoretical uncertainty about the system Hamiltonian, the actual chemical/physical system under study 'knows' its own Hamiltonian precisely! This knowledge would also include any unusual exigencies, perhaps associated with structural or other defects in the particular sample. Furthermore, upon exposure to a control field, the system 'solves' its own Schrödinger equation impeccably accurately and as fast as possible in real time. Considering these points, the aim is to replace the offline arduous digital computer control field design effort with the actual quantum system under study acting as a precise analog computer, solving the true equations of motion.
2. **Control laser technology.** Pulse shaping under full computer control may be carried out using even hundreds of discrete elements in the frequency domain controlling the phase and amplitude structure of the pulse. This technology is readily available and expanding in terms of pulse center frequency flexibility and bandwidth capabilities.

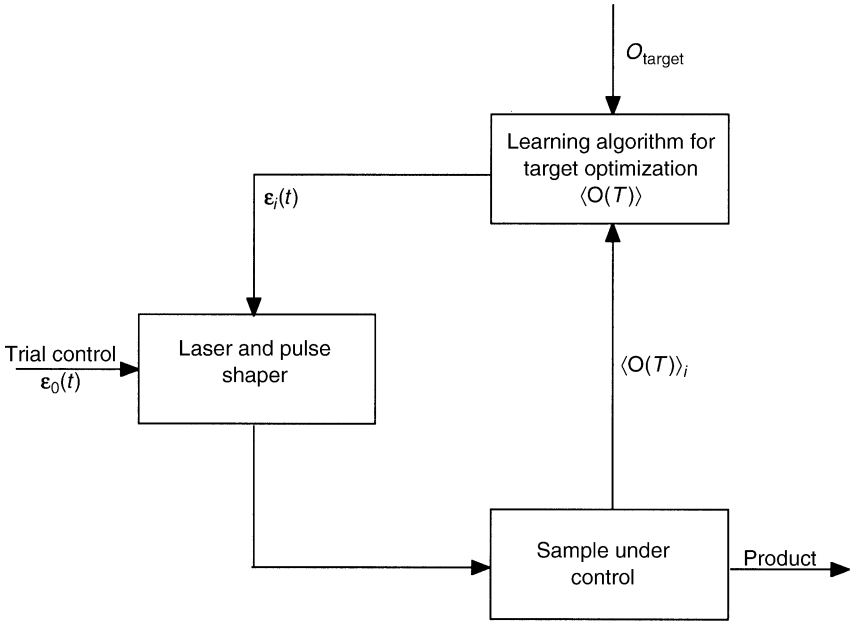


Figure 3 A schematic of the closed-loop concept for allowing the quantum system to teach the laser how to achieve its control. The actual quantum system under control is in a loop with a laser and pulse shaper all slaved together with a pattern recognition algorithm to guide the excursions around the loop. The success of this concept relies on the ability to perform very large numbers of control experiments in a short period of time. In principle, no knowledge of the system Hamiltonian is required to steer the system to the desired final objective, although a good trial design $\varepsilon_0(t)$ may accelerate the process. The ith cycle around the loop attempts to find a better control field $\varepsilon_i(t)$ such that the system response $\langle O(T) \rangle_i$ for the observable operator O comes closer to the desired value O_{target}.

3. **Quantum mechanical objectives**. Many chemical/physical objectives may be simply expressed as the desire to steer a quantum mechanical flux out one clearly defined channel versus another.
4. **Detection of the control action**. Detection of the control outcome can often be carried out by a second laser pulse or any other suitable high duty cycle detection means, such as laser-induced mass spectrometry. The typical circumstance in point (3) implies that little, if any, time-consuming offline data analysis will be necessary beyond that of simple signal averaging.
5. **Fast learning algorithms**. All of the four points above may be slaved together using pattern recognition learning algorithms to identify those control fields which are producing better results, and bias the next sequence of experiments in their favor for further improvement. Although many algorithms may be employed for this purpose, the global search capabilities of genetic algorithms are quite attractive, as they may take full advantage of the high throughput nature of the experiments.

In linking together all of these components into a closed-loop OCE, it is important that no single operational step significantly lags behind any other, for efficiency reasons. Fortunately, this economy is coincident with all the tasks and technologies involved. For achieving control alone, there is no requirement that the actual laser pulse structure be identified on each cycle of the loop. Rather, the laser control 'knob' settings are adequate information for the learning algorithm, as the only criterion is to suitably adjust the knobs to achieve an acceptable value for the chemical/physical objective. The learning algorithm in point (5) operates with a cost functional J, analogous to that used for computational design of fields, but now the cost functional can only depend on those quantities directly observable in the laboratory (i.e., minimally, the current achieved target expectation value). In cases where a physical understanding is sought about the control mechanism, the laser field structure must also be measured. This additional information is typically also not a burden to attain, as it often is only required for the single best control field at the end of the closed-loop learning experiments.

The OCE process in Figure 3 and steps (1)–(5) above constitute the laboratory process for achieving optimal quantum control, and exactly the same reasons put forth for OCT motivate the desire for attaining optimal performance: principally, seeking

the best control result that can possibly be obtained. Seeking optimal performance also ensures some degree of inherent robustness to field noise, as fleeting observations would not likely survive the signal averaging carried out in the laboratory. Additionally, robustness as a specific criterion may also be included in the laboratory cost functional in item (5). When a detailed physical interpretation of the controlled dynamics is desired, it is essential to remove any extraneous control field features that have little impact on the system manipulations. Without the latter clean-up carried out during the experiments, the final field may be contaminated by structures that have little physical impact. Although quantum dynamics typically occurs on femtosecond or picosecond time-scales, the loop excursions in Figure 3 need not be carried out on the latter time-scales. Rather, the loop may be traversed as rapidly as convenient, consistent with the capabilities of the apparatus. A new system sample is introduced on each cycle of the loop. This process is referred learning control, to distinguish it from real-time feedback control.

The closed-loop quantum learning control algorithm can be self-starting, requiring no prior control field designs, under favorable circumstances. Virtually all of the current experiments were carried out in this fashion. It remains to be seen if this procedure of 'going in blind' will be generally applicable in highly complex situations where the initial state is far from the target state. Learning control can only proceed if at least a minimal signal is observed in the target state. Presently, this requirement has been met by drawing on the overwhelmingly large number of exploratory experiments that can be carried out, even in a brief few minutes, under full computer control. However, in some situations, the performance of intermediate observations along the way to the target goal may be necessary in order to at least partially guide the quantum dynamics towards the ultimate objective. Beneficial use could also be made from prior theoretical laser control designs $\boldsymbol{\varepsilon}_0(t)$ capable of at least yielding a minimal target signal.

The OCE closed-loop procedure was based on the growing number of successful theoretical OCT design calculations, even with all of their foibles, especially including less than perfect Hamiltonians. The OCT and OCE processes are analogous, with their primary distinction involving precisely what appears in their corresponding cost functionals. Theoretical guidance can aid in identifying the appropriate cost functionals and learning algorithms for OCE, especially for addressing the needs associated with attaining a physical understanding about the mechanisms governing controlled quantum dynamics phenomena.

A central issue is the rate that learning control occurs upon excursions around the loop in Figure 3. A number of factors control this rate of learning, including the nature of the physical system, the choice of objective, the presence of field uncertainties and measurement errors, the number of control variables, and the capabilities of the learning algorithm employed. Achieving control is a matter of discrimination, and some of these factors, whether inherent to the system or introduced by choice, may work against attaining good-quality discrimination. At this juncture, little is known quantitatively about the limitations associated with any of these factors.

The latter issues have not prevented the successful performance of a broad variety of quantum control experiments, and at present the ability to carry out massive numbers of closed-loop excursions has overcome any evident difficulties. The number of examples of successful closed-loop quantum control is rapidly growing, with illustrations involving the manipulation of laser dye fluorescence, chemical reactivity, high harmonic generation, semiconductor optical switching, fiber optic pulse transmission, and dynamical discrimination of similar chemical species, amongst others. Perhaps the most interesting cases are those carried out at high field intensities, where prior speculation suggested that such experiments would fail due to even modest laser field noise being amplified by the quantum dynamics. Fortunately, this outcome did not occur, and theoretical studies have indicated that surprising degrees of robustness to field noise may accompany the manipulation of quantum dynamics phenomena. Although the presence of field noise may diminish the beneficial influence of constructive and destructive interferences, evidence shows that field noise does not appear to kill the actual control process, at least in terms of obtaining reasonable values for the control objectives. One illustration of control in the high-field regime is shown in Figure 4, demonstrating the learning process for the dissociative rearrangement of acetophenone to form toluene. Interestingly, this case involves breaking two bonds, with the formation of a third, indicating that complex dynamics can be managed with suitably tailored laser pulses.

Operating in the strong-field regime, especially for chemical manipulations, also has the important benefit of producing a generic laser tool to control a broad variety of systems by overcoming the long-standing problem of having sufficient bandwidth. For example, the result in Figure 4 uses a Ti:sapphire laser in the near-infrared regime. Starting with a bandwidth-limited pulse of intensity near

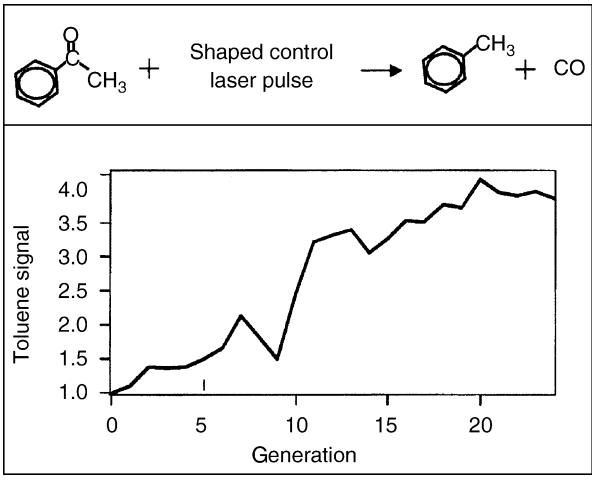


Figure 4 An illustration of dissociative rearrangement achieved by closed-loop learning control in **Figure 3**. The optimally deduced laser pulse broke two bonds in the parent acetophenone molecule and formed a new one to yield the toluene product. The laboratory learning curve is shown for the toluene product signal as a function of the generations in the genetic algorithm guiding the experiments. The fluctuations in the learning curve for optimizing the toluene yield corresponds to the algorithm searching for the optimal yield as the experiments proceed.

$\sim 10^{14}$ W cm^{-2}, in this regime, the dynamic power broadening can easily be on the order of ~ 1 eV or more, thereby effectively converting the otherwise discrete spectrum of the molecule into an effective continuum for ready multiphoton matching by the control laser pulse. Operating under these physical conditions is very attractive, as the apparatus is generic, permitting a single shaped-pulse laser source (e.g., a Ti:sapphire system with phase and amplitude modulation) to be utilized with virtually any chemical system where manipulations are desired. Although every desired physical/chemical goal may not be satisfactorily met (i.e., the goals must be physically attainable!), the means are now available to explore large numbers of systems. Operating under closed loop in the strong-field tailored-pulse regime eliminates the prior serious limitation of first finding a physical system to meet the laser capabilities; now, the structure of the laser pulse can be shaped to meet the molecule's characteristics. Strong field operations may be attractive for these reasons, but other broadband laser sources, possibly working at weaker field intensities, might be essential in some applications (e.g., controlling electron dynamics in semiconductors, where material damage is to be avoided). It is anticipated that additional broadband laser sources will become available for these purposes.

The coming years should be a period of rapid expansion, including a thorough exploration of closed-loop quantum control capabilities. As with the use of any laboratory tool, certain applications may be more amenable than others to attaining successful control. The particular physical/chemical questions need to be well posed, and controls need to have sufficient flexibility to meet the objectives. The experiments ahead should be able to reveal the degree to which drawing on optimality in OCE, combined with the performance of massive numbers of experiments can lead to broad-scale successful control of quantum phenomena. One issue of concern is the richness associated with the large numbers of phase and amplitude control knobs that may be adjusted in the laboratory. Some experiments have already operated with hundreds of knobs, while others have restricted their number in a variety of ways, to simplify the search process. Additional technological and algorithmic advances may be required to manage the high-dimensional control space searches. Fortunately, for typical applications, the search does not reduce to seeking a needle in a haystack, as generally there are multiple control solutions, possibly all of very good quality.

As a final comment on OCE, it is useful to appreciate the subtle distinction of the procedure from closed-loop feedback experiments. This distinction is illustrated in **Figure 5**, pointing to three types of closed-loop quantum control laboratory experiments. The OCE procedure in **Figure 5a** produces a generic laser tool capable of controlling a broad variety of systems with an emphasis in the figure placed on the point that each cycle around the loop starts with a new sample for control. The replacement of the sample on each cycle eliminates a number of difficulties, principally including concerns about sample damage, avoidance of operating the loop at the ultrafast speed of the quantum mechanical processes, and elimination of the effect that 'to observe is to disturb' in quantum mechanics. Thus, learning control provides a generic practical procedure, regardless of the nature of the quantum system. **Figure 5b** has the same structure as that of **Figure 5a**, except the loop is now closed around the same single quantum system, which is followed throughout its evolution. All of the issues mentioned above that are circumvented by operating through laboratory learning control in **Figure 5a** must now be directly faced in the setup of **Figure 5b**. The procedure in **Figure 5b** will likely only be applicable in special circumstances, at least for the reason that many quantum systems operate on time-scales far too fast for opto-electronics and computers to keep up with. However, there are certain quantum mechanical processes that are sufficiently slow to meet the criteria. Furthermore, a period of free evolutionary

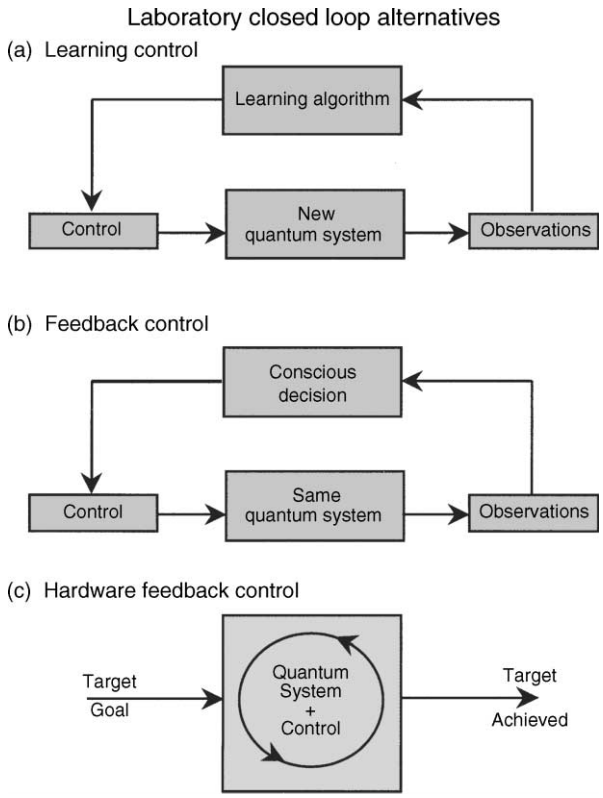


Figure 5 Three possible closed-loop formulations for attaining laser control over quantum phenomena. (a) is a learning control process where a new system is introduced on each excursion around the loop. (b) utilizes feedback control with the loop being traversed with the same sample, generally calling for an accurate model of the system to adjust the controls on each loop excursion. (c) is, at present, a dream machine where the laser and the sample under control are one unit operating without external algorithmic guidance. The closed-loop learning control procedure in (a) appears to form the most practical means to achieve laser control over quantum phenomena, especially for complex systems.

quantum dynamics may be permitted to occur between one control pulse and the next, to make the time issue manageable. While the learning control process in Figures 3 and 5a can be performed model-free, the feedback algorithm in Figure 5b generally must operate with a sufficiently reliable system Hamiltonian to carry out fast design corrections to the control field, based on the previous control outcome. These are very severe demands, but they may be met under certain circumstances. One reason for considering closed-loop feedback control in Figure 5b is to explore the basic physical issue of quantum mechanical limitations inherent in the statement 'to observe is to disturb'. It is suggestive that control over many types of quantum phenomena may fall into the semiclassical regime, lying somewhere between classical engineering behavior and the hard limitations of quantum mechanics. Experiments of the type in Figure 5b will be most interesting for exploring this matter. Finally, Figure 5c introduces a *gedanken* experiment in the sense that the closed-loop process is literally built into the hardware. That is, the laser control and quantum system act as a single functioning unit operating in a stable fashion, so as to automatically steer the system to the desired target. This process may involve engineering the quantum mechanical system prior to control in cases where that freedom exists, as well as engineering the laser components involved. The meaning of such a device in Figure 5c can be understood by considering an analogy with airplane flight, where the aircraft is constructed to have an inherent degree of aerodynamic stability and will essentially fly (glide) on its own accord when pushed forward. It is an open question whether closing the loop in the hardware can be attained for quantum control, and an exploration of this concept may require additional laser technologies.

Conclusions

This article presented an overview of theoretical concepts and algorithmic considerations associated with the control of quantum phenomena. Theory has played a central role in this area by revealing the fundamental principles underlying quantum control, as well as by providing algorithms for designing controls and guiding experiments to discover successful controls in the laboratory. The challenge of controlling quantum phenomena, in one sense, is an old subject, going back at least 40 years. However, roughly the first 30 of these years would best be described as a period of frustration, due to a lack of full understanding of the principles involved and the nature of the lasers needed to achieve success. Thus, from another perspective, the subject is quite young, with perhaps the most notable development being the introduction of closed-loop laboratory learning procedures. At the time of writing this article, these procedures are just beginning to be explored for their full capabilities. To appreciate the young nature of this subject, it is useful to note that the analogous engineering control disciplines presently occupy many thousands of engineers worldwide, both theoreticians and practitioners, and have done so for many years. Yet, engineering control is far from considered a mature subject. Armed now with the basic concepts and proper laboratory tools, one may anticipate a thorough exploration of control over quantum phenomena, including its many possible applications.

Acknowledgments

Support for this work has come from the National Science Foundation and the Department of Defense.

Further Reading

Brixner T, Damrauer NH and Gerber G (2001) Femtosecond quantum control. *Advances in Atomic, Molecular, and Optical Physics* 44: 1–54.

Rabitz H and Zhu W (2000) Optimal control of molecular motion: Design, implementation and inversion. *Accounts of Chemical Research* 33: 572–578.

Rabitz H, de Vivie-Riedle R, Motzkus M and Kompa K (2000) Whither the future of controlling quantum phenomena? *Science* 288: 824–828.

Rice SA and Zhao M (2000) Optical Control of Molecular Processes. NewYork: John Wiley.

Experimental

R J Levis, Temple University, Philadelphia, PA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The experimental control of quantum phenomena in atoms, molecules, and condensed phase materials has been a long-standing challenge in the physical sciences. For example, the idea that a laser beam could be used to selectively cleave a chemical bond has been pursued since the demonstration of the laser in the early 1960s. However, it was not until very recent times that one could realize selective bond cleavage. One key to the rapid acceleration of such experiments is the capability of manipulating the relative phases between two or more paths leading to the same final state. This is done in practice by controlling the phase of two (or two thousand!) laser frequencies coupling the initial state to some desired final state. The present interest in phase-related control stems in part from research focusing on quantum information sciences, controlling chemical reactivity, and developing new optical technologies, such as in biophotonics for imaging cellular materials. Ultimately, one would like to identify coherent control applications having impact similar to linear regime applications like compact disk reading (and writing), integrated circuit fabrication, and photodynamic therapy. Note that in each of these the phase of the electromagnetic field is not an important parameter and only the brightness of the laser at a particular frequency is used. While phase is of negligible importance in any one photon process, for excitations involving two or more photons, phase not only becomes a useful characteristic, it can modulate the yield of a desired process between zero and one hundred percent.

The use of optical phase to manipulate atomic and molecular systems rose to the forefront of laser science in the 1980s. Prior to this there was considerable effort directed toward using the intensity and high spectral resolution features of the laser for various applications. For example, the high spectral resolution character was employed in purifying nuclear isotopes. Highly intense beams were also used in the unsuccessful attempt to pump enough energy into a single vibrational resonance to selectively dissociate a molecule. The current revolution in coherent control has been driven in part by the realization that controlling energy deposition into a single degree of freedom (such as selectively breaking a bond in a polyatomic molecule) can not be accomplished by simply driving a resonance with an ever more intense laser source. In the case of laser-induced photo dissociation, such excitation ultimately results in statistical dissociation throughout the molecule due to vibrational mode coupling. In complex systems, controlling energy deposition into a desired quantum state requires a more sophisticated approach and one scheme involves coherent control.

The two distinct regimes for excitation in coherent control, the time and frequency domain experiments, are formally equivalent. While the experimental approach and level of success for each domain is vastly different, both rely on precisely specifying the phase relations for a series of excitation frequencies. The basic idea behind phase control is simple. One emulates the interference effects that can be easily visualized in a water tank experiment with a wave propagating through two slits. In the coherent control experiment, two or more indistinguishable pathways must be excited that connect one quantum state to another. In either water waves, or coherent control, the probability for observing constructive or destructive interference at a target state depends on the phase difference between the paths connecting the two locations in the water tank, or the initial and final states in the quantum system. In the case of coherent control, the final state may have some technological

value, for instance, a desirable chemical reaction product or an intermediate in the implementation of quantum computation.

Frequency domain methods concentrate on interfering two excitation routes in a system by controlling the phase difference between two distinct frequencies coupling an initial state to a final state. Perhaps the first mention of the concept of phase control in the optical regime can be traced to investigations in Russia in the early 1960s, where the interference between a one- and three-photon path was theoretically proposed as a means to modulate a two-photon absorption. This two-path interference idea lay dormant until the reality of controlling the relative phase of two distinct frequencies was achieved in 1990 using gas phase pressure modulation for frequency domain experiments. The idea was theoretically extended to the control of nuclear motion in 1988. The first experimental example of two-color phase control over nuclear motion was demonstrated in 1991. It is interesting that the first demonstrations of coherent control did not, however, originate in the frequency domain approach, but came from time domain investigations in the mid-1980s. This was work that grew directly out of the chemical dynamics community and will be described below.

While frequency domain methods were developed exclusively for the observation of quantum interference, the experimental implementation of time-domain methods was first developed for observing nuclear motion in real time. In the time domain, a superposition of multiple states is prepared in a system using a short duration ($\sim$50 fs) laser pulse. The superposition state will subsequently evolve according to the phase relationships of the prepared states and the Hamiltonian of the system. At some later time the evolution must be probed using a second, short-time-duration laser pulse. The time-domain methods for coherent control were first proposed in 1985 and were based on wavepacket propagation methods developed in the 1970s. The first experiments were reported in the mid-1980s.

To determine whether the time or frequency domain implementation is more applicable for a particular control context, one should compare the relevant energy level spacing of the quantum system to the bandwidth of the excitation laser. For instance, control of atomic systems is more suited to frequency domain methods because the characteristic energy level spacing in an atom (several electron volts) is large compared to the bandwidth of femtosecond laser pulses (millielectron volts). In this case, preparation of a superposition of two or more states is impractical with a single laser pulse at the present time, dictating that the control scheme must employ the initial state and a single excited eigenstate. In practice, quasi-CW nanosecond duration laser pulses are employed for such experiments because current ultrafast (fs duration) laser sources cannot typically prepare a superposition of electronic states. One instructive exception involves the excitation of Rydberg states in atoms. Here the electronic energy level spacing can be small enough that the bandwidth of a femtosecond excitation laser may span several electronic states and thus can create a superposition state. One should also note that the next generation of laser sources in the attosecond regime may permit the preparation of a wavepacket from low-lying electronic states in an atom. In the case of molecules, avoiding a superposition state is nearly impossible. Such experiments employ pico- to femtosecond duration laser pulses, having a bandwidth sufficient to excite many vibrational levels, either in the ground or excited electronic state manifold of a molecule. To complete the time-dependent measurement, the propagation of the wavepacket to the final state of interest must be observed (probed) using a second ultrafast laser pulse that can produce fluorescence, ionization, or stimulated emission in a time-resolved manner. The multiple paths that link the initial and final state via the pump and the probe pulses are the key to the equivalence of the time and frequency domain methods.

There is another useful way to classify experiments that have been performed recently in coherent control, that is into either open-loop or closed-loop techniques. Open-loop signifies that a calculation may be performed to specify the required pulse shape for control before the experiment is attempted. All of the frequency-based, and most of the time-based experiments fall into this category. Closed-loop, on the other hand, signifies that the results from experimental measurements must be used to assist in determining the correct parameters for the next experiment, this pioneering approach was suggested in 1992. It is interesting that in the context of control experiments performed to date, closed-loop experiments represent the only route to determining the optimal laser pulse shape for controlling even moderately complex systems. In this approach, no input from theoretical models is required for coherent control of complex systems. This is unlike typical control engineering environments, where closed-loop signifies the use of experiments to refine parameters used in a theoretical model.

There is one additional distinction that is of value for understanding the state of coherent control at the present time. Experiments may be classified into systems having Hamiltonians that allow calculation of frequencies required for the desired interference

patterns (thus enabling open-loop experiments) and those having ill-defined Hamiltonians (requiring closed-loop experiments). The open-loop experiments have demonstrated the utility of various control schemes, but are not amenable to systems having even moderate complexity. The closed-loop experiments are capable of controlling systems of moderate complexity but are not amenable to precalculation of the required control fields. Systems requiring closed-loop methods include propagation of short pulses in an optical fiber, control of high harmonic generation for soft X-ray generation, control of chemical reactions and control of biological systems. In the experimental regime, the value of learning control for dealing with complex systems has been well documented.

The remainder of this article represents an overview of the experimental implementation of coherent control. The earliest phase control experiments involved time-dependent probing of molecular wavepackets, and were followed by two-path interference in an atomic system. The most recent, and general, implementation of control involves tailoring a time-dependent electromagnetic field for a desired objective using closed-loop techniques. Since the use of tailored laser pulses appears to be rather general, a more complete description of the experiment is provided.

Time-Dependent Methods

The first experiments demonstrating the possibility of coherent control were performed to detect nuclear motion in molecules in real time. These investigations were the experimental realization of the pump-dump or pulse-timing control methods as originally described in 1985. The experimental applications have expanded to numerous systems including diatomic and triatomic molecules, small clusters, and even biological molecules. For such experiments, a vibrational coherence is prepared by exciting a superposition of states. These measurements implicitly used phase control both in the generation of the ultrashort pulses and in the coherent probing of the superposition state by varying the time delay between the pump and probe pulses, in effect modulating the outcome of a quantum transition. Since the earliest optical experiment in the mid-1980s, there have been many hundreds of such experiments reported in the literature. The motivation for the very first experiments was not explicitly phase control, but rather the observation of nuclear motion of molecules in real time. In such later measurements, oscillations in the superposition states were observable for many cycles, up to many tens of picoseconds in favorable cases, i.e., diatomic molecules in the gas phase. The loss of signal in the coherent oscillations is ultimately due to passage of the superposition state in to surrounding bath states. This may be due to collision with another molecule, dissociation of the system or redistribution of the excitation energy into other modes of the system not originally excited in the superposition. It is notable that coherence can be maintained in solution phase systems for tens of picoseconds and for the case of resonances having weak coupling to other modes, coherence can even be modulated in large biological systems on the picosecond time scale.

Decoherence represents loss of phase information about a system. In this context, phase information represents our detailed knowledge of the electronic and nuclear coordinates of the system. In the case of vibrational coherence, decoherence may be represented by intramolecular vibrational energy transfer to other modes of the molecule. This can be thought of as the propensity of vibrational modes of a molecule to couple together in a manner that randomizes deposited energy throughout the molecule. Decoherence in a condensed phase system includes transfer of energy to the solvent modes surrounding the system of interest. For an electronically excited molecule or atom, decoherence can involve spontaneous emission of radiation (fluorescence), or dissociation in the case of molecules. Certain excited states may have high probability for fluorescence and would represent a significant decoherence pathway and an obstacle for coherent control. These states may be called 'lossy' and are often used in optical pumping schemes as well as for stimulated emission pumping. An interesting question arises as to whether one can employ such states in a pathway leading to a desired final state, or must such lossy states be avoided at all costs in coherent control. This question has been answered to some degree by the method of rapid adiabatic passage. In this experiment, an initial state is coupled to a final state by way of a lossy state. The coupling is a coherent process and the preparation involves dressed states that are eigenstates of the system. In this coherent superposition the lossy state is employed, but no population is allowed to build up and thus decoherence is circumvented.

Two-Path Interference Methods

Interference methods form perhaps the most intuitive example of coherent control. The first experimental demonstration was reported in 1990 and involved the modulation of ionization probability in Hg by interfering a one-photon and three-photon excitation

pathway to an excited electronic state (followed by two-photon ionization). The long delay between the introduction of the two-path interference concept in 1960 and the demonstration in 1990 was the difficulty in controlling the relative phase of the one- and three-photon paths. The solution involved generating the third harmonic of the fundamental in a nonlinear crystal, and copropagating the two beams through a low-density gas. Changing the pressure of the gas changes the optical phase retardance at different rates for different frequencies, thus allowing relative phase control between two colors. The ionization yield of Hg was clearly modulated as a function of pressure. Such schemes have been extended to diatomic molecular systems and the concept of determining molecular phase has been investigated. While changing the phase of more than two frequencies has not been demonstrated using pressure tuning (because of experimental difficulties), a more flexible method has been developed to alter relative phase of up to 1000 frequency bands, as will be described next.

Closed-Loop Control Methods

Perhaps the most exciting new aspect of control in recent years concerns the prospect for performing closed-loop experiments. In this approach a more complex level of control in matter is achieved in comparison to the two-path interference and pump-probe control methods. In the closed-loop method, an arbitrarily complex time-dependent electromagnetic field is prepared that may have multiple subpulses and up to thousands of interfering frequency bands. The realization of closed-loop control over complex processes relies on the confluence of three technologies: (i) production of large bandwidth ultrafast laser pulses; (ii) spatial modulation methods for manipulating the relative phases and amplitudes of component frequencies of the ultrafast laser pulse; and (iii) closed-loop learning control methods for sorting through the vast number of pulse shapes available as potential control fields. In the closed-loop method, the result of a laser molecule interaction is used to tailor an optimized pulse.

In almost all of the experimental systems used for closed-loop control today, Kerr lens mode-locking in the Ti:sapphire crystal is used to phase lock the wide bandwidth of available frequencies (750–950 nm) to create the ultrafast pulse. In this phenomenon, the intensity variation in the short laser pulse creates a transient lens in the Ti:sapphire lasing medium that discriminates between free lasing and mode-locked pulses. As a result, all of the population inversion available in the lasing medium may be coerced into enhancing the ultrashort traveling wave in the cavity in this way. The wide bandwidth available may then be amplified and tailored into a shaped electromagnetic field. Spatial light modulation is one such method for modulating the relative phases and amplitudes of the component frequencies in the ultrafast laser pulse to tailor the shaped laser pulse. Pulse shaping first involves dispersing the phase-locked frequencies on an optical grating, the dispersed radiation is then collimated using a cylindrical lens and the individual frequencies are focused to a line forming the Fourier plane. At the Fourier plane, the time-dependent laser pulse is transformed into a series of phase-locked continuous wave (CW) laser frequencies, and thus Fourier transformed from time to frequency space. The relative phases and amplitudes may be modulated in the Fourier plane using an array of liquid crystal pixels. After altering the spectral phase and amplitude profile, the frequencies are recombined on a second grating to form the shaped laser pulse. With one degree of phase control and 10 pixels there are an astronomic number (360^{10}) of pulse shapes that may be generated in this manner. (At the present time, pulse shapers have up to 2000 independent elements.) Evolutionary algorithms are employed to manage the available phase space. Realizing that the pulse shape is nothing more than the summation of a series of sine waves, each having an associated phase and amplitude, we find that a certain pulse shape can be represented by a genome consisting of an array of these frequency-dependent phases and amplitudes. This immediately suggests that the methods of evolutionary search strategies may be useful for determining the optimal pulse shape for a desired photoexcitation process.

The method of closed-loop optimal control for experiments was first proposed in 1992 and the first experiments were reported in 1997 regarding optimization of the efficiency of a laser dye molecule in solution. Since that time a number of experiments have been performed in both the weak and strong field regime. In the weak field regime, the intensity of the laser does not alter the field free structure of the excited states of the system under investigation. In the strong field, the field free states of the system are altered by the electric field of the laser. Whether one is in the weak or strong field regime depends on parameters including the characteristic level spacing of the states of the system to be controlled and the intensity of the laser coupling into those states. Examples of weak field closed-loop control include optimization of laser dye efficiency, compression of an ultrashort laser pulse, dissociation of alkali clusters, optimization of coherent anti-Stokes

Raman excitation, and optical pulse propagation in fibers. In the strong field, the laser field is used to manipulate the excited states of the molecule as well as induce population transfer among these states. In this sense, the laser is both creating and exciting resonances, in effect allowing universal excitation with a limited frequency bandwidth of 750–850 nm. For chemical applications, it is worth noting that most molecules do not absorb in this wavelength region in the weak field. Strong field control has been used for controlling bond dissociation, chemical rearrangement, photonic reagents, X-ray generation, molecular centrifuges and the manipulation of mass spectral fragmentation intensities. In the latter case, a new sensing technology has emerged wherein each individual pulse shape represents an independent sensor for a molecule. Given the fact that thousands of pulse shapes can be tested per minute, this represents a new paradigm for molecular analysis.

At the time of this writing, the closed-loop method for adaptively tailoring control fields, has far outstripped our theoretical understanding of the process (particularly in the strong field regime). Adaptive control is presently an active field of investigation, encompassing the fields of physics, engineering, optics, and chemistry. In the coming years, there will be as much work done on the mechanism of control as in the application of the methods. Due to small energy level spacings and complexity of the systems, we anticipate the majority of applications to be in the chemical sciences.

See also

Coherence: Coherence and Imaging; Overview. **Interferometry:** Overview. **Microscopy:** Interference Microscopy.

Further Reading

Brumer PW and Shapiro M (2003) *Principles of the Quantum Control of Molecular Processes*. New York: Wiley Interscience.

Chen C and Elliott DS (1990) Measurement of optical phase variations using interfering multiphoton ionization processes. *Physics Review Letters* 65: 1737.

Levis RJ and Rabitz H (2002) Closing the loop on bond selective chemistry using tailored strong field laser pulses. *Journal of Physical Chemistry A* 106: 6427–6444.

Levis RJ, Menkir GM and Rabitz H (2001) Selective bond dissociation and rearrangement with optimally tailored, strong-field laser pulses. *Science* 292(5517): 709–713.

Rice SA and Zhao M (2000) *Optical Control of Molecular Dynamics*. New York: Wiley Interscience.

Stolow A (2003) Femtosecond time-resolved photoelectron spectroscopy of polyatomic molecules. *Annual Review of Physical Chemistry* 54: 89–119.

Zewail AH, Casati G, Rice SA, *et al.* (1997) Femtochemistry: Chemical reaction dynamics and their control. *Chemical Reactions and Their Control on the Femtosecond Time Scale, Solvay Conference on Chemistry* 101: 3–45.

Applications in Semiconductors

H M van Driel and J E Sipe, University of Toronto, Toronto, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Interference phenomena are well-known in classical optics. In the early 19th century, Thomas Young showed conclusively that light has wave properties, with the first interference experiment ever performed. He passed quasi-monochromatic light from a single source through a pair of double slits using the configuration of Figure 1a and observed an interference pattern consisting of a series of bright and dark fringes on a distant screen. The intensity distribution can be explained only if it is assumed that light has wave or phase properties. In modern terminology, if E_1 and E_2 are the complex electric fields of the two light beams arriving at the screen, by the superposition principle of field addition the intensity at a particular point on the screen can be written as

$$I \propto |E_1 + E_2|^2 = |E_1|^2 + |E_2|^2 + |E_1||E_2| \cos(\phi_1 - \phi_2) \qquad [1]$$

where $\phi_1 - \phi_2$ is the phase difference of the beams arriving at the screen. While this simple experiment was used originally to demonstrate the wave properties of light, it has subsequently been used for many other purposes, such as measuring the wavelength of light. However, for our purposes, here the Young's double slit apparatus can also be viewed as a device that redistributes light, or

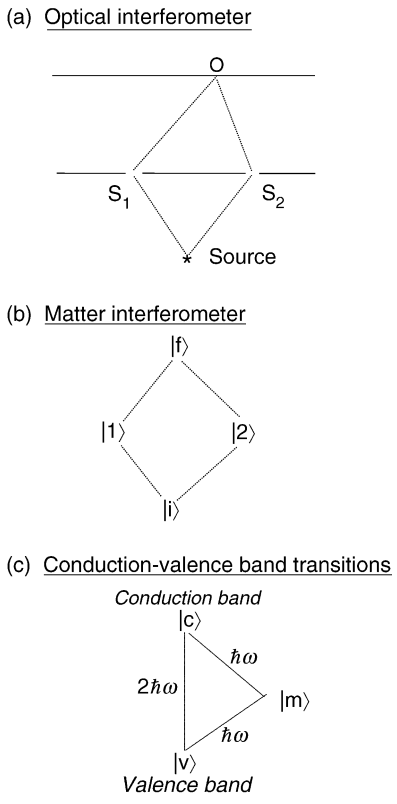


Figure 1 (a) Interference effects in the Young's double slit experiment; (b) illustration of the general concept of coherence control via multiple quantum mechanical pathways; (c) interference of single- and two-photon transitions connecting the same valence and conduction band states in a semiconductor.

controls its intensity at a particular location. For example, let's consider one of the slits to be a source, with the other slit taken to be a gate, with both capable of letting through the same amount of light. If the gate is closed, the distant screen is nearly uniformly illuminated. But if the gate is open, at a particular point on the distant screen there might be zero intensity or as much as four times the intensity emerging from one slit because of interference effects, with all the light merely being spatially redistributed. In this sense the double slit system can be viewed as a device to redistribute the incident light intensity. The key to all this is the superposition principle and the properties of optical phase.

The superposition principle and interference effects also lie at the heart of quantum mechanics. For example, let's now consider a system under the influence of a perturbation with an associated Hamiltonian that has phase properties. In general the system can evolve from one quantum state $|i\rangle$ to another $|f\rangle$ via multiple pathways involving intermediate states $|m\rangle$. Because of the phased perturbation, interference between those pathways can influence the system's final state. If a_m is the (complex) amplitude associated with a transition from the initial to the final state, via intermediate virtual state $|m\rangle$, then for all possible intermediate states the overall transition probability, W, can be written as

$$W = \left| \sum_m a_m \right|^2 \qquad [2]$$

In the case of only two pathways, as illustrated in Figure 1b, W becomes

$$\begin{aligned} W &= |a_1 + a_2|^2 \\ &= |a_1|^2 + |a_2|^2 + |a_1||a_2| \cos(\phi_1 - \phi_2) \end{aligned} \qquad [3]$$

where ϕ_1 and ϕ_2 are now the phases of the two transition amplitudes. While this expression strongly resembles that of eqn [1], here ϕ_1 and ϕ_2 are influenced by the phase properties of the perturbation Hamiltonian that governs the transition process; these phase factors arise, for example, if light fields constitute the perturbation. But overall it is clear that the change in the state of the system is affected by both the amplitude and phase properties of the perturbation. Analogous with the classical interferometer, equality of the transition amplitudes leads to maximum contrast in the transition rate. Although it has been known since the early days of quantum mechanics that the phase of a perturbation can influence the evolution of a system, generally phase has only been discussed in a passive role. Only in recent years has phase been used as a control parameter for a system, on the same level as the strength of the perturbation itself.

Coherence Control

Coherence control, or quantum control as it is sometimes called, refers to the active process through which one can use phase-dependent perturbations originating with, for example, coherent light waves, to control one or more properties of a quantum system, such as state population, momentum, or spin. This more general perspective of interference leads to a picture of interference of matter waves, rather than simply light waves, and one can now speak of an effective 'matter interferometer'. Laser beams, in particular, are in a unique position to play a role in such processes, since they offer a macroscopic 'phase handle' with which to create such effects. This was recognized by the community of atomic and molecular scientists who emphasized the active manifestations of quantum interference effects, and proposed that branching ratios in photochemical reactions

might be controlled through laser-induced interference processes. The use of phase as a control parameter also represents a novel horizon of applications for the laser since most previous applications had involved only amplitude (intensity).

Of course, just as the 'visibility' of the screen pattern for a classical Young's double slit interferometer is governed by the coherence properties of the two slit source, the evolution of a quantum system reflects the coherence properties of the perturbations, and the degree to which one can exercise phase control via external perturbations is also influenced by the interaction of the system of interest with a reservoir – essentially any other degrees of freedom in the system – which can cause decoherence and reduce the effectiveness of the 'matter interferometer'. Since the influence of a reservoir will generally increase with the complexity of a system, one might think that coherence control can only be effectively achieved in simple atomic and molecular systems. Indeed, some of the earliest suggestions for coherence control of a system involved using interference between single and three photon absorption processes, connecting the same initial and final states with photons of frequency 3ω and ω, respectively, and controlling the population of excited states in atomic or diatomic systems. However, it has been difficult to extend this 'two color' paradigm to more complex molecular systems. Since the reservoir leads to decoherence of the system overall, shorter and shorter pulses must be used to overcome decoherence effects. However, short pulses possess a large bandwidth with the result that, for complex systems, selectivity of the final state can be lost. One must therefore consider using interference between multiple pathways in order to control the system, as dictated by the details of the unperturbed Hamiltonian of the system.

For a polyatomic molecule or a solid, the complete Hamiltonian is virtually impossible to determine exactly and it is therefore equally difficult to prescribe the optimal spectral (amplitude and phase) content of the pulses that should be used for control purposes. An alternative approach has therefore emerged, in which one foregoes knowledge of the eigenstates of the Hamiltonian and details of the different possible interfering transition paths in order to achieve control of the end state of a system. This branch of coherence control has come to be known as optimal control. In optimal control one employs an optical source for which one can (ideally) have complete control over the spectral and phase properties. One then uses a feedback process in which experiments are carried out, the effectiveness of achieving a certain result is determined, and the pulse characteristics are then altered to obtain a new result. A key component is the use of an algorithm to select the new pulse properties as part of the feedback system. In this approach the molecule teaches the external control system what it 'requires' for a certain result to be optimally achieved. While the 'best' pulse properties may not directly reveal details of the multiple interference process required to achieve the optimal result, this technique can nonetheless be used to gain some insight into the properties of the unperturbed Hamiltonian and, regardless of such understanding, achieve a desirable result. It has been used to control chemical reaction rates involving several polyatomic molecules, with considerable enhancement in achieving a certain product relative to what can be done using simple thermodynamics.

Coherence Control in Semiconductors

Since coherence control of chemical reactions involving large molecules has represented a significant challenge, it was generally felt that ultrafast decoherence processes would also make control in solids, in general, and semiconductors, in particular, difficult. Early efforts therefore focused on atomic-like situations in which the electrons are bound and associated with discrete states and long coherence times. In semiconductors, the obvious choice is the excitonic system, defect states, or discrete states offered by quantum wells. Population control of excitons and directional ionization of electrons from quantum wells has been clearly demonstrated, similar to related population control of and directional ionization from atoms. Other manifestations of coherence control of semiconductors include control of electron–phonon interactions and intersub-band transitions in quantum wells.

Among the different types of coherence control in semiconductors is the remarkable result that it is possible to coherently control the properties of free electrons associated with continuum states. Although optimal control might be used for some of these processes, we have achieved clear illustrations of control phenomena based on the use of harmonically related beams and interference of single- and two-photon absorption processes connecting the continuum valence and conduction band states as generically illustrated in Figure 1c. Details of the various processes observed can be found elsewhere. Of course, the momentum relaxation time of electrons or holes in continuum states is typically of the order of 100 fs at room temperature, but this time is sufficiently long that it can permit phase-controlled processes. Indeed, with respect to conventional carrier transport, this 'long time' lapse is responsible

for the typically high electrical mobilities in crystalline semiconductors such as Si and GaAs. In essence, a crystalline semiconductor with translational symmetry has crystal momentum as a good quantum number, and selection rules for scattering prevent the momentum relaxation time from being prohibitively small. Since electrons or holes in any continuum state can participate in such control processes, one need not be concerned about pulsewidth or bandwidth, unless the pulses were to be so short that carriers of both positive and negative effective mass were generated within one band.

Coherent Control of Electrical Current Using Two Color Beams

We now illustrate the basic principles that describe how the interference process involving valence and conduction band states can be used to control properties of a bulk semiconductor. We do so in the case of using one or two beams to generate and control carrier population, electrical current, and spin current. In pointing out the underlying ideas behind coherence control of semiconductors, we will skip or suppress many of the mathematical details which are required for a full understanding but which might obscure the essential physics. In particular we consider how phase-related optical beams with frequencies ω and 2ω interact with a direct gap semiconductor such that $\hbar\omega < E_g < 2\hbar\omega$ where E_g is the electronic bandgap. For simplicity we consider exciting electrons from a single valence band via single-photon absorption at 2ω and two-photon absorption at ω. As is well known, within an independent particle approximation, the states of electrons and holes in semiconductors can be labeled by their vector crystal momentum $\mathbf{k}$; the energy of states near the conduction or valence bandedges varies quadratically with $|\mathbf{k}|$. For one-photon absorption the transition amplitude can be derived using a perturbation Hamiltonian of the form $H = e/mc\ \mathbf{A}^{2\omega} \cdot \mathbf{p}$ where $\mathbf{A}^{2\omega}$ is the vector potential associated with the light field and $\mathbf{p}$ is the momentum operator. The transition amplitude is therefore of the form

$$\mathrm{a}^{2\omega} \propto E^{2\omega} e^{i\phi_{2\omega}} p_{cv} \qquad [4]$$

where p_{cv} is the interband matrix element of $\mathbf{p}$ along the field ($E^{2\omega}$) direction; for illustration purposes the field is taken to be linearly polarized. The overall transition rate between two particular states of the same $\mathbf{k}$ can be expressed as $W_1 \propto \mathrm{a}^{2\omega}(\mathrm{a}^{2\omega})^* \propto I^{2\omega}|p_{cv}|^2$ where $I^{2\omega}$ is the intensity of the beam. This rate is independent of the phase of the light beam as well as the sign of $\mathbf{k}$. Hence the absorption of light via single-photon transitions generally populates states of equal and opposite momentum with equal probability or, equivalently, establishes a standing electron wave with zero crystal momentum. This is not surprising since photons possess very little momentum and, in the approximation of a uniform electric field, do not give any momentum to an excited electron. The particular states that are excited depends on the light polarization and crystal orientation. However, the main point is that while single-photon absorption can lead to anisotropic filling of electron states, the distribution in momentum space is not polar (dependent on sign of $\mathbf{k}$). Similar considerations apply to the excited holes, but to avoid repetition we will focus on the electrons only.

For two-photon absorption involving the ω photons and connecting states similar to those connected with single-photon absorption, one must employ the 2nd order perturbation theory using the Hamiltonian $H = e/mc\mathbf{A}^{\omega} \cdot \mathbf{p}$. For two-photon absorption there is a transition from the valence band to an (energy nonallowed) intermediate state followed by a transition from the intermediate state to the final state. To determine the total transition rate one must sum over all possible intermediate states. For semiconductors, by far the dominant intermediate state is the final state itself, so that the associated transition amplitude has the form

$$\mathrm{a}^{\omega} \propto (E^{\omega} e^{i\phi_{\omega}} p_{cv})(E^{2\omega} e^{i\phi_{\omega}} p_{cc}) \propto (E^{\omega})^2 e^{2i\phi_{\omega}} p_{cv} \hbar k \qquad [5]$$

where the matrix element p_{cc} is simply the momentum of the conduction band state ($\hbar k$) along the direction of the field. Note that unlike an atomic system, p_{cc} is nonzero, since Bloch states in a crystalline solid do not possess inversion symmetry. If two-photon absorption acts alone, the overall transition rate between two particular $\mathbf{k}$ states would be $W_2 \propto (I^{\omega})^2 |p_{cv}|^2 k^2$. As with single-photon absorption, because this transition rate is independent of the sign of k, two-photon absorption leads to production of electrons with no net momentum.

When both single- and two-photon transitions are present simultaneously, the transition amplitude is the sum of the transition amplitudes expressed in eqn [3]. The overall transition rate is then found using eqn [1] and yields $W = W_1 + W_2 + Int$ where the interference term Int is given by

$$Int \propto E^{2\omega} E^{\omega} E^{\omega} \sin(\phi_{2\omega} - 2\phi_{\omega}) k \qquad [6]$$

Note, however, that the interference effect depends on the sign of k and hence can be constructive for one part of the conduction band but destructive for other parts of that band, depending on the value of the

relative phase, $\Delta\phi = \phi_{2\omega} - 2\phi_{\omega}$. In principle one can largely eliminate transitions with $+\mathbf{k}$ and enhance those with $-\mathbf{k}$. This effectively generates a net momentum for electrons or holes and hence, at least temporarily, leads to an electrical current in the absence of any external bias. The net momentum of the carriers, which is absent in the individual processes, must come from the lattice. Because the carriers are created with net momentum during the pulses one has a form of current injection that can be written as

$$dJ/dt \propto E^{2\omega}E^{\omega}E^{\omega}\sin(\Delta\phi) \qquad [7]$$

where J is the current density. This type of current injection is allowed in both centrosymmetric and noncentrosymmetric materials. The physics and concepts behind the quantum interference leading to this form of current injection are analogous with the Young's double slit experiment. Note as well that if this 'interferometer' is balanced (single- and two-photon transition rates are similar), then it is possible to control the overall transition rate with great contrast. Roughly speaking, one balances the two arms of the 'effective interferometer' involved in the interference process, one 'arm' corresponding to the one-photon process and the other to the two-photon process. As an example, let's consider the excitation of GaAs, which has a room-temperature bandgap of 1.42 eV (equivalent to 870 nm). For excitation of GaAs using 1550 and 775 nm light under balanced conditions, the electron cloud is injected with a speed close to 500 km s^{-1}. Under 'balanced' conditions, nearly all the electrons are moving in the same direction. Assuming that the irradiance of the 1550 nm beam is 100 MW cm^{-2}, while that of the second harmonic beam is only 15 kW cm^{-2} (satisfying the 'balance' condition), with Gaussian pulse widths of 100 fs, one obtains a surprisingly large peak current of about 1 kA cm^{-2} for a carrier density of only 10^{14} cm^{-3} if scattering effects are ignored. When scattering is taken into account, the value of the peak current is reduced and the transient current decays on a time-scale of the momentum relaxation time. Figure 2a shows an experimental setup which can be used to demonstrate coherence control of electrical current using the above parameters. Figure 2a shows the region between a pair of electrodes on GaAs being illuminated by a train of harmonically related pulses. Figure 2b illustrates how the steady-state experimental voltage across the capacitor changes as the phase parameter $\Delta\phi$ is varied. Transient electrical currents, generated though incident femtosecond optical pulses, have also been detected through the emission of the associated Terahertz radiation.

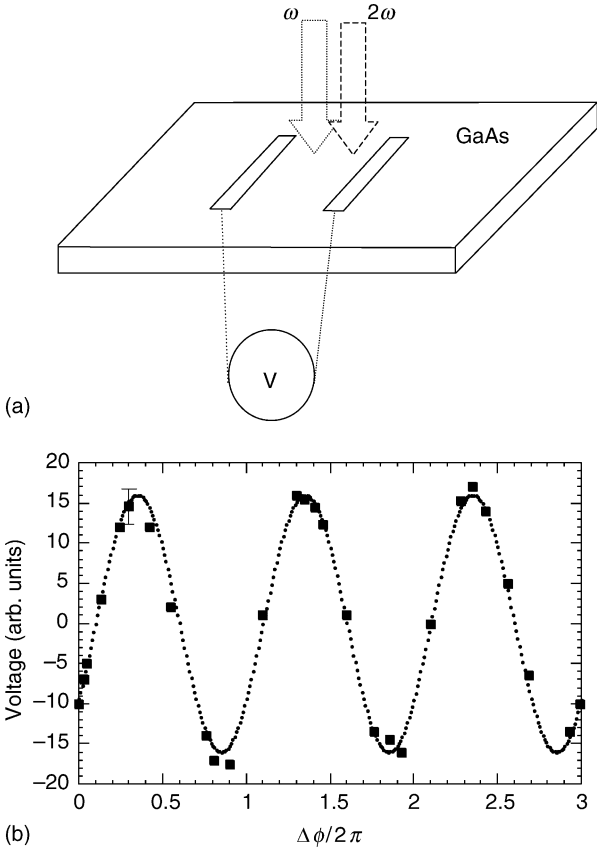


Figure 2 (a) Experimental setup to measure steady-state voltage (V) across a pair of electrodes on GaAs with the intervening region illuminated by phased radiation at ω and 2ω; (b) induced voltage across a pair of electrodes on a GaAs semiconductor as a function of the phase parameter associated with two harmonically related incident beams.

The current injection via coherence control and conventional cases (i.e., under a DC bias) differs with respect to their evolution. The current injected via coherent control has an onset determined by the rise time of the optical pulses. In the case of normal current production, existing carriers are accelerated by an electric field, and the momentum distribution is never far from isotropic. For a carrier density of 10^{14} cm^{-3} a DC field $\sim$80 kV cm^{-1} is required to produce a current density of 1 kA cm^{-2}. In GaAs, with an electron mobility of 8000 cm^2 V^{-1} s^{-1} this current would occur about 1/2 ps after the field is 'instantaneously' turned on. This illustrates that the coherently controlled phenomenon efficiently and quickly produces a larger current than can be achieved with the redirecting of statistically distributed electrons.

A more detailed analysis must take into account the actual light polarization, crystal symmetry, crystal face, and orientation relative to the optical polarization. For given optical intensities of the two beams,

the maximum electrical current injection in the case of GaAs occurs for linearly polarized beams both oriented along the (111) or equivalent direction. However, large currents can also be observed with the light beams polarized along other high symmetry directions.

Coherent Control of Carrier Density, Spin Population, and Spin Current Using Two Color Beams

The processes described above do not exhaust the coherent control effects that can be observed in bulk semiconductors using harmonic beams. Indeed, for noncentrosymmetric materials, certain light polarizations and crystal orientations allow one to coherently control the total carrier generation rate with or without generating an electrical current. In the case of the cubic material GaAs, provided the ω beam has electric field components along two of the three principal crystal axes, with the 2ω beam having a component along the third direction, one can coherently control the total carrier density. However, the overall degree of control here is determined by how 'noncentrosymmetric' the material is.

The spin degrees of freedom of a semiconductor can also be controlled using two color beams. Due to the spin–orbit interaction, the upper valence bands of a typical semiconductor have certain spin characteristics. To date, optical manipulation of electron spin has been largely based on the fact that partially spin-polarized carriers can be injected in a semiconductor via one-photon absorption of circularly polarized light from these upper valence bands. In such carrier injection – where in fact two-photon absorption could be used as well – spins with no net velocity are injected, and then are typically dragged by a bias voltage to produce a spin-polarized current. However, given the protocols discussed above it should not come as a surprise that the two color coherence scheme, when used with certain light polarizations, can coherently control the spin polarization, making it dependent on $\Delta\phi$. Furthermore, for certain polarization combinations it is also possible to generate a spin current with or without an electrical current. Given the excitement surrounding the field of spintronics, where the goal is the use of the spin degree of freedom for data storage and processing, the control of quantities involving the intrinsic angular momentum of the electron is of particular interest.

Various polarization and crystal geometries can be examined for generating spin currents with or without electrical current. For example, with reference to Figure 3a, for both beams propagating in the **z** (*normal*) direction of a crystal and possessing the same circular polarization, an electrical current can be injected in the (*xy*) plane, at an angle from the crystallographic **x** direction dependent on the relative phase parameter, $\Delta\phi$; this current is spin-polarized in the **z** direction. As well, the injected carriers have a $\pm$**z** component of their velocity, as many with one component as with the other; but those going in one direction are preferentially spin-polarized in one direction in the (*xy*) plane, while those in the other direction are preferentially spin-polarized in the opposite direction. This is an example of a spin current in the absence of an electrical current.

Such a pure spin current that is perhaps more striking is observable with the two beams cross linearly polarized, for example with the fundamental beam in the **x** direction and the second harmonic beam in the **y** direction as shown in Figure 3b. Then there is no net spin injection; the average spin in any

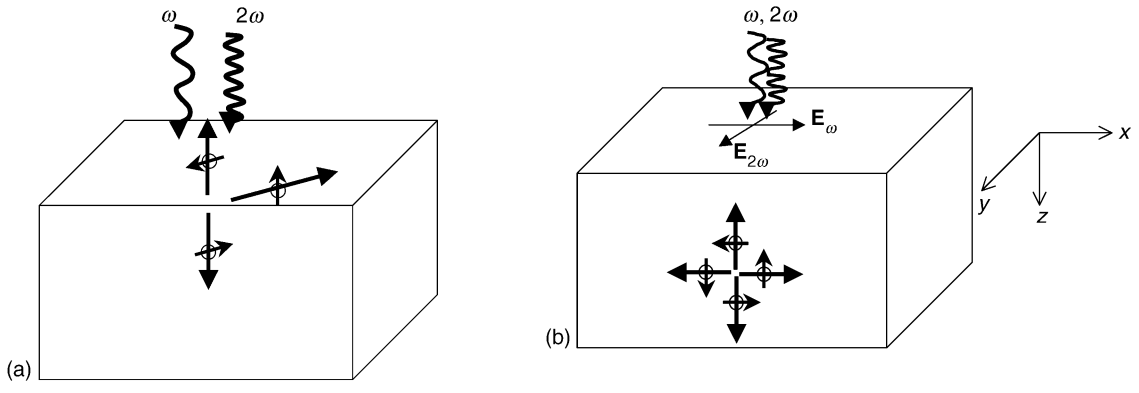


Figure 3 (a) Excitation of a semiconductor by co-circularly polarized ω and 2ω pulses. Arrows indicate that a spin polarized electrical current is generated in the $x-y$ plane in a direction dependent on $\Delta\phi$ while a pure spin current is generated in the beam propagation (z) direction. (b) Excitation of a semiconductor by orthogonally, linearly polarized ω and 2ω pulses. Arrows indicate that a spin polarized electrical current is generated in the direction of the fundamental beam polarization as well as along the beam propagation (z) direction.

direction is zero. And for a vanishing phase parameter $\Delta\phi$ there is no net electrical current injection. Yet, for example, the electrons injected with a $+\mathbf{x}$ velocity component will have one spin polarization with respect to the $\mathbf{z}$ direction, while those injected with a $-\mathbf{x}$ component to their velocity will have the opposite spin polarization with respect to the $\mathbf{z}$ direction.

The examples given above are the spin current analog of the two-color electrical current injection discussed above. There should also be the possibility of injecting a spin current with a single beam into crystals lacking center-of-inversion symmetry.

The simple analysis presented above relies on simple quantum mechanical ideas and calculations using essentially nothing more than Fermi's Golden Rule. Yet the process involves a mixing of two frequencies, and can therefore be thought of as a nonlinear optical effect. Indeed, one of the key ideas to emerge from the theoretical study of such phenomena is that an interpretation of coherence control effects alternate to that provided by the simple quantum interference picture that is provided by the usual susceptibilities of nonlinear optics. These susceptibilities are, of course, based on quantum mechanics, but the macroscopic viewpoint allows for the identification and classification of the effects in terms of 2nd order nonlinear optical effects, 3rd order optical effects, etc. Indeed, one can generalize many of the processes we have discussed above to a hierarchy of frequency mixing effects or high-order nonlinear processes involving multiple beams with frequency $n\omega, m\omega, p\omega\ldots$ with n, m, p being integers. Many of theses schemes require high intensity of one or more of the beams, but can occur in a simple semiconductor.

Coherent Control of Electrical Current Using Single Color Beams

It is also possible to generate coherence control effects using beams at a single frequency (ω), if one focuses on the two orthogonal components (e.g., x and y) polarization states and uses a noncentrosymmetric semiconductor of reduced symmetry such as a strained cubic semiconductor or a wurtzite material such as CdS or CdSe. In this case, interference between absorption pathways associated with the orthogonal components can lead to electrical current injection given by

$$dJ/dt \propto E^{\omega}E^{\omega}\sin(\phi_{\omega}^{x} - \phi_{\omega}^{y}) \qquad [8]$$

Since in this process current injection is linear in the beam's intensity, the high intensities necessary for two-photon absorption are not necessary. Nonetheless, the efficacy of this process is limited by the fact that it relies on the breaking of center-of-inversion symmetry of the underlying crystal. It is also clear that the maximum current occurs for circularly polarized light and that right and left circularly polarized light lead to a difference in sign of the current injection. Finally, for this particular single beam scheme, when circularly polarized light is used the electrical current is partially spin polarized.

Conclusions

Through the phase of optical pulses it is possible to control electrical and spin currents, as well as carrier density, in bulk semiconductors on a time-scale that is limited only by the rise time of the optical pulse and the intrinsic response of the semiconductor. A few optically based processes that allow one to achieve these types of control have been illustrated here. Although at one level one can understand these processes in terms of quantum mechanical interference effects, at a macroscopic level one can understand these control phenomena as manifestations of nonlinear optical phenomena. For fundamental as well as applied reasons, our discussion has focused on coherence control effects using the continuum states in bulk semiconductors, although related effects can also occur in quantum dot, quantum well, and superlattice semiconductors. Applications of these control effects will undoubtedly exploit the all-optical nature of the process, including the speed at which the effects can be turned on or off. The turn-off effects, although not discussed extensively here, are related to transport phenomena as well as momentum scattering and related dephasing processes.

See also

Semiconductor Physics: Spin Transport and Relaxation in Semiconductors; Spintronics. **Interferometry**: Overview. **Coherent Control:** Experimental; Theory.

Further Reading

Brumer P (1995) Laser control of chemical reactions. *Scientific American* 272: 56.

Rabitz H, de Vivie-Riedle R, Motzkus M and Kompa K (2000) Whither the future of controlling quantum phenomena? *Science* 288: 824.

Shah J (1999) *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*. New York: Springer.

van Driel HM and Sipe JE (2001) Coherent control of photocurrents in semiconductors. In: Tsen T (ed.) *Ultrafast Phenomena in Semiconductors*, p. 261. New York: Springer Verlag.

COHERENT LIGHTWAVE SYSTEMS

M J Connelly, University of Limerick, Limerick, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Conventional optical communication systems are based on the intensity modulation (IM) of an optical carrier by an electrical data signal and direct detection (DD) of the received light. The simplest scheme employs on/off keying (OOK), whereby turning on or off, an optical source transmits a binary 1 or 0. The signal light is transmitted down an optical fiber and at the receiver is detected by a photodetector, as shown in Figure 1. The resulting photocurrent is then processed to determine if a 1 or 0 was received. In the ideal case, where a monochromatic light source is used and where no receiver noise (detector dark current and thermal noise) is present, the probability of error is determined by the quantum nature (shot noise) of the received light. In this case, the number of photons/bit required (sensitivity) at the receiver, to achieve a bit-error-rate (BER) of 10^{-9}, is 10. This is called the quantum-limit. In practical receivers without an optical preamplifier, operating in the 1.3 μm and 1.55 μm optical fiber communication windows, the actual sensitivity is typically 10–30 dB less than the theoretical sensitivity. This is due to receiver noise, in particular thermal noise. This means that the actual sensitivity is between 100 to 10 000 photons/bit. The capacity of IM/DD fiber links can be increased using wavelength division multiplexing (WDM). At the receiver the desired channel is selected using a narrowband optical filter. Some advanced commercial IM/DD systems utilize dense WDM with 50 GHz spacing and single-channel bit rates as high as 40 Gb/s. Indeed, IM/DD systems with terabit (10^{12} bit/s) capacity are now possible.

Compared to IM/DD systems, coherent optical communication systems have greater sensitivity and selectivity. In the context of coherent lightwave systems, the term coherent refers to any technique employing nonlinear mixing between two optical waves on a photodetector, as shown in Figure 2. Typically, one of these is an information-bearing signal and the other is a locally generated wave (local oscillator). The result of this heterodyne process is a modulation of the photodetector photocurrent at a frequency equal to the difference between the signal and local oscillator frequencies. This intermediate frequency (IF) electronic signal contains the information in the form of amplitude, frequency, or phase that was present in the original optical signal. The IF signal is filtered and demodulated to retrieve the transmitted information. An automatic frequency control circuit is required to keep the local-oscillator frequency stable.

It is possible for the information signal to contain a number of subcarriers (typically at microwave frequencies), each of which can be modulated by a separate data channel. It is a simple matter to select the desired channel at the receiver by employing electronic heterodyning and low pass filtering, as shown in Figure 3. In IM/DD systems, channel selection can only be carried out using narrowband optical filters.

Coherent optical communications utilize techniques that were first investigated in radio communications. Most of the basic research work on coherent optical communications was carried out in the 1980s and early 1990s, and was primarily motivated by the need for longer fiber links using no repeaters. Improving receiver sensitivity using coherent techniques made it possible to increase fiber link spans. In the mid-1990s, reliable optical fiber amplifiers became available. This made it possible to construct fiber links using optical amplifiers spaced at appropriate intervals to compensate for fiber and other transmission losses.

The sensitivity of an IM/DD receiver can be greatly improved by the use of an optical preamplifier. Indeed the improvement is so marked that coherent techniques are not a viable alternative to IM/DD in the vast majority of commercial optical communication systems due to their complexity and higher cost. Moreover, because semiconductor lasers with very precise wavelengths and narrowband optical filters can be fabricated, the selectivity advantage of coherent optical communications has become less important. This has meant that the research work currently being carried out on coherent optical communications is very limited, and in particular, very few field trials have been carried out since the early 1990s.

We will review the basic principles underlying coherent optical communications, modulation schemes, detection schemes, and coherent receiver sensitivity. The effects of phase noise and polarization on coherent receiver performance will be outlined. A brief summary of pertinent experimental results will also be presented.

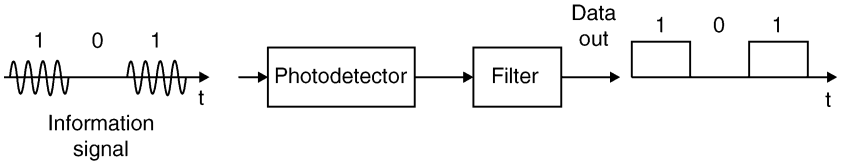


Figure 1 IM/DD optical receiver.

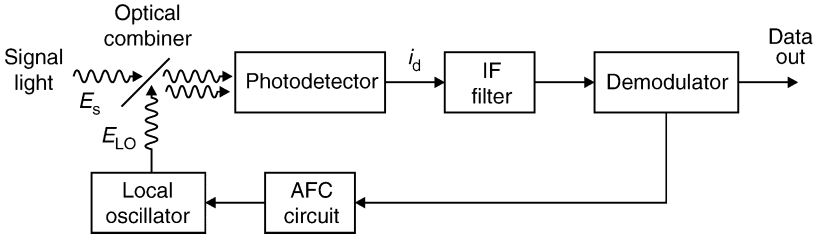


Figure 2 Basic coherent optical receiver. AFC: automatic frequency control.

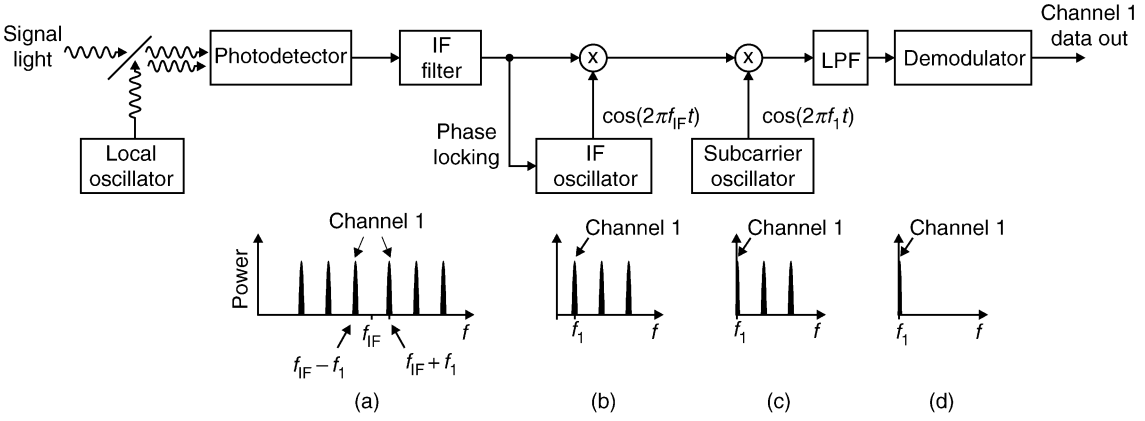


Figure 3 Coherent detection and processing of a subcarrier modulated signal lightwave. The power spectrum of the detected signal is shown: (a) after IF filtering; (b) after multiplication by the IF oscillator; (c) after multiplication with the subcarrier oscillator (in this case channel 1's carrier frequency); and (d) after low pass filtering prior to demodulation and data detection. LPF: lowpass filter.

Basic Principles

In coherent detection, a low-level optical signal field E_s is combined with a much larger power optical signal field E_{LO} from a local oscillator laser. E_s and E_{LO} can be written as

$$E_s(t) = \sqrt{2P_s}\cos[2\pi f_s t + \phi_s(t)] \quad [1]$$

$$E_{LO}(t) = \sqrt{2P_{LO}}\cos[2\pi f_{LO} t + \phi_{LO}(t)] \quad [2]$$

where P_s, f_s, and ϕ_s are the signal power, optical frequency, and phase (including phase noise), respectively. P_{LO}, f_{LO}, and ϕ_{LO} are the equivalent quantities for the local oscillator. For ideal coherent detection, the polarization states of the signal and local oscillator must be equal. If this is the case, the photocurrent is given by

$$i_d(t) = R[E_s(t) + E_{LO}(t)]^2 \quad [3]$$

where R is the detector responsivity. Because the detector cannot respond to optical frequencies, we have

$$i_d(t) = R\left\{P_s + P_{LO} + 2\sqrt{P_s P_{LO}} \times \cos[2\pi f_{IF} t + \phi_s(t) - \phi_{LO}(t)]\right\} \quad [4]$$

The IF signal is centered at frequency $f_{IF} = f_s - f_{LO}$. In homodyne detection $f_s = f_{LO}$ and the homodyne signal is centered at baseband. In heterodyne detection $f_s \neq f_{LO}$ and the signal generated by the photodetector is centered around f_{IF} (typically three to six times the bit rate). The IF photocurrent is proportional to $\sqrt{P_s}$, rather than P_s as is the case in

direct detection, and is effectively amplified by a factor proportional to $\sqrt{P_{LO}}$. If P_{LO} is large enough, the signal power can be raised above the receiver noise, thereby leading to a greater sensitivity than possible with conventional IM/DD receivers using p–i–n or avalanche photodiodes. Theoretically it is possible to reach shot noise limited receiver sensitivity although this is not possible in practice. Homodyne optical receivers usually require less bandwidth and are more sensitive than heterodyne receivers but require that the local oscillator is phase locked to the information signal. This requires an optical phase locked loop (OPLL), which is very difficult to design.

Modulation Schemes

The three principal modulation schemes used in coherent optical communications; amplitude shift keying (ASK), frequency shift keying (FSK) and phase shift keying (PSK), are shown in Figure 4. ASK modulation is essentially the same as the OOK scheme used in IM/DD systems. The optical signal is given by

$$E_s(t) = 2a(t)\sqrt{2P_s}\cos[2\pi f_s t + \phi_s(t)] \qquad [5]$$

where $a(t) = 0$ or 1 for transmission of a 0 or 1, respectively. P_s is the average signal power, assuming that it is equally likely that a 0 or 1 is transmitted. ASK modulation can be achieved by modulating the output of a semiconductor laser using an external modulator. The most common type is based on $LiNbO_3$ waveguides in a Mach–Zehnder interferometer configuration.

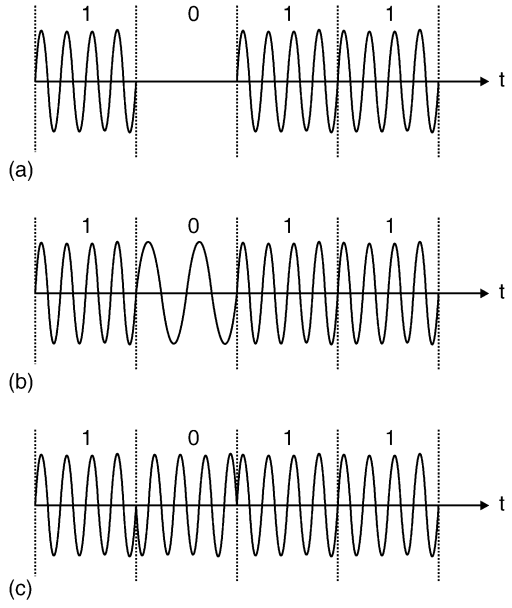


Figure 4 Modulation schemes for coherent optical communications: (a) ASK; (b) FSK; and (c) PSK.

In FSK modulation, the optical frequency of the signal lightwave is changed slightly by the digital data stream. In binary FSK, the optical signal is given by

$$E_s(t) = \sqrt{2P_s}\cos[2\pi(f_s + a(t)\Delta f)t + \phi_s(t)] \qquad [6]$$

where $a(t) = -1$ or 1 for transmission of a 0 or 1, respectively. The choice of the frequency deviation Δf depends on the available bandwidth, the bit rate B_T and the demodulation scheme used in the receiver. The modulation index m is given by

$$m = \frac{2\Delta f}{B_T} \qquad [7]$$

The minimum value of m that produces orthogonality (for independent detection of the 0s and 1s) is 0.5. If $m = 0.5$, the signal is referred to as minimum FSK (MFSK). When $m \geq 2$, the signal is referred to as wide-deviation FSK, and when $m < 2$, as narrow-deviation FSK. Wide-deviation FSK receivers are more resilient to phase noise than narrow-deviation FSK receivers. FSK modulation can be achieved using $LiNbO_3$ external modulators, acoustic optic modulators and distributed feedback semiconductor lasers. The phase of the lightwave does not change between bit transitions.

In PSK modulation, the digital data stream changes the phase of the signal lightwave. In binary PSK the optical signal is given by

$$E_s(t) = \sqrt{2P_s}\cos[2\pi f_s t + a(t)\theta + \frac{\pi}{2} + \phi_s(t)] \qquad [8]$$

where $a(t) = -1$ or 1 for transmission of a 0 or 1, respectively. The phase shift between a 0 and 1 is $\theta + \pi/2$. Most PSK schemes use $\theta = \pi/2$. PSK modulation can be achieved by using an external phase modulator or a multiple quantum well electroabsorption modulator.

Detection Schemes

There are two classes of demodulation schemes: synchronous and asynchronous. The former exploits the frequency and phase of the carrier signal to perform the detection. The latter only uses the envelope of the carrier to perform the detection. In the following we review the principal demodulation schemes.

PSK Homodyne Detection

In an ideal PSK homodyne receiver, shown in Figure 5, the signal light and local oscillator have identical

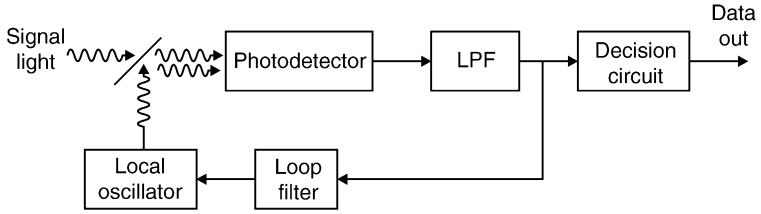


Figure 5 PSK homodyne receiver.

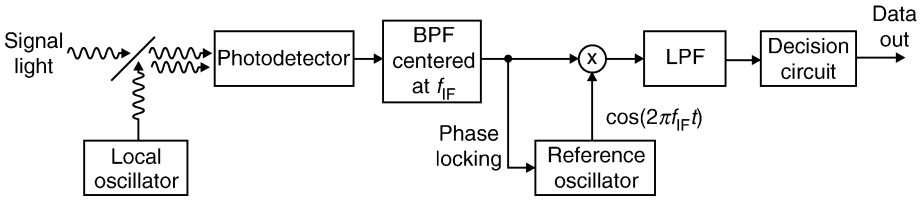


Figure 6 PSK synchronous heterodyne receiver. BPF: bandpass filter.

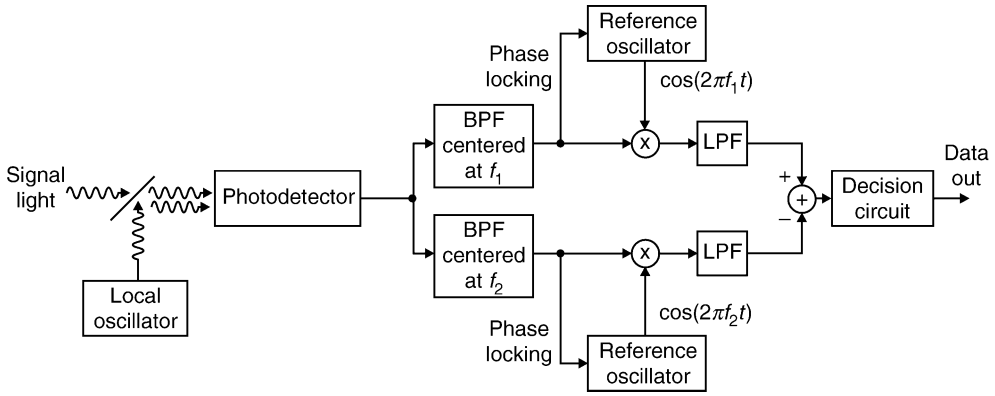


Figure 7 FSK heterodyne synchronous receiver.

optical frequency and phase. If the resulting baseband information signal is positive, then a 1 has been received. If the baseband information signal is negative, then a 0 bit has been received. This scheme has the highest theoretical sensitivity but requires an OPLL and is also highly sensitive to phase noise.

ASK Homodyne Detection

An ideal ASK homodyne receiver is identical to an ideal PSK homodyne receiver. Because the baseband information signal is either zero (0 bit) or nonzero (1 bit), the receiver sensitivity is 3 dB less than the PSK homodyne receiver.

PSK Heterodyne Synchronous Detection

Figure 6 shows an ideal PSK synchronous heterodyne receiver. The IF signal is filtered by a bandpass filter, multiplied by the phase locked reference oscillator to move the signal to baseband, low pass filtered and sent to a decision circuit that decides if the received bit is a 0 or 1. The receiver requires synchronization of the reference oscillator with the IF signal. This is not easy to achieve in practice because of the large amount of semiconductor laser phase noise.

ASK Heterodyne Synchronous Detection

This scheme is basically the same as the PSK heterodyne detection and is similarly difficult to design.

FSK Heterodyne Synchronous Detection

This scheme is shown in Figure 7 for a binary FSK signal consisting of two possible frequencies $f_1 = f_s - \Delta f$ and $f_2 = f_s + \Delta f$. There are two separate branches in which correlation with the two possible IF signals at f_1 and f_2 is performed. The two resulting signals are subtracted from each other and a decision taken on the sign of the difference. Because of the requirement for two electrical phase locked loops, the scheme is not very practical.

ASK Heterodyne Envelope Detection

In this scheme, shown in Figure 8, the IF signal is envelope detected to produce a baseband signal proportional to the original data signal. The output of the envelope detector is relatively insensitive to laser phase noise.

FSK Heterodyne Dual-Filter Detection

In this scheme, shown in Figure 9, two envelope detectors are used to demodulate the two possible IF signals at f_1 and f_2 and a decision taken based on the outputs. The scheme is tolerant to phase noise and can achieve high sensitivity. However, it needs high-bandwidth receiver electronics because a large frequency deviation is required.

FSK Heterodyne Single-Filter Detection

This scheme omits one branch of the FSK heterodyne dual-filter receiver and performs a decision in the same way as for the ASK heterodyne envelope detection. Because half the power of the detected signal is not used, the sensitivity of the scheme is 3 dB worse than the dual-filter scheme.

CPFSK Heterodyne Differential Detection

This scheme, shown in Figure 10, uses the continuous phase (CP) characteristic of an FSK signal, i.e., there is no phase discontinuity at bit transition times. After bandpass filtering, the IF photocurrent is

$$i_{\mathrm{IF}}(t) = A\cos\{2\pi[(f_{\mathrm{IF}} + a(t)\Delta f)]t\} \qquad [9]$$

where $A = R\sqrt{P_s P_{\mathrm{LO}}}$. The output from the delay-line demodulator, after low pass filtering to remove the double IF frequency term, is

$$x(t) = \frac{A^2}{2}\cos\{2\pi[f_{\mathrm{IF}} + a(t)\Delta f]\tau\} \qquad [10]$$

The function of the delay-line demodulator is to act as a frequency discriminator. We require that $x(t)$ is a maximum when $a(t) = 1$ and a minimum when $a(t) = -1$. This is the case if the following relations are satisfied:

$$2\pi(f_{\mathrm{IF}} + \Delta f)\tau = 2\pi k \qquad [11]$$

$$2\pi(f_{\mathrm{IF}} - \Delta f)\tau = (2k-1)\pi, \qquad k \text{ an integer}$$

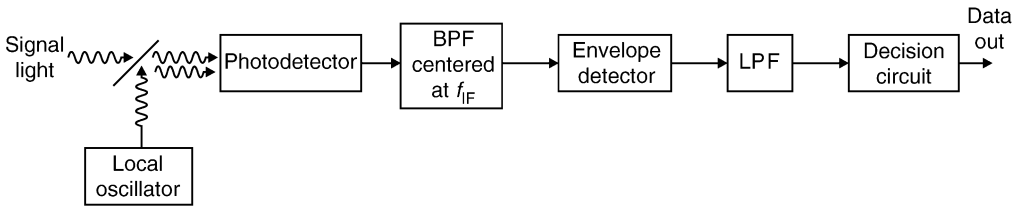


Figure 8 ASK heterodyne envelope receiver.

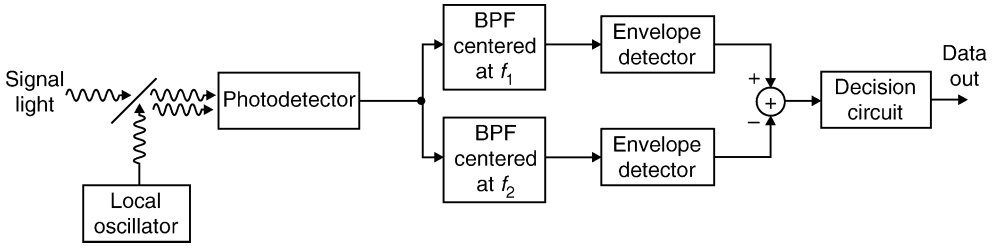


Figure 9 FSK heterodyne dual-filter receiver.

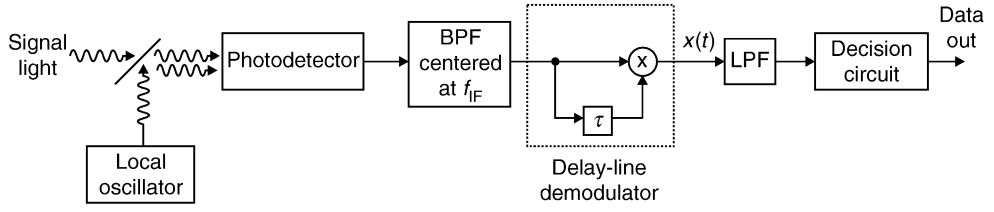


Figure 10 CPFSK heterodyne differential detection receiver.

Hence, we require that

$$\Delta f \tau = \frac{1}{4} \qquad [12]$$

$$f_{IF}\tau = \left[k - \frac{1}{4}\right], \qquad k \text{ integer}$$

In terms of the modulation index we have

$$\tau = \frac{1}{2mB_T} \qquad [13]$$

The above equation shows that it is possible to decrease the value of τ required by increasing m, thereby reducing the sensitivity of the receiver to phase noise. The minimum value of m possible is 0.5. This scheme can operate with a smaller modulation index compared to the dual-filter and single-filter schemes for a given bit rate, thereby relaxing the bandwidth requirements of the receiver electronics.

DPSK Heterodyne Differential Detection

In differential phase shift key (DPSK) modulation, a 0 is sent by changing the phase of the carrier by π with respect to the previous signal. A 1 is sent by not changing the phase. The receiver configuration is identical to the CPFSK heterodyne differential detection receiver, except that the time delay τ is equal to the inverse of the data rate. Compared to other PSK detection schemes, DPSK differential detection is relatively simple to implement and is relatively immune to phase noise.

Coherent Receiver Sensitivity

The signal-to-noise ratio γ_c of an ideal PSK homodyne receiver, in the shot-noise limit, is given by

$$\gamma_c = \frac{2RP_s}{eB_c} \qquad [14]$$

where B_c is the receiver bandwidth. The detector responsivity is given by

$$R = \frac{\eta e}{hf_s} \qquad [15]$$

where η is the detector quantum efficiency (≤ 1). The average signal power $P_s = N_s h f_s B_T$, where N_s is the average number of photons per bit. If it is assumed that $B_c = B_T$, then

$$\gamma_c = 2\eta N_s \qquad [16]$$

The actual detected number of photons per bit $N_R = \eta N_s$, so

$$\gamma_c = 2N_R \qquad [17]$$

If the noise is assumed to have a Gaussian distribution, then the bit error rate (BER) is given by

$$\text{BER} = \frac{1}{2} erfc\left(\sqrt{2N_R}\right) \qquad [18]$$

where *erfc* is the complementary error function; $erfc(x) \approx e^{-x^2}/(x\sqrt{\pi})$ for $x > 5$. The receiver sensitivity for a BER $= 10^{-9}$ is 9 photons/bit.

BER expressions and sensitivity for the demodulation schemes discussed above are listed in Table 1. The sensitivity versus N_R is plotted for some of the demodulation schemes in Figure 11. The received signal power required for a given BER is proportional to B_T. This dependency is shown in Figure 12 for some of the demodulation schemes.

Table 1 Shot-noise limited BER expressions and sensitivity for coherent demodulation schemes

Modulation scheme	*BER*	*Sensitivity (photons/bit)*
Homodyne		
PSK	$\frac{1}{2} erfc\left(\sqrt{2N_R}\right)$	9
ASK	$\frac{1}{2} erfc\left(\sqrt{N_R}\right)$	18 (peak 36)
Synchronous heterodyne		
PSK	$\frac{1}{2} erfc\left(\sqrt{N_R}\right)$	18
ASK	$\frac{1}{2} erfc\left(\sqrt{\frac{N_R}{2}}\right)$	36 (peak 72)
FSK	$\frac{1}{2} erfc\left(\sqrt{\frac{N_R}{2}}\right)$	36
Asynchronous heterodyne		
ASK (envelope)	$\frac{1}{2}\exp\left(-\frac{N_R}{2}\right)$	40 (peak 80)
FSK (dual-filter)	$\frac{1}{2}\exp\left(-\frac{N_R}{2}\right)$	40
FSK (single-filter)	$\frac{1}{2}\exp\left(-\frac{N_R}{4}\right)$	80
Differential detection		
CPFSK	$\frac{1}{2}\exp(-N_R)$	20
DPSK	$\frac{1}{2}\exp(-N_R)$	20

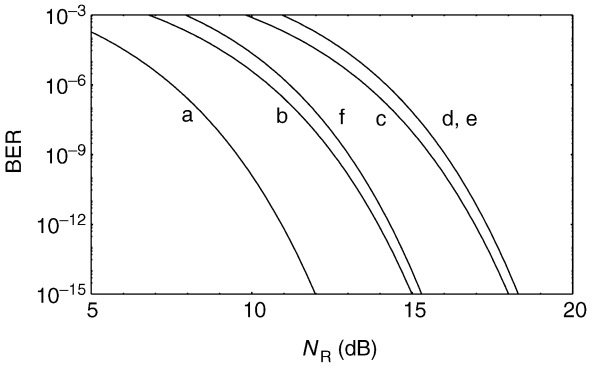


Figure 11 Shot-noise limited BER versus number of received photons/bit for various demodulation schemes: (a) PSK homodyne; (b) synchronous PSK; (c) synchronous FSK; (d) ASK envelope; (e) FSK dual-filter; and (f) CPFSK and DPSK.

Phase Noise

A major influence on the sensitivity possible with practical coherent receivers is laser phase noise, which is responsible for laser linewidth (3 dB bandwidth of the laser power spectrum). The effect of phase noise on system performance depends on the modulation and demodulation scheme used. In general, the effects of phase noise are more severe in homodyne and synchronous heterodyne schemes than in asynchronous schemes. If $\Delta\nu_{IF}$ is the beat linewidth (without modulation present) between the signal and local oscillator lasers, the normalized beat linewidth $\Delta\nu$ is defined as

$$\Delta\nu = \frac{\Delta\nu_{IF}}{B_T} \qquad [19]$$

where $\Delta\nu_{IF}$ is the IF phase noise bandwidth. Semiconductor lasers usually have a Lorentzian spectrum in which case $\Delta\nu_{IF}$ is equal to the sum of the signal and local-oscillator laser linewidths. The requirements placed on $\Delta\nu$ become less severe as the bit rate increases. Table 2 compares the $\Delta\nu$ requirement for a 1 dB power penalty at a BER of 10^{-9} for some of the principal modulation/demodulation schemes.

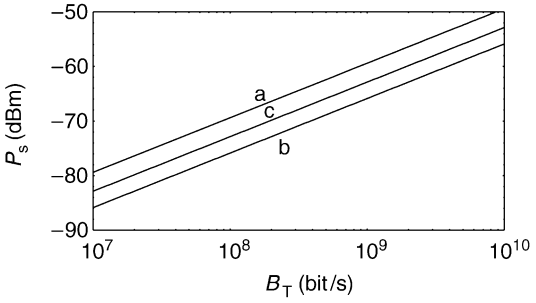


Figure 12 Required signal power versus bit rate for a BER = 10^{-9} for various shot-noise limited demodulation schemes: (a) PSK homodyne; (b) ASK envelope; and (c) CPFSK. The signal wavelength is 1550 nm and the detector is assumed to have unity quantum efficiency.

Table 2 Required normalized beat linewidth at 1 dB power penalty for a BER = 10^{-9} for various modulation/demodulation schemes

	Modulation scheme	*$\Delta\nu$ (%)*
Homodyne	PSK	6×10^{-4} (balanced loop) 0.01 (decision-driven loop)
Synchronous heterodyne	PSK	0.2–0.5
Asynchronous heterodyne	PSK	0.3–0.7
	ASK (envelope)	3
	FSK (dual-filter)	10–20
Differential detection	CPFSK	0.33 ($m = 0.5$) 0.66 ($m = 1.0$)
	DPSK	0.33

PSK Homodyne Detection – Phase Locked Loop Schemes

Practical homodyne receivers require an OPLL to lock the local-oscillator frequency and phase to that of the signal lightwave. There are a number of OPLL schemes possible for PSK homodyne detection, including the balanced PLL, Costas-type PLL and the decision-driven PLL.

The balanced PLL scheme, shown in Figure 13, requires that the PSK signal uses a phase shift of less than π between a 0 and 1, $E_s(t)$ and $E_{LO}(t)$ can be written as

$$E_s(t) = \sqrt{2P_s}\cos\left[2\pi f_0 t + a(t)\theta + \phi_{N,s}(t) + \frac{\pi}{2}\right] \qquad [20]$$

$$E_{LO}(t) = \sqrt{2P_{LO}}\cos[2\pi f_0 t + \phi_{N,LO}(t)] \qquad [21]$$

where f_0 is the common signal and local oscillator optical frequency and $\phi_{N,s}(t)$ and $\phi_{N,LO}(t)$ are the signal and local oscillator phase noise, respectively. When $\theta \neq \pi/2$, $E_s(t)$ contains an unmodulated carrier component (residual carrier) in quadrature with the information-bearing signal. $E_s(t)$ and $E_{LO}(t)$ are combined by an optical 180° hybrid. The output optical signals from the π-hybrid are given by

$$E_1(t) = \frac{1}{\sqrt{2}}[E_s(t) + E_{LO}(t)] \qquad [22]$$

$$E_2(t) = \frac{1}{\sqrt{2}}[E_s(t) - E_{LO}(t)] \qquad [23]$$

The voltage at the output of the summing circuit is given by

$$v_L(t) = 2RR_L\sqrt{P_s P_{LO}}\{a(t)\sin\theta\cos[\phi_e(t)] + \cos\theta\sin[\phi_e(t)]\} \qquad [24]$$

where the photodetectors have the same responsivity and load resistance R_L. The phase error is given by

$$\phi_e(t) = \phi_{N,s}(t) - \phi_{N,LO}(t) - \phi_c(t) \qquad [25]$$

$\phi_c(t)$ is the controlled phase determined by the control signal $v_c(t)$ at the output of the loop filter. v_L contains two terms; an information-bearing signal proportional to sin θ, which is processed by the data detection circuit, and a phase error signal proportional to cos θ used by the PLL for locking. The power penalty, due to the residual carrier transmission, is $10\log_{10}(1/\sin^2\theta)$ dB.

This scheme is not practical because the $\Delta\nu$ needed for a power penalty of <1 dB is typically less than 10^{-5}. This would require that $\Delta\nu_{IF} < 10$ kHz for

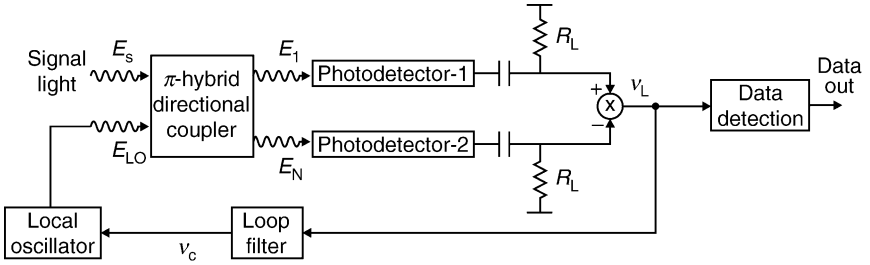


Figure 13 PSK homodyne balanced PLL receiver.

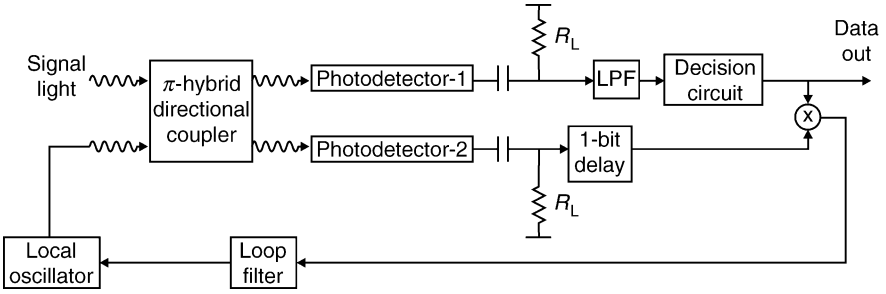


Figure 14 PSK decision-driven PLL.

a data rate of 1 Gb/s. The narrow signal and local-oscillator linewidths required are only achievable with external cavity semiconductor lasers.

The Costas-type and decision-driven PLLs also utilize π-hybrids and two photodetectors but as they have full suppression of the carrier ($\theta = \pi/2$), nonlinear processing of the detected signals is required to produce a phase error signal that can be used by the PLL for locking. In the decision-driven PLL, shown in Figure 14, the signal from one of the photodetector outputs is sent to a decision circuit and the output of the circuit is multiplied by the signal from the second photodetector. The mixer output is sent back to the local oscillator laser for phase locking. The $\Delta\nu$ required for a power penalty of <1 dB is typically less than 2×10^{-4}. This is superior to both the balanced PLL and Costas-type PLL receivers.

Heterodyne Phase Locking

In practical synchronous heterodyne receivers phase locking is performed in the electronic domain utilizing techniques from radio engineering. In the PSK case, electronic analogs of the OPLL schemes used in homodyne detection can be used. Synchronous heterodyne schemes have better immunity to phase noise than the homodyne schemes.

Asynchronous Systems

Asynchronous receivers do not require optical or electronic phase locking. This means that they are less sensitive to phase noise than synchronous receivers. They do have some sensitivity to phase noise because IF filtering leads to phase-to-amplitude noise conversion.

Differential Detection

Differential detection receivers have sensitivities to phase noise between synchronous and asynchronous receivers because, while they do not require phase locking, they use phase information in the received signal.

Phase-Diversity Receivers

Asynchronous heterodyne receivers are relatively insensitive to phase noise but require a much higher receiver bandwidth for a given bit rate. Homodyne receivers only require a receiver bandwidth equal to the detected signal bandwidth but require an OPLL, which is difficult to implement. The phase diversity receiver, shown in Figure 15, is an asynchronous homodyne scheme, which does not require the use of an OPLL and has a bandwidth approximately equal to synchronous homodyne detection. This is at the expense of increased receiver complexity and reduced sensitivity compared to synchronous homodyne detection. The phase-diversity scheme can be used with ASK, DPSK, and CPFSK modulation. ASK modulation uses an squarer circuit for the demodulator component, while DPSK and CPFSK modulation use a delay line and mixer.

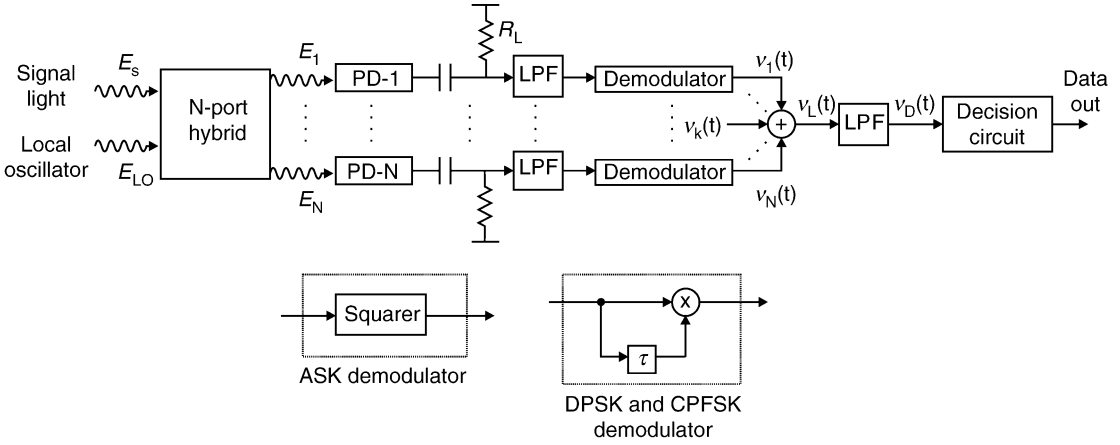


Figure 15 Multiport phase-diversity receiver. PD: photodetector.

In the case of ASK modulation, the outputs of the N-port hybrid can be written as

$$E_k(t) = \frac{1}{\sqrt{N}}\left\{a(t)\sqrt{2P_s}\cos\left[2\pi f_s t + \phi_s(t) + \frac{2\pi k}{N}\right] + \sqrt{2P_{LO}}\cos[2\pi f_{LO}t + \phi_{LO}(t)]\right\}, \quad k = 1,2,\ldots,N,\ N \geq 3 \qquad [26]$$

For $N = 2$, the phase term is $\pi/2$ for $k = 2$. The voltage input to the k-th demodulator (squarer), after the dc terms have been removed by the blocking capacitor, is given by

$$v_k(t) = a(t)\frac{RR_L}{N}2\sqrt{P_s P_{LO}}\cos\left[2\pi(f_s - f_{LO}) + \phi_s(t) - \phi_{LO}(t) + \frac{2\pi k}{N}\right] \qquad [27]$$

These voltages are then squared and added to give an output voltage

$$v_L(t) = \alpha\left(\frac{RR_L}{N}\right)^2 4P_s P_{LO}a(t)\sum_{k=1}^{N}\left\{1 + \cos\left[4\pi \times \left(f_s - f_{LO}\right)t + 2\phi_s(t) - 2\phi_{LO}(t) + \frac{4\pi k}{N}\right]\right\} \qquad [28]$$

where α is the squarer parameter. The lowpass filter at the summing circuit output effectively integrates $v_L(t)$ over a bit period. If $2(f_s - f_{LO}) \ll B_T$ and the difference between the signal and local oscillator phase noise is small within a bit period, then the input voltage to the decision circuit is

$$v_D(t) = \alpha\frac{(RR_L)^2}{N}4P_s P_{LO}a(t) \qquad [29]$$

which is proportional to the original data signal. In shot-noise limited operation, the receiver sensitivity is 3 dB worse than the ASK homodyne case. The scheme is very tolerant of laser phase noise. The $\Delta\nu$ for a power penalty of < 1 dB is of the same order as for ASK heterodyne envelope detection.

Polarization

The mixing efficiency between the signal and local-oscillator lightwaves is a maximum when their polarization states are identical. In practice the polarization state of the signal arriving at the receiver is unknown and changes slowly with time. This means that the IF photocurrent can change with time and in the worst case, when the polarization states of the signal and local-oscillator are orthogonal, the IF photocurrent will be zero. There are a number of possible solutions to this problem.

1. Polarization maintaining (PM) fiber can be used in the optical link to keep the signal polarization from changing. It is then a simple matter to adjust the local-oscillator polarization to achieve optimum mixing. However, the losses associated with PM fiber are greater than for conventional single-mode (SM) fiber. In addition most installed fiber is SM fiber.
2. An active polarization controller, such as a fiber coil using bending-induced birefringence, can be used to ensure that the local-oscillator polarization tracks the signal polarization.
3. A polarization scrambler can be used to scramble the polarization of the transmitted signal at a speed greater than the bit rate. The sensitivity degradation due to the polarization scrambling is 3 dB compared to a receiver with perfect polarization matching. However, this technique is only feasible at low bit rates.
4. The most general technique is to use a polarization-diversity receiver, as shown in Figure 16. In this scheme, the signal light and local-oscillator

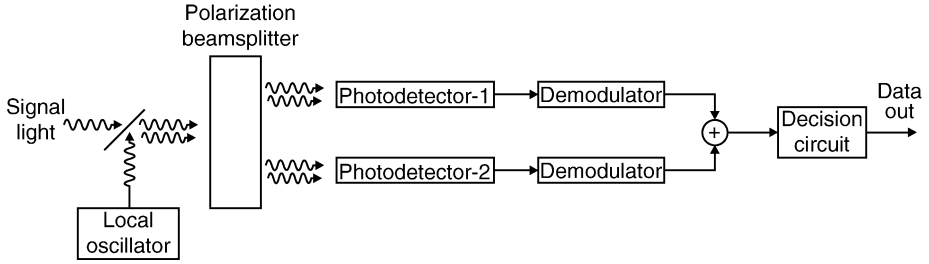


Figure 16 Polarization-diversity receiver.

Table 3 General comparisons between various modulation/demodulation schemes. The sensitivity penalty is relative to ideal homodyne PSK detection

Modulation/demodulation scheme	*Sensitivity penalty (dB)*	*Immunity to phase noise*	*IF bandwidth/bit rate*	*Complexity*
Homodyne PSK	0	Very poor	1	Requires an OPLL
Synchronous heterodyne PSK	3	Poor	3–6	Requires electronic phase locking
Asynchronous heterodyne FSK (dual-filter)	6.5	Excellent	3–6	High bandwidth IF circuits and two envelope detectors required
Differential detection CPFSK	3.5	Moderate	≈2	Relatively simple
DPSK	3.5	Moderate	≈2	Relatively simple

are split into orthogonally polarized lightwaves by a polarization beamsplitter. The orthogonal components of the signal and local-oscillator are separately demodulated and combined. The resulting decision signal is polarization independent. The demodulator used depends on the modulation format.

Comparisons Between the Principal Demodulation Schemes

General comparisons between the demodulation schemes discussed above are given in Table 3. FSK modulation is relatively simple to achieve. FSK demodulators have good receiver sensitivity and low immunity to phase noise. PSK modulation is also easy to achieve and gives good receiver sensitivity. ASK modulation has no real merit compared to FSK or PSK.

In general, homodyne schemes have the best theoretical sensitivity but require optical phase locking and are very sensitive to phase noise. Phase diversity receivers are a good alternative to homodyne receivers, in that they do not require an OPLL. Asynchronous heterodyne receivers have similar sensitivities to synchronous receivers but are much less complex. However, they require an IF bandwidth much greater than the data rate.

Differential detection schemes have good receiver sensitivity, do not require wide bandwidth IF circuits and have moderate immunity to phase noise. The advantage of CPFSK over DPSK is that the immunity to phase noise can be increased by increasing the FSK modulation index.

Polarization diversity receivers can be used with any modulation/demodulation scheme and offer the most general approach to overcoming polarization effects.

Table 4 Coherent systems experiments. The penalty is with respect to the theoretical sensitivity

Receiver	*Bit rate (Mb/s)*	*Sensitivity (dBm)*		*Penalty (dB)*
		dBm	*Photons/bit*	
ASK	400	−47.2	365	9.7
FSK (dual-filter)	680	−39.1	1600	16
CPFSK ($m = 0.5$)	1000	−37	1500	18.6
differential detection	4000	−31.3	1445	18.6
DPSK	400	−53.3	45.1	3.5
	2000	−39.0	480	13.8

System Experiments

Most coherent communication system experiments date from the 1980s and early 1990s and were based on the technology available at that time. This restricted the bit rates possible to typically <4 Gb/s. Some coherent systems experimental results obtained around this time are listed in Table 4. All of the

listed experimental trials used lasers (usually external cavity lasers) with negligible phase noise in the 1550 nm region. It can be seen that the measured sensitivities deteriorate at higher bit rates. This is attributable to the nonuniform response and bandwidth limitations of the modulator, photodetector, and demodulator. There is no doubt that the performance could be improved by the use of current optoelectronic technologies.

Conclusion

Coherent optical communication systems offer the potential for increased sensitivity and selectivity compared to conventional IM/DD systems. However, coherent optical receivers are more complex and have more stringent design parameters compared to IM/DD receivers. Moreover, advances in optoelectronic component design and fabrication along with the advent of reliable optical amplifiers has meant that coherent systems are not a viable alternative to IM/DD systems in the vast majority of optical communication systems.

List of Units and Nomenclature

e	Electronic charge
h	Planck constant

See also

Modulators: Modulation and Demodulation of Optical Signals. **Optical Amplifiers:** Basic Concepts; Erbrium Doped Fiber Amplifiers for Lightwave Systems; Optical Amplifiers in Long-Haul Transmission Systems. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Lightwave Transmitters; Wavelength Division Multiplexing. **Polarization:** Introduction.

Further Reading

Barry JR and Lee EA (1990) Performance of coherent optical receivers. *IEEE Proceedings* 78: 1369–1394.

Cvijetic M (1996) *Coherent and Nonlinear Lightwave Communications*. Boston, MA: Artech House.

Kazovsky LG (1989) Phase- and polarization-diversity coherent optical techniques. *Journal of Lightwave Technology* 7: 279–292.

Kazovsky L, Benedetto S and Willner A (1996) *Optical Fiber Communication Systems*. Boston, MA: Artech House.

Linke RA and Gnauck AH (1988) High-capacity coherent lightwave systems. *Journal of Lightwave Technology* 6: 1750–1769.

Ryu S (1995) *Coherent Lightwave Systems*. Boston, MA: Artech House.

COHERENT TRANSIENTS

Contents

Coherent Transient Spectroscopy in Atomic and Molecular Vapors
Foundations of Coherent Transients in Semiconductors
Ultrafast Studies of Semiconductors

Coherent Transient Spectroscopy in Atomic and Molecular Vapors

P R Berman, University of Michigan, Ann Arbor, MI, USA

R G Brewer, Stanford University, Stanford, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical spectroscopy is a method that allows one to determine atomic and molecular transition frequencies, as well as relaxation rates in atomic and molecular vapors. The transition frequencies serve as 'fingerprints' that can be used to identify and classify atoms and molecules. Optical spectroscopic methods fall into two broad categories, continuous wave (cw) or stationary spectroscopy and time-dependent or transient spectroscopy. In cw spectroscopy, the absorption or emission line shapes are measured as a function of the frequency of a probe field. On the basis of theoretical line shape formulas, it is then possible to extract the relevant transition frequencies and relaxation rates from the line shapes. In transient optical spectroscopy, pulsed optical fields are used to create atomic state populations or coherences between atomic states. Following the

excitation, the time-evolution of the atoms is monitored, from which transition frequencies and relaxation rates can be obtained.

The earliest coherent transient effects were observed with atomic nuclear spins in about 1950 by H C Torrey, who discovered transient spin nutation, and by E L Hahn who detected spin echoes and free induction decay (FID). The optical analogs of these and other nuclear magnetic resonance (NMR) effects have now been detected using either optical pulses or techniques involving Stark or laser frequency switching.

In this brief introduction to coherent optical transients, several illustrative examples of coherent transient phenomena are reviewed. The optical Bloch equations are derived and coupled to Maxwell's Equations. The Maxwell–Bloch formalism is used to analyze free precession decay, photon echoes, stimulated photon echoes, and optical Ramsey fringes. Experimental results are presented along with the relevant theoretical calculations. In addition, coherent transient phenomena involving quantized matter waves are discussed. Although the examples considered in this chapter are fairly simple, it should be appreciated that sophisticated coherent transient techniques are routinely applied as a probe of complex structures such as liquids and solids or as a probe of single atoms and molecules.

Optical Bloch Equations

Many of the important features of coherent optical transients can be illustrated by considering the interaction of a radiation field with a two-level atom. The lower state $|1\rangle$ has energy $-\hbar\omega_0/2$ and upper state $|2\rangle$ has energy $\hbar\omega_0/2$. For the moment, the atom is assumed to be fixed at $\mathbf{R} = 0$ and all relaxation processes are neglected. The incident electric field is

$$\mathbf{E}(\mathbf{R} = 0, t) = \frac{1}{2}\boldsymbol{\varepsilon}[E(t)e^{-i\omega t} + E(t)e^{i\omega t}] \qquad [1]$$

where $E(t)$ is the field amplitude, $\boldsymbol{\varepsilon}$ is the field polarization and ω is the carrier frequency. The time dependence of $E(t)$ allows one to consider pulses having arbitrary shape. A time-dependent phase for the field could have been included, allowing one to consider the effect of arbitrary frequency 'chirps,' but such effects are not included in the present discussion. It is convenient to expand the atomic state wave function in a field interaction representation as

$$|\psi(t)\rangle = c_1(t)e^{i\omega t/2}|1\rangle + c_2(t)e^{-i\omega t/2}|2\rangle \qquad [2]$$

The atom-field interaction potential is

$$V(\mathbf{R}, t) = -\mu \cdot \mathbf{E}(\mathbf{R}, t) \qquad [3]$$

where μ is a dipole moment operator. When eqn [2] is substituted into Schrödinger's equation and rapidly varying terms are neglected (rotating-wave approximation), one finds that the state amplitudes evolve according to

$$i\hbar\frac{\mathrm{d}\mathbf{c}}{\mathrm{d}t} = \tilde{\mathbf{H}}\mathbf{c}; \qquad \tilde{\mathbf{H}} = (\hbar/2)\begin{pmatrix} -\delta_0 & \Omega_0(t) \\ \Omega_0(t) & \delta_0 \end{pmatrix} \qquad [4]$$

where $\mathbf{c}$ is a vector having components (c_1, c_2)

$$\delta_0 = \omega_0 - \omega \qquad [5]$$

is an atom-field detuning

$$\Omega_0(t) = -\frac{\mu E(t)}{\hbar} = -\frac{\mu}{\hbar}\sqrt{\frac{2S(t)}{\varepsilon_0 c}} \qquad [6]$$

is a Rabi frequency, $\mu = \langle 1|\mu\cdot\varepsilon|2\rangle = \langle 2|\mu\cdot\varepsilon|1\rangle$ is a dipole moment matrix element, ε_0 is the permittivity of free space, and $\mathbf{S}(t)$ is the time-averaged Poynting vector of the field. Equation [4] can be solved numerically for arbitrary pulse envelopes.

Expectation values of physical observables are conveniently expressed in terms of density matrix elements defined by

$$\rho_{ij} = c_i c_j^* \qquad [7]$$

which obey equations of motion

$$\begin{aligned} \dot{\rho}_{11} &= -i\Omega_0(t)(\rho_{21} - \rho_{12})/2 \\ \dot{\rho}_{22} &= i\Omega_0(t)(\rho_{21} - \rho_{12})/2 \\ \dot{\rho}_{12} &= -i\Omega_0(t)(\rho_{22} - \rho_{11})/2 + i\delta_0\rho_{12} \\ \dot{\rho}_{21} &= i\Omega_0(t)(\rho_{22} - \rho_{11})/2 - i\delta_0\rho_{21} \end{aligned} \qquad [8]$$

An alternative set of equations in terms of real variables can be obtained if one defines new parameters

$$\begin{aligned} u &= \rho_{12} + \rho_{21}; & \rho_{12} &= (u + iv)/2 \\ v &= i(\rho_{21} - \rho_{12}); & \rho_{21} &= (u - iv)/2 \\ w &= \rho_{22} - \rho_{11}; & \rho_{22} &= (m + w)/2 \\ m &= \rho_{11} + \rho_{22}; & \rho_{11} &= (m - w)/2 \end{aligned} \qquad [9]$$

which obey

$$\begin{aligned} \dot{u} &= -\delta_0 v \\ \dot{v} &= \delta_0 u - \Omega_0(t)w \\ \dot{w} &= \Omega_0(t)v \\ \dot{m} &= 0 \end{aligned} \qquad [10]$$

The last of these equations reflects the fact that $\rho_{11} + \rho_{22} = 1$, while the first three can be rewritten as

$$\dot{\mathbf{U}} = \mathbf{\Omega}(t) \times \mathbf{U} \qquad [11]$$

where the *Bloch vector* $\mathbf{U}$ has components (u, v, w) and the *pseudofield vector* $\mathbf{\Omega}(t)$ has components $[\Omega_0(t), 0, \delta_0]$.

Equations [8] or [10] constitute the optical Bloch equations without decay. The vector $\mathbf{U}$ has unit magnitude and traces out a path on the Bloch sphere as it precesses about the pseudofield vector. The component w is the population difference of the two atomic states, while u and v are related to the quadrature components of the atomic polarization (see below).

Equations [8] and [10] can be generalized to include relaxation. A simplified relaxation scheme that has a wide range of applicability is one in which state 2 decays spontaneously to state 1 with rate γ_2, while the relaxation of the coherence ρ_{12} is characterized by a complex decay parameter Γ_{12}, resulting from phase-changing collisions with a background gas. When such processes are included in eqn [8], they are modified as

$$\dot{\rho}_{11} = -i\Omega_0(t)(\rho_{21} - \rho_{12})/2 + \gamma_2\rho_{22} \qquad [12a]$$

$$\dot{\rho}_{22} = i\Omega_0(t)(\rho_{21} - \rho_{12})/2 - \gamma_2\rho_{22} \qquad [12b]$$

$$\dot{\rho}_{12} = -i\Omega_0(t)(\rho_{22} - \rho_{11})/2 - \gamma\rho_{12} + i\delta\rho_{12} \qquad [12c]$$

$$\dot{\rho}_{21} = i\Omega_0(t)(\rho_{22} - \rho_{11})/2 - \gamma\rho_{21} - i\delta\rho_{21} \qquad [12d]$$

where $\gamma = \gamma_2/2 + \mathrm{Re}(\Gamma_{12})$ and the detuning $\delta = \delta_0 - \mathrm{Im}(\Gamma_{12})$ is modified to include the collisional shift. With the addition of decay, the length of the Bloch vector is no longer conserved. The quantity γ_2 is referred to as the longitudinal relaxation rate. Moreover, one usually refers to $T_1 = \gamma_2^{-1}$ as the longitudinal relaxation time and $T_2 = \gamma^{-1}$ as the transverse relaxation time. In the case of purely radiative broadening, $\gamma_2 = 2\gamma$ and $T_1 = T_2/2$.

Maxwell–Bloch Equations

As a result of atom-field interactions, it is assumed that a polarization is created in the medium of the form

$$\mathbf{P}(\mathbf{R}, t) = \frac{1}{2}\boldsymbol{\varepsilon}\left[P(Z, t)e^{i[kZ-\omega t]} + P^*(Z, t)e^{-i[kZ-\omega t]}\right] \qquad [13]$$

which gives rise to a signal electric field of the form

$$\mathbf{E}_s(\mathbf{R}, t) = \frac{1}{2}\boldsymbol{\varepsilon}\left[E_s(Z, t)e^{i[kZ-\omega t]} + E_s^*(Z, t)e^{-i[kZ-\omega t]}\right] \qquad [14]$$

The z-axis has been chosen in the direction of $\mathbf{k}$. It follows from Maxwell's equations that the quasi-steady-state field amplitude is related to the polarization by

$$\frac{\partial E_s(Z, t)}{\partial Z} = \frac{ik}{2\varepsilon_0}P(Z, t) \qquad [15]$$

To arrive at eqn [15], it is necessary to assume that transverse field variations can be neglected, that the *phase matching* condition $k = \omega/c$ is met, and that the complex field amplitudes $P(Z, t)$ and $E_s(Z, t)$ vary slowly in space compared with e^{ikZ} and slowly in time compared with $e^{i\omega t}$.

The polarization $\mathbf{P}(\mathbf{R}, t)$, defined as the average dipole moment per unit volume, is given by the expression

$$\mathbf{P}(\mathbf{R}, t) = \mathcal{N}\Big[\boldsymbol{\mu}_{21}\langle\rho_{12}(Z, t)\rangle e^{-i[kZ-\omega t]} + \boldsymbol{\mu}_{12}\langle\rho_{21}(Z, t)\rangle e^{i[kZ-\omega t]}\Big] \qquad [16]$$

where $\mathcal{N}$ is the atomic density and $\rho_{12}(Z, t)$ and $\rho_{21}(Z, t)$ are single-particle density matrix elements that depend on the atomic properties of the medium. As such, eqn [16] provides the link between Maxwell's equations and the optical Bloch equations. The $\langle\ \rangle$ brackets in eqn [16] indicate that there may be additional averages that must be carried out. For example, in a vapor, there is a distribution of atomic velocities that must be summed over, while, in a solid, there may be an inhomogeneous distribution of atomic frequencies owing to local strains in the media. By combining eqns [9], [13], [15] and [16], and using the fact that $\mu = \mu_{12}\cdot\varepsilon = \mu_{21}\cdot\varepsilon$, one arrives at

$$\frac{\partial E_s(Z, t)}{\partial Z} = \frac{ik\mathcal{N}\mu}{\varepsilon_0}\langle\rho_{21}(Z, t)\rangle = \frac{ik\mathcal{N}\mu}{2\varepsilon_0}\langle u(Z, t) - iv(Z, t)\rangle \qquad [17]$$

which, together with eqn [12], constitute the Maxwell–Bloch equations. From eqn [17], it is clearly seen that the coherence $\langle\rho_{21}(Z, t)\rangle$ drives the signal field. If $\langle\rho_{21}(Z, t)\rangle$ is independent of Z, the intensity of radiation exiting a sample of length L is proportional to

$$I(L, t) = \frac{k\mathcal{N}\mu L}{\varepsilon_0}^2|\langle\rho_{21}(t)\rangle|^2 \qquad [18]$$

Alternatively, the outgoing field can be heterodyned with a reference field to extract the field amplitude rather than the field intensity.

Coherent Transient Signals

The Maxwell–Bloch equations can be used to describe the evolution of coherent transient signals. In a typical experiment, one applies one or more 'short' radiation pulses to an atomic sample and monitors the radiation field emitted by the atoms following the pulses. A 'short' pulse, is one satisfying

$$|\delta|\tau, ku\tau, \gamma\tau, \gamma_2\tau \ll 1 \qquad [19]$$

where τ is the pulse duration. Conditions [19] allow one to neglect any effects of detuning (including Doppler shifts or other types of inhomogeneous broadening) or relaxation during the pulse's action. The evolution of the atomic density matrix during each pulse is determined solely by the pulse area defined by

$$\theta = \int_{-\infty}^{\infty} \Omega_0(t)\mathrm{d}t \qquad [20]$$

In terms of density matrix elements in a standard interaction representation defined by

$$\rho_{12}^I = \rho_{12}e^{-i[kZ+\delta t]}, \quad \rho_{21}^I = \rho_{21}e^{i[kZ+\delta t]},$$
$$\rho_{22}^I = \rho_{22}, \quad \rho_{11}^I = \rho_{11}$$

the change in density matrix elements for an interaction occurring at time T_j is given by

$$\begin{pmatrix} \rho_{11}^I \\ \rho_{22}^I \\ \rho_{12}^I \\ \rho_{21}^I \end{pmatrix}^+ = \frac{1}{2}\begin{pmatrix} 1+\cos\theta_j & 1-\cos\theta_j & i\sin\theta_j e^{i\Phi(T_j)} & -i\sin\theta_j e^{-i\Phi(T_j)} \\ 1-\cos\theta_j & 1+\cos\theta_j & -i\sin\theta_j e^{i\Phi(T_j)} & i\sin\theta_j e^{-i\Phi(T_j)} \\ i\sin\theta_j e^{-i\Phi(T_j)} & -i\sin\theta_j e^{-i\Phi(T_j)} & 1+\cos\theta_j & (1-\cos\theta_j)e^{-2i\Phi(T_j)} \\ -i\sin\theta_j e^{i\Phi(T_j)} & i\sin\theta_j e^{i\Phi(T_j)} & (1-\cos\theta_j)e^{2i\Phi(T_j)} & 1+\cos\theta_j \end{pmatrix}\begin{pmatrix} \rho_{11}^I \\ \rho_{22}^I \\ \rho_{12}^I \\ \rho_{21}^I \end{pmatrix}^- \qquad [21]$$

where $\mp$ superscripts refer to times just before and after the pulse

$$\Phi(T_j) = \mathbf{k}\cdot\mathbf{R}(T_j) + \delta T_j$$

$$\mathbf{R}(T_j) = \mathbf{R} - \mathbf{v}(t - T_j)$$

θ_j is the pulse area, and $\mathbf{v}$ is the atomic velocity. Between pulses, or following the last pulse, the atoms evolve freely in the absence of any applied fields. The density matrix elements evolve as

$$\dot\rho_{11}^I = \gamma_2\rho_{22}^I; \quad \dot\rho_{22}^I = -\gamma_2\rho_{22}^I;$$
$$\dot\rho_{12}^I = -\gamma\rho_{12}^I; \quad \dot\rho_{21}^I = -\gamma\rho_{21}^I \qquad [22]$$

A coherent transient signal is constructed by piecing together field interaction zones and free evolution periods. There are many different types of coherent transient signals. A few illustrative examples are given below.

Free Polarization Decay

Free polarization decay (FPD), also referred to as optical FID, is the optical analog of free induction decay (FID) in NMR. This transient follows preparation of the atomic sample by any of three methods: (1) an optical pulse resonant with the atomic transition; (2) a Stark pulse that switches the transition frequency of a molecule (atom) into resonance with an incident *cw* laser beam; or (3) an electronic pulse that switches the frequency of a *cw* laser into coincidence with the molecular (atomic) transition frequency. In methods (2) and (3), the FPD emission propagates coherently in the forward direction along with the *cw* laser beam, automatically generating a heterodyne beat signal at a detector that can be orders of magnitude larger than the FPD signal. The FPD field intensity itself exiting the sample is equal to

$$I(L,t) = \left(\frac{k\mathcal{N}\mu L}{\varepsilon_0}\right)^2 |\langle\rho_{21}(t)\rangle|^2$$
$$= \left(\frac{k\mathcal{N}\mu L}{2\varepsilon_0}\sin\theta\right)^2 \left|\int \mathrm{d}\mathbf{v}\, W_0(\mathbf{v}) e^{-[\gamma - i(\delta+\mathbf{k}\cdot\mathbf{v})]t}\right|^2 \qquad [23]$$

For a Maxwellian velocity distribution

$$W_0(\mathbf{v}) = \frac{1}{(\pi u^2)^{3/2}} e^{-(v/u)^2} \qquad [24]$$

where u is the most probable atomic speed, one finds

$$I(L,t) = \left(\frac{k\mathcal{N}\mu L}{2\varepsilon_0}\sin\theta\right)^2 e^{-2\gamma t} e^{-(kut)^2/2} \qquad [25]$$

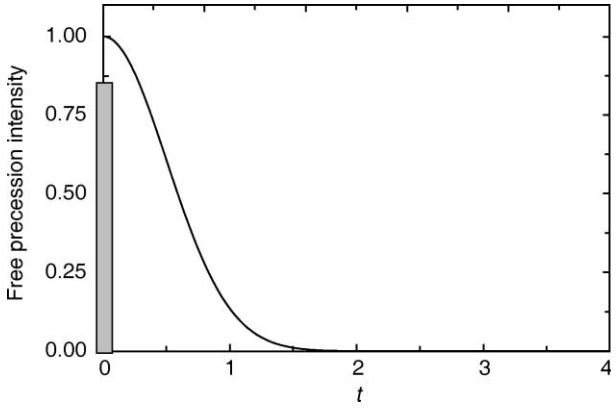


Figure 1 Pulse sequence and signal intensity for free precession decay. In this and subsequent figures, time is measured in units of $2/ku$.

If $ku \gg \gamma$, the signal decays mainly owing to inhomogeneous broadening, as shown in Figure 1. Owing to different Doppler shifts for atoms having different velocities, the optical dipoles created by the pulse lose their relative phase in a time of order $T_2^* = 2/ku$. The FPD signal can be used to measure T_2^*, which can be viewed as an inhomogeneous, transverse relaxation time. At room temperature ku/γ is typically of order of 100.

When FPD signals are observed using laser frequency switching, method 3 above, the *cw* field can modify the molecular velocity distribution, exciting only those atoms having $\mathbf{k}\cdot\mathbf{v} = -\delta \pm \gamma'$, where γ' is a power broadened homogeneous width. In the linear field regime, the FPD signal again decays on a time scale of order $(ku)^{-1}$, since the velocity distribution is unchanged to first order in the field amplitude. When one considers nonlinear interactions with the field, however, the *cw* field excites a narrow velocity subclass that is no longer subject to inhomogeneous broadening. These atoms give rise to a contribution to the FPD signal that decays with rate $T_2^{-1} = (\gamma + \gamma')$. Thus by using nonlinear atom-field interactions, one can extract homogeneous decay rates in situations where there is large inhomogeneous broadening. The price one pays is that only a small percentage of the atoms (those having velocities for which the applied field is resonant) contribute to the signal.

A FPD signal obtained on the D_2 transition in Cs is shown in Figure 2, where $T_2^* = 1.4$ ns. Oscillations in the signal originate from ground state hyperfine splitting. Figure 3 represents a FPD signal obtained in NH_2D at 10.6 μm using the method of Stark switching with *cw* state preparation. The oscillations are the heterodyne beat signal while the slowly varying increase in the signal is the result of optical nutation of molecules switched into resonance by the

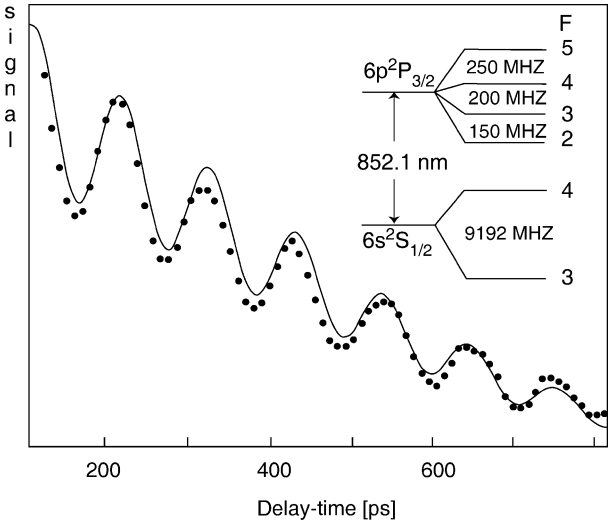


Figure 2 Optical free precession decay in Cs using an excitation pulse having 20 ns duration. Reproduced with permission from Lehmitz H and Harde H (1986) In: Prior Y, Ben-Reuven A and Rosenbluh M (eds) *Methods of Laser Spectroscopy*, pp. 109–112. New York: Plenum.

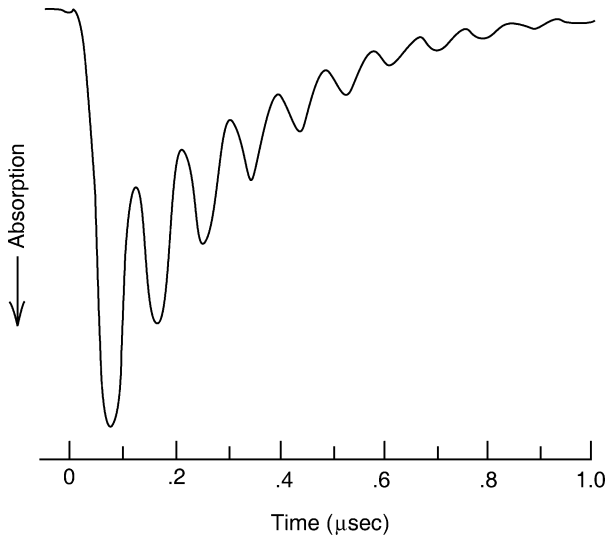


Figure 3 Optical free precession in NH_2D using the Stark switching technique. Reproduced with permission from Brewer RG and Shoemaker RL (1972) Optical free induction decay. *Physical Review A* 6: 2001. Copyright (1972) by the American Physical Society.

Stark pulse. The FPD signal manifests itself as a reduction of the amplitude of the oscillation with time. This amplitude decays with the (power-broadened) homogeneous decay rate γ'.

Photon Echo

Although the FPD signal produced by short excitation pulses decays in a time of order $(ku)^{-1}$, the coherence of individual atoms decays in a much longer time, T_2.

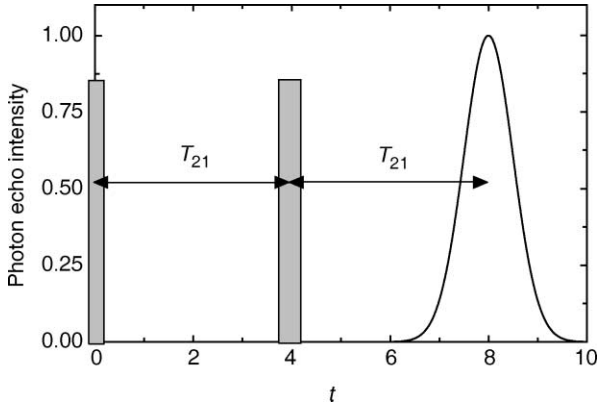


Figure 4 Pulse sequence and signal intensity for the photon echo.

The question arises as to whether it is possible to bring all the atomic dipoles back into phase so they can radiate a coherent signal. The photon echo accomplishes this goal, although it has very little to do with either photons or echoes. The pulse sequence and echo signal are shown in Figure 4. The first pulse creates a phased array of dipoles, which begin to dephase owing to inhomogeneous broadening. At time T_{21} the second pulse provides a nonlinear interaction that couples ρ_{12} and ρ_{21}. As a result of the second pulse, the dipoles begin a rephasing process that is completed at time $t = 2T_{21}$. The 'echo' appears at this time. The echo intensity exiting the sample is

$$\begin{aligned} I(L,t) &= \left(\frac{k\mathcal{N}\mu L}{\varepsilon_0}\right)^2 |\langle\rho_{21}(t)\rangle|^2 \\ &= \left(\frac{k\mathcal{N}\mu L}{2\varepsilon_0}\sin\theta_1 \sin^2(\theta_2/2)\right)^2 \\ &\quad \times e^{-4\gamma T_{21}} e^{-k^2u^2(t-2T_{21})^2/2} \end{aligned} \qquad [26]$$

If the echo intensity is measured as a function of T_{21}, one can extract the homogeneous decay rate $\gamma = T_{21}^{-1}$. It is interesting to note that the echo intensity near $t = 2T_{21}$ mirrors the FPD intensity immediately following the first pulse. For experimental reasons, it is often convenient to use a different propagation vector for the second pulse. If $\mathbf{k}_1$ and $\mathbf{k}_2$ are propagation vectors of the first and second pulses, the echo signal is in the direction $\mathbf{k} = 2\mathbf{k}_2 - \mathbf{k}_1$, provided $|\mathbf{k}_2 - \mathbf{k}_1|/k_1 \ll 1$.

The echo signal is sensitive only to $\rho_{21}(t)$ and $\rho_{12}(t)$ in the entire time interval of interest. Any interaction that results in a degradation of the dephasing–rephasing process for these density matrix elements leads to a decrease in the echo intensity. As such, echo signals can serve as a probe of transverse relaxation. Transverse relaxation generally falls into two broad categories. First, there are dephasing processes which produce an exponential damping of the coherences and contribute to γ. Second there is *spectral diffusion*; in a vapor, spectral diffusion is produced by velocity-changing collisions that change the effective field frequency seen by the atoms owing to the Doppler effect. Such terms enter the optical Bloch equations as *integral* terms, transforming the equations into differentio-integral equations. In general, the phase-changing and velocity-changing aspects of collisions are entangled; however, if the collisional interaction is *state-independent*, as it is for some molecular transitions, then collisions are purely velocity-changing in nature, leading to an echo that decays exponentially as T_{21}^3 for early times and T_{21} for later times. For electronic transitions, collisions are mainly phase-changing in nature, but there is a velocity-changing contribution that persists in the forward diffractive scattering cone. Photon echo techniques are sufficiently sensitive to measure the relaxation caused by such diffractive scattering.

The photon echo was first observed by NA Kurnit and co-workers, using a pulsed ruby laser (Figure 5). In Stark switching experiments, the photon echo follows two Stark pulses as shown in Figure 6. The effects of velocity-changing collisions on the photon echo signal in a CH_3F molecular vapor is represented in Figure 7. The signal exhibits exponential decay varying as T_{21}^3 for early times and T_{21} for later times. A theoretical interpretation of the observed echo decay permits a determination of the binary collision parameters for CH_3F, namely, the characteristic velocity jump $\Delta u_{\text{rms}} = 200\ \text{cm s}^{-1}$ and the cross section of $\sigma = 580\ \text{Å}^2$.

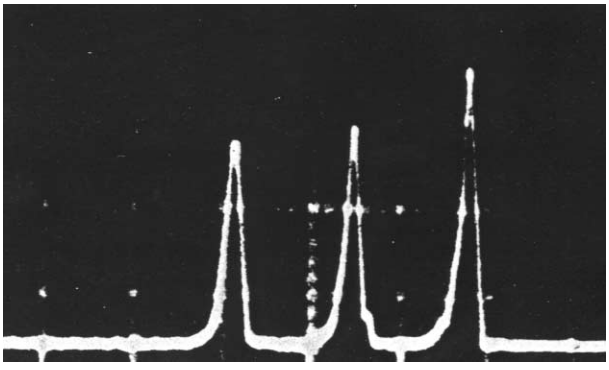


Figure 5 A photon echo signal from ruby. Time increases to the right with a scale of 100 ns/division. The pulse on the right is the echo signal, while the first two pulses are the (attenuated) input pulses. Reprinted with permission from Kurnit NA, Abella ID and Hartmann SR (1964) Observation of a photon echo. *Physical Review Letters* 13: 567. Copyright (1964) by the American Physical Society.

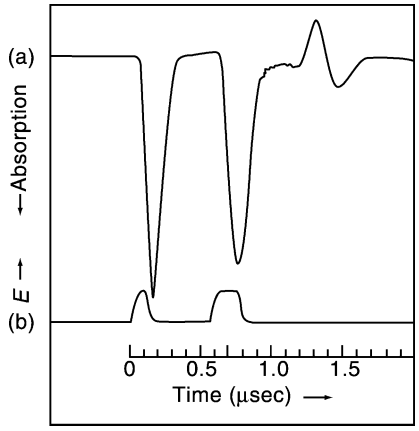


Figure 6 A photon echo signal in $^{13}CH_3F$ using the Stark switching method. The third pulse in trace (a) is the echo heterodyne beat signal. This signal is preceded by 'nutation' transient signals following each of the two Stark pulses shown in trace (b). Reprinted with permission from Brewer RG and Shoemaker RL (1971) Photo echo and optical nutation in molecules. *Physical Review Letters* 27: 631. Copyright (1971) by the American Physical Society.

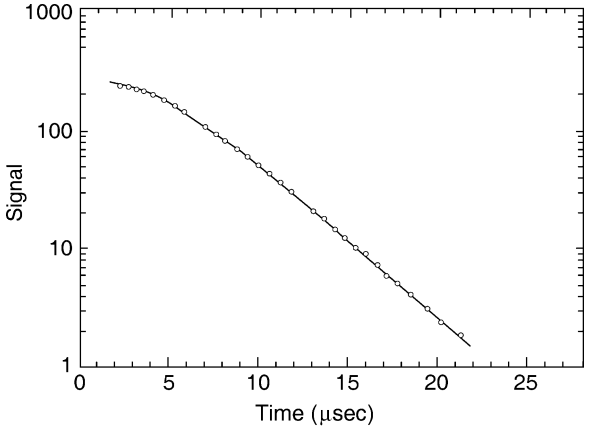


Figure 7 Decay of the photon echo signal in $^{13}CH_3F$ as a function of pulse separation. Reprinted with permission from Schmidt J, Berman PR and Brewer RG (1973) Coherent transient study of velocity-changing collisions. *Physical Review Letters* 31: 1103. Copyright (1973) by the American Physical Society.

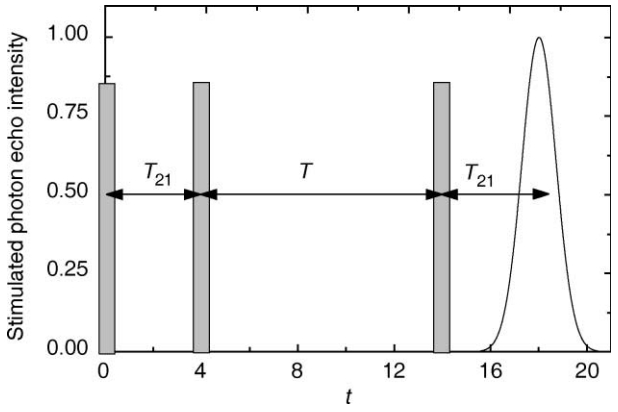


Figure 8 Pulse sequence and signal intensity for the stimulated photon echo.

Stimulated Photon Echo

In contrast to conventional photon echoes, *stimulated photon echoes* can be used to *simultaneously* measure both transverse and longitudinal relaxation times. Stimulated photon echoes have become an important diagnostic probe of relaxation in condensed matter systems as well as atomic vapors. The pulse sequence is shown in Figure 8 and consists of three pulses, having areas $\theta_1, \theta_2, \theta_3$, and equal propagation vectors $\mathbf{k}$. The time interval between the first two pulses is $T_{21} \gg T_2^*$ and pulse 3 occurs at time $t = T_{21} + T$. The maximum echo amplitude is given by

$$I(L, t_e) = \frac{1}{2}\left(\frac{k\mathcal{N}\mu L}{2\varepsilon_0}\sin\theta_1 \sin\theta_2 \sin\theta_3\right)^2 \times e^{-2\gamma_2 T} e^{-4\gamma T_{21}} \quad [27]$$

and occurs when $t_e - (T_{21} + T) = T_{21}$. The optimal pulse sequence consists of three $\pi/2$ pulses.

If slightly different $\mathbf{k}$ vectors are chosen such that $\mathbf{k}_1 = \mathbf{k}_3 \approx \mathbf{k}_2$ the echo intensity varies as

$$e^{-2\gamma_2 T} e^{-4\gamma T_{21}} e^{-|\mathbf{k}_1 - \mathbf{k}_2|^2 u^2 (T + 2T_{21})^2/2}$$

By varying the angle between $\mathbf{k}_1$and $\mathbf{k}_2$, one can determine the Doppler width ku. By monitoring the echo signal as a function of $T_{21}(T)$, one obtains information on the transverse (longitudinal) relaxation.

The situation changes dramatically if there is an *additional* ground state level (level 0) to which the excited state can decay. In that case, the system is 'open' since population can leave the 'two-level' subsystem. The total population of the 1–2 state subsystem is no longer conserved, requiring an *additional* decay rate to account for relaxation. Let us suppose that all states decay with rate Γ_t as a result of their finite time in the laser beams. Moreover, let $\Gamma_{2,1}$ and $\Gamma_{2,0}$ be the decay rates of level 2 to levels 1 and 0, respectively. Assuming that $\gamma_2 T \gg 1$, one finds that the factor $e^{-2\gamma_2 T}$ in eqn [27] must be replaced by $(\Gamma_{2,0}/(\Gamma_{2,1} + \Gamma_{2,0}))^2 e^{-2\Gamma_t T}$. If $\Gamma_{2,0} \neq 0$, there is a long-lived component in the ground state population that contributes to the echo signal. One can exploit this feature of open systems to study spectral diffusion or velocity-changing collisions with very high sensitivity. Echoes can occur for time separations T much greater than the excited state lifetime. The creation of such long-lived stimulated photon echoes can be attributed to the fact that, in open systems, a portion of the velocity modulation created by the first two pulses in the ground state remains following spontaneous decay from the excited state.

The stimulated photon echo observed on the D_1 transition in sodium is shown in Figure 9. This is

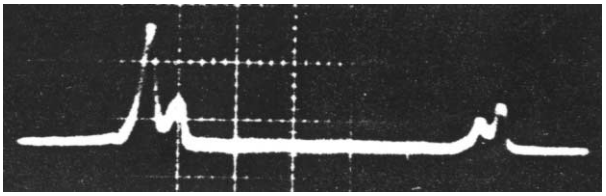


Figure 9 Stimulated photon echo observed on the D_1 transition in sodium. Reprinted with permission from Mossberg T, Flusberg A, Kachru R and Hartmann SR (1979) Total scattering cross section for Na on He measured by stimulated photon echoes. *Physical Review Letters* 42: 1665. Copyright (1979) by the American Physical Society.

an 'open' system, owing to ground state hyperfine structure. For the data shown, T is 17 times the 16 ns lifetime of the excited state. The first three pulses represent light scattered from the three input pulses and the fourth pulse is the echo. The echo appears at $t = T + 2T_{21}$.

Optical Ramsey Fringes

We have seen that coherent transients can be used to measure relaxation rates. With a slight modification of the stimulated photon echo geometry, it is also possible to use coherent transient signals to measure transition frequencies. If one chooses $\mathbf{k}_3$ in the $-\mathbf{k}_1$ direction instead of the $\mathbf{k}_1$ direction, it is possible to generate a phase matched signal in the

$$\mathbf{k} = \mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3 = -\mathbf{k}_2$$

direction. To simplify matters, it is assumed that $\mathbf{k}_1 \approx \mathbf{k}_2 = -\mathbf{k}$ and terms of order $|\mathbf{k}_1 - \mathbf{k}_2|u \times (T + 2T_{21})$ are neglected.

The averaged density matrix element in the vicinity of the echo is

$$\langle \rho_{21}(t) \rangle = (i/8) \sin\theta_1 \sin\theta_2 \sin\theta_3 e^{-\gamma_2 T} e^{-2\gamma T_{21}} \times e^{i\delta[(t-T_{21}-T)+T_{21}]} e^{-k^2u^2|(t-T_{21}-T)-T_{21}|^2/4} \quad [28]$$

An echo is formed at time $t_e = T + 2T_{21}$, just as for the stimulated photon echo, but there is an additional phase factor, given by $e^{2i\delta T_{21}}$ when $t = t_e$, that is absent for the nearly collinear geometry. This phase factor is the optical analog of the phase factor that is responsible for the generation of Ramsey fringes in the microwave domain. The phase factor can be measured directly by heterodyning the signal field with a reference field or, indirectly, by converting the off-diagonal density matrix element into a population by the addition of a *fourth* pulse in the $\mathbf{k}_3$ direction at time $t = T + 2T_{21}$. In either case, the signal varies as $\cos(2\delta T_{21})$. The central fringe, corresponding to $\delta = 0$ can be isolated using one of two methods. If the experiment is carried out using an atomic beam rather than atoms in a cell, $T_{21} = L/u_0$, will be different for atoms having different u_0 (L = spatial separation of the first two pulses and u_0 is the longitudinal velocity of the atoms). On averaging over a distribution of u_0, the fringe having $\delta = 0$ will have the maximum amplitude. Experiments of this type allow one to measure optical frequencies with accuracy of order T_{21}^{-1}. For experiments using temporally-separated pulses acting on atoms in a cell, it is necessary to take data as a function of δ for several values of T_{21}, and then average the data over T_{21}. It might seem remarkable that narrow pulses having large bandwidths can be used to measure optical frequencies with arbitrary accuracy as T_{21} is increased. Of course, it is the *total* duration of the pulse sequence rather than the duration of a single pulse that is the relevant time parameter. For the optical coherence, T_{21} is the appropriate time parameter (during the interval T, it is population rather than coherence that contributes to the signal). The price one pays by using large T_{21} is a signal that decays as $e^{-4\gamma T_{21}}$.

Figure 10 illustrates an optical Ramsey fringe signal on the 657 nm intercombination line in Ca. Four-field zones were used. The most probable value of T_{21} was about 10^{-5} s for an effusive beam having a most probable longitudinal speed equal to 800 ms^{-1}, giving a central fringe width of order 60 kHz.

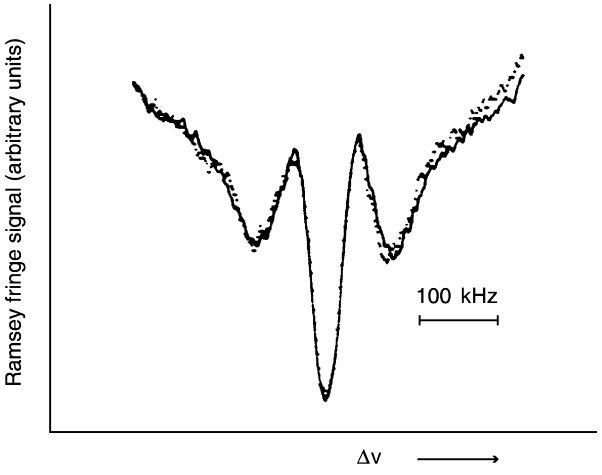


Figure 10 An optical Ramsey fringe signal on the 657 nm intercombination line in Ca. Four field zones were used. The solid and dotted lines represent runs with the directions of the laser field reversed, to investigate phase errors in the signals. Reprinted from Ito N, Ishikawa J and Moringa A (1994) Evaluation of the optical phase shift in a Ca Ramsey fringe stabilized optical frequency standard by means of laser-beam reversal. *Optics Communications* 109: 414, with permission from Elsevier.

Atom Interferometry

Matter wave atom interferometers rely on the wave nature of the center-of-mass motion for their operation. Such interferometers illustrate some interesting features of coherent optical transients not found in NMR. As an example, one can consider the response of a cold atomic ensemble to two, off-resonant standing-wave optical pulses separated in time by T. The electric field amplitude of pulse $j(j = 1, 2)$ is given by $\mathbf{E}_j(\mathbf{Z}, t) = \boldsymbol{\varepsilon} E_j(t)\cos(kZ)\cos(\omega t)$. The net effect of the field is to produce a spatially-modulated, ac Stark or light shift of the ground state energy. As a result, pulse j modifies the ground state amplitude by the phase factor $\exp[i\theta_j \cos(2kZ)]$. In other words, the standing wave field acts as a phase grating for the atoms. The phase factor can be interpreted in terms of momentum exchange between the two traveling wave components of the standing wave field. Owing to the fields' interaction with the atoms, all even integral multiples of $2\hbar k$ can be exchanged by the fields, imparting impulsive momenta of $2n\hbar k$ (n is a positive or negative integer) to the atoms. For an atom having momentum $\mathbf{P}$ and mass M, the frequency change associated with this momentum change is

$$E_{P,\mathbf{P}\pm 2n\hbar k\hat{Z}}/\hbar = \left(\frac{P^2}{2M} - \frac{|\mathbf{P} \pm 2n\hbar k\hat{Z}|^2}{2M}\right)\Big/\hbar = \mp\frac{2nP_Z k}{M} - \omega_{2nk} \qquad [29]$$

where

$$\omega_\kappa = \frac{\hbar\kappa^2}{2M}$$

There are two contributions to the frequency eqn [29]. The first part is independent of $\hbar$ and represents a classical Doppler shift, while the second part is proportional to $\hbar$ and represents a quantum, matter-wave effect. The quantum contribution will become important for times of order ω_{2nk}^{-1}. As with the echoes described previously, the Doppler effect results in a dephasing and rephasing of the optical response. However, in contrast to the previous echoes, the ground state density is no longer constant for times $t > \omega_{2nk}^{-1}$, when matter wave effects begin to play a role. At such times the phase grating created by the pulses can be transformed into amplitude gratings of the atomic density.

Long interaction times are possible, limited only by the time the atoms spend in the field. As such, atom interferometers offer exciting new possibilities for precision measurements of rotation rates, fundamental constants such as $\hbar$, gravitational acceleration, and gravitational gradients. Moreover, one can take advantage of the nonlinear atom-field interaction to create atomic density patterns having period $\lambda/2n$ from a standing wave optical field having wavelength λ. In this manner one has the possibility to construct atom interferometers and atom structures that operate on the nano-scale (spacings on the order of tens of nanometers).

Conclusion

Optical coherent transients are now used routinely as probes of atomic vapors, liquids, and condensed matter. Sophisticated techniques have been developed using fast and ultrafast pulses (having duration of 10 to 100 femtoseconds) to probe such systems and obtain the relevant relaxation rates. The manner in which optical coherent transients can be used to extract information on transition frequencies and relaxation rates in atomic and molecular vapors has been reviewed. In atomic vapors, optical coherent transients provide a means for obtaining both longitudinal and transverse relaxation rates. Moreover, echo experiments serve as an extremely sensitive probe of velocity changing collisions. Similar techniques are being rediscovered as an important tool in the analysis of cold atomic vapors and Bose–Einstein condensates. Only the most basic optical coherent transient phenomena have been described in this chapter. More elaborate, multi-pulse excitation schemes are joining established methods in attempts to understand the complex decay dynamics that one encounters in many branches of physics, chemistry and biology.

Acknowledgments

The research of PRB is supported by the National Science Foundation under Grants No. PHY-0098016 and the FOCUS Center Grant, and by the U. S. Army Research Office under Grant No. DAAD19-00-1-0412.

List of Units and Nomenclature

$\mathbf{E}(\mathbf{R}, t)$	Electric field vector
$E(t)$	Electric field amplitude
$E_s(Z, t)$	Complex signal electric field amplitude
$I(L, t)$	Signal intensity exiting the sample
$\mathbf{k}$	Field propagation vector
L	Sample length
$\mathcal{N}$	Atomic density
$\mathbf{P}$	Center-of-mass momentum
$\mathbf{P}(\mathbf{R}, t)$	Polarization vector
$P(Z, t)$	Complex polarization field amplitude
T_{21}, T	Time interval between pulses
u	Most probable atomic speed

(u, v, w)	Elements of Bloch vector
$\mathbf{U}(t)$	Bloch vector
$\mathbf{v}$	Atomic velocity
$W_0(\mathbf{v})$	Atomic velocity distribution
γ	Transverse relaxation rate
γ_2	Excited state decay rate or longitudinal relaxation rate
δ	Atom-field detuning
θ	Pulse area
μ	Dipole moment matrix element
ρ_{ij}^{I}	Density matrix element in 'normal' interaction representation
$\rho_{ij}(Z, t)$	Density matrix element in a field interaction representation
τ	Pulse durations
ω	Field frequency
ω_k	Recoil frequency
ω_0	Atomic transition frequency
$\Omega(t)$	Generalized Rabi frequency
$\mathbf{\Omega}(t)$	Pseudofield vector
$\Omega_0(t)$	Rabi frequency

See also

Interferometry: Gravity Wave Detection; Overview.

Further Reading

Allen L and Eberly J (1975) *Optical Resonance and Two-Level Atoms*, Chapters 2,4,9, New York: Wiley.

Berman PR and Steel DG (2001) *Coherent optical transients*. In: Bass M, Enoch JM, Van Stryland E and Wolf WL (eds) Handbook of Optics, vol. IV. chap. 24 New York: McGraw-Hill. (Material presented in this chapter is a condensed version of this reference.).

Brewer RG (1977) Coherent optical spectroscopy. In: Balian R, Haroche S and Liberman S (eds) *Frontiers in Laser Spectroscopy, Les Houches, Session XXVII*, vol. 1, pp. 341–396. Amsterdam: North Holland.

Brewer RG (1977) Coherent optical spectroscopy. In: Bloembergen N (ed.) *Nonlinear Spectroscopy*, Proc. Int. School Phys., Enrico Fermi, Course 64, Amsterdam: North-Holland.

Brewer RG (1977) *Coherent optical transients*. In *Physics Today*, vol. 30(5), 50–59.

Brewer RG and Hahn EL (1984) Atomic memory. In *Scientific American*, vol. 251(6), 50–57.

Journal of Optical Society America (1986) Special issue on coherent optical spectroscopy, Vol. A3(4).

Levenson MD and Kano SS (1988) *Introduction to Nonlinear Laser Spectroscopy*, Chapter 6. Boston: Academic Press.

Meystre P and Sargent M III (1999) *Elements of Quantum Optics*, 3rd edn. Chapter 12. Berlin: Springer-Verlag.

Shen YR (1984) *The Principles of Nonlinear Optics*, Chapter 17. New York: Wiley.

Shoemaker RL (1978) Coherent transient infrared spectroscopy. In: Steinfeld JI (ed.) *Laser and Coherence Spectroscopy*, pp. 197–371. New York: Plenum.

Zinth W and Kaiser W (1988) *Ultrafast Coherent Spectroscopy*. In: Kaiser W (ed.) *Ultrashort Laser Pulses and Applications*, vol. 60, pp. 235–277. Berlin: Springer-Verlag.

Foundations of Coherent Transients in Semiconductors

T Meier and S W Koch, Philipps-University, Marburg, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Electromagnetic irradiation of matter generates a coherent superposition of the excited quantum states. The attribute 'coherent' relates to the fact that the material excitations originally have a well defined phase dependence which is imposed by the phase of the excitation source, often a laser in the optical regime. Macroscopically, the generated superposition state can be described as an optical polarization which is determined by the transition amplitudes between the participating quantum states.

The optical polarization is a typical non-equilibrium quantity that decays to zero when a system relaxes to its equilibrium state. Coherent effects are therefore only observable in a certain time window after pulsed photo-excitation, or in the presence of a continuous-wave (cw) beam. As coherent transients one usually refers to phenomena that can be observed during or shortly after pulsed laser excitation and critically depend on the presence of the induced optical polarization.

Many materials such as atoms, molecules, metals, insulators, semiconductors including bulk crystals, heterostructures, and surfaces, as well as organic and biological structures are studied using coherent optical spectroscopy. Depending on the particular system, the states participating in the optical transitions, the interactions among them, and the resulting time-scale for the decay of the induced polarization may be very different. As a result, the time window during which coherent effects are

observable can be as long as seconds for certain atomic transitions, or it may be as short as a few femtoseconds (10^{-15}s), e.g., for metals, surfaces, or highly excited semiconductors.

Coherent spectroscopy and the analysis of coherent transients has provided valuable information on the nature and dynamics of the optical excitations. Often it is possible to learn about the interaction processes among the photoexcitations and to follow the temporal evolution of higher-order transitions which are only accessible if the system is in a non-equilibrium state.

The conceptually simplest experiment which one may use to observe coherent transients is to time resolve the transmission or reflection induced by a single laser pulse. However, much richer information can be obtained if one excites the system with several pulses. A pulse sequence with well-controlled delay times makes it possible to study the dynamical evolution of the photo-excitations, not only by time resolving the signal, but also by varying the delay. Prominent examples of such experiments are pump-probe measurements, which usually are performed with two incident pulses, or four-wave mixing, for which one may use two or three incident pulses.

Besides the microscopic interaction processes, the outcome of an experiment is determined by the quantum mechanical selection rules for the transitions and by the symmetries of the system under investigation. For example, if one wants to investigate coherent optical properties of surface states one often relies on phenomena, such as second-harmonic or sum-frequency generation, which give no signal in perfect systems with inversion symmetry. Due to the broken translational invariance, such experiments are therefore sensitive only to the dynamics of surface and/or interface excitations.

In this article the basic principles of coherent transients are presented and several examples are presented. The basic theoretical description and its generalization for the case of semiconductors are introduced.

Basic Principles

In the absence of free charges and currents, Maxwell's equations show that the electromagnetic field interacts with matter via the optical polarization. This polarization $\mathbf{P}$, or more precisely its second derivative with respect to time $(\partial^2/\partial t^2)\mathbf{P}$, appears as a source term in the wave equation for the electric field $\mathbf{E}$. Consequently, if the system is optically thin such that propagation effects within the sample can be ignored and if measurements are performed in the far field region, i.e., at distances exceeding the characteristic optical wavelength λ, the emitted electric field resulting from the polarization is proportional to its second time derivative, $\mathbf{E} \propto (\partial^2/\partial t^2)\mathbf{P}$. Thus the measurement of the emitted field dynamics yields information about the temporal evolution of the optical material polarization.

Microscopically the polarization is determined by the transition amplitudes between the different states of the system. These may be the discrete states of atoms or molecules, or the microscopic valence and conduction band states in a dielectric medium, such as a semiconductor. In any case, the macroscopic polarization $\mathbf{P}$ is computed by summing over all microscopic transitions p_{cv} via $\mathbf{P} = \sum_{c,v} (\boldsymbol{\mu}_{cv} p_{cv} + c.c.)$, where $\boldsymbol{\mu}_{cv}$ is the dipole matrix element which determines the strength of the transitions between the states v and c, and $c.c.$ denotes the complex conjugate. If ε_c and ε_v are the energies of these states, their dynamic quantum mechanical evolution is described by the phase factors $e^{-i\varepsilon_c t/\hbar}$ and $e^{-i\varepsilon_v t/\hbar}$, respectively. Therefore, each p_{cv} is evolving in time according to $e^{-i(\varepsilon_c - \varepsilon_v)t/\hbar}$. Assuming that we start at $t = 0$ with $p_{cv}(t = 0) = p_{cv,0}$, which may be induced by a short optical pulse, we have for the optical polarization $\mathbf{P}(t) = \sum_{c,v} (\boldsymbol{\mu}_{cv} p_{cv,0} e^{-i(\varepsilon_c - \varepsilon_v)t/\hbar} + c.c.)$. Thus $\mathbf{P}(t)$ is given by a summation over microscopic transitions which all oscillate with frequencies proportional to the energy differences between the involved states. Hence, the optical polarization is clearly a coherent quantity which is characterized by amplitude and phase. Furthermore, the microscopic contributions to $\mathbf{P}(t)$ add up coherently. Depending on the phase relationships, one may obtain either constructive superposition, interference phenomena like quantum beats, or destructive interference leading to a decay (dephasing) of the macroscopic polarization.

Optical Bloch Equations

The dynamics of photo-excited systems can be conveniently described by a set of equations, the optical Bloch equations, named after Felix Bloch (1905–1983) who first formulated such equations to analyze the spin dynamics in nuclear magnetic resonance. For the simple case of a two-level model, the Bloch equations can be written as

$$i\hbar \frac{\partial}{\partial t} p = \Delta\varepsilon p + \mathbf{E}\cdot\boldsymbol{\mu} I \qquad [1]$$

$$i\hbar \frac{\partial}{\partial t} I = 2\mathbf{E}\cdot\boldsymbol{\mu}(p - p^*) \qquad [2]$$

Here, $\Delta\varepsilon$ is the energy difference and I the inversion, i.e., the occupation difference between upper and lower state. The field $\mathbf{E}$ couples the polarization to the product of the Rabi energy $\mathbf{E}\cdot\boldsymbol{\mu}$

and the inversion I. In the absence of the driving field, i.e., $\mathbf{E} \equiv 0$, eqn [1] describes the free oscillation of p discussed above.

The inversion is determined by the combined action of the Rabi energy and the transition p. The total occupation N, i.e., the sum of the occupations of the lower and the upper states, remains unchanged during the optical excitation since light does not create or destroy electrons; it only transfers them between different states. If N is initially normalized to 1, then I can vary between -1 and 1. $I = -1$ corresponds to the ground state of the system, where only the lower state is occupied. In the opposite extreme, i.e., for $I = 1$, the occupation of the upper state is 1 and the lower state is completely depleted.

One can show that eqns [1]–[2] contain another conservation law. Introducing $\mathcal{P} = p + p^* = 2Re[p]$ and $J = i(p - p^*) = -2Im[p]$, which are real quantities and often named polarization and polarization current, respectively, one finds that the relation

$$\mathcal{P}^2 + J^2 + I^2 = 1 \qquad [3]$$

is fulfilled. Thus the coupled coherent dynamics of the transition and the inversion can be described on a unit sphere, the so-called Bloch sphere. The three terms $\mathcal{P}, J$, and I can be used to define a three-dimensional Bloch vector $\mathbf{S} = (\mathcal{P}, J, I)$. With these definitions, one can reformulate eqns [1]–[2] as

$$\frac{\partial}{\partial t}\mathbf{S} = \mathbf{\Omega} \times \mathbf{S} \qquad [4]$$

with $\mathbf{\Omega} = (-2\boldsymbol{\mu}\cdot\mathbf{E}, 0, \Delta\varepsilon)/\hbar$ and $\times$ denoting the vector product.

The structure of eqn [4] is mathematically identical to the equations describing either the angular momentum dynamics in the presence of a torque or the spin dynamics in a magnetic field. Therefore, the vector $\mathbf{S}$ is often called pseudospin. Moreover, many effects which can be observed in magnetic resonance experiments, e.g., free decay, quantum beats, echoes, etc. have their counterparts in photo-excited optical systems.

The vector product on the right-hand side of eqn [4] shows that $\mathbf{S}$ is changed by $\mathbf{\Omega}$ in a direction perpendicular to $\mathbf{S}$ and $\mathbf{\Omega}$. For vanishing field the torque $\mathbf{\Omega}$ has only a z-component. In this case the z-component of $\mathbf{S}$, i.e., the inversion, remains constant, whereas the x- and y-components, i.e., the polarization and the polarization current, oscillate with the frequency $\Delta\varepsilon/\hbar$ on the Bloch sphere around $\mathbf{\Omega}$.

Semiconductor Bloch Equations

If one wants to analyze optical excitations in crystallized solids, i.e., in systems which are characterized by a periodic arrangement of atoms, one has to include the continuous dispersion (bandstructure) into the description. In translationally invariant, i.e., ordered systems, this can be done by including the crystal momentum $\hbar\mathbf{k}$ as an index to the quantities which describe the optical excitation of the material, e.g., p and I in eqns [1]–[2]. Hence, one has to consider a separate two-level system for each crystal momentum $\hbar\mathbf{k}$. As long as all other interactions are disregarded, the bandstructure simply introduces a summation over the uncoupled contributions from the different $\mathbf{k}$ states, leading to inhomogeneous broadening due to the presence of a range of transition frequencies $\Delta\varepsilon(\mathbf{k})/\hbar$.

For any realistic description of optical processes in solids, it is essential to go beyond the simple picture of non-interacting states and to treat the interactions among the elementary material excitations, e.g., the Coulomb interaction between the electrons and the coupling to additional degrees of freedom, such as lattice vibrations (electron–phonon interaction) or other bath-like subsystems. If crystals are not ideally periodic, imperfections, which can often be described as a disorder potential, need to be considered as well.

All these effects can be treated by proper extensions of the optical Bloch equations introduced above. For semiconductors the resulting generalized equations are known as semiconductor Bloch equations, where the microscopic interactions are included at a certain level of approximation. For a two-band model of a semiconductor, these equations can be written schematically as

$$i\hbar\frac{\partial}{\partial t}p_{\mathbf{k}} = \Delta\varepsilon_{\mathbf{k}}p_{\mathbf{k}} + \Omega_{\mathbf{k}}(n^c_{\mathbf{k}} - n^v_{\mathbf{k}}) + i\hbar\frac{\partial}{\partial t}p_{\mathbf{k}}|_{\mathrm{corr}} \qquad [5]$$

$$i\hbar\frac{\partial}{\partial t}n^c_{\mathbf{k}} = (\Omega^*_{\mathbf{k}}p_{\mathbf{k}} - \Omega_{\mathbf{k}}p^*_{\mathbf{k}}) + i\hbar\frac{\partial}{\partial t}n^c_{\mathbf{k}}|_{\mathrm{corr}} \qquad [6]$$

$$i\hbar\frac{\partial}{\partial t}n^v_{\mathbf{k}} = -(\Omega^*_{\mathbf{k}}p_{\mathbf{k}} - \Omega_{\mathbf{k}}p^*_{\mathbf{k}}) + i\hbar\frac{\partial}{\partial t}n^v_{\mathbf{k}}|_{\mathrm{corr}} \qquad [7]$$

Here $p_{\mathbf{k}}$ is the microscopic polarization and $n^c_{\mathbf{k}}$ and $n^v_{\mathbf{k}}$ are the electron populations in the conduction and valence bands (c and v), respectively. Due to the Coulomb interaction and possibly further processes, the transition energy $\Delta\varepsilon_{\mathbf{k}}$ and the Rabi energy $\Omega_{\mathbf{k}}$ both depend on the excitation state of the system, i.e., they are functions of the time-dependent polarizations and populations. This leads, in particular, to a coupling among the excitations for all different values of the crystals momentum $\hbar\mathbf{k}$. Consequently, in the presence of interactions, the optical excitations can no longer be described as independent two-level systems but have to be treated

as a coupled many-body system. A prominent and important example in this context is the appearance of strong exciton resonances which, as a consequence of the Coulomb interaction, show up in the absorption spectra of semiconductors energetically below the fundamental bandgap.

The interaction effects lead to significant mathematical complications since they induce couplings between all the different quantum states in a system and introduce an infinite hierarchy of equations for the microscopic correlation functions. The terms given explicitly in eqns [5]–[7] arise in a treatment of the Coulomb interaction on the Hartree-Fock level. Whereas this level is sufficient to describe excitonic resonances, there are many additional effects, e.g., excitation-induced dephasing due to Coulomb scattering and significant contributions from higher-order correlations, like excitonic populations and biexcitonic resonances, which make it necessary to treat many-body correlation effects that are beyond the Hartree-Fock level. These contributions are formally included by the terms denoted as $|_{\mathrm{corr}}$ in eqns [5]–[7]. The systematic truncation of the many-body hierarchy and the analysis of controlled approximations is the basic problem in the microscopic theory of optical processes in condensed matter systems.

Examples

Radiative Decay

The emitted field originating from the non-equilibrium coherent polarization of a photo-excited system can be monitored by measuring the transmission and the reflection as function of time. If only a single isolated transition is excited, the dynamic evolution of the polarization and therefore of the transmission and reflection is governed by radiative decay. This decay is a consequence of the coupling of the optical transition to the light field described by the combination of Maxwell- and Bloch-equations. Radiative decay simply means that the optical polarization is converted into a light field on a characteristic time scale $2T_{\mathrm{rad}}$. Here, T_{rad} is the population decay time, often also denoted as T_1-time. This radiative decay is a fundamental process which limits the time on which coherent effects are observable for any photo-excited system. Due to other mechanisms, however, the non-equilibrium polarization often vanishes considerably faster.

The value of T_{rad} is determined by the dipole matrix element and the frequency of the transition, i.e., $T_{\mathrm{rad}}^{-1} \propto |\boldsymbol{\mu}|^2 \Delta\omega$, with $\Delta\omega = \Delta\varepsilon/\hbar$. The temporal evolution of the polarization and the emitted field are proportional to $e^{-i\Delta\omega t - t/(2T_{\mathrm{rad}})}$. Usually one measures the intensity of a field, i.e., its squared modulus, which evolves as $e^{-t/T_{\mathrm{rad}}}$ and thus simply shows an exponential decay, the so-called *free-induction decay* (see Figure 1). T_{rad} can be very long for transitions with small matrix elements. For semiconductor quantum wells, however, it is in the order of only 10 ps, as the result of the strong light-matter interaction.

The time constant on which the optical polarization decays is often called T_2. In the case where this decay is dominated by radiative processes, we thus have $T_2 = 2T_{\mathrm{rad}}$.

Superradiance

The phenomenon of superradiance can be discussed considering an ensemble of N two-level systems which are localized at certain positions $\mathbf{R}_i$. In this case

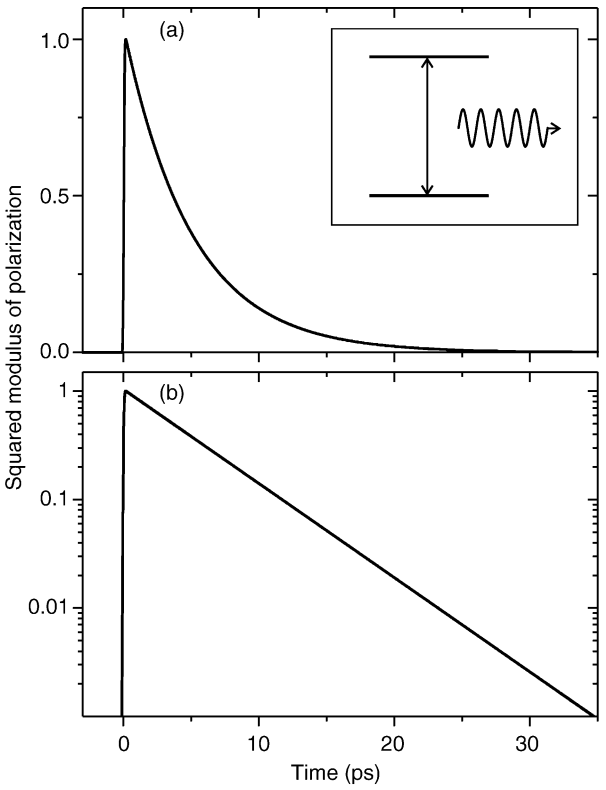


Figure 1 A two-level system is excited by a short optical pulse at $t = 0$. Due to radiative decay (inset) and possibly other dephasing mechanisms, the polarization decays exponentially as function of time with the time constant T_2, the dephasing time. The intensity of the optical field which is emitted as a result of this decay is proportional to the squared modulus of the polarization, which falls off with the time constant $T_2/2$. The decay of the squared modulus of the polarization is shown in (a) on a linear scale and in (b) on a logarithmic scale. The dephasing time was chosen to be $T_2 = 10$ ps which models the radiative decay of the exciton transition of a semiconductor quantum well.

Maxwell's equations introduce a coupling among all these resonances since the field emitted from any specific resonance interacts with all other resonances and interferes with their emitted fields. As a result, this system is characterized by N eigenmodes originating from the radiatively coupled optical resonances.

A very spectacular situation arises if one considers N identical two-level systems regularly arranged with a spacing that equals an integer multiple of $\lambda/2$, where λ is the wavelength of the system's resonance, i.e., $\lambda = c/\Delta\omega$, where c is the speed of light in the considered material (see Figure 2a). In this case all emitted fields interfere constructively and the system behaves effectively like a single two-level system with a matrix element increased by $\sqrt{N}$. Consequently, the radiative decay rate is increased by N and the polarization of the coupled system decays N-times faster than that of an isolated system (see Figure 2b). This effect is called superradiance.

It is possible to observe superradiant coupling effects, e.g., in suitably designed semiconductor heterostructures. Figure 2c compares the measured time-resolved reflection from a single quantum well (dashed line) with that of a Bragg structure, i.e., a multiple quantum-well structure which consists of 10 individual wells that are separated by $\lambda/2$ (solid line), where λ is the wavelength of the exciton resonance. For times greater than about 2 ps the direct reflection of the exciting pulse has decayed sufficiently and one is left with the exponential decay of the remaining signal. Figure 2c shows that this decay is much more rapid for the Bragg structure than for the single quantum well, due to superradiance introduced by the radiative coupling among the quantum wells.

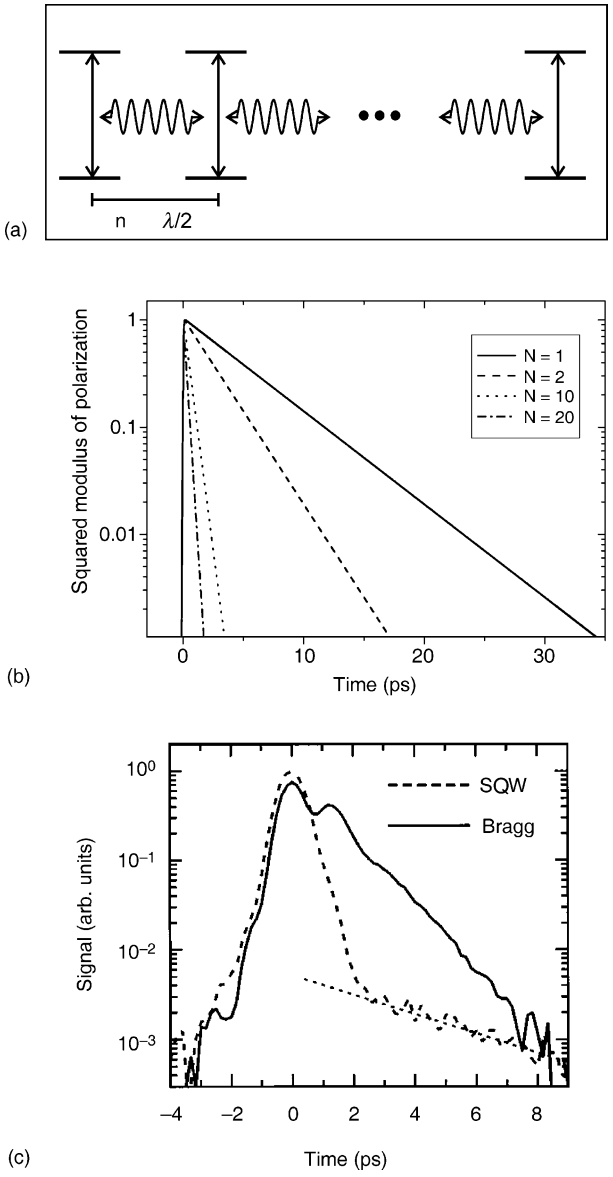


Figure 2 (a) N identical two-level systems are depicted which are regularly arranged with a spacing that equals an integer multiple of $\lambda/2$, where λ is the wavelength of the system's resonance. Due to their coupling via Maxwell's equations all fields emitted from the two-level systems interfere constructively and the coupled system behaves effectively like a single two-level system with an optical matrix element increased by $\sqrt{N}$. (b) The temporal decay of the squared modulus of the polarization is shown on a logarithmic scale. The radiative decay rate is proportional to N and thus the polarization of the coupled system decays N-times faster than that of an isolated system. This effect is called superradiance. (c) Measured time-resolved reflection of a semiconductor single quantum well (dashed line) and a $N = 10$ Bragg structure, i.e., a multi quantum well where the individual wells are separated by $\lambda/2$ (solid line), on a logarithmic scale. [Part (c) is reproduced with permission from Haas S, Stroucken T, Hübner M, *et al.* (1998) Intensity dependence of superradiant emission from radiatively coupled excitons in multiple-quantum-well Bragg structures. *Physical Review B* 57: 14860. Copyright (1998) by the American Physical Society].

Destructive Interference

As the next example we now consider a distribution of two-level systems which have slightly different transition frequencies characterized by the distribution function $g(\Delta\omega - \bar{\omega})$ of the transition frequencies which is peaked at the mean value $\bar{\omega}$ and has a spectral width of $\delta\omega$ (see Figure 3a). Ignoring the radiative coupling among the resonances, the optical polarization of the total system evolves after excitation with a short laser pulse at $t = 0$ proportional to $\int d\omega\, g(\Delta\omega - \bar{\omega}) e^{-i\Delta\omega t} \propto \tilde{g}(t) e^{-i\bar{\omega}t}$, where $\tilde{g}(t)$ denotes the Fourier transform of the frequency distribution function. $\tilde{g}(t)$ and thus the optical polarization decays on a time-scale which is inversely proportional to the spectral width of the distribution function $\delta\omega$. Thus the destructive interference of many different transition frequencies results in a rapid decay of the polarization (see Figure 3b).

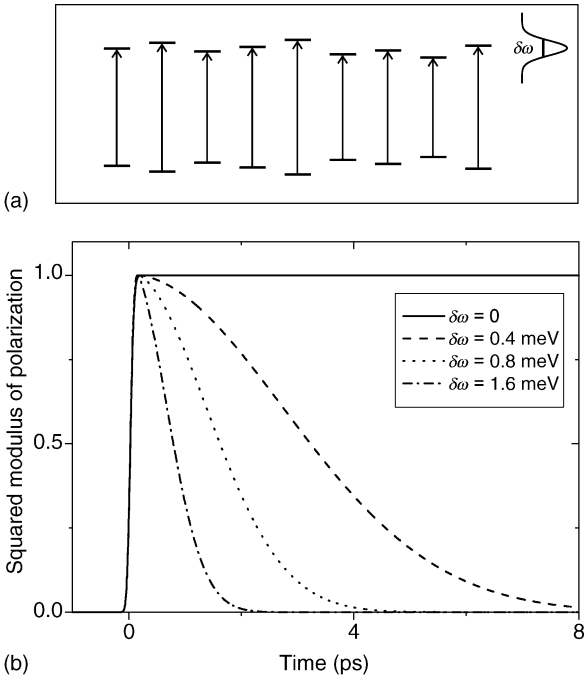


Figure 3 (a) An ensemble of two-level systems is shown where the resonance frequencies are randomly distributed according to a Gaussian distribution function of width $\delta\omega$ around an average value. (b) The temporal dynamics of the squared modulus of the polarization of ensembles of two-level systems after excitation with a short optical pulse at $t = 0$ is shown on a linear scale. Since the dephasing time is set to infinity, i.e., $T_2 \rightarrow \infty$, the polarization of an ensemble of identical two-level system ($\delta\omega = 0$) does not decay. However, for a finite width of the distribution, $\delta\omega > 0$, the individual polarizations of the ensemble oscillate with different frequencies and, therefore, due to destructive interference the polarization of the ensemble decays as function of time. Since the Fourier transform of a frequency-domain Gaussian is a Gaussian in the time domain, the dashed, dotted, and dashed dotted line have a Gaussian shape with a temporal width which is inversely proportional to $\delta\omega$.

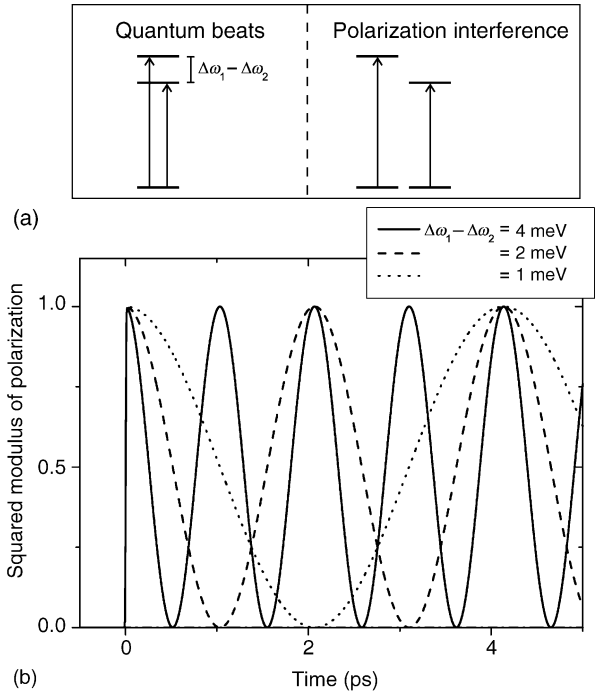


Figure 4 (a) Two optical resonances with frequency difference $\Delta\omega_1 - \Delta\omega_2$ may be realized by either a three-level system or by two uncoupled two-level systems. After impulsive excitation with an optical pulse the linear polarization of both types of systems show a modulation of the squared modulus of the polarization with the time period $2\pi/(\Delta\omega_1 - \Delta\omega_2)$. For the intrinsically coupled three-level system these modulations are called quantum beats, whereas for an interference of uncoupled systems they are named polarization interferences. Using nonlinear optical techniques, e.g., four-wave mixing and pump probe, it is possible to distinguish quantum beats and polarization interferences since coupled and uncoupled systems have different optical nonlinearities. (b) The temporal dynamics of the squared modulus of the polarization is shown for systems with two resonances and frequency difference $\Delta\omega_1 - \Delta\omega_2$, neglecting dephasing, i.e., setting $T_2 \rightarrow \infty$. After excitation with a short optical pulse at $t = 0$ the squared modulus of the polarization is periodically modulated with a time period $2\pi/(\Delta\omega_1 - \Delta\omega_2)$.

In the spectral domain this rapid decay shows up as inhomogeneous broadening. Depending on the system under investigation, there are many different sources for such an inhomogeneous broadening. One example is the Doppler broadening in atomic gases or disorder effects such as well-width fluctuations in semiconductor quantum wells or lattice imperfections in crystals.

In the nonlinear optical regime it is, under certain circumstances, possible to reverse the destructive interference of inhomogeneously broadened coherent polarizations. For example, in four-wave mixing, a second pulse may lead to a rephasing of the contributions with different frequencies, which results in the photon echo (see discussion below).

Quantum Beats

The occurrence of quantum beats can be understood most easily in a system where the total optical polarization can be attributed to a finite number of optical transitions. Let us assume for simplicity that all these transitions have the same matrix element. In this case, after excitation with a short laser pulse at $t = 0$ the optical polarization of the total system evolves proportional to $\sum_i e^{-i\Delta\omega_i t}$. The finite number of frequencies results in a temporal modulation with time periods $2\pi/(\Delta\omega_i - \Delta\omega_j)$ of the squared modulus of the polarization which is proportional to the emitted field intensity. For the case of two frequencies the squared modulus of the polarization is proportional to $[1 + \cos((\Delta\omega_1 - \Delta\omega_2)t)]$, i.e., due to the interference of two contributions the polarization varies between a maximum and zero (see Figure 4b).

In the linear optical regime it is impossible to distinguish whether the optical transitions are

uncoupled or coupled. As shown in Figure 4, the two uncoupled two-level systems give the same linear polarization as a three-level systems where the two transitions share a common state. It is, however, possible to decide about the nature of the underlying transitions if one performs nonlinear optical spectroscopy. This is due to the fact that the so-called quantum beats, i.e., the temporal modulations of the polarization of an intrinsically coupled system, show a different temporal evolution as the so-called polarization interferences, i.e., the temporal modulations of the polarization of uncoupled systems. The former ones are also much more stable in the presence of inhomogeneous broadening than the latter ones.

In semiconductor heterostructures, quantum-beat spectroscopy has been widely used to investigate the temporal dynamics of excitonic resonances. Also the coupling among different optical resonances has been explored in pump-probe and four-wave-mixing measurements.

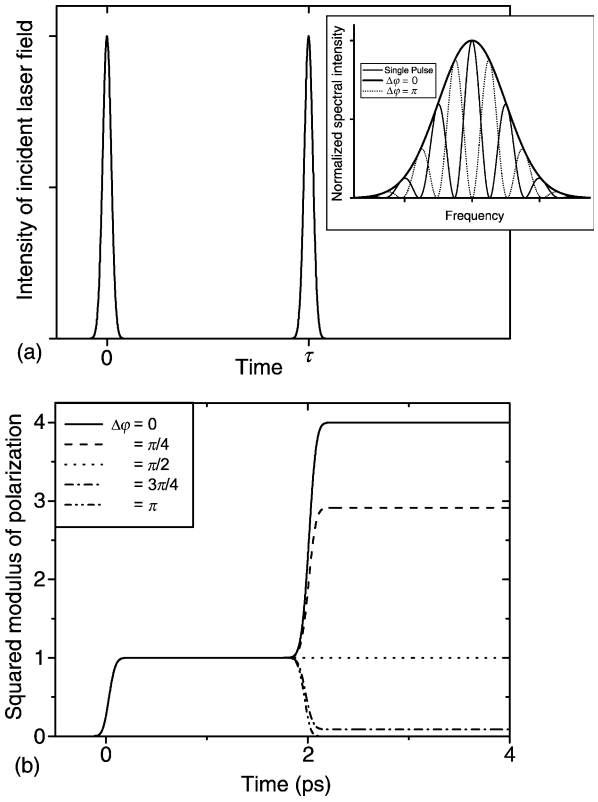


Figure 5 (a) Two Gaussian laser pulses separated by the time delay τ can lead to interferences. As shown by the thick solid line in the inset, the spectral intensity of a single pulse is a Gaussian with a maximum at the central frequency of the pulse. The width of this Gaussian is inversely proportional to the duration of the pulse. The spectral intensity of the field consisting of both pulses shows interference fringes, i.e., is modulated with a spectral period that is inversely proportional to τ. As shown by the thin solid and the dotted lines in the inset, the phase of the interference fringes depends on the phase difference $\Delta\varphi$ between the two pulses. Whereas for $\Delta\varphi = 0$ the spectral intensity has a maximum at the central frequency, it vanishes at this position for $\Delta\varphi = \pi$. (b) The temporal dynamics of the squared modulus of the polarization is shown for a two-level system excited by a pair of short laser pulses neglecting dephasing, i.e., setting $T_2 \to \infty$. The first pulse excites at $t = 0$ an optical polarization with a squared modulus normalized to 1. The second pulse, which has the same intensity as the first one, also excites an optical polarization at $t = 2$ ps. For $t > 2$ ps the total polarization is given by the sum of the polarizations induced by the two pulses. Due to interference, the squared modulus of the total polarization depends sensitively on the phase difference $\Delta\varphi$ between the two pulses. For constructive interference the squared modulus of the total polarization after the second pulse is four times bigger than that after the first pulse, whereas it vanishes for destructive interference. One may achieve all values in between these extremes by using phase differences $\Delta\varphi$ which are no multiples of π. It is thus shown that the second pulse can be used to coherently control the polarization induced by the first pulse.

Coherent Control

In the coherent regime the polarization induced by a first laser pulse can be modified by a second, delayed pulse. For example, the first short laser pulse incident at $t = 0$ induces a polarization proportional to $e^{-i\Delta\omega t}$ and a second pulse arriving at $t = \tau > 0$ induces a polarization proportional to $e^{-i\varphi}e^{-i\Delta\omega(t-\tau)}$. Thus for $t \geq \tau$ the total polarization is given by $e^{-i\Delta\omega t}(1 + e^{-i\Delta\varphi})$ with $\Delta\varphi = \varphi - \Delta\omega\tau$. If $\Delta\varphi$ is an integer multiple of 2π, the polarizations of both pulses interfere constructively and the amplitude of the resulting polarization is doubled as compared to a single pulse. If, on the other hand, $\Delta\varphi$ is an odd multiple of π, the polarizations of both pulses interfere destructively and the resulting polarization vanishes after the action of both pulses. Thus by

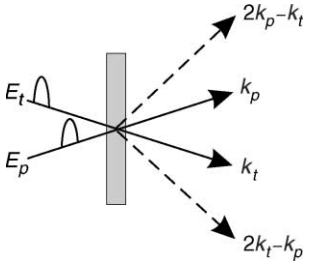


Figure 6 Transient optical nonlinearities can be investigated by exciting the system with two time delayed optical pulses. The pump pulse (E_p) puts the system in an excited state and the test (probe) pulse (E_t) is used to measure its dynamics. The dynamic nonlinear optical response may be measured in the transmitted or reflected directions of the pump and test pulses, i.e., $\pm k_p$ and $\pm k_t$, respectively. In a pump-probe experiment one often investigates the change of the test absorption induced by the pump pulse, by measuring the absorption in the direction k_t with and without E_p. One may also measure the dynamic nonlinear optical response in scattering directions like $2k_p - k_t$ and $2k_t - k_p$ as indicated by the dashed lines. This is what is done in a four-wave-mixing or photon-echo experiment.

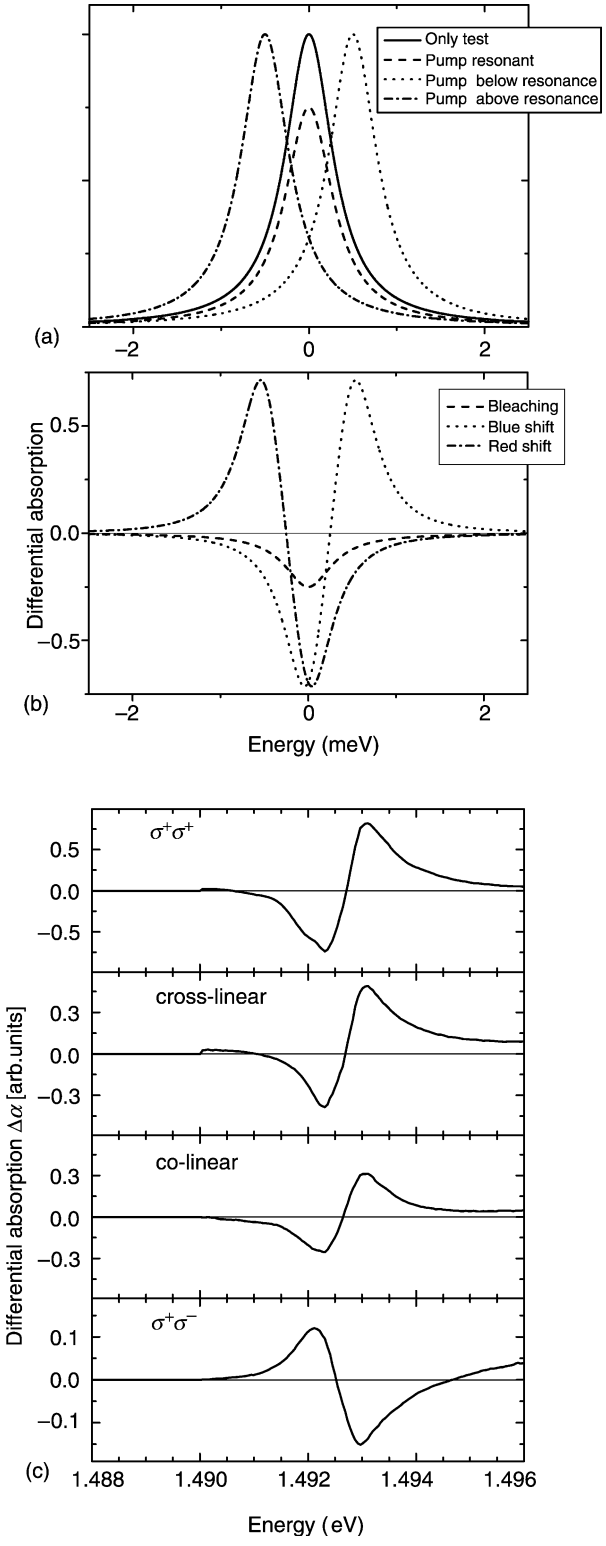


Figure 7 (a) Absorption spectra of a two-level system. The solid line shows the linear optical absorption spectrum as measured by a weak test pulse. It corresponds to a Lorentzian line that is centered at the transition frequency $\Delta\omega$, which has been set to zero. The width of the line is inversely proportional to the dephasing time T_2. In the presence of a pump pulse which resonantly excited the two-level system, the absorption monitored by the test pulse is reduced in amplitude, i.e., bleached, since the pump puts the system in an excited state (dashed line). In the optical Stark effect the frequency of the pump is nonresonant with the transition frequency. If the pump is tuned below (above) the transition frequency, the absorption is shifted to higher (lower) frequencies, i.e., shifted to the blue (red) part of the spectrum, see dotted (dashed-dotted) line. (b) The differential absorption obtained by taking the difference between the absorption in the presence of the pump and the absorption without pump. The dashed line shows the purely negative bleaching obtained using a resonant pump. The dotted and the dashed-dotted lines correspond to the dispersive shape of the blue and red shift obtained when pumping below and above the resonance, respectively. (c) Measured differential absorption spectra of a semiconductor quantum well which is pumped off resonantly 4.5 *meV* below the 1s heavy-hole exciton resonance. The four lines correspond to different polarization directions of the pump and probe pulses as indicated. [Part (c) is reproduced with permission from Sieh C, Meier T, Jahnke F, *et al.* (1999) Coulomb memory signatures in the excitonic optical Stark effect. *Physical Review Letters* 82: 3112. Copyright (1999) by the American Physical Society.]

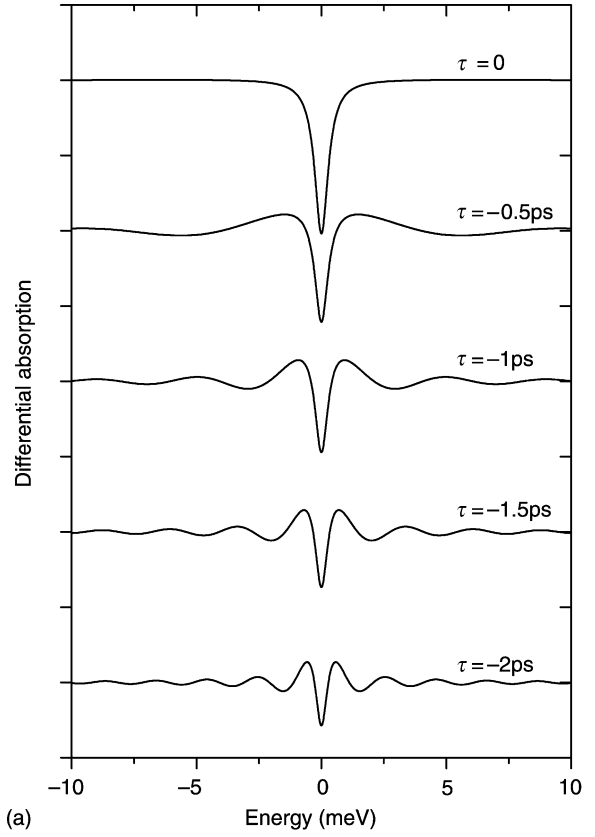
varying the phase difference of the two pulses it is possible to coherently enhance or destroy the optical polarization (see Figure 5b).

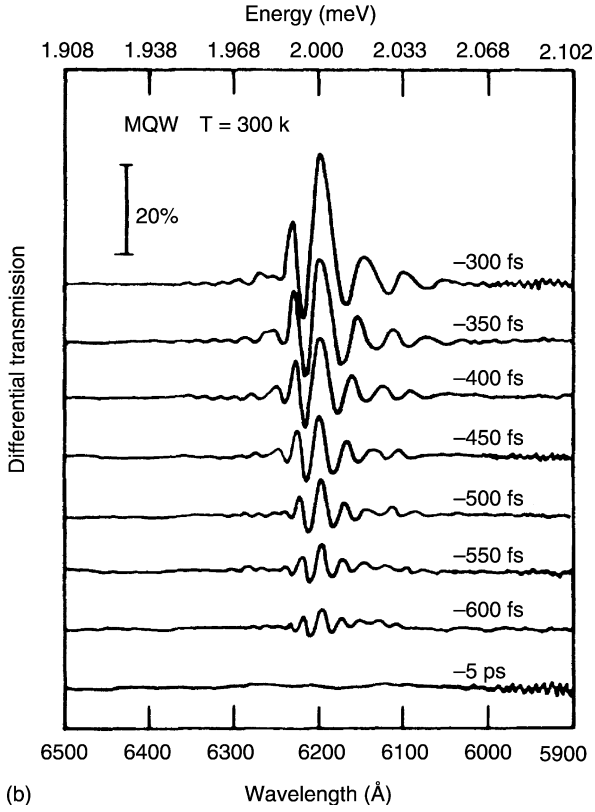
One can easily understand the coherent control in the frequency domain by considering the overlap between the absorption of the system and the pulse spectrum. For excitation with two pulses that are temporally delayed by τ, the spectrum of the excitation shows interference fringes with a spectral oscillation period that is inversely proportional to τ. The positions of the maxima and the minima of the fringes depend on the phase difference between the two pulses. For constructive interference, the excitation spectrum is at a maximum at the resonance of the system, whereas for destructive interference the excitation spectrum vanishes at the resonance of the system (see Figure 5a).

Coherent control techniques have been applied to molecules and solids to control the dynamical evolution of electronic wavepackets and also the coupling to nuclear degrees of freedom. In this context, it is sometimes possible to steer certain chemical reactions into a preferred direction by using sequences of laser pulses which can be chirped, i.e., have a time-dependent frequency. Furthermore, in semiconductors and heterostructures, coherent control has been used to optically inject electronic currents on an ultrafast time-scale.

Transient Absorption Changes

In a typical pump-probe experiment one excites the system with a pump pulse (E_p) and probes its dynamics with a weak test pulse (E_t) (see Figure 6). With such experiments one often measures the differential absorption $\Delta\alpha(\omega)$ which is defined as





the difference between the probe absorption in the presence of the pump $\alpha_{\text{pump on}}(\omega)$ and the probe absorption without the pump $\alpha_{\text{pump off}}(\omega)$.

For resonant pumping and for a situation where the pump precedes the test (positive time delays $\tau > 0$), the absorption change is usually negative in the vicinity of the resonance frequency $\Delta\omega$, indicating the effect of absorption bleaching (see Figure 7a,b). There may be positive contributions spectrally around the original absorption line due to resonance broadening and, at other spectral positions, due to excited state absorption, i.e., optical transitions to energetically higher states which are only possible if the system is in an excited state. The bleaching and the positive contributions are generally present in coherent, and also in incoherent situations, where the polarization vanishes but occupations in excited states are present.

For detuned pumping, the resonance may be shifted by the light field as for example, in the optical Stark effect. Depending on the excitation configuration and the system, this transient shift may be to higher (blue shift) or lower energies (red shift) (see Figure 7a,b). With increasing pump-probe time delay the system gradually returns to its unexcited state and the absorption changes disappear.

As an illustration we show in Figure 7c experimentally measured differential absorption spectra of a semiconductor quantum well which is pumped spectrally below the exciton resonance. For a two-level system one would expect a blue shift of the absorption, i.e., a dispersive shape of the differential absorption with positive contributions above and negative contributions below the resonance frequency. This is indeed observed for most polarization directions of the pump and probe pulses. However, if both pulses are oppositely circularly polarized, the experiment shows a resonance shift in the reverse direction, i.e., a red shift. For the explanation of this behavior one has to consider the optical selection rules of the quantum well system. One finds that the signal should actually vanish for oppositely circularly

Figure 8 (a) Differential absorption spectra are shown for resonant pumping of a two-level system for various time delays τ between the pump and the test pulse. When the pump precedes the test, i.e., $\tau < 0$, the differential absorption exhibits spectral oscillations with a period which is inversely proportional to the time delay. When τ approaches zero, these spectral oscillations vanish and the differential absorption develops into purely negative bleaching. (b) Measured differential transmission spectra of a multi quantum well for different negative time delays as indicated. [Part (b) is reproduced with permission from Sokoloff JK, Joffre M, Fluegel B, *et al.* (1988) Transient oscillations in the vicinity of excitons and in the band of semiconductors. *Physics Review B* 38: 7615. Copyright (1988) by the American Physical Society.]

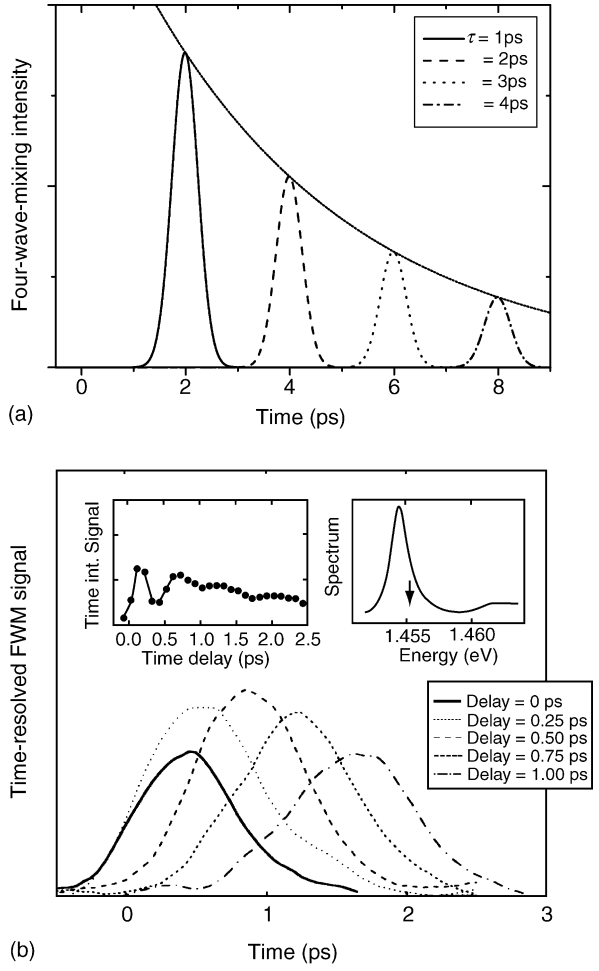


Figure 9 (a) The intensity dynamics as measured in a four-wave-mixing experiment on an ensemble of inhomogeneously broadened two-level systems for various time delays τ of the two incident pulses. The first pulse excites the system at $t = 0$. In the time period $0 < t < \tau$ the individual polarizations of the inhomogeneously broadened system oscillate with their respective frequencies and due to destructive interference the total polarization decays, cp. **Figure 3**. The second pulse arriving at $t = \tau$ leads to a partial rephasing of the individual polarizations. Due to phase conjugation all individual polarizations are in phase at $t = 2\tau$ which results in a macroscopic measurable signal, the photon echo. Due to dephasing processes the magnitude of the photon echo decays with increasing τ, see solid, dashed, dotted and dashed-dotted lines. Describing the dephasing via a T_2 time corresponds to an exponential decay as indicated by the thin dotted line. By measuring the decay of the echo with increasing τ, one thus gets experimental information on the dephasing of the optical polarization. (b) Time-resolved four-wave-mixing signal measured on an inhomogeneously broadened exciton resonance of a semiconductor quantum well structure for different time delays as indicated. The origin of the time axis starts at the arrival of the second pulse, i.e., for a non-interacting system the maximum of the echo is expected at $t = \tau$. The insets show the corresponding time-integrated four-wave-mixing signal and the linear absorption. [Part (b) is reproduced with permission from Jahnke F, Koch M, Meier T, *et al.* (1994) Simultaneous influence of disorder and Coulomb interaction on photon echoes in semiconductors. *Physics Review B* 50: 8114. Copyright (1994) by the American Physical Society.]

polarized pulses, as long as many-body correlations are neglected. Thus in this case the entire signal is due to many-body correlations and it is their dynamics which gives rise to the appearance of the red shift.

Spectral Oscillations

For negative time delays $\tau < 0$, i.e., if the test pulse precedes the pump, the absorption change $\Delta\alpha(\omega)$ is characterized by spectral oscillations around $\Delta\omega$ which vanish as τ approaches zero (see **Figure 8a**). The spectral period of the oscillations decreases with increasing $|\tau|$. **Figure 8b** shows measured differential transmission spectra of a multiple quantum-well structure for different negative time delays. As with the differential absorption, also the differential transmission spectra are dominated by spectral oscillations around the exciton resonance whose period and amplitude decrease with increasing $|\tau|$.

The physical origin of the coherent oscillations is the pump-induced perturbation of the free-induction decay of the polarization generated by the probe pulse. This perturbation is delayed by the time τ at which the pump arrives. The temporal shift of the induced polarization changes in the time domain leads to oscillations in the spectral domain, since the Fourier transformation translates a delay in one domain into a phase factor in the conjugate domain.

Photon Echo

In the nonlinear optical regime one may (partially) reverse the destructive interference of a coherent, inhomogeneously broadened polarization. For example, in four-wave mixing, which is often performed with two incident pulses, one measures the emitted field in a background-free scattering direction (see **Figure 5**). The first short laser pulse excites all transitions at $t = 0$. As a result of the inhomogeneous broadening the polarization decays due to destructive interference (see **Figure 3**). The second pulse arriving at $t = \tau > 0$ is able to conjugate the phases ($e^{i\varphi} \rightarrow e^{-i\varphi}$) of the individual polarizations of the inhomogeneously broadened system. The subsequent unperturbed dynamical evolution of the polarizations leads to a measurable macroscopic signal at $t = 2\tau$. This photon echo occurs since at this point in time all individual polarizations are in phase and add up constructively (see **Figure 9a**). Since this rephasing process leading to the echo is only possible as long as the individual polarizations remain coherent, one can analyze the loss of coherence (dephasing) by measuring the decay of the photon echo with increasing time delay.

Echo experiments are frequently performed using more complicated geometries, e.g., more than

two laser pulses. Further echo phenomena can also be observed as the result of the coupling of the photo-excitation to nuclear degrees of freedom and also due to spatial motion of the electrons, i.e., electronic currents.

Time-resolved four-wave-mixing signal, measured on an inhomogeneously broadened exciton resonance of a semiconductor quantum well structure for different time delays, is presented in Figure 9b. Compared to Figure 9a, the origin of the time axis starts at the arrival of the second pulse, i.e., for a non-interacting system the maximum of the echo is expected at $t = \tau$. Due to the inhomogeneous broadening the maximum of the signal is shifting to longer times with increasing delay. Further details of the experimental results, in particular, the exact position of the maximum and the width of the signal, cannot be explained on the basis of non-interacting two-level systems, but require the analysis of many-body effects in the presence of inhomogeneous broadening.

See also

Coherent Control: Applications in Semiconductors; Experimental; Theory. **Spectroscopy:** Nonlinear Laser Spectroscopy.

Further Reading

Allen L and Eberly JH (1975) *Optical Resonance and Two-Level Atoms*. New York: Wiley.

Baumert T, Helbing J, Gerber G, *et al.* (1997) Coherent control with femtosecond laser pulses. *Advanced Chemical Physics* 101: 47–82.

Bergmann K, Theuer H and Shore BW (1998) Coherent population transfer among quantum states of atoms and molecules. *Review of Modern Physics* 70: 1003–1025.

Bloembergen N (1965) *Nonlinear Optics*. New York: Benjamin.

Brewer RG (1977) *Coherent Optical Spectroscopy*. In: Balian, *et al.* (ed.). Course 4 in Les Houches Session XXVII, *Frontiers in Laser Spectroscopy*, vol. 1. Amsterdam: North-Holland Publ.

Cohen-Tannoudji C, Dupont-Roc J and Grynberg G (1989) *Photons and Atoms*. New York: Wiley.

Haug H and Koch SW (2004) *Quantum Theory of the Optical and Electronic Properties of Semiconductors*, 4th edn. Singapore: World Scientific.

Koch SW, Peyghambarian N and Lindberg M (1988) Transient and steady-state optical nonlinearities in semiconductors. *Journal of Physics* C 21: 5229–5250.

Macomber JD (1976) *The Dynamics of Spectroscopic Transitions*. New York: Wiley.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. New York: Oxford.

Peyghambarian N, Koch SW and Mysyrowicz A (1993) *Introduction to Semiconductor Optics*. Englewood Cliffs, NJ: Prentice-Hall.

Pötz W and Schroeder WA (eds) (1999) *Coherent Control in Atoms, Molecules, and Semiconductors*. Dordrecht: Kluwer.

Schäfer W and Wegener M (2002) *Semiconductor Optics and Transport Phenomena*. Berlin: Springer.

Shah J (1999) *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*, 2nd edn. Berlin: Springer.

Yeazell JA and Uzer T (eds) (2000) *The Physics and Chemistry of Wave Packets*. New York: Wiley.

Ultrafast Studies of Semiconductors

J Shah, Arlington, VA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Linear optical spectroscopy, using the techniques of absorption, transmission, reflection and light scattering, has provided invaluable information about the electronic and vibrational properties of atoms, molecules, and solids. Optical techniques also possess some additional unique strengths: the ability to

1. generate nonequilibrium distributions functions of electrons, holes, excitons, phonons, etc. in solids;
2. determine the distribution functions by optical spectroscopy;
3. determine the nonequilibrium distribution functions on femtosecond timescales;
4. determine femtosecond dynamics of carrier and exciton transport and tunneling; and
5. investigate interactions between various elementary excitations as well as many-body processes in semiconductors.

Researchers have exploited these unique strengths to gain fundamental new insights into nonequilibrium, nonlinear, and transport physics of semiconductors and their nanostructures over the past four decades.

A major focus of these efforts has been devoted to understanding how a semiconductor in

thermodynamic equilibrium, excited by an ultrashort optical pulse, returns to the state of thermodynamic equilibrium. Four distinct regimes can be identified:

(a) the coherent regime in which a well-defined phase relationship exists between the excitation created by the optical pulse and the electromagnetic (optical) field creating it;
(b) the nonthermal regime in which the coherence (well-defined phase relationships) has been destroyed by various collision and interference processes but the distribution function of excitations is nonthermal (i.e., cannot be described by a Maxwell–Boltzmann distribution or its quantum (degenerate) counterparts);
(c) the hot carrier regime in which the distributions for various excitations are thermal, but with different characteristic temperatures for different excitations and thermal bath; and
(d) the isothermal regime in which the excitations and the thermal bath are at the same temperature but there is an excess of excitation (e.g., electron–hole pairs) compared to thermodynamic equilibrium.

Various physical processes take the semiconductor from one regime to the next, and provide information not only about the fundamental physics of semiconductors but also about the physics and ultimate performance limits of electronic, optoelectronic, and photonic devices.

In addition to coherent and relaxation dynamics, ultrafast studies of semiconductors provide new insights into tunneling and transport of carriers, and demonstrate novel quantum mechanical phenomena and coherent control in semiconductors.

The techniques used for such studies have also advanced considerably over the last four decades. Ultrafast lasers with pulse widths corresponding to only a few optical cycles (1 optical cycle ≈ 2.7 fs at 800 nm) have been developed. The response of the semiconductor to an ultrafast pulse, or a multiple of phase-locked pulses, can be investigated by measuring the dynamics of light emitted by the semiconductor using a streak camera (sub-ps time resolution) or a nonlinear technique such as luminescence up-conversion (time resolution determined by the laser pulse width for an appropriate nonlinear crystal). The response of the semiconductor can also be investigated using a number of two or three beam pump-probe techniques such as transmission, reflection, light scattering or four-wave mixing (FWM). Both the amplitude and the phase of the emitted radiation or the probe can be measured by phase-sensitive techniques such as spectral interferometry. The lateral transport of excitation can be investigated by using time-resolved spatial imaging and the vertical transport and tunneling of carriers can be investigated using the technique of 'optical markers'. Electro-optic sampling can be used for transmission/reflection studies or for measuring THz response of semiconductors. More details on these techniques, and indeed many topics discussed in this brief article, can be found in the Further Reading.

Coherent Dynamics

The coherent response of atoms and molecules is generally analyzed for an ensemble of independent (noninteracting) two-level systems. The statistical properties of the ensemble are described in terms of the density matrix operator whose diagonal components relate to population of the eigenstates and off-diagonal components relate to coherence of the superposition state. The time evolution of the density matrix is governed by the Liouville variant of the Schrödinger equation $i\hbar\dot{\rho} = [H, \rho]$ where the system Hamiltonian $H = H_0 + H_{\text{int}} + H_R$ is the sum of the unperturbed, interaction (between the radiation field and the two-level system), and relaxation Hamiltonians, respectively. Using a number of different assumptions and approximations, this equation of motion is transformed into the optical Bloch equations (OBE), which are then used to predict the coherent response of the system, often using iterative procedures, and analyze experimental results.

Semiconductors are considerably more complex. Coulomb interaction profoundly modifies the response of semiconductors, not only in the linear regime (e.g., strong exciton resonance in the absorption spectrum) but also in the coherent and nonlinear regime. The influence of Coulomb interaction may be introduced by renormalizing the electron and hole energies (i.e., introducing excitons), and by renormalizing the field–matter interaction strength by introducing a renormalized Rabi frequency. These changes and the relaxation time approximation for the relaxation Hamiltonian lead to an analog of the optical Bloch equations, the semiconductor Bloch equations which have been very successful in analyzing the coherent response of semiconductors.

In spite of this success, it must be stressed that the relaxation time approximation, and the Boltzmann kinetic approach on which it is based, are not valid under all conditions. The Boltzmann kinetic approach is based on the assumption that the duration of the collision is much shorter than the interval between the collisions; i.e., the collisions are instantaneous. In the non-Markovian regime where these assumptions are not valid, each collision does not strictly

conserve the energy and momentum and the quantum kinetic approach becomes more appropriate. This is true for photo-excited semiconductors as well as for the quantum transport regime in semiconductors. Also, semiconductor Bloch equations have been further extended by including four-particle correlations. Finally, the most general approach to the description of the nonequilibrium and coherent response of semiconductors following excitation by an ultrashort laser pulse is based on nonequilibrium Green's functions.

Experimental studies on the coherent response of semiconductors can be broadly divided into three categories:

1. investigation of novel aspects of coherent response not found in other simpler systems such as atoms;
2. investigation of how the initial coherence is lost, to gain insights into dephasing and decoherence processes; and
3. coherent control of processes in semiconductors using phase-locked pulses.

Numerous elegant studies have been reported. This section presents some observations on the trends in this field.

Historically, the initial experiments focused on the decay of the coherent response (time-integrated FWM as a function of delays between the pulses), analyzed them in terms of the independent two-level model, and obtained very useful information about various collision processes and rates (exciton–exciton, exciton–carrier, carrier–carrier, exciton–phonon, etc.). It was soon realized, however, that the noninteracting two-level model is not appropriate for semiconductors because of the strong influence of Coulomb interactions. Elegant techniques were then developed to explore the nature of coherent response of semiconductors. These included investigation of time- and spectrally resolved coherent response, and of both the amplitude and phase of the response to complement intensity measurements. These studies provided fundamental new insights into the nature of semiconductors and many-body processes and interactions in semiconductors. Many of these observations were well explained by the semiconductor Bloch equations. When lasers with <10 fs pulse widths became laboratory tools, the dynamics was explored on a time-scale much shorter than the characteristic phonon oscillation period (≈ 115 fs for GaAs longitudinal optical phonons), the plasma frequency (≈ 150 fs for a carrier density of $5 \times 10^{17}\,\mathrm{cm}^{-3}$), and the typical electron–phonon collision interval (≈ 200 fs in GaAs). Some remarkable features of quantum kinetics, such as reversal of the electron–phonon collision, memory effects, and energy nonconservation, were demonstrated. More recent experiments have demonstrated the influence of four-particle correlations on coherent nonlinear response of semiconductors such as GaAs in the presence of a strong magnetic field.

The ability to generate phase-locked pulses with controllably variable separation between them provides an exciting opportunity to manipulate a variety of excitations and processes within a semiconductor. If the separation between the two phase-locked pulses is adjusted to be less than the dephasing time of the system under investigation, then the amplitudes of the excitations produced by the two phase-locked pulses can interfere either constructively or destructively depending on the relative phase of the carrier waves in the two pulses. The nature of interference changes as the second pulse is delayed over an optical period. A number of elegant experiments have demonstrated coherent control of exciton population, spin orientation, resonant emission, and electrical current. Phase-sensitive detection of the linear or nonlinear emission provides additional insights into different aspects of semiconductor physics.

These experimental and theoretical studies of ultrafast coherent response of semiconductors have led to fundamental new insights into the physics of semiconductors and their coherence properties. Discussion of novel coherent phenomena is given in a subsequent section. Coherent spectroscopy of semiconductors continues to be a vibrant research field.

Incoherent Relaxation Dynamics

The qualitative picture that emerges from investigation of coherent dynamics can be summarized as follows. Consider a direct gap semiconductor excited by an ultrashort laser pulse of duration τ_L (with a spectral width $h\Delta\nu_L$) centered at photon energy $h\nu_L$ larger than the semiconductor bandgap E_g. At the beginning of the pulse the semiconductor does not know the pulse duration so that coherent polarization is created over a spectral region much larger than $h\Delta\nu_L$. If there are no phase-destroying events during the pulse (dephasing rate $1/\tau_D \ll 1/\tau_L$), then a destructive interference destroys the coherent polarization away from $h\nu_L$ with increasing time during the excitation pulse. Thus, the coherent polarization exists only over the spectral region $h\Delta\nu_L$ at the end of the pulse. This coherence will be eventually destroyed by collisions and the semiconductor will be left in an incoherent (off-diagonal elements of the density matrix are 0) nonequilibrium state with peaked electron and hole population distributions whose central energies and energy widths are

determined by $h\nu_L$, $h\Delta\nu_L$ and E_g, if the energy and momentum relaxation rates $(1/\tau_E, 1/\tau_M)$ are much smaller than the dephasing rates.

The simplifying assumptions that neatly separate the excitation, dephasing and relaxation regimes are obviously not realistic. A combination of all these (and some other) physical parameters will determine the state of a realistic system at the end of the exciting pulse. However, after times $\sim \tau_D$ following the laser pulse, the semiconductor can be described in terms of electron and hole populations whose distribution functions are not Maxwell–Boltzmann or Fermi–Direct type, but nonthermal in a large majority of cases in which the energy and momentum relaxation rates are much smaller than the dephasing rates. This section discusses the dynamics of this incoherent state.

Extensive theoretical work has been devoted to quantitatively understand these initial athermal distributions and their further temporal evolution. These include Monte Carlo simulations based on the Boltzmann kinetic equation approach (typically appropriate for time-scales longer than $\sim$100 fs, the carrier–phonon interaction time) as well as the quantum kinetic approach (for times typically less than $\sim$100 fs) discussed above. We present below some observations on this vast field of research.

Nonthermal Regime

A large number of physical processes determines the evolution of the nonthermal electron and hole populations generated by the optical pulse. These include electron–electron, hole–hole, and electron–hole collisions, including plasma effects if the density is high enough, intervalley scattering in the conduction band and intervalence band scattering in valence bands, intersub-band scattering in quantum wells, electron–phonon, and hole–phonon scattering processes. The complexity of the problem is evident when one considers that many of these processes occur on the same time-scale. Of these myriad processes, only carrier–phonon interactions and electron–hole recombinations (generally much slower) can alter the total energy in electronic systems. Under typical experimental conditions, the redistribution of energy takes place before substantial transfer of energy to the phonon system. Thus, the nonthermal distributions first become thermal distributions with the same total energy. Investigation of the dynamics of how the nonthermal distribution evolves into a thermal distribution provides valuable information about the nature of various scattering processes (other than carrier–phonon scattering) described above.

This discussion of separating the processes that conserve energy in the electronic system and those that transfer energy to other systems is obviously too simplistic and depends strongly on the nature of the problem one is considering. The challenge for the experimentalist is to devise experiments that can isolate these phenomena so each can be studied separately. If the laser energy is such that $h\nu_L - E_g < \hbar\omega_{LO}$, the optical phonon energy, then majority of the photo-excited electrons and holes do not have sufficient energy to emit an optical phonon, the fastest of the carrier–phonon interaction processes. The initial nonthermal distribution is then modified primarily by processes other than phonon scattering and can be studied experimentally without the strong influence of the carrier–phonon interactions. Such experiments have indeed been performed both in bulk and quantum well semiconductors. These experiments have exhibited spectral hole burning in the pump-probe transmission spectra and thus demonstrated that the initial carrier distributions are indeed nonthermal and evolve to thermal distributions. Such experiments have provided quantitative information about various carrier–carrier scattering rates as a function of carrier density. In addition, experiments with $h\nu_L - E_g > \hbar\omega_{LO}$ and $h\nu_L - E_g >$ intervalley separation have provided valuable information about intravalley as well as intervalley electron–phonon scattering rates. Experiments have been performed in a variety of different semiconductors, including bulk semiconductors and undoped and modulation-doped quantum wells. The latter provide insights into intersub-band scattering processes and quantitative information about the rates.

These measurements have shown that under typical experimental conditions, the initial nonthermal distribution evolves into a thermal distribution in times $<$ or $\sim$1 ps. However, the characteristic temperatures of the thermal distributions can be different for electrons and holes. Furthermore, different phonons may also have different characteristic temperatures, and may even be nonthermal even when the electronic distributions are thermal. This is the hot carrier regime that is discussed in the next subsection.

Hot Carrier Regime

The hot carriers, with characteristic temperatures T_c (T_e and T_h for electrons and holes, respectively), have high energy tails in the distribution functions that extend several kT_c higher than the respective Fermi energies. Some of these carriers have sufficient energy to emit an optical phonon. Since the carrier–optical phonon interaction rates and the phonon energies are

rather large, this may be the most effective energy loss mechanism in many cases even though the number of such carriers is rather small. The emitted optical phonons have relatively small wavevectors and nonequilibrium populations larger than expected for the lattice temperature. These optical phonons are often referred to as hot phonons although nonequilibrium phonons may be a better term since they probably have nonthermal distributions. The dynamics of hot phonons will be discussed in the next subsection, but we mention here that in case of a large population of hot phonons, one must consider not only phonon emission but also phonon absorption. Acoustic phonons come into the picture at lower temperatures when the fraction of carriers that can emit optical phonons is extremely small or negligible, and also in the case of quantum wells with sub-band energy separation smaller than optical phonons.

The dynamics of hot carriers have been studied extensively to obtain a deeper understanding of carrier–phonon interactions and to obtain quantitative measures of the carrier–phonon interaction rates. Time resolved luminescence and pump-probe transmission spectroscopy are the primary tools used for such studies. For semiconductors like GaAs, the results indicate that polar optical phonon scattering (longitudinal optical phonons) dominates for electrons whereas polar and nonpolar optical phonon scattering contribute for holes. Electron–hole scattering is sufficiently strong in most cases to maintain a common electron and hole temperature for times $> \sim 1$ ps. Since many of these experiments are performed at relatively high densities, they provide important information about many-body effects such as screening. Comparison of bulk and quasi-two-dimensional semiconductors (quantum wells) has been a subject of considerable interest. The consensus appears to be that, in spite of significant differences in the nature of electronic states and phonons, similar processes with similar scattering rates dominate both types of semiconductors. These studies reveal that the energy loss rates are influenced by many factors such as the Pauli exclusion principle, degenerate (Fermi–Dirac) statistics, hot phonons, and screening and many-body aspects.

Hot carrier distribution can be generated not only by photo-excitation, but also by applying an electric field to a semiconductor. Although the process of creating the hot distributions is different, the processes by which such hot carriers cool to the lattice temperature are the same in two cases. Therefore, understanding obtained through one technique can be applied to the other case. In particular, the physical insights obtained from the optical excitation case can be extremely valuable for many electronic devices that operate at high electric fields, thus in the regime of hot carriers. Another important aspect is that electrons and holes can be investigated separately by using different doping. Furthermore, the energy loss rates can be determined quantitatively because the energy transferred from the electric field to the electrons or holes can be accurately determined by electrical measurements.

Isothermal Regime

Following the processes discussed above, the excitations in the semiconductor reach equilibrium with each other and the thermal bath. Recombination processes then return the semiconductor to its thermodynamic equilibrium. Radiative recombination typically occurs over a longer time-scale but there are some notable exceptions such as radiative recombination of excitons in quantum wells occurring on picosecond time-scales.

Hot Phonons

As discussed above, a large population of nonequilibrium optical phonons is created if a large density of hot carriers has sufficient energy to emit optical phonons. Understanding the dynamics of these nonequilibrium optical phonons is of intrinsic interest.

At low lattice temperatures, the equilibrium population of optical phonons is insignificant. The optical phonons created as a result of photoexcitation occupy a relatively narrow region of wavevector (k) space near $k = 0$ ($\sim 1\%$ of the Brillouin zone). This nonequilibrium phonon distribution can spread over a larger k-space by various processes, or anharmonically decay into multiple large-wavevector acoustic phonons. These acoustic phonons then scatter or decay into small-wavevector, low-energy acoustic phonons that are eventually thermalized.

Pump-probe Raman scattering provides the best means of investigating the dynamics of nonequilibrium optical phonons. Such studies, for bulk and quantum well semiconductors, have provided invaluable information about femtosecond dynamics of phonons. In particular, they have provided the rate at which the nonequilibrium phonons decay. However, this information is for one very small value of the phonon wavevector so the phonon distribution function within the optical phonon branch is not investigated. The large-wavevector acoustic phonons are even less accessible to experimental techniques. Measurements of thermal phonon propagation provide some information on this subject. Finally, the nature of phonons is considerably more complex in

quantum wells, and some interesting results have been obtained on the dynamics of these phonons.

Tunneling and Transport Dynamics

Ultrafast studies have also made important contributions to tunneling and transport dynamics in semiconductors. Time-dependent imaging of the luminescing region of a sample excited by an ultrashort pulse provides information about spatial transport of carriers. Such a technique was used, for example, to demonstrate negative absolute mobility of electrons in p-modulation-doped quantum wells. The availability of near-field microscopes enhances the resolution of such techniques to measure lateral transport to the subwavelength case.

A different technique of 'optical markers' has been developed to investigate the dynamics of transport and tunneling in a direction perpendicular to the surface (vertical transport). This technique relies heavily on fabrication of semiconductor nanostructures with desired properties. The basic idea is to fabricate a sample in which different spatial regions have different spectral signatures. Thus if the carriers are created in one spatial region and are transported to another spatial region under the influence of an applied electric field, diffusion, or other processes, the transmission, reflection, and/or luminescence spectra of the sample change dynamically as the carriers are transported.

This technique has provided new insights into the physics of transport and tunneling. Investigation of perpendicular transport in graded-gap superlattices showed remarkable time-dependent changes in the spectra, and analysis of the results provided new insight into transport in a semiconductor of intermediate disorder. Dynamics of carrier capture in quantum wells from the barriers provided information not only about the fundamentals of capture dynamics, but also about how such dynamics affects the performance of semiconductor lasers with quantum well active regions.

The technique of optical markers has been applied successfully to investigate tunneling between the two quantum wells in an a-DQW (asymmetric double quantum well) structure in which two quantum wells of different widths (and hence different energy levels) are separated by a barrier. The separation between the energy levels of the system can be varied by applying an electric field perpendicular to the wells. In the absence of resonance between any energy levels, the wavefunction for a given energy level is localized in one well or the other. In this nonresonant case, it is possible to generate carriers in a selected quantum well by proper optical excitation. Dynamics of transfer of carriers to the other quantum well by tunneling can then be determined by measuring dynamic changes in the spectra. Such measurements have provided much insight into the nature and rates of tunneling, and demonstrated phonon resonances. Resonant tunneling, i.e., tunneling when two electronic levels are in resonance and are split due to strong coupling, has also been investigated extensively for both electrons and holes. One interesting insight obtained from such studies is that tunneling and relaxation must be considered in a unified framework, and not as sequential events, to explain the observations.

Novel Coherent Phenomena

Ultrafast spectroscopy of semiconductors has provided valuable insights into many other areas. We conclude this article by discussing some examples of how such techniques have demonstrated novel physical phenomena.

The first example once again considers an a-DQW biased in such a way that the lowest electron energy level in the wide quantum well is brought into resonance with the first excited electron level in the narrow quantum well. The corresponding hole energy levels are not in resonance. By choosing the optical excitation photon energy appropriately, the hole level is excited only in the wide well and the resonant electronic levels are excited in a linear superposition state such that the electrons at $t = 0$ also occupy only the wide quantum well. Since the two electron eigenstates have slightly different energies, their temporal evolutions are different with the result that the electron wavepacket oscillates back and forth between the two quantum wells in the absence of damping. The period of oscillation is determined by the splitting between the two resonant levels and can be controlled by an external electric field. Coherent oscillations of electronic wavepackets have indeed been experimentally observed by using the coherent nonlinear technique of four-wave mixing. Semiconductor quantum wells provide an excellent flexible system for such investigations.

Another example is the observation of Bloch oscillations in semiconductor superlattices. In 1928 Bloch demonstrated theoretically that an electron wavepacket composed of a superposition of states from a single energy band in a solid undergoes a periodic oscillation in energy and momentum space under certain conditions. An experimental demonstration of Bloch oscillations became possible only recently by applying ultrafast optical techniques to semiconductor superlattices. The large period of a superlattice (compared to the atomic period in a

solid) makes it possible to satisfy the assumptions underlying Bloch's prediction. The experimental demonstration of Bloch oscillations was also performed using four-wave mixing techniques. This provides another prime example of the power of ultrafast optical studies.

Summary

Ultrafast spectroscopy of semiconductors and their nanostructures is an exciting field of research that has provided fundamental insights into important physical processes in semiconductors. This article has attempted to convey the breadth of this field and the diversity of physical processes addressed by this field. Many exciting developments in the field have been omitted out of necessity. The author hopes that this brief article inspires the readers to explore some of the Further Reading.

See also

Coherence: Coherence and Imaging. **Coherence Control:** Experimental; Theory. **Coherent Transients:** Ultrafast Studies of Semiconductors. **Nonlinear Optics, Basics:** Four-Wave Mixing. **Scattering:** Scattering Theory.

Further Reading

Allen L and Eberly JH (1975) *Optical Resonance and Two-Level Atoms*. New York: Wiley Interscience.

Chemla DS (1999) Chapter 3. In: Garmire E and Kost AR (eds) *Nonlinear Optics in Semiconductors*, pp. 175–256. New York: Academic Press.

Chemla DS and Shah J (2001) Many-body and correlation effects in semiconductors. *Nature* 411: 549–557.

Haug H (2001) Quantum kinetics for femtosecond spectroscopy in semiconductors. In: Willardson RK and Weber ER (eds) *Semiconductors and Semimetals*, vol. 67, pp. 205–229. London: Academic Press.

Haug H and Jauho AP (1996) *Quantum Kinetics in Transport and Optics of Semiconductors*. Berlin: Springer-Verlag.

Haug H and Koch SW (1993) *Quantum Theory of the Optical and Electronic Properties of Semiconductors*. Singapore: World Scientific.

Kash JA and Tsang JC (1992) Chapter 3. In: Shank CV and Zakharchenya BP (eds) *Spectroscopy of Nonequilibrium Electrons and Phonons*, pp. 113–168. Amsterdam: Elsevier.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. New York: Oxford University Press.

Shah J (1992) *Hot Carriers in Semiconductor Nanostructures: Physics and Applications*. Boston, MA: Academic Press.

Shah J (1999) *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*. 2nd edn. Heidelberg, Germany: Springer-Verlag.

Steel DG, Wang H, Jiang M, Ferrio KB and Cundiff ST (1994) In: Phillips RT (ed.) *Coherent Optical Interactions in Solids*, pp. 157–180. New York: Plenum Press.

Wegener M and Schaefer W (2002) *Semiconductor Optics and Transport*. Heidelberg, Germany: Springer-Verlag.

Zimmermann R (1988) *Many-Particle Theory of Highly Excited Semiconductors*. Leipzig, Germany: Teubner.

COLOR AND THE WORLD

G N Rao, Adelphi University, Garden City, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

The human visual system is capable of detecting a small part of the electromagnetic spectrum which is called light. The visible region has a typical wavelength range of about 380–780 nm. Under optimal conditions the range is 360–830 nm. The most important point is that color is a perception. It is how human beings perceive the visual part of the electromagnetic spectrum. The color of an object perceived depends on the physical characteristics of the radiation illuminating the object, spectral scattering and reflectivity of the object, physiology of the eye, photochemistry of the sensors in the retina, the complex data processing that takes place in the visual system, and the psychology of color perception. In this review, we present an overview of color in nature and the world. Initially, color and wavelength is presented, then the human visual system and color perception. Then we go on to discuss measuring color and colorimetry and then color in nature, followed by a brief account of color in art and holography and the structural colors in nature.

Color and Wavelength (Spectral and Nonspectral Colors)

The ingenious yet simple experiments performed by Isaac Newton in the year 1666 and his ideas presented

in his famous treatise 'Opticks', are the foundations of our understanding of color and color perception. Based on his experiments, Newton showed that different regions of the visible spectrum are perceived as different colors and the solar radiation (white light) consists of spectral colors violet, blue, cyan, green, yellow, orange, and red. Thus, natural day light is quite colorful though we perceive it as colorless. Table 1 gives approximate wavelength ranges of the spectral colors. Superposition of all the spectral colors results in a perception of white.

Colors can be produced in a number of ways:

1. *Gas discharges* emit characteristic radiation in the visible region. For example, neon discharge is red in color, argon discharge blue, sodium discharge yellow, helium-neon laser gives red radiation and argon ion laser blue-green radiation. A variety of fluorescent lamps are available and they provide continuous spectral power distribution with characteristic lines of mercury in the blue-green region. Gas-filled sodium and fluorescent lamps offer long life and high efficiency and are popular for indoor and outdoor lighting.
2. *Transition metal* (Cr, Mn, Fe, Co, Ni, Cu) compounds or transition metal impurities are responsible for the colors of the minerals, paints, gems, and pigments. They involve inorganic compounds of transition metal or rare earth ions with unpaired d or f orbitals. About 1% chromium sesquioxide (Cr_2O_3) in colorless aluminum oxide results in beautiful red ruby, a gemstone. The Cr^{3+} impurity in pure beryllium aluminum silicate gives a beautiful emerald green resulting in sapphire, another gemstone. Transition metal impurities are added to glass in the molten state to prepare colored glass. For example, adding Cr^{3+} gives green, Mn^{3+} purple, Fe^{3+} pale yellow, Co^{3+} reddish blue, Ni^{2+} brown, and Cu^{2+} greenish blue color.
3. *Color in Organic Compounds* – Most natural and synthetic dyes and biological (vegetable and animal origin) colorants are complicated organic molecules. They absorb light, exciting molecular orbitals. The dyes are used for food coloration, clothing, photographic, and sensitizer purposes.
4. *Colors of metals, semiconductor materials and color centers* – The colors of these materials are caused by the electronic transitions involving the energy bands.

   Diode lasers emit radiation corresponding to the bandgap of the semiconductor materials used for their fabrication. They are inexpensive, compact, and can be operated using dry cell batteries. They are extensively used in compact disks, barcode readers, and optical communications.

   Light emitting diodes (LED) are also prepared using semiconductor materials and they emit radiation corresponding to the bandgap of the semiconductor materials used for their fabrication. LEDs are available in a variety of colors, red (GaAsP), orange (AlInGaP), yellow (InGaAs), green (CdSSe), blue (CdZnS) and are widely used in clocks, toys, tuners, displays, electronic billboards, and appliances.

   Color centers may be produced by irradiating some alkali halides and glass materials with electromagnetic radiation (gamma rays, X-rays, ultraviolet, and visible radiation) or with charged or uncharged particles (electrons, protons, neutrons). The irradiation produces an electron hole pair and the electron is trapped forming an electron center. The electron or hole, or both, form a color center absorbing part of the visible radiation. The color centers can be reversed by heating to high temperatures.
5. *Optical Phenomenon* lead to spectacular colors in nature and in biological materials. Scattering of light by the atmosphere is responsible for colors of the sky, beautiful sunsets and sunrise, twilight, blue moon, and urban glows. Dispersion and polarization cause rainbows and halos. Interference gives rise to the beautiful colors of thin films on water, foam, bubbles, and some biological colors (butterflies). Diffraction is responsible for the colors of liquid crystals, coronae, color of gems (opal), color of animals (sea mouse), the colors of butterfly (*Morpho rhetenor*), and diffraction gratings.

Table 1 Approximate wavelength ranges of different spectral colors

Color	*Wavelength range (nm)*
Violet	380–435
Blue	435–500
Cyan	500–520
Green	520–565
Yellow	565–597
Orange	597–625
Red	625–780

The Human Visual System and Color Perception

Photoreceptors in the Eye: Rods and Cones

As light enters the eye, the cornea and the eye lens focus the image on the retina. The retina has two kinds of photoreceptors: rods and cones so named

because of their shapes. Typically, there are about 7 million cones and 110–125 million rods. Cones are activated during photopic (daylight) conditions and the rods during scotopic (night) conditions. Cones are responsible for color perception whereas rods can perceive only white and gray shades. Cones are predominantly located near the fovea, the central region of the retina. When light falls on the rods and the cones, the photosensitive material in them, namely rhodopsin, is activated, generating an electrical signal which is further processed in the visual system and transmitted to the visual cortex. Considerable data processing takes place within the retina.

Color Perception

Subjective nature of color

Color is a perception and is subjective. There is no way one can define color in an absolute or quantitative means, even though it can be precisely characterized by the spectral power distribution (SPD) of the radiation. For monochromatic light of different wavelengths we see the spectral colors detailed in Table 1. In general, the SPD of light from an object is complex and to predict the exact color of such a spectrum is quite difficult. However, in the following we discuss some simple color schemes.

Additive color mixing

Newton's experiments demonstrated that using the three primary color light sources, blue, green, and red and by varying their relative intensities, one can generate most of the colors by additive color mixing. Additive color mixing is employed in stage lighting, and large screen color television projection systems. If we are concerned only with the hue and ignore saturation and brightness, the superposition of colored light beams of equal brightness on a white screen gives the following simple additive color scheme:

$$\text{Green} + \text{Red} \Rightarrow \text{Yellow}$$

$$\text{Blue} + \text{Red} \Rightarrow \text{Magenta}$$

$$\text{Blue} + \text{Green} \Rightarrow \text{Cyan}$$

$$\text{Blue} + \text{Green} + \text{Red} \Rightarrow \text{White}$$

Subtractive color mixing

When white light passes through, say, a red filter, the radiation corresponding to the red radiation is transmitted and all other parts of the visible spectrum are absorbed. If we insert more red filters the brightness of the light transmitted becomes less though the hue is likely to be more saturated. If you insert red and blue filters together in a white beam, you will see no light transmitted because the red filter blocks the entire spectrum except the radiation corresponding to red color which in turn is blocked by the blue filter.

Subtractive color scheme is also applicable when paints of various colors are mixed together. The subtractive primary colors are magenta, yellow, and cyan. The artists often call them red, yellow, and blue. By appropriately mixing the primary colored paints one can, in principle, generate most of the colors.

How many colors can we differentiate?

If we do a simple experiment beginning from the lowest wavelength of the visible spectrum, say 380 nm, and gradually move to the long wavelength end of the visible spectrum, one can differentiate about 150 spectral colors. For each spectral color there will be a number of colors that vary in brightness and saturation. Additionally, we can have additive mixtures of these spectral colors, which in turn have variations in brightness and saturation. It is estimated that under optimal conditions we can differentiate as many as 7 to10 million different colors and shades.

Temporal Response of the Human Visual System

Positive afterimages

The visual system responds to changes in the stimulation in time. However, the response of the visual system is both delayed and persists for some time after the stimulation is off. Because of the persistence of vision, one would see positive afterimages. The afterimages last about 1/20 s under normal conditions but are shorter at high light conditions. A rapid succession of images show results in a perception of continuous motion which is the basis for movies and TV.

Close your eyes for about five minutes so that any after images of the objects you have seen before are cleared. Open your eyes to see a bright colored object and then close your eyes again; you will see positive afterimages of the object in its original colors. Even though the positive afterimage at the initial stages will be of the same colors as the original, they gradually change color showing that different cones recover at different rates.

Negative afterimages

If you stare at a color picture intensely for about a minute and look at a white background, such as a sheet of white paper, you will see the complementary colors of the picture. The image of the object falls on a particular part of the retina and if we look at a green colored object, the cones that are more sensitive to the mid-wavelength region are activated

and are likely to be saturated. If you look at a bright sheet of white paper immediately, the cones that are sensitive to the middle wavelength region are inactivated since they need some time to replenish the active chemical whereas the cones that have maximum sensitivity for the low and high wavelengths are active, therefore you will see the object in complementary colors. Yellow will be seen after looking at blue image, cyan after looking at red image and vice versa.

Spatial Resolution of the Human Visual System

Lateral inhibition (color constancy)

By and large the perceived color of an object is independent of the nature of the illuminating white light source. The color of the object does not change markedly under different white light sources of illumination. For example, we see basically the same color of the object in the daytime in the light of the blue sky or under an incandescent light source even though the spectral power distributions of these sources are different and the intensity of light reaching the eye in both cases is also drastically different. The color constancy is a result of chromatic lateral inhibition. The human visual system has a remarkable way of comparing the signals from the object and manipulating them with the signals generated from the surroundings and picking up the correct color independent of the illuminating source. The color constancy and simultaneous lightness contrast are due to lateral inhibition. Lateral inhibition was exploited by a number of artists to provide certain visual effects that would not have been possible otherwise. For example, Georges Seurat in his famous painting 'La Poseuse en Profil' improved the light intensity contrast in different regions of his painting by employing edge enhancement effects resulting from lateral inhibition. Victor Vasarely's art work 'Arcturus' demonstrates the visual effects of lateral inhibition in different colors.

Spatial assimilation

The color of a region assimilates the color of the neighboring region, thus changing the perception of its color. Closely spaced colors, when viewed from a distance, tend to mix partitively and are not seen as distinct colors but as a uniform color due to additive color mixing. Georges Seurat in his painting 'Sunday Afternoon on the Island of La Grande Jatte', uses thousands of tiny dots of different colors to produce a variety of color sensations, a technique which is known as pointillism. The visual effects of spatial assimilation were clearly demonstrated by Richard Anuszkiewicz in his painting 'All Things Do Live in the Three'.

Color Perception Models

Thomas Young, in 1801, postulated that there are three types of sensors thus proposing trichromacy to explain the three attributes of color perception: hue, saturation, and brightness. Data on microspectrophotometry of excised retinas, reflection densitometry of normal eyes, and psychophysical studies of different observers confirm the existence of three types of cones. As shown in Figure 1, the three types of cones have different responsivities to light. The cones that have maximum responsivity in the short wavelength region are often refereed as S-cones, the cones that have maximum in the intermediate wavelength range as I-cones, and the cones that have maximum in the long wavelength region as L-cones.

The three response curves have considerable overlap. The overlap of the response curves and the complex data processing that takes place in the visual system enable us to differentiate as many as 150 spectral colors and about seven to ten million different colors and shades. An important consequence of the existence of only three types of color sensors is that different spectral power distributions can produce identical stimuli resulting in the perception of the same color. Such spectral power distributions are called metamers.

The neural impulses generated by the cones are processed through a complex cross-linking of bipolar cells, horizontal cells, amacrine cells, and communicated to ganglion cells leading to the optic chiasmas through the optic nerve. Information gathered by about 100 million photoreceptors, after considerable processing, is transmitted by about a million ganglion cells.

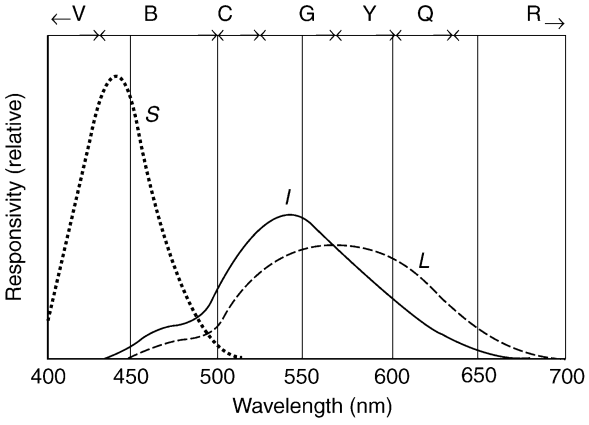


Figure 1 Relative spectral absorption of the three types of cones. *S*, *I*, and *L* respectively stand for the cones that have maximum responsivity for the short, intermediate, and long wavelength regions.

The zone theory of color vision assumes two separate zones where the signals are sequentially processed. In the first zone the signals are generated by the cones and in the second zone, the signals are processed to generate one achromatic and two opponent chromatic signals. Possibly other zones exist where the signals generated in the first and the second zones are further processed and communicated to the visual cortex where they are processed and interpreted based on the temporal and spatial response and comparisons with the information already available in the memory.

Measuring Color and Colorimetry

Since color is a human perception, how do we measure color? Colorimetry is based on the human visual system and how different observers perceive color. Though, the spectral power distribution of light is helpful for characterizing spectral hues, it is very difficult to infer the color from the spectral power distribution.

The three attributes of color perception are hue, saturation, and brightness. Hue is the color such as red, green, and blue. Saturation refers to the purity of the color. A color is said to be more saturated if the whiteness in the color is less and vice versa. Brightness is related to the intensity of light. For the sun at sunrise the hue is red, saturated, and low brightness. About half an hour after sunrise, the hue is yellow, less saturated, and higher brightness. Though a variety of color schemes are available in the literature, in the following we discuss the Munsell system and the C.I.E. diagram only.

Munsell Color System

Albert Munsell, a painter and art teacher, devised a color atlas in 1905 by arranging different colors in an ordered three-dimensional space. He used the three color attributes hue, chroma, and value corresponding to hue, saturation, and brightness. In this system there are 10 basic hues, each of which is divided into 10 equal gradations, resulting in 100 equally spaced hues. Each hue has a chart with a number of small chips arranged in rows and columns. Under daylight illumination, the chips in any column are supposed to have colors of equal chroma and the chips in any row are supposed to be of equal brightness. The brightness increases from the bottom of the chart to the top in steps that are perceptually equal. The saturation of the color increases from the inside edge to the outside edge of the chart in steps that are also perceptually equal. Each chip is identified by hue, value/chroma. In the Munsell scale, black is given a value of 0 and white 10, and the nine grays are uniformly placed in between. Though the Munsell system has gone through many revisions, typically it has a total of about 1450 chips. The Munsell color scheme, though subjective in nature, is widely used because of its simplicity and ease.

The Commission Internationale de l'Echlairage (C.I.E.) Diagram

Even though the visible light reaching the eye has a very complex SPD, color perception is mostly dependent on the signals generated by the three different types of cones. However, we do not know precisely the absorption and sensitivity of the cones and therefore are not in a position to calculate the signal strengths generated by the three different cones. Additionally, the response of the cones is specific to the individual. Therefore, the C.I.E. system is based on the average response of a large sample of normal observers. The C.I.E. system employs three color matching functions $\bar{x}(\lambda), \bar{y}(\lambda)$, and $\bar{z}(\lambda)$ (Figure 2), based on the psychological observations of a large number of standard observers to calculate the 'tristimulus values' X, Y, and Z defined as

$$X = c \int S(\lambda)\bar{x}(\lambda)\mathrm{d}\lambda \qquad [1]$$

$$Y = c \int S(\lambda)\bar{y}(\lambda)\mathrm{d}\lambda \qquad [2]$$

$$Z = c \int S(\lambda)\bar{z}(\lambda)\mathrm{d}\lambda \qquad [3]$$

Here, λ is the wavelength, $S(\lambda)$ the spectral power distribution, and c is the normalization constant. The integration is carried out over the visible region, normally 380 to 780 nm. The system is based on the assumption of additivity and linearity. The sum of the tristimulus values $(X + Y + Z)$ is normalized to 1. We choose $y(\lambda)$ so that the Y tristimulus value is proportional to the luminance which is a quantitative measure of the intensity of light leaving the surface. We further calculate x and y which are called chromaticity coordinates:

$$x = \frac{X}{(X + Y + Z)} \qquad [4]$$

$$y = \frac{Y}{(X + Y + Z)} \qquad [5]$$

The psychological color is then specified by the coordinates (Y, x, y). Relationship between colors is usually displayed by plotting the x and y values on

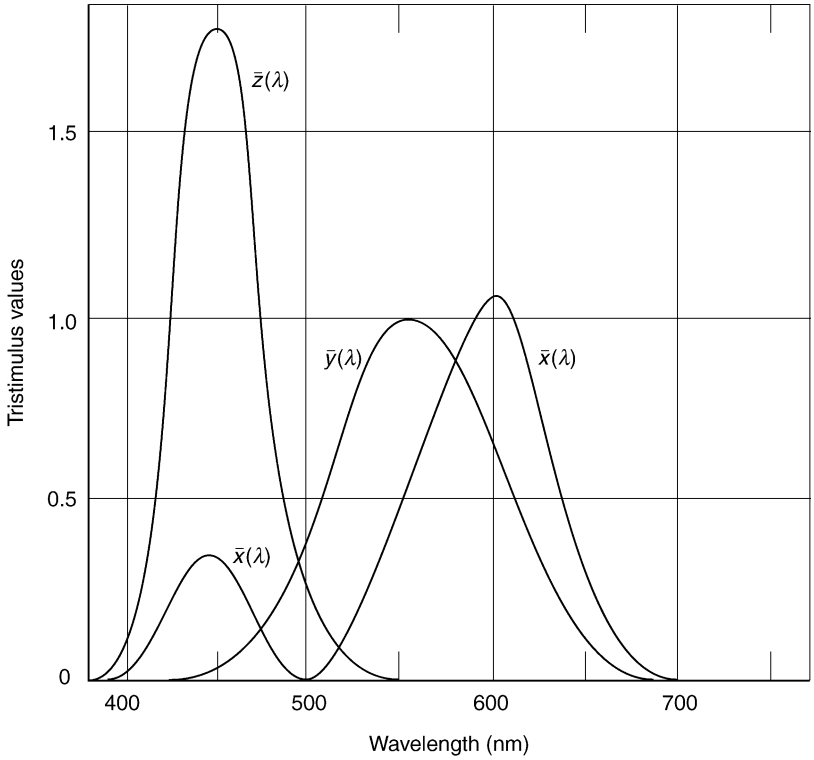


Figure 2 Color-matching functions $\bar{x}(\lambda)$,$\bar{y}(\lambda)$, and $\bar{z}(\lambda)$ of the C.I.E. 1931 standard observer.

a two-dimensional Cartesian Coordinate system with z as an implicit parameter ($z = 1 - x - y$) which is known as the C.I.E. chromaticity diagram (Figure 3). In this plot, the monochromatic hues are on the perimeter of a horseshoe shaped curve and is called spectrum locus. A straight line joining the ends of the spectrum is called the line of purples. Every chromaticity is represented by means of two coordinates (x, y) in the chromaticity diagram. The point corresponding to $x = 1/3$ and $y = 1/3$ (also $z = 1/3$) is the point of 'equal energy' and represents achromatic point. The complementary colors are on the opposite sides of the achromatic point. Complementary wavelength of a color is obtained by drawing a line from the color through the achromatic point to the perimeter on the other side.

Continuum (white) sources

A perfect white source is the one that has constant SPD over the entire visible region. Such a source will not distort the color of the objects they illuminate. A number of sources in general are considered as white sources; they include incandescent lamps, day skylight, a variety of fluorescent lamps, etc. It should be noted that the SPD of these sources is not constant and varies with the source. Typical SPDs of commonly used light sources are given in Figure 4.

Color temperature

The thermal radiation emitted by a black body has a characteristic SPD which depends solely on its temperature. The wavelength λ of the peak of this distribution is inversely proportional to its absolute temperature. As the absolute temperature of the blackbody increases, the peak shifts towards shorter wavelengths (blue region) and also the width of the distribution decreases. Since the peak of the black body radiation is dependent on the absolute temperature, the color of the blackbody can be approximately defined by its temperature.

The color temperature of a body is that temperature at which the SPD of the body best matches with that of the blackbody radiation at that temperature. The concept of color temperature is often used to characterize the sources. Low color temperature sources have a reddish tinge and high color temperature sources tend to be slightly bluish. For example, the color temperature of red looking star Antares is about 5000 K and that of the bluish white looking Sirius is about 11 000 K. One should note that the color temperature of a body is different from its actual temperature. The fluorescent lamps have a

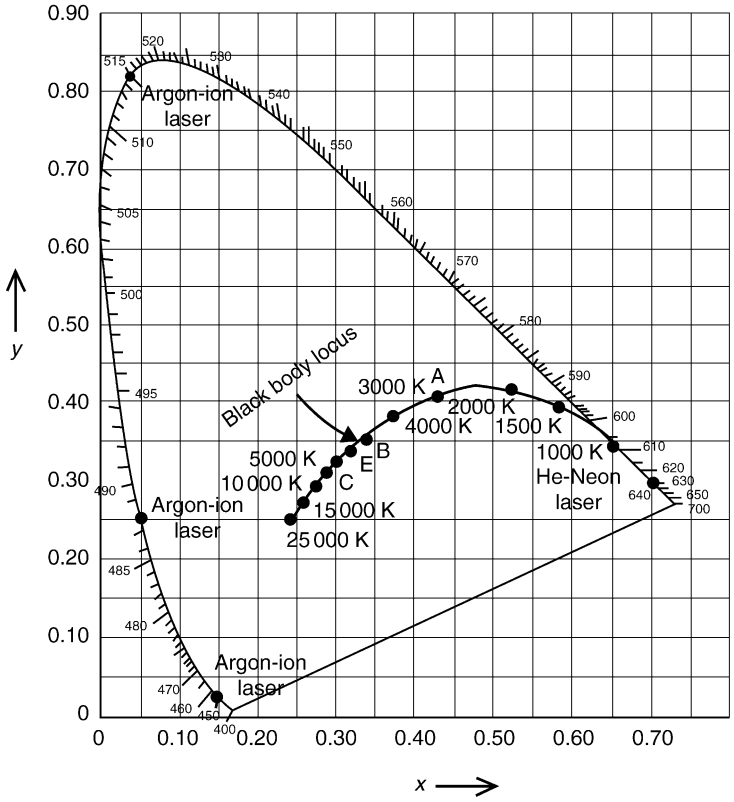


Figure 3 Chromaticity diagram of the C.I.E. 1931 standard observer. The chromaticities of the incandescent sources and their color temperatures are given inside the curve. A, B, and C are the C.I.E. standard illuminants. The chromaticities of the helium-neon laser (red, 633 nm) and the argon-ion laser (green 515 nm, blue 488 nm, and indigo, 458 nm) are also shown.

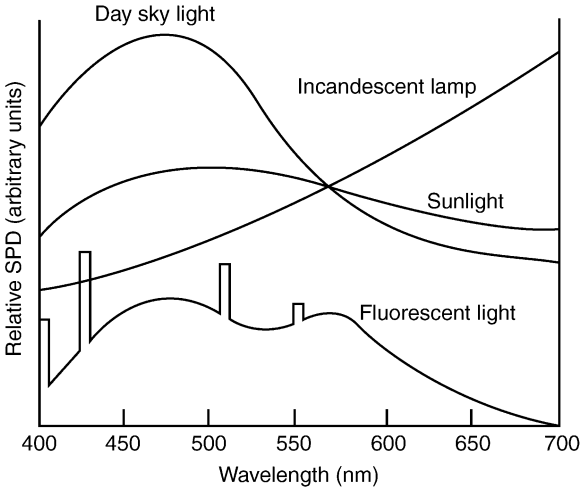


Figure 4 Relative SPDs of commonly used white light sources.

color temperature range of 5000–7500 K, incandescent lamps about 3000 K, direct sunlight ~4870 K and overcast skylight has a color temperature of ~7000 K.

The C.I.E. chromaticity diagram along with the color temperatures of the commonly employed sources, is shown in Figure 3.

Color in the Nature

As the solar radiation passes through the atmosphere, it is partly absorbed and partly scattered by the atmospheric constituents: air, water vapor, and dust particles. Absorption depends on the molecular properties of the air molecules, water vapor, and the dust particles, and is strongly wavelength dependent. Absorbed energy increases the internal energy of the molecule finally manifesting itself as thermal energy. When the sun is at the zenith, the solar radiation passes through one air mass loosing approximately 32% of its power. The SPD of the solar radiation above the atmosphere and at the Earth's surface at air mass 1 is shown in Figure 5. Scattering and absorption of the solar radiation by the atmosphere are responsible for a variety of interesting colorful views in the nature and they are briefly discussed in this section.

Atmospheric Colors

1. *Colors of the sky.* Scattering occurs over most of the visible part of the electromagnetic spectrum and is a function of the frequency of

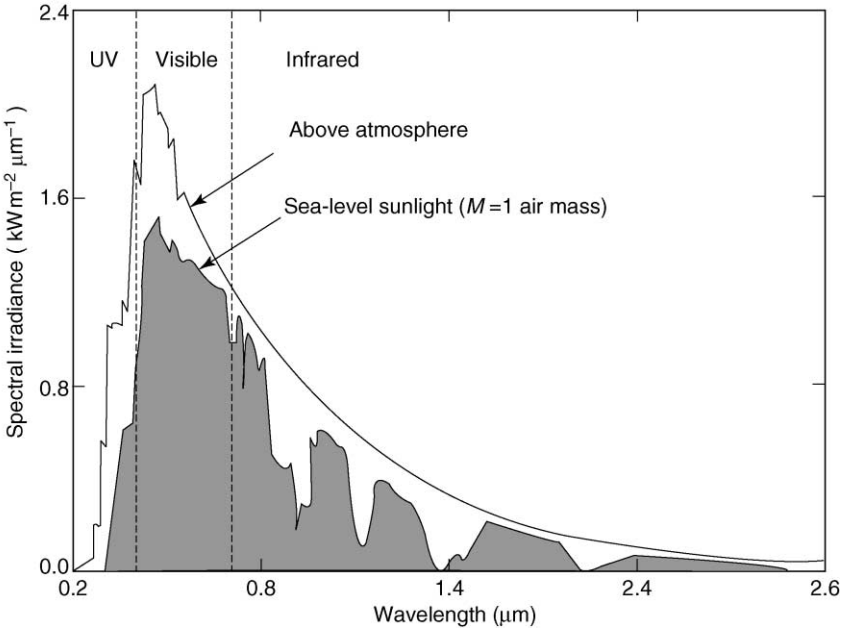


Figure 5 Spectral irradiance (kW m^{-2} μm^{-1}) of the solar radiation measured above the atmosphere and at the Earth's surface at sea level on a clear day and the sun at the zenith.

the radiation. According to Lord Raleigh's theory of scattering, the scattering probability varies inversely as a function of the fourth power of the wavelength of the radiation. If we consider the radiation corresponding to blue (400 nm) and the radiation corresponding to say red (700 nm), the photons corresponding to blue radiation are scattered by about 9.4 times that of photons corresponding to the red radiation. During noontime, when we look at the sky, not directly at the sun, the sky looks unsaturated blue because of the predominant scattering experienced by the high frequency radiation corresponding to blue color. However, it should be noted that the daytime skylight contains all the spectral colors of the solar radiation though the spectral power distribution has changed resulting in an increase in the intensity of the high frequency (blue) radiation. During noontime at the horizon, the sky looks close to white in color with a bluish tinge, because the solar radiation has passed through many air masses. As the thickness of the atmosphere increases, the intensity of the blue radiation decreases and ultimately as the radiation reaches the horizon, all the spectral colors are scattered about the same level resulting in white or bluish white sky.

2. *Polarization of the solar radiation by the atmosphere*. Light scattered by a molecule is polarized. The blue skylight is polarized with a maximum polarization of about 75% to 85% at 90° from the sun.
3. *Distant mountains seem bluish in color.* When an artist wishes to paint a distant mountain in perspective, he/she paints it blue; why? This is called airlight. The low wavelength part of the solar radiation is scattered more prominently, giving it a blue perspective when observed from a distance. However, it should be mentioned that the distant mountains may look any color depending on the scattering and absorption of the solar radiation sometimes resulting in 'Purple Mountain's majesty.'
4. *Sun looks red at the sunrise and sunset*. At sunset as the sun approaches the horizon, the sun changes its color from dazzling white to yellow, orange, bright red, and dull red and the order is reversed at sunrise. At the sunsets and sunrise, the solar radiation travels through many airmasses compared to approximately one airmass when the sun is at the zenith. Because of the large mass of air the radiation has to pass through, first the blue part of the spectrum is removed and the transmitted green to red spectral colors give the appearance of yellow. As this spectrum passes through more airmass, green is scattered most, leaving spectral yellow, orange, and red. As the radiation further passes through air, yellow and orange are also scattered, leaving only red.

5. *Twilight*. After sunset (dusk), or before the sunrise (dawn), the sky is illuminated by light which in turn illuminates the landscape. At sunsets and sunrise the solar radiation passes through a large mass of air. At a solar depression of about 0°, most of the blue is lost by scattering, resulting in a yellow arch and bright sky, and as the solar depression increases to about 10°, the arch will be reddish and the brightness of the sky reduced.
6. *Alpenglow*. Snow-covered mountains glow orange yellow from the low sun and after the sunset the same mountains look purple because of the reflection of the twilight.
7. *Twinkling of stars*. The stars twinkle. The local variations in the density of air are responsible for the twinkling of stars. As the density of air changes, so does the index of refraction and the starlight is continually refracted and dispersed. Twinkling is more prominent when the stars are on the horizon or on cold clear and windy nights.
8. *Mirages*. Mirages are refracted images produced because of the existence of refractive index gradients in the atmosphere. This results in the sight of shimmering water on a highway, inverted view of a car, etc.
9. *Aurora Borealis*. Aurorae are diffused slowly moving lights seen at high latitudes. They are often greenish yellow with different shapes. The lights in the northern hemisphere are called 'Aurora Borealis' and the lights in the southern hemisphere 'Aurora Australis.' They are caused primarily by the high-energy electrons deflected by the Earth's magnetic field towards the Earth's magnetic poles interacting with the atmospheric molecules resulting in molecular excitation and ionization. When the molecules de-excite, the characteristic electromagnetic radiation is commonly emitted. The commonly seen yellow-green light is due to the molecular emission of oxygen at 557 nm.
10. *Urban glows*. The scattered light from cities can be observed from space. The metropolitan Los Angeles can be observed from about 250 km in space. Extensively used sodium lamps with strong spectral yellow and blue-green mercury and fluorescent lights have unique color spectra of their own.

Colors due to Water and Water Droplets

1. *What is the color of pure water?* The transmission coefficient of pure water as a function of wavelength for different thicknesses of water column is shown in Figure 6. Our perception of the color of water changes with the thickness of water. For 1 mm thick water column, all the wavelengths are transmitted, resulting in the perception of white; 1 m thick water column looks unsaturated bluish green. As the thickness of water increases, the bluish green color will get more and more saturated. The color of water perceived by us depends on: (i) light reflected from the water surface; (ii) the light refracted at the water surface and diffused and scattered by the water molecules; and (iii) the refracted light reaches the bottom and gets reflected, which is further diffused and scattered and refracted at the water surface again. What we see is an additive sum of all these components.

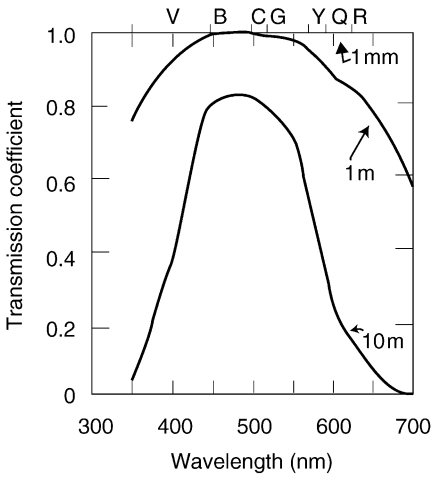


Figure 6 Transmission coefficient of water as a function of wavelength for different thicknesses of water (1 mm, 1 m, and 10 m).

2. *Refraction through water–air interface*. When light travels from an optically denser medium to an optically lighter medium, the light is deviated away from the normal. This is the reason why the apparent depth of a fish located in a pond is less than the true depth and a pencil struck in water seems to bend. If you are a diver when you look at the horizon from under water, the hemisphere seems to be compressed into a circle of diameter 97.2° instead of 180° which is known as 'Optical Manhole.'
3. *Glitter*. When you look at the reflection of the sun or the moon in wavy water, you notice a bright elongated reflected light known as glitter. It has a brilliant reddish orange color when the sun is above the horizon. Since the water is wavy, light reflected by a certain part of the wave at a certain time has the correct angle of reflection to reach your eye, which is known as glint. An ensemble of a large number of glints

coming from different waves and different parts of the waves results in glitter. Since you are looking at the reflected light, it will be polarized.

4. *Sky pools and landpools*. The light coming from a wavy water surface of a lake consists of oval shaped bright colored patterns that continually change in shape. This results from the reflection of skylight by the wavy water. If the waves are not turbulent the reflected light reaching your eye continually changes resulting in bright colored patterns. Landpools are caused by a nearby sloping landscape and the reflected light pattern from the landscape.
5. *Thin films*. Oil floats on water and spreads into thin films because of its low surface tension. When viewed in reflected natural white light, the light reflected from the top surface of the oil and the light reflected from the interfacial surface of oil and water interfere to produce colorful patterns. Brilliant colored patterns are commonly seen on the roads after rain because of gasoline on water. Foam ordinarily consists of bubbles ranging typically 10^{-4} mm to several mm. Even though each bubble produces colorful light due to the interference effects, the light coming out of the foam adds up to white light.
6. *Wet surfaces look darker*. A water spot on a concrete road looks darker. Since concrete is not porous a thin layer of water sticks to the surface. At the water surface part of the light incident is reflected, and a large part is transmitted. The transmitted light is diffusively scattered, part of which goes through total internal reflection and only a small part reaches the eye, making the surface look darker.

   When a fabric is wet it looks darker because of a slightly different mechanism. The surface of the fabric gets coated with a thin layer of water (refractive index = 1.33) which is smaller than the refractive index of the fabric. When the cloth is wet more light is transmitted, and less gets reflected than before. The transmitted light penetrates the fabric and gets scattered, resulting in a darker appearance.
7. *Rainbows*. Water with a refractive index of 1.33 is a dispersive medium. White light consisting of different wavelengths travel at different speeds in the medium and emerge at different directions. If the dispersion is adequate, different spectral colors can be seen by the naked eye. The dew covered lawn on a sunny day displays the spectral colors. If you stand close to the dew drop without blocking the sun's radiation falling on the drop, at some angle one should see the spectral colors. The primary rainbow is produced by the water droplets in the atmosphere after rain. The radius of the primary rainbow is about 42° with a width of about 2°. The color of the primary rainbow would be blue to red from the center of the rainbow. A rainbow is observed at the antisolar point. The brightness of the rainbow depends on the ambient sunlight, the size of the droplets, and whether the sun is low or high in the sky. The secondary rainbow, which is fainter, is often seen outside the primary rainbow at about 51°. The secondary rainbow will have red to blue from the center of the secondary rainbow. The secondary rainbow is approximately half as bright as the primary rainbow. The secondary rainbow is caused by two total internal reflections within the droplet whereas the primary rainbow is the result of one total internal reflection only. Because of the total internal reflection process, light is about 95% tangentially polarized in the primary rainbow.
8. *Heiligenschein*. Heiligenschein or the 'holy light' can be seen around the shadow of your head on a lawn on a sunny morning when dew drops are present. Usually, the dew drops are held by the tiny hairs of the grass at some distance from the blade of the grass. As light falls on the droplet, the spherical drop converges light at its focus and if the grass blade is located at that point, the light gets retroreflected approximately towards the sun resulting in holy light around your shadow.
9. *Coronae*. Colored concentric patchy rings seen around the moon and the sun are called coronae. They are primarily produced by the water droplets in the thin clouds. Even though the water droplets are randomly located, the diffracted light from the droplets has a defined angle and the superposition of the diffraction patterns from different water droplets results in colored ring structures. Coronae can be easily seen around the full moon viewed through thin clouds.
10. *The Glory*. On a foggy day with the sun at your back, one may see bright colored rings in the form of arches of about 5° to 10° at the solar point. The light in the glory is polarized and the ring structure is produced by reflection of the solar radiation by the fog.
11. *Clouds*. Clouds contain suspended water droplets or ice crystals. The size of the droplets ranges from less than a micrometer to about 100 micrometers (10^{-6} m). When the air is supersaturated with water vapor, the water condenses into water droplets, resulting in a definite shape to the cloud. When the sun's radiation falls on

the cloud, the light is scattered by the water droplets. Even though one would expect that the scattered light may have some color distribution, since we look at the scattered light from a very large number of droplets, the summation of all the colors results in white. Some clouds look dark or gray because, either they are in the shadow of other clouds or they are so thick that most of the light is absorbed. During sunset and sunrise, the clouds look bright orange because of the colored light illuminating them. Even though all the clouds are basically white, they may appear colored because of the selective absorption of certain colors in the atmosphere or due to the colored background radiation. Often, you see a silver lining at the edges of a cloud, because at the edges the density of water is low and when the sun illuminates it from the back, the radiation is scattered in the forward direction quite efficiently, resulting in a bright and shiny 'silver lining' to the cloud.

12. *Lightning*. Lightning is caused by an intense electrical discharge due to the high potential differences developed. On a cloudy and rainy day, as the water droplets grow in size, the falling drops get distorted to ellipsoidal shape and get polarized. The top smaller portion becoming negative and the lower larger part becoming positive. As the deformation increases, the drop ultimately breaks and the smaller drops being negatively charged and the larger ones positively charged. The wind currents easily lift the buoyant small drops to higher altitudes and the heavier positively charged droplets fall towards the Earth. The final result is that the upper region of the cloud is negatively charged whereas the lower region of the cloud is positively charged. The portion of the Earth under the cloud becomes negatively charged due to induction. The localized ions in the positively charged cloud produce a strong electric field which is further augmented by the induced negative charges on the Earth below. The strong electric field ionizes the air molecules setting up a current discharge to the earth. This discharge is called stepped leader. Following the stepped leader, the charges in the upper region of the cloud also discharge resulting in what is called return stroke. The sequence of leader and return strokes occur typically about ten times in a second. Since the clouds are charged and they are closer to each other, lightening is more frequent between the oppositely charged clouds than cloud and Earth. It is estimated that the lightning between the clouds is about five times more frequent compared to the lightening events between a cloud and the Earth. In this process, the nitrogen and oxygen molecules in air are excited and ionized. The flash is produced by the de-excitation of these molecular species. Since the temperatures involved are quite high (30 000 K), the flash has a strong unsaturated bluish hue. Some hydrogen lines are also observed in the flash. Hydrogen is produced as a result of the dissociation of water molecules during the discharge process.
13. *Ice and Halos*. Snow reflects most of the light falling with a reflectivity of about 95%. Snowflakes are crystalline and have hexagonal structure. In nature they occur in assorted sizes and shapes. As white light falls on them, light may be simply reflected from their surfaces making them shining white, which are often called glints. The light may also pass through the crystalline material get dispersed and emerge with different colors, which are known as sparkles.

    The halos appear as thin rings around the sun or the moon. They are produced because of the minimum deviation of light in snow crystals. The 22° halo is most common and is often seen a number of times in a year. The halos are circular and have a faint reddish outer edge and a bluish white diffused spot at the center. They often look white because of the superposition of different colors resulting in a more or less white color.

Color in Art and Holography

Sources of Color

There are two major color sources, dyes and paint pigments. Dyes are usually dissolved in a solvent. The dye molecules have selective absorption resulting in selective transmission. Paint pigments are powdered and suspended in oil or acrylic. The paint particles reflect part of the radiation falling on them, selectively absorb part of the radiation and transmit the balance. Lakes are made of translucent materials such as tiny grains of alumina soaked in a dye of appropriate color.

In the case of water colors and printer's inks, part of the light incident is reflected from the surface and the remaining transmitted which is selectively absorbed by the dye. The transmitted light reflected by the paper, passes through the dye again and emerges. Each transmission produces color by subtractive color scheme and the emerging light is mixed partitively by the observer.

In the case of paints, the color depends on the size and concentration of the pigment particles and the

medium in which they are suspended. Part of the light incident on the artwork gets reflected at the surface and the remaining selectively absorbed and transmitted which in turn is further reflected and transmitted by other pigment particles each time following a subtractive color scheme. The light reaching the canvas gets reflected and passes through the medium and emerges. Again, all the emerging beams are partitively mixed by the observer.

Holograms may be broadly categorized as transmission and reflection holograms. Both transmission and reflection (volume) holograms are prepared by producing the interference pattern of the reference and object laser beams. Transmission holograms can be viewed in the presence of the reference laser beam that was employed for its preparation. Spectacular three-dimensional colored images are reconstructed by the reflection holograms depending on the angle of incidence of the white light and the direction of observation. Embossed holograms are popular for advertising, security applications (credit cards), display, and banknotes. Some studios prepare three-dimensional holographic portraits that can be viewed in natural light. In the future, three-dimensional videos and even holographic three-dimensional movies may be a reality.

Structural Colors in Nature

Nanoscale Photonic Lattices

Newton observed that iridescence (change in color with the direction of observation) is caused by optical interference. The colors are produced by the interference or the light diffracting characteristics of the microstructures and not by the selective reflection or absorption of light, as in the case of pigments. Photonic band structures consist of periodic arrangement of dielectric materials even though they are transparent, have a bandgap resulting in inhibition of certain wavelengths. Light with frequencies in the gaps between the bands get reflected resulting in the specific color.

The structural colors of light can be seen in the animal world and minerals. For example, opal, a sedimentary gem does not have crystalline structure, but consists of tiny spheres of silica packed together. The diffraction of light through these structures results in spectacular colors determined by the size of the spheres and their periodicity. The spines of the sea mouse Aphrodite are covered with natural photonic structures that produce iridescence resulting in brilliant colors depending on the direction of the incident light and the direction of observation.

In the case of the butterfly, *Morpho rhetenor*, the metallic blue color is produced by the periodic structure of its wings. Its wings have layers of scales, each of about 200 mm long and 80 mm wide and about 1300 thin parallel structures per mm which form a diffraction grating.

The bird feathers viewed through an optical microscope show the structural components which are responsible for their coloration. The structural details of the peacock's eye pattern employing the scanning electron microscopy show photonic structures on peacock barbules. Barbules are the strap like branches on the peacock feathers. The four principal colors on the peacock feathers blue, green, yellow, and brown were identified wth the specific structures. The interference between the light reflected from the front surface to that reflected from the back was shown to be responsible for the brown color in the peacock's eye. The photonic structure of the green barbule has a periodic structure of melanin cylinders with a spacing of about 150 nm whereas the blue barbule has a periodic structure of 140 nm, and yellow barbules have a periodicity of 165 nm.

See also

Dispersion Management. **Holography, Techniques:** Color Holography. **Photon Picture of Light**. **Polarization:** Introduction. **Scattering:** Scattering Theory.

Further Reading

Blau SK (2004) Light as a feather: structural elements give peacock plumes their color. *Physics Today* 57: 18–21.

Color Science, http://www.physics.sfasu.edu/astro/color.html.

Colourware, Colour FAQ, http://www.colourware.co.uk/cpfaq.htm.

CVRL Color & Vision database, http://www.cvrl.org/.

Hecht E (2002) *Optics*, 4th edn. San Francisco, CA: Addison Wesley.

Lynch DK and Livingston W (2001) *Color and Light in Nature*, 2nd edn. Cambrige, UK: Cambridge University Press.

Minnaert MGJ (1993) *Light and Color in the Outdoors*. Heidelbergh: Springer-Verlag.

Nassau K (2000) *The Physics and Chemistry of Color*. New York: John Wiley and Sons.

Poynton CA, Color technology, http://www.Poynton.com/ColorFAQ.html.

Tayeb G, Gralak B and Enoch S (2003) Structural colors in nature and butterfly-wing modeling. *Optics and Photonics News* 14(2): 38–43.

Wyszecki G and Stiles WS (2000) *Color Science*. New York: John Wiley.

D

DETECTION

Contents

Fiber Sensors
Heterodyning
Image Post-Processing and Electronic Distribution
Smart Pixel Arrays

Fiber Sensors

M J Connelly, University of Limerick, Limerick, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The growth in the fiber optic communications and optoelectronics industries has led to the development and commercialization of many types of optical components. Fiber sensor devices and systems are a major user of such technologies. This has resulted in the development of commercial fiber sensors that can compete with conventional sensors. Fiber sensors are capable of measuring a wide variety of stimuli including: mechanical (displacement, velocity, acceleration, strain, pressure); temperature; electromagnetic (electric field, magnetic field, current); radiation (X-ray, nuclear); chemical composition; flow and turbulence in liquids; and biomedical. The main advantages of fiber sensors are that they are low cost, compact and lightweight, robust, passive, immune to electromagnetic interference, and highly sensitive.

Fiber sensors can be grouped into two basic classes: intrinsic and extrinsic. In an intrinsic fiber sensor the sensing is carried out by the optical fiber itself. In an extrinsic sensor the fiber is simply used to carry light to and from an external optical device where the sensing takes place. Point fiber sensors are localized to discrete regions; quasi-distributed sensors utilize point sensors at various locations along a fiber and distributed sensors are capable of sensing over the entire length of fiber.

A basic fiber sensor system, as shown in Figure 1, consists of a light source, fiber sensor, optical detector, and signal processing electronics. A measurand causes some change (intensity, phase, polarization, spectrum, etc.) in the light propagating through the sensor. This change is detected and processed to give an output signal proportional to the measurand. A large part of fiber sensor research concerns the development of fiber sensors sensitive to particular measurands and appropriate signal processing techniques. We describe some of the most important fiber sensors and also consider quasi-distributed and distributed sensing, including the use of multiplexing techniques.

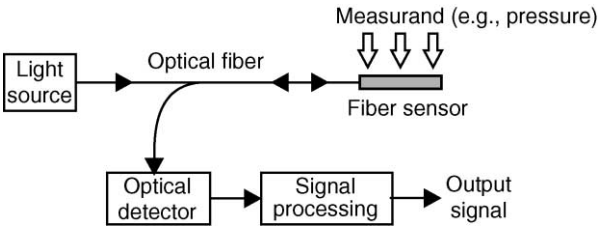


Figure 1 Basic fiber sensor system.

Intensity-Based Fiber Sensors

The intensity modulation (IM) of light is a simple method for optical sensing. There are several mechanisms that can produce a measurand-induced change in the optical intensity propagated by an optical fiber. Perhaps the simplest type of IM fiber sensor is the microbend sensor shown in Figure 2. The sensor consists of two grooved plates between which passes an optical fiber. The upper plate can move in response to pressure. When the fiber is bent sufficiently, light escapes into the fiber cladding and is lost. The greater the pressure on the plates the more loss occurs.

Coupling-based fiber sensors are useful for measurement of displacement or dynamic pressure. Transmission and reflective configurations are possible as shown in Figure 3. The transmission coupling-based sensor consists of two fibers with a small gap between them. The amount of light coupled to the second fiber depends on the fiber acceptance angles and the distance between the fibers. One of the fibers can move in response to vibration or pressure thereby changing the distance between the fibers and hence the coupling loss. The reflection-based sensor operates in a similar fashion, where light is reflected from a flexible diaphragm back into a collecting fiber. The reflected light intensity changes as the diaphragm is flexed. Once the coupling relationship between the input fiber, diaphragm and collecting fiber is known, intensity changes can be related to the applied displacement or pressure.

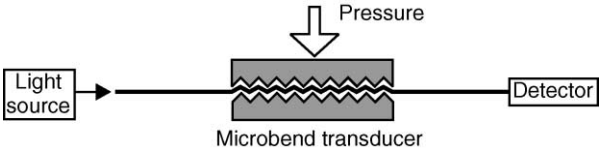


Figure 2 Microbend fiber sensor. The transducer moves in response to pressure and in doing so changes the bending radius of the fiber and thereby the fiber loss.

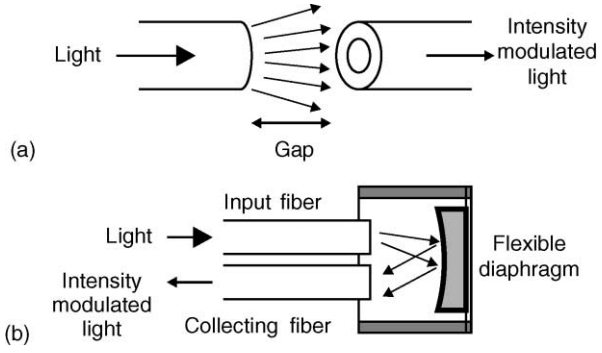


Figure 3 (a) Transmission and (b) reflective coupling-based fiber sensors.

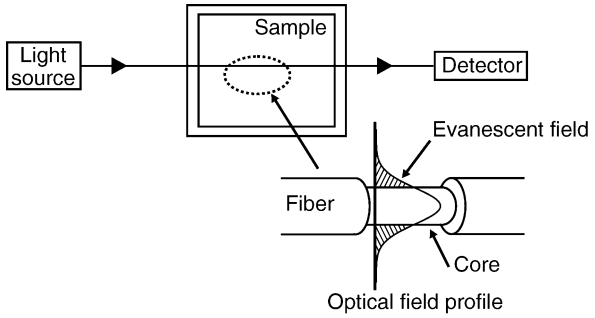


Figure 4 Evanescent wave fiber chemical sensor.

Evanescent wave fiber sensors exploit the fact that some of the energy in the guided mode of an optical fiber penetrates a short distance from the core into the cladding. The penetration of light energy into the cladding is called the evanescent wave. It is possible to design a sensor where energy is absorbed from the evanescent wave in the presence of certain chemicals as shown in Figure 4. This is achieved by stripping the cladding from a section of the fiber and using a light source having a wavelength that can be absorbed by the chemical that is to be detected. The resulting change in light intensity is a measure of the chemical concentration. Measurements can also be performed in a similar fashion by replacing the cladding with a material such as an organic dye whose optical properties can be changed by the chemical under investigation. Evanescent wave fiber sensors have found many applications in the biomedical field, such as blood component meters for detection of cholesterol and uric acid concentrations in blood.

A linear position sensor based on time division multiplexing is shown in Figure 5. It uses a square-wave modulated light source, optical delay loops and an encoded card. The delay loops separate the return signal from the encoded card by a time that is greater than the pulse duration. The encoded return signal can be decoded to determine the card position and velocity.

Interferometric Sensors

Interferometric fiber sensors operate on the principle of interference between two or more light beams to convert phase differences to intensity changes. Common configurations used include Michelson, Mach–Zehnder, and Sagnac interferometers, polarimetric systems, grating and etalon-based interferometers and ring resonators. Interferometric fiber sensors have extremely high sensitivity and are able to resolve path differences of the order of 10^{-6} of the light

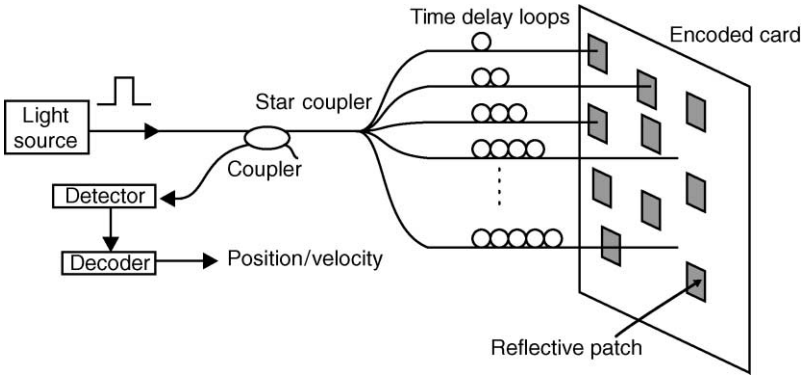


Figure 5 Linear position sensor using time division multiplexing.

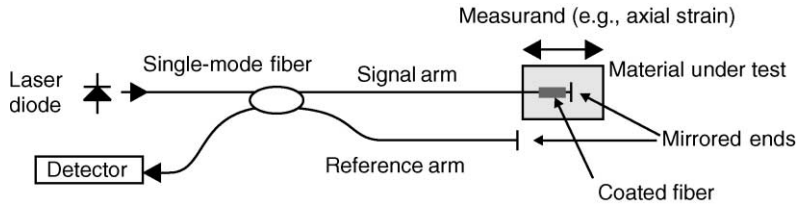


Figure 6 Michelson interferometric strain sensor. Part of the signal arm is embedded in a material, such as concrete (in civil engineering structures) or carbon composite (used in the aerospace industry).

source wavelength. To realize high sensitivity, spectrally pure light sources such as semiconductor lasers must be used. In addition, single mode fibers and components are used to maintain the spatial coherence of the light beams. Factors that affect the performance of interferometric sensors include the optical source phase noise and the polarization states of the interfering light beams.

In two-beam interferometric sensors, such as the Michelson interferometric strain sensor shown in Figure 6, one of the light beams travels through a fiber where its phase can be modulated by the measurand. The two reflected beams recombine on a detector, the output of which is of the form $1 + V \cos \phi$ as shown in Figure 7. ϕ is the relative phase shift between the return beams and V the interferometer visibility. Compensating techniques must be used to ensure that the interferometer sensitivity is maximized. Active techniques involve the insertion of an active device, such as a fiber stretcher, in one of the interferometer arms to control the mean phase difference between the beams. Passive schemes have the advantage in that no active components are required, but usually involve complex signal processing schemes such as active homodyne or synthetic heterodyne demodulation. The sensitivity of the fiber to the measurand can be further improved through the use of specialized coatings.

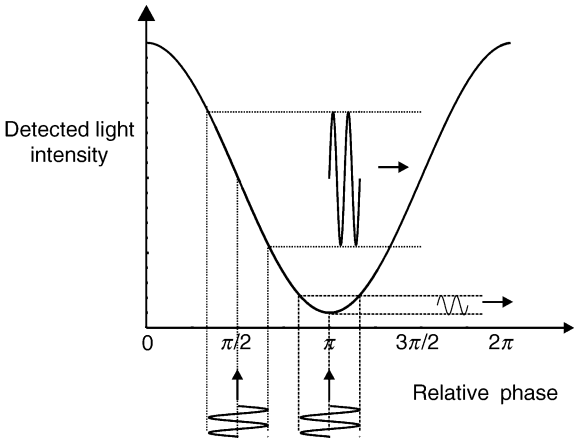


Figure 7 Two-beam interferometer transfer function. The sensitivity is maximized when the mean phase difference between the detected light beams is an odd integer multiple of $\pi/2$. When this is the case the interferometer is said to be operating in quadrature.

Compact interferometric sensors can be constructed using intrinsic or extrinsic Fabry–Perot type configurations, the most common of which are shown in Figure 8. If the Fabry–Perot cavity reflectivities are $\ll 1$, then it can be considered to be a two-beam interferometer.

Polarimetric fiber sensors relate changes induced in the polarization state of light to the measurand.

These sensors usually use high birefringence (Hi-Bi) fiber. The phase difference between the two orthogonally polarized modes of Hi-Bi fiber is given by

$$\phi = \Delta\beta l \qquad [1]$$

where $\Delta\beta$ is the fiber (linear) birefringence and l the fiber length. Both $\Delta\beta$ and l can be changed by the measurand. A polarimetric strain sensor using Hi-Bi fiber embedded in carbon fiber composite, is shown in Figure 9. Light from a linearly polarized He–Ne laser is launched into the Hi-Bi fiber, through a half-waveplate and lens. The waveplate is used to rotate the light plane of polarization so that it is at 45° to the principal axes of the fiber. This ensures that half of the input light power is coupled to each of the fast and slow modes of the Hi-Bi fiber. The output light from the Hi-Bi fiber is passed through a polarization beamsplitter. The beamsplitter separates the light into two orthogonally polarized beams, which are then detected. The output from each of the detectors is given by $V_1 = I_0 \sin^2\phi$ and $V_2 = I_0 \cos^2\phi$, respectively, where I_0 is the input light intensity. The state of polarization (SOP) of the output light from the Hi-Bi fiber is given by

$$\text{SOP} = \frac{V_1 - V_2}{V_1 + V_2} = \cos^2\phi + \sin^2\phi \qquad [2]$$

It is a simple matter to calculate the SOP, from which ϕ can be determined. The form of eqn [2] is very similar to the two-beam interferometer transfer function, shown in Figure 7.

A polarimetric electric current sensor based on the Faraday effect is shown in Figure 10. The Faraday effect provides a rotation of the light's polarization state when a magnetic field is parallel to the optical path in glass. If an optical fiber is closed around a current-carrying conductor, the Faraday rotation is directly proportional to the current. By detecting the polarization rotation of light in the fiber, the current can be measured.

Linearly polarized light is equivalent to a combination of equal intensity right and left circularly polarized components. The linear state of polarization of the polarized laser light rotates in the presence of a magnetic field because the field produces

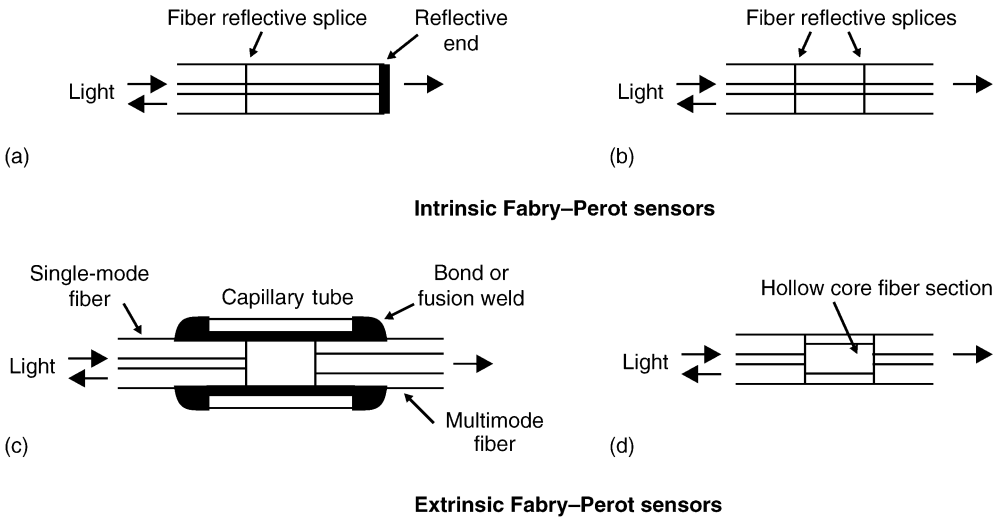


Figure 8 Interferometric fiber Fabry–Perot configurations.

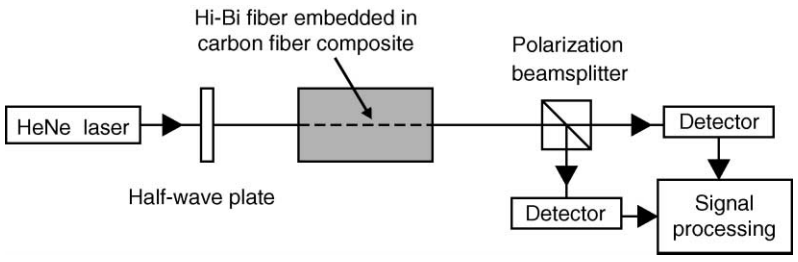


Figure 9 Polarimetric strain sensor system.

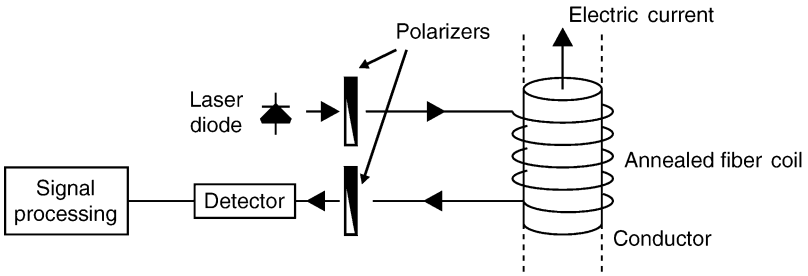


Figure 10 Polarimetric electric current sensor system.

circular birefringence in the fiber. This means that right and left circularly polarized light will travel at different speeds and accumulate a relative phase difference given by

$$\phi = VBl \qquad [3]$$

where V is the Verdet constant (rad Tesla^{-1} m^{-1}) of the glass, B the magnetic field flux density (Tesla), and l the length of the fiber exposed to the magnetic field. In the polarimetric current sensor, the polarization rotation is converted to an intensity change by the output polarizer. The output to the detector is proportional to $1 + \sin(2\phi)$, from which ϕ and the current can be determined.

The linear birefringence of conventional silica fiber is much greater than its circular birefringence. To make a practical current sensor, the linear birefringence must be removed from the fiber and this can be achieved by annealing the fiber. Annealing involves raising the temperature of the fiber, during manufacture, to a temperature above the strain point for a short period of time and slowly cooling back to room temperature. This reduces stresses in the glass, which are the principal cause of linear birefringence, and also cause physical and chemical changes to the glass. Waveguide-induced birefringence cannot be removed by annealing, but can be significantly reduced by twisting the fiber.

White light – or more accurately low-coherence – interferometry utilizes broadband sources such as LEDs and multimode lasers in interferometric measurements. Optical path differences (OPDs) are observed through changes in the interferometric fringe pattern. A processing interferometer is required in addition to the sensing interferometer to extract the fringe information.

The processing of white light interferometry signals relies on two principal techniques. The first technique involves scanning the OPD of the processing interferometer to determine regions of optical path balance. The second technique involves determination of the optical spectrum using an optical spectrum analyzer. The resulting fringe pattern fringe spacing is then related to the OPD of the sensing interferometer.

An example of the optical path scanning technique is shown in Figure 11a. The sensing interferometer is designed such that its OPD is much greater than the coherence length of the light source, so at its output no interference fringes are observed. The output of the sensing interferometer is fed to the processing interferometer. The OPD of the processing interferometer is scanned using a piezoelectric fiber stretcher driven by a suitable scanning voltage. As the processing interferometer is scanned, interference fringes are observed at two distinct points as shown in Figure 11b. The first set of fringes occurs when the OPD of the processing interferometer is within the coherence length of the optical source (close to zero). The second set of fringes occurs when the OPDs in the two interferometers are equal. The OPD in the sensing interferometer can be determined by measuring the OPD in the processing interferometer between the two positions of maximum visibility in the output signal from the detector. This, in turn, can be related to OPD changes due to the action of the measurand, in this case strain.

Sagnac interferometric sensors can be used to create highly sensitive gyroscopes that can be used to sense angular velocity (e.g., in aircraft navigation systems). It is based on the principle that the application of force (e.g., centrifugal force) will alter the wavelength of light as it travels around a coil of optical fiber. A basic open-loop fiber gyroscope is shown in Figure 12. The broadband source (e.g., superluminescent diode) is split into two counter propagating light beams traveling in the clockwise and anticlockwise directions. The polarizer is used to ensure the reciprocity of the counter propagating waves. The inherent nonlinear response of the gyroscope can be overcome by using a phase modulator and signal processing techniques. In the ideal case the detector output is proportional to $1 + \cos\phi_s$. The Sagnac phase shift ϕ_s between the

two returning beams is given by

$$\phi_s = \frac{8\pi AN\Omega}{c\lambda_0} \quad [4]$$

A is the area enclosed by a single loop of the fiber, N the number of turns, Ω the component of the angular velocity perpendicular to the plane of the loop, λ_0 the free-space wavelength of the optical source, and c the speed of light in a vacuum. Sensitivities greater than 10^{-8} rad/s and dynamic ranges in excess of 40 dB have been achieved with open-loop fiber gyroscopes. More advanced gyroscopes can greatly improve on this performance.

Fiber Grating Sensors

Many types of fiber gratings can be used in sensing applications including Bragg, long-period, and

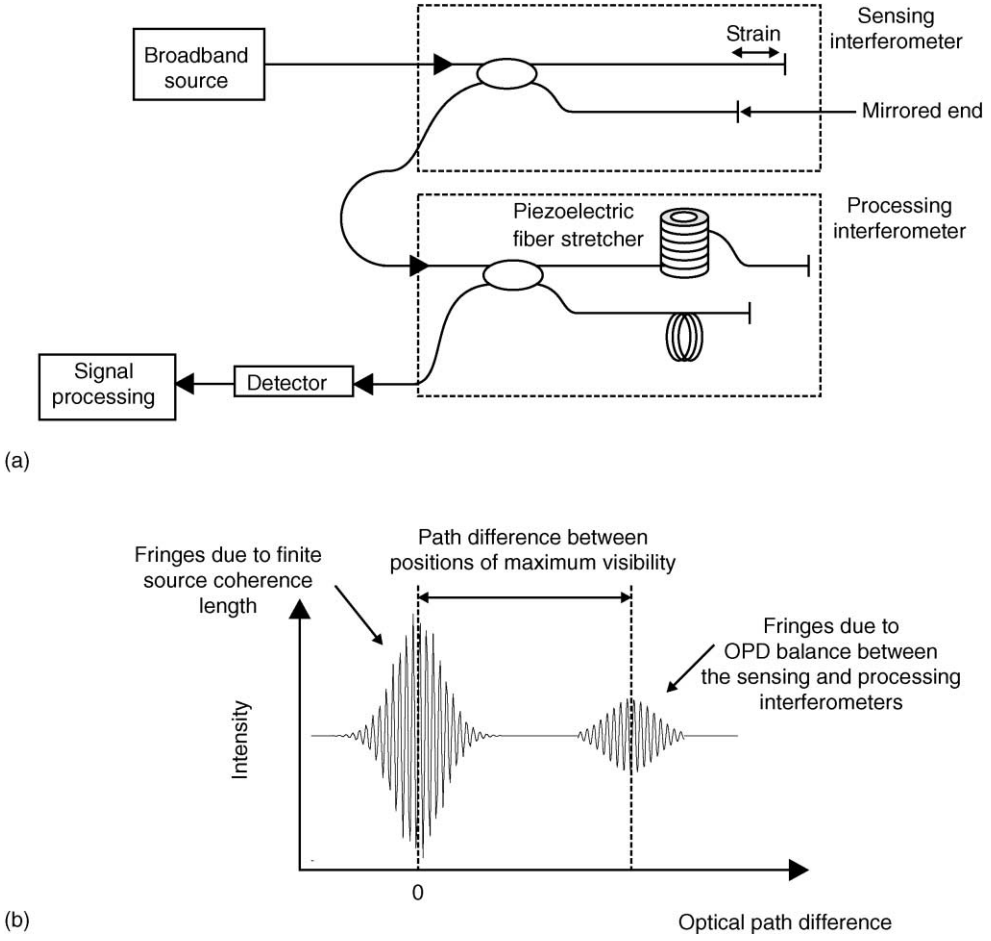


Figure 11 Low-coherence interferometric sensor. (a) Schematic diagram. (b) Output of the scanned processing interferometer.

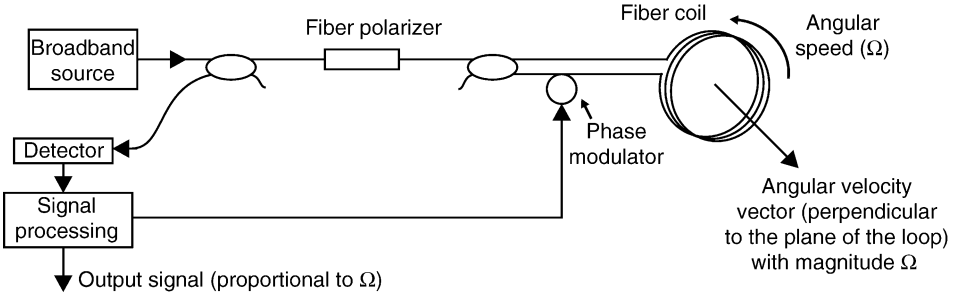


Figure 12 Basic open-loop fiber gyroscope.

chirped gratings. Because fiber gratings are small and have a narrow wavelength response, they can be used for both point and quasi-distributed sensing. They can be embedded in composite materials for smart structure monitoring and also in civil engineering structures such as bridges.

Fiber Bragg gratings (FBGs) are the most popular type of grating sensor. A typical FBG consists of a short section (typically a few mm) of single-mode fiber with a periodic modulation (typically 100s of nm) of the core refractive index. The index modulation causes light in the forward propagating core mode to be coupled into the backward core mode. This causes the FBG to act as a highly wavelength-selective rejection filter. The wavelength of peak reflectivity is the Bragg wavelength $\lambda_B = 2n_{eff}\Lambda$, where n_{eff} is the effective refractive index of the guided mode in the fiber and Λ the index modulation period. Both n_{eff} and Λ can be changed by an external measurand, resulting in a shift in λ_B.

The basic principle of FBG sensors is the measurement of an induced shift in the wavelength of an optical source due to a measurand, such as strain or temperature. A basic reflective FBG sensor system is shown in Figure 13. A broadband light source is used to interrogate the grating, from which a narrowband slice is reflected. The peak wavelength of the reflected spectrum can be compared to λ_B, from which strain or temperature can be inferred. The shift in the Bragg wavelength $\Delta\lambda_B$ with applied microstrain $\Delta(\mu\varepsilon)$ and change in temperature ΔT, for silica fiber is given approximately by

$$\frac{\Delta\lambda_B}{\lambda_B} \approx 0.78 \times 10^{-6}\Delta(\mu\varepsilon) + 6.67 \times 10^{-6}\Delta T \qquad [5]$$

A wavelength resolution of ~1 pm is required (at 1.3 μm) to resolve a temperature change of 0.1 °C or a strain change of 1 $\mu\varepsilon$. The response of the grating to strain can be improved through the use of specialized coatings. Polyimide coatings are commonly used as an efficient strain transducer for gratings embedded in composite materials. Because thermal effects in materials are usually very slow, it is relatively easy to measure dynamic strain (>1 Hz). However, in structural monitoring it can be necessary to distinguish between wavelength shifts due to static strain and those due to temperature. One technique is to use two collocated gratings whose response to strain and temperature is significantly different. In addition to strain and temperature measurement, grating sensors have also been used to measure flow, vibration, electromagnetic fields and chemical effects.

An example of quasi-distributed strain sensing, using a wavelength division multiplexed array of FBGs, is shown in Figure 14. Each FBG in the array has a unique Bragg wavelength. The return light from the FBG array is passed through a tunable narrowband Fabry–Perot filter. As the filter is tuned, the wavelengths returned by the individual FBGs can be analyzed and the strain present at each grating determined.

FBGs can be used as narrowband reflectors for creating fiber laser sensors, capable of measuring temperature, static strain, and very high-resolution dynamic strain. The basic form of an FBG laser sensor system shown in Figure 15 consists of a doped fiber section between two identical FBGs. The doped fiber is optically pumped to provide gain and thereby enable lasing to occur. Single-mode or multi-mode lasing is possible depending on the cavity length. In single-mode operation the FBG laser linewidth can be much smaller than the linewidth of diode lasers. This means that FBG laser sensors have greater sensitivities compared to passive FBG sensors.

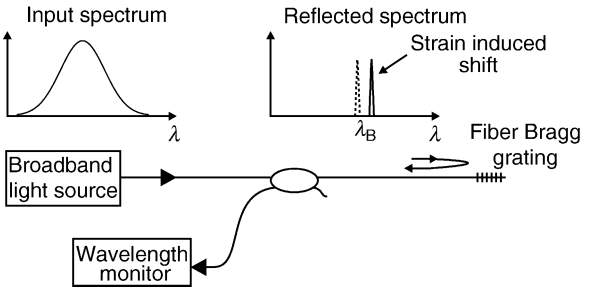


Figure 13 Basic reflective FBG sensor system.

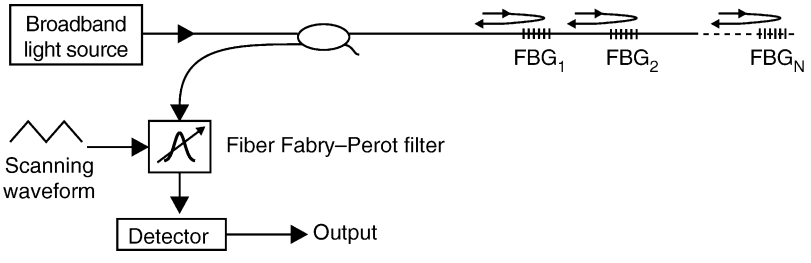


Figure 14 Quasi-distributed strain sensing using a wavelength division multiplexed array of FBGs.

When the cavity is subject to weak dynamic strain, the laser output is frequency modulated. This frequency modulation can be detected by using interferometric techniques. The main advantage of the FBG laser sensor over direct interferometry is that it is possible to obtain comparable strain sensitivities using a much shorter length of fiber.

Long-period fiber gratings (LPFGs) are attracting much interest for use in sensing applications. They are more sensitive to measurands than FBGs and easier to manufacture. A typical LPFG has a length of tens of mm with a grating period of 100s of μm. Its operation is different to an FBG in that coupling occurs between the forward propagating core mode and co-propagating cladding modes. The high attenuation of the cladding modes results in a series of minima occurring in the transmission spectrum of the fiber. This means that the spectral response is strongly influenced by the optical properties of the cladding and surrounding medium. This can be exploited for chemical sensing as shown in Figure 16, where a broadband source is used to interrogate an LPFG. The wavelength shifts of the output spectrum minima can be used to determine the concentration of particular chemicals in the substance surrounding the grating. The longest wavelength attenuation bands are the most sensitive to the refractive index of the substance surrounding the grating. This is because higher order cladding modes extend a greater distance into the external medium. LPFGs can also be used as strain, temperature, refractive index, bend, and load sensors.

Fiber Laser Doppler Velocimeter

Laser Doppler velocimetry (LDV) is a technique used for measuring velocity, especially of fluid flows. A fiber LDV system and associated scattering geometry is shown in Figure 17. In LDV two coherent laser beams intersect in a small measurement volume where they can interfere. The light reflected by a seed particle passing through the measurement volume is modulated at a frequency proportional to the spatial frequency (Doppler difference frequency Δf) of the interference fringes and the component of its velocity

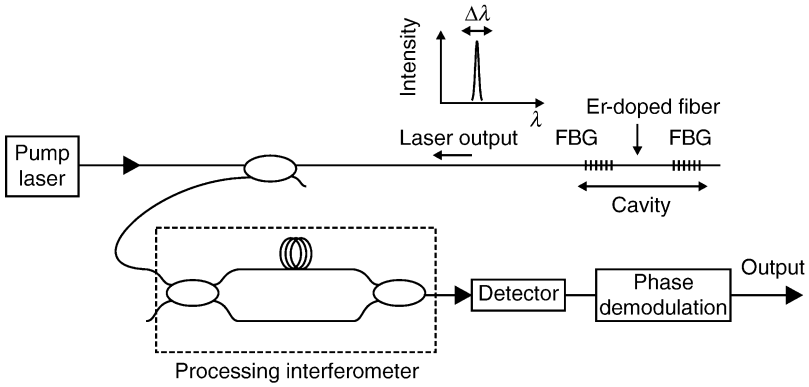


Figure 15 FBG laser sensor with interferometric detection.

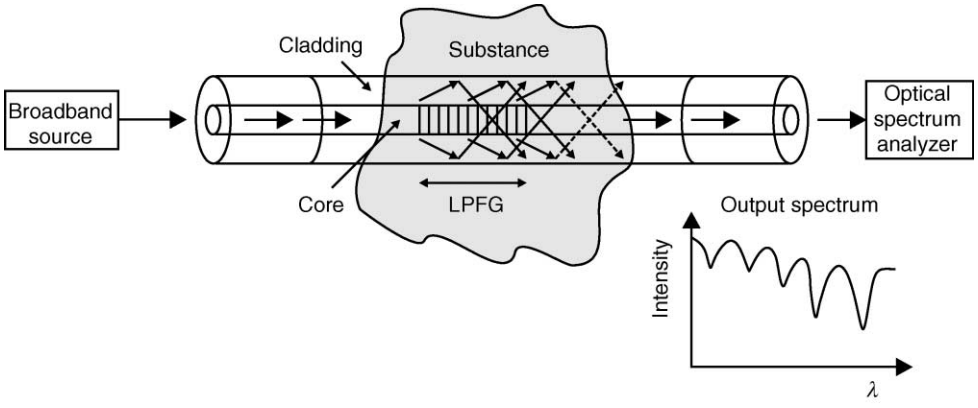


Figure 16 Detail of LPFG sensor.

normal to the interference fringes. Δf is given by

$$\Delta f = \frac{2nV\cos\beta}{\lambda_0}\sin(\theta/2) \qquad [6]$$

where n is the fluid refractive index, V the particle velocity and λ_0 the laser free-space wavelength. The output of the detector is processed to extract Δf and therefore $V\cos\beta$. Δf is independent of the scattered light direction so collection of the scattered light using a lens increases the system sensitivity. The form of eqn [6] indicates that the direction of flow cannot be ascertained. The simplest technique to resolve this ambiguity is to apply a frequency shift Δf_s to one of the input beams. This can be achieved by the use of a piezoelectric frequency shifter. The frequency shift causes a phase shift to appear between the two beams. The phase shift increases linearly with time. This results in a fringe pattern of spacing s, which moves with constant velocity $V_f = s\Delta f_s$. In this case the measured Δf will be less than or greater than Δf_s, depending on whether the particle is moving with or against the fringe motion. There will then be an unambiguous detectable range of velocities from zero to V_f.

Luminescence-Based Fiber Sensors

Luminescence-based fiber sensors are usually based on fluorescence or amplified spontaneous emission occurring in rare earth materials. They can be used in many applications such as chemical, humidity, and temperature sensing. It is possible to connect a fiber to luminescent material or to introduce luminescent dopants into the fiber. An example of the latter, used to detect chemical concentration is shown in Figure 18. A laser pulse causes the doped section of the fiber to luminesce at a longer wavelength than the

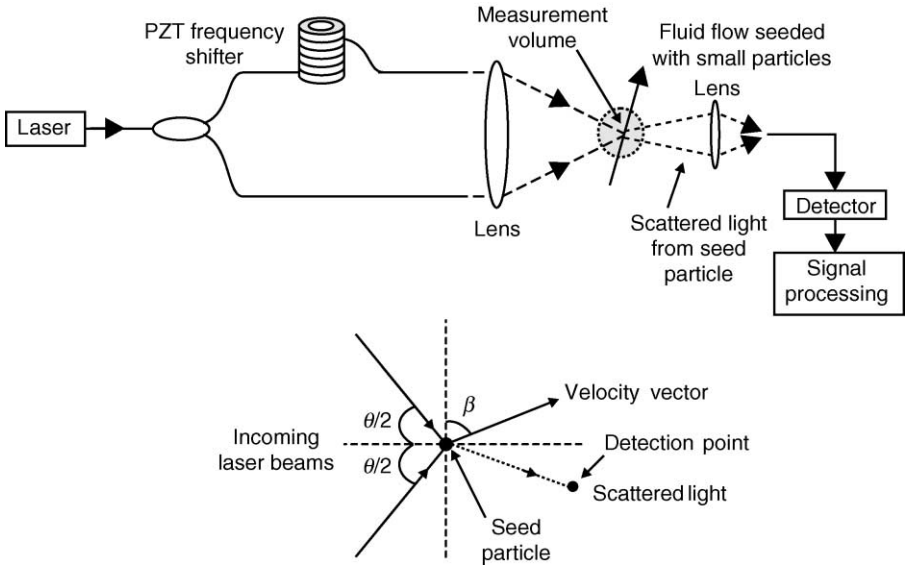


Figure 17 Fiber Doppler velocimeter and scattering geometry.

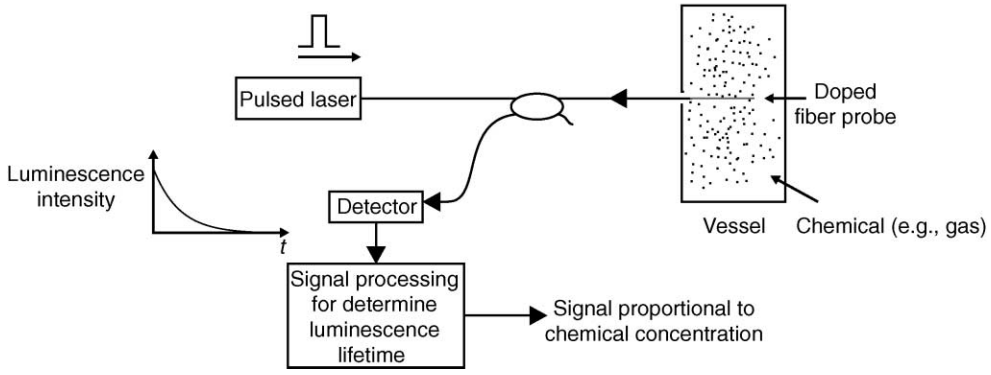


Figure 18 Chemical sensor based on fluorescence in doped optical fiber.

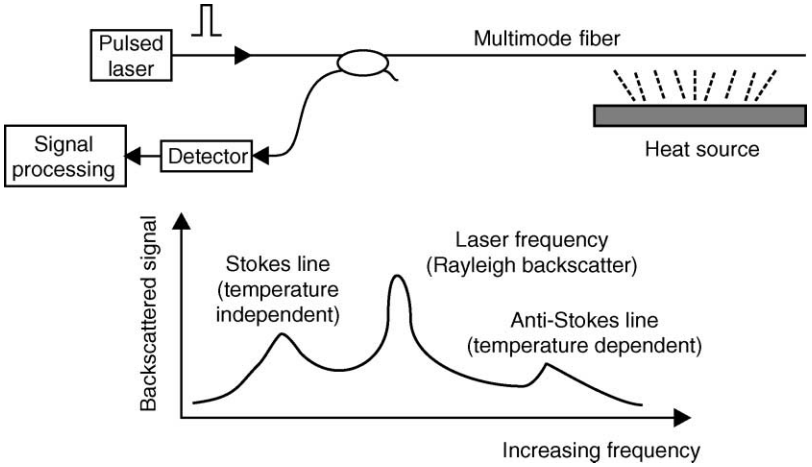


Figure 19 Anti-Stokes Raman thermometry system.

laser. The luminescence intensity $I(t)$ at the detector has an exponential decay profile given by

$$I(t) = I_0 \exp\left(-(k_1 + k_2 C)t\right) \qquad [7]$$

where I_0 is the initial luminescence intensity, t time, k_1, k_2 constants, and C the chemical concentration. The luminescence time constant $1/(k_1 + k_2 C)$ can be determined by comparing the luminescence intensity at various times after excitation by the laser pulse, from which C can be determined. The use of time division multiplexing allows quasi-distributed measurement of chemical concentration. The use of plastic optical fiber for luminescence-based sensors is attracting much interest.

Distributed Fiber Sensing

We have seen that both point and quasi-distributed sensing are possible using fiber sensors. Distributed sensing can be achieved through the use of linear or nonlinear backscattering or forward scattering techniques. In linear backscattering systems, light backscattered from a pulse propagating in an optical fiber is time resolved and analyzed to obtain the spatial distribution of the measurand field, e.g., polarization optical time domain reflectometry analyzes the polarization state of backscattered light to determine the spatial distribution of electromagnetic fields. Nonlinear backscattering schemes use effects such as Raman or Brillouin scattering. An important example of the former is the anti-Stokes Raman thermometry system shown in Figure 19. A light pulse is transmitted down the sensing fiber. Spontaneous Raman scattering causes Stokes and anti-Stokes photons to be generated along the fiber. Some of these photons travel back along the fiber to a fast detector. The intensity of the Stokes line is temperature independent. The anti-Stokes line intensity is a function of temperature. The ratio of the two intensities provides a very accurate measurement of temperature. The measurement location is determined by timing of the laser pulse.

See also

Environmental Measurements: Laser Detection of Atmospheric Gases. **Fiber Gratings**. **Interferometry:** Overview; Phase Measurement Interferometry; White Light Interferometry. **Optical Materials:** Smart Optical Materials.

Further Reading

Culshaw B and Dakin JP (1988, 1989, 1997) *Optical Fiber Sensors*, vols. I, II, III. Norwood, MA: Artech House.

Grattan KTV and Meggitt BT (1994, 1997, 1999, 1999) *Optical Fiber Sensor Technology*, vols. I, II, III and IV. Boston, CA: Kluwer Academic.

Grattan KTV and Meggitt BT (2000) *Optical Fiber Sensor Technology*. Boston, MA: Kluwer Academic.

Kersey AD, Davis MA, Patrick HJ, *et al.* (1997) Fiber grating sensors. *Journal of Lightwave Technology* 15: 1442–1463.

Krohn DA (1999) *Fiber Optic Sensors: Fundamentals and Applications*, 3rd edn. Research Triangle Park, NC: Instrumentation Society of America.

Lopez-Higuera JM (2001) *Handbook of Optical Fiber Sensing Technology*. New York: Wiley.

Othonos A and Kalli K (1999) *Fiber Bragg Gratings: Fundamentals and Applications in Telecommunications and Sensing*. Norwood, MA: Artech House.

Udd E (1991) *Fiber Optic Sensors: An Introduction for Engineers and Scientists*. New York: Wiley.

Udd E (1995) *Fiber Optic Smart Structures*. New York: Wiley.

Yu FTS and Yin S (2002) *Fiber Optic Sensors*. New York: Marcel Dekker.

Heterodyning

T-C Poon, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Heterodyning, also known as frequency mixing, is a frequency translation process. Heterodyning has its root in radio engineering. The principle of heterodyning was discovered in the late 1910s by radio engineers experimenting with radio vacuum tubes. Russian cellist and electronic engineer Leon Theremin invented the so-called Thereminvox, which was one of the earliest electronic musical instruments that generates an audio signal by combining two different radio signals. American electrical engineer Edwin Armstrong invented the so-called super-heterodyne receiver. The receiver shifts the spectrum of the modulated signal, that is, the frequency contents of the modulated signal, so as to demodulate the audio information from it. Commercial amplitude modulation (AM) broadcast receivers nowadays are super-heterodyne type. After illustrating the basic principle of heterodyning by some simple mathematics, we will discuss some examples on how the principle of heterodyning is used in radio and in optics.

For a simple case, when signals of two different frequencies are heterodyned or mixed, the resulting signal produces two new frequencies, the sum and difference of the two original frequencies. Figure 1 illustrates how signals of two frequencies are mixed or heterodyned to produce two new frequencies, $\omega_1 + \omega_2$ and $\omega_1 - \omega_2$, by simply multiplying the two signals $\cos(\omega_1 t + \theta_1)$ and $\cos(\omega_2 t)$, where ω_1 and ω_2 are radian frequencies of the two signals and θ_1 is the phase angle between the two signals. Note that when the frequencies of the two signals to be heterodyned are the same, i.e., $\omega_1 = \omega_2$, the phase information of $\cos(\omega_1 t + \theta_1)$ can be extracted to get $\cos\theta_1$ if we use an electronic lowpass filter (LPF) to filter out the term $\cos(2\omega_1 t + \theta_1)$. This is shown in Figure 2. Heterodyning is often referred to as homodyning for the mixing of two signals of the same frequency.

For a general case of heterodyning, we can have a signal represented by $s(t)$ with its spectrum $S(\omega)$, which is given by the Fourier transform of $s(t)$, and when it is multiplied by $\cos(\omega_2 t)$, the resulting spectrum is frequency-shifted to new locations in the frequency domain as:

$$\mathcal{F}\{s(t)\cos(\omega_2 t)\} = \frac{1}{2}S(\omega - \omega_2) + \frac{1}{2}S(\omega + \omega_2) \quad [1]$$

where $\mathcal{F}\{s(t)\} = S(\omega)$ and $\mathcal{F}\{\cdot\}$ denotes the Fourier transform of the quantity being bracketed. The spectrum of $s(t)\cos(\omega_2 t)$, along with the spectrum of $s(t)$, is illustrated in Figure 3. It is clear that multiplying $s(t)$ with $\cos(\omega_2 t)$ is a process of heterodyning, as we have translated the spectrum of the signal $s(t)$. This process is known as modulation in communication systems.

In order to appreciate the process of heterodyning let us now consider, for example, heterodyning in radio. In particular, we consider AM. While the frequency content of an audio signal (from 0 to around 3.5 kHz) is suitable to be transmitted over a pair of wires or coaxial cables, it is, however, difficult to be transmitted in air. By impressing the audio information onto a higher frequency, say 550 kHz, as one of the broadcast bands in AM radio, i.e., by

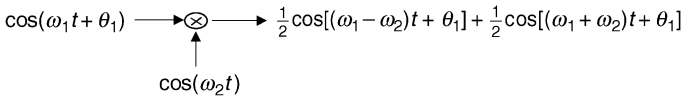


Figure 1 Heterodyning of two sinusoidal signals.

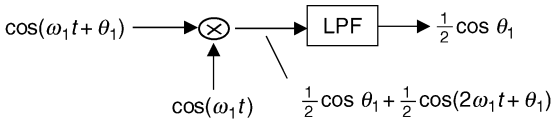


Figure 2 Heterodyning becomes homodyning when the two frequencies to be mixed are the same: homodyning allows the extraction of the phase information of the signal.

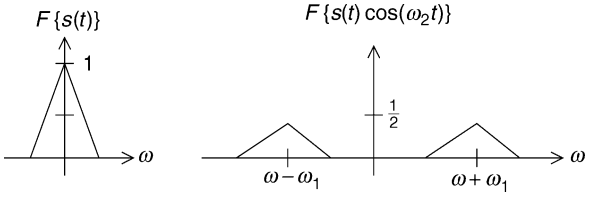


Figure 3 Spectrum of $s(t)$ and $s(t)\cos(\omega_2 t)$. It is clear that spectrum of $s(t)$ has been translated to new locations upon multiplying a sinusoidal signal.

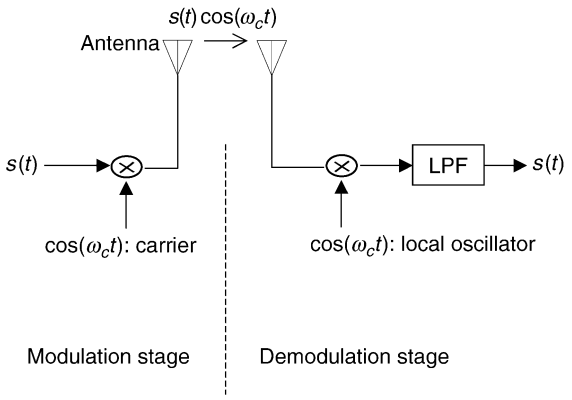


Figure 4 Radio link: heterodyning in the demodulation stage is often known as heterodyne detection.

modulating the audio signal, the resulting modulated signal can now be transmitted using antennas of reasonable dimensions. In order to recover (demodulate) the audio signal from the modulated signal, the modulated signal is first received by an antenna in a radio receiver, multiplied by the so-called local oscillator with the same frequency as the signal used to modulate the audio signal, then followed by a lowpass filter to eventually obtain the audio information back. The situation is illustrated in Figure 4. We shall now describe the process mathematically with reference to Figure 4.

We denote $s(t)$ as the audio signal. After multiplying by $\cos(\omega_c t)$, which is usually called a carrier in radio in the modulation stage, where ω_c is the radian frequency of the carrier. The modulated signal $s(t)\cos(\omega_c t)$. is transmitted via an antenna. When the modulated signal is received in the demodulation stage, the modulated signal is then multiplied by the local oscillator of signal waveform $\cos(\omega_c t)$. The output of the multiplier is given by

$$s(t)\cos(\omega_c t)\cos(\omega_c t) = s(t)\tfrac{1}{2}[1+\cos(2\omega_c t)] \qquad [2]$$

Note that the two signals in the demodulation stage, $s(t)\cos(\omega_c t)$ and $\cos(\omega_c t)$, are heterodyned to produce two new frequencies, the sum $\omega_c + \omega_c = 2\omega_c$ to give the term $\cos(2\omega_c t)$ and the difference $\omega_c - \omega_c = 0$ to give the term $\cos(0) = 1$. Since the two frequencies to be multiplied in the demodulation stage are the same, this is homodyning as explained above. Now by performing lowpass filtering (LPF), we can recover our original audio information $s(t)$. Note that if the frequency of the local oscillator in the demodulation stage within the receiver is higher than the frequency of the carrier used in the modulation stage, heterodyning in the receiver is referred to as superheterodyning, which most receivers nowadays use for amplitude modulation. In general, we have two heterodyning processes in the radio system just described, one in the modulation stage and the other in the demodulation stage. However, it is unusual to speak of heterodyning in the modulation stage, and so we just refer to the process in the demodulation stage as 'heterodyne detection.'

In summary, heterodyne detection can extract the information from a modulated signal and can also extract the phase information of a signal if homodyning is used. We shall see, in the next section, how optical heterodyne detection is employed.

Optical information is usually carried by coherent light such as a laser. Let $\psi_p(x, y)$ be a complex amplitude of the light field, which may physically represent a component of the electric field. We further assume that the light field is oscillating at temporal frequency ω_0. Therefore, we can write the light field as $\psi_p \exp(i\omega_0 t)$. By taking the real part of $\psi_p \exp(i\omega_0 t)$, i.e., $\mathrm{Re}[\psi_p \exp(i\omega_0 t)]$, we recover the usual physical real quantity. For a simple example, if we let $\psi_p = A\exp(-ik_0 z)$, where A is some constant and k_0 is the wavenumber of the light, $\mathrm{Re}[\psi_p \exp(i\omega_0 t)] = \mathrm{Re}[A\exp(-ik_0 z)\exp(i\omega_0 t)] = A\cos(\omega_0 t - k_0 z)$, which is a plane wave propagating along the positive z direction in free space. To detect light energy, we use a photodetector (PD), as shown in Figure 5. Assuming a plane wave for simplicity, as $\psi_p = A$, we have also taken $z = 0$ at the surface of the photodetector. As the photodetector responds to intensity, i.e., $|\psi_p|^2$, which gives the current, i, as output by spatially integrating the intensity:

$$i \propto \int_S |\psi_p \exp(i\omega_0 t)|^2 \mathrm{d}x\mathrm{d}y = A^2 S \qquad [3]$$

where S is the surface area of the photodetector. We can see that the photodetector current is proportional to the intensity, A^2, of the incident light. Hence the output current varies according to the intensity of the

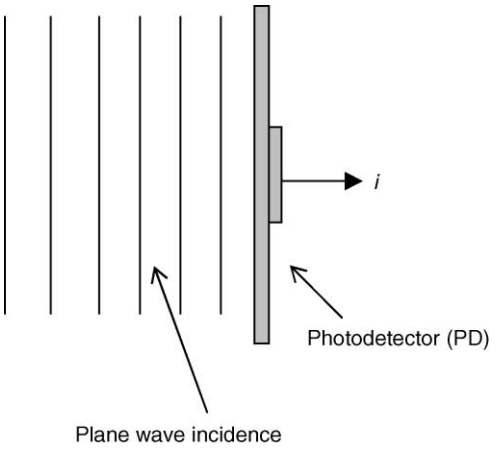


Figure 5 Optical direction detection or optical incoherent detection.

optical signal intensity. This mode of photodetection is called direct detection or incoherent detection in optics.

Let us now consider the heterodyning of two plane waves on the surface of the photodetector. We assume an information-carrying plane wave, also called the signal plane wave, $A_s \exp\{i[(\omega_0 + \omega_s)t + s(t)] \times \exp(-ik_0 x \sin\phi)\}$, and a reference plane wave, $A_r \exp(i\omega_0 t)$, or called a local oscillator in radio. The situation is shown in Figure 6.

Note that the frequency of the signal plane wave is ω_s higher than that of the reference signal, and it is inclined at an angle ϕ with respect to the reference plane wave, which is normal incident to the photodetector. Also, the information content $s(t)$ is in the phase of the signal wave. In the situation shown in Figure 6, we see that the two plane waves are interfered on the surface of the photodetector, giving the total light field ψ_t, given by

$$\psi_t = A_r \exp(i\omega_0 t) + A_s \exp\{i[(\omega_0 + \omega_s)t + s(t)]\} \times \exp(-ik_0 x \sin\phi) \quad [4]$$

Again, the photodetector responds to intensity, giving the current

$$i \propto \int_S |\psi_t|^2 \mathrm{d}x\mathrm{d}y = \int_{-a}^{a}\int_{-a}^{a} [A_r^2 + A_s^2 + 2A_r A_s \cos(\omega_s t + s(t) - k_0 x \sin\phi)]\mathrm{d}x\mathrm{d}y \quad [5]$$

where we have assumed that the photodetector has a $2a \times 2a$ square area. The current can be evaluated to be:

$$i(t) \propto 2a(A_r^2 + A_s^2) + 4A_r A_s \frac{\sin(k_0 a \sin\phi)}{k_0 \sin\phi} \cos(\omega_s t + s(t)) \quad [6]$$

The current output has two parts: the DC current and the AC current. The AC current at frequency ω_s is commonly known as the heterodyne current.

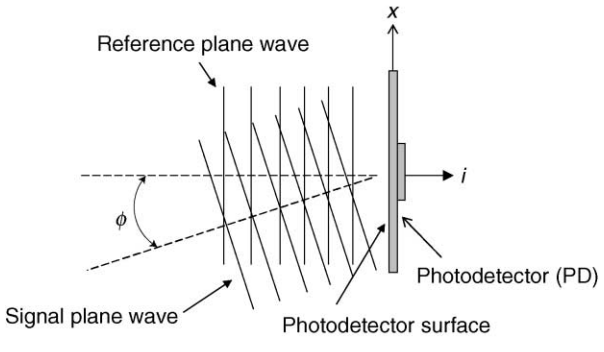


Figure 6 Optical heterodyne detection.

Note that the information content $s(t)$ originally embedded in the phase of the signal plane wave has now been preserved and transferred to the phase of the heterodyne current. The above process is called optical heterodyning. In optical communications, it is often referred to as optical coherent detection. In contrast, if the reference plane wave has not been used for the detection, we have the incoherent detection. The information content carried by the signal plane wave would be lost, as it is evident that for $A_r = 0$, the above equation gives only a DC current at a value proportional to the intensity of the plane wave, A_s^2.

Let us now consider some of the practical issues encountered in coherent detection. Again, the AC part of the current given by the above equation is the heterodyne current $i_{het}(t)$, given by

$$i_{het}(t) \propto A_r A_s \frac{\sin(k_0 a \sin\phi)}{k_0 \sin\phi} \cos(\omega_s t + s(t)) \quad [7]$$

We see that since the two plane waves propagate in slightly different directions, the heterodyne current output is degraded by a factor of

$$\frac{\sin(k_0 a \sin\phi)}{k_0 \sin\phi} = a\,\mathrm{sinc}(k_0 a \sin\phi)$$

where $\mathrm{sinc}(x) = \sin(x)/x$. For small angles, i.e., $\sin\phi \approx \phi$, the current amplitude falls off as $\mathrm{sinc}(k_0 a\phi)$. Hence, the heterodyne current is at a maximum when the angular separation between the signal plane wave and the reference plane wave is zero, i.e., the two plane waves are propagating exactly in the same direction. The current will go to zero when $k_0 a\phi = \pi$, or $\phi = \lambda_0/2a$, where λ_0 is the wavelength of the laser light. To see how critical it is for the angle ϕ to be aligned in order to have any heterodyne output, we assume the size of the photodetector $2a = 1$ cm and the laser used is red, i.e., $\lambda_0 \approx 0.6$ μm; ϕ is calculated to be about 2.3×10^{-3} degrees. Hence to be able to work with coherent detection, we need to have precise optomechanical mounts for angular rotation.

Finally we describe a famous application of heterodyning in optics – holography. As we have seen, the mixing of two light fields with different temporal frequencies will produce heterodyne current at the output of the photodetector. But we can record the two spatial light fields of the same temporal frequency with photographic films instead of using electrical devices. Photographic films respond to light intensity as well. We discuss holographic recording of a point source object as a simple example. Figure 7 shows a collimated laser split into two plane waves and recombined by the use of two mirrors (M) and two beamsplitters (BS). One plane wave is used to illuminate the pinhole aperture (our point source object), and the other

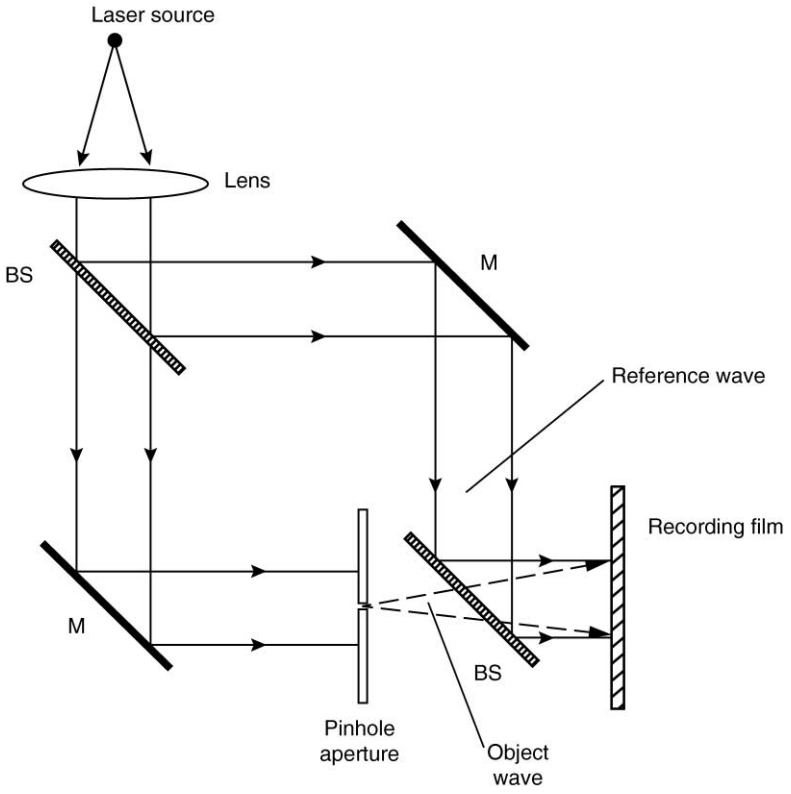


Figure 7 Holographic recording of a point source object.

illuminates the recording film directly. The plane wave that is scattered by the point source object, which is located z_0 away from the film, generates a diverging spherical wave. This diverging wave is known as an object wave in holography. The object wave arising from the point source object on the film is given by, according to Fresnel diffraction:

$$\psi_0 = A_0 \exp(-ik_0 z_0) \frac{ik_0}{2\pi z_0} \times \exp[-ik_0(x^2 + y)^2/2z_0] \exp(i\omega_0 t) \quad [8]$$

This object wave is a spherical wave, where A_0 is the amplitude of the spherical wave.

The plane wave that directly illuminates the photographic plate is known as a reference wave, ψ_r. For the reference plane wave, we assume that the plane wave has the same phase with the point source object at a distance z_0 away from the film. Its field distribution on the film is, therefore, $\psi_r = A_r \exp(-ik_0 z_0) \exp(i\omega_0 t)$, where A_r is the amplitude of the plane wave. The film now records the interference of the reference wave and the object wave, i.e., what is recorded on the film as a 2D pattern is given by $t(x, y) \propto |\psi_r + \psi_0|^2$. The resulting recorded 2D pattern, $t(x, y)$, is called the hologram of the object. This kind of recording is known as holographic recording, distinct from a photographic recording in which the reference wave does not exist and hence only the object wave is recorded. Now:

$$\begin{aligned} t(x,y) &\propto |\psi_r + \psi_0|^2 \\ &= \left| A_r \exp(-ik_0 z_0) \exp(i\omega_0 t) + A_0 \exp(-ik_0 z_0) \right. \\ &\quad \left. \times \frac{ik_0}{2\pi z_0} \exp[-ik_0(x^2 + y)^2/2z_0] \exp(i\omega_0 t) \right|^2 \\ &= A + B \sin\left\{ \frac{k_0}{2z_0} [(x^2 + y)^2/2z_0] \right\} \end{aligned} \quad [9]$$

where A and B are some inessential constants. Note that the result of recording two spatial light fields of the same temporal frequency preserves the phase information (noticeably the depth parameter z_0) of the object wave. This is considered optical homodyning as it is clear from the result shown in Figure 2.

The intensity distribution being recorded on the film, upon being developed, will have transmittance given by the above equation. This expression is called the Fresnel zone plate, which is the hologram of a point source object and we shall call it the point-object hologram. Figure 8 shows the hologram for a particular value of z_0 and k_0. The importance of this phase-preserving recording is that when we

illuminate the hologram with a plane wave, called the reconstruction wave, a point object is reconstructed at the same location of the original point source object if we look towards the hologram as shown in Figure 9, that is the point source is reconstructed at a distance z_0 from the hologram as if it were at the same distance away from the recording film. For an arbitrary 3D object, we can imagine the object as a collection of points, and therefore, we can see that we have a collection of Fresnel zone plates on the hologram. Upon reconstruction, such as the illumination of the hologram by a plane wave, the observer would see a 3D virtual object behind a hologram.

As a final example, we discuss a state-of-the-art holographic recording technique called optical scanning holography, which employs the use of optical heterodyning and electronic homodyning to achieve real-time holographic recording without the use of films. We will take the holographic recording of a point object as an example. Suppose we superimpose a plane wave and a spherical wave of different temporal frequencies and use the resulting intensity pattern as an optical scanning beam to raster scan a point object located z_0 away from the point source that generates the spherical wave. The situation is shown in Figure 10. Point C is the point source that generates the spherical wave (shown with dashed lines) on the pinhole object, our point object. This point source, for example, can be generated by a focusing lens. The solid parallel rays represent the plane wave. The interference of the two waves generates a Fresnel zone plate type pattern on the pinhole object:

$$
\begin{aligned}
&I_s(x, y;\ z_0, t) \\
&= \left| A_r \exp(-ik_0 z_0) \exp(i\omega_0 t) + A_0 \exp(-ik_0 z_0) \frac{ik_0}{2\pi z_0} \right. \\
&\quad \left. \times \exp[-ik_0 (x^2 + y)^2 / 2z_0] \exp[i(\omega_0 + \Omega)t] \right|^2 \\
&= 1 + \left(\frac{1}{\lambda_0 z_0} \right)^2 + \frac{1}{\lambda_0 z_0} \sin\left[\frac{\pi}{\lambda_0 z_0} (x^2 + y^2) - \Omega t \right]
\end{aligned}
\qquad [10]
$$

assuming $A_r = A_0 = 1$ for simplicity. Note that the plane wave is at temporal frequency ω_0, and the spherical wave is at temporal frequency $\omega_0 + \Omega$. The expression $I_s(x, y;\ z_0, t)$ is a temporally modulated Fresnel zone plate and is known as the time-dependent Fresnel zone plate (TDFZP). If we freeze time, say at $t = t_0$, we have the Fresnel zone plate pattern on the pinhole object as shown in Figure 10. However, if we let the time run in the above expression, physically we will have running zones that would be moving away from the center of the pattern. These running zones are the result of optical heterodyning of the plane wave and the spherical wave of different temporal frequencies. It is this TDFZP that is used to raster scan a 3D object to obtain holographic information of the scanned object and such technique is called optical scanning holography.

Upon scanning by the TDFZP, the photodetector, captures all the transmitted light and delivers a current, which consists of a heterodyne current at

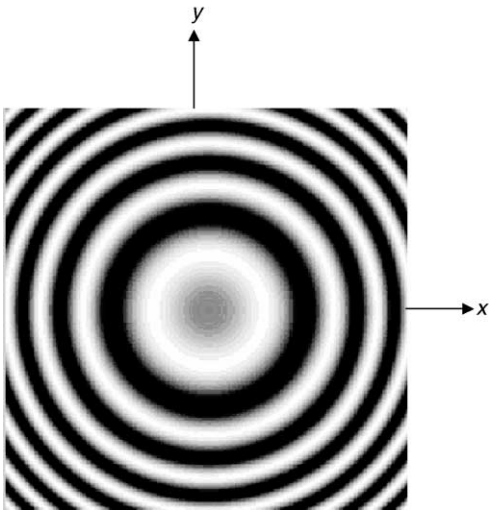


Figure 8 Point-object hologram.

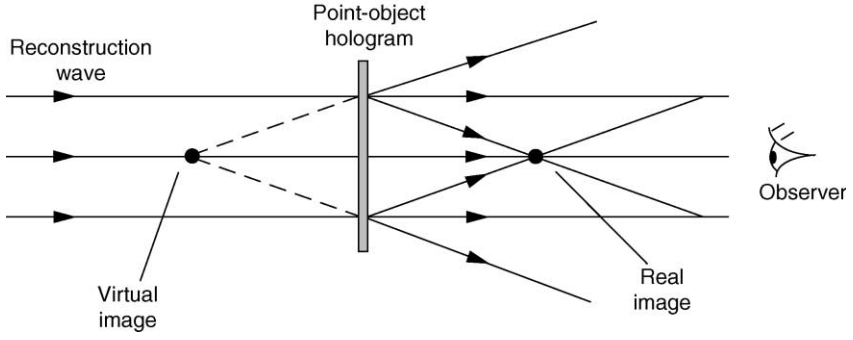


Figure 9 Reconstruction of a point-object hologram.

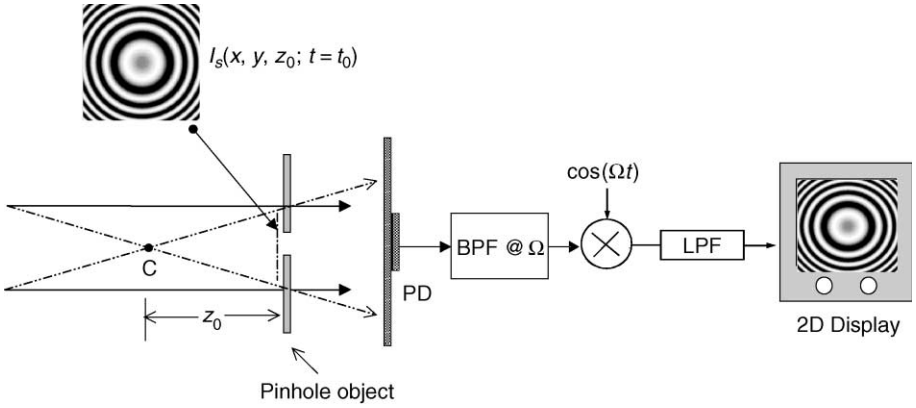


Figure 10 Optical scanning holography (use of optical heterodyning and electronic homodyning to record holographic information upon scanning the object in two dimensions).

frequency Ω. After electronic bandpass filtering (BPF) at Ω, the heterodyne current is homodyned electronically by cos(Ω*t*) and lowpass filtered (LPF) to extract the phase information of the current. When this final scanned and processed current is display in a 2D display, as shown in Figure 10, we have the Fresnel zone plate, which is the hologram of the pinhole object, on the display. We can photograph this 2D display to obtain a transparency and have it illuminated by a plane wave to have the reconstruction as shown in Figure 9, or, since the hologram is now in electronic form, we can store it in a PC and reconstruct it digitally. When holographic reconstruction is performed digitally, we have so-called digital holography.

See also

Diffraction: Fresnel Diffraction. **Holography, Techniques:** Digital Holography.

Further Reading

Banerjee PP and Poon TC (1991) *Principles of Applied Optics*. MA: Irwin.

Palais JC (1988) *Fiber Optic Communications*. New Jersey: Prentice-Hall.

Poon TC and Banerjee PP (2001) *Contemporary Optical Image Processing with MATLAB*. Oxford: Elsevier.

Poon TC and Qi Y (2003) Novel real-time joint-transform correlation using acousto-optic heterodyning. *Applied Optics* 42: 4663–4669.

Poon TC, Wu MH, Shinoda K and Suzuki Y (1996) Optical scanning holography. *Proc. IEEE* 84: 753–764.

Pratt WK (1969) *Laser Communication Systems*. New York: Wiley.

Stremler FG (1990) *Introduction to Communication Systems*. MA: Addison-Wesley.

VanderLugt A (1992) *Optical Signal Processing*. New York: Wiley.

Image Post-Processing and Electronic Distribution

K M Iftekharuddin, The University of Memphis, Memphis, TN, USA

F Ahmed, The Catholic University of America, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The spectacular development of imaging sensors, communication, and internet infrastructure has immensely facilitated the creation, processing, and distribution of electronic digital images. One nontrivial aspect of the digital image is its complete application dependence. The image may be created using a number of different acquisition methods and sensors. It is then processed in different ways, depending upon myriad of digital imaging applications. Consequently, the effectiveness of a useful electronic image processing engine may involve development of robust pre-processing techniques, different types of goal-directed processing such as detection, recognition, identification, classification, tracking, reconstruction and/or registration, as well as post-processing algorithms for different application areas such as multimedia, biomedical, astronomical, defense, consumer, industrial, etc. A notional schematic of this electronic image processing cycle is shown in Figure 1. Note that

while the goal-directed techniques may be different, based on problems at hand, the pre- and post-processing techniques and tools are very similar. Thus, most of the common image pre-processing techniques are also applicable for image post-processing purposes.

Electronic image post-processing techniques involve three broad domains such as: (i) spatial; (ii) frequency; and (iii) time-frequency. Each of these domain techniques may be grouped into different types of processing such as: (i) spatial domain technique to include filtering, spatial processing such as histogram modification, morphological processing, texture processing, etc.; (ii) frequency domain technique to include filter; and (iii) time-frequency domain technique to include short-time Fourier transform, wavelet, and Wigner-ville transforms.

On the other hand, it is useful to identify the typical lifecycle of a digital image from the distribution perspective as shown in Figure 2. The image may be created using a number of different acquisition methods and sensors as mentioned above. The processing block next combines all different types of processing, as mentioned in Figure 1. The image may then be transmitted over the internet or it can leave the digital world to go to the print world. It may then go through wear/tear, subsequently being scanned back to digital domain again. After some processing, it may again be distributed over the internet.

Consequently, the image may be copied legally or fraudulently, processed, and used by a user. Further, the threats and vulnerabilities of the internet have also increased proportionately. The rapid development and availability of high-end computers, printers, and scanners has made the tools of counterfeiting easily accessible and more affordable. With these products, one may counterfeit digital images, which are easy to distribute and modify. As a result, copyright ownership poses a formidable challenge to image post-processing and distribution.

In the subsequent sections, we briefly review relevant background in image post-processing and distribution techniques. We also present a few representative application examples of image post-processing and copyright protection in image distribution.

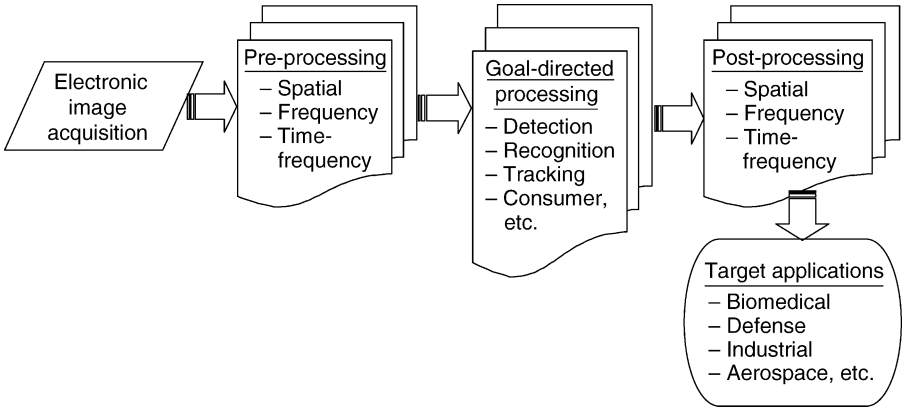


Figure 1 Notional schematic of electronic image processing.

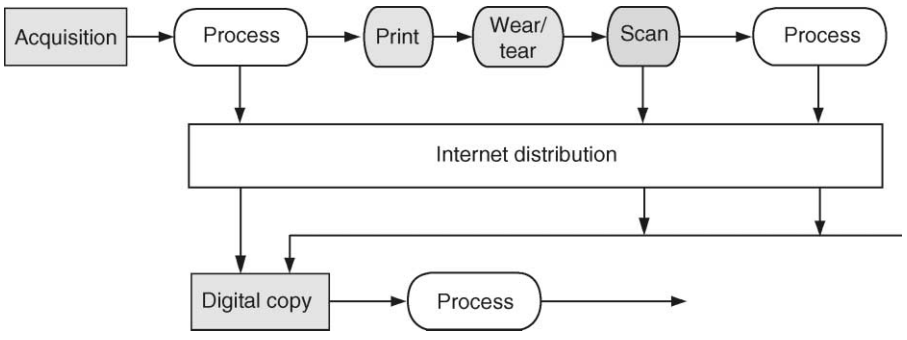


Figure 2 Typical lifecycle of a digital image distribution cycle.

Techniques in Image Post-Processing

Image Formation

We begin by describing the image formation technique briefly. Consider that a three-dimensional (3D) object or scene is imaged onto a 2D imaging plane by means of a recording system such as a camera. If the image formation process is linear, the recorded image may be modeled as the output of the system shown in Figure 3, which is given mathematically by

$$g(x,y) = s\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y;s,t)f(s,t)\mathrm{d}s\,\mathrm{d}t\right\} + n(x,y) \quad [1]$$

where $g(x,y)$ denotes the recorded output image, and $f(x,y)$ is the ideal electronic image. The function $h(x,y;s,t)$ is the 2D impulse response or the point-spread function (PSF) of the image formation system. The PSF determines the radiant energy distribution in the image plane due to a point source located in the object plane. Usually the PSF is normalized such that

$$\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y;s,t)\mathrm{d}s\,\mathrm{d}t\right\} = 1 \quad [2]$$

The noise contribution $n(x,y)$ is shown as an additive random process that is statistically uncorrelated with the image. This is a simplification because noises such as film-grain noise and the noise caused by photon statistics, which often corrupt images, are not uncorrelated with the input. This simplification nonetheless leads to reasonable and useful results. One of the primary goals of image pre-processing (and post-processing as well) is to devise techniques that reduce this additive random noise.

The PSF is a function of only the argument differences $x-s$ and $y-t$ for stationary image and object fields. Thus, it is possible to envision an image formation system that is space invariant; hence the PSF is independent of position. In this case, we may rewrite the superposition integral in eqn [1] into a more familiar convolution integral as follows:

$$g(x,y) = \left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x-s,y-t)f(s,t)\mathrm{d}s\,\mathrm{d}t\right\} + n(x,y) \quad [3]$$

Equivalently, eqn [3] is written as

$$g(x,y) = \{h(x,y)^{*}f(x,y)\} + n(x,y) \quad [4]$$

where $(*)$ is used to denote 2D convolution. Converting from a continuous image $f(x,y)$ to its discrete representation $f(i,j)$ requires the process of sampling. In the ideal sampling system, $f(x,y)$ is multiplied by an ideal 2D impulse train given as

$$f(i,j) = \sum_{m=-\infty}^{+\infty}\sum_{m=-\infty}^{+\infty} f(mX_0,nY_0)\,\delta(x-mX_0,y-Y_0) \quad [5]$$

where X_0 and Y_0 are the sampling distances or intervals, $\delta(^{*},^{*})$ is the ideal impulse function. Note that the square sampling implies that $X_0 = Y_0$. Sampling with an impulse function corresponds to sampling with an infinitesimally small point. If one is interested in image processing, one should choose a sampling density based upon the classical Nyquist sampling theory. Thus, in discrete domain implementation, eqns [3] and [4] are expressed as

$$g(i,j) = \left\{\sum_{\forall(k,l)} h(i,j;k,l)f(k,l)\right\} + n(i,j),\ 0 \le i,j \le N-1 \quad [6]$$

and,

$$g(i,j) = s\left\{\sum_{\forall(k,l)} h(i-k,j-l)f(k,l)\right\} + n(i,j) \quad [7]$$

Thus, we may rewrite eqn [6] in discrete domain as follows:

$$g(i,j) = h(i,j)^{*}f(i,j) + n(i,j) \quad [8]$$

where $(*)$ represents convolution operation. Equation [8] also describes the spatial domain filtering wherein $h(x,y;s,t)$ may be considered as the spatial domain filter mask. The discrete Fourier transform may be used to yield the frequency domain model as follows:

$$G(u,v) = H(u,v)F(u,v) + N(u,v) \quad [9]$$

where G, H, F, and N are corresponding Fourier transforms and u and v are the dummy variables. In eqn [9], $H(u,v)$ is again the frequency domain filter.

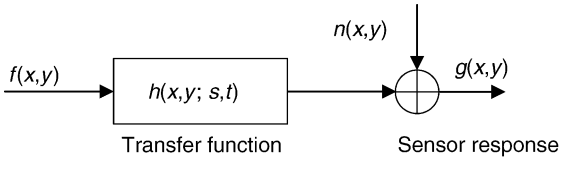


Figure 3 Model for image formation and recording.

We discuss the frequency domain filtering concept further, later in this article.

Spatial Domain Processing

The image pixel domain operations are frequently described as the spatial domain processing. These techniques are important for image pre- as well as post-processing operations. We describe two representative spatial domain processing techniques such as histogram processing and fractal processing for texture analysis below.

Histogram processing

One of the important issues in post-processing is image enhancement. Distortions in image may be introduced due to different noise factors, relative motion between a camera and the object, a camera that is out of focus, or atmospheric turbulence. Noise may be due to thermal measurement errors, recording medium, transmission medium, or digitization process. The histogram processing operation is one of the simplest techniques for enhancing the distorted image for certain type of noises. Histogram equalization is a contrast enhancement technique with the objective to obtain a new enhanced image with a uniform histogram. This can be achieved by using the normalized cumulative histogram as the gray scale mapping function. Histogram modeling is usually introduced using continuous domain functions. Consider that the images of interest contain continuous intensity levels in the interval [0,1] and that the transformation function τ which maps an input image $f(x, y)$ onto an output image $w(x, y)$ is continuous within this interval. We assume that the transfer function, $K_f = \tau(K_w)$, is single-valued and monotonically increasing such that it is possible to define the inverse law $K_w = \tau^{-1}(K_f)$. An example of such a transfer function is illustrated in Figure 4. All pixels in the input image, with densities in the interval D_f to $D_f + dD_f$, are mapped using the transfer function $K_f = \tau(K_w)$, such that they assume an output pixel density value in the interval from D_w to $D_w + dD_w$. Consider that if the histogram h is regarded as a continuous probability density function p, describing the distribution of the intensity levels, then we obtain:

$$p_w(w) = p_f(D_f)/d(D_f) \qquad [10]$$

where

$$D_f = du(f)/df \qquad [11]$$

In the case of histogram equalization, the output probability densities are all an equal fraction of the maximum number of intensity levels in the input image, thus generating a uniform intensity distribution. Sometimes it is desirable to control the shape of the output histogram in order to enhance certain intensity levels in an image. This can be accomplished by the histogram specialization operator which maps a given intensity distribution into a desired distribution using a histogram equalized image as an intermediate stage.

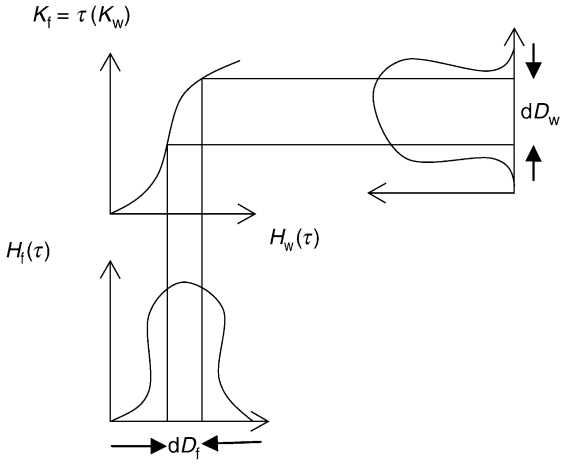


Figure 4 A histogram transformation function.

Fractals for Texture Analysis

The fractal aspect of an image has successfully been exploited for image texture analysis. The fractal concept developed by Mandelbrot, who coined the term fractal from the Latin 'fractus', provides a useful tool to explain a variety of naturally occurring phenomena. A fractal is an irregular geometric object with an infinite nesting of structure at all scales. Fractal objects can be found everywhere in nature such as coastlines, fern trees, snowflakes, clouds, mountains, and bacteria. Some of the most important properties of fractals are self-similarity, chaos, and noninteger fractal dimension (FD). The FD of an object can be correlated with the object properties. Object textures in images are candidates for characterization using fractal analysis because of their highly complex structure. The fractal theory developed by Mandelbrot is based, in part, on the work of mathematicians Hausdorff and Besicovitch. The Hausdorff–Besicovitch dimension, D_H, is defined as:

$$D_H = \lim_{\varepsilon \to 0^+} \frac{\ln N}{\ln 1/r} \qquad [12]$$

where N is the number of elements of box size r required to form a cover of the object. The FD can be defined as the exponent of the number of self-similar pieces (N) with magnification factor ($1/r$) into which a figure may be broken. The FD is a noninteger

value in contrast to objects that lie strictly in Euclidean space. The equation for FD is as follows:

$$\text{FD} = \frac{\ln(\text{number of self-similar pieces})}{\ln(\text{magnification factor})} = \frac{\ln N}{\ln(1/r)} \qquad [13]$$

Frequency Domain Image Filtering

Ignoring noise contribution, we may obtain the basic filtering expression using eqns [7] and [8] as follows:

$$f(i,j) = \Im^{-1}[G(u,v)/H(u,v)] \qquad [14]$$

where $\Im^{-1}$ is the inverse Fourier transform and $[1/H(u,v)]$ is the inverse frequency domain filter. The key importance of this result is that, instead of performing a convolution as shown in eqn [7] to obtain the output of the system, we obtain the inverse of the Fourier transform of the impulse response $h(x,y;s,t)$ and the output image $g(x,y)$, multiply them and then take the inverse Fourier transform of the product to obtain $f(x,y)$. This is the basis for all the filtering as well as other types of image post-processing such as image reconstruction works. The filter function in eqn [9] may have different forms and shapes based on different types of functions such as uniform, Gaussian, lowpass, highpass, etc. While these are all linear filters, there are examples of nonlinear filters such as median, mean, etc. that are equally applicable for image post-processing.

Time-Frequency Domain Processing

One of the fundamental aspects of early image representation is the notion of scale. An inherent property of objects in the world and details in image is that they only exist as meaningful entities over certain ranges of scale. Consequently, a multiscale representation is of crucial importance if one aims at describing the structure of the world, or more specially the structure of projections of the three-dimensional world onto two-dimensional electronic images. The multiscale image post-processing may be facilitated by the wavelet (WT) computation.

A wavelet is, by definition, an amplitude-varying short waveform with a finite bandwidth. Naturally, wavelets are more effective than the sinusoids of Fourier analysis for matching and reconstructing signal features. In wavelet transformation and inversion, all transient or periodic data features can be detected and reconstructed by stretching or contracting a single wavelet to generate the matching building blocks. Consequently, wavelet analysis provides many flexible and effective ways to post-process images that surpass classical techniques – making it very attractive for data analysis, and modeling. The continuous WT is defined as

$$\Phi(a,b) = |a|^{-1/2} \int_{-W}^{W} x(t)\Psi\left[\frac{t-b}{a}\right]\mathrm{d}t \qquad [15]$$

where a and b are scaling and translation factors and $\Psi(t)$ is the mother wavelet. The WT may be viewed as a multiresolution analysis with shifted and scaled versions of the mother wavelet. Corresponding WT coefficients $c_{m,n}$ are

$$c_{m,n} = |a_0|^{-m/2} \int x(t)\Psi([a_0^{-m}t - nb_0])\mathrm{d}t \qquad [16]$$

The reconstruction of the signal is achieved using the admissibility condition, $\int \Psi(t)\mathrm{d}t = 0$, as follows:

$$x(t) = c \int\int_{a>0} \left[\Phi(a,b)\frac{\Psi a, b(t)}{a}\right]\mathrm{d}a\,\mathrm{d}b \qquad [17]$$

where c is a constant. For compact dyadic wavelets, binary scaling and dyadic translation of the mother wavelet form the basis functions. A dyadic translation is a shift by the amount $n/2^m$ that is an integer multiple of the binary scale factor and of the width of the wavelet as well. The discrete WT (DWT) is the natural choice for digital implementation of WT. The DWT basis and 'sub-band coding' process for discretization of a and b are identified. The iterative bandpass filter that finally leads to the wavelet is given as

$$\Psi m, n(x) = 2^{m/2}\Psi(2^{-m}x - n) \qquad [18]$$

The above eqn [18] forms the basis for 'sub-band' implementation for multiresolution analysis (MRA) of wavelet processing.

Image Post-Processing Examples

Some of the important application areas for electronic image post-processing involve the following:

(i) Biomedical image analysis and object detection;
(ii) Image compression for transmission and distribution;
(iii) Satellite image processing and noise removal;
(iv) IR/SAR/detection and clutter removal; and
(v) Image classification and machine vision.

In this article, we provide examples of some of these image-post-processing applications.

Example I: Spatial Domain Processing: Image Post-Processing in Biomedical Image

The mega voltage X-ray images are the X-ray images taken using high voltage X-radiations. These images are often characterized by low resolution, low contrast, and blur. The importance of mega voltage X-ray images arises in the context of radiation therapy for cancer treatment wherein the mega voltage X-radiation is used as a therapeutic agent. Figure 5 shows such mega voltage X-ray images, also known as an electronic portable digital (EPID) image with a very low level of detail. Figure 5a shows poor dynamic range. In order to improve the contrast of this image, without affecting the structure (i.e., geometry) of the information contained therein, we can apply the histogram equalization operation discussed in eqn [10]. Figure 5b presents the result of the histogram image enhancement post-processing process. The low level of detail in the bony anatomy is easily visible in Figure 5b.

Example II: Frequency Domain Processing: Image Post-Processing in Biomedical Tissue Detection

In medical imaging, texture is a fundamental characteristic that can be exploited to analyze tissues and pathologies in a biomedical image. The image texture is often characterized by the random nature of the objects such as tissues and body structures, noise, imaging modality, imaging limitations, and a variety of measurement parameters affecting the image acquisition process. Extensive research suggests that the fractal analysis may be combined with MRA techniques to post-process the texture content of an image in general and tumor detection in particular. In addition, the stochastic fractal Brownian motion (fBm) is well suited to describe texture and tumor characteristics since fBm offers an elegant integration of fractal and multiresolution analysis.

The fBm is a part of the set of $1/f$ processes, corresponding to a generalization of the ordinary Brownian motion $B_{\mathrm{H}}(s)$. It is a nonstationary zero-mean Gaussian random function, defined as

$$B_{\mathrm{H}}(t) - B_{\mathrm{H}}(s) = \frac{1}{\Gamma(H+0.5)} \times \left\{ \int_{-\infty}^{0} [(t-s)^{H-0.5} - s^{H-0.5}] \mathrm{d}B(s) + \int_{0}^{\infty} (t-s)^{H-0.5} \mathrm{d}B(s) \right\}, \quad B_{\mathrm{H}}(0) = 0 \qquad [19]$$

where the Hurst coefficient H, restricted to $0 < H < 1$, is the parameter that characterizes fBm; t and s correspond to different observation times of the process B_{H}, and Γ to the Euler's Gamma function. From the frequency domain perspective, it is known that some of the most frequently seen structures in fractal geometry, generally known as $1/f$ processes, show a power spectrum satisfying the power law relationship:

$$S(\omega) \propto \frac{k}{|\omega|^{\gamma}} \qquad [20]$$

where ω corresponds to the spatial frequency, and $\gamma = 2H + 1$. The observations of stationarity and self-similarity of fBm suggest that the use of time-frequency signal decomposition techniques, such as MRA, are well suited for the analysis of fractal signals. Thus, the fBm analysis framework, along with eqn [13], may be successfully exploited to estimate H, and hence, FD as follows:

$$\mathrm{FD} = D_{\mathrm{E}} + 1 - H \qquad [21]$$

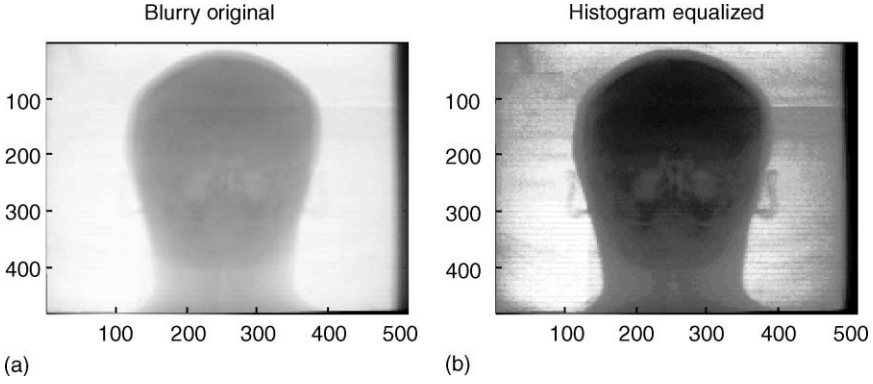


Figure 5 (a) Original EPID image; and (b) Histogram enhanced image.

where D_E is the Euclidean dimension that contains the fBm (i.e., the position of each point of the process is described with the vector $x = (x_1, \ldots, x_E)$).

We exploit the wavelet MRA combined with complementary signal processing techniques such as time-frequency descriptions and parameter estimation of stochastic signals in fBm framework for tumor detection in CT images. We demonstrate an example of image post-processing for tumor segmentation. We test our models to estimate FD in brain MR images as shown in Figures 6a and b, respectively. The images are scanned pixel-by-pixel with a sliding window that defines the subimage for analysis. Pre-processing, such as median filtering, is used to remove noise from the original images.

The results for the spectral analyses of CT image are shown in Figure 7. The results suggest that the FD values do not depend on the mother wavelet used. The window size that offers the best results is the 16×16 pixels, while better discrimination between low and high H values is obtained using biorthogonal wavelet. To improve tumor detection performance of the spectral technique, we apply post-processing filters on the results in Figures 8a and b, respectively. In Figure 8, for an example, we show the filtered fractal analysis results corresponding to a 16×16 pixels window with Daubechies 4. It is possible to obtain an even better description of the location and shape of the tumor after applying a two-dimensional post-processing filter to the matrices derived from the variance estimation algorithms with biorthogonal 1.5 wavelet and an 8×8 pixels window, respectively. The results are shown in Figure 9 and are similar for both averaging and Gaussian filters as well.

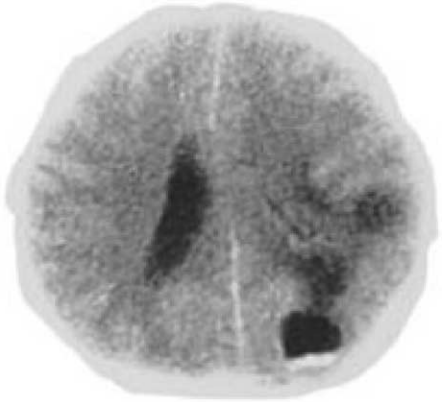


Figure 6 Biomedical images used to test the proposed algorithms.

Example III: Time-Frequency Domain Filtering: Image Post-Processing in Noise Reduction (De-Noising)

This section discusses an example of image post-processing in image noise removal (or de-noising) using a scale-frequency technique such as wavelets. The wavelet transform has been proved to be very successful in reducing noise (de-noising) in an image. We recently propose a new family of wavelets namely Joint Harmonic Wavelet Transform (JHWT) that offers better speckle noise reduction for image post-processing. Unlike wavelets generated by discrete dilation functions, as shown in eqn [18], whose shape cannot be expressed in functional form, JHWT has a simple structure, which can be written, in the

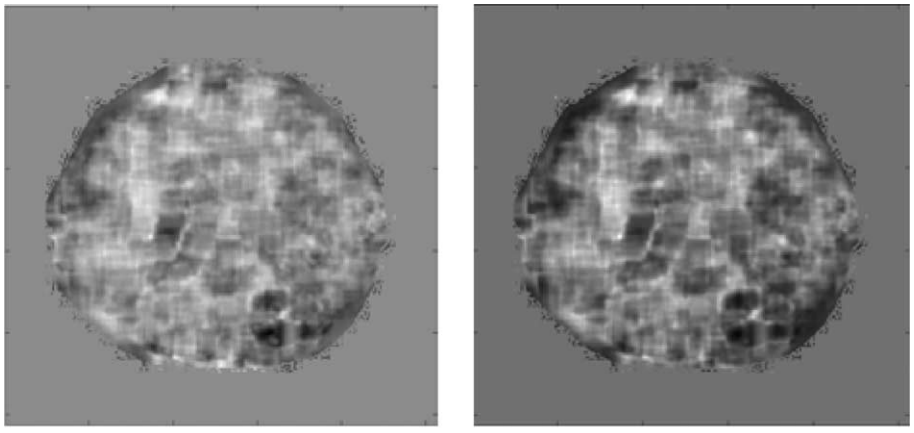


Figure 7 MR-Fractal analysis using spectral method (a) and (b) correspond to a 16×16 analyzing window for Daubechies 4 and biorthogonal 1.5 wavelets, respectively.

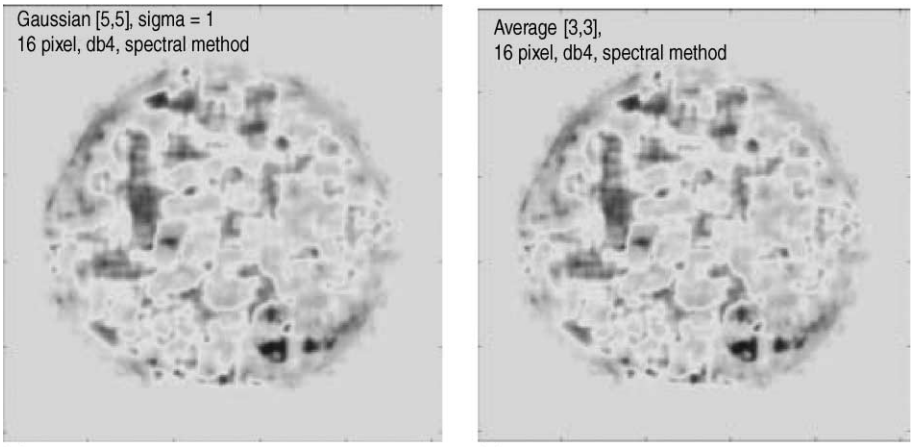


Figure 8 Post-processing results using: (a) 5 × 5 Gaussian filtering with $\sigma = 1$ and (b) 3 × 3 averaging filter.

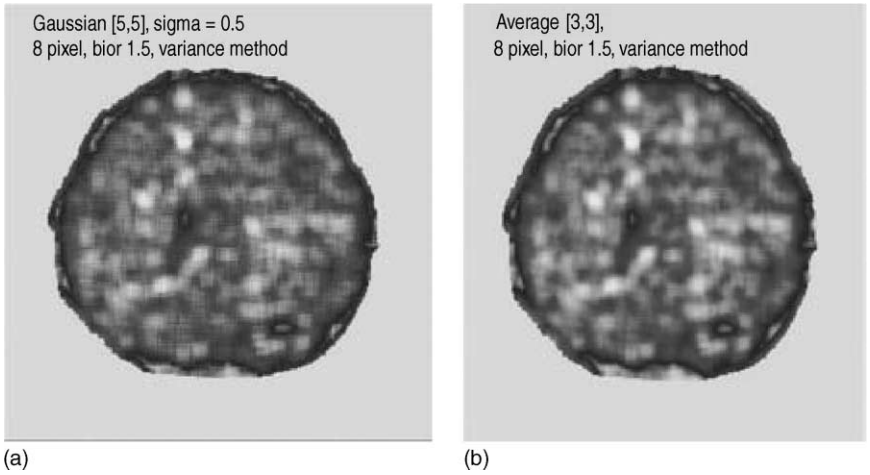


Figure 9 Post-processing on the fractal analysis results of **Figure 8** for (a) 5 × 5 Gaussian filtering with $\sigma = 1$, and (b) 3 × 3 averaging filtering.

frequency domain:

$$W(\omega) = \int_{-\infty}^{\infty} w(x)\exp(-\mathrm{i}\omega x)\mathrm{d}x,$$
$$= \begin{cases} 1 & \text{for } 2\pi \le \omega < 4\pi \\ 0 & \text{elsewhere} \end{cases} \quad [22]$$

Now, at level j, we let $g(x) = w(2^j x)$ and shift it by $k/2^j$ to obtain the harmonic wavelet as:

$$g(x - k/2^j) = \{\exp\{\mathrm{i}4\pi(2^j x - k)\} - \exp\{\mathrm{i}2\pi(2^j x - k)\}\}/\mathrm{i}2\pi(2^j x - k) \quad [23]$$

where integers j and k define the scale and translation of the wavelet $w(x)$, respectively on the x-axis. Now consider a real function $f(t)$ is represented by the sequence f_r of length $(n-1)$, where $r = 0$ to $2^n - 1$. To obtain JHWT coefficients, we take FFT of f_r such that,

$$|F_{2^j+s}|^2 = (1/n^2)\sum_{k=0}^{n-1} |a_{2^j+k}|^2 \exp(-\mathrm{i}4\pi sk/n), \quad [24]$$
where $r = 0$ to $n-1$ and $s = 0$ to $n-1$

where j denotes the level of the wavelet, k is the translation, m and n are adjacent levels of wavelets with $n = 2^j$, and $s = m - n$. The JHWT in eqn [24] offers all the properties and advantages of standard discrete wavelet transform. Putting $n = 2^j$ and $m = s + n$, eqn [24] yields all the JHWT coefficients from



Figure 10 (a) Noisy image; (b) de-noised image with JHWT-filtering; and (c) de-noised image with WP-filtering.

the Fourier coefficients as

$$\sum_{j=0}^{N-1}|F_{2^j+4}|^2=|a_0|^2+\sum_{j=0}^{n-2}(1/2^j)\times\sum_{k=0}^{2^j-1}\left(|a_{2^j+k}|^2+|a_{N-2^j-k}|^2\right)+(a_{N/2})^2=(1/N)\sum_{r=0}^{N-1}f_r^2 \quad [25]$$

Unlike conventional correlation techniques, the perfect octave band limited characteristics of JHWT eliminate the necessity of any multiplication. We show an example of image de-noising in Figure 10 using JHWT.

Image Post-processing, Transmission, and Distribution

One of the areas wherein image post-processing plays a vital role is in image transmission. A significant problem for any digital library is the network bandwidth required to transmit large digital images. The bandwidth consumption depends on the number of images used by a single user, and with the number of users requesting the services of the library. The popularity of database browsers such as Gopher and Mosaic has serious implications for image transmission and distribution. There has been active research in innovative new compression schemes and transmission protocols that reduce bandwidth requirements for many users along with software servers and visualization. Further, compression of digital image and video data plays an important role in reducing the transmission and distribution cost for visual communication.

One of the requirements for effective image transmission and distribution is reduced data size for faster speed. Image compression schemes are often used for image data reduction. Most of the successful image compression schemes exploit subdividing the image into blocks for increased compression. However, this subdivision of image renders a blocky effect on the compressed image. There has been intense research in removing the blocky effect on the compressed image. To achieve high bit rate reduction while maintaining the best possible perceptual quality, post-processing techniques provide one attractive solution. There has been a significant amount of work done in post-processing of compressed image/video. Some of them are developed from the image enhancement point of view while others are from image restoration. In the case of image restoration, the post-processing is considered as a reconstruction problem. Image enhancement methods are performed based on the local contents of the decoded image as discussed above.

The image post-processing algorithms are, generally, very effective in removing blocking artifacts. A straightforward approach to remove blocking artifacts is to apply lowpass filtering to the region, where artifacts occur. Consequently, the filter must be adapted to the block boundaries, i.e., the image partition. Further, if high frequency details such as edges and texture are to be preserved, then lowpass filter coefficients should be selected appropriately in regions where such details occur. Therefore, adaptive image post-processing such as spatial filtering methods may be applied to deal with this problem. In low-bit-rate image and video compression and distribution systems, high frequency transformed coefficients are often lost due to aggressive quantization or uneven error protection schemes. However, with the block-based encoding and transmission methods, the amount of high frequency texture loss is uneven between adjacent blocks. Thus, it is possible to exploit this texture-level correlation between adjacent image blocks to reconstruct the lost texture.

A relevant problem of robust transmission and distribution of compressed electronic images involves unreliable networks. The pre- and post-processing of the image data may play a vital role in addressing this problem. With the rapid growth of mobile wireless

communication systems, wireless multimedia has recently undergone enormous development. However, to design an efficient multimedia communication system over wireless channels, there still exist many challenges due to severe wireless channel conditions, special characteristics of the compressed multimedia data, and resource limitation, such as power supply. Since multimedia data, such as an image, contain redundancy, to efficiently utilize limited resources, source compression may be necessary. Further, as mentioned above, the ease of image creation, processing, transmission, and distribution has also magnified the associated counterfeiting problems. We discuss the relevant issues of image distribution in the following sections.

Techniques in Image Distribution

The spectacular development of imaging sensors, communication, and internet infrastructure has also made the counterfeiting tools easily accessible and more affordable. In addition, the threats and vulnerabilities of the internet have also increased manyfold. All of these together pose a formidable challenge to the secure distribution and post-processing of images.

Requirements of Secure Digital Image Distribution

In light of the above-mentioned vulnerabilities of image distribution channels and processes, below are a few requirements of secure image distribution:

- *Data Integrity and Authentication*: Digital image is easy to tamper with. Powerful publicly available image processing software packages such as Adobe Photoshop or PaintShop Pro make digital forgeries a reality. Integrity of the image content and authentication of the source and destination of the image distribution are, therefore, of utmost importance.
- *Digital Forensic Analysis*: Consider a litigation involving a medical image that is used for diagnosis and treatment of a disease, as shown in the post-processing example in Figures 7 or 8 above. Integrity of the image again is vital for the proper forensic analysis.
- *Copyright Protection*: Digital copyright protection has become an essential issue to keep the healthy growth of the internet and to protect the rights of on-line content providers. This is something that a textual notice alone cannot do, because it can so easily be forged. An adversary can steal an image, use an image processing program to replace the existing copyright notice with a different notice and then claim to own the copyright himself. This requirement has resulted in a generalized digital rights management initiative that involves copyright protection technologies, policies, and legal issues.
- *Transaction Tracking/Fingerprinting*: In the continuous (re)distribution of digital images through internet and other digital media (CD, camera, etc.), it is important to track the distribution points to track possible illegal distribution of images.

Next, we will discuss how the digital watermarking technology can help enforce these security requirements.

Example of secure image distribution: digital watermarking

Digital watermarking is the process of embedding a digital code (watermark) into a digital content like image, audio, or video. The embedded information, sometimes called a watermark, is dependent on the security requirements mentioned above. For example, if it is copyright application, the embedded information could be the copyright notice. Figure 11 shows a block diagram of the digital watermarking process.

Consider that we want to embed a message M in an image, f_o, also known as the original image. For the sake of brevity, we drop the indices representing the pixel coordinates. The message is first encoded using source coding and optionally with the help of error correction and detection coding, represented in Figure 11 as $e(M)$ such that

$$W_m = e(M) \quad [26]$$

The encoded message W_m is then combined with the key-based reference pattern W_k, in addition to the scaling factor α, to result in W_a which is the signal that is actually added to the original image, given as

$$W_a = \alpha(f_o)[W_M \otimes W_K] \quad [27]$$

Note that the scaling coefficient α may be a constant or image dependent or derived from a model of image perceptual quality. When the scaling factor α is dependent on the original image, we can have informed embedding which may result in more perceptually adaptive watermarking. The operation involving W_m and W_k is the modulation part, which is dependent on the actual watermarking algorithm used. As an example, in a typical spread-spectrum based watermarking, this operation is usually an exclusive NOR (XNOR) operation. The final operation is an additive or multiplicative embedding of W_a with the original image. We show a simple additive watermarking process as

$$f_w = f_o + W_a \quad [28]$$

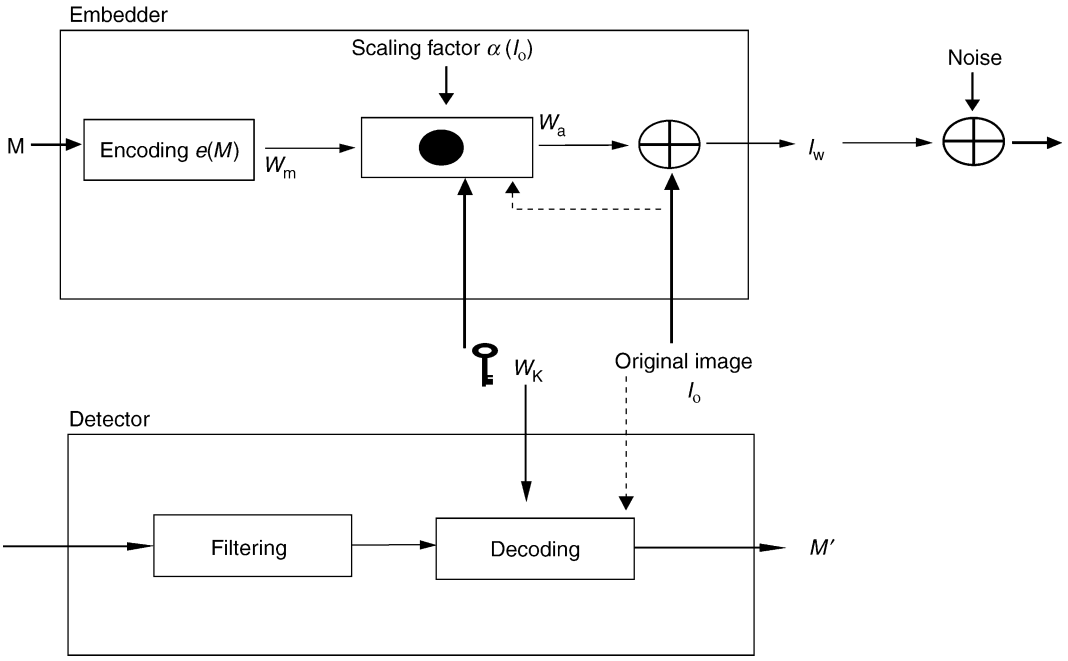


Figure 11 A simple watermarking process: embedder and detector.



Figure 12 (a) Original image, (b) watermarked image, and (c) the difference image.

The result of this process is the watermarked image f_w which may actually need some clipping and clamping operations to have the image pixel values in the desired range.

The detector, as shown in Figure 11, does the inverse operations of the embedder. If the original image is known, as in the case of nonblind watermarking, it is first subtracted from the watermarked image. Blind watermarking detector, on the other hand, uses some filtering with a goal to estimate and subtract the original image component. The result of filtering is then passed to the decoder, which collects the message bits with the help of the shared key, and finally gets back the decoded message M'.

Figure 12 shows the original image, watermarked image, and the difference image using a spread-spectrum based watermarking algorithm, respectively. The enhanced difference image represents a noise-like image, wherein the scaling coefficient α is clearly dependent on the original image. As evident from Figure 12b, the hidden message is imperceptible as well as the change in the original image. The imperceptibility is one of the primary criteria of digital watermarking.

There are other criteria such as unobtrusiveness, capacity, robustness, security, and detectability. Thus, digital watermarking is a multidimensional problem, which attempts to make a trade-off among a number of different criteria. The inserted

watermark should not be obtrusive to the intended use of the original image. Because the watermarking process does not increase the size of the original image, it may be desirable to add as much information as possible. However, generally speaking, the more information one adds, the more severe will be the impact on the perceptual quality of the image. From the detection point of view, watermark should be robust enough to tolerate a range of image post-processing and degradation. In addition, the embedded watermark should also be secure enough so that it may not be removed from the watermarked image. Finally, detectability of the embedded watermark is an important criterion that places some constraints on the embedding algorithm.

In order to fully appreciate the use of watermarking for image distribution, it is instructive to look at different classes of watermark, which is delineated in Figure 13.

For an example, depending on the robustness criterion of watermarking, they can be classified into three categories such as robust, fragile, and semi-fragile. Note that while the robustness of watermark is important, it is not always equally desirable in all applications. Let us take the authentication as an example. If you, as the owner of a digital image, embed some authentication information into the image using watermarking, you might want to see the authentication fail, when some one steals your image. This is an example of nonrobust watermark, where the fragility of the embedded watermark is required to detect any forging done on the digital image.

Now, consider how watermark provides the security requirements for secure image distribution. In particular, we look into the authentication of watermark. Figure 14 shows the use of watermark for image authentication, which has some similarities with the cryptographic authentication using digital signature. However, there are potential benefits to using watermarks in content authentication and verification. Unlike cryptography, watermarking offers both in-storage and in-transit authentication. Watermarking is also faster compared to encryption, which is very important in internet-based image distribution. By comparing a watermark against a known reference, it may be possible to infer not just that an alteration occurred but what, when and where changes have occurred. As shown in Figure 14, first we identify some features from the image and then we compute the digital signature, which is then embedded into the image. The detector takes a test image, computes the signature and extracts the embedded

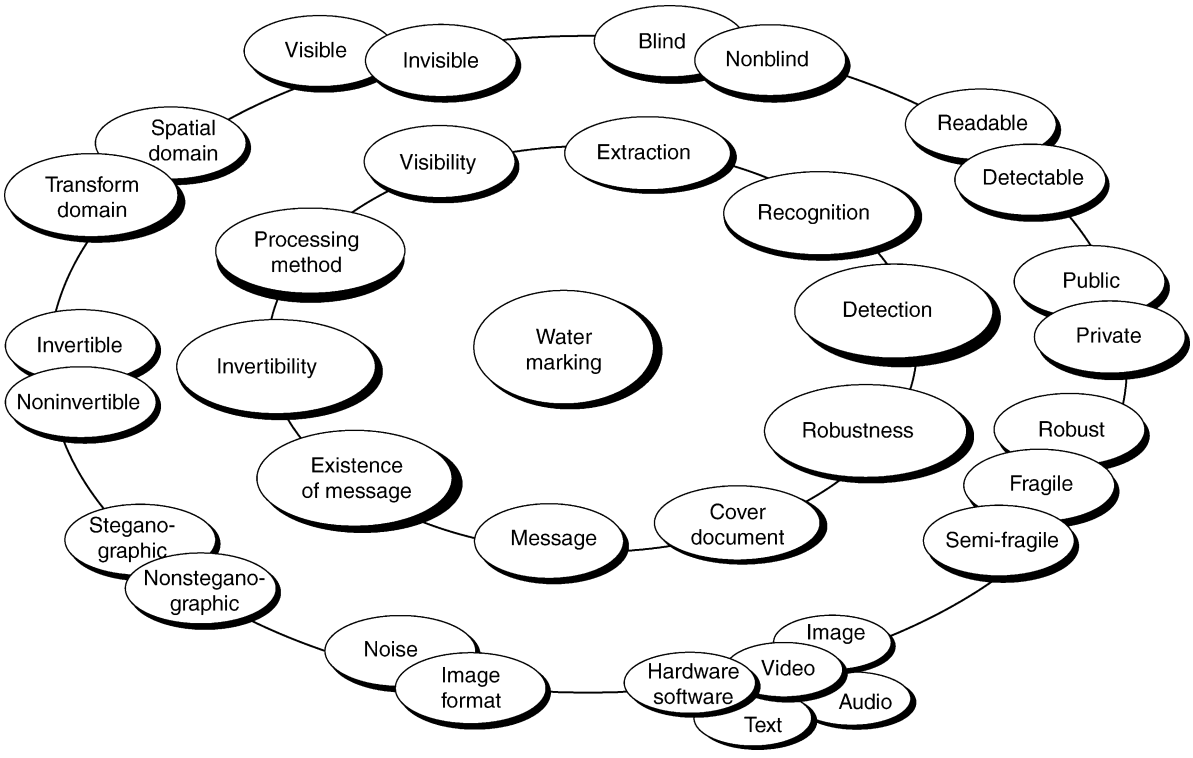


Figure 13 Watermark classifications.

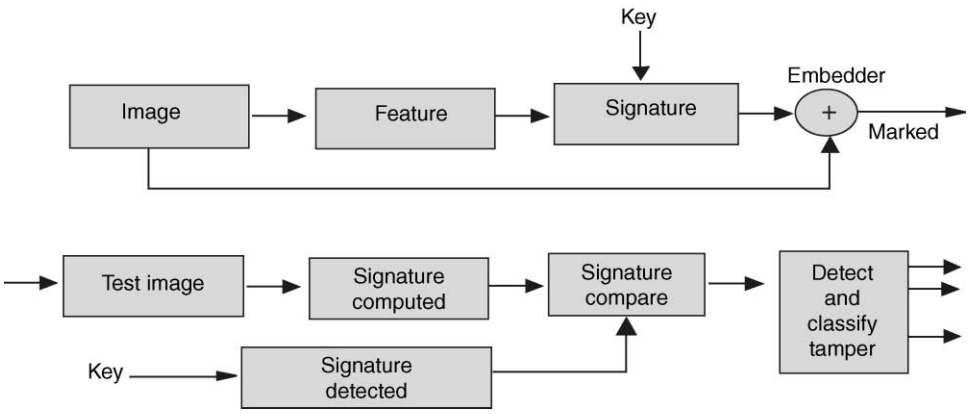


Figure 14 Image authentication using watermark.

signature, and then compares the two. If they match, the image is authenticated. If they do not match, then the image may have gone through some forging/tampering effects and the algorithm may optionally detect the tamper area as well.

One of the limitations of the watermarking approach in secure image distribution is that the technology is not yet widely deployed, and nor is the protocol satisfactorily standardized. Other major limitation is that the watermarking process cannot be made sufficiently robust to arbitrary types of different attacks. In reality, therefore, the hybrid combination of cryptography and watermarking is expected to improve the secure distribution of images across the unreliable internet.

Concluding Remarks

In this chapter, we discussed the principles of image post-processing and distribution. We showed a few representative examples for post-processing and distribution. We also explained the relevant issues in image distribution and present digital watermarking as one of the solutions for secure image distribution.

Acknowledgment

One of the authors (KMI) wishes to acknowledge the partial support of this work through a grant (RG-01-0125) from the Whitaker Foundation.

See also

Imaging: Information Theory in Imaging.

Further Reading

Ahmed F and Moskowitz IS (2004) A correlation-based watermarking method for image authentication applications, accepted for publication in *Optical Engineering*.

Barlaud M (1994) *Wavelets in Image Communication, Advances in Image Communication 5*. Amsterdam: Elsevier.

Chui A (1992) *An Introduction to Wavelets*. San Diego: Academic Press.

Cox NI, Bloom J and Miller M (2001) *Digital Watermarking: Principles & Practice*. Francis, CO: Morgan Kauffman Publishers.

Daubechies I (1990) The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory* 36(5): 961–1005.

Gonzalez RC and Woods RE (1993) *Digital Image Processing*. New Jersey: Addison-Wesley Publishing Company.

Iftekharuddin KM and Parra C (2004) Multiresolution-fractal analysis of medical images. Under Review. *Neuroimaging*, vol. 1. New York: Wiley/IEEE Press, Chapter 4.

Jain AK (1989) *Fundamental of Digital Image Processing*. New Jersey: Prentice-Hall.

Lu N (1997) *Fractal Imaging*. San Diego: Academic Press.

Mallat S (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11(7): 674–693.

Mandelbrot BB (1983) *The Fractal Geometry of Nature*. San Francisco, CA: Freeman.

Schmucker AM and Wolthusen SD (2003) *Techniques and Applications of Digital Watermarking and Content Protection*. Norwood, MA: Artech House.

Serra J (1982) *Image Analysis and Mathematical Morphology*. San Diego: Academic Press.

Stallings W (2003) *Cryptography and Network Security – Principles and Practices*, 3rd edn. New Jersey: Prentice-Hall.

Wu M and Liu B (2002) *Multimedia Data Hiding*. New York: Springer.

Smart Pixel Arrays

P Seitz, Centre Suisse d'Electronique et de Microtechnique SA, Zurich, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

For more than 50 years silicon has been the semiconductor on which the enormous progress of microelectronics is based. Since silicon is also an excellent detector for electromagnetic radiation, in particular in the visible spectrum between 400 and 700 nm, the advances in semiconductor technology have directly influenced the progress of solid-state image sensing. Once the relentlessly shrinking dimensions of transistors have reached a critical minimum limit, it has become economically feasible to co-integrate analog and digital circuitry with each light-sensitive pixel. This capability is at the heart of the technological revolution called smart pixel arrays.

The first section relates how the shrinkage of minimum feature size on silicon chips paralleled the reduction of solid-state pixel geometry, as a function of time. This development ceased once the pixel size reached the optical diffraction limit, which marked the birth of the field of smart pixel arrays. In the second section, the basic functionality and the most important physical limitations of conventional pixels and photosensors are given. The third section deals with opto-electronic functionality that is preferred today, of smart pixel arrays; in particular, the capability of spatio-temporal demodulation is very much in demand. In the fourth section, some of the most useful modular building blocks for smart pixel arrays are presented, together with a few key applications. An outlook on photonic systems-on-chip (SoC) is given in the fifth section, with a discussion of single-chip machine vision systems (so-called seeing chips), organic semiconductors for even higher levels of integration of photonic microsystems, and smart pixel arrays that are sensitive to physical parameters other than electromagnetic radiation.

Silicon Technology for Image Sensing

Long before it was realized that silicon would be the material on which modern semiconductor industry and the information technology revolution is based, in 1940 Russell Ohl discovered that silicon is an excellent detector of electromagnetic radiation. In particular, silicon is an excellent detector for light in the visible and near infrared part of the spectrum, for wavelengths between 400 and 1,000 nm. This is very fortunate because it allows designers of image sensors to profit directly from the amazing progress of silicon technology. Since the 1970s, the minimum feature size on silicon chips showed a reliable exponential reduction of about 10% per year, as illustrated in Figure 1. This is a direct consequence of the observation by Intel's Gordon Moore, sometimes called 'Moore's law', that the number of transistors per integrated circuit doubles about every two years. The evolution of minimum feature size is paralleled by the evolution of minimum pixel size of solid-state image sensor, as illustrated in Figure 2. Once the pixel size reached a record low value of 2.4 μm in the year 1996, no further reduction in pixel size seemed desirable, because the optical diffraction limit was essentially reached.

The combination of technological availability of smaller and smaller electronic components and pixel size that demanded no further reduction, inspired researchers to provide each pixel with an increasing amount of electronic functionality, all of which is available simultaneously for massively parallel information processing. This represents the heart of the smart pixel array revolution; pixels that not only sense the incident electromagnetic radiation but that are capable of pre-processing the acquired information in a relevant fashion.

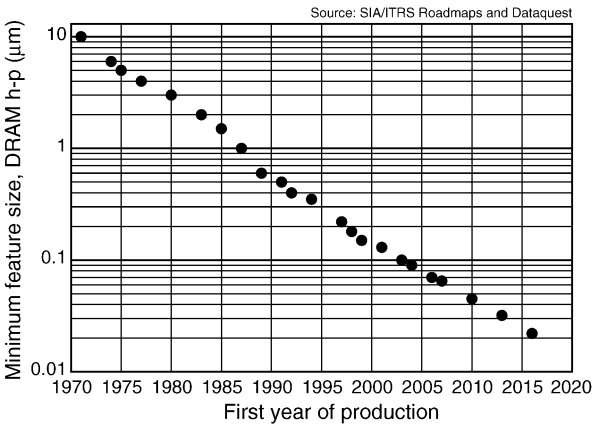


Figure 1 Evolution of minimum feature size (DRAM half-pitch) in state-of-the-art silicon technology, showing an exponential reduction for the past 30 years, which is expected to continue for at least another decade.

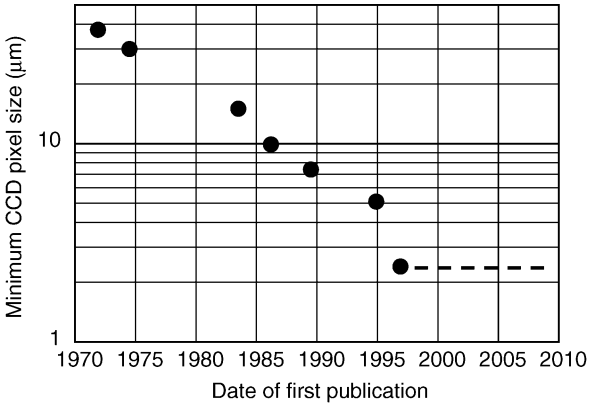


Figure 2 Evolution of minimum pixel pitch in CCD image sensors, tracking the exponential reduction of minimum feature size shown in **Figure 1**, until a lower limit was reached in 1997, imposed by optical diffraction. Since then, no smaller pixel pitches have been described in the literature; this is indicated graphically with the hatched 'diffraction limit' line.

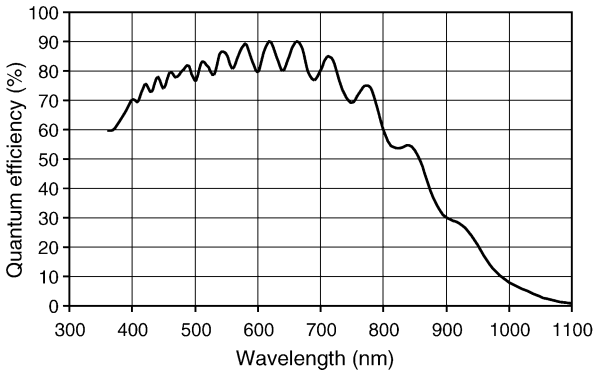


Figure 3 External quantum efficiency of a photosensor realized with an industry standard CMOS process, showing quantum efficiencies above 50% in the whole visible spectral region.

Basic Functionality and Physical Limitations of Conventional Solid-State Photosensors

The prime task of a pixel is the detection of incident electromagnetic radiation. The detection process consists of six basic steps:

1. The incident photon impinges on the surface of the image sensor, and it has to reach the bulk of the semiconductor.
2. In the bulk of the semiconductor the photon has to interact with the semiconductor, converting its energy into a number of electron–hole pairs.
3. The photogenerated charge carriers must be transported to the sensor surface.
4. At the surface, the photogenerated charge needs to be collected and accumulated in the storage part of the pixel.
5. The accumulated photocharge must be transferred to the input of an electronic circuit capable of detecting this charge.
6. The charge detection circuit converts photocharge into a corresponding voltage, exhibiting a monotonous (but not necessarily linear) transfer function.

Since the interaction length of electromagnetic radiation in silicon is in the tens of micrometers for incident photon energies between 1 eV and 10 keV (corresponding to a wavelength range of about 0.1 to 1,000 nm), silicon is an excellent detector for this large spectral range. This is illustrated for the visible and near infrared spectral range in **Figure 3**, showing the absolute external quantum efficiency of a pixel realized with an industry standard silicon process. In the visible spectral range, absolute external quantum efficiencies close to 100% can be reached, even with standard silicon processes for the microelectronics industry that were not optimized for photosensing applications.

Photocharge transport to the surface of the semiconductor is mainly provided by diffusion, i.e. random thermal motion of the charge carriers that exhibit a velocity of about 10^5 m/s in silicon at room temperature. The mean distance L traveled by a diffusing charge carrier during time t is given by

$$L = \sqrt{Dt} \qquad [1]$$

with the diffusivity D, which has a value of around 40 cm^2/s for electrons in low-doped silicon at room temperature.

Once the photocharge carriers are close to the surface of the semiconductor, they are transported by drift in an electric field E, exhibiting a linear dependence of the transport distance L with time t:

$$L = \mu E t \qquad [2]$$

with the mobility μ which has a value of around 1300 cm^2/Vs for electrons in low-doped silicon at room temperature.

There are two basic principles for creating the electric field with which the photocharge carriers are transported to their storage site within the pixel. A photodiode consists of a semiconductor junction between a p-doped and an n-doped region, as illustrated in **Figure 4**. Such a photodiode also represents a capacitance on which photocharge can be accumulated and stored. Photodiodes are used as photosensors in one

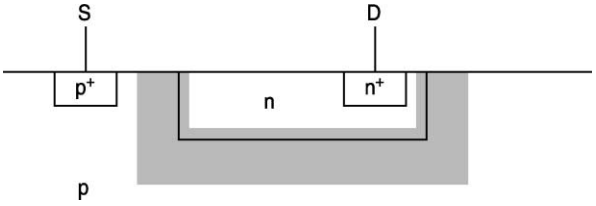


Figure 4 Cross-section of a photodiode, consisting of an n-doped region in a p-doped substrate. The substrate is contacted using a heavily p-doped diffusion S, and the n-region is contacted using a heavily n-doped diffusion D. The space–charge region is illustrated as hatched area at the pn-junction.

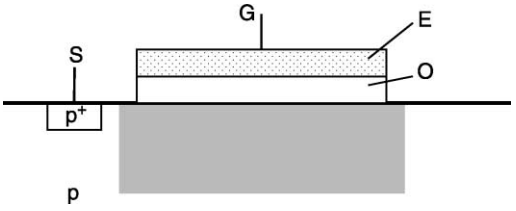


Figure 5 Cross-section of an MOS capacitance, the basic element of a CCD image sensor. On top of a p-substrate a thin transparent insulator O and a transparent electrode E are fabricated. The substrate is contacted with a heavily p-doped diffusion S, and the electrode E is contacted with the gate contact G. The hatched area illustrates the space–charge region under the electrode.

popular type of image sensor, so-called APS (active pixel sensor) or CMOS (complementary metal oxide semiconductor) imagers.

An alternative to a photodiode is an MOS (metal oxide semiconductor) capacitor, consisting of a transparent electrode that covers an insulating layer on top of semiconducting material, as illustrated in Figure 5. An MOS capacitor is the basic element of a charge coupled device (CCD), the other widespread type of image sensor.

The charge transfer between photodiodes and photodetector circuits is based on electrical conduction, occurring essentially with the speed of light. In CCDs, the charge transfer is effectuated with clocks, and excellent charge transfer efficiencies can be observed at clock frequencies of several tens of MHz.

The major limitation in semiconductor photosensors is the last element in the line, i.e. the electronic charge detection circuit. In Figure 6, a so-called source follower is illustrated, which is by far the most frequently used photocharge detection circuit. Its performance is limited by the thermal (Johnson) noise in the channel of the first field effect transistor (FET). This corresponds to a statistical uncertainty with which an amount of photocharge can be measured at the gate of the FET. The input referred charge noise

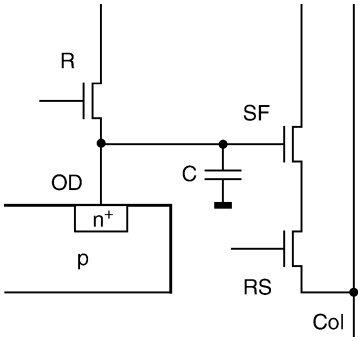


Figure 6 Schematic diagram of a photosensor's output diffusion OD connected to a reset transistor R and a source follower transistor SF, the most widely used charge detection circuit for image sensors and smart pixel arrays. The photocharge is stored on the effective capacitance C at the gate of transistor SF, which is read out using a row select transistor RS, connecting the source follower to the readout column Col.

variance s_Q is given approximately by

$$s_Q \cong C^2 \frac{4kTB}{g} \qquad [3]$$

where C is the effective capacitance seen by the gate of the FET, k is Boltzmann's constant, T is the absolute temperature, B is the effective bandwidth at the FET's channel, and g is the transconductance of the FET.

Desired Optoelectronic Functionality in Smart Pixel Arrays

From the previous section it becomes clear that even conventional pixel and photosensors can be improved in several desirable respects:

- The spectral response can be improved in the ultraviolet (50–400 nm), or it can be expanded significantly into the X-ray region (below 0.1 nm) and into the infrared region (above 1,000 nm).
- The speed of the photosensors can be increased by replacing diffusion based photocharge transfer mechanisms with the directed, more efficient drift in electric fields. This is true for the vertical as well as for the lateral charge transport direction.
- The sensitivity of the electronic charge detection process can be improved in two ways. It can be achieved by decreasing the effective detection capacitance C at the input of the charge detection circuit, as can easily be seen from eqn [3]. An alternative is to make use of a physical charge amplification process (i.e. avalanche multiplication in a high electric field), that has low intrinsic noise.

- A derivative improvement that is desired is the increase of the dynamic range D/R of the complete detection process, as defined, for example, by

$$D/R = 10^{10} \log \frac{Q_{\max}^2}{s_Q} \qquad [4]$$

where $Q_{\max}$ denotes the maximum (or saturation) charge and s_Q is the total charge noise variance of the electronic detection circuit.

In addition to the above described improvements of conventional pixel functionality, a huge demand exists for the capability to measure the parameters of modulated electromagnetic signals, i.e. to demodulate such signals. This demand is related to the large number of very powerful electronic and opto-electronic measurement principles, capable of improving the sensitivity of the measurement process significantly by employing modulated signals, either in the temporal or in the spatial domain. For this reason, smart pixel arrays are sought that offer this combination of photodetection and demodulation.

Modular Building Blocks for Smart Pixel Arrays

Towards the end of the last century, a large body of literature has been created, addressing various approaches for the realization of one of the above described desired opto-electronic functionalities. The result is a powerful 'toolbox' of modular building blocks for custom smart pixel arrays, with which optimized solutions for optical measurement problems and practical applications can be constructed.

High-Sensitivity Charge Detection

From the noise formula (eqn [3]) two possibilities for the improvement of the sensitivity of image sensors can be deduced: A first approach is the reduction of the capacitance value. An excellent example for this is the invention of the 'double gate' FET, in which the charge to be measured is not placed on a gate electrode at the surface but resides in the semiconductor bulk. In this way the effective capacitance of the pixel can be reduced by at least one order of magnitude compared to conventional pixels. The resulting capacitance of less than 1 fF leads to a readout noise of about one electron at room temperature and for a readout bandwidth of several MHz. Unfortunately the small capacitance allows the storage of only around ten thousand electrons. For such small charge packets, quantum noise is still relevant, leading to significantly degraded image quality, even when the images are taken under favorable lighting conditions.

A second approach to high-sensitivity photosensing exploits the dependence of the charge detection variance on the measurement bandwidth, as indicated in eqn [3]. One method for exploiting this consists of reading out the image sensor at a significantly reduced bandwidth of a few 10 kHz. At the same time, the image sensor is cooled to temperatures of around −100 °C. This becomes necessary because readout at such slow speeds increases the time, to read one complete image, to many minutes, requiring a substantial reduction of the dark current. Instead of working at low readout bandwidths, it is also possible to read out the same information repeatedly and non-destructively at high speed, and to average many measurements of the same amount of photocharge. Both methods have been shown to lead to readout noise levels of about one electron; on the other hand, both approaches require very long times for the acquisition of complete images.

This problem can be resolved by exploiting the massive parallelism that smart pixel arrays offer: low-bandwidth techniques can be employed without losing effective readout speed. Pixels and columns are read out with low-bandwidth amplifiers that are all active simultaneously, and only the final off-chip readout amplifier needs to operate at a large bandwidth. In practice, this approach has been used successfully in single-chip digital cameras with Megapixel resolution, exhibiting a readout noise of only a very few electrons.

A completely different approach is based on the use of a physical amplification mechanism with low intrinsic noise, before detecting the amplified charge packets with a conventional electronic circuit; the avalanche multiplication effect in high electric fields is very well suited to this approach. Each pixel is supplied with its own combination of stabilized avalanche multiplier and conventional charge detection circuit. This approach has been used for the realization of a fully integrated 2D array of single photon detectors that can be fabricated with industry standard CMOS processes. Similar performance is obtained with the Impactron™ series of high-sensitivity CCD image sensors offered by Texas Instruments Inc., making use of a single multistage avalanche multiplier at the output of the image sensor. For both high-sensitivity smart pixel arrays, single-photoelectron performance is reported at room temperature and at readout frequencies of several MHz, as required for video applications.

Photosensing with Improved Dynamic Range

Conventional solid-state image sensors and cameras usually have a very limited dynamic range, as defined in eqn [4]. Typical values are between 40 and 80 dB, rarely exceeding four orders of magnitude. In natural scenes, however, much higher D/R values can be encountered. To cope with this, the human eye is capable of spanning a dynamic range of more than 150 dB. This realization has spurred the invention of several methods with which the D/R of smart pixel arrays can be increased.

One approach consists of acquiring several images of the same scene with different exposure times, with the subsequent combination into one high-dynamic-range composite picture. D/R values exceeding 120 dB have been obtained with such multi-exposure methods. Fast-moving objects, however, will have relocated from one exposure to the next, resulting in false, blurred representations of these objects in the composite picture.

This can be overcome using nonlinear response characteristics similar to the ones found in natural vision. Impinging light produces a photocurrent that is highly linear with the incident light intensity. Because the voltage across the photodiode is logarithmically dependent on the photocurrent, a logarithmic voltage-illumination response is obtained, with a D/R exceeding 120 dB. Such logarithmic pixels are well-suited for applications providing high illumination levels but they suffer from excessive noise and slow response speed at low light levels.

This disadvantage led to the development of linear-logarithmic pixels, exhibiting a high-sensitivity linear response at low light levels, and a gradual changeover to logarithmic response once a certain, programmable light level is reached. The design principle is illustrated in Figure 7. An otherwise conventional CMOS/APS pixel is supplied with a device that exhibits logarithmic voltage–current response, such as a diode. As long as the cathode is at a higher voltage than the anode, no current flows across the diode, and the pixel response is linear. Once the photocurrent has lowered the cathode voltage below the anode voltage, the pixel response becomes logarithmic, as described above. The D/R of such a linear-logarithmic pixel reaches unprecedented high values; almost 200 dB have already been obtained in practice. Because of the simplicity and robustness of this principle, such linear-logarithmic smart pixel arrays have found their way into commercial sensor and camera products that are available commercially, as illustrated in Figure 8. A circuit board and a light bulb are imaged in linear mode (Figure 8a), showing highly saturated areas, and in linear-logarithmic mode (Figure 8b), in which even the light bulb's filament can clearly be seen.

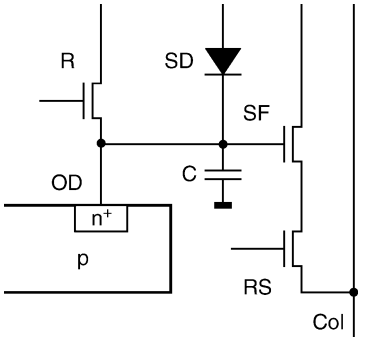


Figure 7 Schematic diagram of a high-dynamic-range pixel, consisting of a conventional APS/CMOS pixel, as illustrated in Figure 6, in combination with a shunt diode SD. The photocurrent lowers the voltage at the gate of transistor SF continuously. Below a certain value of this voltage, the diode starts to become forward biased, leading to a linear-logarithmic illumination response, with a dynamic range that can be as high as 200 dB.

Extension of the Spectral Sensitivity Range

The spectral sensitivity of silicon is essentially restricted to the spectral wavelength range between 0.1 and 1,100 nm. With a suitable choice of materials that cover the silicon sensor's surface, this sensitive wavelength range can be extended. For high-energy radiation, scintillator materials are employed, capable of converting incident high-energy photons into a proportional number of visible photons. Scintillators render smart image sensors particularly sensitive in the UV (300–400 nm), deep UV (below 300 nm), X-ray (0.002–1 nm), and gamma ray (below 0.002 nm) spectral region. Typically, the scintillator materials are fabricated as platelets that are glued on the surface of the produced CMOS/CCD image sensor.

The extension of the sensitive spectral range to the infrared can be achieved with semiconductor materials with smaller energy bandgaps than silicon. Examples include Ge, SiGe, SiC, PtSi, IrSi, InAs, InGaAs, or InGaAsP. Typically, these materials are deposited on the produced CMOS/CCD image sensor in an ultra-high vacuum environment. An alternative is the use of bump-bonding, with which the CMOS/CCD image sensor can be connected to a small-bandgap material chip at each pixel site. With these technologies, silicon-based smart pixel arrays become sensitive to wavelengths of several microns.

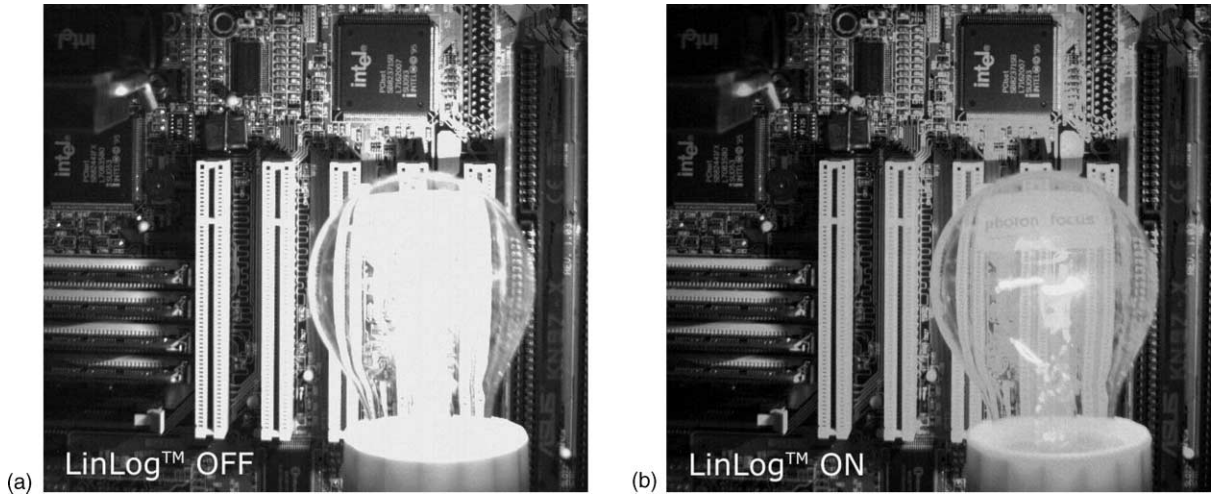


Figure 8 Images of a scene with high dynamic range, taken with an image sensor exhibiting linear illumination response (a) and linear-logarithmic illumination response (b). A similar pixel architecture as illustrated in Figure 7 is employed. (Courtesy of Photonfocus AG, Lachen, Switzerland.)

High-Speed Image Sensing

Increasing the speed of the photosensing process necessitates improvements in two key areas. Photocharges need to be collected and transported faster, and the acquired information must be read out faster.

A comparison of eqns [1] and [2] immediately reveals that the replacement of diffusion by drift transport provides the desired speed improvement. In the vertical direction, the electric field of photodiodes or MOS capacitances reaches more deeply into the semiconductor bulk if lower doping levels of the substrate are used or the bias voltage is increased. Improving the lateral transport speed is more difficult. A highly successful approach in so-called drift-field pixels is illustrated in Figure 9. On top of a thin (a few tens of nanometers) transparent insulator, a highly resistive layer – often of polysilicon – is placed. When a voltage difference is applied at the two ends, a small current flows across the resistive layer, inducing a linear potential variation at the semiconductor surface. A simple calculation shows that very small transit times for photocharges can be obtained in this way. Assuming a (quite large) pixel size of 25 μm and a voltage difference of 5 V, the maximum transit time of photoelectrons that have to move from one end of the pixel to the other, is less than 1 ns. This is more than two orders of magnitude faster than can be achieved with diffusion transport.

Smart pixel arrays provide a simple solution to the problem of speeding up the readout of the acquired picture information; a large degree of parallelism. This is illustrated with the architecture of a split-field high-speed image sensor, shown schematically in Figure 10. The photosensitive area is split into two

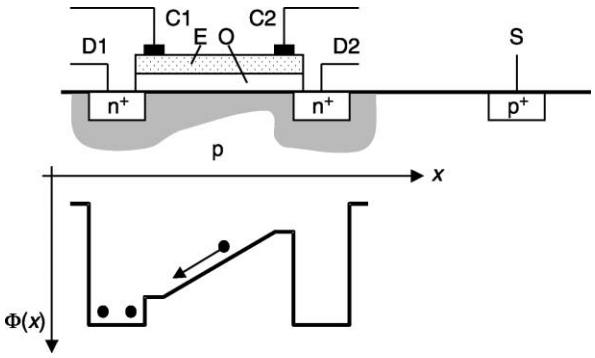


Figure 9 Cross-section of a drift field pixel, exploiting the lateral drift field at the semiconductor–oxide interface. The corresponding sloped potential distribution $\Phi(x)$ is created with a highly resistive transparent electrode E on top of a thin transparent insulator O; this conductor has two contacts C1 and C2 at its ends, to which a voltage difference is applied. Depending on the sign of the slope, the photoelectrons are collected either in diffusion D1 or D2. The hatched area indicates the space-charge region.

parts, so that the column amplifiers of one half have time to settle to their proper voltages, while the columns of the other half are being read out. No time is lost, therefore, when another row of the image is selected for readout. A large number of output amplifiers are provided for highly parallel readout of the image data, each with a typical readout rate of 40–100 MHz. In this way, sustainable pixel rates of several GHz can be achieved in practice.

Static Spatial Demodulation

The spatially varying geometry of the photosensitive part of a pixel can be used for the static spatial

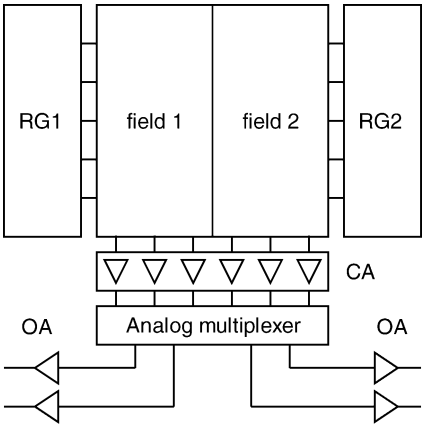


Figure 10 Block diagram of a high-speed image sensor, making use of the large amount of parallelism possible in smart pixel arrays. The light-sensitive area is split into two fields whose rows can be addressed independently with two separate row address generators RG. In this way, one row of one field can be selected and the corresponding column amplifiers CA can reach equilibrium, while the columns of the other field are being read out using several parallel output amplifiers OA.

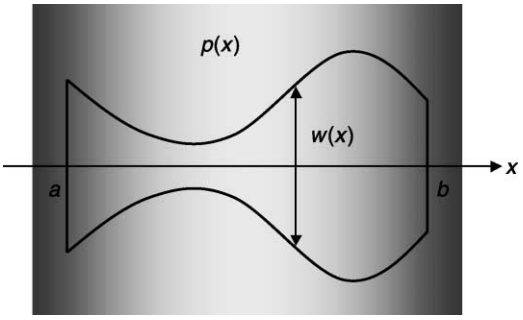


Figure 11 Illustration of a one-dimensional intensity distribution $p(x)$ of light which is incident on a shaped photodiode with position-dependent width $w(x)$, between the two boundaries a and b.

demodulation of an impinging light field. When a one-dimensional intensity distribution $p(x)$ is incident on a photodiode with spatially varying width $w(x)$ between the limits a and b, as illustrated in Figure 11, an output signal r is produced:

$$r = \int_a^b p(x)w(x)\mathrm{d}x \qquad [5]$$

By shaping the photodiode appropriately, any type of parametrized linear transform can be realized, as illustrated by the photomicrograph in Figure 12. In this practical example, four photodiodes are shaped sinusoidally and cosinusoidally, weighted with a slowly varying windowing function. Their signals are used for determining the Fourier coefficients

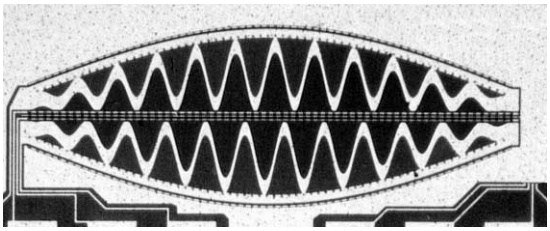


Figure 12 Photomicrograph of two photodiode pairs that are shaped as spatial sine and cosine functions, weighted with a slowly varying windowing function. The four photosignals can be employed to determine the phase and amplitude of a one-dimensional intensity signal that has the same spatial frequency as the sine and cosine. In practice, the accuracy of the phase determination is as high as 0.1% of the spatial wavelength. The width of the photomicrograph corresponds to about 1 mm in reality.

(the sine and cosine transform) of an incident light distribution for a fixed spatial frequency. This is employed, for example, in cost-effective optical encoders, capable of measuring linear displacement with an accuracy below 50 nm.

Dynamic Spatial Convolution

Many image-processing tasks require the convolution of the pixel data $P_{i,j}$ at the position (i, j) in an image with a two-dimensional convolution kernel $w_{k,l}$:

$$R_{k,l} = \sum_{i=-N}^{N} \sum_{j=-N}^{N} P_{i,j} w_{k-i,l-j} \qquad [6]$$

Instead of employing a purely geometrical solution, as described in the previous section, an alternate approach is found when interpreting the convolution eqn [6] in the following way. The resulting image R is obtained by displacing the original image P laterally by an amount (k,l), by multiplying the shifted pixel values with the corresponding weight $w_{k,l}$, and by adding up all these products.

The lateral displacement is realized with a CCD structure, implementing the 2D lateral movement of the acquired image charge packets, the multiplications with the convolution kernel values are realized by using various exposure times for each lateral shift position, proportional to the corresponding weight value, and the summation is just the natural charge accumulation of CCD structures. The problem of negative weights is solved by separately determining the positive and negative parts of the weight function, each stored separately in the pixel, and by calculating the difference of the two values during readout, either on-chip or off-chip. In this way, such convolution image sensors

can be used for the determination of complete, programmable, real-time image convolutions, with arbitrary kernels that are fully under the dynamic control of digital logic or microprocessors. Practical experience with such convolution imagers shows that the convolution results exhibit inaccuracies of less than 1%.

Temporal Demodulation

In many relevant optical measurement techniques, temporally modulated light fields are employed. For this reason, smart pixel arrays are desired, providing synchronous demodulation of these signals. An electromagnetic signal $S(t)$ that is harmonic as a function of time with given frequency ω, having unknown modulation amplitude A, offset B and phase delay φ, is described by

$$S(t) = A \sin(\omega t + \varphi) + B \qquad [7]$$

The three unknowns A, B and φ can be measured if three or more sampling points are sampled synchronously per period in so-called 'lock-in' pixels. An example of a lock-in pixel employing the drift-field transport principle with four accumulation taps, is illustrated in Figure 13. Demodulation frequencies of several tens of MHz have already been demonstrated in practice.

The most important application of such synchronous demodulation pixels is in optical time-of-flight range cameras. Such a camera incorporates a light source, usually an array of infrared or visible LEDs, that is modulated at a few tens of MHz. The modulated light is reflected diffusely by the various objects, and the camera's lens images it onto a lock-in image sensor. Each of its pixels carries out synchronous sampling/accumulation of photocharges. At the end of the exposure time, the accumulated photocharge samples are read out, and the three modulation parameters of the reflected light are determined for each pixel. The phase shift is a direct measure of the local distance. It has been shown that for a wide range of operating conditions, the distance resolution of such optical range cameras is already limited just by the photo shot noise. A typical result of such a depth image is shown in Figure 14.

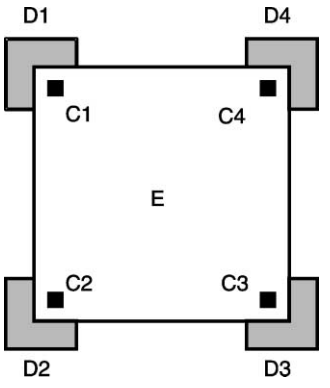


Figure 13 Schematic top view of a lock-in pixel based on the lateral drift-field approach illustrated in Figure 9. A square electrode is fabricated with a highly resistive transparent material on top of a thin insulator. The electrode contacts C1 to C4 are employed to create a 'reeling' potential distribution that rotates with the modulation frequency of the incident light that requires synchronous demodulation, so that the photocharges are collected at diffusions D1 to D4, respectively, at the appropriate time during each period.

Photonic Systems-on-chip

The microelectronics revolution is based on the key notion of integration. More functionality integrated monolithically on the same chip leads to faster, smaller, better performing, and cheaper products. The same is true for smart pixel arrays, for which progress through integration is attained in several respects.

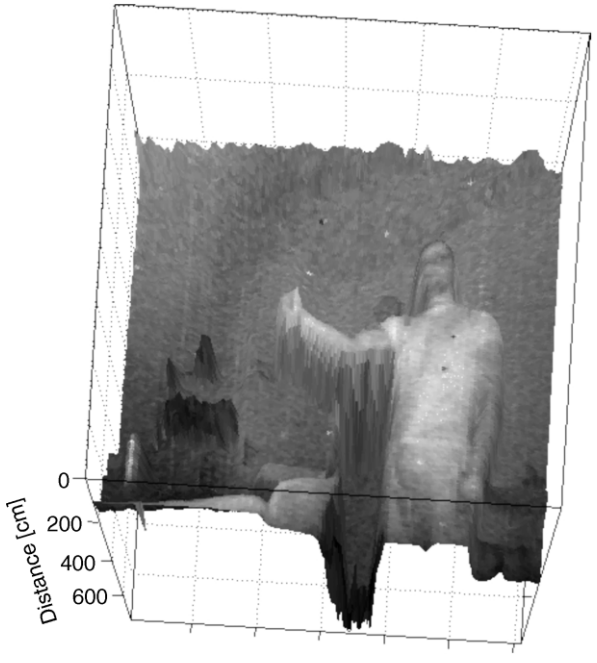


Figure 14 Example of a distance picture (a person) out of a 3D range image sequence, taken with an optical time-of-flight distance camera making use of lock-in pixels. Depending on the available light level, the background illumination and the modulation frequency, such solid-state range cameras can have a resolution of a few mm.

Seeing Chips

Already, in the mid-1980s, researchers were excited about the possibility of integrating local signal processing capabilities in each pixel, implementing some functions of biological vision systems. It was envisioned that an increasing amount of bio-inspired functionality would lead to smart image sensors, exhibiting a primitive sense of vision. The idea of a 'seeing chip', with the capability to interpret a scene, was born.

Unfortunately, even today, our understanding of the enormously successful biological vision systems has not yet reached a mature enough state allowing us to implement signal processing functions that approximate any 'image understanding'. Published algorithms are still incapable of interpreting a scene and 'perceiving' its contents. Obviously, no set of vision algorithms is yet known with which seeing chips could be realized. Although progress is slower than envisioned, single-chip machine vision systems of growing complexity and increasing functionality are being developed.

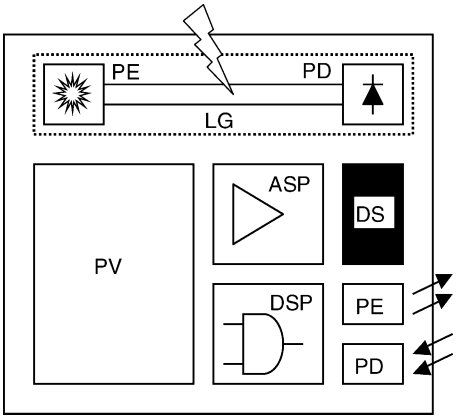


Figure 15 Schematic block diagram of a monolithically integrated photonic microsystem, that is expected to be producable in the near future with organic semiconductors. All optoelectronic devices are fabricated on the same planar surface: photovoltaic power supplies (PV); analog and digital signal processing (ASP and DSP); photodetectors (PD); photoemitters (PE) – used for bidirectional communication or as parts of a photonic detector subsystem – a light guiding structure (LG) that can be sensitized for specific sensing applications, and a display subsystem (DS).

Organic Semiconductors for Monolithic Photonic Microsystems

The realization that organic materials with conjugated double bonds, in particular noncrystalline polymers, are direct bandgap semiconductors, has recently led to a flurry of R&D activities in the domain of organic photonics. Organic light-emitting diodes (OLEDs), color displays, lighting panels, photodiodes, image sensors, photovoltaic cells, RF transponders, and tags, and several other organic photonic devices have not only been demonstrated but many are already commercially available. The special appeal of organic semiconductors stems from their potential for monolithic, low-cost integration of complete, large-area photonic microsystems, even on thin, flexible substrates. In Figure 15, such a monolithic photonic microsystem is illustrated schematically, combining a photovoltaic power supply, a photonic detector subsystem, analog and digital signal processing, as well as bidirectional optical communications with the outside world, possibly including also an on-chip display. The photonic detector subsystem typically consists of a light source, a light guiding structure that is sensitized – for example chemically – and a photodetector/image sensor subsection. In such a monolithic photonic microsystem, the term 'smart pixel array' obtains a much broader meaning, opening up a huge number of practical applications in essentially all domains of our daily and professional lives. This domain of organic semiconductors for integrated photonic microsystems is still largely unexplored, and it is offering exciting opportunities for research and product development.

Smart Pixel Arrays Beyond Electromagnetic Field Imaging

The sensing capabilities of smart pixel arrays are not restricted to electromagnetic radiation; this technology can also be used for the acquisition of other two- or three-dimensional physical fields. As an example, electrostatic imaging can be used for the temporally and spatially resolved acquisition of surface potential distributions in living cells. Electrostatic potential values of several hundred mV can be acquired with a precision of 1 mV and a temporal resolution below 1 millisecond. Other acquisition modalities include magnetostatic field imaging, for example by employing the Hall effect, or imaging of local electrostatic charge or current density distributions.

Summary

The relentless advances of semiconductor technology make it possible to integrate an increasing amount of photosensitive, analog, and digital functionality in the same smart pixel and on the same smart pixel array. Unusual photosensor devices can be fabricated with commercially available

semiconductor processes, implementing innovative designs and layouts, as well as novel electronic circuits, with which the physical limits of photosensing and photocharge processing may be reached. Novel opto-electronic microsystems can be developed, based on application-specific smart pixel arrays, either by realizing known optical measurement techniques more efficiently or by implementing new techniques, whose realization requires the availability of nonconventional image sensing functionality.

Already, smart pixel arrays are opening up a large number of new opportunities in photonics research and product development. The continuously increasing levels of integration lead to new forms of system-on-chip solutions for a wide range of practical applications. Smart pixel arrays can perform at the basic limits imposed by physics, and they have become indispensable tools in almost any type of research laboratory and any field of science and engineering.

List of Units and Nomenclature

Bandwidth	B	s^{-1}
Boltzmann's constant	k	J/K
Capacitance	C	F
Charge	Q	C
Charge carrier mobility	μ	$m^2/(V\,s)$
Charge noise variance	s_Q	C^2
Diffusivity	D	m^2/s
Distance	L	m
Dynamic range	D/R	dB
Field effect transistor transconductance	g	A/V
Integrated power measured with shaped photodiode	r	W
Intensity of electromagnetic wave	S, A, B	W/m^2
One-dimensional light intensity distribution	$p(x)$	W/m^2
One-dimensional width distribution	$w(x)$	m
Phase delay	φ	–
Pixel value at position (i,j) in an image	$P_{i,j}; R_{i,j}$	–
Temperature	T	K
Temporal frequency	ω	s^{-1}
Time	t	s
Two-dimensional convolution kernel	$w_{k,l}$	–

See also

Electromagnetically Induced Transparency. **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguides. **Semiconductor Materials:** Group IV Semiconductors, Si/SiGe.

Further Reading

Forrest SR (2000) Active optoelectronics using thin-film organic semiconductors. *IEEE Journal on Selected Topics in Quantum Electronics* 6: 1072–1083.

Fossum ER (1997) CMOS image sensors: electronic camera-on-a-chip. *IEEE Transactions on Electron Devices* 44: 1689–1698.

Hosticka BJ, Seitz P and Simoni A (2004) Optical time-of-flight range imaging for solid-state 3D vision. In: Grimes G and Dickey EC (eds) *Encyclopedia of Sensors*. Stevenson Ranch, CA: American Scientific Publishers (in print).

Koch C (1989) Seeing chips: analog VLSI circuits for computer vision. *Neural Computation* 1: 184–200.

Lutz G (1999) *Semiconductor Radiation Detectors*. Berlin: Springer.

Mead C (1989) *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: John Wiley and Sons.

Seitz P (1999) Solid-state image sensing. In: Jaehne B (ed.) *Handbook of Computer Vision and Applications*. London: Academic Press.

Seitz P and Blanc N (2003) CCD and APS/CMOS technology for smart pixels and image sensors. *Proceedings of SPIE* 5251 (in print).

Sze SM (1994) *Semiconductor Sensors*. New York: John Wiley and Sons.

Theuwissen AJP (1995) *Solid State Imaging with Charge Coupled Devices*. Dordrecht, Germany: Kluwer.

DIFFRACTION

Contents

Diffraction Gratings
Fraunhofer Diffraction
Fresnel Diffraction

Diffraction Gratings

J Turunen and T Vallius, University of Joensuu, Joensuu, Finland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The classic textbook example of a diffraction grating is a periodic arrangement of parallel opaque wires: if a plane wave is incident upon such a structure, it is divided into a multitude of plane waves propagating in well-defined directions. In particular, the dramatic wavelength-dependence of these directions has attracted great interest ever since the potential of diffraction gratings in spectroscopy was realized in the late 1700s. Today spectroscopy is only one, though important, application of diffraction gratings. Gratings of many different forms belong to the basic building blocks in modern wave-optics-based design of optical elements and systems.

Grating Equations

Some examples of the wide variety of possible grating structures are illustrated in Figure 1, which shows the profiles of some commonly encountered diffraction gratings that are invariant in the y direction. These are known as linear gratings. Each grating may be of reflection type (the refractive index n_{III} is complex) or of transmission type (n_{III} is real). In all cases n_1 and n_2 can both be real or one of them can be complex. Here we assume that the incident field is a plane wave with wave vector $\boldsymbol{k}$ in the x–z plane; more general illumination waves can be represented as superpositions of plane waves.

Figure 1a illustrates a binary grating: in the modulated region $0 < z < H$ the refractive index alternates between n_1 and n_2, changing from one value to another at certain transition points within the period d. The structure in Figure 1b represents a sinusoidal surface-relief grating as an example of smooth grating profiles. Finally, Figure 1c shows a triangular grating profile.

The periodicity of a grating, illuminated by a plane wave, usually leads to the appearance of a set of reflected and transmitted plane waves (diffraction orders of the grating) that propagate in directions determined by the grating period d, the illumination wavelength λ, and the refractive indices n_{I} and n_{III} (see Figure 2a). These directions are obtained from the transmission grating equation:

$$n_{\mathrm{III}} \sin\theta_m = n_{\mathrm{I}} \sin\theta + m\lambda/d \qquad [1]$$

where m is the index of the diffraction order and θ_m is its propagation angle. For reflected orders (superscript r) one simply replaces n_{III} by n_{I} in eqn [1].

Obviously only a limited number of orders satisfy the condition $|\theta_m| < \pi/2$ and thereby represent conventional propagating plane waves. Higher, so-called evanescent orders are inhomogeneous waves that propagate along the grating surface and decay exponentially in the z direction. However, they cannot be ignored in grating analysis, as we shall see.

In the case of a biperiodic grating with a period $d_x \times d_y$ the propagation direction of the incident beam is defined by two angles θ and ϕ illustrated in Figure 2b. The grating equations now read as:

$$n_{\mathrm{III}} \sin\theta_{mn} \cos\phi_{mn} = n_{\mathrm{I}} \sin\theta \cos\phi + m\,\lambda/d_x \qquad [2]$$

$$n_{\mathrm{III}} \sin\theta_{mn} \sin\phi_{mn} = n_{\mathrm{I}} \sin\theta \sin\phi + n\lambda/d_y, \qquad [3]$$

where θ_{mn} and ϕ_{mn} define the propagation directions of the diffraction order with indices m and n. For reflected orders n_{III} is again replaced by n_{I}. These expressions apply also to the special case of linear gratings (limit $d_y \to \infty$) illuminated by a plane

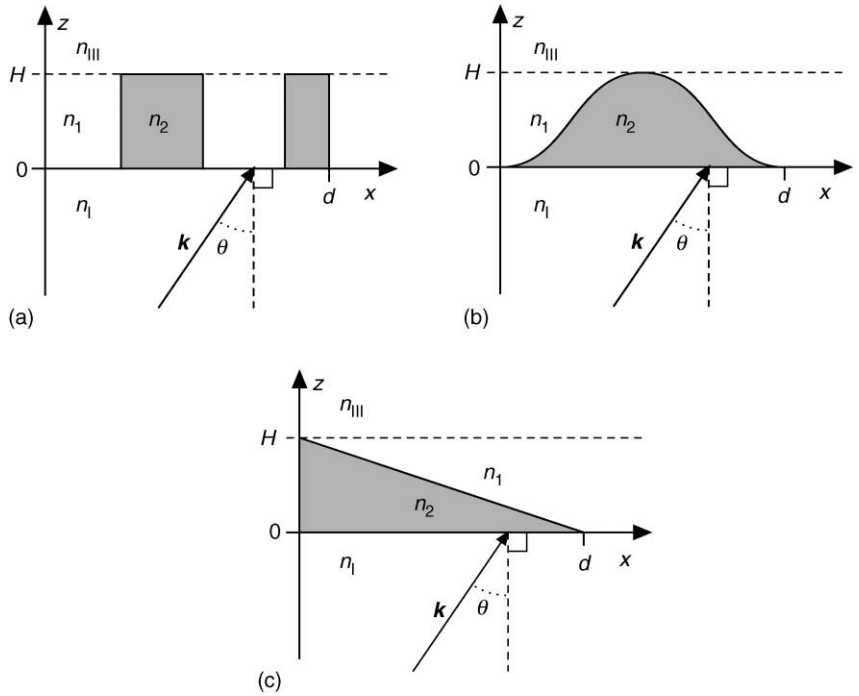


Figure 1 Examples of linear gratings. (a) Binary grating; (b) sinusoidal grating; (c) triangular grating. Here d is the grating period, H is the thickness of the modulated region, θ is the angle of incidence, and n denotes refractive index.

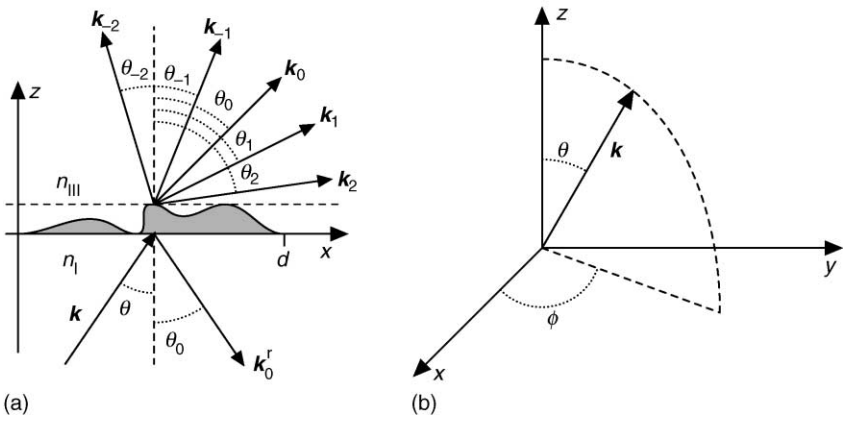


Figure 2 Diffraction by gratings. (a) Linear grating; (b) definitions of direction angles in the case of a biperiodic grating.

wave not propagating in the x–z plane (conical incidence).

Some two-dimensionally periodic gratings are illustrated in Figure 3. The structure illustrated in Figure 3a is an extension of a binary grating to the biperiodic geometry: here a single two-dimensional feature is placed within the grating period but, of course, more features with different outlines could be included as well. Moreover, each period could contain a continuous 'mountain-range' or consist of a 'New-York'-style pixel structure. Figure 3b illustrates a structure consisting of an array of rectangular-shaped holes pierced in a flat screen, while Figure 3c defines an array of isolated particles on a flat substrate. If the grid or the particles are metallic, surrounded by dielectrics from all sides, one speaks of inductive and capacitive structures, respectively: in the former case relatively free-flowing currents are induced in the metallic grid structure, but in the latter case the isolated metal particles become polarized (and therefore capacitive) when illuminated by an electromagnetic field.

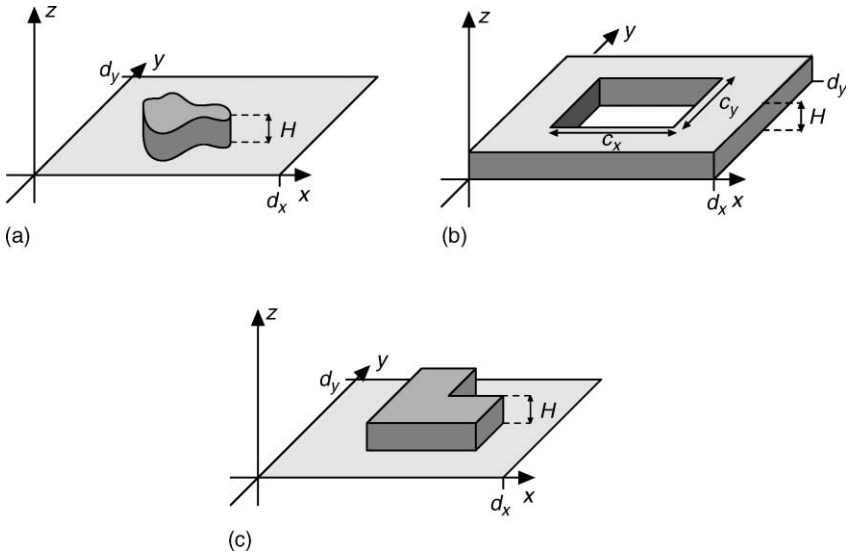


Figure 3 Examples of biperiodic gratings. (a) General binary grating; (b) inductive structure; (c) capacitive structure.

Grating Theory

Gratings equations provide exact knowledge of the propagation directions of the diffraction orders, but they give no information about the distribution of light among different orders. To obtain such information, one must in general solve Maxwell's equations exactly (rigorous diffraction theory), but within certain limits simpler methods are available. In all cases, however, the transmitted and reflected fields are represented as superpositions of plane waves that propagate in the directions allowed by the grating equations (these are known as Rayleigh expansions).

The most popular method for rigorous grating analysis is the Fourier Modal Method (FMM), though many alternative analysis methods exist. To understand the principle of FMM, consider first binary gratings. The field inside the modulated region $0 < z < H$ is represented as a superposition of waveguide modes, determined by imagining that the modulated region would extend from $x = -\infty$ to $x = \infty$. With given transition points the electric permittivity, its inverse, and the electric field inside the grating, can be represented in the form of Fourier series and one ends up with a matrix eigenvalue problem with finite dimensions when the matrices are truncated. A numerical solution of this problem provides a set of orthogonal modes that can propagate in the structure with certain propagation constants. The relative weights of the modes are then determined by applying the electromagnetic boundary conditions at the planes $z = 0$ and $z = H$ to match the modal expansions inside the grating with the plane-wave (Rayleigh) expansions of the fields in the homogeneous materials on both sides of the grating. This matching procedure, which is numerically implemented by solving a set of simultaneous linear equations, yields the complex amplitudes of the electric and magnetic field components associated with the reflected and transmitted diffraction orders of the grating.

Of main interest are the diffraction efficiencies, η_m (or η_{mn} for biperiodic gratings) of the orders, which tell the distribution of energy between different propagating orders. In electromagnetic theory the diffraction efficiency is defined as the z-component of the time-averaged Poynting vector associated with the particular order, divided by that of the incident plane wave.

Gratings with z-dependent surface profiles can be handled similarly if they are 'sliced' into a sufficient number of layers, in each of which the grating is modeled as a binary structure. The electromagnetic boundary conditions are applied at each slice boundary and the approach models the real continuous structure with arbitrary degree of accuracy when the number of the slices is increased; in practice it is usually sufficient to use approximately 20–30 slices for sufficiently accurate modeling of any continuous grating profile by FMM.

The main problem with rigorous techniques such as FMM is bad scaling. If, in the case of 1D modulated (or y-invariant) gratings, the period is increased by a factor of two, the computation time required to solve the eigenvalue problem is increased by a factor of around eight. Thus we have a d^3 type

scaling law. The situation is much worse for bi-periodic gratings, for which the exponent is around six. Thus no foreseeable developments in computer technology will solve the problem, and approximate analysis methods have to be sought. A multitude of such methods are under active development, but in this context we discuss the simplest example only.

If $d \gg \lambda$ and the grating profile does not contain wavelength-scale features, one can describe the effects of a grating on the incident field merely by considering it as a complex-amplitude-modulating object, which imposes a position-dependent phase shift and amplitude transmission (or reflection) coefficient on the incident field. This model is known as the thin-element approximation (TEA).

At normal incidence the plane-wave expansion of any scalar field component U immediately behind the grating (at $z = H$) takes the form:

$$U(x, y, H) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} T_{mn} \times \exp[\mathrm{i}2\pi(mx/d_x + ny/d_y)] \quad [4]$$

where

$$T_{mn} = \frac{1}{d_x d_y} \int_0^{d_x} \int_0^{d_y} U(x, y, H) \times \exp[-\mathrm{i}2\pi(mx/d_x + ny/d_y)]\mathrm{d}x\,\mathrm{d}y \quad [5]$$

are the complex amplitudes of the transmitted diffraction orders. In TEA the field at $z = H$ may be connected to the field at $z = 0$ (a unit-amplitude plane wave is assumed here) by optical path integration along a straight line though the modulated region of the grating:

$$U(x, y, H) = \int_0^H \exp\left[\mathrm{i}\frac{2\pi}{\lambda}\hat{n}(x, y, z)\right]\mathrm{d}z \quad [6]$$

where $\hat{n}(x, y, z) = n(x, y, z) + \mathrm{i}\kappa(x, y, z)$ is the complex refractive-index distribution in the region $0 < z < H$. With TEA the diffraction efficiencies can be calculated from a simple formula:

$$\eta_{mn} = |T_{mn}|^2 \quad [7]$$

which, however, ignores Fresnel reflections.

Equations [5]–[7] allow one to obtain closed-form expressions for the diffraction efficiencies η_{mn} in many grating geometries. For example, a y-invariant triangular profile with $H = \lambda/|n_\mathrm{I} - n_\mathrm{III}|$ gives $\eta_{-1} = 100\%$ if $n_1 = n_\mathrm{III}$, $n_2 = n_\mathrm{I}$, and $\theta = 0$. For a Q-level stair-step-quantized version of this grating we obtain:

$$\eta_m = \sin c^2(1/Q) \quad [8]$$

where $\sin cx = \sin(\pi x)/(\pi x)$. It should be noted, however, that these results are valid only for gratings with $d \gg \lambda$, even if they are corrected by Fresnel transmission losses at the interface between the two media.

Fabrication of Gratings

Gratings can be manufactured in a number of different ways, perhaps the simplest being to plot an equidistant array of straight black and white lines and to reduce the scale of the pattern photographically. This method can be readily adapted to the generation of more complex grating structures and diffractive elements, but it is not suitable for the production of high-quality gratings required in, for example, spectroscopy.

Ruling engines represent the traditional way of making gratings with periods of the order of the wavelength of light, especially for spectroscopic applications. These are machines equipped with a triangularly shaped diamond ruling edge and an interferometrically controlled two-dimensional mechanical translation stage that moves the substrate with respect to the ruling edge. Triangular-profile gratings of high quality can be produced with such machines, but the fabrication process is slow and the cost of the manufactured gratings is therefore high. In addition, even with real-time interferometric control of the ruling-edge position, it is impossible to avoid small local errors in the grating geometry. Such errors are particularly harmful if they are periodic since, according to the grating equation, these errors will generate spurious spectral lines known as ghosts. Such lines generated by the grating 'super-period' are weak but they may be confused with real spectral lines in the source spectrum. If the ruling errors are random, they result in 'pedestal-type' background noise. While such errors do not introduce ghosts, they nevertheless reduce the visibility of weak original spectral lines.

The introduction of lasers in the early 1960s made it possible to form stable and high-contrast interference patterns over a sufficient time-scale to allow the exposure of photographic films and other photosensitive materials such as polymeric photoresists. This interferometric technique offers a direct method for the fabrication of large diffraction gratings without appreciable periodic errors. With this technique it is possible to make highly regular

gratings for spectroscopy, but it does not permit easy generation of arbitrarily shaped grating structures. The two intersecting plane waves can also be generated with a linear grating using diffraction orders $m = +1$ and $m = -1$, suppressing the zero order by appropriate design of the grating profile and choosing the period in such a way that all higher transmitted orders are evanescent. This so-called phase mask technique is suitable, for example, for the exposure of Bragg gratings in fibers.

To generate arbitrary gratings one usually employs lithographic techniques. Some of the possible processes are illustrated in Figure 4. The exposure can be performed with scanning photon, electron, or ion beams, etc. The process starts with substrate preparation, i.e., coating it with a beam-sensitive layer known as resist, and in the case of electron beam exposure with a thin conductive layer to prevent charging. Following the exposure the resist is developed chemically, and in this process either the exposed or nonexposed parts of the resist are dissolved (depending on whether the resist is of positive or negative type), leading to a surface-relief profile. Several alternative fabrication steps may follow the development, depending on the desired grating type. If a high-contrast resist is used, a slightly 'undercut' profile as illustrated in Figure 4c is obtained. This can be coated with a metal layer and the resist can be 'lifted off' chemically. Thus a binary metallic grating is formed. To produce a binary dielectric profile one bombards the substrate with ions either purely mechanically (ion beam milling) or with chemical assistance (reactive ion etching), and finally the residual metal is dissolved.

Alternatively, the particle beam dose can be varied with position to obtain an analog profile in the resist after development. This resist profile can be transformed into the substrate by proportional etching, which leads to an analog dielectric surface-relief structure (see Figure 4f). This profile can be made

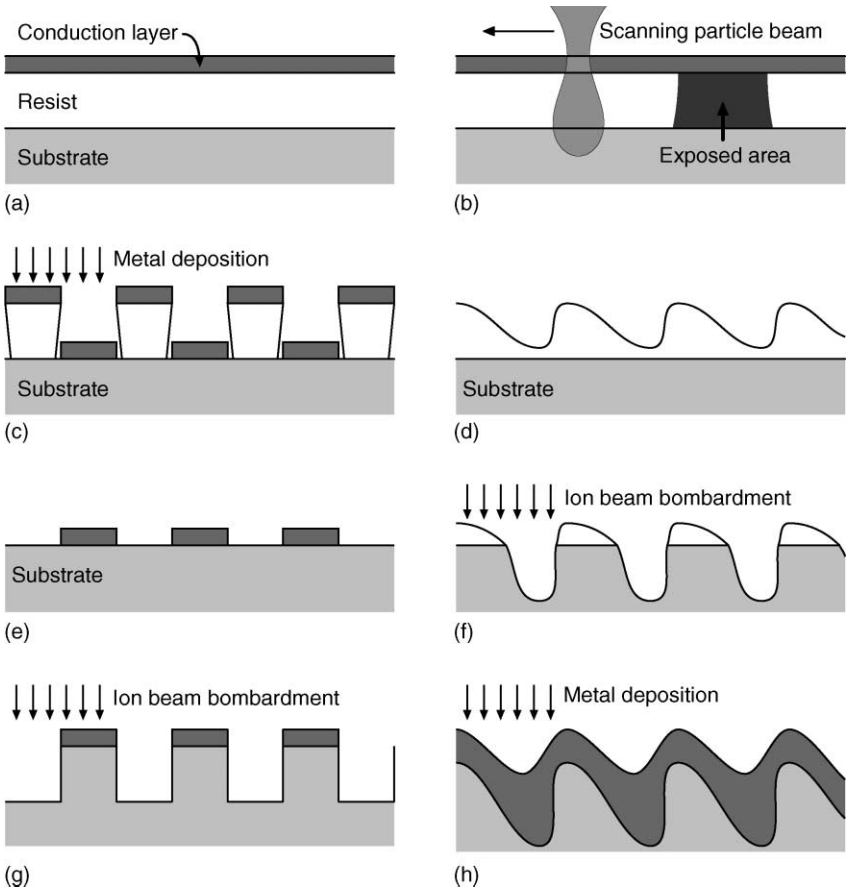


Figure 4 Lithographic processes for grating fabrication. (a) Substrate after preparation for electron beam exposure; (b) exposure by a scanning electron beam; (c) slightly undercut profile in hight-contrast resist after development, under metal film evaporation; (d) analog resist profile after variable-dose exposure and resist development; (e) finished binary metallic grating after resist lift-off; (f) transfer of the analog surface-relief profile into the substrate by proportional etching; (g) etching of a binary profile into the substrate with a metal mask; (h) growth of a metal film on an analog surface profile.

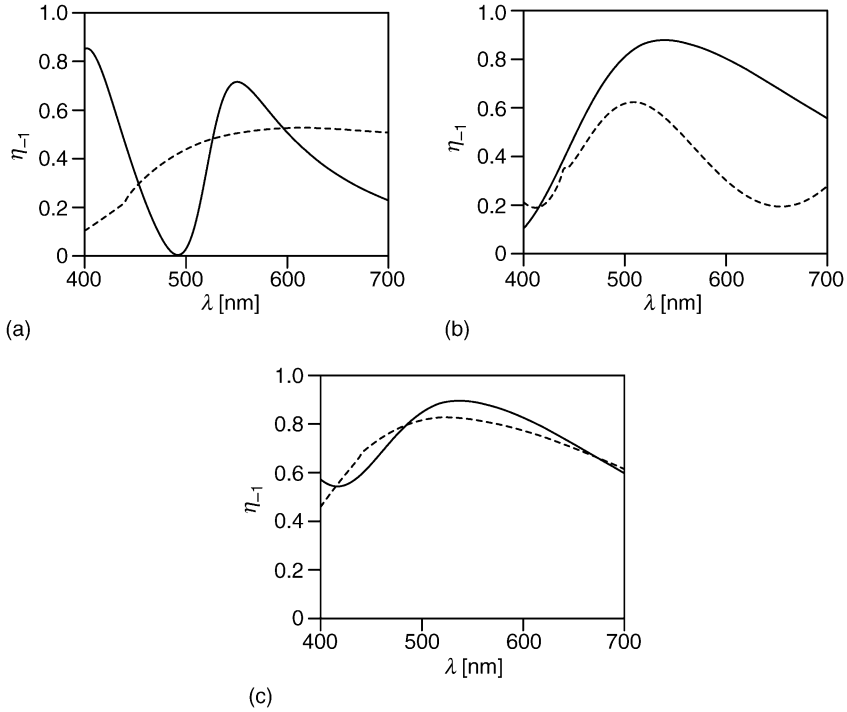


Figure 5 Spectral diffraction efficiency curves of spectroscopic gratings in TE polarization. (a) Binary; (b) sinusoidal; and (c) triangular reflection gratings illuminated at Bragg angle. Solid lines: $d = \lambda_0$. Dashed lines: $d = 2\lambda_0$.

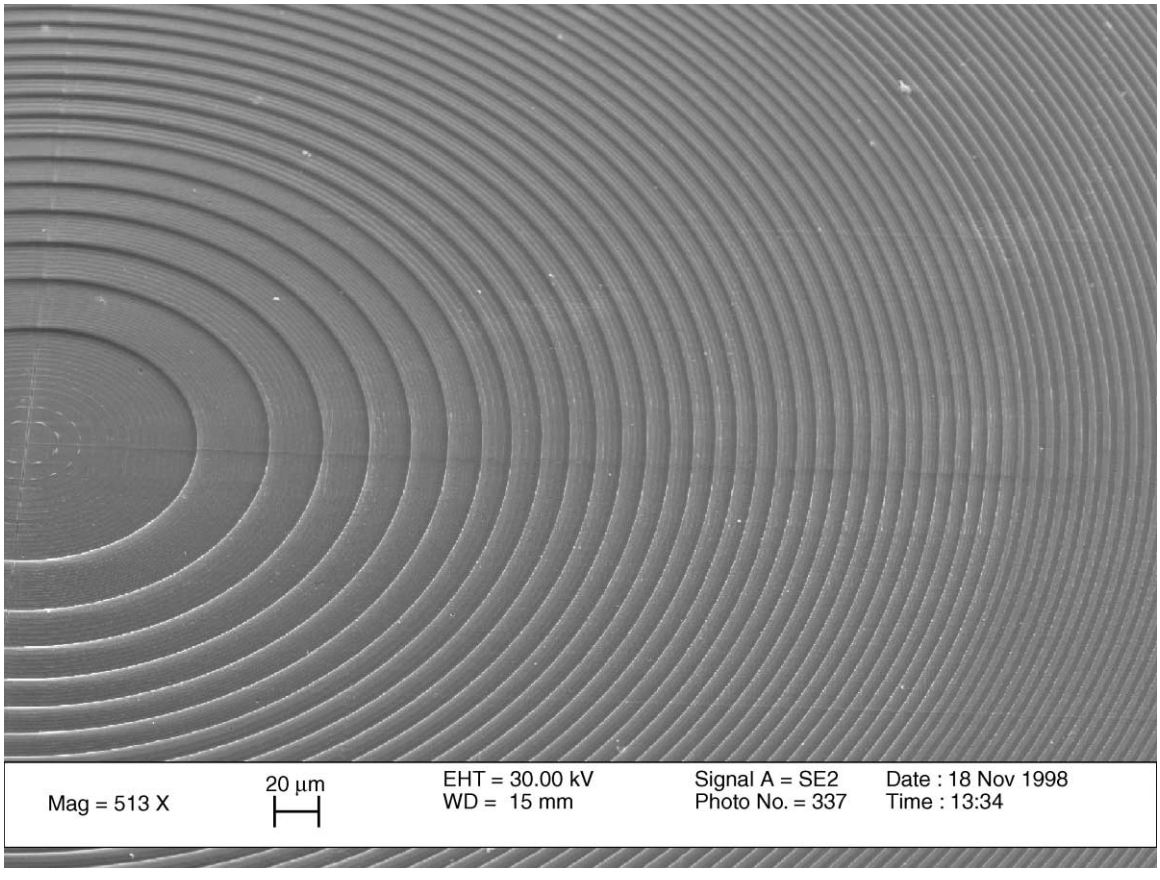


Figure 6 A scanning electron micrograph of a part of a diffractive lens.

reflective by depositing a thin metal film on top of it. If large-scale fabrication of identical gratings is the goal, a thick ($\sim 100\ \mu m$) metal layer (for example nickel) can be grown on the surface by electroplating. After separation from the substrate, the electroplated metal film, which is a negative of the original profile, can be used as a stamp in replicating the original profile in plastics using methods such as hot embossing, ultraviolet radiation curing, and injection molding. In particular, the latter method provides the key to high-volume and low-cost production of various types of diffraction gratings and other grating-based diffractive structures in the industrial scale.

Some Grating Applications

We now proceed to describe a collection of representative grating applications. More exhaustive coverage of grating applications can be found in the Further Reading section at the end of this article.

Spectroscopic Gratings

The purpose of a spectroscopic grating is to divide the incident light into a spectrum, i.e., to direct different frequency components of the incident beam into different directions given by the grating (eqn [1]). When the grating period is reduced, the angular dispersion $\partial\theta_m/\partial\lambda$, and therefore the resolution of the spectrograph, increases.

In the numerical examples presented in Figure 5 we assume a reflection grating in a Littrow mount, i.e., $\theta_m^r = -\theta$. This mount is common in spectroscopy; it obviously requires that the grating is rotated when the spectrum is scanned. We see from eqn [1] that in the Littrow mount:

$$\sin\theta = \frac{m\lambda}{2d} \qquad [9]$$

By inserting this result back into eqn [1] and differentiating we obtain a quantitative measure for dispersion in the Littrow mount:

$$\frac{\partial\theta_m}{\partial\lambda} = -\frac{m}{2d\cos\theta} \qquad [10]$$

We consider TE polarization (electric field is linearly polarized in the y-direction) and order $m = -1$, optimize H to give maximum efficiency at $\lambda_0 = 550$ nm, and plot the diffraction efficiency η_{-1} over the visible wavelength range for two values of the grating period, $d = \lambda_0$ and $d = 2\lambda_0$. Rigorous diffraction theory is used with $n_I = n_2 = 1$, and the metal is assumed to be aluminum (n_1 and n_{III} are complex-valued).

In Figure 5a binary gratings with one groove (width $d/2$) per period is considered. Parametric optimization gives $H = 422$ nm if $d = 1\lambda$, and $H = 168$ nm if $d = 2\lambda$. Figure 5b illustrates the results for sinusoidal gratings with $H = 345$ nm if $d = 1\lambda$, and

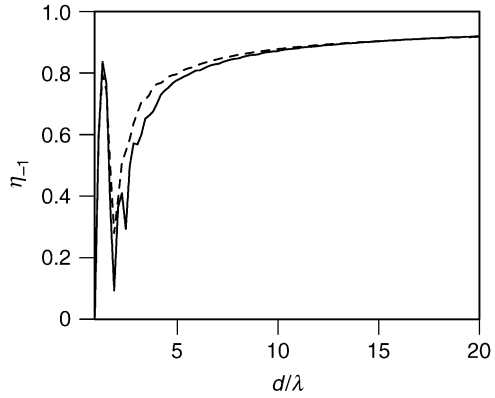


Figure 7 Diffraction efficiency versus wavelength for triangular transmission gratings illuminated at normal incidence. Solid lines: TE polarization. Dashed lines: TM polarization.

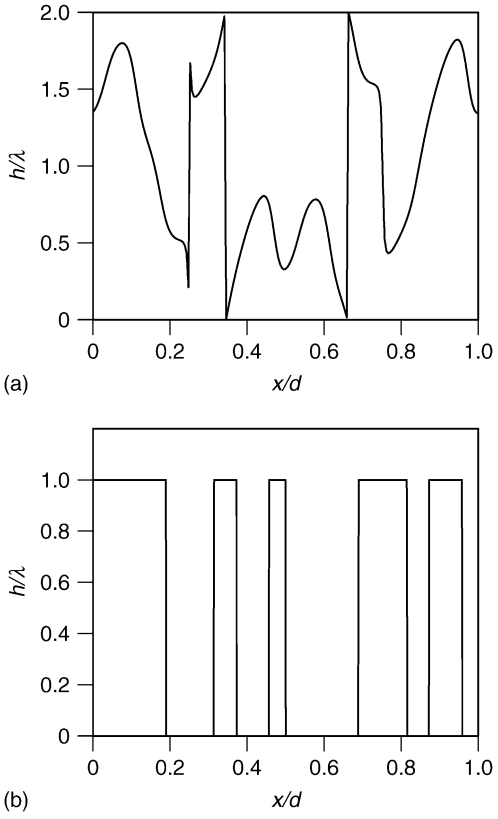


Figure 8 Surface-relief profiles of (a) continuous and (b) binary $1 \rightarrow 8$ beamsplitter gratings.

$H = 520$ nm if $d = 2\lambda$. Finally, triangular gratings with $H = 431$ nm if $d = 1\lambda$, and $H = 316$ nm if $d = 2\lambda$, are considered. The triangular grating profile is seen to be the best choice in general.

Diffractive Lenses

Diffractive lenses are gratings with locally varying period and groove orientation. Figure 6 illustrates the central part of a diffractive microlens fabricated by lithographic methods. It is seen to consist of concentric zones with radially decreasing width.

The phase transmission function of an ideal diffractive focusing lens can be obtained by a simple optical path calculation, and is given by

$$\phi(r) = \frac{2\pi}{\lambda_0}(f - \sqrt{f^2 + r^2}) \approx -\frac{\pi}{\lambda_0 f} r^2 \qquad [11]$$

where f is the focal length, r is the radial distance from the optical axis, and the approximate expression is valid in the paraxial region $r \ll f$. Here λ_0 is the design wavelength, which determines the modulation depth of the locally blazed grating profile:

$$H = \frac{\lambda_0}{n(\lambda_0) - 1} \qquad [12]$$

and $n(\lambda_0)$ is the refractive index of the dielectric material at $\lambda = \lambda_0$.

The zone boundaries r_n are defined by the condition $\phi(r_n) = -2\pi m$, and therefore we obtain from eqn [11]:

$$r_n = \sqrt{2nf\lambda_0 + (n\lambda_0)^2} \approx \sqrt{2nf\lambda_0} \qquad [13]$$

Obviously, when the numerical aperture of the lens is large, the zone width in the outer regions of the lens is reduced to values of the order of λ. Figure 7 illustrates the local diffraction efficiency in different part of a diffractive lens: we use the order $m = -1$ of a locally triangular diffraction grating of the

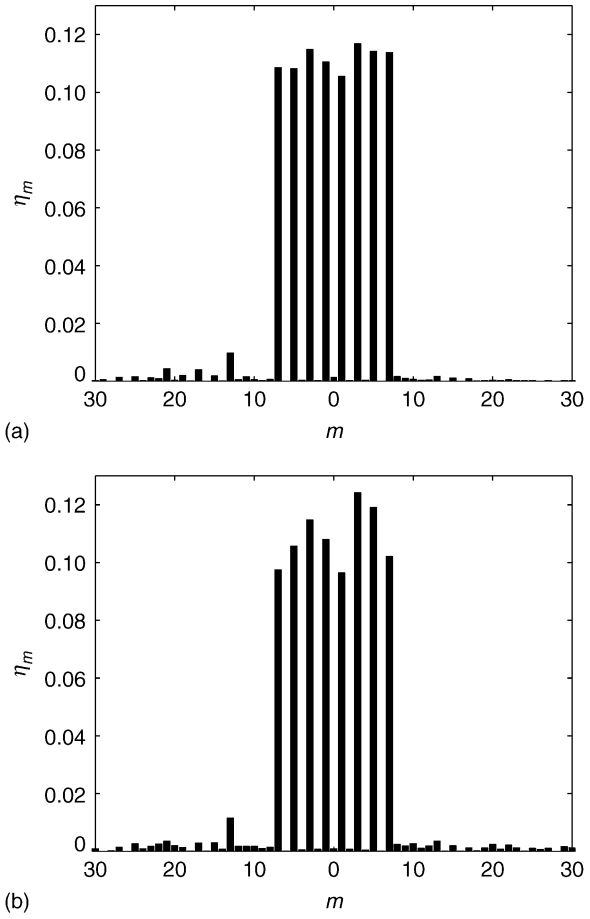


Figure 9 Diffraction pattern produced by a continuous-profile beamsplitter grating when (a) $d = 200\lambda$ and (b) $d = 50\lambda$.

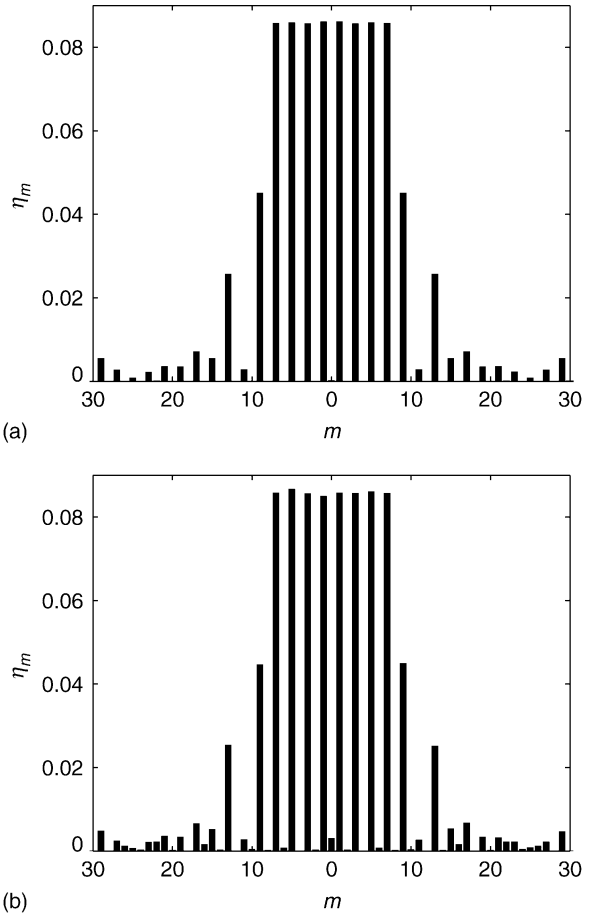


Figure 10 Diffraction pattern produced by a binary beamsplitter grating grating when (a) $d = 200\lambda$ and (b) $d = 50\lambda$.

type shown in Figure 1c with $\theta = 0$, assuming that $n_{\mathrm{I}} = n_2 = 1.5$ (fused silica) and $n_1 = n_{\mathrm{III}} = 1$. The efficiency η_{-1} is calculated as a function of λ using rigorous diffraction theory for both TE and TM polarization (the electric vector points in the groove direction in the TE case and is perpendicular to it in the TM case). It is seen that the local efficiency of the lens becomes poor in regions where the local grating period is small, and thus diffractive lenses operate best in the paraxial region (it is, however, possible to obtain improved non-paraxial performance by optimizing the local grating profile).

A factor that limits the use of diffractive lenses with polychromatic light is their large chromatic aberration: in the paraxial limit the focal length depends on the wavelength according to the formula:

$$f(\lambda) = \frac{\lambda_0}{\lambda} f \qquad [14]$$

However, since the dispersion of a grating has an opposite sign compared to the diffraction of a prism, a weak positive diffractive lens can be combined with a relatively strong positive refractive lens to form a single-element achromatic lens. Such a hybrid lens is substantially thinner and lighter than a conventional all-refractive achromatic lens made of crown and flint glasses.

Beam Splitter Gratings

Multiple beamsplitters, also known as array illuminators, are gratings with sophisticated periodic structure that are capable of transforming an incident plane wave into a set of diffraction orders with a specified distribution of diffraction efficiencies. Most often one wishes to obtain a one- or two-dimensional array of diffracted plane waves with equal efficiency. This goal can be achieved by appropriate design of the grating profile, which may be of binary, multilevel, or continuous form.

Figure 8 illustrates binary and continuous grating profiles capable of dividing one incident plane wave into eight diffracted plane waves (central odd-numbered orders of the grating) with equal efficiency in view of TEA. Figure 9 shows the distribution of energy among the different diffraction orders in and around the array produced by the continuous profile, evaluated by rigorous theory for two different grating periods, while Figure 10 gives corresponding results for the binary grating.

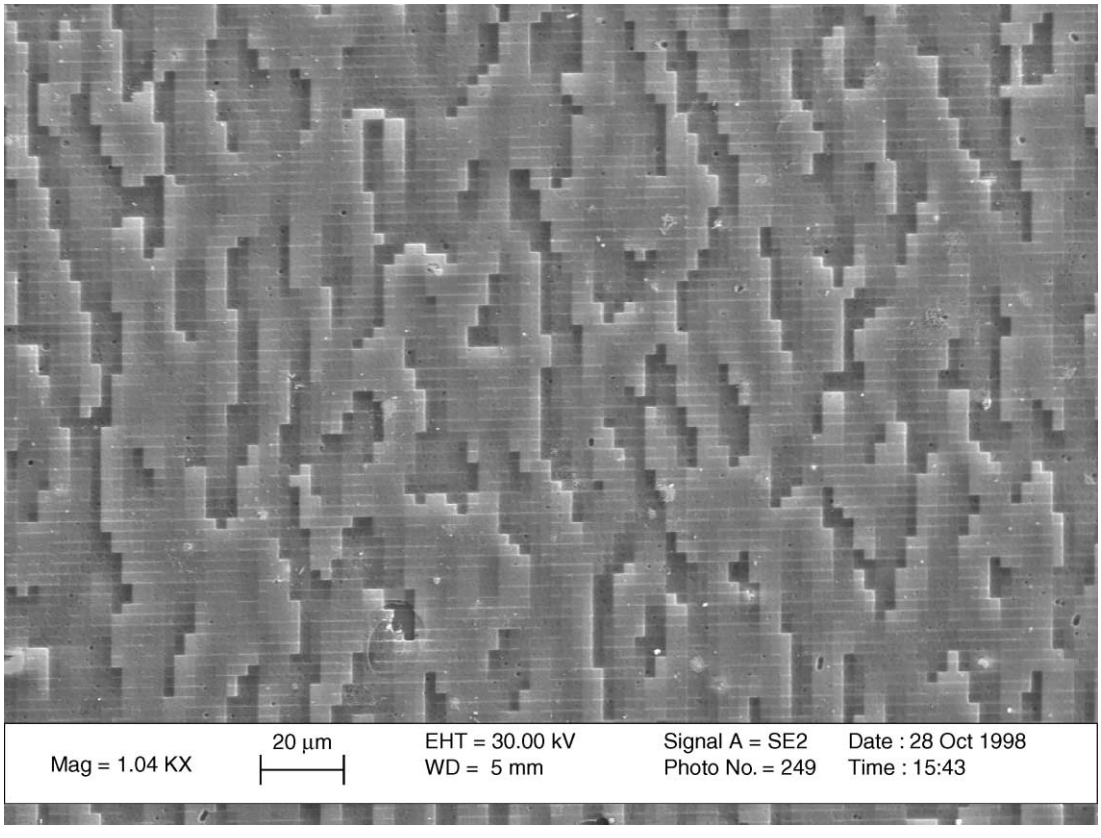


Figure 11 Scanning electron micrograph of a sophisticated beamsplitter grating.

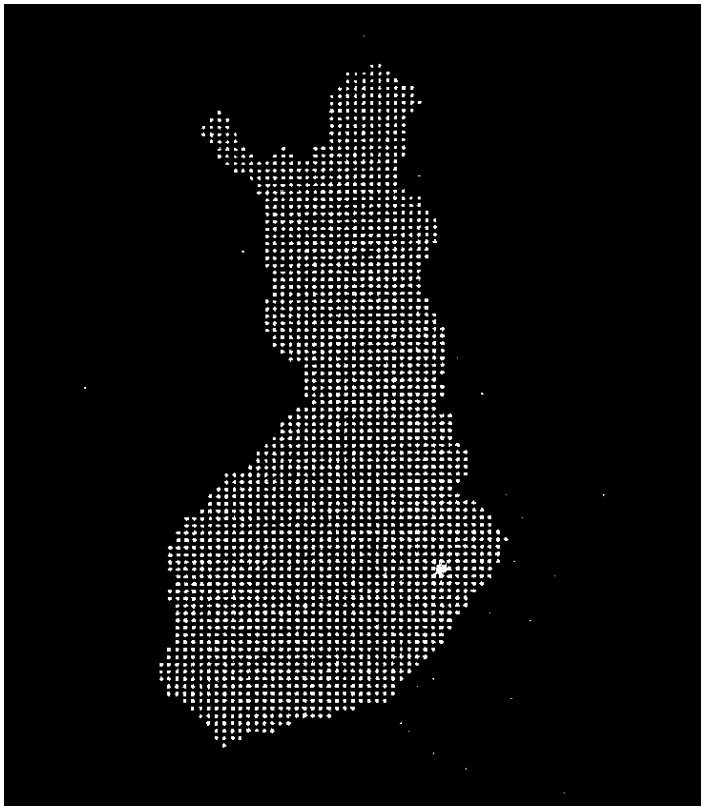


Figure 12 The far-field diffraction pattern produced by the grating in Figure 11.

The fraction of incident energy that ends up in the desired eight orders is 72% for the binary grating and 96% for the continuous grating according to TEA, making the latter more attractive from a theoretical viewpoint. However, the continuous profile is more difficult to fabricate accurately than the binary one, especially for small values of the ratio d/λ. Moreover, the array uniformity produced by the continuous profile degrades faster than that of the binary profile when the period is reduced.

An example of the structure of a biperiodic grating that is capable of forming a large, shaped array of diffraction orders with equal efficiency is presented in Figure 11. Figure 12 shows the far-field diffraction pattern produced by this element: the shape of Finland with the location of the town of Joensuu highlighted by placing the zero diffraction order on the appropriate position on the map (the apparent nonuniformity in the array is mostly due to the finite resolution of the CCD detector).

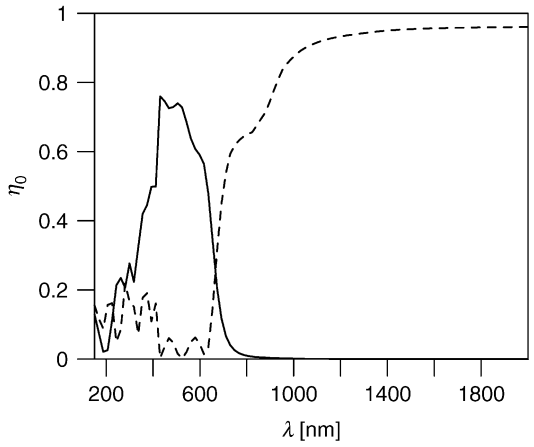


Figure 13 Zero order spectral transmission (solid line) and reflection (dashed line) of an aluminium inductive grid filter with 400 nm period.

Inductive Grid Filters

As a final example we consider a scaled-down version of a device well known from the door of a microwave oven: one can see in through the holes pierced in the metallic screen in the oven door, but the microwaves are trapped inside because their wavelengths are large in comparison with the dimensions of the holes. Another example of these inductive grid filters is found in large radio telescope mirrors for centimeter-scale

wavelengths, which are wide-grid structures rather than smoothly curved metal surfaces.

Figure 13 illustrates the spectral transmission and reflection curves of an inductive grid filter of the type illustrated in Figure 3b. We assume that $d_x = d_y = d = 400$ nm, $c_x = c_y = c = 310$ nm, $H = 500$ nm, and that the grid is a self-supporting structure ($n_{\mathrm{I}} = n_{\mathrm{III}} = 1$) made of aluminum. Only the zero reflected and transmitted orders propagate when $\lambda > d$, but other orders (with rather low efficiency) appear for smaller wavelengths. We see that the filter reflects infrared radiation effectively, but transmits a substantial part of radiation at visible and shorter wavelength regions.

List of Units and Nomenclature

Complex field amplitude	U
Diffraction angles [rad]	θ, ϕ
Diffraction efficiency	η
Focal length [m]	f, f_0
Grating period [m]	d
Grating thickness [m]	H
Number of quantization levels	Q
Phase function [rad]	$\phi(r)$
Refractive index	n
Wavelength [m]	λ, λ_0
Wave vector [m^{-1}]	$\mathbf{k}$

See also

Fiber Gratings.

Further Reading

Chandezon J, Maystre D and Raoult G (1980) A new theoretical method for diffraction gratings and its numerical application. *Journal of Optics* 11: 235–241.

Gaylord TK and Moharam MG (1985) Analysis and applications of optical diffraction by gratings. *Proceedings of IEEE* 73: 894–937.

Herzig HP (ed.) (1997) *Micro-optics: Elements, Systems and Applications*. London: Taylor & Francis.

Hutley MC (1982) *Diffraction Gratings*. London: Academic Press.

Li L (2001) Mathematical reflections on the fourier modal method in grating theory. In: Bao G, Cowsar L and Masters W (eds) *Mathematical Modelling in Optical Science*, pp. 111–139. Philadelphia: SIAM.

Li L, Chandezon J, Granet G and Plumey J-P (1999) Rigorous and efficient grating-analysis method made easy for optical engineers. *Applied Optics* 38: 304–313.

Petit R (ed.) (1980) *Electromagnetic Theory of Gratings*. Berlin: Springer-Verlag.

Rayleigh JWS (1907) On the dynamical theory of gratings. *Proceeding of the Royal Society A* 79: 399.

Solymar L and Cooke DJ (1981) *Volume Holography and Volume Gratings*. London: Academic Press.

Turunen J, Kuittinen M and Wyrowski F (2000) Diffractive optics: electromagnetic approach. In: Wolf E (ed.) *Progress in Optics XL*. Amsterdam: Elsevier.

Turunen J and Wyrowski F (eds) (1997) *Industrial and Commercial Applications of Diffractive Optics*. Berlin: Akademie Verlag.

van Renesse RL (1993) *Optical Document Security*. Boston, MA: Artech House.

Fraunhofer Diffraction

B D Guenther, Duke University, Durham, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Geometrical optics does a reasonable job of describing the propagation of light in free space unless that light encounters an obstacle. Geometrical optics predicts that the obstacle would cast a sharp shadow. On one side of the shadow's edge would be a bright uniform light distribution, due to the incident light, and on the other side there would be darkness. Close examination of an actual shadow edge reveals, however, dark fringes in the bright region and bright fringes in the dark region. This departure from the predictions of geometrical optics was given the name diffraction by Grimaldi.

There is no substantive difference between diffraction and interference. The separation between the two subjects is historical in origin and is retained for pedagogical reasons. Since diffraction and interference are actually the same physical process, we expect the observation of diffraction to be a strong function of the coherence of the illumination. With incoherent sources, geometrical optics is usually all that is needed to predict the performance of an optical system and diffraction can be ignored except at dimensions on the scale of a wavelength. With light sources having a large degree of spatial coherence, it is impossible to neglect diffraction in any analysis.

Even though diffraction produced by most common light sources and objects is a small effect, it is possible to observe diffraction without special equipment. By viewing a light source through a slit formed by two fingers that are nearly touching, fringes can be observed. Diffraction is used by photographers and television cameramen to provide artistic highlights in an image containing small light sources. They accomplish this by placing a screen in front of the camera lens. Light from point sources in the field of view is diffracted by the screen and produces 'stars' in the image. A third place where diffraction effects are easily observed is when a mercury or sodium streetlight, a few hundred meters away, is viewed through a sheer curtain.

Fresnel originally developed an approximate scalar theory based on Huygens' principle which states that:

> Each point on a wavefront can be treated as a source of a spherical wavelet called a secondary wavelet or a Huygens' wavelet. The envelope of these wavelets, at some later time, is constructed by finding the tangent to the wavelets. The envelope is assumed to be the new position of the wavefront.

Rather than simply using the wavelets to construct an envelope, Fresnel's scalar theory assumes that the Huygens' wavelets interfere to produce a new wavefront (*see* **Diffraction:** Fresnel Diffraction).

A rigorous diffraction theory is based on Maxwell's equations and uses the boundary conditions associated with the obstacle to calculate a field scattered by the obstacle. The origins of this scattered field are currents induced in the obstacle by the incident field. The scattered field is allowed to interfere with the incident field to produce a resultant diffracted field. The application of the rigorous theory is very difficult and for most problems an approximate scalar theory developed by Kirchhoff and Sommerfeld is used.

Kirchhoff Theory of Diffraction

Kirchhoff's theory was based on the elastic theory of light but can be reformulated into a vector theory. Here we will limit our discussion to the scalar formulation. Given an incident wave, φ, we wish to calculate the optical wave at point P_0, in Figure 1, located at $\mathbf{r}_0$, in terms of the wave's value on a surface, S, that we construct about the observation point.

To obtain a solution to the wave equation at the point, P_0, we select as a Green's function the one proposed by Kirchhoff (Box 1), a unit amplitude spherical wave, expanding about the point at $\mathbf{r}_0$, denoted by Ψ, and then apply Green's theorem (see Box 2). We will run into boundary condition problems as we proceed with the derivation of the solution. Sommerfeld removed the problems by assuming a slightly more complex Green's function (see Box 1). We will limit our discussion to the Kirchhoff Green's function.

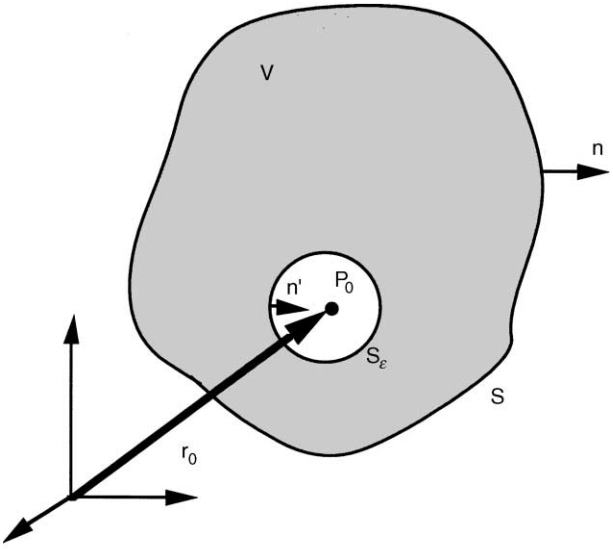


Figure 1 Region of integration for solution of Kirchhoff's diffraction integral. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Green's theorem requires that there be no sources (singularities) inside the surface S but our Green's function [1] is based on a source at r_0. We eliminate the source by constructing a small spherical surface S_ε, of radius ε, about $\mathbf{r}_0$, excluding the singularity at $\mathbf{r}_0$ from the volume of interest, shown in gray in Figure 1. The surface integration that must be performed in Green's theorem is over the surface $S' = S + S_\varepsilon$.

Within the volume enclosed by S′, the Green's function, Ψ, and the incident wave, φ, satisfy the scalar Helmholtz equation so that the volume integral can be written as

$$\iiint_V (\Psi\nabla^2\varphi - \varphi\nabla^2\Psi)\mathrm{d}\nu = -\iiint_V (\Psi\varphi k^2 - \varphi\Psi k^2)\mathrm{d}\nu \qquad [7]$$

The right side of [7] is identically equal to zero. This fact allows us to use [6] to produce a simplified statement of Green's theorem:

$$\iint_{S'} \left(\Psi\frac{\partial\varphi}{\partial n} - \varphi\frac{\partial\Psi}{\partial n}\right)\mathrm{d}s = 0 \qquad [8]$$

Box 1

Green's Function

1. **Kirchhoff Green's function**: Assume that the wave, Ψ, is due to a single point source at $\mathbf{r}_0$. At an observation point, $\mathbf{r}_1$, the Green's function is given by

$$\Psi(\mathbf{r}_1) = \frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}, \qquad [1]$$

   where $r_{01} = |\mathbf{r}_{01}| = |\mathbf{r}_1 - \mathbf{r}_0|$ is the distance from the source of the Green's function, at $\mathbf{r}_0$, to the observation position, at $\mathbf{r}_1$.
2. **Sommerfeld Green's function**: Assume that there are two point sources, one at P_0 and one at P_0' (see Figure 2 where the source, P_0', is positioned to be the mirror image of the source at P_0 on the opposite side of the screen, thus

$$r_{01} = r_{01}' \qquad \cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) = -\cos(\hat{\mathbf{n}}, \mathbf{r}_{01}') \qquad [2]$$

   A Green's function that could be used is

$$\Psi' = \frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}} - \frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{01}'}}{r_{01}'} \qquad [3a]$$

   or

$$\Psi' = \frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}} + \frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{01}'}}{r_{01}'} \qquad [3b]$$

Box 2

Green's Theorem

To calculate the complex amplitude, $\tilde{E}(\mathbf{r})$, of a wave at an observation point defined by the vector $\mathbf{r}$, we need to use a mathematical relation known as Green's Theorem. Green's theorem states that, if φ and Ψ are two scalar functions that are well behaved, then

$$\iint_S (\varphi \nabla\Psi\cdot\hat{\mathbf{n}} - \Psi\nabla\varphi\cdot\hat{\mathbf{n}})\mathrm{d}s = \iiint_V (\varphi\nabla^2\Psi - \Psi\nabla^2\varphi)\mathrm{d}\nu \qquad [4]$$

The vector identities

$$\nabla\Psi\cdot\hat{\mathbf{n}} = \frac{\partial\Psi}{\partial n} \qquad \nabla\varphi\cdot\hat{\mathbf{n}} = \frac{\partial\varphi}{\partial n} \qquad [5]$$

allow Green's theorem to be written as

$$\iint_S \left[\varphi\frac{\partial\Psi}{\partial n} - \Psi\frac{\partial\varphi}{\partial n}\right]\mathrm{d}s = \iiint_V [\varphi\nabla^2\Psi - \Psi\nabla^2\varphi]\mathrm{d}\nu \qquad [6]$$

This equation is the prime foundation of scalar diffraction theory but only the proper choice of Ψ, φ and the surface, S, allows direct application to the diffraction problem.

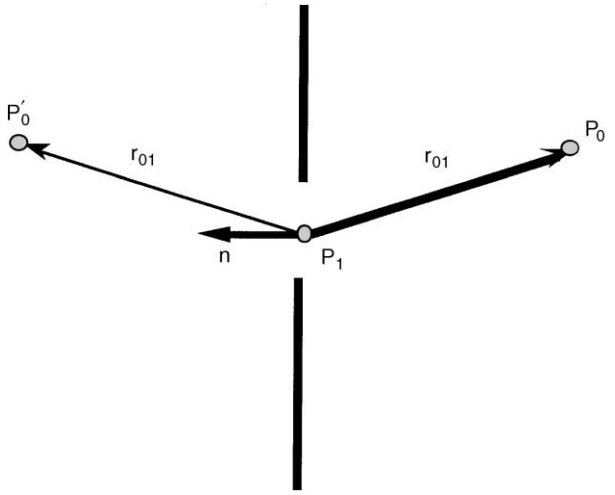


Figure 2 Geometry for the Sommerfeld Green's function. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Because the integral over the combined surfaces, S and S_ε, is equal to zero, the integral over the surface S must equal the negative of the integral over the surface, S_ε:

$$-\iint_{S_\varepsilon}\left(\Psi\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial\Psi}{\partial n}\right)\mathrm{d}s=\iint_{S}\left(\Psi\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial\Psi}{\partial n}\right)\mathrm{d}s \quad [9]$$

To perform the surface integrals, we must evaluate the Green's function on the surface, S′. We therefore select $\mathbf{r}_1$, the vector defining the observation point, to be on either of the two surfaces that make up S′. The derivative, $\partial\Psi/\partial n$, to be evaluated is

$$\frac{\partial\Psi}{\partial n}=\frac{\partial\Psi}{\partial r}\frac{\partial r}{\partial n}$$
$$=\left(-\mathrm{i}k\frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}-\frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}^2}\right)\cos(\mathrm{n},\mathbf{r}_{01}) \quad [10]$$

where $\cos(\hat{\mathbf{n}},\mathbf{r}_{01})$ is the cosine of the angle between the outward normal $\hat{\mathbf{n}}$ and the vector $\mathbf{r}_{01}$, the vector between points P_0 and P_1. (Note from Figure 1 that the outward normal on S_ε is inward, toward P_0, while the normal on S is outward, away from P_0.)

For example if $\mathbf{r}_1$ were on S_ε, then

$$\cos(\mathbf{n},\mathbf{r}_{01})=-1, \quad [11]$$

$$\frac{\partial\Psi(\mathbf{r}_1)}{\partial n}=\left[\frac{1}{\varepsilon}+\mathrm{i}k\right]\frac{\mathrm{e}^{-\mathrm{i}k\varepsilon}}{\varepsilon}, \quad [12]$$

where ε is the radius of the small sphere we constructed around the point at P_0.

We now use [10] to rewrite the two integrals in [9]. The integral over the surface S_ε is

$$\iint_{S_\varepsilon}\left(\Psi\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial\Psi}{\partial n}\right)\mathrm{d}s$$
$$=\iint_{S_\varepsilon}\left[\frac{\partial\varphi}{\partial n}\frac{\mathrm{e}^{-\mathrm{i}k\varepsilon}}{\varepsilon}-\varphi\frac{\mathrm{e}^{-\mathrm{i}k\varepsilon}}{\varepsilon}\left(\frac{1}{\varepsilon}+\mathrm{i}k\right)\right]$$
$$\times\varepsilon^2\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \quad [13]$$

while the integral over the surface S is

$$\iint_{S}\left[\frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial}{\partial n}\left(\frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}\right)\right]r^2\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \quad [14]$$

The omitted volume contained within the surface S_ε is allowed to shrink to zero by taking the limit as $\varepsilon\to 0$. Equation [14] will not be affected by taking the limit. The first and last terms of the right side of [13] go to zero as $\varepsilon\to 0$ because they contain $\varepsilon\mathrm{e}^{-\mathrm{i}k\varepsilon}$. The second term in [13] contains $\mathrm{e}^{-\mathrm{i}k\varepsilon}$ which goes to 1 as $\varepsilon\to 0$. Therefore, in the limit as $\varepsilon\to 0$, [9] becomes

$$\varphi(\mathbf{r}_0)=\frac{1}{4\pi}\iint_{S}\left[\frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial}{\partial n}\left(\frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}\right)\right]\mathrm{d}s \quad [15]$$

which is sometimes called the integral theorem of Kirchhoff. By carefully selecting the surface of integration in [15], we can calculate the diffraction field observed at P_0 produced by an aperture in an infinite opaque screen.

Assume that a source at P_2 in Figure 3 produces a spherical wave that is incident on an infinite, opaque screen from the left. To find the field at P_0 in Figure 3, we apply [15] to the surface $S_1+S_2+\Sigma$, where S_1 is a plane surface adjacent to the screen, Σ is that portion of S_1 in the aperture and S_2 is a large spherical surface, of radius R, centered on P_0, see Figure 3. The first question to address is how to calculate the contribution from S_2 or better yet how do we show we can ignore contributions from S_2. As R increases, Ψ and φ will decrease as $1/R$. However, the area of integration increases as R^2 so the $1/R$ fall-off is not a sufficient reason for neglecting the contribution of the integration over S_2.

On the surface S_2, the Green's function and its derivative are

$$\Psi=\frac{\mathrm{e}^{-\mathrm{i}kR}}{R} \quad [16]$$

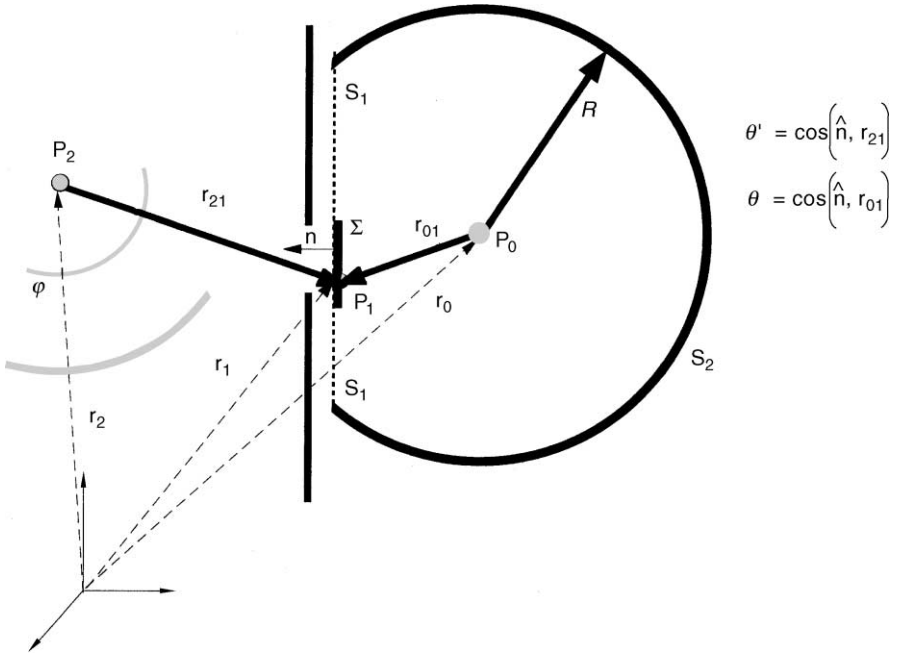


Figure 3 Integration surface for diffraction calculation. P_2 is the source and P_0 is the point where the field is to be calculated. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

$$\frac{\partial \Psi}{\partial n} = \left(-ik - \frac{1}{R} \right) \frac{e^{-ikR}}{R} \approx -ik\Psi \qquad [17]$$

where the approximate equality is for large R. Thus, the integral over S_2 for large R is

$$\int\int_{S_2} \left[\Psi \frac{\partial \varphi}{\partial n} + \varphi(ik\Psi) \right] ds$$

$$= \int\int_{S_2} \Psi \left(\frac{\partial \varphi}{\partial n} + ik\varphi \right) R^2 \sin\theta \, d\theta \, d\phi \qquad [18]$$

The quantity $R\Psi = e^{-ik \cdot R}$ is finite as $R \to \infty$ so for the integral to vanish, we must have

$$\lim_{R \to \infty} R \left(\frac{\partial \varphi}{\partial n} + ik\varphi \right) = 0 \qquad [19]$$

This requirement is called the Sommerfeld radiation condition and is satisfied if $\varphi \to 0$ as fast as $1/R$ (for example, a spherical wave). Since the illumination of the screen will be a spherical wave, or at least a linear combination of spherical waves, we should expect the integral over S_2 to make no contribution (it is exactly zero).

Another way to insure that surface S_2 makes no contribution is to assume that the light turns on at time t_0. At a later time, t, when we desire to know the field at P_0, the radius of S_2 is $R > c\ (t - t_0)$ which physically means that the light has not had time to reach S_2. In this situation, there can be no contribution from S_2. This is not a perfect solution because when the wave is of finite duration it can no longer be considered monochromatic.

We should expect the major contribution in the integral over S_1 to come from the portion of the surface in the aperture, Σ. We make the following assumptions about the incident wave, φ (these assumptions are known as St. Venaut's hypothesis or Kirchhoff's boundary conditions)

1. We assume that, in the aperture, φ and $\partial\varphi/\partial n$ have the values they would have if the screen were not in place.
2. On the portion of S_1 not in the aperture, φ and $\partial\varphi/\partial n$ are identically zero.

The Kirchhoff boundary conditions allow the screen to be neglected, reducing the problem to an integration only over the aperture. It is surprising that the assumptions about the contribution of the surface S_2 and the screen yield very accurate results (as long as polarization is not important, the aperture is large with respect to the wavelength, and we don't get too close to the aperture).

Mathematically the Kirchhoff boundary conditions are incorrect. The two Kirchhoff boundary conditions imply that the field is zero everywhere behind the screen, except in the aperture, which makes Ψ and $\partial\Psi/\partial n$ discontinuous on the boundary of the aperture. Another problem with the boundary conditions is

that Ψ and $\partial\Psi/\partial n$ are known only if the problem has already been solved.

The Sommerfeld Green's functions will remove the inconsistencies of the Kirchhoff theory. If we use the Green's function [3a] then Ψ vanishes in the aperture while $\partial\Psi/\partial n$ can assume the value required by Kirchhoff's boundary conditions. If we use the Green's functions [3b] then the converse holds, i.e. $\partial\Psi/\partial n$ is zero in the aperture. This improved Green's function makes the problem harder with little gain in accuracy so we will retain the Kirchhoff formalism.

As a result of the above discussion, the surface integral [15] is only performed over the surface, Σ, in the aperture. The evaluation of the Green's function on this surface can be simplified by noting that normally $r_{01} \gg \lambda$, i.e., $k \gg 1/r_{01}$, thus on Σ,

$$\begin{aligned}\frac{\partial\Psi}{\partial n} &= \cos(\mathbf{n}, \mathbf{r}_{01})\left(-ik - \frac{1}{r_{01}}\right)\frac{e^{-ik\cdot r_{01}}}{r_{01}} \\ &\approx -ik\cos(\hat{\mathbf{n}}, \mathbf{r}_{01})\frac{e^{-ik\cdot r_{01}}}{r_{01}}\end{aligned} \quad [20]$$

Substituting this approximate evaluation of the derivative into [15] yields

$$\varphi(\mathbf{r}_0) = \frac{1}{4\pi}\iint_{\Sigma}\frac{e^{-ik\cdot r_{01}}}{r_{01}}\left[\frac{\partial\varphi}{\partial n} + ik\varphi\cos(\hat{\mathbf{n}}, \mathbf{r}_{01})\right]ds \quad [21]$$

The source of the incident wave is a point source located at P_2, with a position vector, $\mathbf{r}_2$, measured with respect to the coordinate system and a distance $|\mathbf{r}_{21}|$ away from P_1, a point in the aperture (see Figure 3). The incident wave is therefore a spherical wave of the form

$$\varphi(\mathbf{r}_{21}) = A\frac{e^{-ik\cdot r_{21}}}{r_{21}} \quad [22]$$

which fills the aperture. Here also we will assume that $r_{21} \gg \lambda$ so that the derivative of the incident wave assumes the same form as [20]. Then [21] can be written

$$\varphi(\mathbf{r}_0) = \frac{iA}{\lambda}\iint_{\Sigma}\frac{e^{-ik\cdot(r_{21}+r_{01})}}{r_{21}r_{01}} \times\left[\frac{\cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_{21})}{2}\right]ds \quad [23]$$

This relationship is called the Fresnel–Kirchhoff diffraction formula. It is symmetric with respect to r_{01} and r_{21}, making the problem identical when the source and measurement point are interchanged.

A physical understanding of [23] can be obtained if it is rewritten as

$$\varphi(\mathbf{r}_0) = \iint_{\Sigma}\Phi(\mathbf{r}_{21})\frac{e^{-ik\cdot r_{01}}}{r_{01}}ds \quad [24]$$

The field at P_0 is due to the sum of an infinite number of secondary Huygens sources in the aperture Σ. The secondary sources are point sources radiating spherical waves of the form

$$\Phi(\mathbf{r}_{21})\frac{e^{-ik\cdot r_{01}}}{r_{01}} \quad [25]$$

with amplitude $\Phi(r_{21})$, defined by

$$\Phi(\mathbf{r}_{21}) = \frac{i}{\lambda}\left[A\frac{e^{-ik\cdot r_{21}}}{r_{21}}\right]\left[\frac{\cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_{21})}{2}\right] \quad [26]$$

The imaginary constant, i, causes the wavelets from each of these secondary sources to be phase shifted with respect to the incident wave. The obliquity factor, in the amplitude [26]

$$\tfrac{1}{2}[\cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_{21})] \quad [27]$$

causes the secondary sources to have a forward directed radiation pattern.

If we had used the Sommerfeld Green's function, the only modification to our result would be a change in the obliquity factor to $\cos(\mathbf{n},\mathbf{r}_{01})$. In our discussions below we will assume that the angles are all small, allowing the obliquity factor to be replaced by 1 so that in the remaining discussion the choice of Green's function has no impact.

We will assume the source of light is at infinity, $z_1 = \infty$ in Figure 3, so that the aperture is illuminated by a plane wave, traveling parallel to the z-axis. With this assumption

$$\theta' = \cos(\mathrm{n}, \mathbf{r}_{21}) \approx -1 \quad [28]$$

We will also make the paraxial approximation which assumes that the viewing position is close to the z-axis, leading to

$$\theta = \cos(\mathbf{n}, \mathbf{r}_{01}) \approx 1 \quad [29]$$

With these assumptions eqn [24] becomes identical to the Huygens–Fresnel integral discussed in the section on Fresnel diffraction:

$$\tilde{E}(\mathbf{r}_0) = \frac{i}{\lambda}\int\int_{\Sigma}\frac{\tilde{E}_i(\mathbf{r})}{R}e^{-ik\cdot R}\,ds \quad [30]$$

The modern interpretation of the Fresnel–Kirchhoff (or now in the small-angle approximation, the Fresnel–Huygens) integral is to view it as a convolution integral. By considering free space to be a linear system, the result of propagation can be calculated by convolving the incident (input) wave with the impulse response of free space (our Green's function),

$$\frac{i}{\lambda}\frac{e^{-ik\cdot R}}{R} \qquad [31]$$

The job of calculating diffraction has only begun with the derivation of the Fresnel–Kirchhoff integral. In general, an analytic expression for the integral cannot be found because of the difficulty of performing the integration over R. There are two approximations that will allow us to obtain analytic expressions of the Fresnel–Kirchhoff integral. In both approximations, all dimensions are assumed to be large with respect to the wavelength. In one approximation, the viewing position is assumed to be far from the obstruction; the resulting diffraction is called Fraunhofer diffraction and will be discussed here. The second approximation, which leads to Fresnel diffraction, assumes that the observation point is nearer the obstruction, to be quantified below. This approximation, which is the more difficult mathematically, is discussed in the article on **Diffraction:** Fresnel Diffraction.

Fraunhofer Approximation

In Fraunhofer diffraction, we require that the source of light and the observation point, P_0, be far from the aperture so that the incident and diffracted wave can be approximated by plane waves. A consequence of this requirement is that the entire waveform passing through the aperture contributes to the observed diffraction.

The geometry to be used in this derivation is shown in Figure 4. The wave incident normal to the aperture, Σ, is a plane wave and the objective of the calculation is to find the departure of the transmitted wave from its geometrical optical path. The calculation will provide the light distribution, transmitted by the aperture, as a function of the angle the light is deflected from the incident direction. We assume that diffraction makes only a small perturbation on the predictions of geometrical optics. The deflection angles encountered in this derivation are, therefore, small, as assumed above, and we will be able to use the paraxial approximation.

The distance from a point P, in the aperture, to the observation point P_0, of Figure 4, is

$$R^2 = (x-\xi)^2 + (y-\eta)^2 + Z^2 \qquad [32]$$

From Figure 4 we see R_0 is the distance from the center of the screen to the observation point, P_0,

$$R_0^2 = \xi^2 + \eta^2 + Z^2 \qquad [33]$$

The difference between these two vectors is

$$\begin{aligned} R_0^2 - R^2 &= (R_0 - R)(R_0 + R) \\ &= \xi^2 + \eta^2 + Z^2 - Z^2 - (x^2 - 2x\xi + \xi^2) \\ &\quad - (y^2 - 2y\eta + \eta^2) \\ &= 2(x\xi + y\eta) - (x^2 + y^2) \end{aligned} \qquad [34]$$

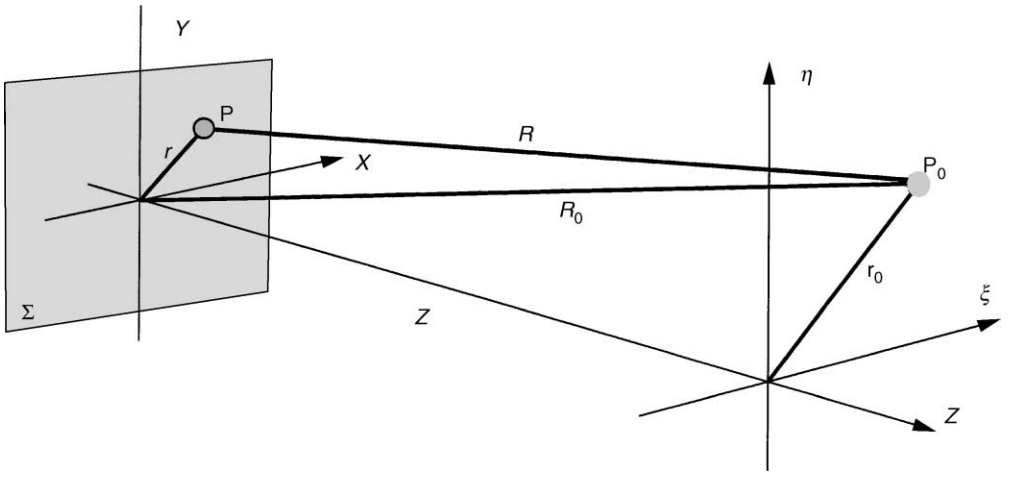


Figure 4 Geometry for Fraunhofer diffraction. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The equation for the position of point P in the aperture can be written in terms of [34]:

$$r = R_0 - R = \frac{R_0^2 - R^2}{R_0 + R},$$
$$= [2(x\xi + y\eta) - (x^2 + y^2)]\frac{1}{R_0 + R} \qquad [35]$$

The reciprocal of $(R_0 + R)$ can be written as

$$\frac{1}{R_0 + R} = \frac{1}{2R_0 + R - R_0} = \frac{1}{2R_0}\left(1 + \frac{R - R_0}{2R_0}\right)^{-1} \qquad [36]$$

Now using [36]

$$R_0 - R = \left[\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right]\left[1 - \frac{R_0 - R}{2R_0}\right]^{-1}$$
$$= \left[\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right]\left[1 - \frac{r}{2R_0}\right]^{-1} \qquad [37]$$

If the diffraction integral [30] is to have a finite (nonzero) value, then

$$k|R_0 - R| \ll kR_0 \qquad [38]$$

This requirement insures that all the Huygens's wavelets, produced over the aperture from the center out to the position r, will have similar phases and will interfere to produce a nonzero amplitude at P_0. This requirement results in

$$\frac{1}{1 - \dfrac{r}{2R_0}} \approx 1 \qquad [39]$$

From this equation we are tempted to state that the aperture dimension must be small relative to the observation distance. However, for the exponent to provide a finite contribution in the integral, it is kr that is small, not the aperture dimension, r.

Using the approximation for $(R_0 - R)$, we obtain the diffraction integral

$$\tilde{E}_P = \frac{i\alpha\, e^{-i\mathbf{k}\cdot\mathbf{R}_0}}{\lambda R_0}\iint_{\Sigma} f(x, y) e^{+ik\left(\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right)}\, dx\, dy \qquad [40]$$

The change in the amplitude of the wave due to the change in R, as we move across the aperture, is neglected, allowing R in the denominator of the Huygens–Fresnel integral to be replaced by R_0 and moved outside of the integral. We have introduced the complex transmission function, $f(x, y)$, of the aperture to allow a very general aperture to be treated. If the aperture function described the variation in absorption of the aperture as a function of position, as would be produced by a photographic negative, then $f(x, y)$ would be a real function. If the aperture function described the variation in transmission of a biological sample, it might be entirely imaginary.

The argument of the exponent in [40] is

$$ik\left(\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right)$$
$$= i2\pi\left(\frac{x\xi + y\eta}{\lambda R_0} - \frac{x^2 + y^2}{2\lambda R_0}\right) \qquad [41]$$

If the observation point, P_0, is far from the screen, we can neglect the second term and treat the phase variation across the aperture as a linear function of position. This is equivalent to assuming that the diffracted wave is a collection of plane waves. Mathematically, the second term in [41] can be neglected if

$$\frac{x^2 + y^2}{2\lambda R_0} \ll 1 \qquad [42]$$

This is called the far-field approximation and the theory yields Fraunhofer diffraction. If the quadratic term of [41] is on the order of 2π then the fraction is

$$\frac{x^2 + y^2}{2\lambda R_0} \approx \Theta(1) \qquad [43]$$

and we must retain the quadratic term and the theory yields Fresnel diffraction.

We define spatial frequencies in the x and y direction by

$$\omega_x = \frac{2\pi}{\lambda}\sin\theta_x = -\frac{2\pi\xi}{\lambda R_0}$$
$$\omega_y = \frac{2\pi}{\lambda}\sin\theta_y = -\frac{2\pi\eta}{\lambda R_0} \qquad [44]$$

We define the spatial frequencies with a negative sign to allow equations involving spatial frequencies to have the same form as those involving temporal frequencies. The negative sign is required because ωt and $\mathbf{k}\cdot\mathbf{r}$ appear in the phase of the wave with opposite signs and we want the Fourier transform in space and in time to have the same form.

With the variables defined in [44], the integral becomes

$$\tilde{E}_P(\omega_x, \omega_y) = \frac{i\alpha\, e^{-i\mathbf{k}\cdot\mathbf{R}_0}}{\lambda R_0}\iint_{\Sigma} f(x, y) e^{-i(\omega_x x + \omega_y y)}\, dx\, dy \qquad [45]$$

The result of our derivation is that the Fraunhofer diffraction field, $\tilde{E}_P$, can be calculated by performing the two-dimensional Fourier transform of the aperture's transmission function.

Definition of Fourier Transform

In this discussion the function $F(\omega)$ is defined as the *Fourier transform* of $f(t)$:

$$\mathcal{F}\{f(t)\} \equiv F(\omega) \equiv \int_{-\infty}^{\infty} f(\tau)\mathrm{e}^{-\mathrm{i}\omega\tau}\,\mathrm{d}\tau \qquad [46]$$

The transformation from a temporal to a frequency representation given by [46] does not destroy information; thus, the inverse transform can also be defined

$$\mathcal{F}^{-1}\{F(\omega)\} = f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}\omega \qquad [47]$$

By assuming that the illumination of an aperture is a plane wave and that the observation position is in the far field, the diffraction of an aperture is found to be given by the Fourier transform of the function describing the amplitude transmission of the aperture. The amplitude transmission of the aperture, $f(x, y)$, may thus be interpreted as the superposition of mutually coherent plane waves.

Diffraction by a Rectangular Aperture

We will now use Fourier transform theory to calculate the Fraunhofer diffraction pattern from a rectangular slit and point out the reciprocal relationship between the size of the diffraction pattern and the size of the aperture.

Consider a rectangular aperture with a transmission function given by

$$f(x, y) = \begin{cases} 1 & |x| \le x_0, |y| \le y_0 \\ 0 & \text{all other } x \text{ and } y \end{cases} \qquad [48]$$

Because the aperture is two-dimensional, we need to apply a two-dimensional Fourier transform but it is not difficult because the amplitude transmission function is separable, in x and y. The diffraction amplitude distribution from the rectangular slit is simply one-dimensional transforms carried out individually for the x and y dimensions:

$$\tilde{E}_P = \frac{\mathrm{i}\alpha}{\lambda R_0}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}\int_{-x_0}^{x_0} f(x)\mathrm{e}^{-\mathrm{i}\omega_x x}\,\mathrm{d}x\int_{-y_0}^{y_0} f(y)\mathrm{e}^{-\mathrm{i}\omega_y y}\,\mathrm{d}y \qquad [49]$$

Since both $f(x)$ and $f(y)$ are defined as symmetric functions, we need only calculate the cosine transforms to obtain the diffracted field:

To calculate the Fourier transform defined by [46] we can rewrite the transform as

$$F(\omega) = \int_{-\infty}^{\infty} f(\tau)\cos\,\omega\tau\,\mathrm{d}\tau - \mathrm{i}\int_{-\infty}^{\infty} f(\tau)\sin\,\omega\tau\,\mathrm{d}\tau \qquad [50]$$

If $f(\tau)$ is a real, even function then the Fourier transform can be obtained by simply calculating the cosine transform:

$$\int_{-\infty}^{\infty} f(\tau)\cos\,\omega\tau\,\mathrm{d}\tau \qquad [51]$$

$$\tilde{E}_P = \mathrm{i}\frac{4x_0y_0\alpha}{\lambda R_0}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}\frac{\sin\,\omega_x x_0}{\omega_x x_0}\frac{\sin\,\omega_y y_0}{\omega_y y_0} \qquad [52]$$

The intensity distribution of the Fraunhofer diffraction produced by the rectangular aperture is

$$I_P = I_0\frac{\sin^2\omega_x x_0}{(\omega_x x_0)^2}\frac{\sin^2\omega_y y_0}{(\omega_y y_0)^2} \qquad [53]$$

The maximum intensities in the x and y directions occur at

$$\omega_x x_0 = \omega_y y_0 = 0 \qquad [54]$$

The area of this rectangular aperture is defined as $A = 4x_0y_0$, resulting in an expression for the maximum intensity of

$$I_0 = \left[\frac{\mathrm{i}4x_0y_0\alpha\,\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}}{\lambda R_0}\right]^2 = \frac{A^2\alpha^2}{\lambda^2 R_0^2} \qquad [55]$$

The minima of [53] occur when $\omega_x x_0 = n\pi$ or when $\omega_y y_0 = m\pi$. The location of the zeros can be specified as a dimension in the observation plane or, using the paraxial approximation, in terms of the angle defined by [44]

$$\begin{aligned} \sin\,\theta_x \approx \theta_x \approx \frac{\xi}{R_0} = \frac{n\lambda}{2x_0} \\ \sin\,\theta_y \approx \theta_y \approx \frac{\eta}{R_0} = \frac{m\lambda}{2y_0} \end{aligned} \qquad [56]$$

The dimensions of the diffraction pattern are characterized by the location of the first zero, i.e., when $n = m = 1$, and are given by the observation plane coordinates, ξ and η. The dimensions of the diffraction pattern are inversely proportional to the dimensions of the aperture. As the aperture dimension expands, the width of the diffraction pattern decreases until, in the limit of an infinitely wide aperture, the diffraction pattern becomes a delta function.

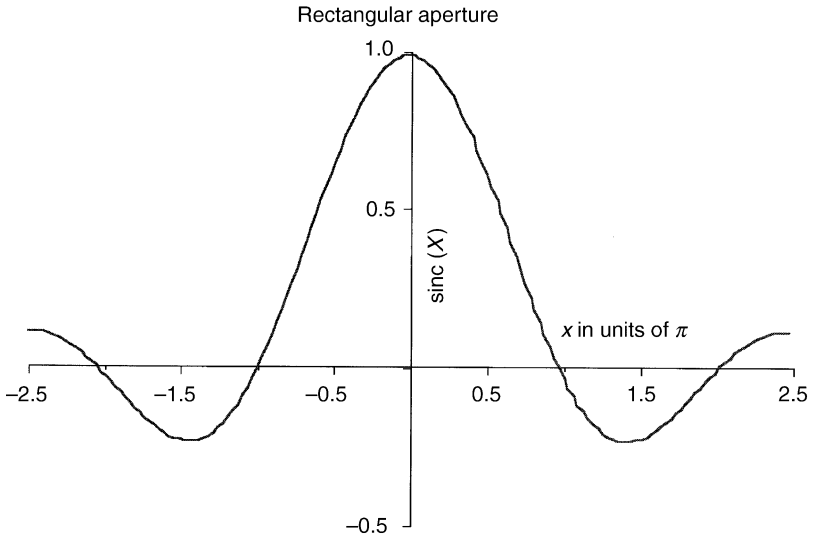


Figure 5 The amplitude of the wave diffracted by a rectangular slit. This sinc function describes the light wave's amplitude that would exist in both the *x* and *y* directions. The coordinate would have to be scaled by the dimension of the slit. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Figure 5 is a theoretical plot of the amplitude of the diffraction wave from a rectangular slit.

Diffraction from a Circular Aperture

To obtain the diffraction pattern from a circular aperture we first convert from the rectangular coordinate system we have used to a cylindrical coordinate system. The new cylindrical geometry is shown in Figure 6. At the aperture plane

$$\begin{aligned} x = s{\cdot}\cos\varphi \qquad & y = s{\cdot}\sin\varphi \\ f(x,y) = f(s,\varphi) \qquad & \mathrm{d}x\,\mathrm{d}y = s\,\mathrm{d}s\,\mathrm{d}\varphi \end{aligned} \quad [57]$$

At the observation plane

$$\xi = \rho{\cdot}\cos\theta \qquad \eta = \rho{\cdot}\sin\theta \quad [58]$$

In the new, cylindrical, coordinate system at the observation plane, the spatial frequencies are written as

$$\begin{aligned} \omega_x &= -\frac{k\xi}{R_0} = -\frac{k\rho}{R_0}\cos\theta \\ \omega_y &= -\frac{k\eta}{R_0} = -\frac{k\rho}{R_0}\sin\theta \end{aligned} \quad [59]$$

From Figure 6 we see that the observation point, P, can be defined in terms of the angle ψ, where

$$\sin\psi = \frac{\rho}{R_0} \quad [60]$$

This allows an angular representation of the size of the diffraction pattern if it is desired.

Using [57] and [59], we may write

$$\begin{aligned} \omega_x x + \omega_y y &= -\frac{ks\rho}{R_0}(\cos\theta\cos\varphi + \sin\theta\sin\varphi), \\ &= -\frac{ks\rho}{R_0}\cos(\theta-\varphi) \end{aligned} \quad [61]$$

The Huygens–Fresnel integral for Fraunhofer diffraction can now be written in terms of cylindrical coordinates:

$$\tilde{E}_P = \frac{\mathrm{i}\alpha}{\lambda R_0}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}\int_0^{\frac{a}{2}}\int_0^{2\pi} f(s,\varphi)\mathrm{e}^{-\mathrm{i}k\frac{s\rho}{R_0}\cos(\theta-\varphi)}\,s\,\mathrm{d}s\,\mathrm{d}\varphi \quad [62]$$

We will use this equation to calculate the diffraction amplitude from a clear aperture of diameter *a*, defined by the equation

$$f(s,\varphi) = \begin{cases} 1 & s \le \frac{a}{2},\ \text{all}\ \varphi \\ 0 & s > \frac{a}{2} \end{cases} \quad [63]$$

The symmetry of this problem is such that

$$f(s,\varphi) = f(s)g(\varphi) = f(s) \quad [64]$$

$$\mathcal{F}\{f(s,\phi)\} = \mathcal{F}\{f(s)\} = F(\rho) \quad [65]$$

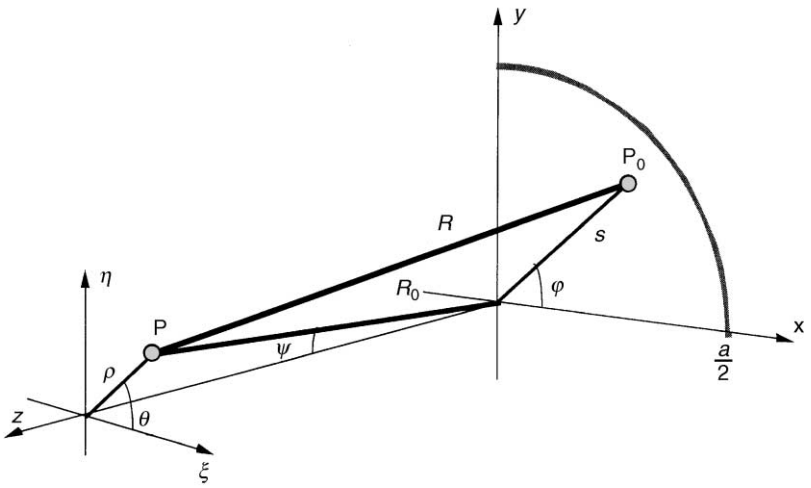


Figure 6 Geometry for calculation of Fraunhofer diffraction from a circular aperture. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

$$F(\rho,\theta)=\int_0^{\infty} f(s)s\,\mathrm{d}s\int_0^{2\pi}\mathrm{e}^{-\mathrm{i}\rho s\cos(\varphi-\theta)}\,\mathrm{d}\varphi \qquad [66]$$

$$F(\rho)=\int_0^{\infty} f(s)\,s\,\mathrm{d}s\int_0^{2\pi}\mathrm{e}^{-\mathrm{i}\rho s\cos\varphi}\,\mathrm{d}\varphi \qquad [67]$$

The second integral in [67] belongs to a class of functions called the Bessel function defined by the integral

$$\mathrm{J}_n(s\rho)=\frac{1}{2\pi}\int_0^{2\pi}\mathrm{e}^{\mathrm{i}[s\rho\sin\varphi-n\varphi]}\mathrm{d}\varphi \qquad [68]$$

In [67] the integral corresponds to the $n=0$, zero-order, Bessel function. Using this definition, we can rewrite [67] as

$$F(\rho)=\int_0^{\infty} f(s)\mathrm{J}_0(s\rho)s\,\mathrm{d}s \qquad [69]$$

This transform is called the Fourier–Bessel transform or the Hankel zero-order transform.

Using these definitions, the transform of the aperture function, $f(s)$, is

$$F(\rho)=\int_0^{\frac{a}{2}}\mathrm{J}_0(s\rho)s\,\mathrm{d}s \qquad [70]$$

We use the identity

$$x\mathrm{J}_1(x)=\int_0^{x}\chi\mathrm{J}_0(\chi)\,\mathrm{d}\chi \qquad [71]$$

to obtain

$$F(\rho)=\frac{a}{2\rho}\mathrm{J}_1\left(\frac{\rho a}{2}\right) \qquad [72]$$

We can now write the amplitude of the diffracted wave as

$$\tilde{E}_P=\frac{\mathrm{i}\alpha}{\lambda}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}\left[\frac{\pi a}{k\rho}\mathrm{J}_1\left(\frac{ka\rho}{2R_0}\right)\right] \qquad [73]$$

A plot of the function contained within the bracket of [73] is given in Figure 7. If we define

$$u=\frac{ka\rho}{2R_0} \qquad [74]$$

then the spatial distribution of intensity in the diffraction pattern can be written in a form known as the Airy formula,

$$I=I_0\left[\frac{2\mathrm{J}_1(u)}{u}\right]^2 \qquad [75]$$

where we have defined

$$I_0=\left(\frac{\alpha A}{\lambda R_0}\right)^2 \qquad [76]$$

with A representing the area of the aperture, $A=\pi(\frac{a}{2})^2$.

The intensity pattern described by [75] is called the Airy pattern. The intensity at $u=0$ is the same as was obtained for a rectangular aperture of the same area, because, in the limit,

$$\lim_{u\to 0}\left[\frac{2\mathrm{J}_1(u)}{u}\right]=1 \qquad [77]$$

For the Airy pattern, 84% of the total area is contained in the disk between the first zeros of [75]. Those zeros occur at $u=\pm 1.22\pi$, which corresponds

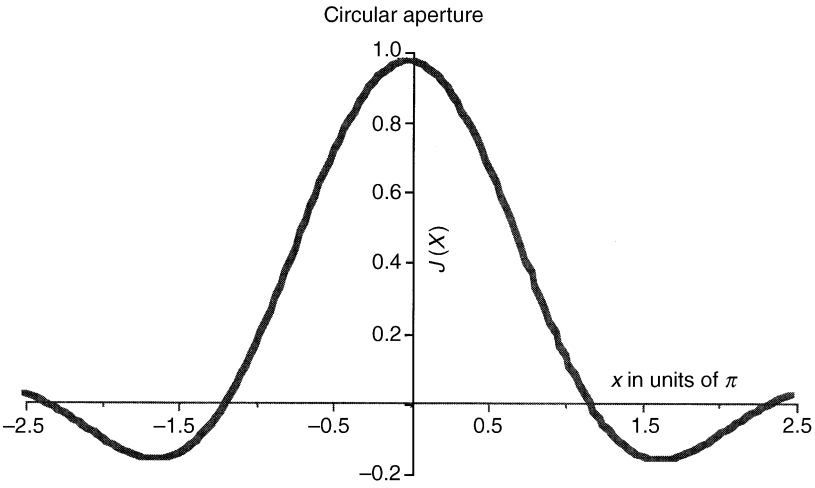


Figure 7 Amplitude of a wave diffracted by a circular aperture. The observed light distribution is constructed by rotating this Bessel function around the optical axis. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

to a radius in the observation plane of

$$\rho = \frac{1.22(\lambda R_0)}{a} \qquad [78]$$

91% of the light intensity is contained within the circle bounded by the second minimum at $u = 2.233\pi$. The intensities in the secondary maxima of the diffraction pattern of a rectangular aperture [53] are much larger than the intensities in the secondary maxima of the Airy pattern of a circular aperture of equal area. The peak intensities, relative to the central maximum, of the first three secondary maxima of a rectangular aperture are 4.7%, 1.6%, and 0.8%, respectively. For a circular aperture, the same quantities are 1.7%, 0.04%, and 0.02%, respectively.

Array Theorem

There is an elegant mathematical technique for handling diffraction from multiple apertures, called the array theorem. The theorem is based on the fact that Fraunhofer diffraction is given by the Fourier transform of the aperture function and utilizes the convolution integral.

The convolution of the functions $a(t)$ and $b(t)$ is defined as

$$g(t) = a(t) \otimes b(t) = \int_{-\infty}^{\infty} a(t)b(\tau - t)\,\mathrm{d}t \qquad [79]$$

The Fourier transform of a convolution of two functions is the product of the Fourier transforms of the individual functions:

$$\mathcal{F}\{a(t) \otimes b(t)\} = \mathcal{F}\left\{\int_{-\infty}^{\infty} a(t)b(\tau - t)\,\mathrm{d}t\right\} = A(\omega)B(\omega) \qquad [80]$$

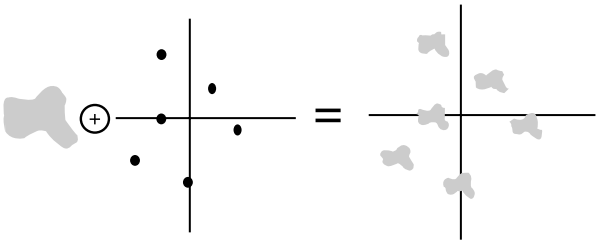


Figure 8 The convolution of an aperture with an array of delta functions will produce an array of identical apertures, each located at the position of one of the delta functions. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

We will demonstrate the array theorem for one dimension, where the functions represent slit apertures. The results can be extended to two dimensions in a straightforward way.

Assume that we have a collection of identical apertures, shown on the right of Figure 8. If one of the apertures is located at the origin of the aperture plane, its transmission function is $\psi(x)$. The transmission function of an aperture located at a point, x_n, can be written in terms of a generalized aperture function, $\psi(x - \alpha)$, by the use of the sifting property of the delta function

$$\psi(x - x_n) = \int \psi(x - \alpha)\delta(\alpha - x_n)\,\mathrm{d}\alpha \qquad [81]$$

The aperture transmission function representing an array of apertures will be the sum of the

distributions of the individual apertures, represented graphically in Figure 8 and mathematically by the summation

$$\Psi(x) = \sum_{n=1}^{N} \psi(x - x_n) \qquad [82]$$

The Fraunhofer diffraction from this array is $\Phi(\omega_x)$, the Fourier transform of $\Psi(x)$,

$$\Phi(\omega_x) = \int_{-\infty}^{\infty} \Psi(x) e^{-i\omega_x x}\, dx \qquad [83]$$

which can be rewritten as

$$\Phi(\omega_x) = \sum_{n=1}^{N} \int_{-\infty}^{\infty} \psi(x - x_n) e^{-i\omega_x x}\, dx \qquad [84]$$

We now make use of the fact that $\psi(x - x_n)$ can be expressed in terms of a convolution integral. The Fourier transform of $\psi(x - x_n)$ is, from the convolution theorem [80], the product of the Fourier transforms of the individual functions that make up the convolution:

$$\begin{aligned}\Phi(\omega_x) &= \sum_{n=1}^{N} \mathcal{F}\{\psi(x - \alpha)\}\mathcal{F}\{\delta(\alpha - x_n)\} \\ &= \mathcal{F}\{\psi(x - \alpha)\} \sum_{n=1}^{N} \mathcal{F}\{\delta(\alpha - x_n)\} \\ &= \mathcal{F}\{\psi(x - \alpha)\}\mathcal{F}\left\{\sum_{n=1}^{N} \delta(\alpha - x_n)\right\} \end{aligned} \qquad [85]$$

The first transform in [85] is the diffraction pattern of the generalized aperture function and the second transform is the diffraction pattern produced by a set of point sources with the same spatial distribution as the array of identical apertures. We will call this second transform the array function.

To summarize, the array theorem states that the diffraction pattern of an array of similar apertures is given by the product of the diffraction pattern from a single aperture and the diffraction (or interference) pattern of an identically distributed array of point sources.

N Rectangular Slits

An array of N identical apertures is called a diffraction grating in optics. The Fraunhofer diffraction patterns produced by such an array have two important properties; a number of very narrow beams are produced by the array and the beam positions are a function of the wavelength of illumination of the apertures and the relative phase of the waves radiated by each of the apertures.

Because of these properties, arrays of diffracting apertures have been used in a number of applications.

- At radio frequencies, arrays of dipole antennas are used to both radiate and receive signals in both radar and radio astronomy systems. One advantage offered by diffracting arrays at radio frequencies is that the beam produced by the array can be electrically steered by adjusting the relative phase of the individual dipoles.
- An optical realization of a two-element array of radiators is Young's two-slit experiment and a realization of a two-element array of receivers is Michelson's stellar interferometer. Two optical implementations of arrays, containing more diffracting elements, are diffraction gratings and holograms. In nature, periodic arrays of diffracting elements are the origin of the colors observed on some invertebrates.
- Many solids are naturally arranged in three-dimensional arrays of atoms or molecules that act as diffraction gratings when illuminated by X-ray wavelengths. The resulting Fraunhofer diffraction patterns are used to analyze the ordered structure of the solids.

The array theorem can be used to calculate the diffraction pattern from N rectangular slits, each of width a and separation d. The aperture function of a single slit is equal to

$$\psi(x, y) = \begin{cases} 1 & |x| \le \frac{a}{2}, |y| \le y_0 \\ 0 & \text{all other } x \text{ and } y \end{cases} \qquad [86]$$

The Fraunhofer diffraction pattern of this aperture has already been calculated and is given by [52]:

$$\mathcal{F}\{\psi(x, y)\} = K \frac{\sin \alpha}{\alpha} \qquad [87]$$

where the constant K and the variable α are

$$\begin{aligned} K &= \frac{i2ay_0\alpha}{\lambda R_0} e^{-i\mathbf{k}\cdot\mathbf{R}_0} \frac{\sin \omega_y y_0}{\omega_y y_0} \\ \alpha &= \frac{ka}{2} \sin \theta_x \end{aligned} \qquad [88]$$

The array function is

$$A(x) = \sum_{n=1}^{N} \delta(x - x_n) \quad \text{where } x_n = (n - 1)d \qquad [89]$$

and its Fourier transform is given by

$$\mathcal{F}\{A(x)\} = \frac{\sin N\beta}{\sin \beta} \qquad [90]$$

$$\beta = \frac{kd}{2}\sin\theta_x$$

The Fraunhofer diffraction pattern's intensity distribution in the x-direction is thus given by

$$I_\theta = I_0 \underbrace{\frac{\sin^2\alpha}{\alpha^2}}_{\text{shape factor}} \underbrace{\frac{\sin^2 N\beta}{\sin^2\beta}}_{\text{grating factor}} \qquad [91]$$

We have combined the variation in intensity in the y-direction into the constant I_0 because we assume that the intensity variation in the x-direction will be measured at a constant value of y.

A physical interpretation of [91] views the first factor as arising from the diffraction associated with a single slit; it is called the shape factor. The second factor arises from the interference between light from different slits; it is called the grating factor. The fine detail in the spatial light distribution of the diffraction pattern is described by the grating factor and arises from the coarse detail in the diffracting object. The coarse, overall light distribution in the diffraction pattern is described by the shape factor and arises from the fine detail in the diffracting object.

Young's Double Slit

The array theorem makes the analysis of Young's two-slit experiment a trivial exercise. This application of the array theorem will demonstrate that the interference between two slits arises naturally from an application of diffraction theory. The result of this analysis will support a previous assertion that interference describes the same physical process as diffraction and the division of the two subjects is an arbitrary one.

The intensity of the diffraction pattern from two slits is obtained from [91] by setting $N = 2$:

$$I_\theta = I_0 \frac{\sin^2\alpha}{\alpha^2}\cos^2\beta \qquad [92]$$

The sinc function describes the energy distribution of the overall diffraction pattern, while the cosine function describes the energy distribution created by interference between the light waves from the two slits. Physically, α is a measure of the phase difference between points in one slit and β is a measure of the phase difference between similar points in the two slits. Zeros in the diffraction intensity occur whenever $\alpha = n\pi$ or whenever $\beta = \frac{1}{2}(2n+1)\pi$. Figure 9 shows the interference maxima, from the grating factor, under the central diffraction peak, described by the shape factor. The number of interference maxima contained under the central maximum is given by

$$\frac{2d}{a} - 1 \qquad [93]$$

In Figure 9 three different slit spacings are shown with the ratio d/a equal to 3, 6, and 9, respectively.

The Diffraction Grating

In this section we will use the array theorem to derive the diffraction intensity distribution of a large number of identical apertures. We will discover that the positions of the principal maxima are a function of the illuminating wavelength. This functional relationship has led to the application of a diffraction grating to wavelength measurements.

The diffraction grating normally used for wavelength measurements is not a large number of diffracting apertures but rather a large number of reflecting grooves cut in a surface such as gold or aluminum. The theory to be derived also applies to these reflection gratings but a modification must be made to the theory. The shape of the grooves in the reflection grating can be used to control the fraction of light diffracted into a principal maximum and we will examine how to take this into account. A grating whose groove shape has been controlled to enhance the energy contained in a particular principal maximum is called a blazed grating. The use of special groove shapes is equivalent to the modification of the phase of individual elements of an antenna array at radio frequencies.

The construction of a diffraction grating for use in an optical device for measuring wavelength was first suggested by David Rittenhouse, an American astronomer, in 1785, but the idea was ignored until Fraunhofer reinvented the concept in 1819. Fraunhofer's first gratings were fine wires spaced by wrapping the wires in the threads of two parallel screws. He later made gratings by cutting (*ruling*) grooves in gold films deposited on the surface of glass. H.A. Rowland made a number of well designed ruling machines, which made possible the production of large-area gratings. Following a suggestion by Lord Rayleigh, Robert Williams Wood (1868–1955) developed the capability to control the shape of the individual grooves.

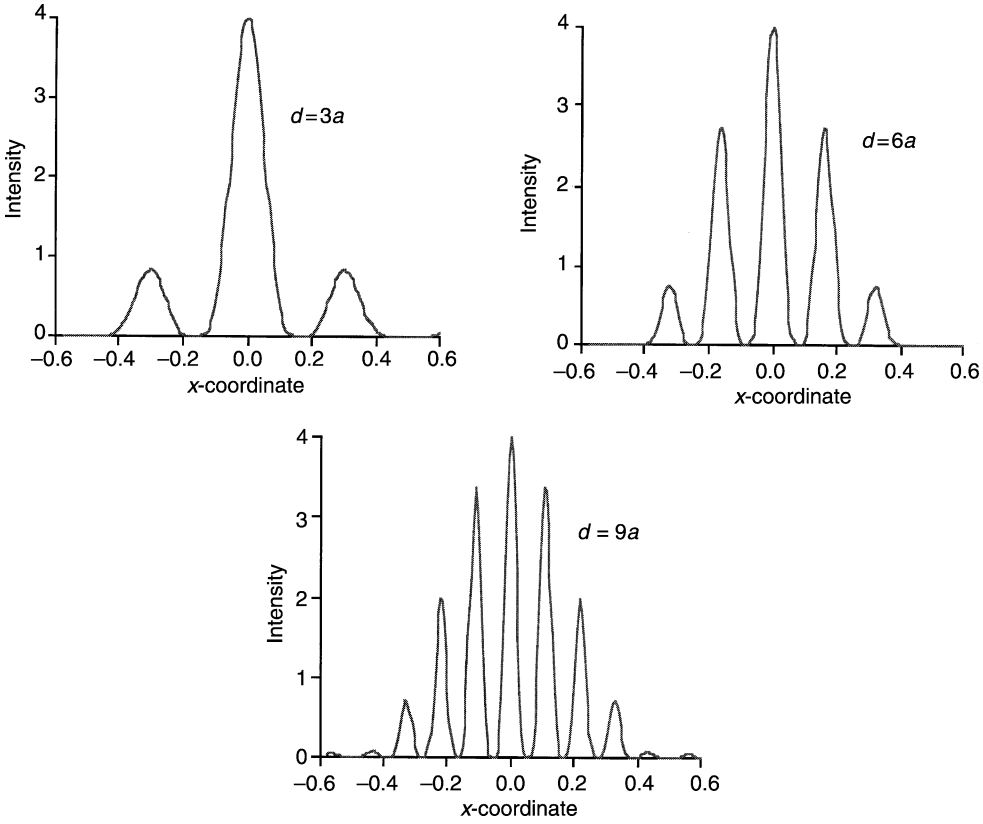


Figure 9 The number of interference fringes beneath the main diffraction peak of a Young's two-slit experiment with rectangular apertures. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

If N is allowed to assume values much larger than 2, the appearance of the interference fringes, predicted by the grating factor, changes from a simple sinusoidal variation to a set of narrow maxima, called principal maxima, surrounded by much smaller, secondary maxima. To calculate the diffraction pattern we must evaluate [91] when N is a large number.

Whenever $N\beta = m\pi$, $m = 0, 1, 2, \ldots$, the numerator of the second factor in [91] will be zero, leading to an intensity that is zero, $I_\theta = 0$. The denominator of the second factor in [91] is zero when $\beta = l\pi$, $l = 0, 1, 2, \ldots$ If both conditions occur at the same time, the ratio of m and N is equal to an integer, $m/N = l$, and instead of zero intensity, we have an indeterminate value for the intensity, $I_\theta = 0/0$. To find the actual value for the intensity, we must apply L'Hospital's rule

$$\lim_{\beta \to l\pi} \frac{\sin N\beta}{\sin \beta} = \lim_{\beta \to l\pi} \frac{N \cos N\beta}{\cos \beta} = N \qquad [94]$$

L'Hospital's rule predicts that whenever

$$\beta = l\pi = \frac{m}{N}\pi \qquad [95]$$

where l is an integer, a principal maximum in the intensity will occur with a value given by

$$I_{\theta P} = N^2 I_0 \frac{\sin^2 \alpha}{\alpha^2} \qquad [96]$$

Secondary maxima, much weaker than the principal maxima, occur when

$$\beta = \left(\frac{2m+1}{2N}\right)\pi \qquad m = 1, 2, \ldots \qquad [97]$$

(When $m = 0$, m/N is an integer, thus the first value that m can have in [97] is $m = 1$.) The intensity of each secondary maximum is given by

$$\begin{aligned} I_{\theta S} &= I_0 \frac{\sin^2 \alpha}{\alpha^2} \frac{\sin^2 N\beta}{\sin^2 \beta} \\ &= I_0 \frac{\sin^2 \alpha}{\alpha^2} \left[\frac{\sin^2\left(\frac{2m+1}{2}\right)\pi}{\sin^2\left(\frac{2m+1}{2N}\right)\pi} \right] \\ &= I_0 \frac{\sin^2 \alpha}{\alpha^2} \left[\frac{1}{\sin\left(\frac{2m+1}{2N}\right)\pi} \right]^2 \end{aligned} \qquad [98]$$

The quantity $(2m+1)/2N$ is a small number for large values of N, allowing the small angle approximation to be made:

$$I_{\theta S} \approx I_0 \frac{\sin^2\alpha}{\alpha^2}\left[\frac{2N}{\pi(2m+1)}\right]^2 \qquad [99]$$

The ratio of the intensity of a secondary maximum and a principal maximum is given by

$$\frac{I_{\theta S}}{I_{\theta P}} = \left[\frac{2}{\pi(2m+1)}\right]^2 \qquad [100]$$

The strongest secondary maximum occurs for $m=1$ and, for large N, has an intensity that is about 4.5% of the intensity of the neighboring principal maximum.

The positions of principal maxima occur at angles specified by the grating formula.

$$\beta = \frac{kd\sin\theta}{2} = l\pi = \frac{m}{N}\pi \qquad [101]$$

The angular position of the principal maxima (the diffracted angle θ_d) is given by the Bragg equation

$$\sin\theta_d = \frac{l\lambda}{d} \qquad [102]$$

where l is called the grating order. This simple relationship between the angle and the wavelength can be used to construct a device to measure wavelength, the grating spectrometer.

The model we have used to obtain this result is based on a periodic array of identical apertures. The transmission function of this array is a periodic square wave. If, for the moment, we treat the grating as infinite in size, we discover that the principal maxima in the diffraction pattern correspond to the terms of the Fourier series describing a square wave.

The zero order, $m=0$, corresponds to the temporal average of the periodic square wave and has an intensity proportional to the spatially averaged transmission of the grating. Because of its equivalence to the temporal average of a time-varying signal, the zero-order principal maximum is often called the dc term.

The first-order principal maximum corresponds to the fundamental spatial frequency of the grating and the higher orders correspond to the harmonics of this frequency.

The dc term provides no information about the wavelength of the illumination. Information about the wavelength of the illuminating light can only be obtained by measuring the angular position of the first or higher order principal maxima.

Grating Spectrometer

The curves shown in Figure 10 display the separation of the principal maxima as a function of $\sin\theta_d$. The angular separation of principal maxima can be converted to a linear dimension by assuming a distance, R, from the graing to the observation plane. In the lower right-hand curve of Figure 10, a distance of 2 meters was assumed. Grating spectrometers are classified by the size in meters of R used in their design. The larger R, the easier it is to resolve wavelength differences. For example, a 1 meter spectrometer is a higher-resolution instrument than a 1/4 meter spectrometer.

The fact that the grating is finite in size causes each of the orders to have an angular width that limits the resolution with which the illuminating wavelength can be measured. To calculate the resolving power of the grating, we first determine the angular width of a principal maximum. This is accomplished by measuring the angular change, of the principal maximum's position, when β changes from $\beta = l\pi = m\pi/N$ to $\beta = (m+1)\pi/N$, i.e., $\Delta\beta = \pi/N$. Using the definition of β

$$\beta = \frac{kd\sin\theta_d}{2} \qquad [103]$$

$$\Delta\beta = \frac{\pi d\cos\theta_d\Delta\theta}{\lambda}$$

and the angular width is

$$\Delta\theta = \frac{\lambda}{Nd\cos\theta_d} \qquad [104]$$

The derivative of the grating formula gives

$$\Delta\lambda = \frac{d}{l}\cos\theta_d\Delta\theta \qquad [105]$$

$$\Delta\theta = \frac{l\Delta\lambda}{d\cos\theta_d}$$

Equating this spread in angle to eqn [104] yields

$$\frac{l\Delta\lambda}{d\cos\theta} = \frac{\lambda}{Nd\cos\theta} \qquad [106]$$

The resolving power of a grating is therefore

$$\frac{\lambda}{\Delta\lambda} = Nl \qquad [107]$$

The improvement of resolving power with N can be seen in Figure 10. A grating 2 inches wide and containing 15 000 grooves per inch would have a resolving power in second order ($l=2$) of 6×10^4. At a wavelength of 600 nm this grating could resolve two waves, differing in wavelength by 0.01 nm.

The diffraction grating is limited by overlapping orders. If two wavelengths, λ and $\lambda+\Delta\lambda$, have successive orders that are coincident, then

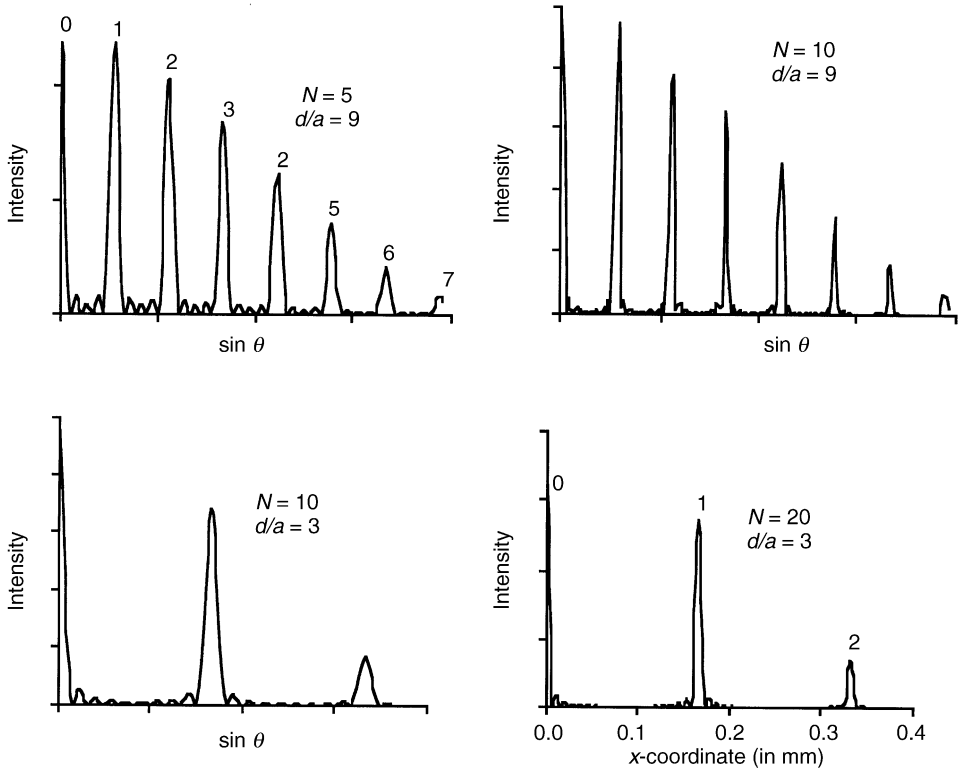


Figure 10 Decrease in the width of the principal maxima of a transmission grating with an increasing number of slits. The various principal maxima are called orders, numbering from zero, at the origin, out to as large as seven in this example. Also shown is the effect of different ratios, d/a, on the number of visible orders. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

from eqn [102]

$$(l+1)\lambda = l(\lambda + \Delta\lambda) \qquad [108]$$

The minimum wavelength difference for which this occurs is defined as the free spectral range of the diffraction grating

$$(\Delta\lambda)_{SR} = \frac{\lambda}{l} \qquad [109]$$

Blazed Gratings

We have been discussing amplitude transmission gratings. Amplitude transmission gratings have little practical use because they waste light. The light loss is from a number of sources.

1. Light is diffracted simultaneously into both positive and negative orders (the positive and negative frequencies of the Fourier transform). The negative diffraction orders contain redundant information and waste light.
2. In an amplitude transmission grating, light is thrown away because of the opaque portions of the slit array.
3. The width of an individual aperture leads to a shape factor for a rectangular aperture of $\text{sinc}^2\alpha = (\sin^2\alpha)/\alpha^2$, which modulates the grating factor and causes the amplitude of the orders to rapidly decrease. This can be observed in **Figure 10**, where the second order is very weak. Because of the loss in intensity at higher orders, only the first few orders ($l = 1, 2$ or 3) are useful. The shape factor also causes a decrease in diffracted light intensity with increasing wavelength for higher-order principal maxima.
4. The location of the maximum in the diffracted light, i.e., the angular position for which the shape factor is a maximum, coincides with the location of the principal maximum due to the zero-order interference. This zero-order maximum is independent of wavelength and is not of much use.

One solution to the problems created by transmission gratings would be the use of a grating that modified only the phase of the transmitted wave. Such gratings would operate using the same physical processes as a microwave phased array antenna, where the location of the shape factor's maximum is controlled by adding a constant phase shift to each antenna element. The construction of an optical

transmission phase grating with a uniform phase variation across the aperture of the grating is very difficult. For this reason, a second approach, based on the use of reflection gratings, is the practical solution to the problems listed above.

By tilting the reflecting surface of each groove of a reflection grating, Figure 11, the position of the shape factor's maximum can be controlled. Problems 1, 3, and 4 are eliminated because the shape factor maximum is moved from the optical axis out to some angle with respect to the axis. The use of reflection gratings also removes Problem 2 because all of the incident light is reflected by the grating.

Robert Wood, in 1910, developed the technique of producing grooves of a desired shape in a reflective grating by shaping the diamond tool used to cut the grooves. Gratings, with grooves shaped to enhance their performance at a particular wavelength, are said to be blazed for that wavelength. The physical properties on which blazed gratings are based can be understood by using Figure 11. The groove faces can be treated as an array of mirror surfaces. The normal to each of the groove faces makes an angle θ_B with the normal to the grating surface. We can measure the angle of incidence and the angle of diffraction with respect to the grating normal or with respect to the groove normal, as shown in Figure 11. From Figure 11, we can write a relationship between the angles

$$\theta_i = \varphi_i - \theta_B \qquad -\theta_d = -\varphi_d + \theta_B \qquad [110]$$

(The sign convention used defines positive angles as those measured in a counterclockwise rotation from the normal to the surface. Therefore, θ_B is a negative angle.) The blaze angle provides an extra degree of freedom that will allow independent adjustment of the angular location of the principal maxima of the grating factor and the zero-order, single-aperture, diffraction maximum. To see how this is accomplished, we must determine, first, the effect of off-axis illumination of a diffraction grating.

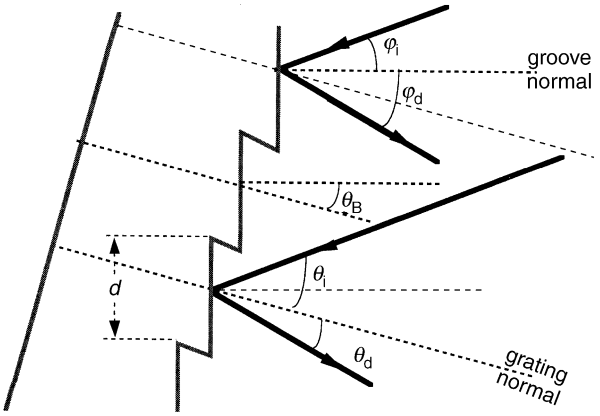


Figure 11 Geometry for a blazed reflection grating. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Off-axis illumination is easy to incorporate into the equation for the diffraction intensity from an array. To include the effect of an off-axis source, the phase of the illuminating wave is modified by changing the incident illumination from a plane wave of amplitude, E, traveling parallel to the optical axis, to a plane wave with the same amplitude, traveling at an angle θ_i to the optical axis: $\tilde{E}\, e^{-ikx \sin \theta_i}$. (Because we are interested only in the effects in a plane normal to the direction of propagation, we ignore the phase associated with propagation along the z-direction, $kz \cos \theta_i$.) The off-axis illumination results in a modification of the parameter for single-aperture diffraction from

$$\alpha = \frac{ka}{2} \sin \theta_d \qquad [111]$$

to

$$\alpha = \frac{ka}{2} (\sin \theta_i + \sin \theta_d) \qquad [112]$$

and for multiple aperture interference from

$$\beta = \frac{kd}{2} \sin \theta_d \qquad [113]$$

to

$$\beta = \frac{kd}{2} (\sin \theta_i + \sin \theta_d) \qquad [114]$$

The zero-order, single-aperture, diffraction peak occurs when $\alpha = 0$. If we measure the angles with respect to the groove face, this occurs when

$$\alpha = \frac{ka}{2} (\sin \varphi_i + \sin \varphi_d) = 0 \qquad [115]$$

The angles are therefore related by

$$\sin \varphi_i = -\sin \varphi_d \qquad [116]$$

$$\varphi_i = -\varphi_d$$

We see that the single-aperture, diffraction maximum (the shape factor's maximum) occurs at the same angle that reflection from the groove faces occurs. We can write this result in terms of the angles measured with respect to the grating normal

$$\theta_i = -(\theta_d + 2\theta_B) \qquad [117]$$

The blaze condition requires the single aperture diffraction maximum to occur at the lth principal

maximum, for wavelength λ_B. At that position

$$l\pi = \frac{2\pi d}{\lambda_B}(\sin\theta_i + \sin\theta_d) \qquad [118]$$

$$l\lambda_B = 2d\sin\tfrac{1}{2}(\theta_i + \theta_d)\cos\tfrac{1}{2}(\theta_i - \theta_d)$$

For the special geometrical configuration called the Littrow condition, where $\theta_i = \theta_d$, we find that [117] leads to the equation

$$l\lambda_B = 2d\sin\theta_B \qquad [119]$$

By adjusting the blaze angle, the single-aperture diffraction peak can be positioned on any order of the interference pattern. Typical blaze angles are between 15° and 30° but gratings are made with larger blaze angles.

See also

Diffraction: Fresnel Diffraction.

Further Reading

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.

Guenther RD (1990) *Modern Optics*. New York: Wiley.

Haus HA (1984) *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall.

Hecht E (1998) *Optics*. 3rd edn. Reading, MA: Addison-Wesley.

Jackson JD (1962) *Classical Electrodynamics*. New York: Wiley.

Klein MV (1970) *Optics*. New York: Wiley.

O'Neill EL (1963) *Introduction to Statistical Optics*. Reading, MA: Addison-Wesley.

Rossi B (1957) *Optics*. Reading, MA: Addison-Wesley.

Rubinowicz A (1984) Tahe Miyamoto–Wolf diffraction wave. In: Wolf E (ed.) *Progress in Optics*, vol. IV, p. 201. Amsterdam: North-Holland.

Sommerfeld A (1964) *Optics*. New York: Academic Press.

Stone JM (1963) *Radiation and Optics*. New York: McGraw-Hill.

Fresnel Diffraction

B D Guenther, Duke University, Durham, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Fresnel was a civil engineer who pursued optics as a hobby, after a long day of road construction. Based on his own very high-quality observations of diffraction, Fresnel used the wave propagation concept developed by Christiaan Huygens (1629–1695) to develop a theoretical explanation of diffraction. We will use a descriptive approach to obtain the Huygens–Fresnel integral by assuming an aperture can be described by N pinholes which act as sources for Huygens' wavelets. The interference between these sources will lead to the Huygens–Fresnel integral for diffraction.

Huygens' principle views wavefronts as the product of wavelets from various luminous points acting together. To apply Huygens' principle to the propagation of light through an aperture of arbitrary shape, we need to develop a mathematical description of the field from an array of Huygens' sources filling the aperture. We will begin by obtaining the field from a pinhole that is illuminated by a plane wave,

$$\tilde{\mathbf{E}}_i(\mathbf{r}, t) = \tilde{\mathbf{E}}_i(\mathbf{r})e^{i\omega t}$$

Following the lead of Fresnel, we will use theory of interference to combine the fields from two pinholes and then generalize to N pinholes. Finally by letting the areas of each pinhole approach an infinitesimal value, we will construct an arbitrary aperture of infinitesimal pinholes. The result will be the Huygens–Fresnel integral.

We know that the wave obtained after propagation through an aperture must be a solution of the wave equation,

$$\nabla^2\tilde{\mathbf{E}} = \mu\varepsilon\frac{\partial^2\tilde{\mathbf{E}}}{\partial t^2}$$

We will be interested only in the spatial variation of the wave so we need only look for solutions of the Helmholtz equation

$$(\nabla^2 + k^2)\tilde{\mathbf{E}} = 0$$

The problem is further simplified by replacing this vector equation with a scalar equation,

$$(\nabla^2 + k^2)\tilde{E}(x, y, z) = 0$$

This replacement is proper for those cases where $\mathbf{n}E(x,y,z)$ [where $\mathbf{n}$ is a unit vector] is a solution of the vector Helmholtz equation. In general, we cannot substitute $\mathbf{n}E$ for the electric field $\mathbf{E}$ because of Maxwell's equation

$$\nabla\cdot\mathbf{E} = 0$$

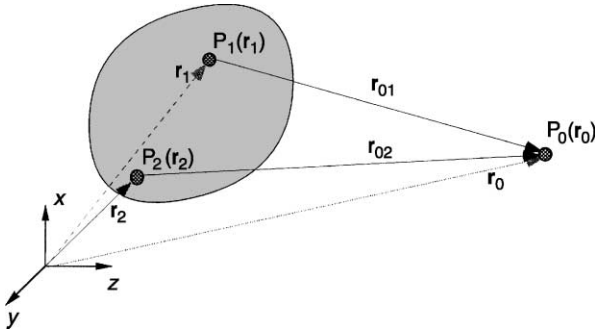


Figure 1 Geometry for application of Huygens' principle to two pinholes. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

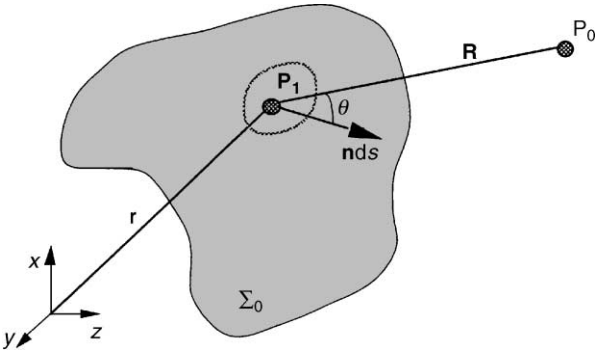


Figure 2 The geometry for calculating the field at P_0 using [9]. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Rather than working with the magnitude of the electric field, we should use the scalar amplitude of the vector potential. We will neglect this complication and assume the scalar, E, is a single component of the vector field, **E**. A complete solution would involve a scalar solution for each component of **E**.

The pinhole is illuminated by a plane wave, and the wave that leaves the pinhole will be a spherical wave which is written in complex notation as

$$\tilde{E}(\mathbf{r})e^{i\omega t} = A\frac{e^{-i\delta}e^{-ik\cdot r}}{r}e^{i\omega t}$$

The complex amplitude

$$\tilde{E}(\mathbf{r}) = A\frac{e^{-i\delta}e^{-ik\cdot r}}{r} \qquad [1]$$

is a solution of the Helmholtz equation. The field at P_0, in Figure 1, from two pinholes: one at P_1, located a distance $\mathbf{r}_{01} = |\mathbf{r}_0 - \mathbf{r}_1|$ from P_0, and one at P_2, located a distance $\mathbf{r}_{02} = |\mathbf{r}_0 - \mathbf{r}_2|$ from P_0, is a generalization of Young's interference. The complex amplitude is

$$\tilde{E}(\mathbf{r}_0) = \frac{A_1}{r_{01}}e^{-ik\cdot r_{01}} + \frac{A_2}{r_{02}}e^{-ik\cdot r_{02}} \qquad [2]$$

We have incorporated the phase δ_1 and δ_2 into the constants A_1 and A_2 to simplify the equations.

The light emitted from the pinholes is due to a wave, E_i, incident onto the screen from the left. The units of E_i are per unit area so to obtain the amount of light passing through the pinholes, we must multiply E_i by the areas of the pinholes. If $\Delta\sigma_1$ and $\Delta\sigma_2$ are the areas of the two pinholes, respectively, then

$$A_1 \propto \tilde{E}_i(\mathbf{r}_1)\Delta\sigma_1 \qquad A_2 \propto \tilde{E}_i(\mathbf{r}_2)\Delta\sigma_2$$

$$\tilde{E}(\mathbf{r}_0) = C_1\frac{\tilde{E}_i(\mathbf{r}_1)}{r_{01}}e^{-ik\cdot r_{01}}\Delta\sigma_1 + C_2\frac{\tilde{E}_i(\mathbf{r}_2)}{r_{02}}e^{-ik\cdot r_{02}}\Delta\sigma_2 \qquad [3]$$

where C_i is the constant of proportionality. The constant will depend on the angle that $\mathbf{r}_{0i}$ makes with the normal to $\Delta\sigma_i$. This geometrical dependence arises from the fact that the apparent area of the pinhole decreases as the observation angle approaches 90°.

We can generalize [3] to N pinholes

$$\tilde{E}(\mathbf{r}_0) = \sum_{j=1}^{N} C_j\frac{\tilde{E}_i(\mathbf{r}_j)}{r_{0j}}e^{-ik\cdot r_{0j}}\Delta\sigma_j \qquad [4]$$

The pinhole's diameter is assumed to be small compared to the distances to the viewing position but large compared to a wavelength. In the limit as $\Delta\sigma_j$ goes to zero, the pinhole becomes a Huygens source. By letting N become large, we can fill the aperture with these infinitesimal Huygens' sources and convert the summation to an integral.

It is in this way that we obtain the complex amplitude, at the point P_0, see Figure 2, from a wave exiting an aperture, Σ_0, by integrating over the area of the aperture. We replace $\mathbf{r}_{0j}$, in [4], by **R**, the position of P_1, the infinitesimal Huygens' source of area, ds, measured with respect to the observation point, P_0. We also replace $\mathbf{r}_j$ by **r**, the position of the infinitesimal area, ds, with respect to the origin of the coordinate system. The discrete summation [4] becomes the integral

$$\tilde{E}(\mathbf{r}_0) = \int\int_A C(\mathbf{r})\frac{\tilde{E}_i(\mathbf{r})}{R}e^{-ik\cdot\mathbf{R}}ds \qquad [5]$$

This is the Fresnel integral. The variable $C(\mathbf{r})$ depends upon θ, the angle between **n**, the unit vector normal to the aperture, and **R**, shown in Figure 2. We now need to determine how to treat $C(\mathbf{r})$.

The Obliquity Factor

When using Huygens' principle, a problem arises with the spherical wavelet produced by each

Huygens' source. Part of the spherical wavelet propagates backward and results in an envelope propagating toward the light source, but such a wave is not observed in nature. Huygens neglected this problem. Fresnel required the existence of an obliquity factor to cancel the backward wavelet but was unable to derive its functional form. When Kirchhoff placed the theory of diffraction on a firm mathematical foundation, the obliquity factor was generated quite naturally in the derivation. In this intuitive derivation of the Huygens–Fresnel integral, we will present an argument for treating the obliquity factor as a constant at optical wavelengths. We will then derive the constant value it must be assigned.

The obliquity factor, $C(\mathbf{r})$, in [5], causes the amplitude per unit area of the transmitted light to decrease as the viewing angle increases. This is a result of a decrease in the effective aperture area with viewing angle. When Kirchhoff applied Green's theorem to the scalar diffraction problem, he found the obliquity factor to have an angular dependence given by

$$\frac{\cos(\hat{\mathbf{n}}, \mathbf{R}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_s)}{2} \qquad [6]$$

which includes the geometrical effect of the incident wave arriving at the aperture, at an angle with respect to the normal to the aperture from a source located at a position $\mathbf{r}_s$, with respect to the aperture. If the source is at infinity, then the incident wave can be treated as a plane wave, incident normal to the aperture, and we may simplify the angular dependence of the obliquity factor to

$$\frac{1 + \cos\theta}{2}$$

where $\theta \rightarrow (\hat{\mathbf{n}}, \mathbf{R})$ is the angle between the normal to the aperture and the vector $\mathbf{R}$. This is the configuration shown in Figure 2. The obliquity factor provides an explanation of why it is possible to ignore the backward-propagating wave that occurs in application of Huygens' principle. For the backward wave, $\theta = \pi$, and the obliquity factor is zero.

The obliquity factor increases the difficulty of working with the Fresnel integral and it is to our benefit to be able to treat it as a constant. We can neglect the angular contribution of the obliquity factor by making an assumption about the resolving power of an optical system operating at visible wavelengths.

Assume we are attempting to resolve two stars that produce plane waves at the aperture of a telescope with an aperture diameter, a. The wavefronts from the two stars make an angle ψ with respect to each other, Figure 3:

$$\tan\psi \approx \psi = \frac{\Delta x}{a}$$

The smallest angle, ψ, that can be measured is determined by the smallest length, Δx, that can be measured. We know we can measure a fraction of wavelength with an interferometer but, without an interferometer, we can measure a length no smaller than λ, leading to the assumption that $\Delta x \geq \lambda$. This reasoning leads to the assumption that the smallest angle we can measure is

$$\psi \geq \frac{\lambda}{a} \qquad [7]$$

The resolution limit established by the above reasoning places a limit on the minimum separation that can be produced at the back focal plane of the telescope. The minimum distance on the focal plane between the images of star 1 and 2 is given by

$$d = f\psi \qquad [8]$$

From [7]

$$d = \frac{\lambda f}{a}$$

The resolution limit of the telescope can also be expressed in terms of the cone angle produced when the incident plane wave is focused on the back focal plane of the lens. From the geometry of Figure 3, the cone angle is given by

$$\tan\theta = \frac{a}{2f}$$

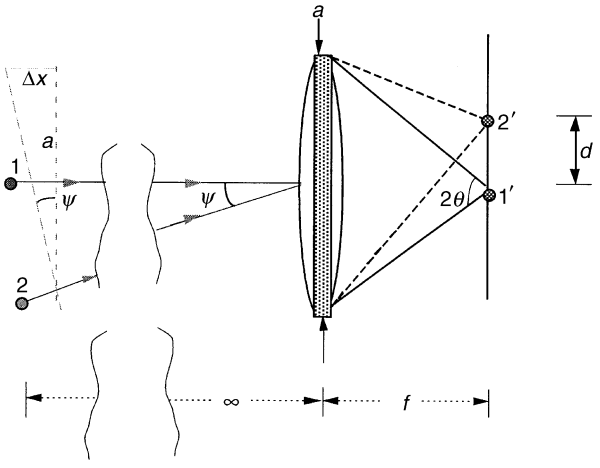


Figure 3 Telescope resolution. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The separation between the stars 1 and 2 at the back focal plane of the lens in Figure 3 is thus

$$d = \frac{\lambda}{2\tan\theta} \qquad [9]$$

If we assume that the minimum d is 3λ (this is four times the resolution of a typical photographic film at visible wavelengths), then the largest obliquity angle that should be encountered in a visible imaging system is

$$\tan\theta = \frac{\lambda}{2d} = \frac{\lambda}{6\lambda} = \tfrac{1}{6} = 0.167$$

$$\theta = 9.5^\circ \approx 10^\circ$$

yielding a value for $\cos\theta = 0.985$ and $(1 + cos\ \theta)/2 = 0.992$. The obliquity factor has only changed by 0.8% over the angular variation of 0° to 10°. The obliquity factor as a function of angle is shown in Figure 4 for an incident plane wave. The obliquity factor undergoes a 10% change when θ varies from 0° to about 40°; therefore, the obliquity factor can safely be treated as a constant in any optical system that involves angles less than 40°.

While we have shown that it is safe to ignore the variability of C, we still have not assigned a value to the obliquity factor. To find the proper value for C, we will compare the result obtained using [5] with the result predicted by using geometric optics.

We illuminate the aperture Σ_0 in Figure 5 with a plane wave of amplitude α, traveling parallel to the z-axis. Geometrical optics (equivalently the propagation of an infinite plane wave) predicts a field at P_0, on the z-axis, a distance z_0 from the aperture, given by

$$\tilde{E}_{\text{geom}} = \alpha e^{-ikz_0} \qquad [10]$$

The area of the infinitesimal at P_1 (the Huygens' source) is

$$ds = r\,dr\,d\phi$$

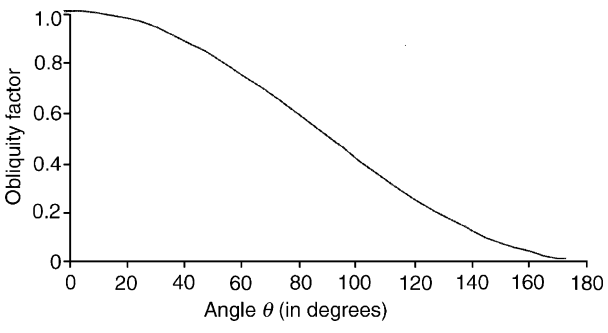


Figure 4 The obliquity factor as a function of the angle defined in eqn [6] for an incident plane wave. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

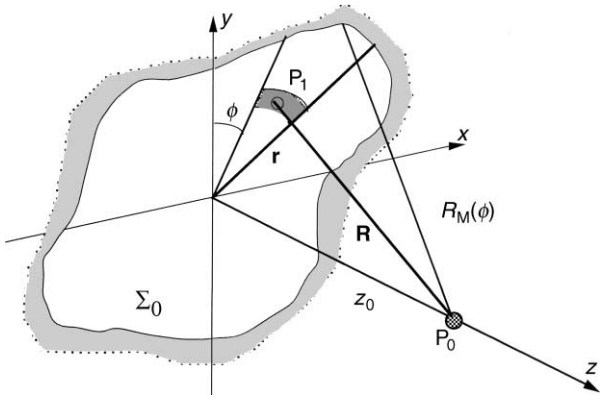


Figure 5 Geometry for evaluating the constant in the Fresnel integral. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Based on our previous argument, the obliquity factor can be treated as a constant, C, that can be removed from under the integral. The incident wave is a plane wave whose value at $z = 0$ is $E(r) = \alpha$. Using these parameters, the Fresnel integral [5] can be written as

$$\tilde{E}(z_0) = C\alpha \int\int \frac{e^{-ik\cdot R}}{R} r\,dr\,d\phi \qquad [11]$$

From the geometry in Figure 5, the distance from the Huygens source to the observation point is

$$z_0^2 + r^2 = R^2$$

where z_0 is a constant equal to the distance from the observation plane to the aperture plane. The variable of integration can be written in terms of R:

$$r\,dr = R\,dR$$

The limits of integration over the aperture extend from $R = z_0$ to the maximum value of R, $R = R_M(\phi)$.

$$\tilde{E}(z_0) = C\alpha \int_0^{2\pi} \int_{z_0}^{R_M(\phi)} e^{-ikR} dR\,d\phi \qquad [12]$$

The integration over R can now be carried out to yield

$$\tilde{E}(z_0) = \underbrace{\frac{C\alpha}{ik} e^{-ikz_0} \int_0^{2\pi} d\phi}_{\text{Geometrical Optics}} - \underbrace{\frac{C\alpha}{ik} \int_0^{2\pi} e^{-ikR_M(\phi)} d\phi}_{\text{Diffraction}} \qquad [13]$$

The first integration in [13] is easy to perform and contains the amplitude at the observation point due to geometric optics. The second term may be interpreted as interference of the waves diffracted by the boundary of the aperture. An equivalent statement is that the second term is the interference of the waves scattered from the aperture's boundary, an interpretation of diffraction that was first suggested by Young.

The aperture is irregular in shape, at least on the scale of a wavelength, thus in general, kR_M will vary over many multiples of 2π as we integrate around the aperture. For this reason, we should be able to ignore the second integral in [13] if we confine our attention to the light distribution on the z-axis. After neglecting the second term, we are left with only the geometrical optics component of [13]:

$$\tilde{E}(z_0) = \frac{2\pi C}{ik}\alpha e^{-ikz_0} \qquad [14]$$

For [14] to agree with the prediction of geometric optics [10] the constant C must be equal to

$$C = \frac{ik}{2\pi} = \frac{i}{\lambda} \qquad [15]$$

The Huygens–Fresnel integral can be written, using the value for C just derived,

$$\tilde{E}(\mathbf{r}_0) = \frac{i}{\lambda}\int\int_{\Sigma}\frac{\tilde{E}_i(\mathbf{r})}{R}e^{-ik\cdot R}\,ds \qquad [16]$$

The job of calculating diffraction has only begun with the derivation of the Huygens–Fresnel integral [16]. Rigorous solutions exist for only a few idealized obstructions. To allow discussion of the general properties of diffraction, it is necessary to use approximate solutions. To determine what type approximations we can make let's look at the light propagation path shown in Figure 6.

For the second term in [13] to contribute to the field at point P_0, in Figure 6, the phase of the exponent that we neglected in [13] must not vary over 2π when the integration is performed over the aperture, i.e.,

$$\Delta = \mathbf{k}\cdot(\mathbf{R}_1 - \mathbf{R}'_1 + \mathbf{R}_2 - \mathbf{R}'_2) < 2\pi$$

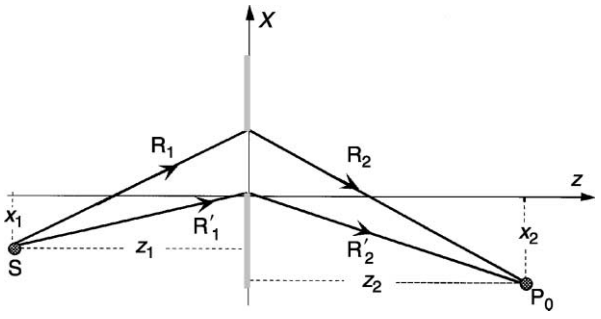


Figure 6 One-dimensional slit used to establish Fresnel approximations. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

From the geometry of Figure 6, the two paths from the source, S, to the observation point, P_0, are

$$R_1 + R_2 = \sqrt{z_1^2 + (x_1 + b)^2} + \sqrt{z_2^2 + (x_2 + b)^2}$$

$$R'_1 + R'_2 = \sqrt{z_1^2 + x_1^2} + \sqrt{z_2^2 + x_2^2}$$

By assuming that the aperture, of width b, is small compared to the distances z_1 and z_2, the difference between these distances, Δ, can be rewritten as a binomial expansion of a square root

$$\sqrt{1+b} = 1 + \frac{b}{2} - \frac{b^2}{8} + \ldots \qquad [17]$$

thus

$$\Delta \approx k\left(\frac{x_1}{z_1} + \frac{x_2}{z_2}\right)b + \frac{k}{2}\left(\frac{1}{z_1} + \frac{1}{z_2}\right)b^2 + \ldots \qquad [18]$$

If we assume that the second term of this expansion is small, i.e.,

$$\frac{k}{2}\left(\frac{1}{z_1} + \frac{1}{z_2}\right)b^2 \ll 2\pi$$

$$\frac{b^2}{2}\frac{1}{z_1} + \frac{1}{z_2} < \lambda$$

we see that the phase of the wavefront in the aperture is assumed to have a quadratic dependence upon aperture coordinates. Diffraction predicted by this assumption is called Fresnel diffraction.

To discover the details of the approximation consider Figure 7. In the geometry of Figure 7, the Fresnel integral [16] becomes

$$\tilde{E}_{P_0} = \frac{i\alpha}{\lambda}\int\int_{\Sigma} f(x,y)\frac{e^{-ik\cdot(R+R')}}{RR'}dx\,dy \qquad [19]$$

As we mentioned in our discussion of [13] the integral that adds diffractive effects to the geometrical optics field is nonzero only when the phase of the integrand is stationary. For Fresnel diffraction, we can insure that the phase is nearly constant if R does not differ appreciably from Z, or R' from Z'. This is equivalent to stating that only the wave in the aperture, around a point in Figure 7 labeled **S**, called the stationary point, will contribute to E_P. The stationary point, **S**, is the point in the aperture plane where the line, connecting the source and observation positions, intersects the plane. Physically, only light propagating over paths nearly equal to the path predicted by geometrical optics (obtained from Fermat's principle) will contribute to E_{P_0}.

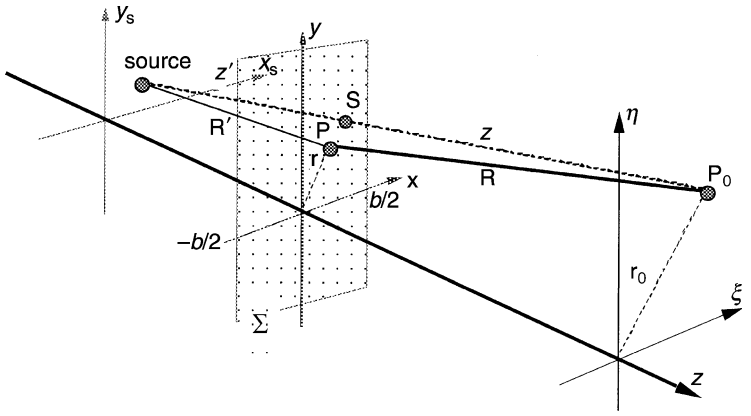


Figure 7 Geometry for Fresnel diffraction. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The geometry of Figure 7 allows the distances R and R' to be written as

$$R^2 = (x-\xi)^2 + (y-\eta)^2 + Z^2 \qquad [20]$$
$$R'2 = (x_s - x)^2 + (y_s - y)^2 + Z'2$$

To solve these expressions for R and R', we apply the binomial expansion [17] and retain the first two terms:

$$R + R' \approx Z + Z' + \left[\frac{(x-\xi)^2}{2Z} + \frac{(x_s-x)^2}{2Z'}\right] + \left[\frac{(y-\eta)^2}{2Z} + \frac{(y_s-y)^2}{2Z'}\right] \qquad [21]$$

In the denominator of [19], R is replaced by Z and R' by Z'. (By making this replacement, we are implicitly assuming that the amplitudes of the spherical waves are not modified by the differences in propagation distances over the area of integration. This is a reasonable approximation because the two propagation distances are within a few hundred wavelengths of each other). In the exponent, we must use [21], because the phase changes by a large fraction of a wavelength, as we move from point to point in the aperture.

If the wave, incident on the aperture, were a plane wave we can assume $R' = \infty$ and [19] becomes

$$\tilde{E}_{P_0} = \frac{i\alpha}{\lambda}\frac{e^{-ikZ}}{Z}\int\int_{\Sigma} f(x,y)e^{-\frac{ik}{2Z}[(x-\xi)^2+(y-\eta)^2]}dx\,dy \qquad [22]$$

The physical interpretation of [22] states that when a plane wave illuminates the obstruction, the field at point P_0 is a spherical wave, originating at the aperture, a distance Z away from P_0:

$$\frac{e^{-ikZ}}{Z}$$

The amplitude and phase of this spherical wave are modified by the integral containing a quadratic phase dependent on the obstruction's spatial coordinates.

By defining three new parameters,

$$\rho = \frac{ZZ'}{Z+Z'} \quad \text{or} \quad \frac{1}{\rho} = \frac{1}{Z} + \frac{1}{Z'} \qquad [23a]$$
$$x_0 = \frac{Z'\xi + Zx_s}{Z+Z'} \qquad [23b]$$
$$y_0 = \frac{Z'\eta + Zy_s}{Z+Z'} \qquad [23c]$$

the more general expression for Fresnel diffraction of a spherical wave can be placed in the same format as [22]. The newly defined parameters x_0 and y_0 are the coordinates, in the aperture plane, of the stationary point, **S**. The parameters in [23] can be used to express, after some manipulation, the spatial dependence of the phase, in the integrand of [19]

$$R + R' = Z + Z' + \frac{(\xi - x_s)^2 + (\eta - y_s)^2}{2(Z+Z')} + \left[\frac{(x-x_0)^2 + (y-y_0)^2}{2\rho}\right]$$

A further simplification to the Fresnel diffraction integral can be made by obtaining an expression for the distance, D, between the source and the observation point:

$$D = Z + Z' + \frac{(\xi - x_s)^2 + (\eta - y_s)^2}{2(Z+Z')} \qquad [24]$$

We use this definition of D to also write

$$\frac{1}{ZZ'} = \frac{Z+Z'}{ZZ'}\frac{1}{Z+Z'} \approx \frac{1}{\rho D}$$

Using the parameters we have just defined, we may rewrite [19] as

$$\tilde{E}_{P_0} = \frac{i\alpha}{\lambda\rho D} e^{-ikD} \int\int f(x,y) e^{-\frac{ik}{2\rho}[(x-x_0)^2+(y-y_0)^2]} dx\,dy \qquad [25]$$

With the use of the variables defined in [23], the physical significance of the general expression of the Huygens–Fresnel integral can be understood in an equivalent fashion as was [22]. At point P_0, a spherical wave

$$\frac{e^{-ikD}}{D}$$

originating at the source, a distance D away, is observed, if no obstruction are present. Because of the obstruction, the amplitude and phase of the spherical wave are modified by the integral in [25]. This correction to the predictions of geometrical optics is called Fresnel diffraction.

Mathematically, [25] is an application of the method of stationary phase, a technique developed in 1887 by Lord Kelvin to calculate the form of a boat's wake. The integration is nonzero only in the region of the critical point we have labeled **S**. Physically, the light distribution, at the observation point, is due to wavelets from the region around **S**. The phase variations of light, coming from other regions in the aperture, are so rapid that the value of the integral over those spatial coordinates is zero. The calculation of the integral for Fresnel diffraction is complicated because, if the observation point is moved, a new integration around a different stationary point in the aperture must be performed.

Rectangular Apertures

If the aperture function, $f(x,y)$, is separable in the spatial coordinates of the aperture, then we can rewrite [25] as

$$\tilde{E}_{P_0} = \frac{i\alpha}{\lambda\rho D} e^{-ikD} \int_{-\infty}^{\infty} f(x) e^{-ig(x)} dx \int_{-\infty}^{\infty} f(y) e^{-ig(y)} dy$$

$$\tilde{E}_{P_0} = A[C(x) - iS(x)][C(y) - iS(y)]$$

A represents the spherical wave from the source, a distance D away,

$$A = \frac{i\alpha}{2D} e^{-ikD}$$

If we treat the aperture function as a simple constant, C and S are integrals of the form

$$C(x) = \int_{x_1}^{x_2} \cos[g(x)]dx \qquad [26a]$$

$$S(x) = \int_{x_1}^{x_2} \sin[g(x)]dx \qquad [26b]$$

The integrals, $C(x)$ and $S(x)$, have been evaluated numerically and are found in tables of Fresnel integrals. To use the tabulated values for the integrals, [26] must be written in a general form

$$C(w) = \int_0^w \cos\left(\frac{\pi u^2}{2}\right) du \qquad [27]$$

$$S(w) = \int_0^w \sin\left(\frac{\pi u^2}{2}\right) du \qquad [28]$$

The variable u is an aperture coordinate, measured relative to the stationary point, $\mathbf{S}(x_0,y_0)$, in units of

$$\sqrt{\frac{\lambda\rho}{2}}$$

$$u = \sqrt{\frac{2}{\lambda\rho}}(x - x_0) \text{ or } u = \sqrt{\frac{2}{\lambda\rho}}(y - y_0) \qquad [29]$$

The parameter w in [27] and [28] specifies the location of the aperture edge relative to the stationary point **S**. The parameter w is calculated through the use of [29] with x and y replaced by the coordinates of the aperture edge.

The Cornu Spiral

The plot of $S(w)$ versus $C(w)$, shown in Figure 8, is called the Cornu spiral in honor of M. Alfred Cornu (1841–1902) who was the first to use this plot for graphical evaluation of the Fresnel integrals.

To use the Cornu spiral, the limits w_1 and w_2 of the aperture are located along the arc of the spiral. The length of the straight line segment drawn from w_1 to w_2 gives the magnitude of the integral. For example, if there were no aperture present, then, for the x-dimension, $w_1 = -\infty$ and $w_2 = \infty$. The length of the line segment from the point $(-1/2,1/2)$ to $(1/2,1/2)$ would be the value of E_{P_0}, i.e., $\sqrt{2}$. An identical value is obtained for the y-dimension, so that

$$\tilde{E}_{P_0} = 2A = \frac{\alpha e^{-ikD}}{D}$$

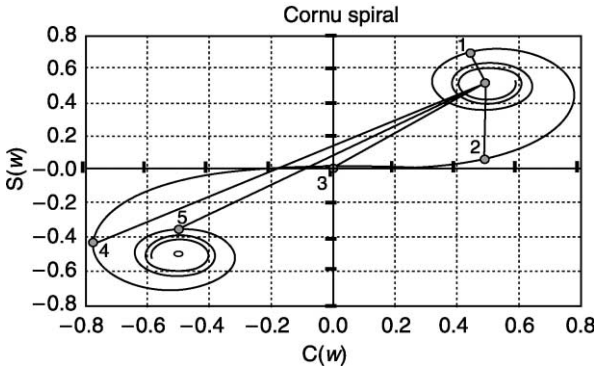


Figure 8 Cornu spiral obtained by plotting the values for $C(w)$ versus $S(w)$ from a table of Fresnel integrals. The numbers indicated along the arc of the spiral are values of w. The points 1–5 correspond to observation points shown in **Figure 9** and are used to calculate the light intensity diffracted by a straight edge. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

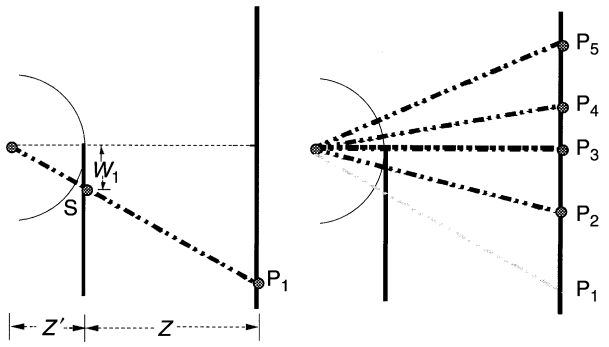


Figure 9 A geometrical construct to determine w_1 for the stationary point S associated with five different observation points. The values of w are then used in **Figure 8** to calculate the intensity of light diffracted around a straight edge. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

This is the spherical wave that is expected when no aperture is present.

As an example of the application of the Fresnel integrals, assume the diffracting aperture is an infinitely long straight edge reducing the problem to one dimension:

$$f(x, y) = \begin{Bmatrix} 1 & x \geq x_1 \\ 0 & x < x_1 \end{Bmatrix}$$

The value of E_{P_0} in the y-dimension is from above $\sqrt{2}$ since $w_1 = -\infty$ and $w_2 = \infty$. In the x-dimension, the value of w at the edge is

$$w_{x_1} = \sqrt{\frac{2}{\lambda\rho}}(x_1 - x_0)$$

We will assume that the straight edge blocks the negative half-plane so that $x_1 = 0$. The straight edge is treated as an infinitely wide slit with one edge located at infinity so that $w_2 = \infty$ in the upper half-plane. When the observation point is moved, the coordinate x_0 of **S** changes and the origin of the aperture's coordinate system moves. New values for the w's must therefore be calculated for each observation point.

Figure 9 shows the assumed geometry. A set of observation points, P_1–P_5, on the observation screen are selected. The origin of the coordinate system in the aperture plane is relocated to a new position (given by the position of **S**) when a new observation point is selected. The value of w_1, the position of the edge with respect to **S**, must be recalculated for each observation point. The distance from the origin to the straight edge, w_1, is positive for P_1 and P_2, zero for P_3, and negative for P_4 and P_5.

Table 1 Fresnel integrals for straight edge

w_1	$C(w_1)$	$S(w_1)$	I_P/I_0	P_j
∞	0.5	0.5	0	
2.0	0.4882	0.3434	0.01	
1.5	0.4453	0.6975	0.021	P_1
1.0	0.7799	0.4383	0.04	
0.5	0.4923	0.0647	0.09	P_2
0	0	0	0.25	P_3
−0.5	−0.4923	−0.0647	0.65	
−1.0	−0.7799	−0.4383	1.26	P_4
−1.2	−0.7154	−0.6234	1.37	
−1.5	−0.4453	−0.6975	1.16	
−2.0	−0.4882	−0.3434	0.84	P_5
−2.5	−0.4574	−0.6192	1.08	
$-\infty$	−0.5	−0.5	1.0	

Figure 8 shows the geometrical method used to calculate the intensity values at each of the observation points in Figure 9 using the Cornu spiral. The numbers labeling the straight line segments in Figure 8 are associated with the labels of the observation points in Figure 9.

To obtain an accurate calculation of Fresnel diffraction from a straight edge, a table of Fresnel integrals provides the input to the following equation

$$I_P = I_0\left\{\left[\tfrac{1}{2} - C(w_1)\right]^2 + \left[\tfrac{1}{2} - S(w_1)\right]^2\right\}$$

where $I_0 = 2A^2$. Table 1, shows the values extracted from the table of Fresnel integrals to find the relative intensity at various observation points. The result obtained by using either method for calculating the light distribution in the observation plane, due to the straight edge in Figure 9, is plotted in Figure 10. The relative intensities at the observation points

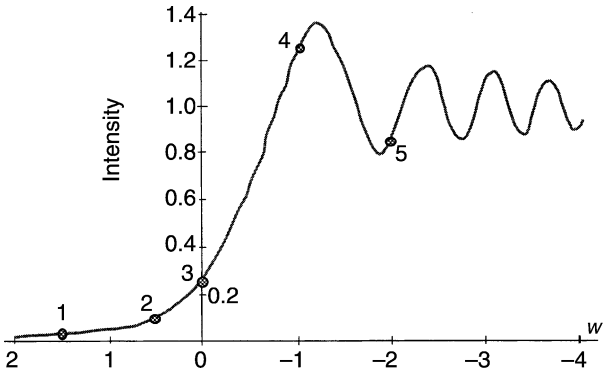


Figure 10 Light diffracted by a straight edge with its edge located at $w = 0$. The points labeled 1–5 were found using the construction in **Figure 9** to obtain w. This was then used in **Figure 8** to find a position on the Cornu spiral. The length of the lines shown in **Figure 8** from the (1/2,1/2) point to the numbered positions led to the intensities shown. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

depicted in **Figure 9** are labeled on the diffraction curve of **Figure 10**.

Fresnel Zones

Fresnel used a geometrical construction of zones to evaluate the Huygens–Fresnel integral. The Fresnel zone is a mathematical construct, serving the role of a Huygens source in the description of wave propagation. Assume that at time t, a spherical wavefront from a source at P_1 has a radius of R'. To determine the field at the observation point, P_0, due to this wavefront, a set of concentric spheres of radii, $Z, Z + (\lambda/2), Z + 2(\lambda/2), \ldots, Z + j(\lambda/2), \ldots$ are constructed, where Z is the distance from the wavefront to the observation point on the line connecting P_1 and P_0 (see **Figure 11**). These spheres divide the wavefront into a number of zones, $\zeta_1, \zeta_2, \ldots, \zeta_j \ldots$, called Fresnel zones, or half-period zones.

We treat each zone as a circular aperture illuminated from the left by a spherical wave of the form

$$\tilde{E}_j(\mathbf{R}') = \frac{A e^{-i\mathbf{k}\cdot\mathbf{R}'}}{R'} = \frac{A e^{-ikR'}}{R'}$$

R' is the radius of the spherical wave. The field at P_0 due to the jth zone is obtained by using [5]

$$\tilde{E}_j(P_0) = \frac{A}{R'} e^{-i\mathbf{k}\cdot\mathbf{R}'} \int\int_{\zeta_j} C(\varphi) \frac{e^{-i\mathbf{k}\cdot\mathbf{R}}}{R} ds \qquad [30]$$

For the integration over the jth zone, the surface element is

$$ds = R' 2\sin\theta d\theta \, d\phi \qquad [31]$$

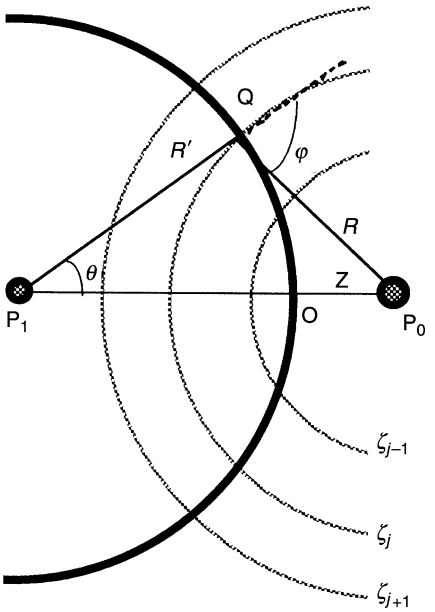


Figure 11 Construction of Fresnel zones observed from a position P_0 on a spherical wave originating from a source at point P_1. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

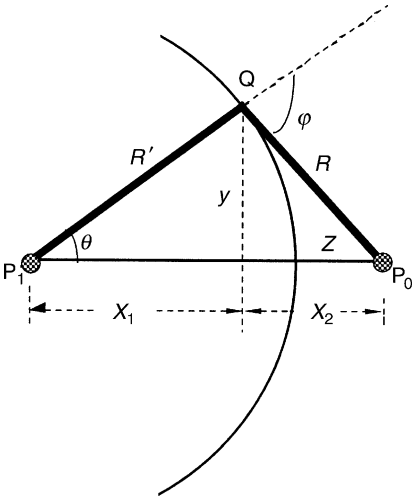


Figure 12 Geometry for finding the relationship between R and R' yielding eqn [32]. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The limits of integration extend over the range

$$Z + (j-1)\tfrac{\lambda}{2} \le R \le Z + j\tfrac{\lambda}{2}$$

The variable of integration is R; thus, a relationship between R' and R must be found. This is accomplished by using the geometrical construction shown in **Figure 12**.

A perpendicular is drawn from the point Q on the spherical wave to the line connecting P_1 and P_0, in **Figure 12**. The distance from the source to the

observation point is $x_1 + x_2$ and the distance from the source to the plane of the zone is the radius of the incident spherical wave, R'. The distance from P_1 to P_0 can be written

$$x_1 + x_2 = R' + Z$$

The distance from the observation point to the zone is

$$R^2 = x_2^2 + y^2 = [(R' + Z) - x_1]^2 + (R'\sin\theta)^2$$

$$= R'2 + (R' + Z)^2 - 2R'(R' + Z)\cos\theta$$

The derivative of this expression yields

$$R\,\mathrm{d}R = R'(R' + Z)\sin\theta\,\mathrm{d}\theta \qquad [32]$$

Substituting [32] into [31] gives

$$\mathrm{d}s = \frac{R'}{R' + Z} R\mathrm{d}R\mathrm{d}\phi \qquad [33]$$

The integration over ϕ is accomplished by rotating the surface element about the P_1P_0 axis. After integrating over ϕ between the limits of 0 and 2π, we obtain

$$\tilde{E}_j(P_0) = \frac{2\pi A}{R' + Z} e^{-ik\cdot R'} \int_{Z+(j-1)\frac{\lambda}{2}}^{Z+\frac{j\lambda}{2}} e^{-ik\cdot R} C(\varphi)\mathrm{d}R \quad [34]$$

We assume that $R',\ Z \gg \lambda$ so that the obliquity factor is a constant over a single zone, i.e., $C(\varphi) = C_j$. The obliquity factor is not really a constant as we earlier assumed but rather a very slowly varying parameter of j, changing by less than two parts in 10^4 when 'j' changes by 500 (for $Z = 1$ meter and $\lambda = 500$ nm). In fact it only changes by 2 parts in 10^{-7} across one zone. We are safe in applying the assumption that the obliquity factor is a constant over a zone and this allows the integral in [34] to be calculated:

$$\tilde{E}_j(P_0) = \frac{2\pi i C_j A}{k(R' + Z)} e^{-ik(R'+Z)} e^{-ikj\frac{\lambda}{2}} \left[1 - e^{-ik\frac{\lambda}{2}}\right] \quad [35]$$

Using the identity $k\lambda = 2\pi$ and the definition for the distance between the source and observation point [24] modified for this geometry, $D = R' + Z$ [35] can be simplified to

$$\tilde{E}_j(P_0) = 2i\lambda(-1)^j \frac{C_j A}{D} e^{-ikD} \qquad [36]$$

The physical reasons for the behavior predicted by [36] are quite easy to understand. The distance from P_0 to a zone changes by only $\lambda/2$ as we move from zone to zone and the area of a zone is almost a constant, independent of the zone number; thus, the amplitudes of the Huygens' wavelets from each zone should be approximately equal. The alternation in sign, from zone to zone, is due to the phase change of the light wave from adjacent zones because the propagation paths for adjacent zones differ by $\lambda/2$.

To find the total field strength at P_0 due to N zones, the collection of Huygens' wavelets is added:

$$E(P_0) = \sum_{j=1}^{N} \tilde{E}_j(P_0) = \frac{2i\lambda A}{D} e^{-ikD} \sum_{j=1}^{N} (-1)^j C_j \quad [37]$$

To evaluate the sum, the elements of the sum are regrouped and rewritten as

$$-\sum_{j=1}^{N} (-1)^j C_j = \frac{C_1}{2} + \left(\frac{C_1}{2} - C_2 + \frac{C_3}{2}\right) + \left(\frac{C_3}{2} - C_4 + \frac{C_5}{2}\right) + \ldots$$

Because the C's are very slowly varying functions of j, even out to 500 zones, we are justified in setting the quantities in parentheses equal to zero. With this approximation, the summation can be set equal to one of two values, depending upon whether there is an even or odd number of terms in the summation

$$-\sum_{j=1}^{N} (-1)^j C_j = \left\{ \begin{array}{ll} \frac{1}{2}(C_1 + C_N) & N \text{ odd} \\ \frac{1}{2}(C_1 - C_N) & N \text{ even} \end{array} \right\}$$

For very large N, the obliquity factor approaches zero, $C_N \to 0$, as was demonstrated in **Figure 4**. Thus, the theory has led us to the conclusion that the total field produced by an unobstructed wave is equal to one half the contribution from the first Fresnel zone, i.e., $E = E_1/2$. Stating this result in a slightly different way, we obtain a surprising result – the contribution from the first Fresnel zone is twice the amplitude of the unobstructed wave!

The zone construction can be used to analyze the effect of the obstruction of all or part of a zone. For example, by constructing a circular aperture with a diameter equal to the diameter of the first Fresnel zone, we have just demonstrated that it is possible to produce an intensity at the point P_0 equal to four times the intensity that would be observed if no aperture were present. To analyze the effects of a smaller aperture, we subdivide a half-period zone into a number of subzones such that there is a constant phase difference between each subzone. The individual vectors form a curve know as the vibrational curve. **Figure 13** shows a vibrational curve produced by the addition of waves from nine subzones of the first Fresnel zone. The vibrational curve in **Figure 13** is an arc with the appearance of a half-circle.

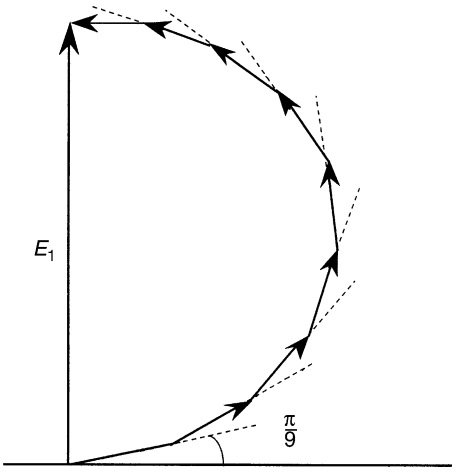


Figure 13 Vector addition of waves from nine subzones constructed in the first Fresnel zone. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

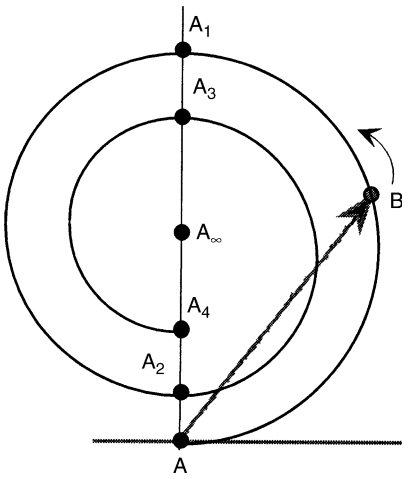


Figure 14 Vibration curve for determining the Fresnel diffraction from a circular aperture. The change of the diameter of the half-circles making up the spiral has been exaggerated for easy visualization. The actual changes are one part in 10^{-5}. The position of point B is determined by the aperture size. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

If the radius of curvature of the arc were calculated, we would discover that it is a constant, except for the contribution of the obliquity factor,

$$\rho = \frac{\lambda R'}{R' + Z} C(\varphi)$$

Because the obliquity factor for a single zone is a constant to 2 parts in 10^{-7}, the radius of curvature of the vibration curve can be considered a constant over a single zone. If we let the number of subzones approach infinity, the vibrational curve becomes a semicircle whose chord is equal to the wavelet produced by zone one, i.e., E_1. If we subdivide additional zones, and add the subzone contributions, we create other half-circles whose radii decrease at the same rate as the obliquity factor. The vibrational curve for the total wave is a spiral, constructed of semicircles which converges to a point halfway between the first half-circle, see Figure 14. The length of the vector from the start to the end of the spiral is $E = E_1/2$, as we derived above. When this same construction technique is applied to a rectangular aperture, the vibrational curve generated is the Cornu spiral.

Circular Aperture

The vector addition technique, described in Figure 13, can be used to evaluate Fresnel diffraction, at a point P_0, from a circular aperture and yields the intensity distribution along the axis of symmetry of the circular aperture. The zone concept will also allow a qualitative description of the light distribution normal to this axis.

To develop a quantitative estimate of the intensity at an observation point on the axis of symmetry of the circular aperture, we construct a spiral, Figure 14, to represent the Fresnel zones of a spherical wave incident on the aperture. The point B on the spiral shown in Figure 14 corresponds to the portion of the spherical wave unobstructed by the screen. The length of the chord, AB, represents the amplitude of the light wave at the observation point P_0. As the diameter of the aperture increases, B moves along the spiral, in a counterclockwise fashion, away from A. The first maximum occurs when B reaches the point labeled A_1 in Figure 14; the aperture has then uncovered the first Fresnel zone. At this point, the amplitude is twice what it would be with no obstruction. Four times the intensity!

If the aperture's diameter continues to increase, B reaches the point labeled A_2 in Figure 14 and the amplitude is very nearly zero; two zones are now exposed in the aperture. Further maxima occur when an odd number of zones are in the aperture and further minima when an even number of zones are exposed. Figure 15 shows an aperture containing four exposed Fresnel zones. The amplitude at the observation point would correspond to the chord drawn from A to A_4 in Figure 14.

The aperture diameter can be fixed and the observation point P_0 can move along, or perpendicular to, the axis of symmetry of the circular aperture. As P_0 is moved away from the aperture, along the symmetry axis, i.e., as Z increases, the radius of the Fresnel zones increase without limit. For small values of n, the radius of the nth zone can be approximated by

$$r_n \approx \sqrt{nZ\lambda} \qquad [38]$$

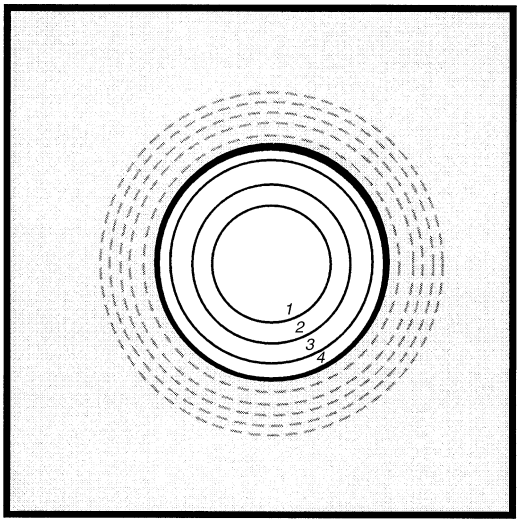


Figure 15 Aperture with four Fresnel zones exposed. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

At Z_{max}, the light intensity is a maximum, given by the chord length from A to A_1 in Figure 14. If $\lambda = 500$ nm and $a = 0.5$ mm then this maximum occurs when $Z = 0.5$ m

$$Z_{max} = \frac{a^2}{\lambda} \qquad [39]$$

If we start at Z_{max} and move toward the aperture, along the axis, as Z decreases in value, a point will be reached when the intensity on the axis becomes a minimum. The value of Z where the first minimum in intensity is observed is equal to

$$Z_{min} = \frac{a^2}{2\lambda}$$

In Figure 14 the chord would extend from A to A_2.

As the observation point, P_0, is moved along the axis toward the aperture, and Z assumes values less than Z_{min}, the point B in Figure 14 spirals inward, toward the center of the spiral and the intensity cycles through maximum and minimum values. The cycling of the intensity, as the observation point moves toward the aperture, will not continue indefinitely. At some point, the field on the axis will approach the field observed without the aperture because the distance between Z_{max} and Z_{min} shrinks to the wavelength of light making the intensity variation unobservable.

For values of Z that exceed Z_{max} of [39], the aperture radius, a, will be smaller than the radius of the first zone, and Fraunhofer diffraction will be observed because the aperture contains only one zone and the phase in the aperture is a constant.

Opaque Screen

If the screen containing a circular aperture, of radius a, is replaced by an opaque disk of radius a, the intensity distribution on the symmetry axis, behind the disk, is found to be equal to the value that would be observed with no disk present. This prediction was first derived by Poisson, to demonstrate that wave theory was incorrect; however, experimental observation supported the prediction and verified the theory.

We construct a spiral, shown in Figure 16, similar to the one for the circular aperture in Figure 14. The point B on the spiral represents the edge of the disk. The shaded portion of the spiral from A to B does not contribute because that portion of the wave is covered by the disk and the zones, associated with that portion of the wave, cannot be seen from the observation point. The amplitude at P_0 is the length of the chord from B to A_∞ shown in Figure 16. If the observation point moves toward the disk, then B moves along the spiral toward A_∞. There is always intensity on the axis for this configuration, though it slowly decreases until it reaches zero when the observation point reaches the disk; this corresponds to point B reaching point A_∞ on the spiral. Physically, zero intensity occurs when the disk blocks the entire light wave. There are no maxima or minima observed as the disk diameter, a, increases or as the observation distance changes. If the observation point is moved perpendicular to the symmetry axis, a set of concentric bright rings are observed. The origin of these bright rings can be explained using Fresnel zones, in a manner similar to the one used to explain the bright rings observed in a Fresnel diffraction pattern from a circular aperture.

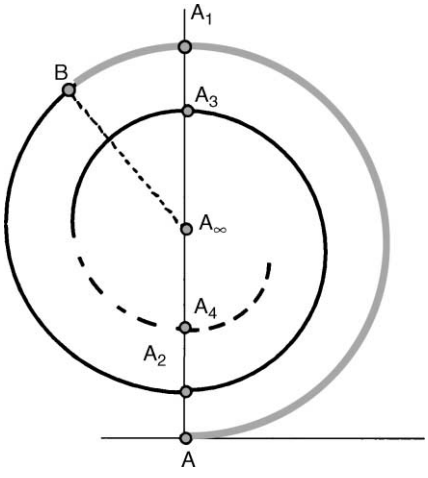


Figure 16 Vibration curve for opaque disk. The shaded region makes no contribution because it is associated with the portion of the wave obstructed by the opaque object. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Zone Plate

In the construction of Fresnel zones, each zone was assumed to produce a Huygens wavelet, out of phase with the wavelets produced by its nearest neighbors. If every other zone were blocked, then there would be no negative contributions to [37]. The intensity on-axis would be equal to the square of the sum of the amplitudes produced by the in-phase zones – exceeding the intensity of the incident wave. An optical component made by the obstruction of either the odd or the even zones could therefore be used to concentrate the energy in a light wave.

The boundaries of the opaque zones, used to block out-of-phase wavelets, are seen from [38] to increase as the square root of the integers. An array of opaque rings, constructed according to this prescription is called a zone plate, see Figure 17.

A zone plate will perform like a lens with focal length

$$f = \pm \frac{r_1^2}{\lambda} \qquad [40]$$

The zone plate, shown in Figure 17, will act as both a positive and negative lens. What we originally called the source can now be labeled the object point, O, and what we called the observation point can now be labeled the image point, I. The light passing through the zone plate is diffracted into two paths, labeled C and D in Figure 17. The light waves, labeled C, converge to a real image point I. For these waves the zone plate performs the role of a positive lens with a focal length given by the positive value of [40]. The light waves, labeled D in Figure 17, appear to originate from the virtual image point labeled $\mathbb{I}$. For these waves the zone plate performs the role of a negative lens with a focal length given by the negative value of [40].

The zone plate will not have a single focus, as is the case for a refractive optical element, but rather will have multiple foci. As we move toward the zone plate, from the first focus, given by [40], the effective Fresnel zones will decrease in diameter. The zone plate will no longer obstruct out-of-phase Fresnel zones and the light intensity on the axis will decrease. However, additional maxima, of the on-axis intensity, will be observed at values of Z for which the first zone plate opening contains an odd number of zones. These positions can also be labeled as foci of the zone plate; however, the intensity at each of these foci will be less than the intensity at the primary focus.

Lord Rayleigh suggested that an improvement of the zone plate design would result if, instead of blocking every other zone, we shifted the phase of alternate zones by 180°. The resulting zone plate, called a phase-reversal zone plate, would, more efficiently, utilize the incident light. R.W. Wood was the first to make such a zone plate. Holography provides an optical method of constructing the phase-reversal zone plates in the visible region of the spectrum. Semiconductor lithography has been used to produce zone plates in the x-ray region of the spectrum.

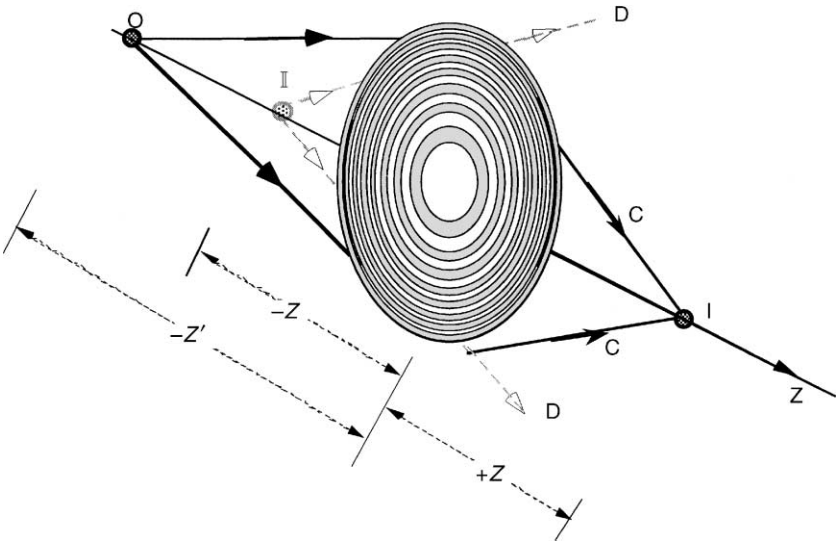


Figure 17 A zone plate acts as if it were both a positive and a negative lens. Light from the object, O, is diffracted into waves traveling in both the D and the C directions. The light traveling in the C direction produces a real image of O at I. The light traveling in the D direction appears to originate from a virtual image at I. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

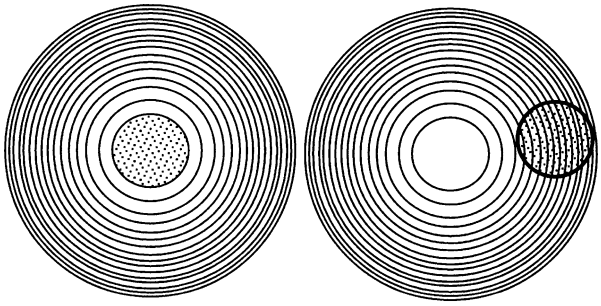


Figure 18 (a) A set of Fresnel zones has been constructed about the optical path taken by some light ray. If the optical path of the ray is varied over the cross-hatched area shown in the figure, then the optical path length does not change. This cross-hatched area is equal to the first Fresnel zone and is described as the neighborhood of the light ray. (b) The neighborhood defined in (a) is moved so that it surrounds an incorrect optical path for a light ray. We see that this region of space would contribute no wave amplitude at the observation point because of the destructive interference between the large number of partial zones contained in the neighborhood. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The resolving power of a zone plate is a function of the number of zones contained in the plate. When the number of zones exceeds 200, the zone plate's resolution approaches that of a refractive lens.

Fermat's Principle

The Fresnel zone construction provides physical insight into the interpretation of Fermat's principle which states that, if the optical path length of a light ray is varied in a neighborhood about the true path, there is no change in path length. By constructing a set of Fresnel zones about the optical path of a light ray, we can discover the physical meaning of a neighborhood.

The rules for constructing a Fresnel zone require that all rays passing through a given Fresnel zone have the same optical path length. The true path will pass through the center of the first Fresnel zone constructed on the actual wavefront. A neighborhood must be the area of the first Fresnel zone, for it is over this area that the optical path length does not change. Figure 18a shows the neighborhood, about the true optical path, as a cross-hatched region equal to the first Fresnel zone.

Light waves do travel over wrong paths but we do not observe them because the phase differences for those waves that travel over the 'wrong' paths are such that they destructively interfere. By moving the neighborhood, defined in the previous paragraph as the first Fresnel zone, to a region displaced from the true optical path, we can use the zone construction to see that this statement is correct. In Figure 18b the neighborhood is constructed about a ray that is improper according to Fermat's principle. We see that this region of space would contribute no energy at the observation point because of destructive interference between the large number of partial zones contained in the neighborhood.

This leads us to another interpretation of diffraction. If an obstruction is placed in the wavefront so that we block some of the normally interfering wavelets, we see light not predicted by Fermat's principle, i.e., we see diffraction.

Further Reading

Backer BB and Copson EJ (1969) *The Mathematical Theory of Huygens' Principle*. London: Oxford University Press.

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.

Guenther RD (1990) *Modern Optics*. New York: Wiley.

Haus HA (1984) *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall.

Hecht E (1998) *Optics*, 3rd edn. Reading, MA: Addison-Wesley.

Klein MV (1970) *Optics*. New York: Wiley.

Rubinowicz A (1984) *Tahe Miyamoto–Wolf diffraction wave*. In: Wolf E (ed.) Progress in Optics, vol. IV.

Sayanagi I (1967) Pinhole Imagery. *Journal of the Optical Society of America*. 57: 1091–1099.

Sears FW (1949) *Optics*. Reading, MA: Addison-Wesley.

Sommerfeld A (1964) *Optics*. New York: Academic Press.

Stegliani DJ, Mittra R and Semonin RG (1967) Resolving power of a zone plate. *Journal of the Optical Society of America* 57: 610–613.

Stone JM (1963) *Radiation and Optics*. New York: McGraw-Hill.

DIFFRACTIVE SYSTEMS

Contents

Aberration Correction with Diffractive Elements
Applications of Diffractive and Micro-Optics in Lithography
Design and Fabrication of Diffractive Optical Elements
Diffractive Laser Resonators
Diffractives in Animals
Microstructure Fibers
Omnidirectional Surfaces and Fibers
Wave Optical Modeling and Design

Aberration Correction with Diffractive Elements

N Lindlein, University of Erlangen-Nürnberg, Erlangen, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Diffractive optical elements (DOEs) can be used for the correction of aberrations in optical systems because of two fundamental properties:

1. The phase function Φ of a DOE can be chosen quite arbitrarily so that for a given wavelength λ the DOE can act like an aspheric optical element.
2. The dispersion of a DOE due to the local grating equation is very high and its sign is opposite to that of the dispersion of a refractive optical element with normal dispersion of the refractive index.

The first property allows the correction of monochromatic aberrations in an optical system similar to the case of using an aspheric surface in a refractive optical system. The second property allows the construction of hybrid achromatic optical elements that consist of a refractive lens and a diffractive lens whose optical powers have the same sign (e.g., two positive lenses for a positive achromatic element or two negative lenses for a negative achromatic element).

This article will first describe two methods for the simulation of DOEs by ray tracing because ray tracing is the basic simulation tool for aberration correction. Afterwards, the properties of DOEs for the correction of monochromatic and chromatic aberrations and some applications will be discussed. The name DOE will always mean either binary, or multiple-step or blazed surface relief diffractive optical elements which change only the phase of the incident wave and not the amplitude. All examples will be given for transmission elements but in principle reflective elements can also be described with quite similar formulas.

Ray Tracing Simulation of DOEs

First of all it seems straightforward to use a diffractive analysis to simulate a diffractive optical element. However, the global diffractive analysis of a general DOE with perhaps 1000 or more grating structures and locally changing grating period and especially the diffractive analysis of a complete optical system containing DOEs and other elements is so complex that it can be done only for very simple DOEs and systems. Therefore, the diffractive analysis of a DOE is mostly restricted to a local analysis using some periods (e.g., 10–50 periods) to calculate, e.g., the different local diffraction orders, their diffraction efficiencies and possibly local phase shifts. For the global analysis another simulation method has to be used.

The most important simulation method for the correction of aberrations in traditional optical systems is ray tracing. Therefore, ray tracing methods are also used for DOEs. These methods can calculate the path of a ray for a given diffraction order, but they cannot calculate the diffraction efficiency for this ray. For the calculation of the diffraction efficiency local diffractive analysis methods (scalar or rigorous methods depending on the local grating frequency and the angle of the incident ray) have to be used

in addition. In particular, for applications with a broad wavelength spectrum (e.g., an achromatic lens) or a broad spectrum of angles of the incident rays, the diffraction efficiency has to be analyzed carefully to obtain the energy distribution in the different diffraction orders. But this is beyond the scope of this article about aberration correction and we assume that we know the local diffraction efficiency for each ray.

Sweatt Model for the Ray Tracing Simulation of DOEs

An aspheric refractive surface with height h as a function of the lateral coordinates x and y and a refractive index n below the surface and air (refractive index 1) above the surface can be approximately replaced by a DOE with the phase function (see Figure 1)

$$\Phi(x, y) = 2\pi \frac{(n-1)h(x, y)}{\lambda} \qquad [1]$$

Here, λ is the wavelength of the light for which the DOE is designed and the phase function is a continuous representation of the structure of the DOE. To encode the phase function of the DOE the wrapped phase function $\Phi \bmod 2\pi$ has to be considered. The encoding itself can be done by building a blazed, a multiple-step or a binary structure (see Figure 1), whereby the first-order diffraction efficiency decreases from the blazed to the binary structure (at maximum 100% diffraction efficiency for the blazed structure and about 40% for the binary structure). The replacement of an aspheric refracting surface by a DOE using eqn [1] is only valid in the paraxial approximation since the phase function of the DOE is encoded in a plane whereas the refracting surface of the aspheric is typically curved. But, the agreement between the aspheric surface and the DOE with a phase function calculated with the help of eqn [1] increases with increasing refractive index n since then the curvature of the aspheric is quite small and the refraction also nearly takes place in a plane. Of course, the refractive index of real materials cannot exceed a value of more than about 3.5 (for silicon used in the near infrared). But, in a numerical simulation the refractive index can be arbitrarily high, e.g., several thousand. This is the reason why the so called Sweatt model for tracing rays through DOEs uses a weakly curved aspheric surface with height h_{sweatt} and a very high refractive index n_{sweatt} to simulate a diffractive phase function Φ:

$$h_{\text{sweatt}}(x, y) = \frac{\lambda}{n_{\text{sweatt}} - 1} \frac{\Phi(x, y)}{2\pi} \qquad [2]$$

To take into account the wavelength dispersion of the DOE the refractive index n_{sweatt} in the Sweatt model has to be proportional to the wavelength λ

$$n_{\text{sweatt}}(\lambda) = n_0 \frac{\lambda}{\lambda_0} \qquad [3]$$

where n_0 is the refractive index at the wavelength λ_0.

The validity of the Sweatt model increases with increasing refractive index and becomes nearly exact for $n_{\text{sweatt}} \to \infty$ (in practice values of several thousand for the refractive index are normally sufficient and a rule of thumb is to take the numerical value of the utilized wavelength specified either in nanometers for infrared light or angstroms for visible light). The Sweatt model was introduced as a means to simulate DOEs in ray tracing programs unable to implement DOEs directly but capable of simulating aspheric refractive surfaces. However, a disadvantage of this model is the numerical error generated by the finite value of the refractive index.

Modern ray tracing programs therefore use a more appropriate model based on the local grating equation.

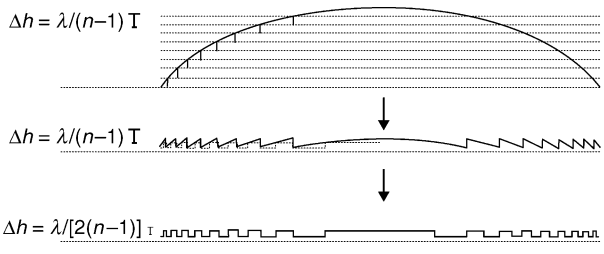


Figure 1 Transition from an aspheric refracting surface to a blazed diffractive optical element and finally to a binary diffractive optical element.

Local Grating Model for the Ray Tracing Simulation of DOEs

The basic equation for the simulation of DOEs by the local grating model is that the phase Φ_{out} of a diffracted ray in the plane of the DOE can be calculated from the phase Φ_{in} of the incident ray by adding the phase function Φ of the DOE multiplied by the diffraction order m:

$$\Phi_{\text{out}}(x, y, z_0) = \Phi_{\text{in}}(x, y, z_0) + m\Phi(x, y) \qquad [4]$$

Here, z_0 is the z-coordinate of the DOE plane which is assumed to be perpendicular to the z-axis (see Figure 2). The diffraction order m will be only one for an ideal blazed DOE but it can have an arbitrary integer value for binary DOEs. In practice, the optical designer will calculate the system for one diffraction order and analyze the system for other diffraction orders afterwards if necessary.

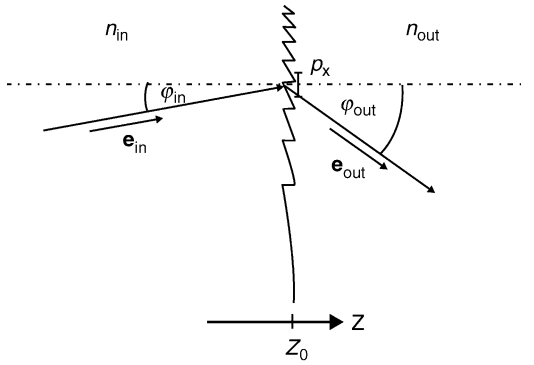


Figure 2 Schematic view in two-dimensions of the path of a ray which is diffracted at a DOE.

For the phases Φ_{in} of the incident and Φ_{out} of the diffracted ray the eikonal equation of geometrical optics is valid,

$$|\nabla\Phi_j|^2 = k_j^2 = \left(\frac{2\pi n_j}{\lambda}\right)^2 \qquad [5]$$

where the index j can have the two values $j \in \{\text{in}, \text{out}\}$. n_{in} or n_{out} are the refractive indices of the materials in which the incident or diffracted rays propagate. λ is the wavelength in vacuum. This form of the eikonal equation is obtained from the more common form $|\nabla L_j|^2 = n_j^2$, which is derived in any textbook of optics, by multiplying it with $(2\pi/\lambda)^2$ and using the definition of the optical phase $\Phi_j = 2\pi L_j/\lambda$. Here, L_j is the optical path length which is often called the eikonal. The curves of constant phase $\Phi_j(x, y, z) = \text{constant}$ are the wavefronts and the gradient $\nabla\Phi_j(x, y, z)$ points in the direction of the local ray which is perpendicular to the wavefront at the point (x, y, z) and describes its direction of propagation. Furthermore, the unit vectors $\mathbf{e}_{in}$ and $\mathbf{e}_{out}$ along the ray direction of the incident and diffracted ray, respectively, can be calculated by normalizing the gradient:

$$\mathbf{e}_j = \frac{\nabla\Phi_j}{|\nabla\Phi_j|} = \frac{\lambda}{2\pi n_j}\nabla\Phi_j\,; \qquad j \in (\text{in}, \text{out}) \qquad [6]$$

Since $\mathbf{e}_j$ is a unit vector, i.e., $e_{x,j}^2 + e_{y,j}^2 + e_{z,j}^2 = 1$, it can be seen from eqn [6] (and also from the original eqn [5]) that the z-component $e_{z,j}$ of the ray direction vector can be calculated (apart from its sign) if the two partial derivatives of the phase with respect to x and y are known. This means that it is sufficient to know the phase in a plane as stated in eqn [4].

By calculating the partial derivatives of eqn [4] with respect to x and y the equations for the ray tracing at a DOE are derived:

$$\begin{aligned}
\frac{\partial\Phi_{out}(x,y)}{\partial x} &= \frac{\partial\Phi_{in}(x,y)}{\partial x} + m\frac{\partial\Phi(x,y)}{\partial x} \\
&\Rightarrow n_{out}e_{x,out}(x,y) \\
&= n_{in}e_{x,in}(x,y) + m\frac{\lambda}{2\pi}\frac{\partial\Phi(x,y)}{\partial x} \\
\frac{\partial\Phi_{out}(x,y)}{\partial y} &= \frac{\partial\Phi_{in}(x,y)}{\partial y} + m\frac{\partial\Phi(x,y)}{\partial y} \\
&\Rightarrow n_{out}e_{y,out}(x,y) \\
&= n_{in}e_{y,in}(x,y) + m\frac{\lambda}{2\pi}\frac{\partial\Phi(x,y)}{\partial y}
\end{aligned} \qquad [7]$$

The local grating frequencies ν_x and ν_y in x and y direction are calculated from the phase function Φ by:

$$\begin{aligned}
\nu_x(x,y) &= \frac{1}{2\pi}\frac{\partial\Phi(x,y)}{\partial x} \\
\nu_y(x,y) &= \frac{1}{2\pi}\frac{\partial\Phi(x,y)}{\partial y}
\end{aligned} \qquad [8]$$

Therefore, given $\mathbf{e}_{in}$ and eqn [6], the ray direction vector $\mathbf{e}_{out}$ of the diffracted ray is

$$\begin{aligned}
n_{out}\mathbf{e}_{out} &= n_{out}\begin{pmatrix} e_{x,out} \\ e_{y,out} \\ e_{z,out} \end{pmatrix} \\
&= \begin{pmatrix} n_{in}e_{x,in} + m\lambda\nu_x \\ n_{in}e_{y,in} + m\lambda\nu_y \\ \sqrt{n_{out}^2 - (n_{in}e_{x,in} + m\lambda\nu_x)^2 - (n_{in}e_{y,in} + m\lambda\nu_y)^2} \end{pmatrix}
\end{aligned} \qquad [9]$$

The dependence of the lateral coordinates x and y is left out in eqn [9] to simplify notation. It can be easily checked that this equation fulfills the required condition for the unit vector: $e_{x,out}^2 + e_{y,out}^2 + e_{z,out}^2 = 1$.

As stated above, for a blazed DOE fabricated to the exact depth and illuminated by the design wavelength only the first diffraction order $m = 1$ is generated. But in practice other diffraction orders m are generated due to fabrication errors or illuminating with other wavelengths. In particular, by using binary DOEs there are inherently different diffraction orders. Therefore, it may be necessary to analyze the optical system for several values m one after the other. Each value m defines one wavefront generated by the DOE and the wavefront can be calculated by evaluating the optical path lengths along the rays at different points. In practice, the optical system has to be analyzed for those values m with sufficiently high

diffraction efficiency. Then, it can be checked whether undesired diffraction orders of the DOE are blocked in the optical system by aperture stops or whether they will disturb the output of the optical system. The local diffraction efficiency of each ray for a value m can be calculated to a good approximation by using a diffractive analysis of the corresponding local grating, i.e., that of an infinitely extended grating which would have the same grating period and the same structure shape (e.g., depth) as the DOE at the specific point.

If the y component of the incident ray and the local grating frequency ν_y are both zero the ray components in the x direction can be expressed via the sine of the angle of incidence φ_{in} or the angle of the diffracted ray φ_{out} (see **Figure 2**). In this case the first row of eqn [9] can be written as

$$n_{\text{out}} \sin \varphi_{\text{out}} = n_{\text{in}} \sin \varphi_{\text{in}} + m\frac{\lambda}{p_x} \qquad [10]$$

where the local grating period $p_x = 1/\nu_x$ has been introduced. This is the well-known grating equation.

In this article eqns [4]–[10] assume that the DOE is on a planar substrate and that x and y are the lateral coordinates on the substrate. However, the equations can also be written in a vector formulation which is valid for DOEs on curved substrates (see Further Reading).

So, we have now a powerful simulation method to design and analyze in the following several optical systems containing DOEs to correct monochromatic or chromatic aberrations.

Correction of Monochromatic Aberrations

For practical applications it is important to have an efficient algorithm to calculate the phase function of a DOE to correct the aberrations in a given optical system. Several different cases have to be distinguished. An easy case is the design of a DOE to produce a desired output wavefront Φ_{out} from a known input Φ_{in}. In this case, eqn [4] can be used to calculate the phase function Φ of the DOE directly. Two examples are given in the following. Afterwards, some aspects of a more general numerical optimization procedure are described.

Aberration Correction for Interferometrically Recorded Holograms

Although the following example of the aberration correction of interferometrically recorded holograms by using DOEs seems to be quite special it shows exemplarily several principal steps which are necessary to design a DOE for correcting monochromatic aberrations.

In this article the name DOE is used for diffractive elements with a microstructured surface (binary, multistep or blazed). Such elements have the property that their diffraction efficiency depends only weakly on the wavelength or the angle of incidence of the illuminating light. Another interesting class of diffractive elements are holograms in volume media, which are recorded interferometrically using two coherent recording waves. The refractive index in a volume hologram is modulated periodically and the thickness has to amount to several periods of the interference pattern along the z-direction in order to have a so-called thick hologram with a high diffraction efficiency. If there are only a few periods along the z-direction the quite small modulation of the refractive index (typically about 0.02) will result in a low diffraction efficiency. Therefore, if the angle between the two recording waves is small the resulting interference pattern will have large periods and only a few periods are recorded in the photosensitive layer along the z-direction. But this means that the holographic optical element (HOE) can only have a small diffraction efficiency. From this fact, it is clear that also by using the recorded HOE the deflection angle between the incident and the diffracted wave must not be smaller than a certain value because otherwise the thickness has to be very big or the diffraction efficiency decreases. Therefore, such elements with high diffraction efficiency, which will be named in the following holographic optical elements (HOE), are always off-axis elements because for on-axis elements all rays in the neighborhood of the axis are only deflected by a small angle and then the diffraction efficiency would be low. For the correct wavelength and angle of incidence, so that the Bragg condition of holography is fulfilled, HOEs, like blazed DOEs, exhibit only one dominant diffraction order $m = m_{\text{dom}}$ (normally $m_{\text{dom}} = 1$). But, if the angle of incidence or the wavelength change, the diffraction efficiency in the order m_{dom} decreases rapidly and light passes through the HOE without being deflected, i.e., most of the light is then in the order $m = 0$. For this reason, a HOE can be used for example for a head-up display to fade in some information under the Bragg angle without hindering the normal view onto the environment.

Traditional materials for HOEs like dichromated gelatin (DCG) are sensitive only in the blue–green spectral region (e.g., for light of an argon ion laser with $\lambda = 488$ nm). However, once recorded it will be illuminated in many applications (e.g., optical interconnects) by a wavelength in the red–infrared spectral region. Unfortunately, the wavelength mismatch between recording and reconstruction illumination introduces aberrations.

If the two desired reconstruction waves at the wavelength λ_2 have the phases $\Phi_{r,\lambda 2}$ and $\Phi_{o,\lambda 2}$ in the surface plane of the HOE (the subscripts r and o refer to interferometry reference and object, respectively), the phase function Φ of the HOE which has to be recorded is

$$\Phi(x, y) = \Phi_{o,\lambda 2}(x, y) - \Phi_{r,\lambda 2}(x, y) \qquad [11]$$

If our discussion here is restricted to holographic lenses, the two reconstruction waves will be spherical (or plane) waves (see Figure 3b) with the center of curvature at the points $\mathbf{r}_{r,\lambda 2}$ and $\mathbf{r}_{o,\lambda 2}$ and eqn [11] is:

$$\Phi(x,\ y) = \frac{2\pi\sigma_{o,\lambda 2}}{\lambda_2}\sqrt{(x - x_{o,\lambda 2})^2 + (y - y_{o,\lambda 2})^2 + z_{o,\lambda 2}^2} - \frac{2\pi\sigma_{r,\lambda 2}}{\lambda_2}\sqrt{(x - x_{r,\lambda 2})^2 + (y - y_{r,\lambda 2})^2 + z_{r,\lambda 2}^2} \qquad [12]$$

The parameters $\sigma_{o,\lambda 2}$ and $\sigma_{r,\lambda 2}$ are either $+1$ for a divergent spherical wave or -1 for a convergent spherical wave. The surface of the HOE itself is in the plane at $z = 0$.

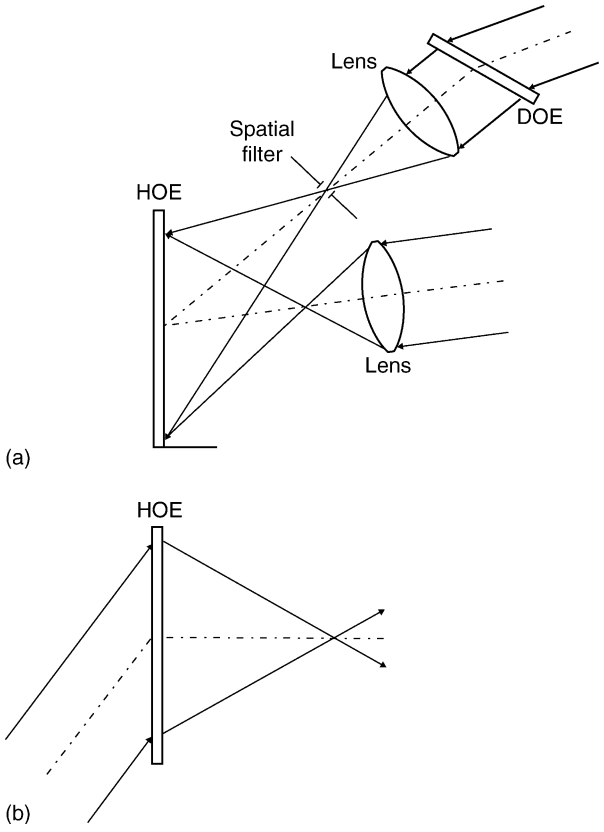


Figure 3 Schematic view of the (a) recording and (b) reconstruction setup for a holographic lens. The recording is made at wavelength λ_1 and the reconstruction at wavelength λ_2.

The task is now to find two recording waves which construct nearly the same three-dimensional interference pattern in the hologram at the recording wavelength λ_1. This means that eqn [12] has to be fulfilled exactly for the difference of the phases of the two recording waves in the plane $z = 0$ and, additionally, the Bragg condition has to be fulfilled at least approximately. So, the next step is to calculate one of the recording waves with phase $\Phi_{r,\lambda 1}$ so that the Bragg condition is fulfilled. This can be done approximately with a third-order theory resulting in a spherical wave or nearly exactly with an iterative numerical procedure resulting generally in an aspheric recording wave. If $\Phi_{r,\lambda 1}$ is known, the other recording wave with phase $\Phi_{o,\lambda 1}$ is calculated by

$$\Phi_{o,\lambda 1}(x, y) = \Phi(x, y) + \Phi_{r,\lambda 1}(x, y) \qquad [13]$$

At least this recording wave will be aspheric. To generate such an aspheric wave it is traced by ray tracing to a plane where the two recording waves are spatially separated. There, either a DOE or a combination of a DOE and a lens are used to generate the aspheric wave. The calculations of the phase function of the DOE are done again by using an equation like [11] whereby one wave is the propagated aspheric wave (possibly propagated also through a lens) and the other is normally a plane off-axis wave with which the DOE will be illuminated in the real recording setup (see Figure 3a). Here, the off-axis angle is necessary to allow spatial filtering of disturbing diffraction orders of the DOE which will be normally fabricated as a binary DOE.

Experiments of our own showed (see the literature) that nearly diffraction-limited holographic lenses with a numerical aperture of NA = 0.4 can be fabricated by correcting the aberrations of about 28 wavelengths peak-to-valley (for the reconstruction at 633 nm wavelength) due to the wavelength mismatch between recording at 488 nm wavelength and reconstruction at 633 nm wavelength by using a DOE in the recording setup.

Testing of Aspheric Surfaces by Using a DOE as Aberration Compensating Element

An important application for DOEs is testing aspheric surfaces and nowadays more and more companies use DOEs to test aspheric surfaces which are used for example in the objectives of lithographic wafer steppers to produce highly integrated microelectronic circuits.

In this testing setup the DOE is illuminated by a wave with phase Φ_{in} to generate a wave with phase Φ_{out} that is incident perpendicular to the aspheric surface (see Figure 4). After reflection at the asphere

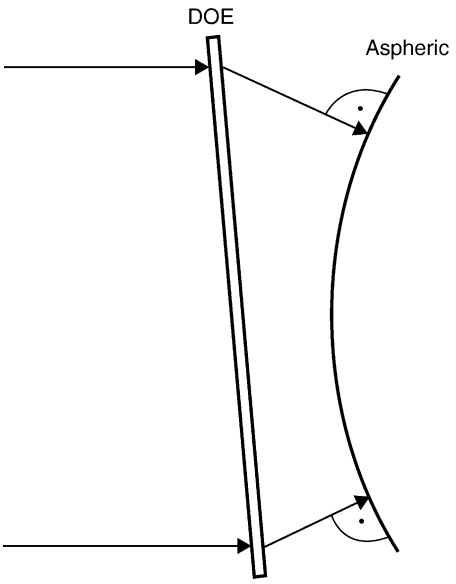


Figure 4 Scheme showing the basic elements for testing an aspheric surface.

the wave is retro-reflected and takes the same path if the asphere has no surface deviations and no misalignments. This means that the setup is completely symmetric for a reflection at the aspheric surface. Therefore, the optical path lengths have to be identical for all rays on the asphere. To calculate the phase function of the DOE the following steps are made:

- A grid of N rays starting perpendicularly on the aspheric surface and having all phase zero (due to the fact that the phase has to be identical it can be set to zero) is traced to the plane of the DOE. N depends on the shape and numerical aperture of the asphere and the grid will be in most cases two dimensional and regular or one dimensional (along one radial section of the asphere) and regular if the asphere and the system are completely rotationally symmetric. As a rule of thumb the number N of rays has to be more than twice the number of polynomial coefficients if the asphere is described by a polynomial.
- The phase of these rays is the function $-\Phi_{\text{out}}$ (the negative sign is necessary because the rays take the opposite direction) and is now known at some points (x_i, y_i) with $i \in \{1, 2, \ldots, N\}$.
- The phase Φ_{in} of the incident wave which is in practice either a plane wave or a spherical wave is calculated in the plane of the DOE. If the incident wave is generated by a lens having aberrations it is again calculated by ray tracing, otherwise it can be calculated analytically.
- Now, the phase function Φ of the DOE can in principle be calculated with the help of eqn [4].

However, the phase Φ_{out} (and possibly also Φ_{in}) is known only at some points (x_i, y_i), but it has to be known at all points of the DOE. Therefore in practice an analytical representation of the phase functions of the incident and output wave has to be calculated. This can be done by fitting a polynomial to the phase values using a least squares fit. In this case the coordinates, at which the phase values are known, can be quite arbitrary as long as the sampling is dense enough. For a rotationally symmetric system the polynomial should also be rotationally symmetric, i.e., a polynomial P_{rot} depending only on the radius coordinate r

$$r = \sqrt{x^2 + y^2} \qquad [14]$$

$$P_{\text{rot}}(r) = \sum_{j=0}^{G} a_j r^j \qquad \text{or} \qquad P_{\text{rot}}(r) = \sum_{j=0}^{G/2} a_{2j} r^{2j} \qquad [15]$$

where G is the degree of the polynomial, i.e., the highest power in r. In these two cases the number N of sampling points lying along one radial section of the rotationally symmetric optical system should be larger than $2G$ or G, respectively.

For general non-symmetric systems a Cartesian polynomial P_{cart} can be used, e.g.,

$$P_{\text{cart}}(x, y) = \sum_{j=0}^{G} \sum_{k=0}^{G-j} a_{jk} x^j y^k \qquad [16]$$

Here, the number of coefficients of the polynomial is $(G + 1)(G + 2)/2$ and the number N of sampling points should again be larger than twice the number of coefficients, i.e., $N > (G + 1)(G + 2)$.

For both types of polynomials the coefficients a_j or a_{jk} have to be calculated so that the merit function M

$$M(\mathbf{a}) = \sum_{i=1}^{N} (P(x_i, y_i) - \Phi_j(x_i, y_i))^2; \quad j \in \{\text{in, out}\} \qquad [17]$$

is minimized. The symbol P means either P_{rot} or P_{cart}. The vector $\mathbf{a}$ symbolizes the coefficients $\mathbf{a}_j$ or $\mathbf{a}_{jk}$. By differentiating M with respect to all coefficients and setting the partial derivatives to zero a system of linear equations is obtained which can be solved by standard procedures. As mentioned before, the number N of independent sampling points has to be at least double the size of the number of coefficients of the polynomial. Otherwise, the polynomial can 'swing' between the sampling points and the polynomial fits well to the desired phase function only at the sampling points but not at other points in between.

For very irregular aspherics with a fast change of the local curvature it might by necessary not to take a

global polynomial fit but a spline fit, i.e., a set of adjacent local polynomials with several continuity conditions at the border area of each local polynomial. In this case the total number of coefficients used to describe the phase functions and Φ_{in} and Φ_{out} of the DOE will be increased compared to the description by a global polynomial. Then, the number N of rays for the calculation has also to be increased accordingly.

So, in this example the DOE acts as a null-compensator for the asphere and at first sight it is not a typical example for aberration correction. However, by regarding the complete system the DOE compensates the aberrations, i.e., deviations from a spherical wave, generated by the asphere which are primary and higher-order spherical aberrations if the asphere is rotationally symmetric and more general aberrations otherwise. This example shows how the phase function of a DOE can be designed straightforwardly for correcting aberrations if there is only one well-defined wavefront propagating through the optical system. This design method can also be used for other optical systems where the input and the output wavefront of the DOE are well defined.

Numerical Optimization of DOEs in Optical Systems

In a general optical system, where a DOE is not only used to compensate the aberrations for one specific incident wave but for different incident waves (optical system with a field), it is not possible to calculate the optimum phase function in such a straightforward way as given above. However, a DOE is in principle similar to an asphere in an optical system, so that the design rules for including aspherics in optical systems can be used in some cases as long as monochromatic aberrations are considered. Nevertheless, one should keep in mind that a DOE is, according to the Sweatt model, an asphere with high refractive index and small height, i.e., the deflection of the light rays occurs in a plane (or on a curved surface if the DOE is situated on a curved substrate). Therefore, the design rules for aspheric surfaces can be used only if the high refractive index of the Sweatt model is used in the design.

In the general case only a numerical optimization can be used. For this, it is very important to select an appropriate mathematical representation of the phase function of the DOE.

For rotationally symmetric optical systems also the phase function of the DOE should be rotationally symmetric. Therefore, in such a case it is useful to write the phase function Φ as a polynomial in the radius coordinate r as given in eqns [14] and [15]. Then, the $(G + 1)$ or $(G/2 + 1)$ coefficients a_j have to be optimized, whereby a_2 is proportional to the focal power of the DOE, whereas the other coefficients represent aspheric terms. Depending on the numerical aperture and on the application the degree G of the polynomial can be increased in order to achieve a better solution. Nevertheless, the degree in the optimization should be as small as necessary in order to avoid 'swinging effects' and numerical instabilities of the polynomial. To determine the smallest adequate degree an iterative method has to be used by starting with a quite small degree (e.g., 4). If the result of the optimization is not good enough the degree has to be increased until the result of the optimization is satisfying.

For non-rotationally symmetric optical systems a Cartesian polynomial like in eqn [16] can be taken and all the $(G + 2)(G + 1)/2$ coefficients a_{jk} have to be optimized. Since the number of free parameters is in this case larger than the number of free parameters in the rotationally symmetric case the convergence of the optimization may take a long time and the risk to break down in a local minimum of the multidimensional parameter space increases.

In principle, the calculation of a DOE used, e.g., for testing an aspheric surface can also be done by optimizing the coefficients of the polynomial representing the phase function of the DOE. Nevertheless, as a general rule it can be said that a direct calculation of the phase function is always more efficient than a numerical optimization, if the direct calculation is possible. In practice, it can be seen that the maximum degree of a polynomial, which is directly calculated by the method described in the last section, is always smaller than the degree which is necessary in a numerical optimization to achieve convergence of the optimization.

Of course, in such cases where several different types of aberrations (on-axis aberrations like spherical aberration, off-axis aberrations like astigmatism and coma or field aberrations like field curvature or distortion) have to be corrected in an optical system a direct calculation of the phase function is not possible so that only the numerical optimization can provide a solution.

Up to now only monochromatic aberrations have been discussed. But as stated at the beginning of this article the strong dispersion of a DOE having opposite sign to the dispersion of a refractive element offers additional applications of DOEs by correcting chromatic aberrations. The principle of doing this is described in the following.

Correction of Chromatic Aberrations

The focal power $1/f$ (f is the focal length of the lens) of a thin refractive lens consisting of two spherical

surfaces with radii of curvature R_1 and R_2 and made of a material (glass) with refractive index n is

$$\frac{1}{f(\lambda)} = (n(\lambda) - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \qquad [18]$$

where λ is the wavelength of the illuminating light.

This equation is valid for all wavelengths and so also for the commonly used wavelength $\lambda_d = 587.6$ nm (yellow d-line of helium) which is near the center of the visible spectrum. Then $1/R_1 - 1/R_2$ can be expressed as $1/((n_d - 1)f_d)$ and the focal power depends on the wavelength as

$$\frac{1}{f(\lambda)} = \frac{n(\lambda) - 1}{n_d - 1}\frac{1}{f_d} \qquad [19]$$

where n_d and f_d are the refractive index and the focal length at $\lambda_d = 587.6$ nm, respectively. So, by combining two thin lenses made of materials with refractive indices n_1 and n_2 and focal lengths f_1 and f_2 the focal power of the doublet assuming a zero distance between the two lenses is:

$$\begin{aligned}\frac{1}{f_{ges}(\lambda)} &= \frac{1}{f_1(\lambda)} + \frac{1}{f_2(\lambda)} \\ &= \frac{n_1(\lambda) - 1}{n_{1,d} - 1}\frac{1}{f_{1,d}} + \frac{n_2(\lambda) - 1}{n_{2,d} - 1}\frac{1}{f_{2,d}}\end{aligned} \qquad [20]$$

For an achromatic doublet the focal power should be equal for two different wavelengths which are commonly the wavelengths $\lambda_F = 486.1$ nm (blue F-line of hydrogen) and $\lambda_C = 656.3$ nm (red C-line of hydrogen) which mark the limits of the well visible spectrum (of course the visible spectrum itself is broader but the sensitivity of the human eye to wavelengths below λ_F or beyond λ_C is quite bad). Therefore, the following equations result:

$$\begin{aligned}\frac{1}{f_{ges}(\lambda_F)} = \frac{1}{f_{ges}(\lambda_C)} &\Rightarrow \frac{n_1(\lambda_F) - n_1(\lambda_C)}{n_{1,d} - 1}\frac{1}{f_{1,d}} \\ &= -\frac{n_2(\lambda_F) - n_2(\lambda_C)}{n_{2,d} - 1}\frac{1}{f_{2,d}}\end{aligned} \qquad [21]$$

By introducing the Abbe number V_d

$$\begin{aligned}V_d &= \frac{n_d - 1}{n_F - n_C} \\ &= \frac{n(\lambda_d = 587.6\text{ nm}) - 1}{n(\lambda_F = 486.1\text{ nm}) - n(\lambda_C = 656.3\text{ nm})}\end{aligned} \qquad [22]$$

equation [21] can be simplified to

$$V_{1,d}f_{1,d} = -V_{2,d}f_{2,d} \qquad [23]$$

Now, the wavelength dispersion of all conventional materials (glasses, polymers, etc.) is such that the refractive index decreases with increasing wavelength, i.e. the material has a so-called normal dispersion. This means that the Abbe number V_d is always positive and ranges from about 20 for high-dispersive glasses (e.g., SF59) to more than about 80 for low-dispersive glasses (e.g. Ultran 20). So, to fabricate a refractive achromatic doublet with a positive focal power it is necessary to take a positive lens with focal length f_1 made of a low-dispersive glass with a large Abbe number $V_{1,d}$ and a negative lens with focal length f_2 made of a high-dispersive glass with a small Abbe number $V_{2,d}$ (see **Figure 5a**). Therefore, the resulting focal power of the achromatic doublet

$$\frac{1}{f_{ges,d}} = \frac{1}{f_{1,d}} + \frac{1}{f_{2,d}} = \frac{1}{f_{1,d}}\left[1 - \frac{V_{2,d}}{V_{1,d}}\right] \qquad [24]$$

is always smaller than the focal power of the positive lens with focal power $1/f_{1,d}$ alone.

The situation is completely different if a refractive and a diffractive lens are used. From the Sweatt model

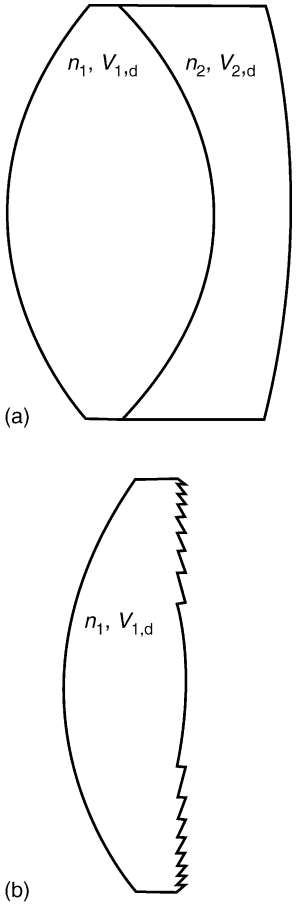


Figure 5 Structure of achromats: (a) refractive achromatic doublet consisting of a positive lens with low dispersion and a negative lens with high dispersion; (b) hybrid achromatic lens consisting of a positive refractive lens and a positive diffractive lens.

of a DOE (eqn [3]) it is known that the refractive index n_{sweatt} is proportional to the wavelength λ

$$n_{\text{sweatt}}(\lambda) = \frac{n_0}{\lambda_0}\lambda = \frac{n_d}{\lambda_d}\lambda \qquad [25]$$

Therefore, the Abbe number of a DOE is

$$\begin{aligned} V_d &= \frac{n_d - 1}{n_F - n_C} \\ &= \frac{\lambda_d}{\lambda_F - \lambda_C}\left(1 - \frac{1}{n_d}\right) \overset{n_d \to \infty}{=} \frac{\lambda_d}{\lambda_F - \lambda_C} \\ &= -3.452 \end{aligned} \qquad [26]$$

Here, we used the fact that the accuracy of the Sweatt model increases with increasing refractive index so that the Abbe number of a DOE is achieved in the limit of an infinite refractive index. So, it can be seen that the Abbe number of a DOE is a small negative constant. This means that the dispersion of a DOE is high and opposite in sign to that of a refractive element. Therefore, according to eqns [23] and [26] a positive refractive lens with a large focal power (i.e., small focal length) and a positive diffractive lens with a small focal power (i.e., long focal length) can be combined to form an achromatic doublet with a focal power larger than the focal power of one of the two lenses alone (see Figure 5b). By using eqns [19] and [25] the focal power of a diffractive lens can be calculated with the Sweatt model (n very big $\rightarrow n - 1 \approx n$):

$$\frac{1}{f_{\text{DOE}}(\lambda)} = \frac{\lambda}{\lambda_d f_d} \qquad [27]$$

In practice, the lenses are neither thin nor is the distance between them zero. In this case the equations for the design of a hybrid achromatic doublet are more complex (see Further Reading). Nevertheless, the given equations can be used for a first design and to understand the principal behavior.

Another advantage of a hybrid achromatic doublet is that the DOE can in addition behave like an asphere. Therefore, e.g., the spherical aberration can be corrected with the aspheric terms of the DOE provided that the spherical aberration is considerably smaller than the phase term encoded in the DOE to correct the chromatic aberrations. Otherwise, the correction of the monochromatic (spherical) aberration will disturb the correction of the chromatic aberrations. A comparison of a simple hybrid achromat, consisting of a planoconvex lens with the diffractive structure made directly onto the plane surface of the refractive lens and a conventional refractive achromatic doublet shows that the optical performance of the hybrid achromat is a little bit worse than that of the refractive achromatic doublet regarding, for example, off-axis aberrations. A simple explanation of this is that a refractive achromatic doublet has three spherical surfaces and two glasses, i.e., in total five parameters, whereas a hybrid achromat has one spherical surface, one glass and the DOE with a paraxial term and an aspheric term, i.e. in total about four parameters. Of course, the merit of these different types of parameters is not identical so that this assessment is only qualitative.

In the case of hybrid achromatic elements for a broad wavelength range (e.g., for the whole visible wavelength range) the dependence of the diffraction efficiency on the wavelength may be a limiting factor because the other diffraction orders behave like stray light and reduce the maximum contrast of the image. Therefore, special laminated blazed diffractive optical elements made of laminae of two or more different materials are described in the literature which show a nearly achromatic behavior of the diffraction efficiency but the 'normal' dispersion of DOEs. DOEs are in the meantime used in zoom lenses (see the cited patent in Further Reading) to correct chromatic aberrations.

Summary

In this article the possibilities of correcting monochromatic and chromatic aberrations in optical systems by using diffractive optical elements have been discussed. There are several advantages and disadvantages of using DOEs.

Advantages:

- DOEs offer many degrees of freedom in the design. By using modern lithographic fabrication techniques the fabrication of a DOE with a strongly aspheric phase function is just as expensive as fabricating a simple Fresnel zone lens with a quadratic phase function. So, DOEs can be used in some cases instead of aspheric surfaces which are more expensive. In addition, the lithographic fabrication technique guarantees a quite good accuracy concerning the phase function of the DOE compared to the more difficult fabrication of well-defined aspheric surfaces.
- The strong chromatic dispersion with the opposite sign to that of a refractive element allows the correction of chromatic aberrations. DOEs also give in this case an additional degree of freedom in the design because aspheric terms of the phase function can correct in addition monochromatic aberrations of the system as long as these aberrations are considerably smaller than the quadratic term of the phase function which corrects the chromatic aberrations.

Disadvantages:

- DOEs often exhibit not only one diffraction order but several diffraction orders due to fabrication errors, a broad spectrum of application wavelengths or because some DOEs naturally have several diffraction orders like, for example, binary DOEs. This can produce disturbing light in the optical system and limit the application.
- The strong dispersion can restrict the application in monochromatic systems because it requires in some cases (e.g., for testing aspheric surfaces) not only a relative constancy, i.e., constant but the absolute value does not have to be hit exactly, of the illuminating wavelength but an absolute constancy because otherwise systematic aberrations are introduced by the DOE.

So, there are many advantages of DOEs compared to other optical elements but on the other hand a DOE is not a magic bullet in correcting aberrations in optical systems. Depending on the system a careful assessment of the advantages and disadvantages has to be performed in order to find the best solution.

List of Units and Nomenclature

Abbe number	real number	V_d
Angle	[rad]	φ
Cartesian coordinates	[m]	x, y, z
Coefficients of a polynomial	different negative powers of m	a
Components of the ray vector	real numbers	e_x, e_y, e_z
Degree of a polynomial	integer number	G
Diffraction order	integer number	m
Focal length	[m]	f
Grating frequencies in x and y direction	[m^{-1}]	ν_x, ν_y
Height of a lens surface	[m]	h
Local grating periods in x and y direction	[m]	p_x, p_y
Merit function	real valued function	M
Modulus of the wave vector	[m^{-1}]	k
Nabla differential operator	[m^{-1}]	∇ $\begin{pmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{pmatrix}$
Number of sampling points/rays	integer number	N
Optical phase, phase function	[rad]	Φ
Polynomial	real valued function	P
Radial coordinate	[m]	r
Refractive index	real number	n
Sign variable	either +1 or −1	σ
Symbol for partial differentiation		∂
Unit vector in ray direction	vector with three real components	$\mathbf{e}$
Wavelength	[m]	λ

See also

Geometrical Optics: Aberrations.

Further Reading

Born M and Wolf E (1997) *Principles of Optics*, 6th edn. Cambridge, UK: Cambridge University Press.

Fairchild RC and Fienup JR (1982) Computer-originated aspheric holographic optical elements. *Optical Engineering* 21: 133–140.

Falkenstörfer O, Lindlein N, Keinonen T and Schwider J (1996) Diffraction-limited holographic lenses in dichromated gelatine. *Optical Engineering* 35: 2026–2033.

Herzig HP (1997) *Micro-Optics*. London: Taylor & Francis.

Kogelnik H (1969) Coupled wave theory for thick hologram gratings. *The Bell System Technology Journal* 48: 2909–2947.

Kress B and Meyrueis P (2000) *Digital Diffractive Optics*. Chicester, UK: Wiley.

Latta JN (1971) Computer-based analysis of holography using ray tracing. *Applied Optics* 10: 2698–2710.

Lindlein N, Schwider J, Schrader M, *et al.* (1993) Wavelength mismatch correcting elements for volume holograms written on a laser pattern generator. *Journal of Modern Optics* 40: 647–661.

MacGovern AJ and Wyant JC (1971) Computer generated holograms for testing optical elements. *Applied Optics* 10: 619–624.

Meier RW (1965) Magnification and third-order aberrations in holography. *Journal of the Optical Society of America* 55: 987–992.

Nishio A (1999) Zoom lens with diffractive optical elements. United States Patent 5978153.

Offner A (1966) Ray tracing through a holographic system. *Journal of the Optical Society of America* 56: 1509–1512.

Sinzinger S and Jahns J (1999) *Microoptics*. Weinheim: Wiley-VCH.

Stone Th and George N (1988) Hybrid diffractive-refractive lenses and achromats. *Applied Optics* 27: 2960–2971.
Sweatt WC (1977) Describing holographic optical elements as lenses. *Journal of the Optical Society of America* 67: 803–808.
Upatnieks J, Vander Lugt A and Leith E (1966) Correction of lens aberrations by means of holograms. *Applied Optics* 5: 589–593.
Welford WT (1975) A vector raytracing equation for hologram lenses of arbitrary shape. *Optical Communications* 14: 322–323.
Winick K (1982) Designing efficient aberration-free holographic lenses in the presence of a construction-reconstruction wavelength shift. *Journal of the Optical Society of America* 72: 143–148.

Applications of Diffractive and Micro-Optics in Lithography

V Kettunen, Heptagon, Zurich, Switzerland

H P Herzig, University of Neuchâtel, Neuchâtel, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Most aspects of modern life are thoroughly influenced by the micro-electronics industry through microchips, such as microprocessors and memory chips. Optical lithography has been a key factor in the rapid pace of integration that drives the development of the micro-electronics industry. The photolithography process not only facilitates the fabrication of the chips, but also directly influences the limits of the obtainable chip performance in terms of speed or storage capacity, through the minimum feature size that can be created in the lithography steps. The increasing resolution demands of the IC industry, outlined in the International Technology Roadmap for Semiconductors as technology nodes, mean continuing demand for improvements in the photolithography systems.

The challenges to the improvements are encapsulated in the equation for the resolution of an optical lithography system:

$$R = k_1 \lambda / NA \qquad [1]$$

which expresses the resolution R in terms of the smallest resolvable half-pitch as a function of the wavelength λ, the numerical aperture NA of the exposure system and the unit less constant k_1, which is used as a measure of lithographic difficulty, with smaller values indicating a more challenging process. Traditionally, the increase in the resolution has been obtained by decreasing the used wavelength, with the current mark for volume production systems currently standing at 193 nm. Within each wavelength, further reduction has been traditionally obtained by increasing the numerical aperture towards the mathematical limit of 1. Finally, several so-called Resolution Enhancement Techniques (RET) have been introduced to reduce k_1, in order to reach subwavelength technology nodes within each lithography wavelength generation. It is in connection with these techniques that micro-optics has found a role in modern lithography systems.

In this chapter we discuss the ways micro-optics are being utilized in lithography systems. We will start with a brief overview of conventional lithography systems, of the essential parts of such systems as well as the special resolution enhancement techniques utilized in the systems. We will then discuss the utilization of micro-optics within the context of lithography systems, along with the unique challenges in terms of design and fabrication that are associated with such applications. Finally, we will briefly discuss nonconventional lithography applications, especially those based on utilization of micro-optics. It should be noted that the motivation of this article is to describe the issues related to the use of micro-optical elements in a modern lithography system and is not intended as a review of the modern lithography systems, nor is it a review of micro-optics.

Overview of Conventional Lithography Systems

Photolithography is a process whereby a specified pattern or mask is transferred onto a wafer using photons. In the process a photosensitive material, commonly known as a resist, is initially used to record the pattern that is generated when the mask is illuminated. After exposure, the resist is developed. The resist is referred to as positive or negative, depending on whether it is removed or remains, respectively, after the development of the irradiated regions. After its development, the pattern in the resist is then transferred onto the wafer using suitable chemical processes, such as wet chemical etching or dry plasma etching.

Several different lithographical techniques exist. The oldest and simplest is contact printing, where a

photomask is brought into contact with a layer of photoresist on the wafer. Although this technique can reproduce high-resolution features, the contact between the wafer and the mask can lead to problems with cross contamination, and can damage either the wafer or the mask. Contact printing is mainly used nowadays in laboratory environments for small series of photolithography steps. A slightly more complicated technique is proximity printing, in which the photomask is held in close proximity to the surface of the wafer without actual contact. This avoids the problems of contamination and damage, but due to diffraction effects, reduces the maximum resolution of the process. Nevertheless, proximity printing remains a valuable technique for lithography applications, that do not require the highest possible resolution.

Modern volume production photolithography systems are based on the use of a projection system to image a mask onto a wafer through a complex system of lenses. There exist several variations of the basic idea. In some systems the entire mask is imaged at once onto the wafer. This approach works only when the feature size on the mask is reasonably large, in the range of a few microns. Reproducing smaller features with this approach is generally not possible, as the demands on the optical quality of the projection system, in the form of allowed aberrations, become extremely difficult to meet, especially when the physical size of the optical components increases with the reducing feature size. Therefore, so-called wafer steppers are used in modern microchip fabrication. The basic idea of a wafer stepper is to image only a small region of the mask onto the wafer, and pattern the entire surface area in consecutive steps.

The centerpiece of a wafer stepper is the exposure system. Its task is to image desired patterns from the photomask to their proper positions on the surface of the wafer. In many systems this process includes a size reduction between the patterns on the photomask and on the wafer. This reduces the dimensional accuracy requirements, such as feature size and positioning for the mask fabrication by whatever is the de-magnification of the lithography system. However, it also increases the complexity of the exposure system, as significance of aberrations is increased. The requirements of microchip production mean that the image of the photomask with submicron features must be reproduced on the wafer with a dimensional accuracy of only a few tens of nanometers and aligned to a specific position within a fraction of the linewidth. Furthermore, these tolerances must be guaranteed over the entire exposure field that is typically several square centimeters. It is not surprising that the resulting complex piece of optics costs several millions dollars.

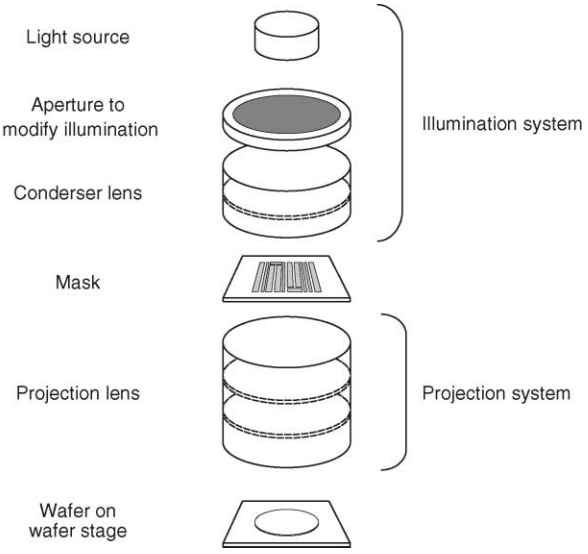


Figure 1 Schematic illustration of the exposure system.

The exposure system consists of a projection lens, illumination system, and wafer management system, as illustrated schematically in Figure 1.

A typical projection lens is a complicated lens system. As the resolution of a lens system is inversely proportional to the diameter of the lenses, the lenses in photolithographic systems can have diameters up to 250 mm, and the whole system is arranged inside a massive barrel-frame. The aim of the projection lens in the entire exposure system is to image the pattern from the photomask to the wafer. The projection lens is primarily responsible for the resolution and must therefore fulfill the highest optical requirements. To produce the smallest possible image, diffraction effects alone must essentially limit the resolution of the system. This means that the optical aberrations over the exposure field of several square centimeters must be virtually eliminated, by the wavefront deviations being less than a fraction of the wavelength.

The illumination system transfers the illumination source through the photomask to the entrance of the projection system. The main requirements for the illumination system are to collect most of the light into the system and to ensure that the irradiance of the photomask is uniform to within a few percent. In addition, resolution enhancement techniques, based on off-axis illumination, require that the illumination pattern on the photomask has a specific shape. We will discuss these requirements later in more detail. In general terms, the illumination system is less complicated than the projection system, especially with regard to the elimination of aberrations, due to the fact that precise imaging is not required in the illumination system.

Although the wafer management system is outside the scope of this article, we note in passing that it maintains the position, orientation, and movement of the wafers through the exposure system. Again, high precision is required to ensure the successful exposure of the desired pattern to the wafer.

Resolution Enhancement Techniques

In most modern lithography systems the achievable maximum resolution is increased via the use of one or more special resolution enhancement techniques (RET). These are essentially creative refinements of the lithography technology, based on theoretical optics papers published often decades before their first application in the lithography process in the late 1980s and early 1990s. There are essentially three RET approaches: Off-Axis Illumination (OAI), Optical Proximity Correction (OPC), and Phase Shifted Masks (PSM).

In optical proximity correction the basic idea is to predistort the mask pattern in order to overcome the proximity effects caused by the optics and the different processes. This is done either using a set of pre-existing rules describing the connection between a mask pattern and its realization on the wafer, using advanced models to predict the proximity effects during the lithography process, or with some combination of both. The key difficulties with OPC are related to the fabrication of the predistorted mask and to the accuracy of the used distortion rules or proximity effect models. Finally, it should be noted that OPC is strictly speaking not a true resolution enhancement technique, as it only enhances the effective resolution and not the theoretical resolution limit of the process defined by the optics.

In phase shifted masks the improvement in the resolution is obtained by altering the phase of light passing through different portions of the mask, creating regions of destructive interference in the image. This leads to higher contrast, which helps to improve the effective patterning resolution. However, as with OPC, there is no true resolution enhancement in terms of the theoretical resolution limit. There are two main difficulties with application of PSM. First, it is not straightforward to introduce required phase shifts into an arbitrary design layout, because some geometrical shapes, such as junctions are harder to realize than others. Another problem lies with the mask fabrication, which requires careful calibration of the material layers and can be very complicated and costly compared to fabrication of conventional binary masks.

For the purposes of this article, only off-axis illumination is discussed, as both OPC and PSM are essentially mask-based techniques and thus provide no opportunity for micro-optics. Off-axis illumination refers to any illumination scheme that significantly reduces or eliminates the light hitting the mask at near normal incidence, leading to the diffraction pattern of the mask being shifted on the projection lens. The principle behind the resolution enhancement, obtained by introducing such a shift, is schematically illustrated in Figure 2 for periodic pattern.

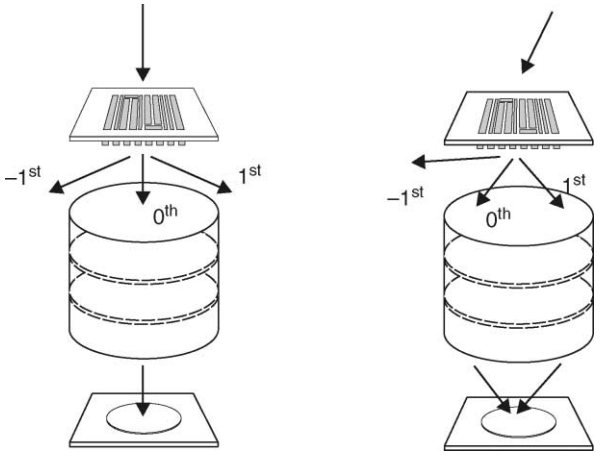


Figure 2 Schematic illustration of effect of off-axis illumination on image formation.

The mask pattern can be imaged if at least two diffraction orders generated by the mask pass through the pupil of the projection lens. Using on-axis illumination, the minimum feature size of the system is reached when the period of the mask pattern is so short that the ± 1st diffraction orders can no longer pass through the pupil. If the illumination is moved off-axis, the diffraction pattern is shifted and the second diffraction order (in this case $+1$st) again passes the pupil. Thus the resolution has been increased. A secondary improvement in contrast may be obtained by shaping the diffraction orders falling onto the pupil, such that the filling of the pupil is optimized, i.e., such that less unwanted light contributing to the background illumination passes the pupil. Finally, the depth of focus of the system can be optimized by ensuring that the contributing orders (in this case 0th and $+1$st) are symmetric with regards to the optical axis.

The conventional way to realize off-axis illumination schemes is to introduce a suitable aperture to the illumination optics between the source and the mask. In order to effectively use off-axis illumination, the shape and size of the aperture, i.e., the illumination pattern, must be optimized for the specific mask shape and pitch. This optimization leads to four basic shapes, as illustrated in Figure 3. In most modern lithography systems these shapes

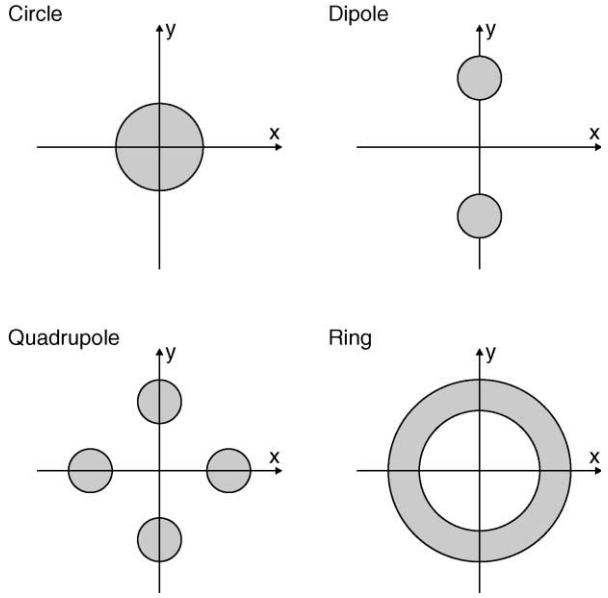


Figure 3 The four typical illumination patterns: disk, ring or annular, quadrupole, and dipole.

are available with different values of partial coherence factor, governed by the ratio of the numerical aperture of the illumination lens and projection lens, allowing control of proximity effects and edge sharpness.

Overview of Micro-Optics

There exists no generally agreed definition of micro-optics, but it typically means optical systems built to significantly smaller scales than conventional table-top systems, using components such as microlenses and microprisms with physical dimensions in the range of a couple of millimeters at most. Micro-optical systems also include so-called diffractive optical elements, which are usually optical components whose operation cannot be explained by Snell's law or the law of reflection. Such elements are usually macroscopically planar microstructures, consisting of features with dimensions from a couple of wavelengths to a few tens of microns, and are designed by advanced numerical algorithms based on diffraction theory.

In the past decade, micro-optics has emerged as a powerful new addition to the field of optics and optical systems, following improvements in micro-optics fabrication techniques, both via adaptation from established microelectronics fabrication processes and development of novel approaches, such as variable dose direct writing using either UV light or an electron beam. Additionally, emergence of replication techniques, such as injection molding or UV casting, allows the low-cost mass production of micro-optical elements, making micro-optics economically feasible for a wide range of consumer applications. Consequently, micro-optical elements, utilizing either refractive or diffractive surfaces, are now found in industrial applications such as material processing, in the optical telecom field, and in consumer electronics such as the digital video camera.

One of the major strengths of micro-optics, as compared to conventional optics, is their ability to realize complex optical functions into a compact form. This is especially true for micro-optics using diffractive optical elements, as diffractive elements may be designed to realize several optical functions such as focusing, filtering, or beam splitting at once allowing integration of several classical optical components into a single diffractive element.

Micro-Optics in Conventional Lithography Systems

The simplicity of contact or proximity printing means that there is little need or opportunity to incorporate additional functionality through inclusion of micro-optical elements. The only exception to this will be discussed in the final section of this article, where we will consider a nonconventional technique, which utilizes micro-optics and can be incorporated into proximity printing systems. However, in this section we will limit the discussion to projection-based lithography systems such as wafer steppers.

Let us consider the possible uses of micro-optics in the three main parts of the exposure system identified in the earlier section. It is obvious that the wafer management system, which is essentially mechanical, cannot be enhanced by inclusion of micro-optics. In principle, micro-optics could be incorporated into the projection system to reduce its weight and, to a lesser degree, complexity. However, it is important to note that using micro-optics to reduce the physical size, i.e., the diameter of the lenses in the projection system is not possible, as such reduction would increase the theoretical minimum feature size the system can resolve which is inversely proportional to the lens diameter. In practical terms, even the most advanced fabrication approaches used to realize micro-optical elements cannot yet produce elements with the precision and quality sufficient to meet the optical requirements associated with a lithographic projection lens. In general, the state of the art in micro-optics is such that it does not make sense to replace refractive lenses with diffractive ones unless the weight of the optics is a critical parameter in

determining the merit of the system. This is not the case for wafer steppers.

Even though the optical requirements for the lenses in the illumination system are not as high as in the projections system, they are in practice still far too demanding for micro-optical elements. Therefore, the only area where micro-optics could be used in the illumination system is the generation of various illumination patterns needed for the off-axis illumination schemes used for resolution enhancement, as well as overall homogenization of light. An element meeting these goals is essentially a custom-designed diffuser, where both the generated angular range and the relative intensity of illumination in any given angular direction are controlled to a high degree.

Custom Diffusers for Illumination

As mentioned earlier, one of the strengths of micro-optics lies in its ability to combine different optical functions into a single element. When used in lithographic illumination systems, micro-optical elements can serve a dual purpose. The element can serve as an artificial aperture, effectively distributing the light over some predetermined angular range and at the same time minimizing light outside the range, allowing generation of various illumination patterns needed for the resolution enhancement via off-axis illumination. Unlike conventional solutions based on the use of actual apertures to generate the desired illumination patterns, distribution of light using micro-optical elements can be done with efficiencies up to 95%, with typical target efficiency being in the range of 90%, depending on the pattern complexity. At the same time, the element can be used to smooth the light distribution, improving the uniformity of the illumination. Typically the required uniformity error, i.e., allowed variations from the average illumination, is in the range of 5%, but can be as low as 1%.

In addition to the above-mentioned high-efficiency and -uniformity targets, optical quality requirements of modern lithography systems introduce unique challenges to the micro-optics used in such systems. First, micro-optical elements used in lithography systems must also exhibit low amounts of stray light, especially in formation of hot-spots. Typically, hot-spots are allowed to be only a few percent of the average illumination level. Because of the fact that the size, shape, and profile of the illumination beam can vary, partly due to variations in the source and partly because such variations improve the functionality of the lithography system, by, for example, controlling the partial coherence factor, the micro-optical elements must be virtually space-invariant. The elements must perform the same function regardless of where and how the illumination coming from the source hits the element. Furthermore, the elements themselves must have very low absorption, because the high power used could otherwise lead to destruction of the elements. This means that the materials must be carefully selected. Finally, the optical wavelengths used in advanced lithography systems are 248 nm or less. Because the feature size of a diffractive element scales down with the wavelength, the features in elements used are much smaller than in the case of elements used for applications in the visible region. Consequently the accurate fabrication of such elements becomes a formidable challenge, because most fabrication errors, such as mask misalignment, etch depth error or the way the sharp corners in diffractive elements get rounded during fabrication, increase in significance when the feature size is decreased.

There exist two distinct approaches for obtaining such custom-designed diffusers. We will now discuss the benefits and weaknesses of these two techniques in terms of efficiency, pattern uniformity, stray light, and ease of fabrication, in order to illustrate the trade-off a designer creating micro-optical elements for a lithography system has to consider.

Aperture Modulated Diffusers

One approach to acquire custom diffusers for illumination systems is to use an array of refractive or diffractive micro-lenses. More generally, the lens phase can be replaced by any phase obtained by geometrical considerations, i.e., by a geometrical map transform between the input and output planes of the optical system. The basic principle of an aperture modulated diffuser is to use the phase function to accurately control the angular divergence of the diffuser, while the shape of the desired intensity distribution is created by introducing a properly shaped aperture in which the phase function is contained. This basic principle is shown in Figure 4.

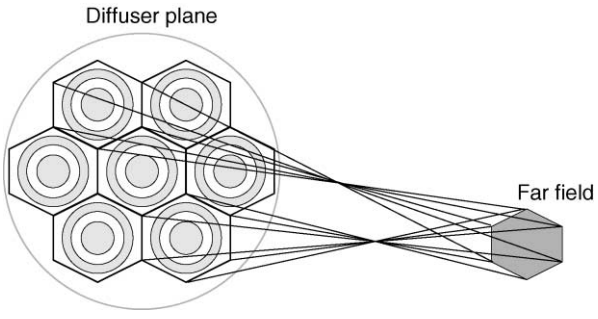


Figure 4 Basic principle of aperture modulated diffuser. The lens f-number determines the angular extend of the far field pattern, while the lens aperture shape determines the shape of the pattern.

The main benefit of using a lens phase lies in the ease of the design process, which essentially consists of selecting the numerical aperture of the lens to meet the angular size of the desired pattern, while more complicated phase profiles can be used to improve some aspects of the performance of the diffuser.

In terms of optical performance, aperture modulated diffusers typically achieve efficiency of up to 95% and low levels of stray light, easily fulfilling the target specifications outlined earlier, especially when refractive surfaces are used. However, since the lens aperture in the element plane controls the shape of the intensity distribution, the range of intensity distribution shapes is limited. This is necessary because it must be possible to tile the required aperture shape in order to realize an array of elements with high fill factor, i.e., to realize a space-invariant diffuser. Furthermore, introducing an aperture into the element plane means that intensity oscillations, due to boundary diffraction at the edge of the apertures, cannot be avoided. Consequently, the uniformity of the generated intensity distributions suffers. This is especially true in cases where the desired intensity pattern has sharp corners, where the intensity oscillations due to boundary diffraction, can easily be up to 50% of the average intensity of the desired intensity pattern. Figure 5 illustrates a typical far field intensity distribution obtainable with aperture modulated diffusers.

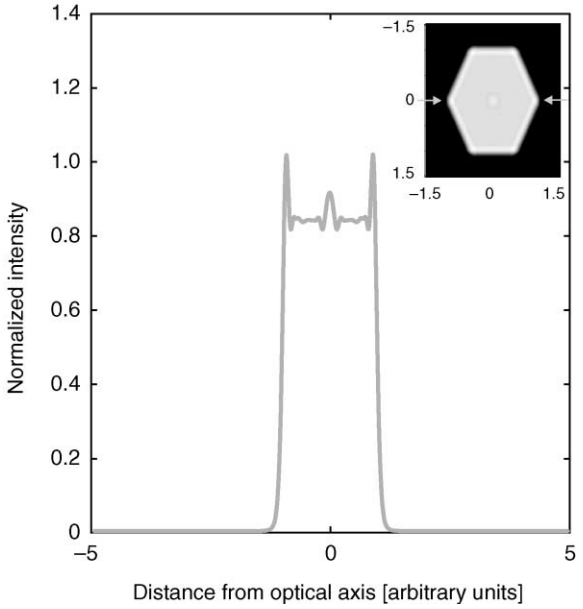


Figure 5 Typical intensity distribution obtainable with an aperture modulated diffuser. The overall distribution is smooth apart from the slow intensity oscillations at the edges resulting from diffraction at the aperture boundary.

One needs to consider several issues with regard to the fabrication of aperture modulated diffusers. For refractive aperture modulated diffusers, the main issue is the accurate realization of the required surface profile. The fabrication of large arrays of refractive elements having arbitrary surface profiles is still challenging. Resist reflow techniques, where the micro-lenses are realized by melting small islands of resist, lead to a spherical lens profile. Complicated and hard-to-control profile reshaping techniques are required if dramatic alterations from the spherical lens profiles are needed. Even though fabrication through direct writing techniques, such as laser writing offers more freedom, realization of arbitrary refractive surface remains a challenge. Therefore, use of refractive elements is mainly limited to cases where the lens profile can be close to the spherical shape. The most important of such cases are patterns with uniform circular or rectangular distributions. The former can be realized by proper selection of the numerical aperture of the lens, while the latter can be realized by combining two properly selected cylindrical lenses on opposite sides of the substrate.

For diffractive aperture modulated diffusers there are considerably less limitations in terms of what phase profiles can be attained. As long as the opening angles, i.e., the angular size of the pattern is below 10 degrees, the phase can be realized as a diffractive element. With diffractive elements, the main issue to consider is therefore the effect of typical fabrication errors such as profile depth and shape errors. First, the typical consequence of fabrication errors in the element is loss of efficiency via increase of stray light. Fortunately the errors do not generate sharp stray light peaks, with the obvious exception of the axis hot-spot. This results from the fact that both lens functions and more complicated phase functions designed by geometrical considerations, exhibit locally periodic structures, and the introduction of common fabrication errors then leads to redistribution of light away from the desired signal orders of these local structures to higher or lower orders. On the level of the whole intensity distribution, such redistribution appears as a constant rise in the background noise level, but will not lead to sharp intensity fluctuations.

Another issue related to the fabrication of aperture modulated diffusers appears when diffuser elements are fabricated using a series of photomasks, as it is typically the case with a relatively large size of the element. Although mask misalignment is essentially a random error, i.e., the direction and amount of the error are difficult to predict before hand, it easily leads to systematic and partly symmetric uniformity problems with aperture modulated diffusers.

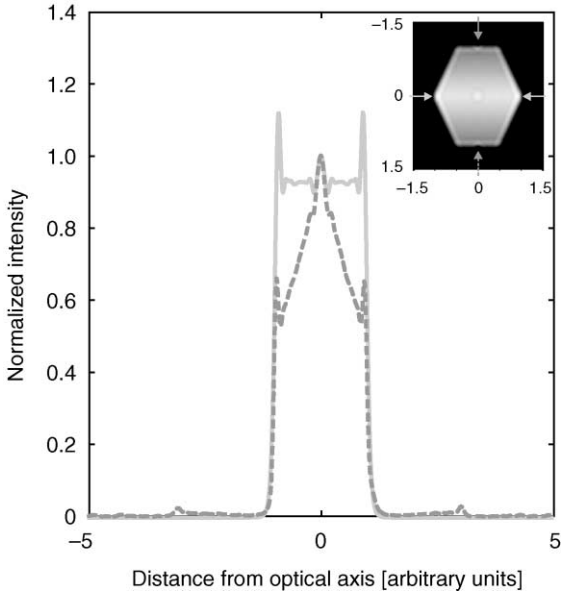


Figure 6 Effect of mask misalignment on the intensity distribution with aperture modulated diffusers. Essentially random error leads to symmetric and systematic uniformity error due to locally periodic nature of the design.

Again the reason for this lies in the locally periodic nature of the elements. A single mask misalignment along some direction results in an erroneous profile shape in the structures, which are periodic in the direction of the alignment error. Consequently, the relative intensity distribution along that direction is changed. On the other hand, structures, where the period is perpendicular to the misalignment, are realized correctly. Furthermore, the local period of the phase function is often changing from the center of the element to the edge, meaning that the relative magnitude of the mask misalignment is different in different parts of the element. These facts lead to asymmetry in the overall intensity distribution, which is especially harmful with off-axis illumination patterns, as the asymmetry results in loss of the resolution gain obtained with off-axis illumination. Furthermore, as the source of the asymmetry is essentially a random error, compensation in the design is very difficult. Figure 6 illustrates the above-mentioned phenomena.

Fan-Out Based Diffusers

An alternative technique uses specially designed fan-out elements to produce a large number of closely spaced diffraction orders arranged in a desired pattern, as illustrated in Figure 7. Such an approach offers all the flexibility of diffractive elements, allowing generation of nearly arbitrary intensity distributions, especially if the elements can be realized

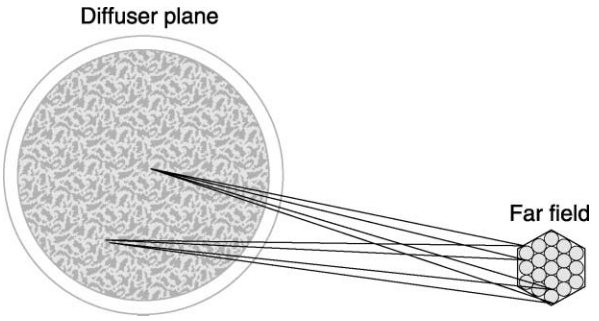


Figure 7 Basic principle of fan-out based diffusers. A grating is used to generate a high number of closely spaced diffraction orders arranged in the shape of the desired pattern.

as so-called multilevel elements, i.e., if the surface profile can have more than two distinct levels.

The flexibility in the design comes at a cost, where specially developed design tools, such as the iterative Fourier transform algorithm, are required. Fortunately such tools are starting to become commercially available as emphasis and interest in wave-optical engineering, an extension of the classical optical engineering approaches, increases.

In terms of optical performance, fan-out based diffusers offer high uniformity as there are no boundary diffraction effects as with aperture modulated diffusers. Furthermore, by choosing the beam separation of the fan-out to be clearly smaller than the divergence of the incidence beam, it is possible to average out any local intensity variations from order to order, leading to a smooth overall intensity distribution. The remaining variations in the intensity distribution generated by a fan-out based diffuser can be as small as 1% of the average intensity.

As a trade-off for the improved performance in terms of uniformity compared to aperture modulated diffusers, fan-out based diffusers typically exhibit lower efficiency and, consequently, higher levels of stray light, even when no fabrication errors are present. This is a result of additional design freedoms being used to remove unwanted intensity oscillations from the far field regions, belonging to the desired intensity distribution, by introducing some noise outside these regions. Typically the efficiency of a fan-out based diffuser is in the range of 85–90%, or 5–10 percentage points lower than the efficiency of a comparable aperture modulated diffuser (should one exist). Furthermore, unlike aperture modulated diffusers, fan-out based diffusers can have stray light that forms hot-spots. If care is not taken during the design, these hot-spots may be up to 10% of the average intensity of the illumination pattern, even in the absence of fabrication errors. Figure 8 illustrates

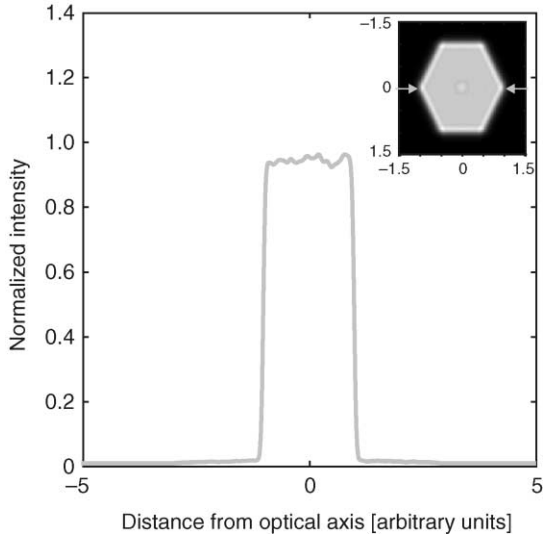


Figure 8 Typical intensity distribution obtainable with a fan-out based diffuser. Overall distribution is smooth, but stray light appears outside the main pattern.

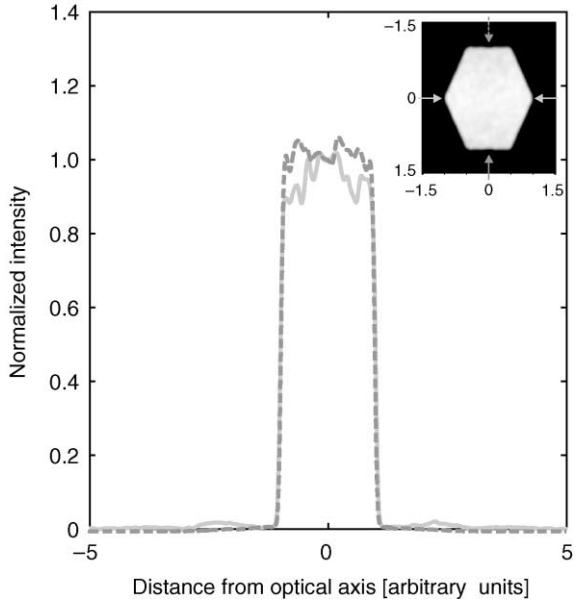


Figure 9 Effect of mask misalignment on the intensity distribution with fan-out based diffusers. The uniformity error appears randomly, allowing statistical compensation schemes.

a typical far field intensity distribution obtainable with fan-out based diffusers.

Fan-out based diffusers offer several advantages in terms of fabrication compared to mapping based diffusers. First, random errors, such as mask misalignment, do not lead to symmetric or systematic uniformity problems, but instead result in random local oscillations in the intensity pattern as illustrated in Figure 9. This allows special compensation schemes where the element is divided into segments, each of which has different fabrication tolerances but generates the same far field intensity distribution. If the fabrication tolerances of the individual designs are balanced correctly against the typical fabrication errors of the used fabrication methods, the loss of uniformity due to fabrication errors can be eliminated almost entirely. Furthermore, the flexibility of diffractive design allows for fabrication error-tolerant designs. This follows from the ability to combine several optical functions in a single optical element. The typical trade-off with designs with relaxed fabrication tolerances is that the optical performance of the element in terms of efficiency is worse than that of a conventional design if fabrication errors are not present.

Micro-Optics and Non-Conventional Lithography

In the previous section we have considered micro-optics in connection with conventional lithography approaches and shown that micro-optical elements can find a significant role in such systems. We conclude this article by considering briefly nonconventional lithography systems, focusing especially on a technique called micro-lens lithography, which is fundamentally based on the use of micro-optical elements.

The most straightforward way to improve existing lithography approaches is to further reduce the wavelength used in the exposure process. In X-ray lithography, the light source used in conventional lithography approaches is replaced by a cyclotron emitting light with wavelengths deep in the X-ray region. This leads to minimum resolution far lower than achievable by conventional techniques, but with the additional difficulty and cost of maintaining and running such a source.

The techniques that are currently considered as serious alternatives for conventional photolithography approaches are based on the use of particles other than photons in the fundamental exposure process. Techniques such as electron- or ion-beam lithography are routinely used to realize structures with feature size and positioning resolution of some tens of nanometers for use in both micro-electronics and micro-optics. The main drawback of these approaches is the cost involved in patterning large surface areas. Unlike optical lithography techniques, particle beam lithography cannot expose large surface areas at once, instead requiring painstaking scanning of the surface area with a relatively small writing beam. This leads to processing times that scale linearly with the required surface area, and consequently the throughput of such techniques is small. Therefore, particle beam lithography is not yet widely used in commercial fabrication of micro-electronics. However, it is worth noting that especially

electron beam lithography is widely used in the origination of micro-optical components, as the relatively high cost of creating the master component can be offset in the final product by the inexpensive replication process. Such an approach allows the replicated end-product to meet the high optical requirements achievable only by high resolution surface control, yet remain affordable for use in, for example, consumer electronics. This has lead some in the micro-optics community to conclude that electrons are better suited for fabrication of components controlling photons, while photons are better suited for fabrication components controlling electrons.

In addition to particle beam lithography, other noteworthy techniques exist. For our purposes, the most interesting technique is the so-called micro-lens array lithography, where one or more micro-lens arrays are used to image the pattern of the photomask to the resist layer on the substrate. Although the resolution of this approach is limited to 3–5 microns, for applications such as with flat panel displays, micromechanics, or multichip module fabrication, this technique is of considerable interest.

It is well known that the geometrical aberrations scale down proportionally with the physical lens size. Thus, large areas can be imaged favorably by use of arrays of micro-lenses, where each lens images only a small section of the whole field. Such a setup is shown in Figure 10.

With this setup it is possible to expose a comparatively large area with a single exposure, thus eliminating the need for costly stepping operations and therefore increasing the throughput of the system. Analysis of the basic system shows that an array based on spherical micro-lenses of 1 mm focal length and 300 micron aperture can produce a nearly diffraction limited resolution on the order of 3–5 micrometers over a 300 micron field.

In terms of conventional lithography systems, the micro-lens array lithography has most in common with proximity printing. Both techniques achieve comparable resolutions, yet avoid damaging the mask or substrate by keeping the two separated by some finite working distance. However, micro-lens array lithography offers two distinct advantages over conventional proximity printing. First, the working distance between the imaging system and the substrate surface can be much larger, which further minimizes the risk of mask or substrate damage. Furthermore, added working distance allows lithography steps on nonflat surfaces, e.g., inside holes or grooves where conventional proximity printing would not be possible. Another advantage of the micro-lens array lithography is the larger depth of focus. With proximity printing, the resolution rapidly decreases with increasing distance. In micro-lens array lithography, the optimum resolution is obtained at the image plane, while a region of reasonable resolution extends forward and backward from this plane (Figure 11). This extension means that in the depth of field, resolution is less sensitive to changes in the working distance, which makes it better suited for practical applications.

In addition to the optical benefits of micro-lens array lithography, there also exist other practical benefits. Many micro-devices have a repetitive and

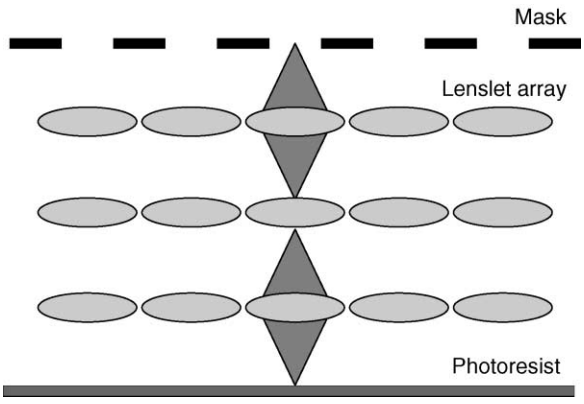


Figure 10 Basic geometry of micro-lens array lithography.

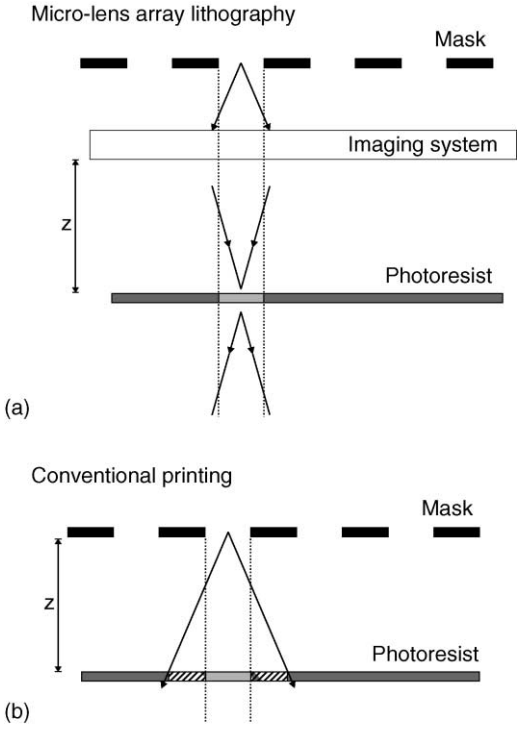


Figure 11 Depth of focus with micro-lens array lithography and conventional printing. (a) With micro-lens array lithography, the image of the mask is extended both forward and backward from the working plane at distance z. (b) In conventional printing, the resolution is rapidly reduced with increasing working distance z.

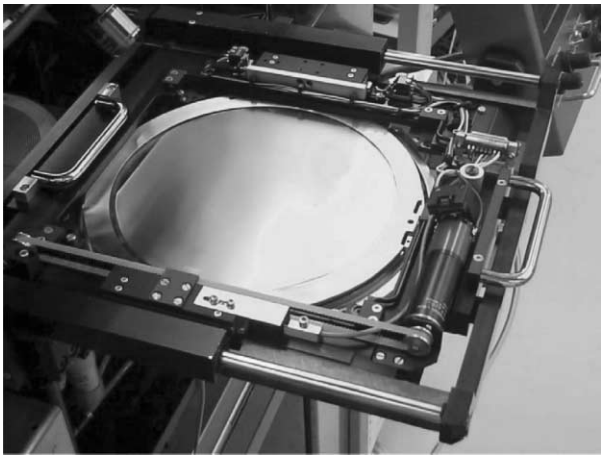


Figure 12 Ultra-flat micro-lens projection system (courtesy of R. Völkel, SUSS MicroOptics, Neuchâtel, Switzerland).

discrete nature with respect to some unit element. Such a unit element lends itself well to the imaging properties particular to lens arrays, as each lens of the array can be made to correspond to the unit cell of the particular micro-device making the imaging highly efficient. A photograph of a micro-lens projection system is shown in Figure 12.

One could also imagine that instead of an array of imaging systems, an array of pattern generators is used that generates a specific pattern, such as lines or points. In the ideal case such a system would replace the mask and the imaging system altogether. However, this approach is only possible for simple patterns and large production series.

Outlook

As long as optical lithography stays attractive, micro-optics will be used as beam-homogenizers and sophisticated diffusers to generate optimized illumination patterns. How far micro-optics will be used in the imaging system is difficult to predict. However, it can be assumed that the difference between micro-optics and macro-optics fabrication technology will disappear.

Recent research is also investigating near-field optics to further reduce the resolution limit. Micro-optics technology is well suited to realize such systems. A serious problem of near-field optics is the short distance (10 nm–100 nm) between the optical element and the interface to be patterned. In addition, near-field optics is based on scanning systems, which increases the exposure time considerable. However, there is still room for improvements or as Richard Feynman said 'there is plenty of room at the bottom'.

Further Reading

Dändliker R, Gray S, Clube F, Herzig HP and Völkel R (1995) Non-conventional techniques for optical lithography. *Microelectronic Engineering* 27: 205–211.

Herzig HP (1998) *Micro-Optics: Elements, Systems and Applications*. London: Taylor & Francis.

Herzig HP and Kipfer P (1999) Aperture modulated diffusers (AMDs). In: Asakura T (ed.) *International Trends in Optics and Photonics*. Berlin: Springer.

Naumann H and Schröder G (1992) *Bauelemente der Optik*. München, Germany: Carl Hanser Verlag.

Sheats JR and Smith BW (eds) (1998) *Microlithography: Science and Technology*. New York: Marcel Decker.

Turunen J and Wyrowski F (eds) (1997) *Diffractive Optics for Industrial And Commercial Applications*. Berlin: Wiley-VCH.

Völkel R, Herzig HP, Nussbaum P, Dändliker R and Hugle WB (1996) Microlens array imaging system for photolithography. *Optical Engineering* 35: 3323–3330.

Design and Fabrication of Diffractive Optical Elements

D W Prather, T Dillon, A Sure and X Gao, University of Delaware, Newark, DE, USA

J N Mait, US Army Research Laboratory, Adelphi, MD, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Diffractive optical elements (DOEs) are passive devices that redirect and focus light through the division and mutual interference of a propagating electromagnetic wave. This is in contrast to refractive elements, which redirect and focus light through variations in indices of refraction. From a historical perspective, diffraction was traditionally considered a detriment in optics and caused images to blur. However, we now know that it is possible to use diffraction to enhance the performance of optical systems. For example, DOEs can be used to correct chromatic aberrations and to implement optical functions that conventional refractive and reflective optics cannot, such as multiple beamsplitting.

Further, recent advances in computing and fabrication technologies have increased the number of applications in which diffractive elements are used, which has established diffractive optics as yet another tool in an optical designer's kit.

The first modern diffractive element was the hologram. A hologram is a photographic record of the constructive and destructive interference between two coherent beams known as the object and write beams. The object beam contains information about the object recorded. The write beam is necessary to record the hologram but contains no information itself. When the hologram is played back with a third beam, the read beam, the recorded interference pattern diffracts it in such a way that it reconstructs the object beam.

Unfortunately, the functions that a natural hologram can produce are limited to those whose wavefields can be produced in a laboratory. Thus, although it is possible to record a hologram that can focus or deflect a single beam, it is difficult to record a hologram that splits a single beam into multiple foci. However, the advent of computing technology made it possible to design such holograms analytically. The first diffractive element designed with the aid of a computer was demonstrated in 1966. Since then the ability to create elements that respond arbitrarily has had widespread application in imaging systems, telecommunications, and sensing.

The two most basic diffractive elements are the grating and the Fresnel zone plate. The diffractive behavior of a grating is predominately transverse to the optic axis. It causes light to split into discrete beams, referred to as diffraction orders, each with a different angle of propagation relative to the optic axis. The diffractive behavior of a Fresnel zone plate also divides incoming light into discrete orders, however, the orders are produced along the optic axis. This is useful in producing the diffractive equivalent of a refractive lens. Arbitrary transverse and longitudinal behavior is realized by combining characteristics of these basic two diffractive elements. In the remainder of this article, we discuss the design and fabrication of general diffractive elements.

Design

Once an optical system designer determines the functional behavior required for a particular application, for example, focusing or multiple beamsplitting, he or she needs to design a diffractive element that realizes this function. However, unlike refractive and reflective optical elements, whose material properties and surface curvatures determine their functionality, the functionality of a diffractive element is determined by a pattern of wavelength scale features. Determining this pattern is the goal of diffractive design but to do so one must first understand the physical relationship between the DOE and the wavefield it produces.

Scalar diffraction theory is one of the simpler models that relates the effect of a surface relief, or diffractive profile, $d(x, y)$, on an incident wavefield $U_{\text{in}}(x, y)$:

$$U_{\text{out}}(x, y) = \exp[\mathrm{i}\theta(x, y)]U_{\text{in}}(x, y) \qquad [1]$$

where the phase transformation $\theta(x, y)$ is related to the diffractive profile $d(x, y)$ by

$$\theta(x, y) = 2\pi(n - 1)d(x, y)/\lambda \qquad [2]$$

The constant n is the refractive index of the optical substrate and λ is the wavelength of illumination in a vacuum.

The effects of the phase change are not apparent until the wavefield $U_{\text{out}}(x, y)$ propagates over some distance z. In scalar diffraction theory the relationship between $U_{\text{out}}(x, y)$ and the wavefield $U_z(x, y)$ at a plane a distance z away, is described by the Rayleigh–Sommerfeld diffraction equation:

$$U_z(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U_{\text{out}}(\zeta,\eta) \times \frac{\exp\left\{\mathrm{i}(2\pi z/\lambda)\sqrt{1+(x-\zeta/z)^2+(y-\eta/z)^2}\right\}}{\mathrm{i}z\lambda[1+(x-\zeta/z)^2+(y-\eta/z)^2]}\mathrm{d}\zeta\,\mathrm{d}\eta \qquad [3]$$

Huygens' principle is reflected in this formulation by the convolution between the transmitted field $U_{\text{out}}(x,y)$ and a spherical wave $h(x,y)$, which represents free space propagation over the distance z:

$$U_z(x,y) = U_{\text{out}}(x,y)^{**}h(x,y) \qquad [4]$$

where

$$h(x,y) = \frac{\exp\left\{\mathrm{i}(2\pi z/\lambda)\sqrt{1+(x/z)^2+(y/z)^2}\right\}}{\mathrm{i}z\lambda[1+(x/z)^2+(y/z)^2]} \qquad [5]$$

The Rayleigh–Sommerfeld formulation is valid so long as the angles at which energy diffracts from the aperture are small. If the distance z is sufficiently large that the angles stay clustered about the optic axis, one can replace the spherical wave response of freespace

with a parabolic wave:

$$h(x, y) \approx \frac{\exp(\mathrm{i}2\pi z/\lambda)}{\mathrm{i}z\lambda}\exp[\mathrm{i}(\pi/\lambda z)(x^2 + y^2)] \qquad [6]$$

This simplification is referred to as the paraxial, or Fresnel, approximation. The diffraction integral is now given by:

$$\begin{aligned} U_z(x, y) \approx & \frac{\exp(\mathrm{i}2\pi z/\lambda)}{\mathrm{i}z\lambda}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U_{\text{out}}(\zeta, \eta) \\ & \times \exp\{\mathrm{i}(\pi/\lambda z)[(x-\zeta)^2 + (y-\eta)^2]\}\mathrm{d}\zeta\,\mathrm{d}\eta \\ = & \frac{\exp(\mathrm{i}2\pi z/\lambda)}{\mathrm{i}z\lambda}\exp[\mathrm{i}(\pi/\lambda z)(x^2 + y^2)] \\ & \times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U_{\text{out}}(\zeta, \eta)\exp[\mathrm{i}(\pi/\lambda z)(\zeta^2 + \eta^2)] \\ & \times \exp\{-\mathrm{i}2\pi[\zeta(x/\lambda z) + \eta(y/\lambda z)]\}\mathrm{d}\zeta\,\mathrm{d}\eta \end{aligned} \qquad [7]$$

The second representation results from an expansion of the exponent and allows the diffraction integral to be written as a Fourier transform. The quadratic phase terms can be removed if a positive lens with focal length z is placed next to the diffractive element and a second positive lens with focal length z is placed in the image plane (see **Figure 1**). Under these conditions, $U_{\text{out}}(x, y)$ and $U_z(x, y)$ form a Fourier transform pair:

$$\begin{aligned} U_z(x, y) = & \ \mathrm{FT}[U_{\text{out}}(x, y)]|_{u=x/\lambda z, v=y/\lambda z} \\ = & \ \mathrm{FT}\{\exp[\mathrm{i}\theta(x, y)]\}^{**}\mathrm{FT}[U_{\text{in}}(x, y)]|_{u=x/\lambda z, v=y/\lambda z} \end{aligned} \qquad [8]$$

The objective of diffractive design is to find a surface relief $d(x, y)$ that generates a desired response $q(x, y)$ somewhere in the response $U_z(x, y)$. However, eqns [3] and [7] neglect the fabrication process, which affects the fields that can actually be realized. A diffractive optic designer can incorporate fabrication constraints into their design either directly or indirectly. In a direct approach, the designer represents the diffractive structure parametrically, for example, a multilevel phase element can be represented by a discrete set of allowed phase values and a binary phase element can be represented by the contours that separate one phase value from the other. A designer defines the diffractive element by specifying and modifying these parameters.

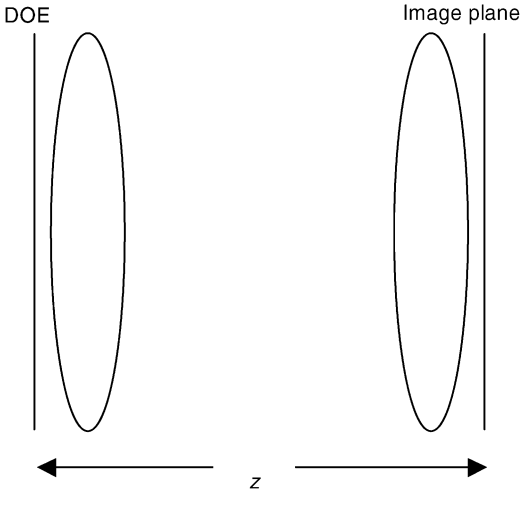


Figure 1 Optical Fourier transform module.

In an indirect approach, the designer first determines an intermediate function without consideration for fabrication and afterward applies the fabrication constraints to this function. For example, if $f(u, v)$ is the desired response one wishes to generate in the z-plane using $U_{\text{in}}(x, y)$ as the input, from eqn [8], the optical function required to do is

$$Q(x, y) = F(x, y)/U_{\text{in}}(x, y) \qquad [9]$$

where $F(x, y)$ is the Fourier transform of $f(u, v)$. The designer must then find a phase $\theta(x, y)$ that satisfies the fabrication constraints and insures that $\exp[\mathrm{i}\theta(x, y)] \approx Q(x, y)$.

For example, fabrication using multiple binary masks yields a phase structure quantized to 2^N levels. An effective design should, therefore, yield a quantized phase structure that minimizes reconstruction error. A naïve approach is to set the magnitude of $Q(x, y)$ in eqn [9] to unity and quantize the phase. However, this introduces considerable error. A direct approach to design searches only over the set of quantized phase structures for the one that, for example, generates the desired response with maximum energy and minimum error. An indirect design seeks first a phase structure whose response produces a minimum amount of error between it and $Q(x, y)$ and then quantizes this structure.

Performance of the element can be improved if one exploits certain freedoms available in design. For example, all systems have a finite-sized detector. It is, therefore, unnecessary to specify the performance of the diffractive element everywhere in the image plane. Its behavior within the detector window is all that is important. If only the intensity of the response is specified, then the phase of the response within the window is also a design freedom. Finally, one can also control the amount of energy contained in the signal window by scaling the amplitude of the response. This allows a designer to balance diffraction efficiency with error in the signal window.

Two general nonlinear optimization algorithms for the design of DOEs are represented in **Figure 2**. Indicative of the nature of data flow, **Figure 2a** is

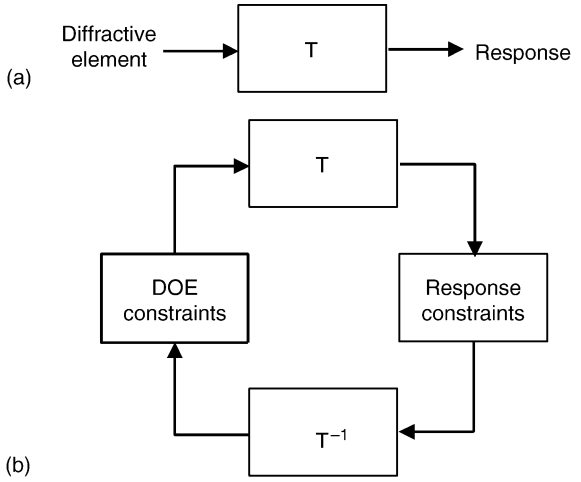


Figure 2 Algorithms used for diffractive design. (a) Unidirectional and (b) bidirectional algorithms.

referred to as a unidirectional algorithm and Figure 2b as bidirectional. In a unidirectional algorithm, changes are made only to the diffractive element. The changes are driven by how well the element performed in a previous iteration. If models exist for both forward and inverse propagation ($\mathbf{T}$ and $\mathbf{T}^{-1}$, respectively), a designer can use a bidirectional algorithm, which allows modification of both the diffractive element and its response. This allows bidirectional algorithms to converge to a solution, in general, more rapidly than unidirectional algorithms. In scalar theory, $\mathbf{T}$ is a Fourier transform but more rigorous models exist. If the model of the optical system cannot be inverted, a unidirectional algorithm must be used.

Fabrication

Regardless of which design algorithm is used to determine $d(x, y)$, ultimately it needs to be realized physically. Although numerous techniques exist for doing so, by far the most common are based on microlithographic processes. The process requires exposing a layer of photoresist deposited on top of an optical substrate and etching the photoresist to transfer the pattern into the substrate. These methods are especially appealing because they leverage the existing manufacturing infrastructure of the microelectronics industry and, as a consequence, have enabled the efficient and economical mass manufacture of DOEs.

Two basic methods exist for patterning photoresist, direct writing using a modulated beam and indirect patterning using masks. Further, the mask process can be either a single or multistep process. The multistep approach uses N masks and N exposures to fabricate a multilevel diffractive element that has 2^N phase levels (see Figure 2). However, profiles that contain very small features are difficult to fabricate in this manner. As a consequence, alternative processes using single-step grayscale lithography have also been developed.

Although grayscale photolithography is more time-efficient and less prone to mask alignment errors than multistep lithography, the photoresist must respond linearly to illumination. Unfortunately, most photoresists used to fabricate microelectronic circuits are highly nonlinear. However, one can obtain a linear response by adjusting the exposure and development process. In the following, we review fabrication using both multiple binary masks and a single grayscale mask.

Multilevel Fabrication Using Binary Masks

In this fabrication method, one must first generate a sequence of N masks that are related, in a recursive manner, to the quantized phase $\theta(x, y)$:

$$m_k(x, y) = \mathrm{mod}(M_k(x, y), 2), \quad k = [1, N] \qquad [10]$$

where

$$M_{k+1}(x, y) = \frac{M_k(x, y) - m_k(x, y)}{2} \qquad [11]$$

and

$$M_1(x, y) = \theta(x, y)\frac{N}{2\pi} \qquad [12]$$

The masks can be produced using either optical or electron beam (e-beam) pattern generators from integrated circuit fabrication. The pattern generators expose a thin layer of photoresist on the surface of a chrome-covered quartz substrate. Development of the photoresist leaves the desired diffractive pattern in the unexposed areas. The diffractive pattern is transferred to the chrome layer by etching any chrome not covered by photoresist. After the chrome etch process is completed, any remaining photoresist is washed away to yield the finished lithographic mask. This process is repeated for each mask.

The completed mask set is used to fabricate the DOE using techniques similar to those used to generate the mask set. An optical substrate is coated with a thin layer of photoresist, placed in intimate contact with the first lithographic mask $m_1(x, y)$, and illuminated with an ultraviolet (UV) exposure lamp. During photoresist development, exposed resist is washed away, which leaves a binary pattern in the remaining developed photoresist.

When the pattern is transferred into the substrate material via etching, the photoresist acts as an etch stop. For the first mask, the etch depth into the substrate corresponds to a π-phase shift. The process is repeated for each subsequent mask, that is, the substrate is coated with a layer of photoresist, UV illumination exposes the photoresist through the mask, the photoresist is developed, and the substrate etched by the remaining pattern. The etch depth of each mask is half that of the previous mask. Figure 3 illustrates this process.

The substrate can be etched using either wet or dry techniques. In wet etching, one uses reactive chemicals, such as hydrofluoric acid, to etch the substrate. However, wet etching is isotropic, that is, substrate etching occurs equally in all directions. As a result, it is difficult to preserve vertical sidewalls and to achieve high fidelity in etching high-aspect ratio features. As a consequence, dry etching is more commonly used.

A dry etch uses a beam of ions to remove substrate material. The process for removal can be either chemical or kinetic. The process used is determined by the value of the plasma power that creates the source of ions relative to the bias power used to accelerate the ions toward the substrate. If the bias power is large and the plasma power small, a small stream of ions are accelerated rapidly toward the substrate. In this case, the operational mode is similar to a micromilling machine, wherein inert particles chip away at the surface by virtue of their kinetic energy. Alternatively, a small bias power and a large plasma power produce a large population of ions with no preferential directionality. This yields an isotropic process similar to wet etching.

Two methods for performing dry etching are reactive ion etching (RIE) and inductively coupled plasma (ICP) etching. However, in RIE systems the bias and plasma power are not independent due to capacitive coupling. Thus, adjusting one affects the other. ICP, on the other hand provides independent control over the bias and plasma powers. Independent control allows one to control precisely the anisotropic nature of an etch process. As mentioned above, this is achieved by varying the bias power relative to the plasma power, which controls the directionality of the acceleration.

Independent control is important because the ideal process for a given DOE depends on the substrate material and the characteristics of the diffractive profile. Independent control allows one to adjust the respective powers, as well as gas flow rates, and their partial pressures, to determine an optimal etch process. Dry etching allows one to incorporate endpoint detection, which provides precise control

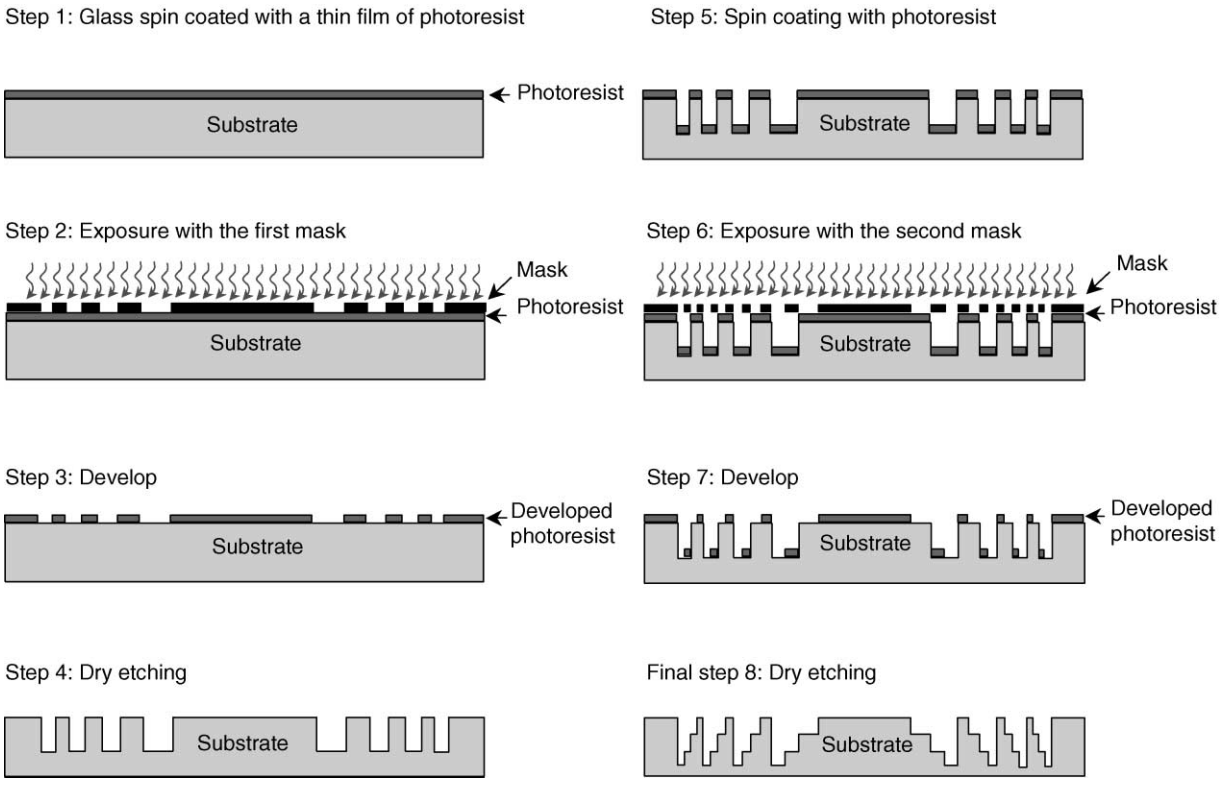


Figure 3 Fabrication of a multilevel quantized phase DOE using binary masks.

over the etch depth. As presented below, achieving the proper etch depth is important to generating responses with high diffraction efficiency.

Care must be used when determining the number of phase levels appropriate for a particular application. The number of phase levels has a direct impact on the portion of input energy one can hope to diffract into the desired response. This is referred to as the diffraction efficiency. According to scalar diffraction theory, the relationship between the maximum diffraction efficiency η and the number of masks N is:

$$\eta = \mathrm{sinc}^2\left(\frac{1}{N}\right) \qquad [13]$$

Thus, using two masks to create a 4-level phase element, as opposed to a single mask to create a binary-phase element, doubles the upper bound on the diffraction efficiency from 41% to 81%. The maximum diffraction efficiency for 8-level and 16-level phase elements is 95 and 99%, respectively.

However, this increase in diffraction efficiency incurs fabrication costs. After the first etch step, each subsequent exposure requires careful mask alignment. Alignment is accomplished using a standard tool from integrated circuit fabrication known as a mask aligner. The quality of a mask aligner is determined by its precision and the majority of commercially available aligners can register a mask and substrate to within less than a micrometer. This degree of alignment accuracy is sufficient to fabricate many useful multilevel diffractive elements.

To summarize, a multilevel surface relief profile uses a set of N masks to generate a surface profile that exhibits 2^N phase levels. The etch depth for the kth mask is $d_k = \lambda/2^k(n-1)$. In the next section, we describe a process for fabricating continuous DOE profiles in a single expose and etch cycle.

Continuous DOE Profiles Using Grayscale Lithography

Grayscale photolithography has become increasingly popular because of its flexibility and convenience in fabricating continuous diffractive profiles. Methods including directly writing partially exposed e-beam resist, resist reflow lithography, and using grayscale masks in a single step exposure. Here we discuss only the latter method because the processes are similar to those used in multiple mask fabrication. Thus, the same facilities can be used to produce both. In particular, we consider the application of high-energy-beam-sensitive (HEBS) glass to grayscale lithography and the fabrication of continuous profile DOEs.

The process for creating a diffractive element using HEBS glass is similar to fabricating a binary-phase element from a single mask. However, mask creation is markedly different. HEBS glass is a white crown glass within which a thin buried layer of silver ions is incorporated. The layer of silver ions is created via ion exchange in an aqueous acidic solution and essentially serves the same function as silver grains in photographic film. When the glass is exposed to a beam of electrons, the silver-containing crystal complexes are chemically reduced and the glass darkens. By controlling the electron dose, one controls the opacity of the glass to produce a grayscale pattern.

Although the attraction of single-step grayscale lithography is reduced costs in fabrication, the process incurs high calibration costs to insure quality pattern transfer through several steps. The parameter that has the greatest impact on the overall process is the desired size of the smallest diffractive feature. The minimum feature size affects the selection of the accelerating potential, beam current, addressing grid, and electron dose of the e-beam used to expose the HEBS glass.

The e-beam accelerating potential sets the resolution limit of the process and affects the sensitivity of the glass to the electron dose, the total charge per area incident upon the glass. For example, for a constant electron dose, increasing the accelerating potential increases the darkening response of the material. The large accelerating potential produces a large flux, which achieves the desired electron dose in a short period using energetic electrons. This, in turn, enhances the sensitivity of the glass. However, increasing the accelerating potential also increases scattering, which limits the minimum feature size achievable. Thus, low beam energies are generally used to write small features. But this, unfortunately, reduces the range of available optical densities because the material is less sensitive at low beam energies. To compensate for the reduced sensitivity, it is possible to increase the electron dose by increasing the exposure time. However, this undermines to some extent using a low energy beam to achieve small features due to broadening that results from long exposure times.

Once the accelerating potential for writing the mask is chosen, the beam current and addressing grid need to be specified. The beam current is related to the spot size and, therefore, affects directly the length of time to write the mask. High current levels produce large spots. Since write times can easily reach into

hours for large structures, high current levels are desirable to reduce them.

The addressing grid represents the physical grid that the electron beam is translated over in order to complete the exposure. This grid should be chosen as finely as possible, however, too fine a grid can significantly increase the required dwell time on each grid point. In this respect, a compromise is required. High beam currents produce short dwell times, which may necessitate using a coarse address grid to insure the beam is capable of performing the exposure.

Once the process parameters are set, calibration between the HEBS transmission and the electron dose is necessary. Analyzing the optical density of written test structures by some experimental means and determining its functional dependence on the electron dose completes the calibration. This allows one to specify the electron doses required to achieve the features dictated by the device design.

If resolution requirements cannot be met through judicious selection of the accelerating potential, then some of the contrast in the mask can be sacrificed for an increase in available resolution. As we have mentioned, reducing the maximum optical density, or contrast, can alleviate scattering-induced broadening. It is possible, however, to compensate to some degree for the reduced contrast during the photolithography process, where UV exposure and develop times provide additional latitude in the pattern transfer process.

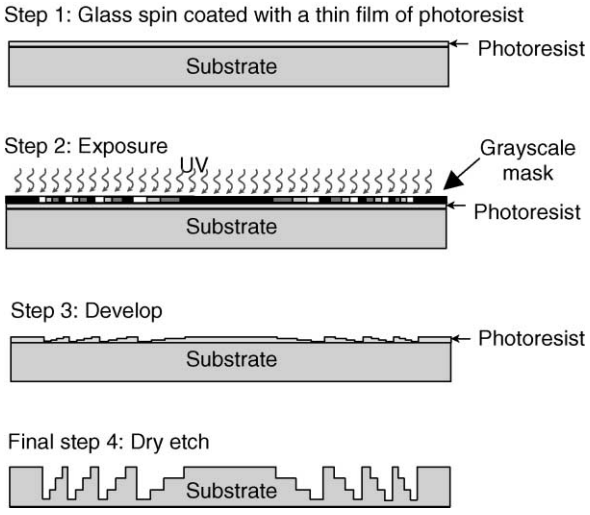


Figure 4 Fabrication of a continuous phase DOE using a grayscale mask.

Figure 4 represents the fabrication of a continuous profile DOE using a grayscale mask. The substrate is spin-coated with a thin film of photoresist, placed in intimate contact with the grayscale mask, and UV illuminated to expose the resist film. The resist is developed and the pattern transferred into the substrate using either wet or dry etching techniques. Because the solubility of the resist is proportional to exposure, developing the photoresist produces a variable height profile.

Grayscale lithography is challenging because modern photoresists are also designed to be highly nonlinear yet linear response is required. Traditional photoresists are also designed to amplify mask contrast and produce the sharpest resist profile, thus they either dissolve away quickly and completely, or remain impervious to developer. The exposure threshold of traditional resists is, therefore, steep and considerable effort is required to

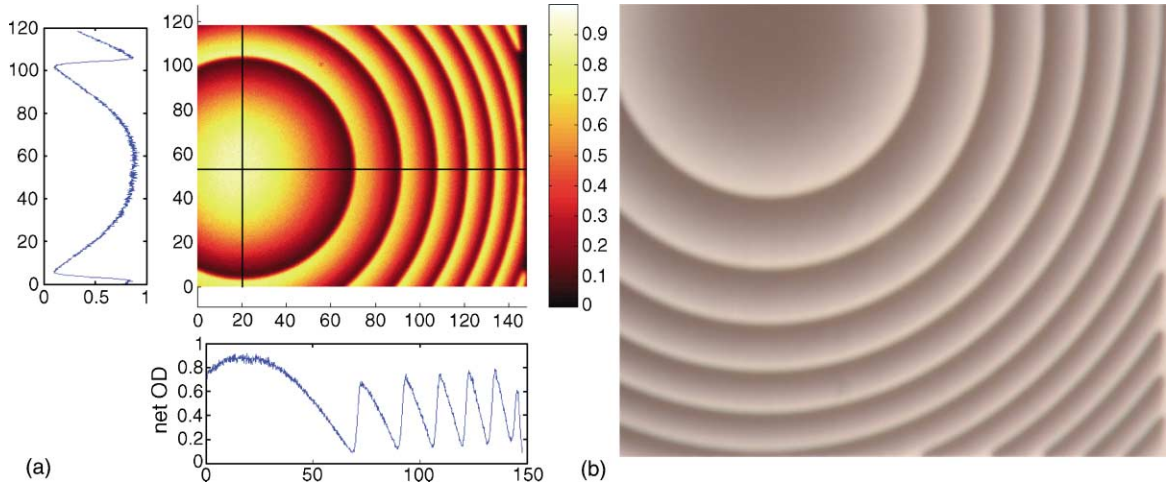


Figure 5 HEBS glass fabrication. (a) Data map of the optical density cross-section of an exposed HEBS grayscale mask and (b) its corresponding visible image.

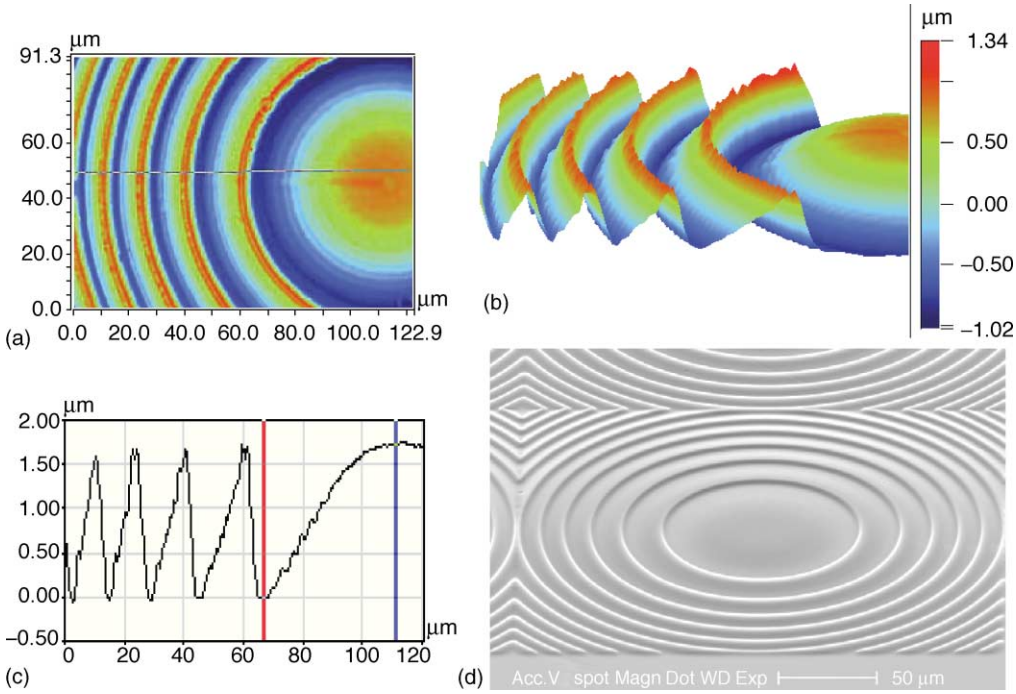


Figure 6 Characterization of a continuous phase diffractive lens. (a) Two-dimensional profile. (b) Three-dimensional profile. (c) Line scan of the two-dimensional plot. (d) Image of element from a scanning electron microscope.

insure that resist exposure is on the linear portion of the curve.

Although the resist exposure process is sensitive to all parameters, the behavior of a resist can be characterized using simulation models. The details of the photoresist models can be found elsewhere, but the simulations predict that the greatest utilization of the available resist occurs for short exposure times (seconds), and that these short exposures require long developing times (minutes). Further, the models predict that low mask contrast does not adversely affect the usable resist height. It is, therefore, not unreasonable to use a low contrast mask to achieve small features.

Through simulation it is possible to determine process parameters that maximize the linear response range of the photoresist. However, these parameters must be validated in the laboratory. The results presented in Figures 5 and 6 provide convincing evidence of the efficacy of these techniques.

Final Remarks

Advances in passive photonic devices have been made possible by leveraging fabrication processes from the electronics industry. Modifying standard photolithographic techniques and photoresist processes has produced an industrial capacity for fabricating diffractive elements. This capability will no doubt lead to future advances in telecommunications, signal processing, and imaging.

See also

Diffraction: Fresnel Diffraction. **Diffractive Systems:** Aberration Correction with Diffractive Elements. **Holography, Techniques:** Overview.

Further Reading

Dascher W, Long P, Stein R, Wu C and Lee SH (1997) Cost-effective mass fabrication of multilevel diffractive optical elements by use of a single optical exposure with a gray-scale mask on high-energy beam-sensitive glass. *Applied Optics* 36: 4675–4680.

Dillon T, Balcha A, Murakowski JA and Prather DW (2004) Continuous-tone grayscale mask fabrication using high-energy-beam-sensitive (HEBS) glass. *Journal of Microlithography, Microfabrication, and Microsystems*.

Fienup JR (1980) Iterative method applied to image reconstruction and to computer-generated holograms. *Optical Engineering* 19: 297–306.

Goodman JW (1968) *Introduction to Fourier Optics*. New York: McGraw-Hill.

Kohler M (1998) Etching in microsystem technology. In: Turunen J and Wyrowski F (eds) *Diffractive Optics for Industrial and Commercial Applications*. Berlin: Wiley-VCH.

LeCompte M, Gao X and Prather DW (2001) Photoresist characterization and linearization procedure for the grayscale fabrication of diffractive optical elements. *Applied Optics* 40: 5921–5927.

Leith EN and Upatnieks J (1962) Reconstructed wavefronts and communication theory. *Journal of the Optical Society of America A* 52: 1123–1130.

Lohmann AW and Paris DP (1967) Binary Fraunhofer holograms generated by computer. *Applied Optical* 6: 1739–1748.

Mait JN (1995) Understanding diffractive optical design in the scalar domain. *Journal of the Optical Society of America A* 12: 2145–2158.

Stern MB and Rubico-Jay T (1994) Dry etching for coherent refractive microlens arrays. *Optical Engineering* 33: 3547–3551.

Swanson GJ (1989) Binary optics technology: the theory and design of multi-level diffractive optical elements, Technical Report 854, MIT.

Swanson GJ (1991) Binary optics technology: theoretical limits on the diffraction efficiency of multi-level diffractive optical elements, Technical Report 914, MIT.

Swanson GJ and Veldkamp WB, (1990) High-efficiency, multilevel diffractive optical elements. United States Patent 4,895,790.

Wang MR and Su H (1998) Laser direct-write gray-level mask and one-step etching for diffractive microlens fabrication. *Applied Optics* 37: 7568–7576.

Wu CK (1994) High Energy Beam Sensitive Glasses. U.S. Patent 5,285,517.

Wyrowski F (1991) Digital holography as part of diffractive optics. *Report on Progress in Physics* : 1481–1571.

Diffractive Laser Resonators

U D Zeitner, Fraunhofer Institute of Applied Optics and Precision Engineering, Jena, Germany

F Wyrowski, Friedrich-Schiller University, Jena, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

All lasers consist of three basic functional components: an active medium, an optical arrangement to pump the active medium, and a resonator, which we consider in detail here. The basic function of the resonator is to back-couple a portion of the emitted laser light into the active medium for further amplification. Simultaneously, in steady-state laser operation, the geometry of the resonator defines the spatial field distribution of the generated light field.

In the most simple, and most common, case the laser resonator consists of two planar or spherical mirrors. This configuration was proposed in 1958 and has been treated extensively in the literature. However, these configurations are only special cases of a more general resonator concept in which arbitrary mirror surface profiles are used. Historically, resonators consisting of surface profiled mirrors played only a small role in laser physics. This was due, not only to difficulties in their numerical design and analysis, but also to the lack of suitable fabrication methods for the mirrors.

However, since the 1990s, these difficulties have been overcome and in 1992 it was demonstrated for the first time experimentally a stable CO_2 laser resonator that has a user-defined mode shape generated by surface structured mirrors.

In the following we derive the basic design rules for stable and unstable diffractive laser resonators and discuss the advantages and disadvantages of the technology.

Modes in Laser Resonators

Before presenting the details of diffractive laser resonators we look first at some general properties of the electromagnetic field inside a laser resonator. As is typical in laser resonator theory, we use scalar and paraxial descriptions of the fields.

The complex amplitude $U(x, y, z)$ represents one scalar, monochromatic component of the electromagnetic field. To handle wave propagation inside the resonator, we use a general operator formalism in which the influence of each optical component on the wave $U(x, y, z)$ is represented by an operator. One of the most important operators is the so-called round trip operator $\mathbf{Z}$, which describes the effect of a round trip inside the resonator on the wave $U(x, y, z)$.

If $U(x, y, z)$ denotes the wave at the beginning of a resonator round trip, the resulting wave $\tilde{U}(x, y, z)$ at its end is

$$\tilde{U}(x, y, z) = \mathbf{Z}U(x, y, z) \qquad [1]$$

With reference to Figure 1, the round trip operator can be decomposed into a series of operators that represent the different elements in the resonator. For the common two-mirror resonator a round trip from mirror 1 to mirror 2, reflection at mirror 2, back propagation to mirror 1, and final reflection at

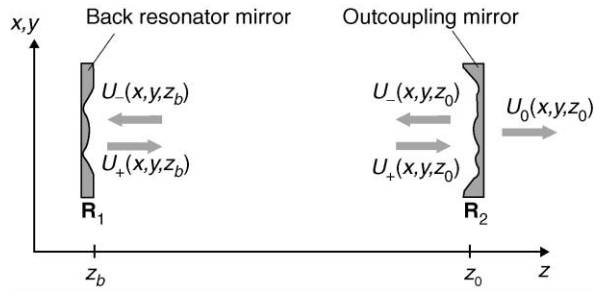


Figure 1 General diffractive laser resonator arrangement.

mirror 1, is represented by

$$\mathbf{Z} = \mathbf{R}_1\mathbf{P}_{\Delta z}\mathbf{R}_2\mathbf{P}_{\Delta z} \qquad [2]$$

The operators $\mathbf{R}_1$ and $\mathbf{R}_2$ represent the reflection at mirrors 1 and 2, respectively, and the propagation along the resonator length Δz is represented by the propagation operator $\mathbf{P}_{\Delta z}$.

In steady-state laser operation, the wave circulating inside the resonator reproduces itself, up to a constant factor γ, after a complete round trip:

$$\tilde{U}(x, y, z) = \gamma U(x, y, z) \qquad [3]$$

Together eqns [1] and [3] give the eigenvalue equation:

$$\gamma U(x, y, z) = \mathbf{Z}U(x, y, z) \qquad [4]$$

which, in general, has a set of solutions $U_p(x, y, z)$.

Thus, in steady-state, the waves $U_p(x, y, z)$ circulating inside the resonator are eigenfunctions of the round trip operator $\mathbf{Z}$ with the connected eigenvalue γ_p. The eigenfunctions of the resonator are called modes and the real-valued eigenvalues γ_p are connected with the round trip losses V, according to the equation:

$$V = 1 - |\gamma_p|^2 \qquad [5]$$

As we can see, the spatial properties of the modes are determined by the optical elements in the resonator. In the simple two-mirror resonator, the mirror profiles, which are related to the operators $\mathbf{R}_1$ and $\mathbf{R}_2$, can be used to specify the shape of the modes.

To derive the rules for resonator design we must look closer at the orthogonality of the resonator modes. The normalized eigenfunctions of the round trip operator $\mathbf{Z}$ do not obey the usual power-orthogonality relationship:

$$\int\int U_m(x, y, z)U_n^*(x, y, z)\,\mathrm{d}x\,\mathrm{d}y = \delta_{mn} \qquad [6]$$

but instead satisfy the biorthogonality relationship:

$$\int\int U_{m;+}(x, y, z)U_{n;-}(x, y, z)\,\mathrm{d}x\,\mathrm{d}y = \delta_{mn} \qquad [7]$$

The δ_{mn} is the so-called Kronecker symbol and $*$ denotes the complex conjugate. In eqn [7], $U_{m;+}(x, y, z)$ is the mode in plane z traveling in the $+z$ direction and $U_{m;-}(x, y, z)$ is the mode in the same plane z propagating in the opposite direction. The physical reason for the difference between eqns [6] and [7] are diffractive mode losses at limiting apertures. However, if a mode $U_{m;+}(x, y, z)$ travels without losses through the whole resonator, a power normalization:

$$\int\int U_m(x, y, z)U_m^*(x, y, z)\,\mathrm{d}x\,\mathrm{d}y = 1 \qquad [8]$$

is valid in all planes z. Equations [7] and [8] will be used in the next section to derive the design strategy for resonators with predefined fundamental modes.

If the mirrors, either planar or spherical, are assumed to be infinite in extent, it is possible to derive analytic solutions for the modes. This yields either the Gauss–Hermite or Gauss–Laguerre modes, which are orthogonal in the sense of eqn [6]. For resonators whose mirror shape differs from either a plane or sphere, the analytical calculation of its eigenfunctions or resonator modes is difficult and in most cases even impossible.

However, in such cases the analysis can be performed numerically. One of the most common methods is the Fox-Li-algorithm which uses the physical propagation of the modes through the resonator. The algorithm begins with a random field distribution in some plane of the resonator. This field is then numerically propagated through the resonator until a round trip is completed. The effect of the resonator optical elements must be accounted for in this procedure. At the end of the resonator round trip, the starting field will have changed due to diffraction and serve as a new starting wave for a succeeding round trip calculation. This procedure is repeated until a static field distribution is obtained, which is then an eigenfunction of the resonator.

Design Principles

The condition in eqn [7] on the biorthogonality of resonator modes, together with the eqn [8] valid for lossless modes, can be used to develop principles to design resonators with an arbitrary fundamental mode shape. Consider a stable laser resonator characterized by negligible diffraction loss of the fundamental mode. Following the discussion above, its lossless modes must simultaneously satisfy

eqns [7] and [8]. A comparison of both equations yields:

$$U_+(x, y, z) = U_-^*(x, y, z) \qquad [9]$$

For reasons of simplicity we neglect here, and in the following, the subscript m for the mode order if only one mode is considered. Because this equation must hold in all planes z, it is satisfied in the mirror resonator planes z_0 and z_b. Further, by definition of the operators $\mathbf{R}_1$ and $\mathbf{R}_2$:

$$U_+(x, y, z_b) = \mathbf{R}_1 U_-(x, y, z_b) \qquad [10]$$

$$U_-(x, y, z_0) = \mathbf{R}_2 U_+(x, y, z_0) \qquad [11]$$

The combination of eqns [10] and [11] with eqn [9] yields the following expressions for the mirror operators:

$$\mathbf{R}_1 = \frac{U_+(x, y, z_b)}{U_+^*(x, y, z_b)} \qquad [12]$$

$$\mathbf{R}_2 = \frac{U_-(x, y, z_0)}{U_-^*(x, y, z_0)} \qquad [13]$$

Equations [12] and [13] imply that lossless modes are phase-conjugated on reflection at the resonator mirrors. Because we have not made any assumptions on the shape of the mode, these equations represent a general property of lossless resonator modes. Furthermore, they provide a general design rule for the construction of stable laser resonators with predefined fundamental mode shape.

If one considers limiting apertures inside the resonator, an equivalent statement for lossy modes can be made by making assumptions about phase conjugation in a lossy system and showing that they lead to a contradiction. Without loss of generality, we assume an aperture on the outcoupling resonator mirror that clips the mode. We now consider the wave $U_-(x, y, z_0)$ after reflection at this mirror and assume this wave is still phase-conjugated after propagation through the resonator during its reflection at the back resonator mirror according to:

$$U_+(x, y, z_b) = U_-^*(x, y, z_b) \qquad [14]$$

If now the wave $U_+(x, y, z_b)$ is back propagated to the outcoupling mirror, we end up with the wave $U_+(x, y, z_0)$ which must, because of the phase conjugation in z_b, have the same amplitude as the wave $U_-(x, y, z_0)$ we started with. As a result, because both waves are connected to each other via the reflection operator of the outcoupling mirror by eqn [11], $\mathbf{R}_2$ cannot influence the amplitude distribution which, however, is a contradiction to the assumption of a limiting aperture. Thus, we finally can state that lossy modes do not have the property to be phase conjugated during their reflection at both mirrors. As a consequence, the design of diffractive mirrors for unstable laser resonators, in general characterized by energy outcoupling via diffractive losses, is more complex than simply calculating a phase-conjugating mirror structure for stable laser resonators.

Up to now, design methods described in the literature for unstable laser resonators with user-defined output beam shapes have all been based on an iterative calculation of the mirror reflection function. Because a detailed description of these methods is rather complex, we restrict ourselves to a sketch of the principal ideas behind them.

Without loss of generality, we consider an unstable laser resonator from which the light is coupled out by a mirror with spatially varying transmission and reflection. The second mirror is fully reflective within its aperture. The particular application defines only few design constraints which are in most cases the light wavelength, the dimensions of the resonator, and the shape of the outcoupled beam $U_0(x, y, z_0)$. As a consequence of the outcoupling principle, the shape of the mode inside the resonator is not finally defined by the outcoupled beam. In regions where the outcoupling mirror is reflective and not transmissive, the mode amplitude can be chosen freely. Also the amplitude of the mode in the plane of the back resonator mirror is not fixed but influenced by the amplitude and phase of the mode in the plane of the outcoupling mirror.

The general starting situation for an unstable resonator design is sketched in Figure 2. The design procedure must yield the complex amplitudes of four waves: the mode directly in front of the outcoupling mirror $U_-(x, y, z_0)$, the mode after reflection at the outcoupling mirror $U_+(x, y, z_0)$, and the complex amplitudes of the forward and reverse modes in the plane of the back resonator mirror $U_+(x, y, z_b)$ and $U_-(x, y, z_b)$, respectively. Since we assume no losses at

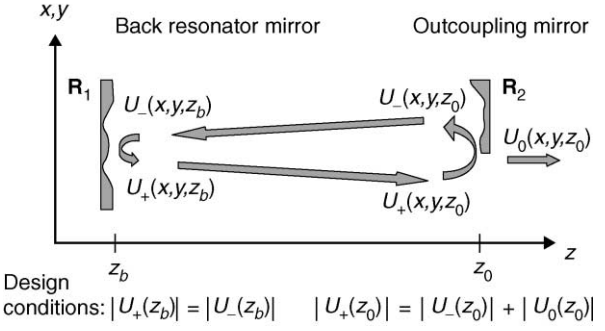


Figure 2 Starting situation for design of an unstable diffractive laser resonator.

this mirror, the amplitudes must satisfy the condition:

$$|U_+(x, y, z_b)| = |U_-(x, y, z_b)| \quad [15]$$

If we further assume no absorption at the outcoupling mirror, the amplitudes of the three waves in z_0 are related according to:

$$|U_-(x, y, z_0)| + |U_0(x, y, z_0)| = |U_+(x, y, z_0)| \quad [16]$$

To design the resonator one starts with the choice of initial intensity distributions for the wavefields $U_-(x, y, z_0)$ and $U_+(x, y, z_b)$. In order to obtain a good filling of the available mode volume, a homogeneous intensity across the mirror apertures is recommendable. To avoid strong diffraction effects at the mirror edges, the corresponding intensity distributions should be smoothed out. Together with the desired outcoupled beam amplitude $|U_0(x, y, z_0)|$ one can calculate the two remaining amplitudes for the wave fields $U_-(x, y, z_b)$ and $U_+(x, y, z_0)$ according to eqns [15] and [16].

An iterative algorithm, such as the well known Gerchberg–Saxton algorithm, can be used to find the phases of the beams $U_-(x, y, z_0)$ and $U_+(x, y, z_0)$ that insure the wavefields have almost equal amplitudes once they are propagated or back-propagated, respectively, to the plane z_b. In order to find wavefields that satisfy eqns [15] and [16], two successive iteration runs with different constraints on the complex amplitudes have to be performed. During the first iteration run no change in the initial amplitudes in plane z_0 are allowed. The iteration only look for phases for $U_-(x, y, z_0)$ and $U_+(x, y, z_0)$ which result in the amplitude $|U_+(x, y, z_b)|$ once they are propagated and back-propagated, respectively, to plane z_b. Usually the results of a Gerchberg–Saxton algorithm, with such constraints, are wavefields $U_+(x, y, z_b)$ and $U_-(x, y, z_b)$ with similar but not equal amplitude. Thus, in order to determine wavefields that satisfy eqn [15], a second iteration run with the Gerchberg–Saxton algorithm has to be added in which slight variations in wavefield amplitudes also in plane z_0 are allowed. Because the application of the constraints to ensure convergence of the iteration is rather complex, we do not go into details here but refer the reader to the Further Reading at the end of this article.

Upon completion of the iterations, the mirror reflectance operators $\mathbf{R}_1$ and $\mathbf{R}_2$ are calculated according to:

$$\mathbf{R}_1 = \frac{U_+(x, y, z_b)}{U_-(x, y, z_b)} \quad [17]$$

$$\mathbf{R}_2 = \frac{U_-(x, y, z_0)}{U_+(x, y, z_0)} \quad [18]$$

The design method described so far yields a mathematical expression for the resonator mirror reflection operator. Practical realization of the resonator requires this expression be converted into a description of a physical property of the mirror. In most practical cases this is the surface profile. Since the design method considers only phase-only functions, the physical description of the reflection operators are:

$$\begin{aligned} \mathbf{R}_1 &= \exp[-\mathrm{i}\phi_1(x, y)] \text{ and} \\ \mathbf{R}_2 &= \exp[-i\phi_2(x, y)] \end{aligned} \quad [19]$$

Using the so-called thin-element-approximation, the surface profiles $h_1(x, y)$ and $h_2(x, y)$ of the two mirrors follow directly from the phases $\phi_1(x, y)$ and $\phi_2(x, y)$:

$$\begin{aligned} h_1(x, y) &= -\frac{\lambda}{4\pi}\phi_1(x, y) \text{ and} \\ h_2(x, y) &= -\frac{\lambda}{4\pi}\phi_2(x, y) \end{aligned} \quad [20]$$

with λ being the laser wavelength. The expressions in eqn [20] contain a factor of 1/2 due to reflection. These surface profiles are commonly fabricated by microlithographic techniques. For stable laser resonators, where the mirrors perform a static phase conjugation, the mirror profile is a surface of equal phase of the fundamental mode.

Use of General Resonator Modes

Applications of diffractive laser resonators can be classified into beam shaping for a particular task external to the laser and resonator optimization via modifying the eigen-space of the round trip operator.

Internal beam shaping for external tasks is relevant to applications where it is advantageous to use a fixed beam profile other than a Gaussian, as is the case for instance in materials processing. Shaping the beam internally, avoids the energy losses usually associated with external beam shaping techniques. However, lasers that have a nonGaussian outcoupled beam profile are often less flexible in their general applicability.

However, there is considerable potential for diffractive laser resonators to enhance resonator characteristics and thereby improve the lasing operation itself. The conventional Gaussian-shaped fundamental mode in customary laser resonators varies considerably in intensity from its central maximum to its outer wings. The transfer of energy from the pumped active medium into the propagating mode is dependent upon the local transverse intensity and is

optimal for homogeneous illumination. As a result, the efficiency with which energy is transferred into the fundamental mode can be increased by using diffractive resonator mirrors to match the shape of the lateral fundamental mode to that of the cross-section of the active medium.

An important resonator property, especially in high power laser systems, is the difference of round trip losses V between the fundamental mode and the higher, usually undesired, laser modes. This difference is referred to as modal discrimination. With no limiting apertures, the wavefronts of all Gauss–Hermite or Gauss–Laguerre modes of common laser resonators are phase conjugated upon reflection at the spherical mirrors. As we have shown above, there is a tight connection between the phase conjugation property of resonator modes and their property of having no round trip losses. Thus, the phase conjugation property of Gauss–Hermite and Gauss–Laguerre modes directly implies small round trip losses. In fact, in resonators without apertures they all have an eigenvalue $\gamma_p = 1$ and according to eqn [5] the round trip losses are zero. High-order modes, whose diameters are large, can be suppressed by decreasing the size of the mirror aperture, however, because the difference in mode diameter between modes of neighboring order m is small, the achievable discrimination is limited to a maximum ratio $V_{m+1}/V_m \approx 10$. The structured surface profile of the mirrors in a diffractive laser resonator whose fundamental mode is nonGaussian is tightly adapted to the phase of the design mode. The wavefront of the higher-order modes in these resonators often do not fit the mirror profile and it can be observed that their round trip losses are significantly higher. Thus, diffractive laser resonators usually show an improved modal discrimination compared to resonators with spherical mirrors. This effect can be further improved by inserting into the resonator additional phase plates with diffuser-like properties.

Shaping the surface profile of resonator mirrors or insertion of additional phase plates significantly improves the flexibility for the resonator design. As is demonstrated by the examples below, there are resonator configurations in which high modal discrimination or adaptation of the mode volume to the cross-section of the active medium are combined with a Gaussian shape of the outcoupled beam.

In the design procedure described so far, the resonators were treated as passive optical systems. The propagation of the mode through the resonator has not been influenced by the active medium. However, in real laser systems, this simplification is often not applicable. Due to the pumping process and the connected heat insertion, the refractive index distribution inside the active medium is not homogeneous, which affects the propagation of the modes. Furthermore, changes or fluctuations in the pump energy and cooling mechanisms introduce temporal changes in the optical function associated with the refractive index of the active medium. Further, although static or averaged, optical power of the active medium are included in the round trip operator and calculation of the mirror profiles, experiments with lasers that have diffractive resonator mirrors indicate that the temporal fluctuations can strongly influence the shape and round trip losses of the fundamental resonator mode. In general, diffractive laser resonators are more sensitive to deviations from design than conventional resonators.

Examples

We consider as an example the adaptation of fundamental mode shape to the cross-section of the active medium. In solid state lasers, such as Nd:YAG lasers, the active medium is often a crystal rod with a circular cross-section. Instead of using this sharp-edged cross-section shape as intensity distribution for the desired fundamental mode, a smoothed version, which is approximated by a super-Gaussian intensity distribution in the center plane of the rod, is chosen to be the fundamental mode. The super-Gaussian intensity distribution has a more desirable propagation characteristic than a sharp circular one. However, even these modes do not exhibit the advantageous property of propagation-invariant beam shape characteristic of Gaussian beams.

In our first example, we consider the design of a stable diffractive laser resonator with a super-Gaussian fundamental mode shape inside the active medium and a Gaussian beam coupled out from the resonator. The resonator configuration is based upon a beam shaping arrangement for a Gauss to super-Gauss beam transformation. Its optical setup is sketched in Figure 3. In the plane of the outcoupling

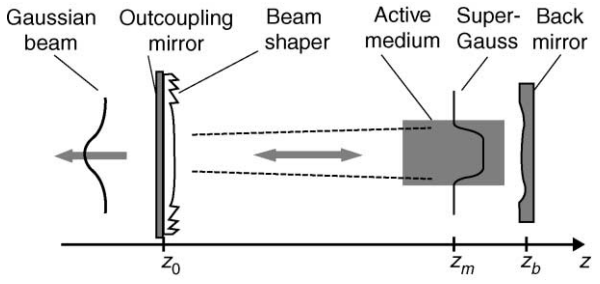


Figure 3 Setup for the Gauss-to-super-Gauss resonator.

mirror z_0, the fundamental mode has a Gaussian amplitude $U_g(x, y, z_0)$. A phase element, which is located directly in front of the outcoupling mirror, shapes the mode amplitude into a super-Gaussian one $U_{sg}(x, y, z_m)$ in the center plane z_m of the active medium. For this special beam shaping problem it is possible to calculate a phase plate surface profile that transforms the beams without energy losses. The corresponding calculation procedure is described in the Further Reading section at the end of this article.

The mirror profiles have to be realized to satisfy the previously derived phase conjugation property for the described beam. Since the Gaussian beam has its minimum waist at the outcoupling mirror, its phase is planar in that plane. Thus, to insure phase conjugation a plane mirror is all that is required in front of the beam shaping phase plate. In the plane of the back resonator mirror the beam has a modulated wavefront due to its transformation from a Gaussian to a super-Gaussian amplitude distribution. Thus, a surface modulated mirror profile is required to phase conjugate this beam during reflection. Figures 4a and b represent the phase profile of the beam shaping element and the back resonator mirror, respectively, for the following set of parameters: laser wavelength $\lambda = 1064$ nm, radius of the Gaussian and super-Gaussian beams in the plane z_0 and z_m, respectively, $w_g = 1$ mm and $w_{sg} = 1$ mm, order of the super-Gauss $n = 10$, distance $z_m - z_0 = 200$ mm, and resonator length $z_b - z_0 = 250$ mm. The amplitude distribution for the fundamental mode in this resonator simulated with the Fox–Li-algorithm is shown in Figure 4c for the plane of the outcoupling mirror and in Figure 4d for the center of the active medium.

In our second example, the flexibility of diffractive laser resonators is demonstrated by a resonator with a user-defined far-field intensity distribution. In our case the far-field intensity $|U_f(x, y, z_f)|^2$ is shaped like the Greek letter lambda, λ. Because we are concerned only with the intensity of the wavefield, the phase $\phi_f(x, y, z_f)$ of the far-field $U_f(x, y, z_f)$ can be chosen freely. As shown in the following, this phase freedom can be used to improve the interaction of the mode with the active medium, as was the case in our first design example. Again, the mode shape in the center of the active medium has a super-Gaussian intensity distribution, which insures high extraction efficiency. We used an iterative Fourier-transform algorithm to calculate a far-field phase that yields a nearly super-Gaussian intensity distribution in the outcoupling plane of the resonator. The resulting wave:

$$U_0(x, y, z_0) = a_0(x, y, z_0) \exp[i\varphi_0(x, y, z_0)] \qquad [21]$$

is generated directly as the laser output. The amplitude $a_0(x, y, z_0)$ is generated as a mode inside the cavity using a surface-modulated back resonator mirror. This performs a static phase conjugation of the desired mode in the mirror plane. The phase $\varphi_0(x, y, z_0)$ is provided by an additional element on the outside of the outcoupling mirror. A sketch of the resonator configuration is shown in Figure 5.

For the experimental realization of this example the thermal index gradient of the active medium was included into the resonator design and the two diffractive elements were fabricated on a quartz substrate using electron beam lithography and ion-beam etching. The measured mode shape in the plane of the outcoupling mirror and the far-field

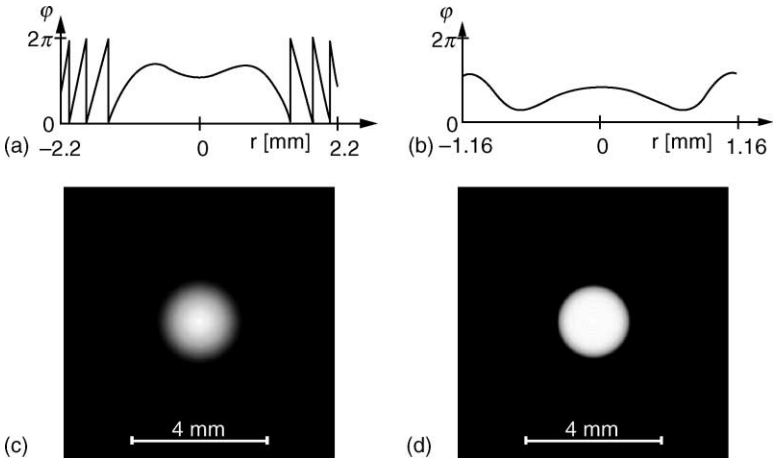


Figure 4 Gauss-to-super-Gauss resonator: (a) phase of beam shaping element; (b) phase of back resonator mirror, simulated amplitude distributions for the fundamental resonator mode; (c) in the plane of the outcoupling mirror; and (d) in the center of the active medium.

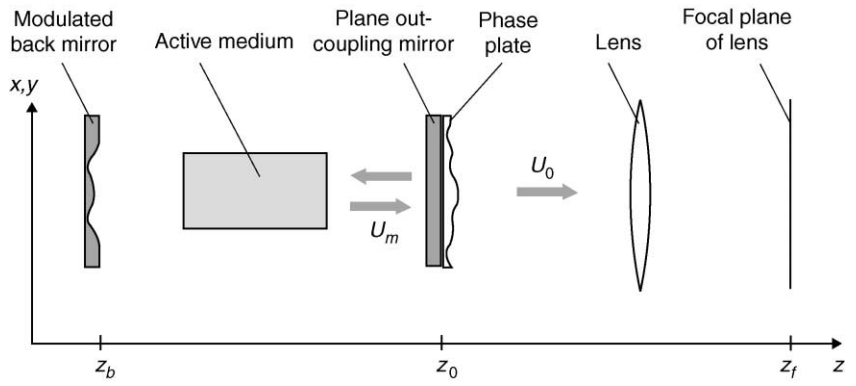


Figure 5 Configuration for resonator internal generation of a user defined far-field intensity distribution. (With permission from Zeitner UD, Wyrowski F and Zellmer H (2000) External design freedom for optimization of resonator originated beam shaping. *IEEE Journal of Quantum Electronics* 36: 1105–1109. © 2001 IEEE.)

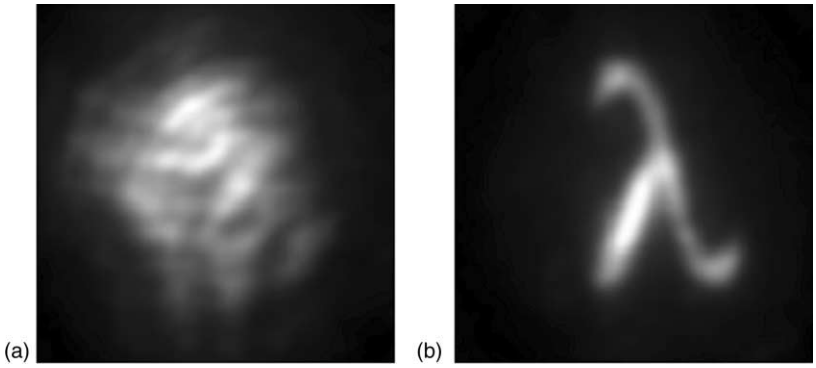


Figure 6 Measured mode intensities for the resonator of **Figure 5**. (a) In the plane of the outcoupling mirror; and (b) in the far-field of the resonator. (With permission from Zeitner UD, Wyrowski F and Zellmer H (2000) External design freedom for optimization of resonator originated beam shaping. *IEEE Journal of Quantum Electronics* 36: 1105–1109. © 2000 IEEE.)

intensity distribution are shown in **Figures 6a** and **b**, respectively. The mode in the outcoupling plane of the resonator fills an area of about 2 mm diameter with nearly homogeneous intensity. The far-field shows the desired λ-shaped intensity distribution. The variations in intensity are due to the nonlinear response of the active medium. As mentioned above, the response can be improved if these effects are included in the design of the resonator mirrors.

Conclusion

Diffractive laser resonators are a generalization of common spherical mirror resonators. The use of arbitrary resonator mirror profiles and additional internal phase plates offer a huge flexibility for the generation of user defined mode shapes and optimization of resonator properties like extraction efficiency and modal discrimination. The drawback of the use of structured mirrors is an increased adjustment effort, because the diffractive mirrors have more degrees of freedom to be aligned. Furthermore, diffractive resonators tend to be more sensitive against changes of the optical properties of the active medium than spherical mirror resonators. Such changes may occur due to temporal fluctuations in pump power or cooling power.

Most promising applications for diffractive resonators are:

- lasers for materials processing which require a fixed intensity pattern for the sample treatment and a fixed laser power;
- high efficient homogeneous laser illumination, for example, for data projection or sensing applications; and
- suppression of higher order modes and maintaining single-mode operation for high brightness lasers.

Some example applications and solutions have been presented here which cannot be achieved by conventional resonator optics. For stable laser resonators, the phase-conjugation property of the

modes is used for a straightforward calculation of the required diffractive mirror profiles. Unstable resonator geometries with diffractive outcoupling require more complex iterative design procedures. However, especially for high-power and high-gain lasers, the latter sometimes have more advantageous properties, such as better discrimination of undesired higher-order modes and more simple mirror cooling options.

See also

Lasers: Carbon Dioxide Lasers. **Phase Control:** Phase Conjugation and Image Correction.

Further Reading

Aagedal H, Wyrowski F and Schmid M (1997) Paraxial beam splitting and shaping. In: Turunen J and Wyrowski F (eds) *Diffractive Optics for Industrial and Commercial Applications*, pp. 165–188. Berlin: Akademie Verlag.

Belanger PA, Lachance RL and Pare C (1992) Super-Gaussian output from a CO_2 laser by using a graded-phase mirror resonator. *Optics Letters* 17: 739–741.

Fox A and Li T (1961) Resonant modes in a maser interferometer. *Bell Systems Technical Journal* 40: 453–488.

Gerchberg R and Saxton W (1972) A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* 35: 237–246.

Herzig H (1997) *Micro Optics*. London: Taylor & Francis.

Leger JR (1997) Diffractive laser resonators. In: Turunen J and Wyrowski F (eds) *Diffractive Optics for Industrial and Commercial Applications*, pp. 165–188. Berlin: Akademie Verlag.

Leger JR, Chen D and Dai K (1994) High modal discrimination in a Nd:YAG laser resonator with internal phase gratings. *Optics Letters* 19: 1976–1978.

Makki S and Leger JR (2001) Mode shaping of a graded-reflectivity-mirror unstable resonator with an intracavity phase element. *IEEE Journal of Quantum Electronics* 37: 80–86.

Siegman AE (1986) *Lasers*. Mill Valley, CA: Univ. Science Books.

Zeitner UD and Wyrowski F (2001) Design of unstable laser resonators with user-defined mode shape. *IEEE Journal of Quantum Electronics* 37: 1594–1599.

Zeitner UD, Wyrowski F and Zellmer H (2000) External design freedom for optimization of resonator originated beam shaping. *IEEE Journal of Quantum Electronics* 36: 1105–1109.

Diffractives in Animals

A R Parker, University of Oxford, Oxford, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Structural coloration involves the selective reflectance of incident light by the physical nature of a structure. Although the color effects often appear considerably brighter than those of pigments, structural colors often result from completely transparent materials. In addition, animal structures can be designed to provide the opposite effect to a light display: they can be antireflectors, causing 'all' of the incident light to penetrate a surface (like glass windows, smooth surfaces cause some degree of reflection).

The study of natural structural colors began in the seventeenth century, when Hooke and Newton correctly explained the structural colors of silverfish (Insecta) and peacock feathers (Figure 1), respectively. Nevertheless, accurate, detailed studies of the mechanisms of structural colors did not begin until the introduction of the electron microscope in 1942.

Structural colors generally may be formed by one of three mechanisms: thin-film reflectors, diffraction gratings, or structures causing scattering of light waves. This is a simplistic characterization resulting from the application of geometric optics to the reflective elements. Some structures, however, rather fall between the above categories, such as photonic crystals (Figure 2) – these become apparent when quantum optics is considered. In some cases, this has led to confusion over the reflector type. For example, the reflectors in some scarab beetles have been categorized by different authors as multilayer reflectors, three-dimensional diffraction gratings, photonic crystals, and liquid crystal displays. Perhaps all are correct! Since the above categories are academic, I place individual cases of structural colors in their most appropriate, not unequivocal, category. Also, in this article I will employ geometric optical treatments only.

The array of structural colors found in animals today results from millions of years of evolution. Structures that produce metallic colors have also been identified in extinct animals. Confirmation of this



Figure 1 Newton's subject.

fact, from ultrastructural examination of exceptionally well-preserved fossils, such as those from the Burgess Shale (Middle Cambrian, British Columbia), 515 million years old, permits the study of the role of light in ecosystems throughout geological time, and consequently its role in evolution. In some cases, such as for ostracod crustaceans, diffraction gratings have been found to increase in their optical sophistication throughout 350 million years of evolution.

Multilayer Reflectors

Light may be strongly reflected by constructive interference between reflections from the different interfaces of a stack of thin films (of actual thickness d) of alternately high and low refractive index (n). For this to occur, reflections from successive interfaces must emerge with the same phase and this is achieved when the so-called 'Bragg Condition' is fulfilled. The optical path difference between the light reflected from successive interfaces is an integral number of wavelengths and is expressed in the equation:

$$2nd \cos \Theta = (m + 1/2)\lambda \qquad [1]$$

from which it can be seen that the effect varies with angle of incidence (Θ, measured to the surface normal), wavelength (λ), and the optical thickness of the layers (nd). There is a phase change of half a wavelength in waves reflected from every low to high refractive index interface only. The optimal narrowband reflection condition is therefore achieved where the optical thickness (nd) of every layer in the stack is a quarter of a wavelength. In a multilayer consisting of a large number of layers with a small variation of index the process is more selective than one with a smaller number of layers with a large difference of index. The former therefore gives rise to more saturated colors corresponding to a narrow spectral bandwidth and these colors therefore vary more with a change of angle of incidence. Both conditions can be found in animals – different colored effects are appropriate for different functions under different conditions. For an oblique angle of incidence, the wavelength of light that interferes constructively will be shorter than that for light at normal incidence. Therefore, as the angle of the incident light changes, the observed color also changes (Figure 3).

Single layer reflectors are found in nature, where light is reflected, and interferes, from the upper and lower boundaries (Figure 4). A difference in the thickness of the layer provides a change in the color observed from unidirectional polychromatic light. The wings of some houseflies act as a single thin film and appear different colors as a result of this phenomenon. A single quarter-wavelength film of guanine in cytoplasm, for example, reflects about 8% of the incident light. However, in a multilayer reflector with ten or more high index layers, reflection efficiencies can reach 100%. Thus, animals possessing such reflectors may appear highly metallic.

The reflectance of the multilayer system increases very rapidly with increasing number of layers. If the dimensions of the system deviate from the quarter-wave condition (i.e., nd is not equal for all layers), then the reflector is known as 'nonideal' in a theoretical sense (may be 'ideal' for some natural situations). 'Nonideal' reflectors have a reduced proportional reflectance (not always a significant reduction) for a given number of layers and this

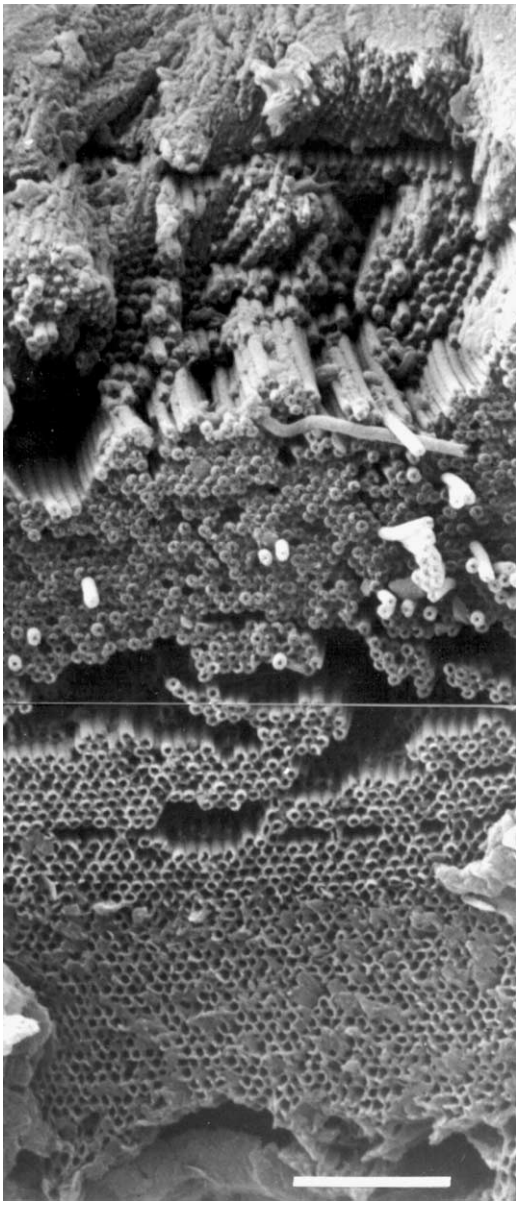


Figure 2 Scanning electron micrograph of a cross-section of the wall of a cylindrical spine of the sea mouse *Aphrodita* sp. (Polychaeta). The wall is composed of small cylinders with varying internal diameters (increasing with depth in the stack), arranged in a hexagonal array, that form a photonic crystal fiber. The opal-type photonic crystal – hexagonally packed nano-spheres (250 nm diameter) – are also found in animals (e.g. weevils). Scale bar represents 8 μm.

reflectance has a narrower bandwidth. A narrow bandwidth, less conspicuous reflection is sometimes selected for in animals, this will be discussed later in this article. Multilayer reflectors polarize light incident at Brewster's angles. This is about 54° for a quarter-wave stack of guanine and cytoplasm. At very oblique angles, all wavelengths are strongly reflected.

Multilayer reflectors are common in animals. They are usually extra-cellular, produced by periodic secretion and deposition, but sometimes occur within cells. Guanine ($n = 1.83$) is a common component in invertebrate reflectors because it is one of very few biological materials with a high refractive index and is readily available to most invertebrates as a nitrogenous metabolite. However, arthropods, including insects, crustaceans, and spiders, have largely ignored guanine in favor of pteridines. Also surprising is that the reflector material of closely related species, e.g., the molluscs *Pecten* (scallop) and *Cardium* (cockle), may differ.

Multilayers produce effects in beetle cuticle from highly metallic colors ('ideal' system) to rather dull greens ('nonideal' system in combination with scattering; Figure 5), and colors from the wings of many butterflies. Often in butterflies, layers of chitin ($n =$ about 1.56) are supported by vertical vanes of the scales. Air ($n = 1.0$) fills in the spaces and provides the alternate layers of the system. Under white light at normal incidence, the blue color of the butterfly *Arhopala micale* turns to green when the air is replaced by acetone. This is due to an increase in the refractive index of the low n layer (the actual thickness of this layer remains the same but the optical thickness increases) that effectively makes the system less 'ideal'. If the refractive indices of both layers were made equal, there would be no internal interference. A layer of melanin often underlies the reflector and intensifies the metallic colored effect by absorbing the transmitted portion of incident light. For example, in beetles, the elytra of *Anopognathus parvulus* appears metallic gold, green, or yellow in reflected light, and diffuse brown in transmitted light. Individual butterfly scales have been examined in detail to reveal a number of variations of quarter-wave stacks, sometimes in combination with other optical structures, to provide a range of colored effects.

The crustaceans *Limnadia* (Conchostraca), *Tanais tennicornis* (Tanaidacea), *Ovalipes molleri* (Decapoda), and the males of *Sapphirina* (Copepoda) all bear multilayer reflectors in their cuticles, in different forms (Figure 6). In contrast to the usual continuous thin layers, male sapphirinids have ten to fourteen layers of interconnecting hexagonal platelets within the epidermal cells of the dorsal integument. The reflector of *O. molleri* comprises layers that are corrugated and also slightly out of phase (Figure 7). At close to normal incidence this structure reflects red light, but at an angle of about 45° blue light is reflected. The corrugation, however, functions to broaden the reflectance band, at the expense of reducing the intensity of reflection.

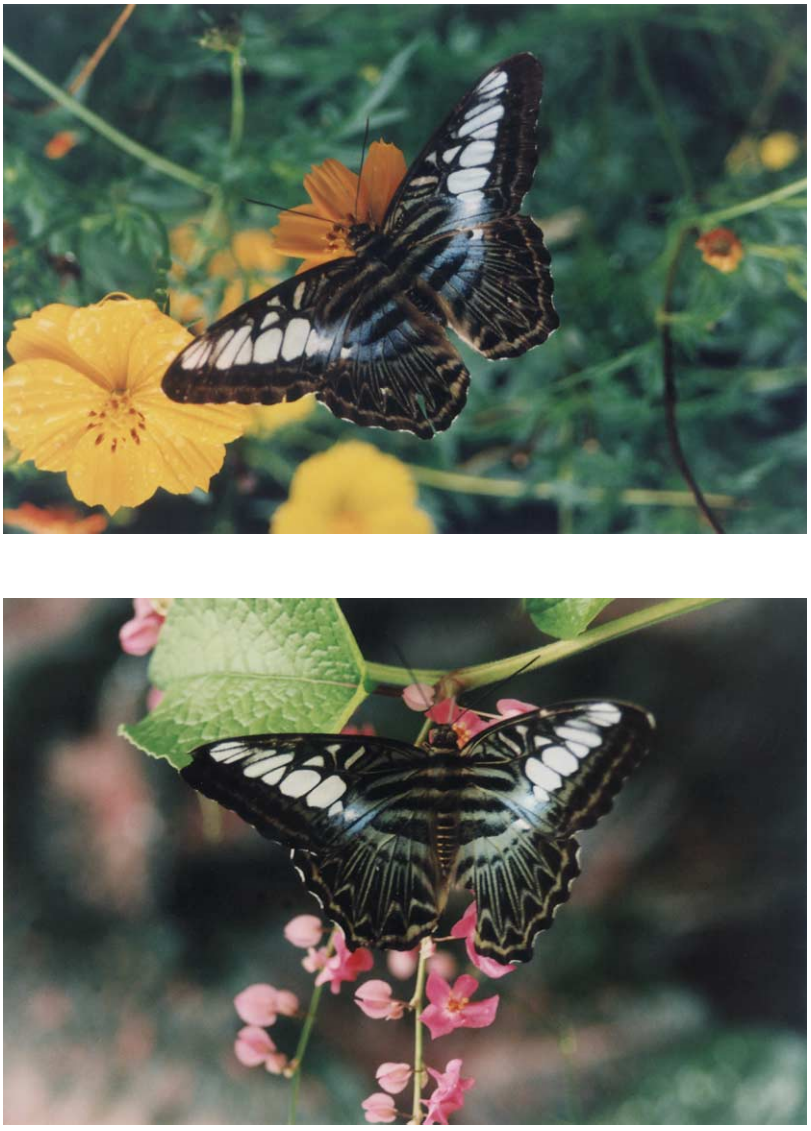


Figure 3 Butterfly (*Parthenos sylvia* (Nymphalidae) from Malaysia), showing the change in color with angle of wings. This animal may have evolved a temporal signal rather than a static one.

A broadband wavelength-independent reflectance, appearing silver or mirror-like to the human eye, can be achieved in a multilayer stack in at least three ways in invertebrates (Figure 8). These are (a) a composite of regular multilayer stacks each tuned to a specific wavelength, (b) a stack with systematically changing optical thicknesses with depth in the structure, termed a 'chirped' stack, and (c) a disordered arrangement of layer thicknesses about a mean value, termed a 'chaotic' stack (Figure 8). The nauplius eye of the copepod *Macrocyclops* (Crustacea) has regularly arranged platelets ~100 nm thick in stacks of 20–60, achieving the first condition. Silver beetles and the silver and gold chrysalis of butterflies in the genera *Euoplea* and *Amauris*, owe their reflection to the second condition. The mirror-like reflectors in the scallop *Pecten* eye comprise alternating layers of cytoplasm ($n = 1.34$) and guanine crystals ($n = 1.83$) and approximate an 'ideal' quarter-wave system in the same manner as within fish skin using the third mechanism. The ommatidia of the superposition compound eyes of *Astacus* (Crustacea) are lined with a multilayer of isoxanthopterin (a pteridine) crystals, which again fall into the third category. Multilayer reflectors can also be found in the eyes of certain spiders, butterflies, and possibly flies, where they assist vision, as discussed further below.

Squid and cuttlefish, for example, possess mirror-like reflectors in photophores (light organs) and

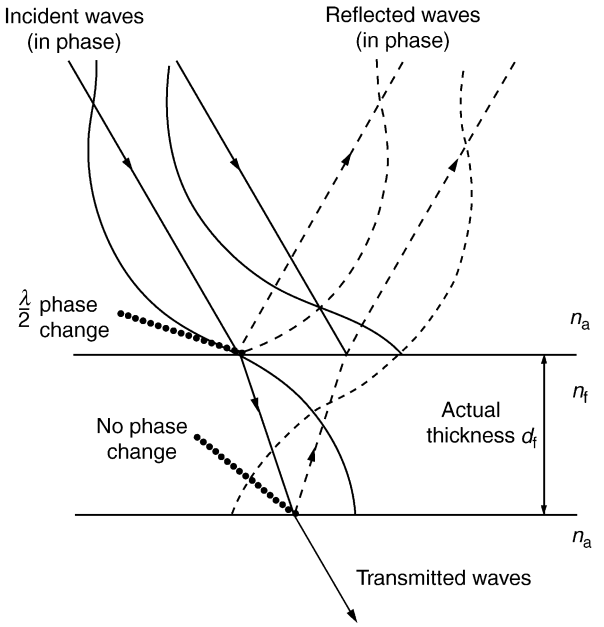


Figure 4 Schematic diagram of thin-film reflection. The direction of wave (straight line) and profile of electric (or magnetic) component are illustrated. Incident waves are indicated by a solid line, reflected waves by broken lines. Refraction occurs at each media interface. The refractive index of the film (n_f) is greater than the refractive index of the surrounding medium (n_a). Constructive interference of the reflected waves is occurring. As the angle of incidence changes, different wavelengths constructively interfere. At normal incidence constructive interference occurs where $n_f \times d_f = \lambda/4$.

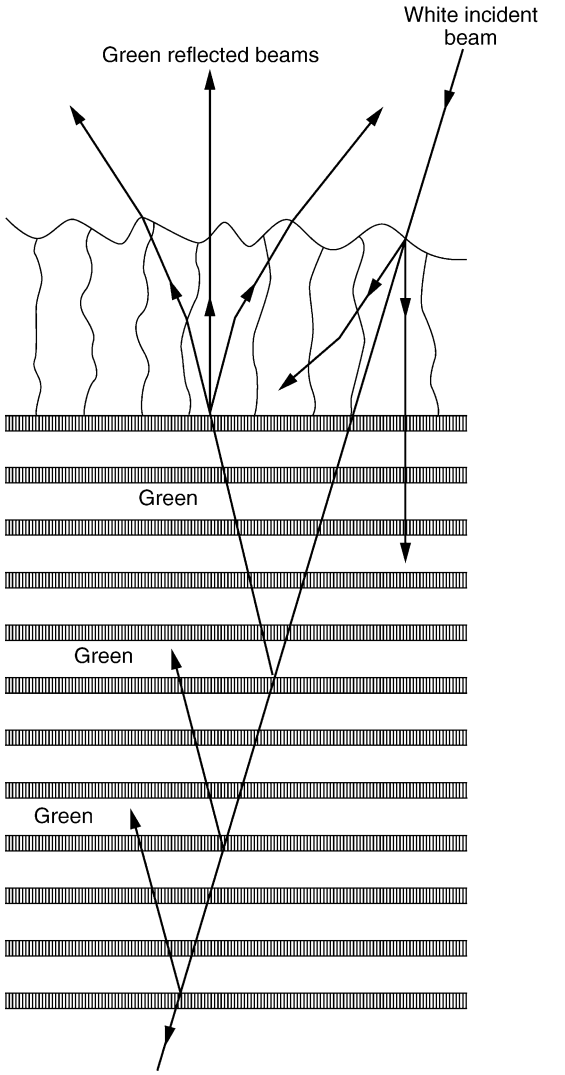


Figure 5 Generalized diagram of a multilayer reflector in the cuticle of the green beetle *Calloodes grayanus* (high refractive index material is shown shaded). Outer layer causes scattering.

iridophores. Iridophores are cells that, in this case, contain groups (iridosomes) of flexible layers of thin lamellae with cytoplasm between, forming a quarter-wave stack. The platelets of both squids and octopods develop from the rough endoplasmic reticulum and are separated by extracellular space. Euphausiid crustaceans possess photophores with very elaborate mirror-like reflectors. Up to 60 dense layers, about 70 nm thick and 75–125 nm apart, are formed from the aggregation of granules (probably a type of chitin) and surround the main photogenic mass. An intricate ring consisting of very flattened cells, forming the dense layers of a multilayer reflector (about 175 nm thick, separated by 90 nm) surrounds the lens of the photophore and reflects blue light at acute angles of incidence.

Dead invertebrates may not display their original colors. Following death, one (or both) of the layers in a multilayer reflector may become gradually reduced. For example, water may be lost from the system. This occurs in beetles of the genus *Coptocycla*; their brassy yellow color quickly changes through green, blue, and violet until the brown of melanin is finally observed. The color progression may subsequently be reversed by water uptake. This is an important consideration when examining fossils for multilayer reflectors. Multilayer reflectors have been identified in the shells of some ammonites, preserved as fossils.

Diffraction Gratings

When light interacts with a periodic surface consisting for example of a series of parallel grooves, it may be deviated from the direction of simple transmission or reflection. For this to happen, light which is scattered or diffracted from successive grooves should be out of phase by integral values of 2π. This occurs when, for a given direction of propagation, the optical path difference via successive grooves is $M\lambda$, where M is an integer known as the circle number.

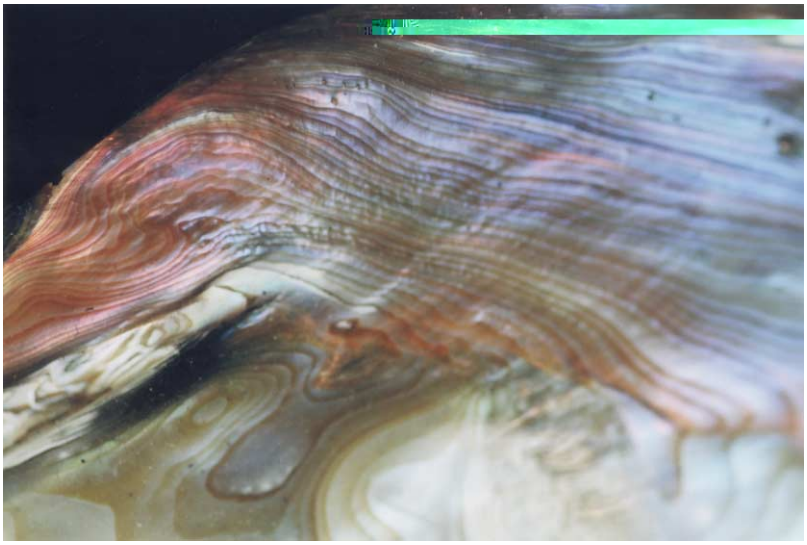


Figure 6 Bivalve shell with outer pigment removed to reveal an incidental multilayer reflector. Layers evolved to provide strength and prevent crack propagation.

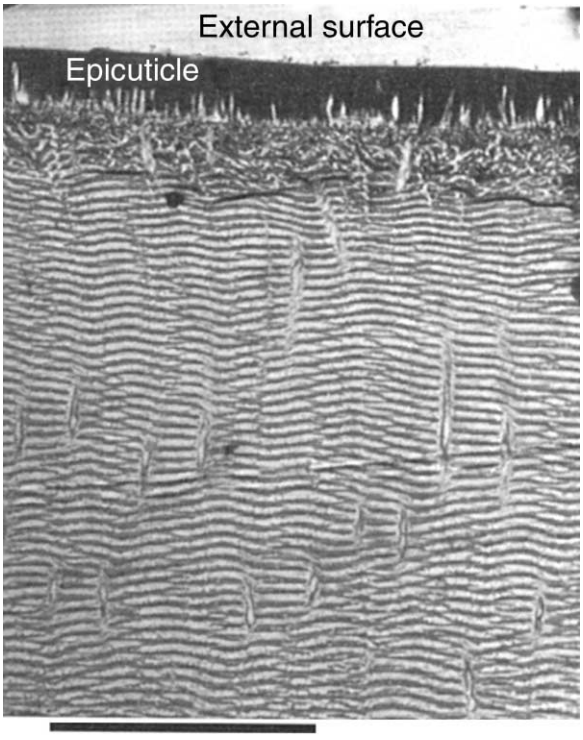


Figure 7 Transmission electron micrograph of a multilayer reflector in the cuticle of a swimming paddle of the crab *Ovalipes molleri* (Crustacea: Decapoda). Layers of reflector are slightly sinuous and out of phase; note the unusual side branches of each high refractive index (dark) layer, which provide support for these solid layers within a liquid matrix. Scale bar represents 5 μm.

This may be expressed by the grating equation:

$$2d(\sin\alpha - \sin\beta) = M\lambda \qquad [2]$$

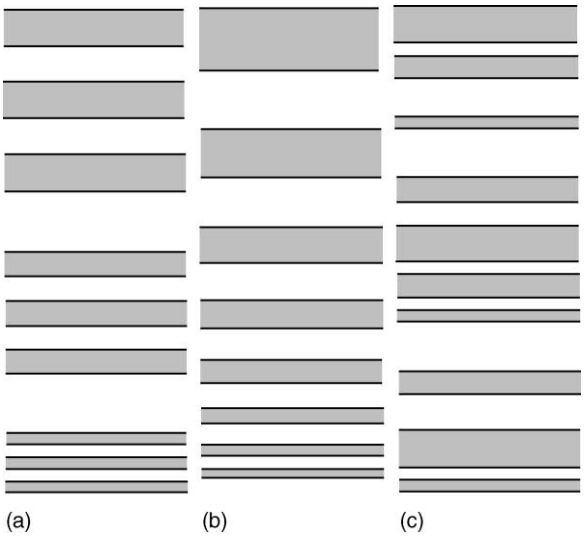


Figure 8 Three ways of achieving a broadband, wavelength-independent reflector in a multilayer stack (high refractive index material is shown shaded). (a) Three quarter-wave stacks, each tuned to a different wavelength. (b) A 'chirped' stack. (c) A 'chaotic' stack.

where α and β are angles of incidence and diffraction, and d is the period.

As with multilayers, the effect gives rise to coloration because different wavelengths are diffracted into different directions. Although the effect changes with angle of incidence, it is less critical than it is with thin films and the visual appearance is different. For a parallel beam of white light incident upon a multilayer, one wavelength will be reflected as determined by the Bragg Condition. The same beam

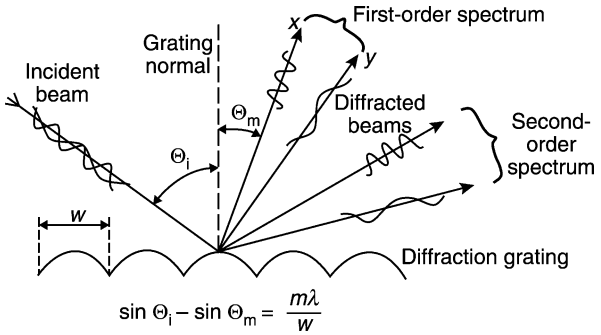


Figure 9 Reflection-type diffraction grating (w = periodicity, m = order of reflected beam). For white light of fixed angle of incidence (Θ_i), the color observed is dependent on the point of observation (e.g., violet light can be seen at point *x*, red at point *y*, within the first order spectrum).

incident upon a grating will be dispersed into spectra (e.g., Figure 9). The complete spectrum reflected nearest to the perpendicular (grating normal) is the first order. The first-order spectrum is reflected over a smaller angle than the second-order spectrum, and the colors are more saturated and appear brighter within the former. Diffraction gratings have polarizing properties, but this is strongly dependent on the grating profile.

Diffraction gratings were believed to be extremely rare in nature, but have recently been revealed to be common among invertebrates. They are particularly common on the seate or setules (hairs) of Crustacea, such as on certain first antennal setules of male Myodocopina ostracods or 'seed shrimps' (Crustacea) (Figure 10). Here, the grating is formed by the external surface of juxtaposed rings with walls circular in cross-section (Figure 11). The width of the rings, and consequently the periodicity of the grating, is about 700 nm in *Azygocypridina lowryi*. Different colors are viewed with varying angles of observation under a fixed light source. The ostracod *Euphilomedes carcharodonta*, for example, additionally houses a diffraction grating on the rostrum, a continuous flattened area of the carapace that is corrugated to form periodic ridges. The dark brown beetle *Serica sericea*, bears gratings on its elytra with 800 nm periodicity, which causes a brilliant iridescence in sunlight.

Many polychaetes possess gratings on their setae (hairs). For example, the opheliid *Lobochesis longiseta* bear gratings with periodicities in the order of 500 nm (Figure 12), appearing iridescent. The wings of the neurochaetid fly *Neurotexis primula* bear diffraction gratings only on their dorsal surfaces, and the iridescent effect remains after the insect is gold coated for electron microscopy. These gratings cause iridescence with a higher reflectance than the

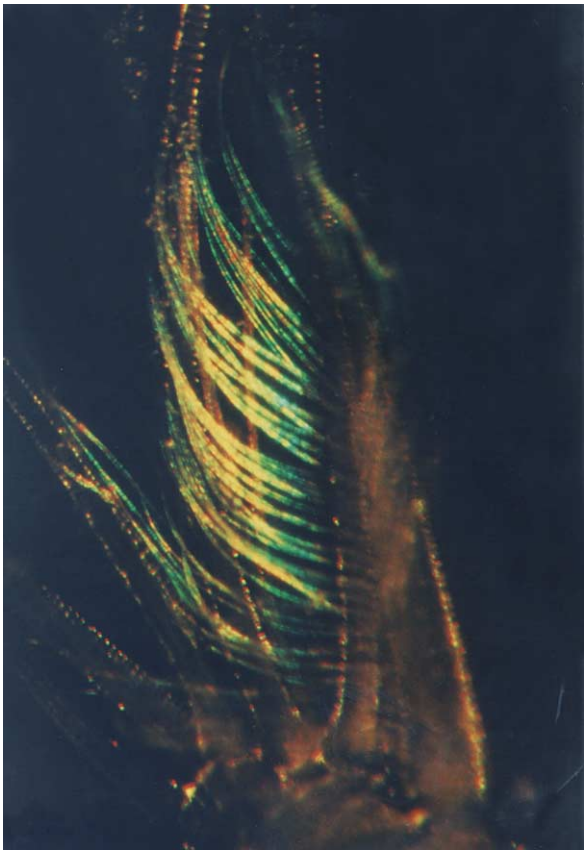


Figure 10 Diffraction gratings on the halophores (hairs) on the first antenna of the seed shrimp *Azygocypridina lowyri* (Ostracoda).

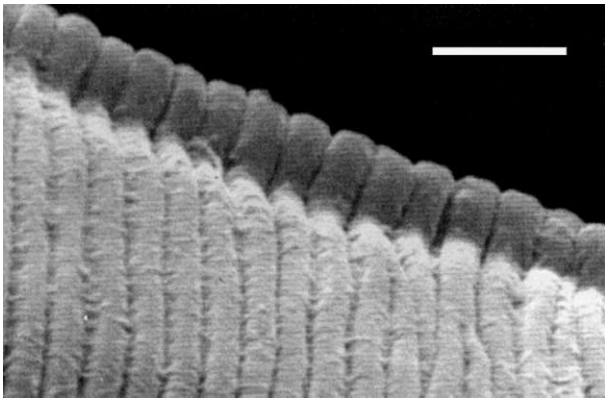


Figure 11 Scanning electron micrograph of diffraction grating on a halophore (setule) of *Azygocypridina lowryi* (Ostracoda). Scale bar represents 2 μm.

iridescence of the membranous wings of other insects, which reflect light by interference. Iridescence caused by interference disappears after gold coating because transmission of light through the outer surface is prevented.

Very closely spaced, fine setules may also form the ridges of a diffraction grating. Cylindroleberidid

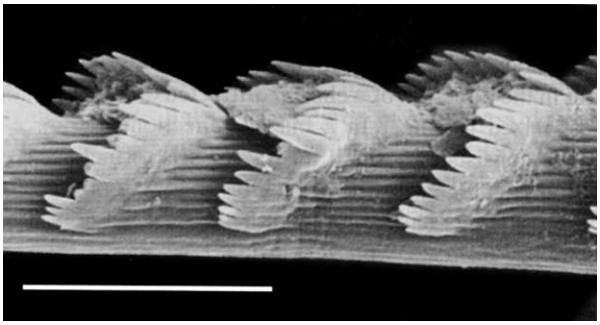


Figure 12 Scanning electron micrograph of diffraction gratings on a hair of *Lobochesis longiseta* (Polychaeta). Scale bar represents 10 μm.

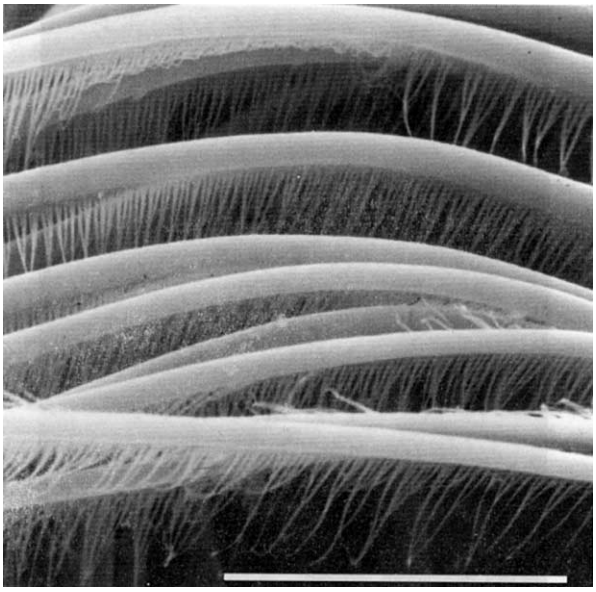


Figure 13 Scanning electron micrograph of *Tetraleberis brevis* (Ostracoda), hairs of comb of maxilla; setules (orientated vertically) form the ridges of a grating. Scale bar represents 2 μm.

ostracods (seed-shrimps) possess a comb on their maxilla bearing numerous setules on each seta (hair), collectively forming a grating with a periodicity of about 500 nm (Figure 13).

The 'helicoidal' arrangement of the microfibrils comprising the outer 5–20 μm of the cuticle (exocuticle) of certain scarabeid beetles, such as *Plusiotis resplendens*, also gives rise to metallic colors. Here, the fibrils are arranged in layers, with the fibril axis in each layer arranged at a small angle to the one above, so that after a number of layers the fibrillar axis comes to lie parallel to the first layer. Thus, going vertically down through the cuticle, two corresponding grating layers will be encountered with every 360° rotation of the fibrils – the 'pitch' of the system. Polarized light encounters an optically reinforcing plane every half turn of the helix. The system can be treated as a three-dimensional diffraction grating, with a peak reflectance at $\lambda = 2nd$, where d is the separation of analogous planes, or half the pitch of the helix. The diffracted light resembles that from a linear grating except for the polarization; the three-dimensional grating reflects light that is circularly or elliptically polarized. It should be noted that the diffracted color does not depend on the total film thickness as it does in interference, but on the layer repeat distance within the film as in the diffraction grating (analogous to 'liquid crystals').

When each groove of a grating is so formed that it independently, by means of geometrical optics, redirects the light in the direction of a chosen diffracted order, it is known as a blazed grating. In a blazed reflection grating, each groove consists of a small mirror (or prism) inclined at an appropriate angle (i.e., the grating has a 'saw-tooth' profile). Blazed gratings have been identified on the wing scales of the moth *Plusia argentifera* (Plusinae).

When the periodicity of a grating reduces much below the wavelength of light, it becomes a zero-order grating and its effect on light waves changes. This difference in optical effect occurs because when the periodicity of the grating is below the wavelength of light the freely propagating diffracted orders are suppressed and only the zero-order is reflected when the illumination is normal to the plane of the grating. To describe accurately the optical properties of a zero-order grating, rigorous electromagnetic theory is required. In contrast to gratings with freely propagating orders, zero-order structures can generate saturated colors even in diffuse illumination. Such structures occur on the setae (hairs) of the first antenna of some isopod crustaceans, such as the giant species of *Bathynomus*. Here, there are diffracted orders and the spectral content of the light within the grating is controlled by the groove profile. In an optical system that only accepts the zero-order, what is seen is white light minus that diffracted into the ± 1 orders.

A zero-order grating can cause total transmission (i.e., there is no reflection). Such antireflective structures are found on the corneal surfaces of each ommatidium (visual unit) in the eye of *Zalea minor* (Diptera) (Figure 14). The periodicity of the corneal gratings of this fly is 242 nm. Another form of antireflection grating is formed on the transparent wings of the hawkmoth *Cephonodes hylas* on the corneal surface of each visual unit (ommatidium) of the eyes of moths and butterflies (Figure 15). Here, optical-impedance matching is achieved by means of a hexagonal array of tapered cylindrical protuberances, each of about 250 nm diameter, thus forming a 'tri-grating' with grooves transecting at 120°.

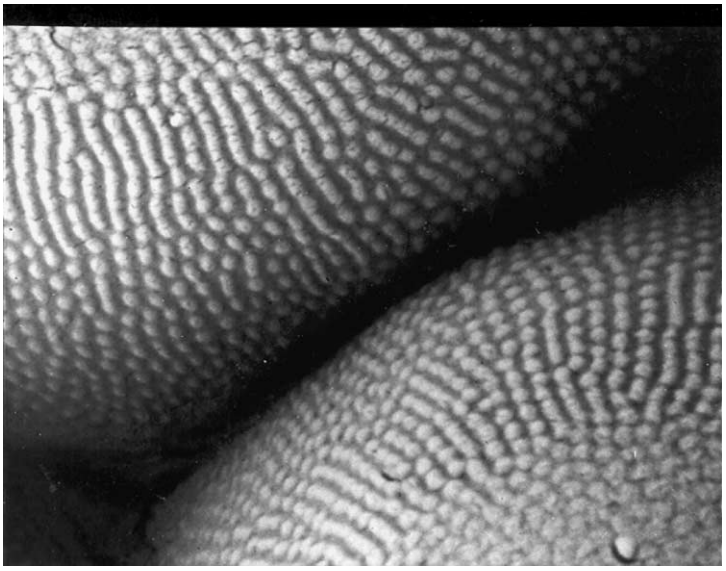


Figure 14 Scanning electron micrograph of the compound eye of *Zalea minor* (Diptera). The corneal surface at the junction of two ommatidia is shown with antireflection gratings (positioned on sloping regions, i.e., at high angles to the normal of the eye) phasing into moth-eye-type protuberances. Scale bar represents 2 μm.

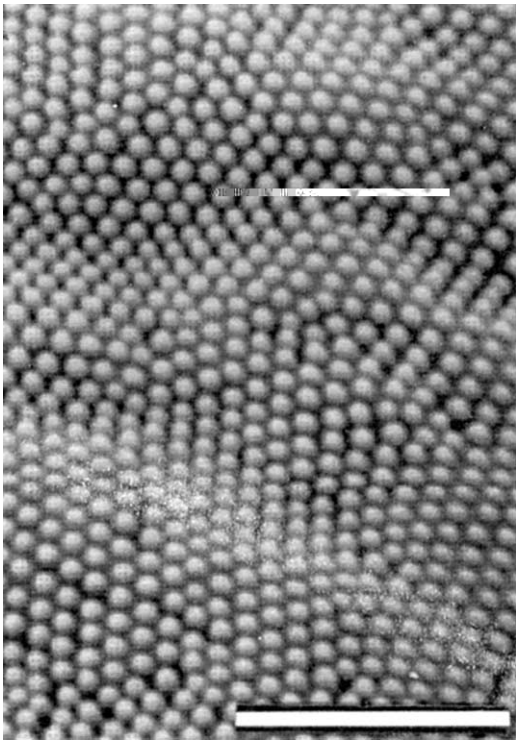


Figure 15 Tri-grating (antireflector) on the corneal surface of a butterfly eye. Scale bar represents 2 μm.

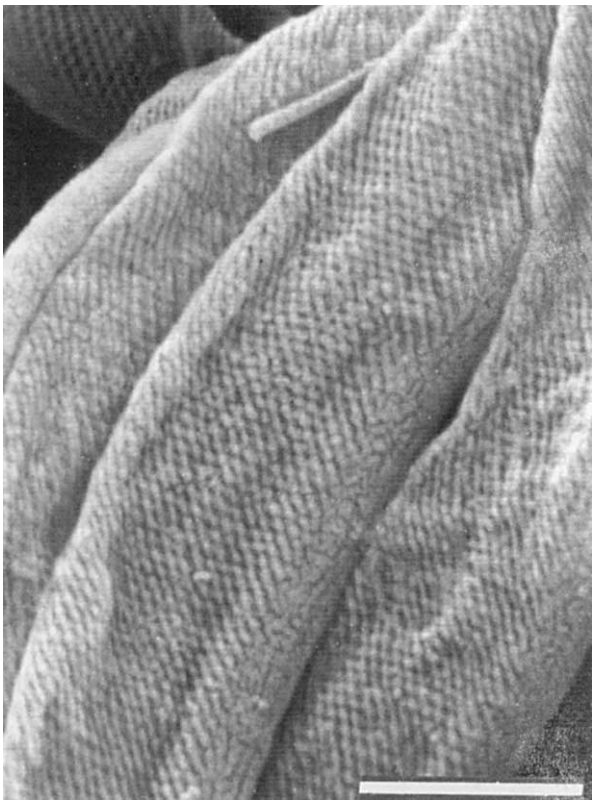


Figure 16 Bi-grating on the callynophores (hairs) of the amphipod crustacean *Waldeckia australiensis*. Scale bar represents 2 μm.

The protuberances provide a graded transition of refractive index between the air and the cornea/wing. Hence the refractive index at any depth is the average of that of air and the corneal/wing material.

The grooves of a grating may also create parallel rows in two directions, forming a bi-grating. Bi-gratings can be found in some crustaceans and flies. In the amphipod crustacean *Waldeckia australiensis*, two effectively superimposed gratings subtend angles of about 60/120° (Figure 16).

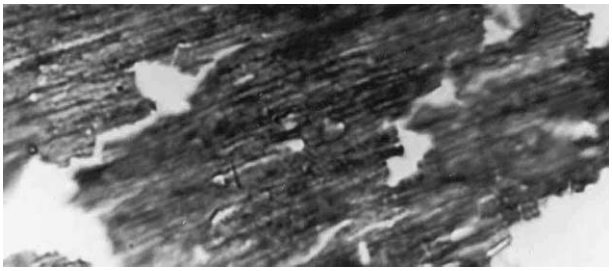


Figure 17 Light micrograph of the surface of a hair of *Canadia spinosa* (Middle Cambrian, 515 million years old, British Columbia), isolated by acid maceration of the rock matrix, showing gratings running longitudinally. Periodicity 0.9 μm.

In terms of fossil colors, diffraction gratings are responsible for the nacreous luster of brachiopod shells, such as those from the Devonian, around 360 million years old. Here, tabular aragonite platelets averaging 600 nm in thickness, each comprising a linear diffraction grating, form layers and consequently a three-dimensional diffraction grating. Antireflective, zero-order gratings have been identified on the eye of an Eocene fly, 45 million years old, preserved in Baltic amber. Linear diffraction gratings causing color have been discovered on the sclerites of *Wiwaxia corrugata* from the Burgess Shale (Middle Cambrian, 515 million years old) (Figure 17). This finding has led to the 'light switch theory', in that the evolution of the first eye caused the big bang in animal evolution.

Scattering (Incoherent Reflectors)

Simple, equal scattering of all spectral wavelengths results in the observation of a diffuse white effect. This commonly arises from the effects of a nonperiodic arrangement of colloidally dispersed matter where the different materials involved have different refractive indices (Figure 18), or from solid colorless materials in relatively concentrated, thick layers. In the colloidal system, the particles are larger than the wavelength of light and can be thought of as mirrors oriented in all directions. The reflection is polarized unless the incident light is at normal incidence on the system and, in the colloidal system, spherical or randomly arranged particles are involved.

The colloidal system involves either a gas-in-solid, gas-in-liquid, liquid-in-liquid (emulsions), or solid-in-liquid. For example, the gas-in-liquid system is partly responsible for the white body and/or tentacles of certain anemones. Light is reflected and refracted at the surfaces of the particles of matter or spaces (with dimensions >1 μm), regardless of the color of the materials involved (except opaque brown and black compounds, such as melanin). In insects, the

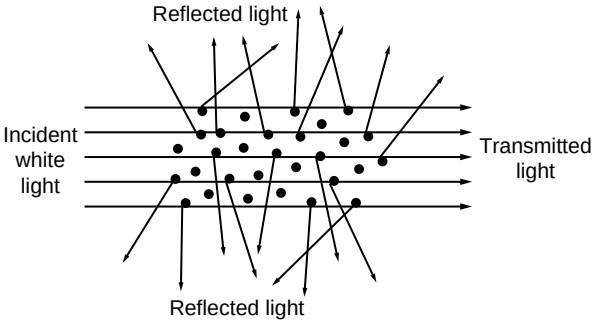


Figure 18 Scattering of white light by small particles (represented by black circles). The reflected light is randomly directed at different angles.

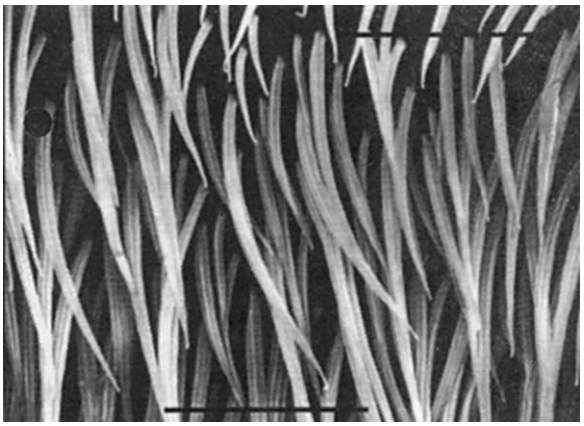


Figure 19 Scanning electron micrograph of a section of a white reflecting patch on the abdomen of the fly *Amenia* sp., showing closely packed, randomly arranged hairs that scatter incident light in all directions (Figure 11). Scale bar represents 20 μm.

materials involved typically have very low transparencies.

An unordered (as opposed to periodic) group of closely spaced setae (hairs), such as those in patches on the fly *Amenia* sp., may form a white reflection via random scattering or reflection (Figure 19). However, if the arrangement becomes periodic to some degree, a diffraction grating may be formed, such as the grating of *Tetraleberis brevis* (Ostracoda) (Figure 13).

From some scales of butterfly wings, light is scattered uniformly and completely in all directions, due to the chaotic disposition of the surfaces. Matt or pearly whites may be observed depending on the complexity of the arrangement of the structures, which affects the relative degree of scattering. The structures may be so small that the molecular topography of the surface has an effect. The chromatic effects of the butterfly scales are greatly intensified if a dark, absorbing pigment screen lies beneath. This screen prevents reflection of white or

colored light from the background that would dilute or alter the color. Additionally, if a dark pigment such as melanin is interspersed with the scattering elements, the reflection will appear a shade of gray or brown. The cells of the reflector in the photophore of a beetle ('fire-fly') are packed with sphaerocrystals of urate that cause a diffuse reflection (Figure 20).

Reflection and refraction that occur at the interfaces of strata with different refractive indices may result in the display of white light. The degree of whiteness depends upon the difference in refractive indices. This mechanism is evident in the shells of many lamellibranch molluscs. Between the outer, often pigmented layer and the mantle, is a thick middle layer of crystalline calcium carbonate. The inner surface of this (nacreous) layer is lined with multiple laminations of the same salt. In most species these laminations are sufficiently thick (>1 μm) to render the inner lining white, although in some species they become thin so as to form a multilayer reflector. Calcium carbonate similarly produces whiteness in Foraminifera and in calcareous sponges, corals, echinoderms, and crustacean cuticles. Also, in the class of white solids is silica in diatom tests and skeletons of hexactinellid sponges.

Other forms of scattering also exist and result in a blue colored effect (red when the system is viewed in transmission). Tyndall or Mie scattering occurs in a colloidal system where the particle size approximates the wavelength of light. Here, diffraction is important. Rayleigh scattering occurs in molecules in a two-photon process by which a photon is absorbed and raises the molecule to an excited electronic state from which it re-radiates a photon when it returns to the ground state. Diffraction is not involved here.

Tyndall scattered light is polarized under obliquely incident light. The intensity of the resultant blue is increased when it is viewed against a dark background, such as melanin. The relative sizes of particles determine the shade of blue. If the particles responsible for the scattering coalesce to form particles with a diameter greater than about 1 μm, then white light is observed (see above). A gradation from blue to white scattering ('small' to 'large' particles) occurs on the wings of the dragonfly *Libellula pulchella*.

Scattered blues can also be found in other dragonflies. In the aeschnids and agrionids, the epidermal cells contain minute colorless granules and a dark base. The males of libellulids and agrionids produce a waxy secretion that scatters light similarly over their dark cuticle. The green of the female *Aeschna cyanea* is the combined result of Tyndall scattering and a yellow pigment, both within the epidermal cells (degradation of the yellow pigment turns the dead dragonfly blue).

Scattered blues are also observed from the skin of the cephalopod *Octopus bimaculatus*, where a pair of ocelli are surrounded by a broad blue ring. Blue light is scattered from this region as a result of fine granules of purine material within cells positioned above melanophore cells. The color and conspicuousness of the ring are controlled by the regulation of the melanophores, by varying the distribution of melanin

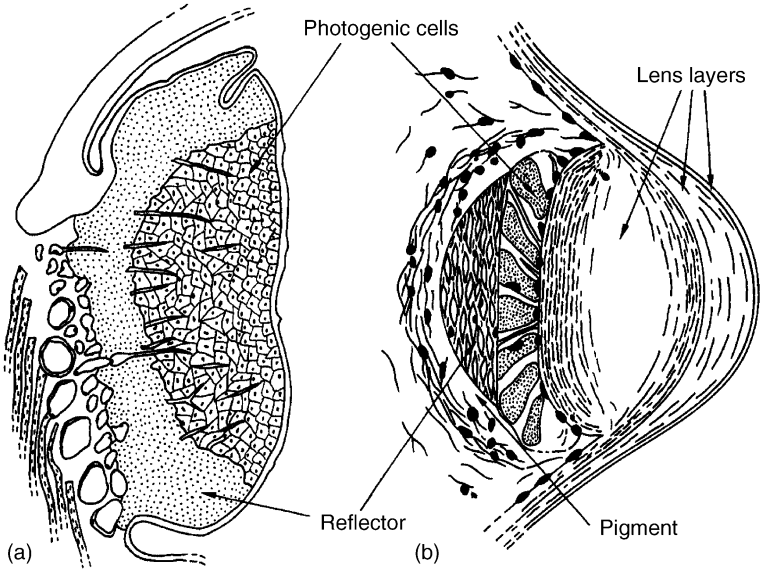


Figure 20 Sections of photophores (bioluminescent organs), external surface of animals on the right. (a) The 'fire-fly' beetle *Pyrophorus* sp., abdominal photophore of male – reflector is based on a scattering system. (b) The shrimp *Sergestes prehensilis*, showing lens layers and absorptive pigment – reflector is based on a multilayer type.

and consequently the density of the absorbing screen. The squid *Onychia caribbaea* can produce rapidly changing blue colors similarly. The bright blue patterns produced by some nudibranch molluscs result from reflecting cells containing small vesicular bodies, each composed of particles about 10 nm in diameter and therefore appropriate for Rayleigh scattering.

See also

Diffraction: Diffraction Gratings. **Scattering:** Scattering Theory.

Further Reading

Fox DL (1976) *Animal Biochromes and Structural Colours*. Berkeley, CA: University of California Press.

Fox HM and Vevers G (1960) *The Nature of Animal Colours*. London: Sidgwick and Jackson Ltd.

Gale M (1989) Diffraction, beauty and commerce. *Physics World* 2: 24–28.

Harvey EN (1952) *Bioluminescence*. New York: Academic Press Inc.

Herring PJ, Cambell AK, Whitfield M and Maddock L (1990) *Light and Life in the Sea*. Cambridge, UK: Cambridge University Press.

Hutley MC (1982) *Diffraction Gratings*. London: Academic Press.

Land MF and Nilsson D-E (2002) *Animal Eyes*. Oxford, UK: Oxford University Press.

Pain S (1998) Seeing the light. *New Scientist* 2161: 42–46.

Parker AR (1999) Light-reflection strategies. *American Scientist* 87: 248–255.

Parker AR (2003) *In the Blink of an Eye*. London: Simon and Schuster/Cambridge (USA): Perseus Publishing.

Parker AR, Welch VL, Driver D and Martini N (2003) An opal analogue discovered in a weevil. *Nature* 426: 786–787.

Parker AR, McPhedran RC, McKenzie DR, Botten LC and Nicorovici N-AC (2001) Aphrodite's iridescence. *Nature* 409: 36–37.

Microstructure Fibers

R S Windeler, OFS Laboratories, Murray Hill, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Microstructure fibers have unique properties that cannot be obtained from traditional fibers (i.e. all glass, doped silica fibers) and can deliver functionalities superior to many of today's best transmission and specialty fibers. Their unique properties are obtained from an intricate cross-section of high and low index regions that traverse the length of the fiber. The vast majority of fibers consist of silica for the high-index region and air for the low-index region. These fibers are known by several different names including microstructure fiber, holey fiber, and photonic crystal fiber. The term microstructure fiber is used in this chapter because it is the most generic, encompassing all the fiber types.

Microstructure fiber properties vary greatly and are determined by the size and arrangement of air holes in the fiber. For example, fibers have been fabricated such that the numerical aperture of the core or cladding approaches one. They have been fabricated with unique dispersion profiles, such as a zero-group velocity dispersion in the near visible regime or an ultraflat dispersion over hundreds of nanometers wide. They have been fabricated to generate a super continuum over two octaves wide. Some guide light with an air core via bandgap guidance. In addition, the air holes in these microstructure fibers can be filled with highly tunable materials, giving one the capability of controlling the fiber properties for use in energy efficient devices.

Microstructure fibers can have doped regions like traditional fibers. These hybrid fibers combine the benefits of traditional and microstructure fibers. The doped core typically guides most of the light, but its guidance properties can be strongly influenced by the air regions. Applications for hybrid fibers include dispersion compensation, polarization maintaining, and bend insensitive fibers.

Loss is always an important factor when determining whether a microstructure fiber will compete with or replace a traditional optical fiber. The index profiles that make these fibers so special can also lead to fibers that have an inherently high loss at connections or along the length of the fiber. Loss at connections typically occurs due to mode mismatch between the two fibers, undesired hole collapse, or poor alignment. Loss along the fiber length can occur due to impurities, hole surface roughness, or poor confinement.

The dispersion profile (zero dispersion wavelength, dispersion slope, normal or anomalous) of microstructure fibers can be optimized more than a traditional fiber, because the hole size and placement

can be arranged to make the cladding index strongly wavelength dependent. Dispersion causes a pulse to spread because the phase velocity is wavelength dependent. Normal dispersion is when longer wavelengths travel faster than shorter wavelengths. Anomalous dispersion is just the opposite – shorter wavelengths travel faster than longer wavelengths. Dispersion determines the amount of interaction between different wavelengths. For applications such as transmission fiber, one wants little interaction and for high nonlinear applications, one wants significant interaction.

This chapter examines the methods for fabricating microstructure fibers and reviews the different types of microstructure fibers and their properties. Fibers that guide by total internal reflection are reviewed separately from bandgap guided fibers because their properties are significantly different. Lastly, a description and example of an active device is presented in which the air regions of a microstructure fiber are filled with controllable material.

Microstructure Fiber Fabrication

At first glance, fabrication of microstructure fibers looks similar to traditional fibers in that the fibers are fabricated in a two-step process. First a preform is fabricated and then it is drawn (stretched) into a fiber. However, when the process is examined in more detail, both the preform fabrication and draw depart significantly from traditional methods. First, a novel preform fabrication method is used to incorporate air holes that run the length of the preform. Second, novel drawing procedures are used to keep the air holes open as the preform diameter is reduced by several orders of magnitude during draw.

Preform Fabrication

Microstructure preforms are cylinders of amorphous material usually less than a few centimeters in diameter that have an index profile running the length of the preform similar to the desired fiber. The most common preform consists of air and a single material such as silica, polymer, or high nonlinear glass. These air-containing preforms are relatively easy to fabricate but are difficult to draw because the structures can deform.

The basic parameters of the fiber are usually determined by the geometry of the holes in the preform. These include hole diameter (d), hole position, pitch (center-to center hole spacing, Λ), core diameter (a), and number of layers. However, changes in the size and shape of the holes in the fiber can be made purposely or accidentally during draw, which will cause deviations between the fiber and preform profiles.

Several methods are used to fabricate microstructure preforms including stacking, casting, extrusion, and drilling. Each process has its advantages and disadvantages. By far, the most common method is stacking in which tubes, rods, and core rods (rods containing doped cores fabricated by traditional optical fiber techniques) are stacked in a close packed pattern such as a triangular or hexagonal lattice (Figure 1). The bundle of tubes and rods is then bound together with a large tube, called an overclad tube. Fibers can be made with complicated or asymmetric index profiles by strategically placing tubes with the same outer diameter (for good stacking) but a different inner diameter to change the index in that location. For example, a fiber with smaller holes on opposite sides of the core produces high birefringence.

The main advantages of stacking are that no special equipment is needed for fabricating microstructure preforms and that doped cores are easily added. However, there are several disadvantages. First, the stacking method is limited to simple geometries of air holes because the tubes are stacked in a close pack arrangement. Second, unless hexagonal tubes are used, interstitial areas are created between the tubes that may not be desired in the final fiber. Third, the method is labor intensive and requires significant glass handling.

A variation of the stacking method is used for fabricating air-clad fibers. Air-clad preforms are created by placing a layer of small tubes around the perimeter of a large core rod. The assembly is then overcladed for strength (Figure 2).

Extrusion and casting of glass powder, polymer, or sol–gel slurry are also used for making single material

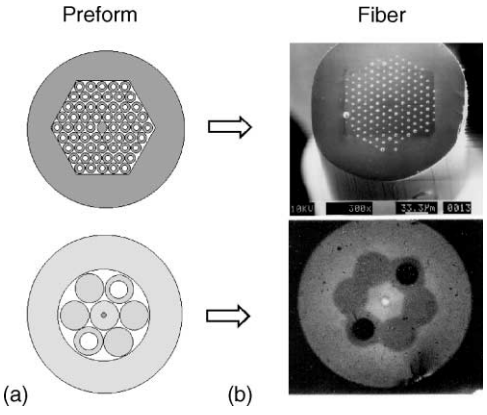


Figure 1 (a) Tubes, rods, and core rods are stacked together in a close packed arrangement and held together with an overclad tube. (b) During draw the desired air region stay open forming a microstructure fiber.

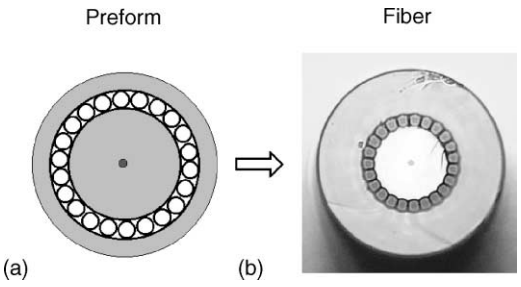


Figure 2 (a) An air-clad preform is fabricated by stacking a ring of small, thin walled tubes around a large core rod. (b) Photograph of the resulting fiber.

microstructure preforms. The main advantage of extrusion and casting is that complicated structures can be fabricated in which the position, size, and shape of the air regions are independent of one another. The disadvantage of this method is that preform fabrication is more difficult compared to the other fabricate methods. This method may become more common as complex air structures are needed to make advanced microstructure fibers.

The last method consists of drilling holes in a preform or rod. Drilling is well understood and used for other specialty fibers. The advantage of this method is that it is easy to drill holes of various sizes in any position in a preform, including doped regions. The disadvantages of drilling are that the holes cannot be drilled very deep compared to a traditional preform length; the distance between holes may be limited due to cracking; and the fiber may experience high loss due to surface roughness of the holes and impurities incorporated during drilling.

Fiber Draw

It is significantly more difficult to draw a microstructure preform with air holes than a traditional preform. This is because the air holes tend to close due to surface tension. This force can be overcome or minimized using two different techniques. The first is to draw the preform under very high tension (low temperature). Minimal hole collapse occurs when the draw tension is significantly larger than the surface tension. The problem with this method is that the break rate increases substantially at higher tension.

The second method is to apply an external pressure to the holes to counteract the surface tension. If a single pressure source is used, the holes can be made larger or smaller during draw by changing the pressure as needed. However, the process can become unstable because a large hole needs a lower pressure to maintain its size than a small hole. If holes of different sizes exist, a larger hole will grow at the expense of a smaller one regardless of the pressure used. To minimize the instability, this method is typically performed in conjunction with high-tension drawing.

To alleviate the draw problem, a second solid material can be used in the preform to obtain the desired index profile. With the air regions removed from the preform, the drawing process becomes significantly easier. In addition, when a preform consists of two solid materials, one of the materials does not have to be continuous, as in single material preforms. This allows for simpler designs such as concentric rings, which may be easier to fabricate and model (Figure 13c). The disadvantage of this approach is finding two materials that have compatible physical and chemical properties, have low loss, and have a large index of refraction difference.

Microstructure Fiber Types

Microstructure fibers are categorized by their method of guidance, properties, and function. Index guided microstructure fibers most closely resemble traditional fibers and are described first. These fibers guide light by total internal reflection and have a core index of refraction greater than the cladding. Bandgap fibers are examined next. The unique guidance of bandgap fibers allows the core index to be lower than the cladding index. Tunable microstructure fibers are described last. These microstructure fibers are filled with tunable materials so that the fiber's properties can be actively manipulated.

Index Guided Fibers

Most microstructure fibers guide light by total internal reflection. The size and spacing of air holes determine the guiding properties. For example, a solid silica core surrounded by a cladding consisting of small air holes that are spread apart creates a small core-to-cladding index difference similar to what can be achieved in traditional fibers. At the other extreme, a solid silica core surrounded by a cladding of large holes, which are closely spaced together, forms a large core-to-cladding index difference. This structure creates a large numerical aperture in the core and optically isolated regions within the fiber. The ability to create regions of significantly different indices in various parts of the fiber distinguishes the index guided microstructure fibers from traditional fiber. Different types of microstructure fiber are formed by varying the size, spacing and pattern of the air holes.

Different techniques are used to model these fibers, depending on spacing and size of the air holes relative to the wavelength of the guided light. If the hole diameter and spacing is much smaller than the

wavelength of the guided light, the cladding index can be modeled as an average index weighted by the volume fraction of the two materials (typically air and silica). As the hole diameter or spacing approaches the wavelength of the guided light, complicated models are used such as finite difference method, finite element method, beam propagation method, or multipole method.

Endlessly single-mode fiber

The endlessly single-mode fiber consists of a solid core surrounded by a two dimensional array of small air holes in the cladding (Figure 3). The size and position of the air holes are arranged such that only the fundamental mode exists in the core regardless of the wavelength of the guided light.

The prevention of higher order modes can be understood by examining the V number. The V number is a dimensionless parameter that describes the guiding nature of the waveguide. In traditional fiber, when the V number is less than 2.405, the fiber will guide only the fundamental mode. The V number is given by

$$V = \frac{2\pi a}{\lambda}(n_{\text{core}}^2 - n_{\text{clad}}^2)^{1/2} \qquad [1]$$

where a is the core diameter, λ is the wavelength of the guided light, and n_{core} and n_{clad} are the index of refractions of the core and cladding, respectively. In traditional fiber, the core and cladding index (n_{core} and n_{clad}) are for the most part constant, so the V number increases as the wavelength decreases.

However, for the endlessly single mode microstructure fiber, the V number does not increase indefinitely with decreasing wavelength, but approaches a constant value. Obtaining a V number independent of wavelength is understood by examining the distance the light propagates into the cladding as a function of wavelength. As the wavelength becomes shorter, light penetrates a shorter distance into the cladding, interacting with less of the air regions. This causes the effective cladding index (n_{clad}) to increase with decreasing wavelength, such that V approaches a constant value.

Microstructure fibers are made endlessly single-moded over all wavelengths by properly designing the hole diameter-to-pitch ratio (d/Λ). When the d/Λ is low enough, the microstructure fiber guides light only in the single mode regardless of wavelength and core size (Figure 4). However, these large core diameter fibers are limited by long wavelength bend loss in the same way as large core single mode traditional fibers.

High delta core fiber

High delta core microstructure fibers (HDCMF) consist of a small (typically less than 2 microns) solid silica core surrounded by one or more layers of large air holes closely spaced together (Figure 5). This creates a very large index difference between the core and cladding. A core-to-cladding index difference of 0.4 is easily achieved in a HDCMF, which is an order of magnitude larger than can be achieved with traditional fiber.

Even though the core is very small, these fibers are almost always multimoded due to the large index difference between the core and cladding. However, because of its small core, it is difficult to launch anything other than the fundamental mode into the fiber. Once a mode is guided in the fiber it remains in that mode due to the large difference in effective indices between modes (Figure 6). When higher-order modes are purposely generated in the fiber, they also

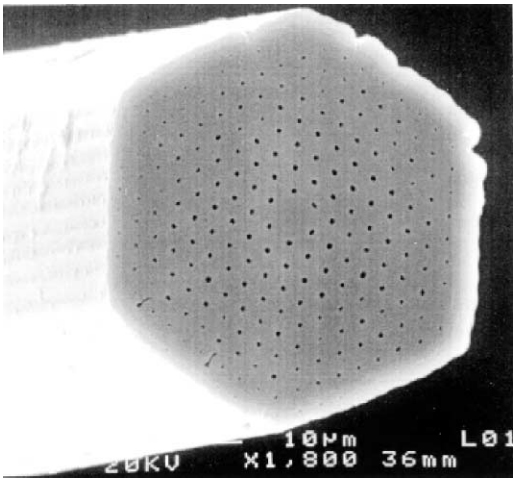


Figure 3 Photograph of an endlessly single mode fiber (Reprinted with permission from Birks TA, Knight JC and Russell P (1997) Endlessly single mode photonic crystal fiber. *Optics Letters* 22: 961–963).

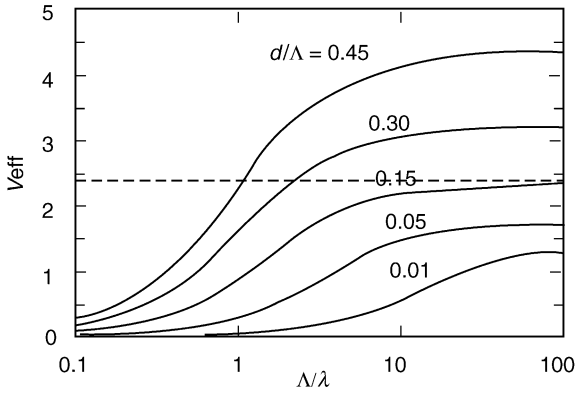


Figure 4 The effective V number as a function of Λ/λ for various d/Λ (Reprinted with permission from Birks TA, Knight JC and Russell P (1997) Endlessly single mode photonic crystal fiber. *Optics Letters* 22: 961–963).

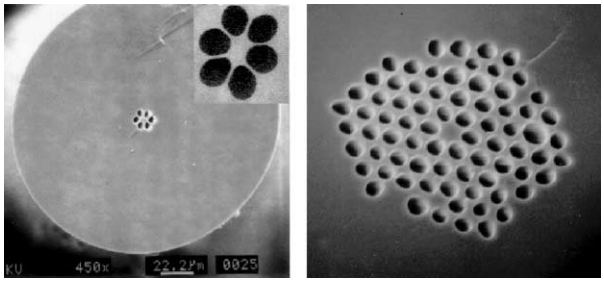


Figure 5 High delta core fibers consist of a small solid core surrounded by at least one layer of large air holes closely spaced together. Right-hand panel reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798.

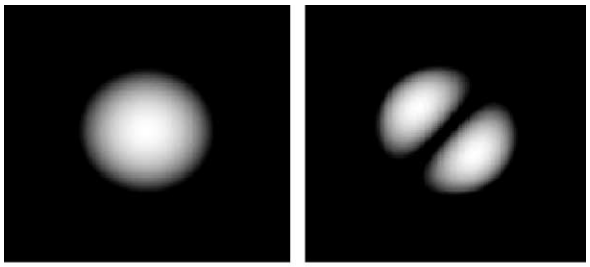


Figure 6 Individual modes do not couple well due to large differences in effective indices between modes. The figure shows the mode profile and effective index of the fundamental and the next higher mode for a fiber with a two-micron diameter core with a core-to-cladding index difference of 0.45 (Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798).

guide over long lengths without coupling to other modes (Figure 7). This strong guidance makes the fiber extremely bend insensitive.

HDCMF can have a unique dispersion profile because it can guide the fundamental mode exclusively in the multimode regime. This allows greater flexibility in controlling the shape of the dispersion profile, especially the ability to move the zero dispersion to shorter wavelengths. The HDCMF is the first silica fiber with anomalous dispersion of single-mode light below 1270 nm (the wavelength at which the material dispersion of silica is zero). These fibers are often made to have a zero-group velocity dispersions around 770 nm for use with Ti-sapphire lasers.

HDCMF is ideal for performing nonlinear experiments in the near-visible regime because the fiber's small, high-index core creates high intensity light, and the fiber's low dispersion creates long interaction times. For example, a broadband continuum can be generated over two octaves using less than a meter of HDCMF fiber (Figures 8 and 9).

Tapered microstructure fiber (TMF)

Coupling light in and out of HDCMF is difficult and results in a large loss due to its small core. When free space optics are used to couple the light, frequent realignment of the beam is required to minimize loss due to drifts in the equipment. Splicing HDCMF fiber to traditional fiber also results in high loss due to a large modal mismatch. These coupling problems are solved with tapered microstructure fibers.

Tapered microstructure fibers (TMF) consist of a traditional germanium single-mode core surrounded by a small silica cladding (Figure 10). The inner

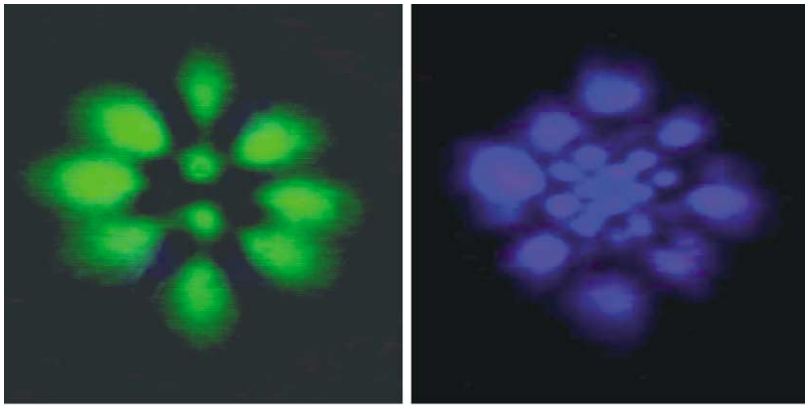


Figure 7 Far-field patterns show that higher order modes propagate long distances in the HDCMF without coupling to other modes (Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798).

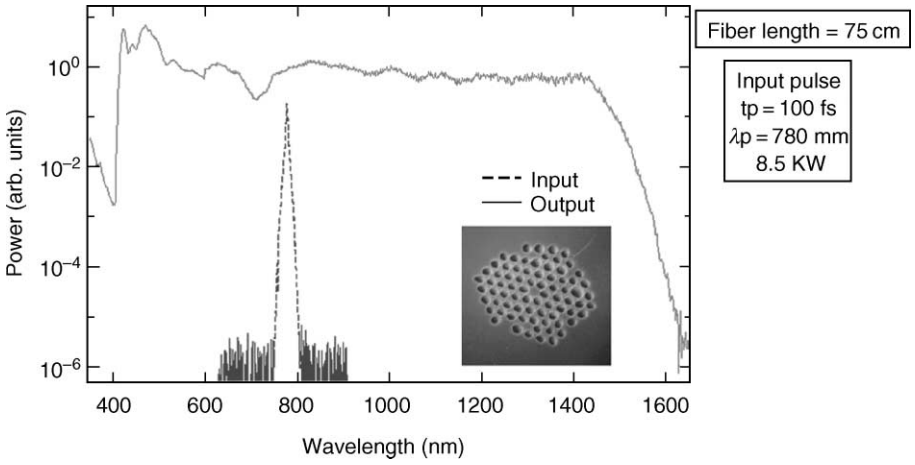


Figure 8 Plot of a supercontinuum spectrum over two octaves wide generated from a 75 cm piece of HDCMF (Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optics Letters* 25: 25–27).

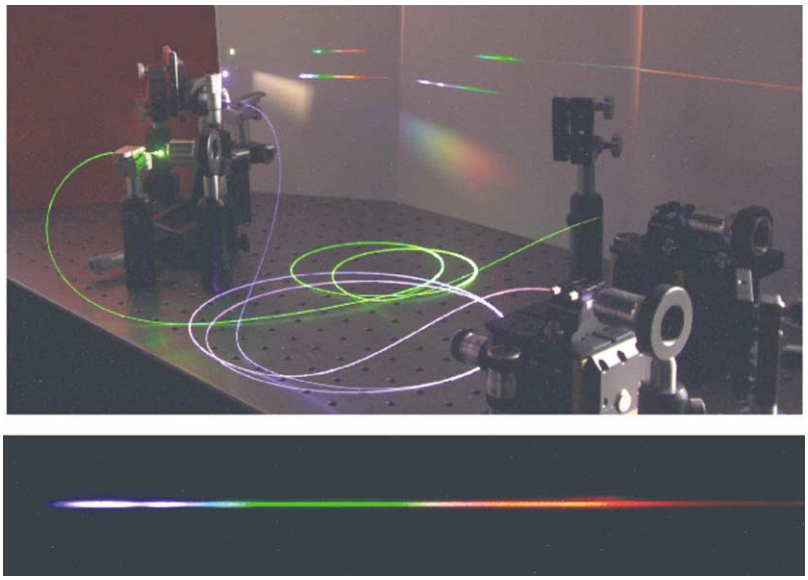


Figure 9 Photograph of fiber generating the supercontinuum. The bottom photo shows the continuum after passing through a prism.

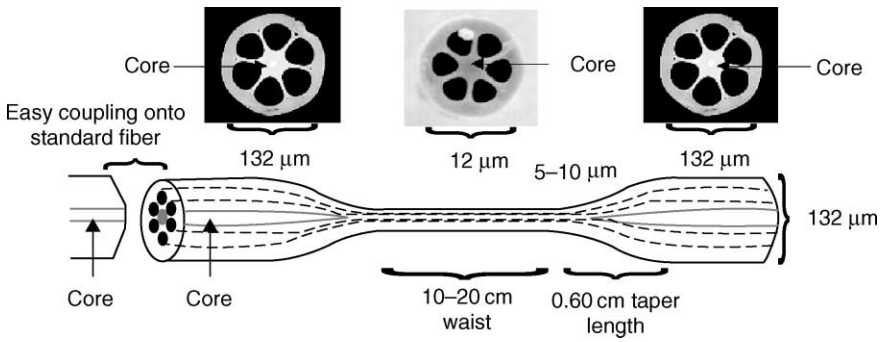


Figure 10 Schematic drawing of the tapered microstructure fiber (Reprinted with permission from Chandalia JK, Eggleton BJ, Windeler RS, Kosinski SG, Liu X and Xu C (2001) Adiabatic coupling in tapered air-silica microstructures optical fiber. *IEEE Photonics Tech. Letters* 13(1) (©2001 IEEE)).

cladding is surrounded by a layer of air holes and then a protective outer silica cladding for strength. Since the core of the TMF is very similar to a traditional single-mode fiber core, the splice loss is low. Splice losses below 0.075 dB are typical.

Properties similar to the HDCMF are obtained by adiabatically tapering the TMF while maintaining the cross-sectional profile of the fiber. In the taper region, the fundamental mode is no longer guided by the germanium core and evolves into the fundamental mode of the silica region, where it is confined by the ring of air holes. The taper waist is typically ten to twenty centimeters long and has properties identical to the HDCMF. When the fiber is adiabatically expanded back to its original size, the light is guided back into the standard diameter germanium core, making it easy to splice to traditional fiber with low loss.

The same effect can be obtained by stretching a traditional fiber from 125 microns down to 2 microns. However, there are several disadvantages compared to the tapered microstructure fiber. The taper region is much longer in the traditional fiber; the optimal diameter of the silica core is harder to maintain; the taper region must remain clean and uncoated; and the taper region is much weaker.

Air-clad fiber

Air-clad fibers consist of a doped core surrounded by a silica inner cladding. The inner cladding is surrounded by a ring of air, which in turn is surrounded by an outer silica cladding for strength (Figure 2). Thin silica webs connect the inner and outer claddings to hold the inner fiber region in place. The region consisting of air and thin silica webs creates an optical barrier, preventing most of the light from escaping the inner cladding. Such fibers are referred to as double clad fiber.

In traditional double-clad fibers, a material of lower index (typically another layer of glass or a layer of polymer) immediately surrounds the inner silica cladding. A key optical parameter describing the light guiding properties of the inner cladding is the numerical aperture (NA). The NA is defined as the sine of the largest (acceptance) angle of light that will be guided in the inner cladding. The upper NA values of traditional double-clad fibers (~0.22 and ~0.45 for glass and polymer outer claddings, respectively) are limited by the refractive indices of the available cladding materials.

For the air-clad fiber, the NA is determined by the ability of light to escape from the inner cladding through the silica webs. The NA of the air-clad fiber can be calculated by modeling the web as an infinitely long slab waveguide. As seen in Figure 11, the NA increases as the webs become thinner or the wavelength of light becomes longer. To obtain NA similar to what is achieved by coating a traditional fiber with a low index polymer (NA ~0.45), the web thickness must be about equal to the wavelength of the light. The NA of the air-clad fiber increases dramatically as the web thickness becomes less than half the wavelength of the guided light. Numerical apertures above 0.90 have been experimentally demonstrated.

The large NA values achievable with air-clad fibers are an important advantage for advanced double-clad

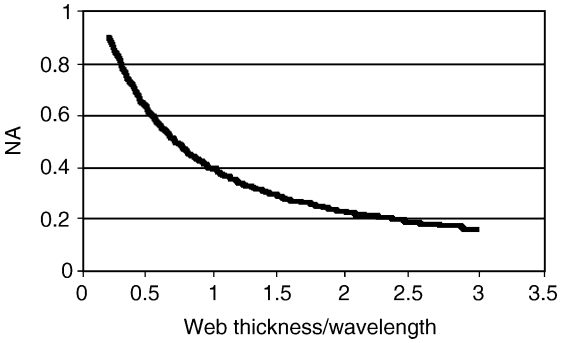


Figure 11 Predicted NA as a function of the web thickness divided by the wavelength using an infinite slab model.

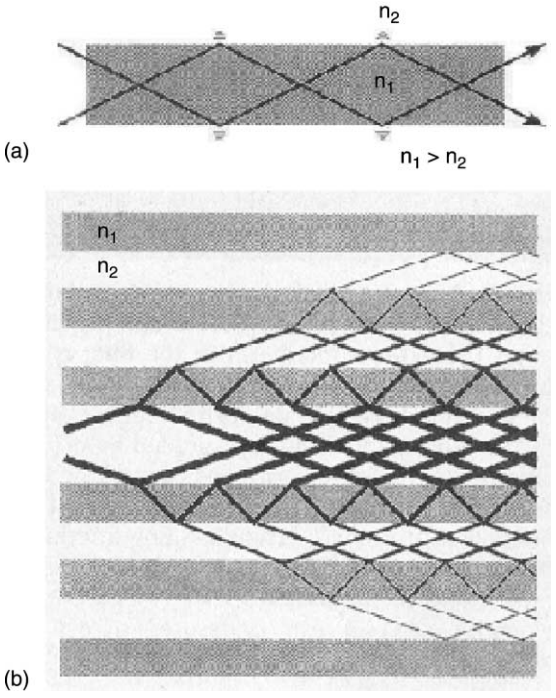


Figure 12 Schematic of the mechanism of (a) index guided light and (b) band gap guided light (Reprinted with permission from Cregan RF, Mangan BJ, Knight JC, Birks TA, Russell PJ, Roberts PJ and Allan DC (1999) Single-mode photonic band gap guidance of light in air. *Science* 285. Copyright 1999 American Association for the Advancement of Science).

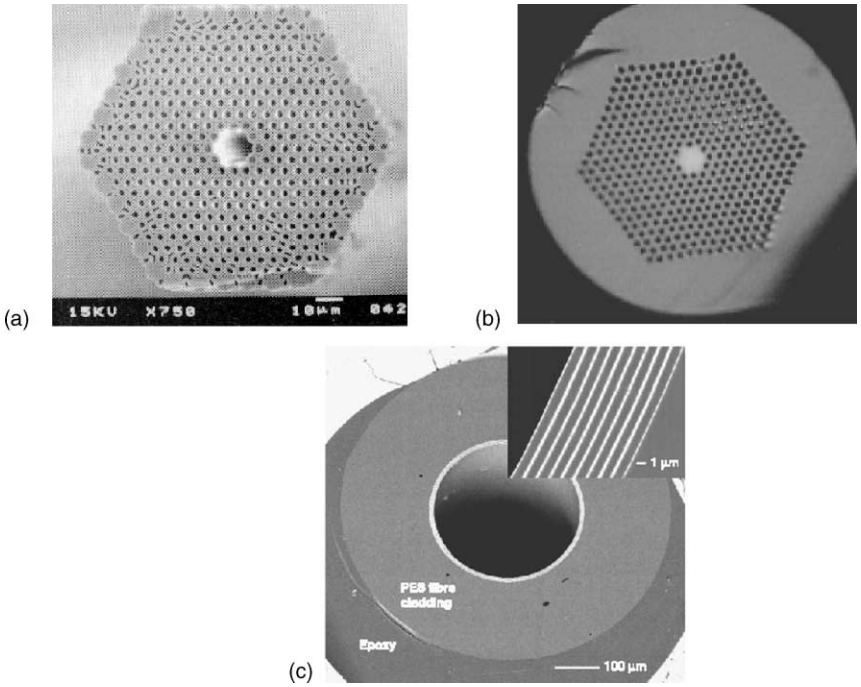


Figure 13 Photographs and drawing of three types of bandgap fibers: (a) triangular array cladding with an air core (Cregan RF, Mangan BJ, Knight JC, Birks TA, Russell PJ, Roberts PJ and Allan DC (1999) Single-mode photonic band gap guidance of light in air. *Science* vol. 285), (b) tunable with a silica core, (c) concentric rings with an air core (Reprinted with permission from Temelkuran *et al*. *Nature* 420. Copyright 1999 American Association for the Advancement of Science).

amplifiers and lasers. A double-clad amplifier or laser (also known as cladding pumped amplifier or laser) contains a double-clad fiber with a rare earth doped core. High-power pump light, launched into and guided within the inner cladding, excites the rare earth ions as the pump light crosses through the core. The excited ions, which emit light with the same wavelength and phase as the signal, amplify the signal. The fiber's efficiency and the rare earth inversion level increase with the amount of pump light absorbed by the core. The light absorbed by the core, in turn, increases with the core-to-inner cladding area ratio and the pump light intensity in the inner cladding.

Traditional fibers have a practical limit to their core-to-cladding ratio. The maximum core diameter is limited by bend loss. And, while ultimately the minimum fiber diameter is limited by draw capabilities, in practice it is limited by the ability to couple pump light into a fiber having a limited inner-cladding NA. For a pump source of given brightness, the product of beam diameter and NA cannot be increased. Such light can be successfully coupled into a larger inner cladding with smaller NA or a smaller inner cladding with larger NA.

Air-clad fibers have other advantages over traditional double-clad fibers. The outer glass cladding in air-clad fibers is optically inactive. This allows the inner cladding diameter to be decoupled from the physical fiber diameter, eliminating draw-related constraints on minimum inner cladding size. Since only the inner cladding is optically active, its size can be optimized (made smaller) while keeping the total fiber diameter at the standard 125 microns. In addition, the pump light does not interact with the polymer coating, which can be important if the polymer properties are affected by the surroundings.

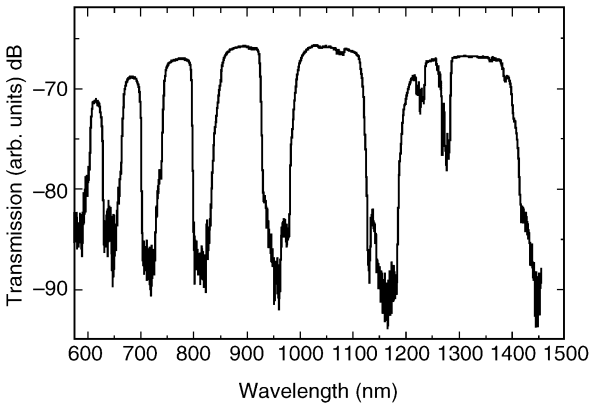


Figure 14 Spectrum of a typical bandgap fiber.

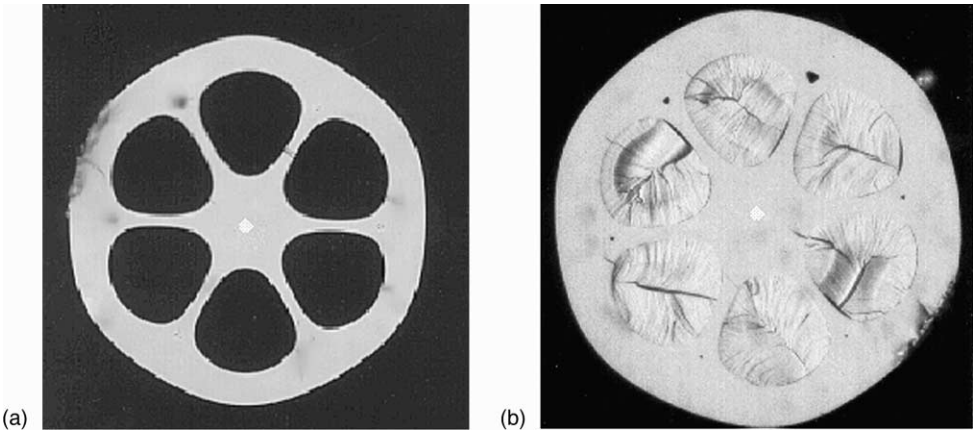


Figure 15 Holes in the microstructure fibers are filled with controllable materials to affect the fiber properties actively. Photos show (a) original fiber and (b) fiber after polymer is inserted and cured in the microstructure fiber.

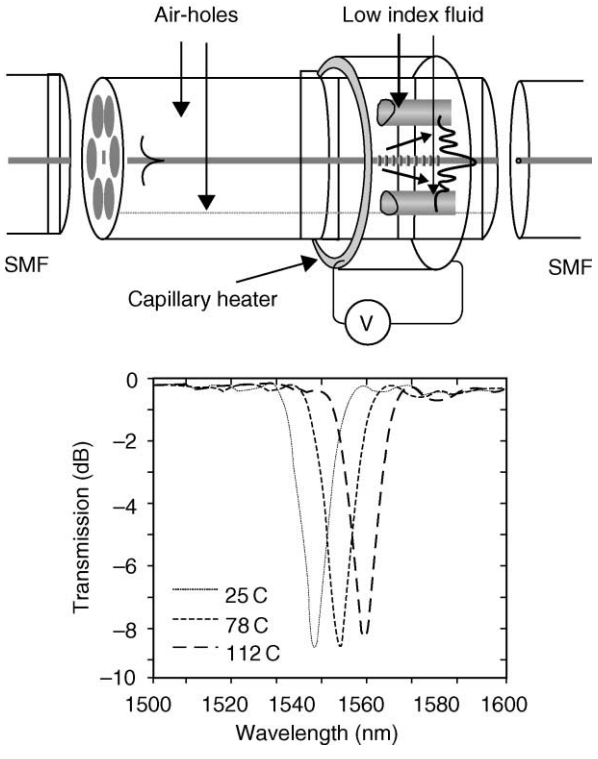


Figure 16 Schematic of a tunable device in which low index liquid is positioned over the grating. The plot shows the shift in the spectrum as a function of temperature (Reproduced from Kerbage C, Windeler RS, Eggleton BJ, Dolinskoi M, Mach P and Rogers JA (2002) Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber. *Optics Communications* 204: 179–184, with permission from Elsevier).

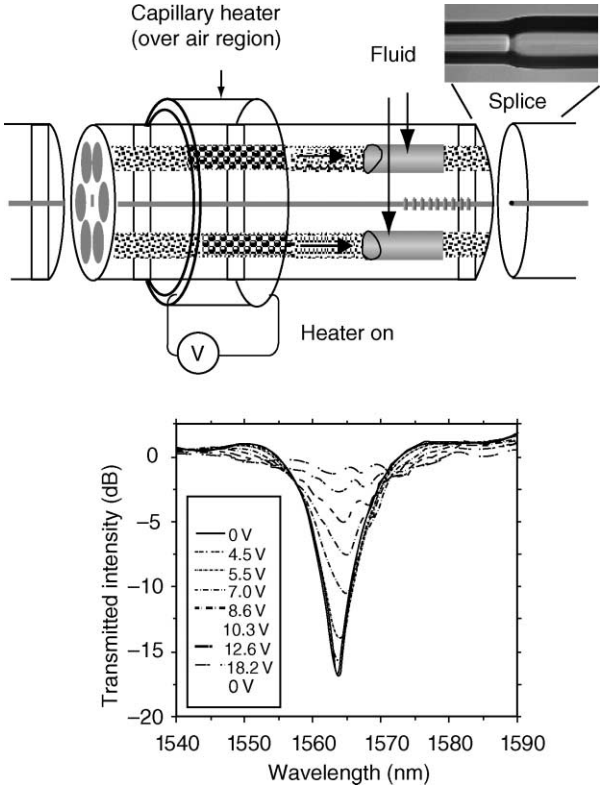


Figure 17 Schematic of a tunable device in which a high index liquid plug is moved over the grating. The plot shows the intensity of the spectrum as a function of temperature (Reproduced from Kerbage C, Windeler RS, Eggleton BJ, Dolinskoi M, Mach P and Rogers JA (2002) Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber *Optics Communications* 204: 179–184, with permission from Elsevier).

Bandgap Guided Fibers

Bandgap fibers guide light in the core (also referred to as a defect) by confining the light through constructive interference due to Bragg scattering (Figure 12). Unlike traditional fibers, this mechanism allows light to propagate in a core that has an index lower than the cladding. Bandgap fibers with an air core could theoretically be very low loss, propagate high powers, have a large effective area core, and

exhibit very low nonlinearities. In addition, the fibers can be used in novel devices by filling the core with special gasses or liquids for nonlinear processes.

Bandgap fibers can be generalized into three types (Figure 13). The first type of fiber has an air core surrounded by a cladding that consists of a periodic array of air holes in silica. Fibers of this type have been designed to guide a single mode and hold the greatest promise of guiding light over long distances with very low loss. The second type of bandgap fiber consists of a silica core and a silica cladding with a periodic array of holes which are filled with a high-index, tunable liquid. This design allows the bandgap properties to be tuned and is of interest for use in devices. The third fiber type consists of rings of alternating high- and low-index dielectric material with a center air core. They typically have a large, multimode core which is ideal for sending high powers. In addition, very little light is in the dielectric materials so the fiber can be designed to guide any wavelength with relatively little loss.

The spectrum of a bandgap fiber is quite different than an index guided fiber (Figure 14). Frequencies that lie within the bandgap cannot propagate in the cladding and are confined to the defect in the lattice (the core). Bandgap fibers with ideal geometry are predicted to have losses over an order of magnitude lower than can be obtained in an ideal traditional fiber. Although the loss of bandgap fibers has not been demonstrated to be less than traditional fiber at telecommunication wavelengths, bandgap fibers have been shown to have much lower loss than traditional fibers at other wavelengths.

Tunable Microstructure Fiber Devices

Unique tunable devices are made by filling the air holes of microstructure fibers with materials whose properties can be controlled actively (Figure 15). The materials can be positioned in the core or cladding to affect the corresponding modes. Most commonly, the index of refraction is changed using temperature sensitive liquids or polymers, although

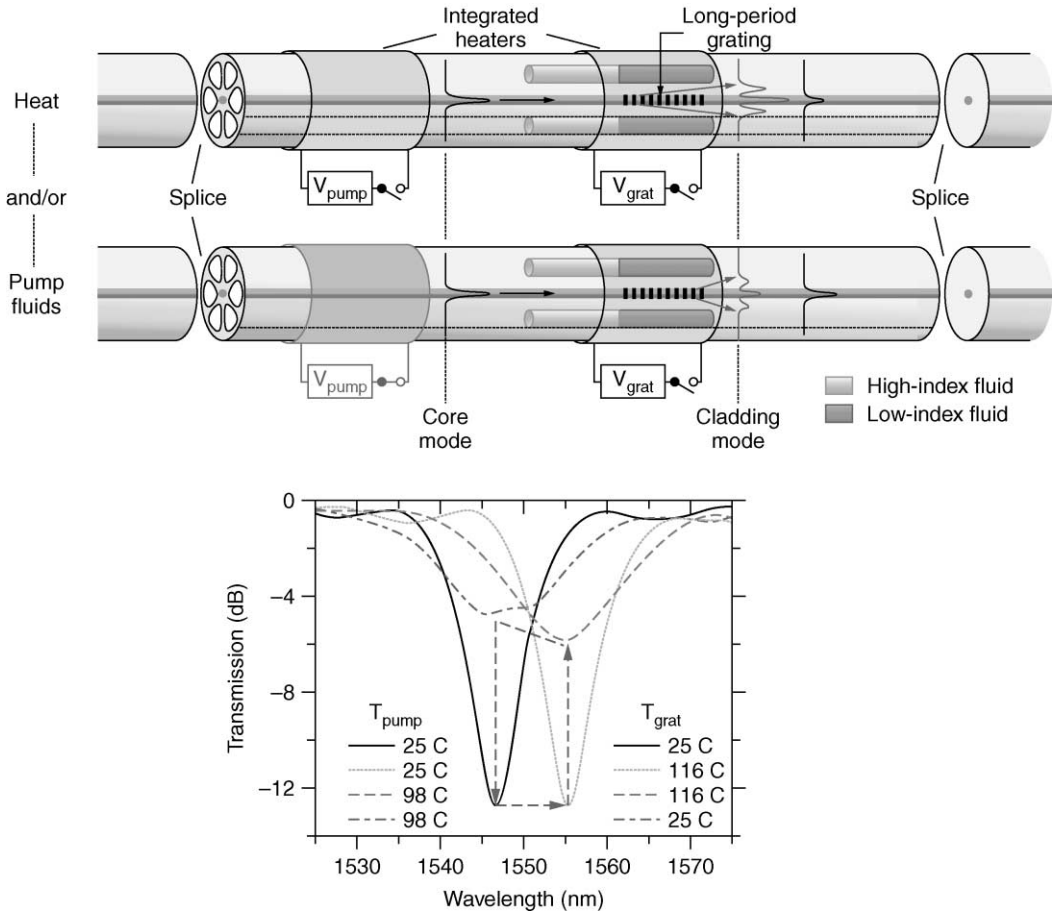


Figure 18 Schematic of a tunable device using both high and low index liquids with two heaters. The plot shows that the position and strength of the spectrum are adjusted independently (Reproduced with permission from Mach P, Dolinski M, Baldwin KW, Rogers JA, Kerbage C, Windeler RS and Eggleton BJ (2002) Tunable microfluidic optical fiber. *Applied Physics Letters* 80(23). Copyright (2002) by the American Physical Society).

materials that respond to electric or magnetic fields enable a quicker response time. These material-filled active fibers can be used to fabricate devices like tunable filters, switches, broadband attenuators, and fibers with tunable birefringence. Below, a few examples of tunable filters are presented in which the device properties depend on the type and position of the liquid inserted in the fiber holes.

A microstructure tunable filter works by controlling the spectral position, shape, or strength of a longitudinal long-period grating written in the fiber. A grating is written in the fiber before filling the holes with tunable liquid. The position of the peak absorption wavelength is controlled by placing a liquid, with an index close to silica and strongly temperature dependent, in air holes in the cladding region over the grating (**Figure 16**). The cladding index is tuned by varying the temperature of the liquid, using a small thin film heater located over the liquid. The position of the peak absorption wavelength moves as the cladding index is changed.

The strength or depth of the filter is controlled by inserting a moveable plug of high-index fluid in a sealed air section of a microstructure fiber (**Figure 17**). The strength of the grating is determined by the fraction of the grating surrounded by high index fluid. The position of the liquid plug relative to the grating can be finely adjusted with a thin film heater located over the air region adjacent to the material plug. As heat is applied to the air, the air expands and pushes the liquid plug over the grating, decreasing the coupling of the fundamental mode to higher order cladding modes. **Figure 18** shows an example in which the position and strength of a filter are varied independently using two heaters and two adjacent materials.

See also

Photonic Crystals: Photonic Crystal Lasers, Cavities and Waveguide.

Further Reading

Birks TA, Knight JC and Russell P (1997) Endlessly single mode photonic crystal fiber. *Optics Letters* 22(13).

Bjarklev A, Broeng J and Bjarklev AS, *Photonic Crystal Fibers*, Boston: Kluwer Academic Publishers.

Broeng J, Mogilevstev D, Bjarklev A and Barkou SE (1999) Photonic crystal fibers: A new class of optical waveguides. *Optical Fiber Technology*, vol. 5.

Chandalia JK, Eggleton BJ, Windeler RS, Kosinski SG, Liu X and Xu C (2001) Adiabatic coupling in tapered air-silica microstructures optical fiber. *IEEE Photonics Tech. Letters* 13(1).

Cregan RF, Mangan BJ, Knight JC, Birks TA, Russell PJ, Roberts PJ and Allan DC (1999) Single-mode photonic band gap guidance of light in air. *Science* 285.

Eggleton BJ, Westbrook PS, White CA, Kerbage C, Windeler RS and Burdge GL (2000) Cladding-mode resonance in air-silica microstructure optical fiber. *Journal Lightwave Technology* 18(8).

Joannopoulos JD, Meade RD and Winn JN (1995) *Photonic Crystals*. New Jersey: Princeton University Press.

Johnson SG, Ibanesc M, Skorobogatiy M, Weisberg M, Engeness O, Soljacic M, Jacobs SA, Joannopoubs JD and Fink Y (2001) Low-loss asymptotically single-mode propagation in large-core OmniGuide fibers. *Optics Express* 9(13).

Kerbage C, Hale A, Yablon A, Windeler RS and Eggleton BJ (2001) Integrated all fiber variable attenuator based on hybrid microstructure fiber. *Applied Physics Letters* 79(19).

Kerbage C, Windeler RS, Eggleton BJ, Dolinskoi M, Mach P and Rogers JA (2002) Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber. *Optics Communications* 204: 179–184.

Knight JC, Birks TA, Russell PJ and Atkin DM (1996) All-silica single-mode optical fiber with photonic crystal cladding. *Optics Letters* 21(19).

Knight JC, Birks TA and Russell PJ (2001) 'Holey' silica fibers. In: Markel VA and George TF (eds). *Optics of nanostructured materials*. New York: John Wiley & Sons.

Mach P, Dolinski M, Baldwin KW, Rogers JA, Kerbage C, Windeler RS and Eggleton BJ (2002) Tunable microfluidic optical fiber. *Applied Physics Letters* vol. 80.

Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optics Letters* 25(1).

Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798.

Omnidirectional Surfaces and Fibers

S Hart, G Benoit and Y Fink, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Omnidirectional Reflecting Mirrors

Mirrors are probably the most prevalent of optical devices. Known to the ancients and used by them as objects of worship and beauty, mirrors are currently employed for imaging, solar energy collection, and also in laser cavities. Their intriguing optical properties have captured the imagination of scientists as well as artists and writers.

One can distinguish between two types of mirrors, the age-old metallic, and the more recent multilayer dielectric. Metallic mirrors reflect light over a broad range of frequencies incident from arbitrary angles (i.e., omnidirectional reflectance). However, at infrared and optical frequencies a few percent of the incident power is typically lost due to absorption.

Multilayer dielectric mirrors are used primarily to reflect a narrow range of frequencies incident from a particular angle or particular angular range. Unlike their metallic counterparts, dielectric reflectors can be extremely low loss. The ability to reflect light of arbitrary angle of incidence for all-dielectric structures has been associated with the existence of a complete photonic bandgap, which can exist only in a system with a dielectric function that is periodic along three orthogonal directions. In fact, a sufficient condition for the achievement of omnidirectional reflection in a periodic system with an interface, is the existence of an overlapping bandgap regime in phase space above the light cone of the ambient media.

Consider a system that is constructed of alternating dielectric layers coupled to a homogeneous medium – characterized by n_0 (such as air with $n_0 = 1$), at the interfaces. Electromagnetic waves are incident upon the multilayer film from the homogeneous medium. While such a system was analyzed extensively in the literature the possibility of omnidirectional reflectivity was not recognized until recently.

The generic system is described by the index of refraction profile in Figure 1, where h_1 and h_2 are the layer thickness, and n_1 and n_2 are the indices of refraction of the respective layers. The incident wave has a wavevector $\vec{k} = k_x\hat{e}_x + k_y\hat{e}_y$, and frequency of $\omega = c|k|$. The wavevector together with the normal to the periodic structure defines a mirror plane of symmetry which allows us to distinguish between two independent electromagnetic modes: transverse electric (TE) modes and transverse magnetic (TM) modes. For the TE mode the electric field is perpendicular to the plane, as is the magnetic field for the TM mode.

General features of the transport properties of the finite structure can be understood when the properties of the infinite structure are elucidated. In a structure with an infinite number of layers, translational symmetry along the direction perpendicular to the layers leads to Bloch wave solutions of the form:

$$u_K(x, y) = E_K(y)\, e^{iKy}\, e^{ik_x x} \qquad [1]$$

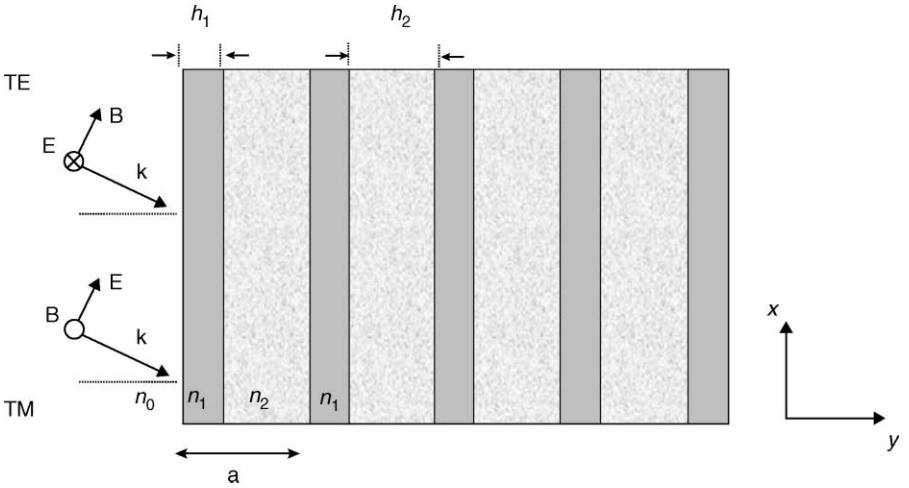


Figure 1 Schematic of the multilayer system showing the layer parameters (n_α, h_α – index of refraction and thickness of layer α), the incident wavevector $\vec{k}$ and the electromagnetic mode convention.

where $E_K(y)$ is periodic, with a period of length a, and K is the Bloch wave number. These waves represent solutions to an eigenvalue problem and are completely and uniquely defined by the specification of K, k_x and ω.

Solutions can be propagating or evanescent, corresponding to real or imaginary Bloch wave numbers, respectively. It is convenient to display the solutions of the infinite structure by projecting the $\omega(K, k_x)$ function onto the $\omega - k_x$ plane; Figure 2a,b are examples of such projected structures.

The gray background areas highlight phase space where K is strictly real, that is, regions of propagating states, while the white areas represent regions containing evanescent states. The shape of the projected bandstructures for the multilayer film can be understood intuitively. At $k_x = 0$, the bandgap for waves traveling normal to the layers is recovered. For $k_x > 0$, the bands curve upward in frequency. As $k_x \to \infty$, the modes become largely confined to the slabs with the high index of refraction and do not couple between layers (and are therefore independent of k_x).

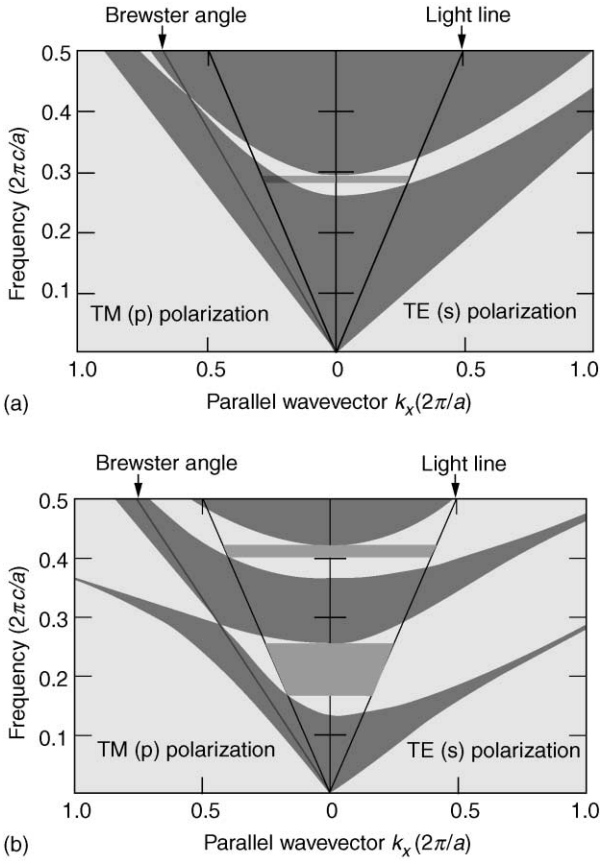


Figure 2 (a) Projected bandstructure of a multilayer film with the light line and Brewster line, exhibiting a reflectivity range of limited angular acceptance with ($n_1 = 2.2$ and $n_2 = 1.7$ and a thickness ratio of $h_2/h_1 = 2.2/1.7$). (b) Projected bandstructure of a multilayer film together with the light line and Brewster line, showing an omnidirectional reflectance range at the first and second harmonic (propagating states – dark gray, evanescent states – white, omnidirectional reflectance range – light grey). The film parameters are $n_1 = 4.6$ and $n_2 = 1.6$ with a thickness ratio of $h_2/h_1 = 1.6/0.8$. These parameters are similar to the actual polymer–tellurium film parameters measured in the experiment. Reproduced with permission from Fink Y, Winn JN, Fan S, *et al.* (1998) A dielectric omnidirectional reflector. *Science* 282: 1679–1682. Copyright 1998 American Association for the Advancement of Science.

In a finite structure, the translational symmetry in the directions parallel to the layers is preserved, hence k_x remains a conserved quantity and can be used to label solutions even though these solutions will no longer be of the Bloch form. The relevance of the band diagram to finite structures is that it allows for the prediction of regions of phase space where waves are evanescently decaying in the multilayer structure.

Since we are primarily interested in waves originating from the homogeneous medium external to the periodic structure, we will focus only on the portion of phase space lying above the light line. Waves originating from the homogeneous medium satisfy the condition $\omega \geq ck_x/n_0$, where n_0 is the refractive index of the homogeneous medium, and therefore they must reside above the light line. States of the homogeneous medium with $k_x = 0$ are normal incident, and those lying on the $\omega = ck_x/n_0$ line with $k_y = 0$ are incident at an angle of 90°.

The states in Figure 2a, that are lying in the restricted phase space defined by the light line and that have a (ω, k_x) corresponding to the propagating solutions (gray areas) of the crystal can propagate in both the homogeneous medium and in the structure. These waves will partially or entirely transmit through the film. Those with (ω, k_x) in the evanescent regions (white areas) can propagate in the homogeneous medium but will decay in the crystal – waves corresponding to this portion of phase space will be reflected off the structure.

The multilayer system leading to Figure 2a represents a structure with a limited reflectivity cone, since for any frequency one can always find a k_x vector for which a wave at that frequency can propagate in the crystal – and hence transmit through the film. The necessary and sufficient criterion for omnidirectional reflectivity at a given frequency is that there exists no transmitting state of the structure inside the light cone – this criterion is satisfied by frequency ranges marked in light gray in Figure 2b. In fact, the system leading to Figure 2b exhibits two omnidirectional reflectivity ranges.

The omnidirectional range is defined from above by the normal incidence bandedge $\omega_h(k_y = \pi/a, k_x = 0)$ (Figure 2b) and below by the intersection of the top of the TM allowed bandedge with the light line $\omega_l(k_y = \pi/a, k_x = \omega_l/c)$ (Figure 2b).

A dimensionless parameter used to quantify the extent of the omnidirectional range is the range to mid-range ratio defined as $(\omega_h - \omega_l)/\frac{1}{2}(\omega_h + \omega_l)$.

Figure 3 is a plot of this ratio as a function of n_2/n_1 and n_1/n_0, where ω_h and ω_l are determined by solutions of eqn [2] with quarter wave layer thickness. The contours in this figure represent various equi-omnidirectional ranges for different material index parameters and could be useful for design purposes.

At normal incidence, there is no distinction between TM and TE modes. At increasingly oblique angles, the gap of the TE mode increases, whereas the gap of the TM mode decreases. In addition, the center of the gap shifts to higher frequencies. Therefore, the criterion for the existence of omnidirectional reflectivity can be restated as the occurrence of a frequency overlap between the gap at normal incidence and the gap of the TM mode at 90°. Analytical expressions for the range to mid-range ratio can be obtained by setting:

$$\begin{aligned}\omega_h &= \frac{2c}{h_2 n_2 + h_1 n_1}\cos^{-1}\left(-\left|\frac{n_1 - n_2}{n_1 + n_2}\right|\right)\\ \omega_l &= \frac{2c}{h_2\sqrt{n_2^2 - 1} + h_1\sqrt{n_1^2 - 1}}\\ &\quad\times\cos^{-1}\left(\left|\frac{n_1^2\sqrt{n_2^2 - 1} - n_2^2\sqrt{n_1^2 - 1}}{n_1^2\sqrt{n_2^2 - 1} + n_2^2\sqrt{n_1^2 - 1}}\right|\right)\end{aligned} \quad [2]$$

Moreover, the maximum range width is attained for thickness values that are not equal to the quarter wave stack though the increase in bandwidth gained by deviating from the quarter wave stack is typically only a few percent.

In general, the TM mode defines the lower frequency edge of the omnidirectional range; an example can be seen in Figure 2b for a particular choice of the indices of refraction. This can be proven by showing that:

$$\left.\frac{\partial\omega}{\partial k_y}\right|_{TM} \geq \left.\frac{\partial\omega}{\partial k_y}\right|_{TE} \quad [3]$$

in the region that resides inside the light line. The physical reason for eqn [3] lies in the vectorial nature of the electric field. In the upper portion of the first band the electric field concentrates its energy in the high dielectric regions. Away from normal incidence, the electric field in the TM mode has a component in the direction of periodicity, this component forces a larger portion of the electric field into the low dielectric regions. The group velocity of the TM mode is therefore enhanced. In contrast, the electric field of the TE mode is always perpendicular to the direction of periodicity and can concentrate its energy primarily in the high dielectric region. While the omnidirectional reflection criteria can be used to confine light in various geometries, it is the cylindrical fiber configuration that appears to present significant application opportunities.

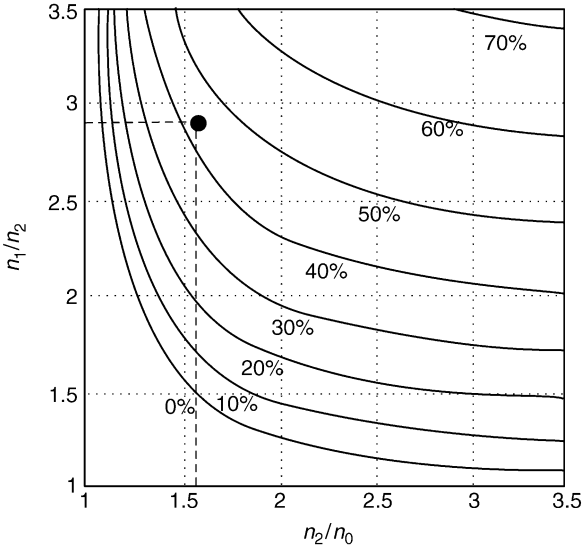


Figure 3 The range to mid-range ratio $(\omega_h - \omega_l)/\frac{1}{2}(\omega_h + \omega_l)$, for the fundamental frequency range of omnidirectional reflection, plotted as contours. Here, the layers were set to quarter wave thickness and $n_1 > n_2$. The ratio for our materials is approximately 45%. ($n_1/n_2 = 2.875$, $n_2/n_0 = 1.6$) is located at the intersection of the dashed lines (black dot). Reproduced with permission from Fink Y, Winn JN, Fan S, *et al.* (1998) A dielectric omnidirectional reflector. *Science* 282: 1679–1682. Copyright 1998 American Association for the Advancement of Science.

Mesostructured Fibers for External Reflection Applications

Polymer fibers are ubiquitous in applications such as textile fabrics, due to their excellent mechanical properties and the availability of low-cost, high-volume processing techniques; however, the control over their optical properties has so far remained relatively limited. Conversely, dielectric mirrors are used to precisely control and manipulate light in high performance optical applications, but the fabrication of these typically fragile mirrors has been mostly restricted to planar geometries and typically involves multiplicity of deposition sequences in a high vacuum thin film deposition system.

Fabrication Approach

The fabrication of extended lengths of omnidirectionally reflecting fibers with large bandgaps and

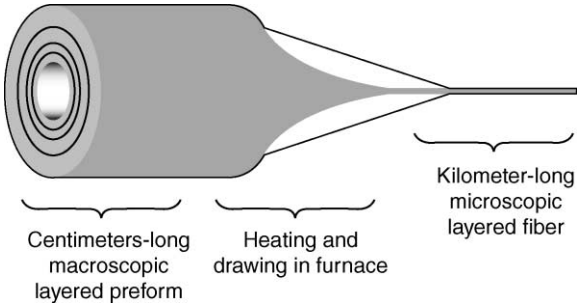


Figure 4 Conceptual preform based fabrication process for meso-structured fibers.

high layer counts poses considerable challenges. To illustrate the nature of this formidable task, one needs to consider the necessity of maintaining the uniformity of sub-100 nm layer thicknesses over kilometer-length scales; creating continuous layers with an aspect ratio of $\sim 10^{10}$ in a single process! To meet this and other challenges associated with the fabrication of mesostructured fibers, we have developed a preform-based fabrication approach (Figure 4). A scaled-up version of the final fiber, called a preform, is fabricated which shares the geometry and materials of the final fiber but exhibits macroscopic lateral features. The preform is heated up and drawn under tension into the fiber using a simple cylindrical furnace. The macroscopic layers are reduced in the process to microscopic dimensions while maintaining the overall geometry and symmetry of the original preform. Conservation of mass determines the final length of the fiber – typically this length is equal to the lateral reduction factor squared multiplied by the length of the preform. The nature of the process and requirements on the fiber's optical properties lead to the definition of materials selection rules. First, in order to achieve omnidirectional reflectivity, one needs to identify two solid materials exhibiting an index contrast given by the plot in Figure 3. Second, to enable the codrawing of the two dissimilar materials both will need to have viscosities that are lower than $\sim 10^8$ poise at the drawing temperature. In order to maintain high draw speeds, the majority component needs to be amorphous and the adhesion between these two materials needs to be sufficient to prevent delamination. Finally, their thermal expansion coefficients need to be close or alternatively at least one of the materials needs to be capable of relieving the stress due to CTE mismatch.

Materials Selection Criteria

Pairs of materials that are compatible with the process and fiber property requirements have been

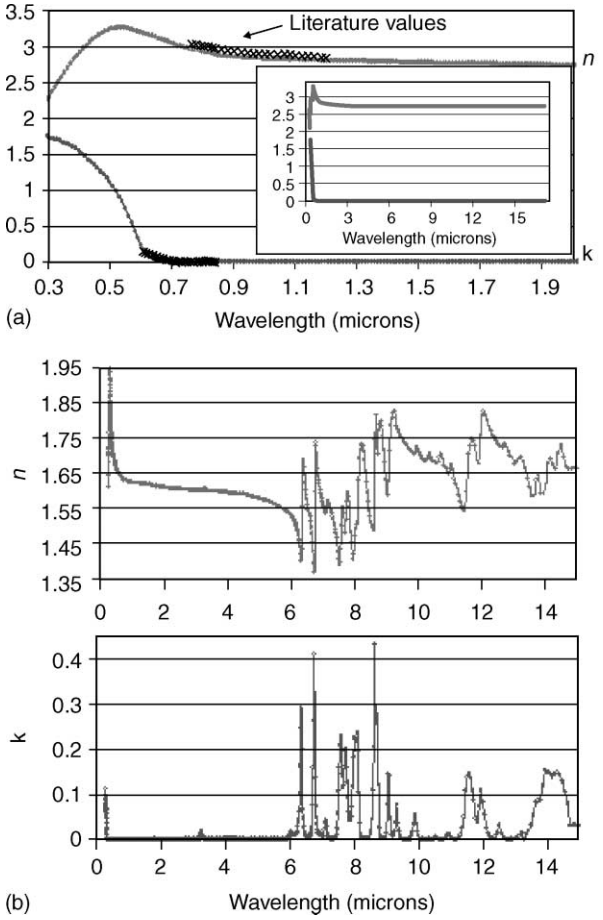


Figure 5 (a) Real part (n – light gray) and imaginary part (k – dark gray) of the refractive index of annealed As_2Se_3. The black crosses correspond to literature values. (b) Real part (n – upper) and imaginary part (k – lower) of the refractive index of PES.

identified, including: high glass transition temperature (T_g), thermoplastic polymers such as poly(ether-sulfone) (PES), poly (ether-imide) (PEI) and members of the chalcogenide glass family, such as arsenic triselenide (As_2Se_3) or arsenic trisulfide. Pairs of these materials will have substantially different refractive indices, as shown in Figure 5, which is a measurement of the real and imaginary indices of refraction of PES and As_2Se_3, obtained using a broadband spectroscopic ellipsometer (SOPRA GES5). They nevertheless exhibit similar thermomechanical properties within a certain thermal processing window.

Adhesion and extensional viscosity in the fluid state are difficult to measure in general, and the measurement of high-temperature surface tension is quite involved. Thus, limited data on these properties are available and it was necessary to empirically identify materials that could be used to draw out mirror-fibers. Various high-index chalcogenide (S, Se, and Te containing) glasses and low-index polymers were

identified as potential candidates based on their optical properties and overlapping thermal softening regimes. Adhesion and viscosity matching were tested by thermal evaporation of a chalcogenide glass layer on top of a polymer film or rod and elongation of the coated substrate at elevated temperatures. The choice of a high-temperature polymer, PES, and a simple chalcogenide glass, As_2Se_3, resulted in excellent thermal co-deformation without film cracking or delamination. Approximate matching of extensional viscosity in this manner was also demonstrated using As_2Se_3 and PEI. The properties, processing, and applications of chalcogenide glasses have been explored extensively elsewhere. One advantage in choosing As_2Se_3 for this application is that not only is it a stable glass, but it is a stoichiometric compound that can be readily deposited in thin films through thermal evaporation or sputtering without dissociation. Additionally, As_2Se_3 is transparent to IR radiation from approximately 0.8 to 17 μm as shown in the ellipsometric data (Figure 5a), and has a refractive index of ~2.8 in the mid-IR. PES is a high-performance, dimensionally stable thermoplastic with a refractive index of ~1.6 and good transparency to EM waves in a range extending from the visible regime into the mid-IR (~5 μm), as shown in Figure 5b.

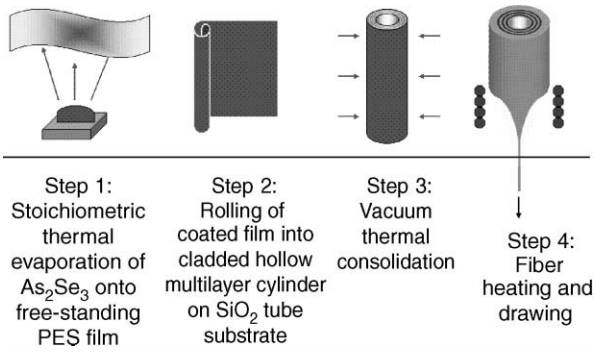


Figure 6 Multilayer preform fabrication sequence.

Preform Construction Process and Fiber Draw

The selected materials were used to construct a multilayer preform rod, which essentially is a macroscale version of the final fiber. In order to fabricate the dielectric mirror fiber preform, an As_2Se_3 film was deposited through thermal evaporation on either side of a free-standing PES film which was then rolled on top of a PES tube substrate, forming a structure having 21 alternating layers of PES and As_2Se_3, as shown in Figure 6.

The resulting multilayer fiber preform was subsequently drawn down using an optical fiber draw tower into hundreds of meters of multilayer fiber with a precisely controlled submicron layer thickness, creating a photonic bandgap in the mid-IR. Fibers of outer diameters varying from 175–500 μm with a typical standard deviation of 10 μm from target, were drawn from the same preform to demonstrate adjustment of the reflectivity spectra through thermal deformation. The spectral position of the photonic bandgap was controlled by the optical monitoring of the outer diameter (OD) of the fiber during draw, which was later verified by reflectivity measurements on single and multiple fibers of different diameters. Scanning electron micrographs (SEMs) of the cross-section of these fibers are depicted in Figure 7.

Bandstructure for Multilayer Fibers for External Reflection Applications

In theoretically predicting the spectral response of these fibers, it is helpful to calculate the photonic

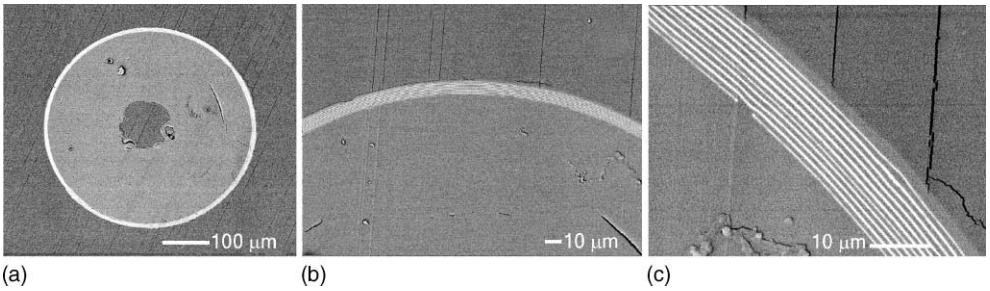


Figure 7 SEM micrographs of 400 μm OD fiber cross-section. The entire fiber is embedded in epoxy. (a) shows the entire fiber cross-section, with mirror structure surrounding the PES core; (b) demonstrates that the majority of the fiber exterior is free of significant defects and that the mirror structure adheres well to the fiber substrate; and (c) reveals the ordering and adhesion within the alternating layers of As_2Se_3 (bright layers) and PES. Stresses developed during sectioning caused some cracks in the mounting epoxy that are deflected at the fiber interface. Fibers from this batch were used in the reflectivity measurements recorded below in Figure 9a. Reproduced with permission from Hart SD, Maskaly GR, Temelkuran B, Prideaux PH, Joannopoulos JD and Fink Y (2002) External reflection from omnidirectional dielectric mirror fibers. *Science* 296: 510–513. Copyright 2002 American Association for the Advancement of Science.

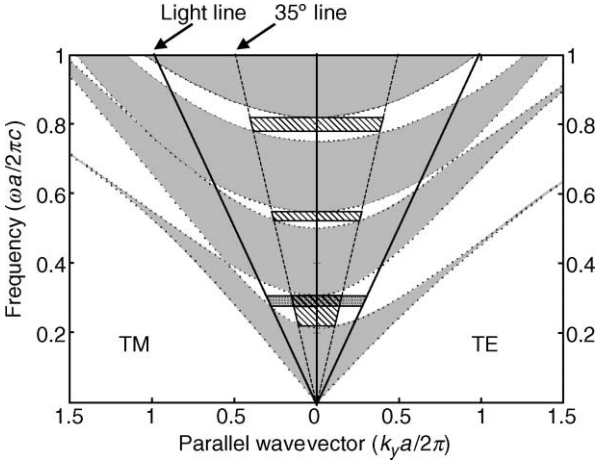


Figure 8 Photonic band diagram for a one-dimensional photonic crystal having a periodic refractive index alternating between 2.8 and 1.55. Gray regions represent propagating modes within the structure, while white regions represent evanescent modes. Hatched regions represent photonic bandgaps where high reflectivity can be expected for external EM waves over an angular range extending from normal to 35° incidence. The lower dark shaded trapezoid represents a region of external omnidirectional reflection.

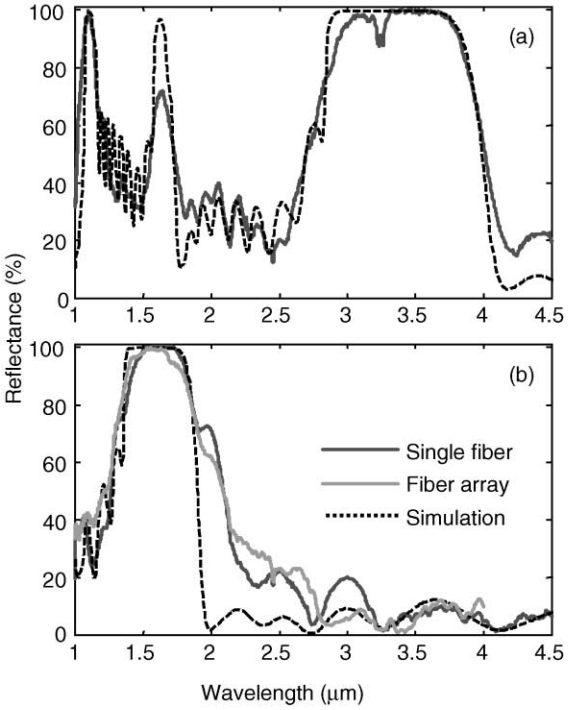


Figure 9 Measured reflectance spectra for 400 μm OD (a) and 200 μm OD (b) dielectric mirror fibers relative to gold-coated fibers of the same diameter. (a) shows a single-fiber reflectivity measurement, while (b) compares single-fiber reflectivity to that measured from a multifiber array. Simulations were performed using the transfer matrix method.

bandstructure that corresponds to an infinite one-dimensional photonic crystal. This allows for the analysis of propagating and evanescent modes in the structure, corresponding to real or imaginary Bloch wave number solutions.

The electric or magnetic field vector is parallel to the mirror layer interfaces for the TE and TM polarized modes, respectively. The parallel wavevector (k_y) is the component of the incident electromagnetic (EM) wavevector that is parallel to the layer interfaces. The phase space accessible from an external ambient medium is contained between the light lines (defined by the glancing-angle condition $\omega = ck_y/n_0$), and the modes between the 35° lines correspond to those sampled experimentally. Axes are normalized to the thickness a of one mirror bilayer (a period consisting of one high- and one low-index layer). Figure 8 depicts the photonic band diagram for an infinite structure having similar periodicity and refractive indices to the mirror structures fabricated here. Three photonic bandgaps are present where high reflectivity is expected within the 0–35° angular range, and the fundamental gap contains a region of external omnidirectional reflectivity.

Optical Characterization of 'Mirror Fibers'

Mirror fiber reflectivity was measured from both single fibers and parallel fiber arrays using a Nicolet/SpectraTech NicPlan Infrared Microscope and Fourier Transform Infrared Spectrometer (Magna 860). The microscope objective (SpectraTech 15×, Reflachromat) used to focus on the fibers had a numerical aperture (NA) of 0.58. This results in a detected cone where the angle of reflection with respect to the surface normal of the structure could vary from normal incidence to ~35°, which is determined by the NA of the microscope objective. As a background reference for the reflection measurements, we used gold-coated PES fibers of matching diameters. Dielectric mirror fibers, drawn to 400 μm OD, exhibited a very strong reflection band centered at 3.4 μm wavelength (Figure 9a). Measured reflectivity spectra agree well with planar-mirror transfer matrix method (TMM) simulations, where the reflectivity was averaged across the aforementioned angular range for both polarization modes. Fibers drawn down to 200 μm OD show a similar strong fundamental reflection band centered near 1.7 μm (Figure 9b). This shifting of the primary photonic bandgap clearly illustrates the precise tuning of the reflectivity spectra over wide frequency ranges through thermal deformation processing. Strong optical signatures are measurable from single fibers as small as 200 μm OD. Fiber array measurements,

simultaneously sampling reflected light from multiple fibers, agree quite well with single-fiber data (Figure 9b).

These reflectivity results are strongly indicative of uniform layer thickness control, good interlayer adhesion, and low interdiffusion through multiple thermal treatments. This was confirmed by SEM inspection of fiber cross-sections (Figure 7). The layer thicknesses observed ($a = 0.90$ μm for the 400 μm fibers; $a = 0.45$ μm for the 200 μm fibers) correspond well to the measured reflectivity spectra. The fibers have a hole in the center, due to the choice of a hollow rod as the preform substrate, which experienced some nonuniform deformation during draw. The rolled-up mirror structure included a double outer layer of PES for mechanical protection, creating a noticeable absorption peak in the reflectivity spectrum at ~3.2 μm (Figure 9a).

A combination of spectral and direct imaging data demonstrates excellent agreement with the photonic band diagram. The measured gap width (range to mid-range ratio) of the fundamental gap for the 400 μm OD fiber is 27%, compared to 29% in the photonic band diagram.

Tunable 'Fabry–Perot' Fibers

The fabrication of fibers surrounded by or lined with alternating layers of materials with a large disparity in their refractive indices, presents interesting opportunities for passive and active optical devices. While a periodic multilayer structure, such as the one reported above, leads to the formation of photonic bandgaps and an associated range of high reflectivity, it is the incorporation of intentional deviations from periodicity, also called 'defects', which allows for the creation of localized electromagnetic modes in the vicinity of the defect. These structures, sometimes called optical cavities, in turn can provide the basis for a large number of interesting passive and active optical devices such as vertical cavity surface emitting lasers (VCSELs), bi-stable switches, tuneable dispersion compensators, tuneable drop filters, etc. Here we report on the fabrication of a fiber surrounded by a Fabry–Perot cavity structure and demonstrate that by application of axial mechanical stress, the spectral position of the resonant Fabry–Perot mode can be reversibly tuned.

Structure and Optical Properties of the Fabry–Perot Fibers

The fabrication technique described above allows for the accurate placement of optical cavities that can encompass the entire or partial fiber circumference. The fibers discussed above are made of As_2Se_3 and PES and have a low index Fabry–Perot cavity (Figure 10).

This structure was achieved by introducing an extra polymer layer in the middle of the periodic multilayer structure of the preform, thus generating a defect mode in the photonic bandgaps of the drawn fibers. The position of the bandgap center is linearly related to the optical thickness of the layers by the Bragg condition. Through accurate outer diameter control afforded by the laser micrometer mounted on the draw tower we have been able to place gaps (Figure 11) at wavelengths ranging from 11 microns to below 1 micron. Such cost-effective tuneable optical filters could lead to applications such as optical

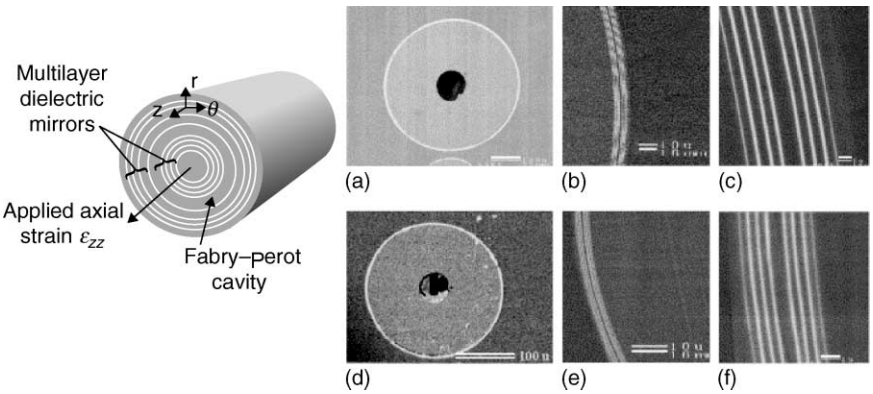


Figure 10 Schematic of the structure of dielectric mirror fibers made of As_2Se_3 (light gray) and PES (dark gray) with low index Fabry–Perot cavity. The local cylindrical coordinate system is represented as well as the applied axial strain ε_{zz}. Typical radii are ~150 μm ± 100 μm. Backscattered SEM micrographs of the cross-sections of a 460 micron diameter fiber (a, b, c) and of a 240 micron diameter fiber (d, e, f) embedded in epoxy and microtomed. (a) and (d) show the entire cross-section of the fibers, (b) and (e) demonstrate long-range layer uniformity, and (c) and (f) reveal the ordering and adhesion of the Fabry–Perot cavity structure. Bar scales have been redrawn for clarity. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

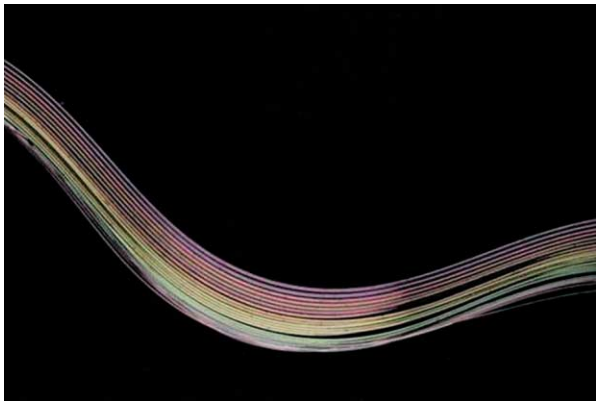


Figure 11 Array of parallel fibers with outer diameter ranging from ~420 μm (bottom) to ~100 μm (top). The colors are due to the narrow 4th order photonic bandgap in the visible.

switches for wavelength-division-multiplexing (WDM) systems and sensors.

The cross-sectional structures of a 240 micron and a 460 micron diameter fiber were observed by an SEM using a backscattered electron detector (Figure 10). The structure is composed of a hollow core polymer rod surrounded by six bilayers of As_2Se_3 and PES, separated in the middle by an extra polymer layer, which forms the Fabry–Perot cavity. An extra polymer layer protects the fiber surface. For the 240 micron (460 micron) diameter fiber, the glass layers are ~135 nm (250 nm) thick, except for the first and the last ones which are half as thick due to the fabrication technique; the polymer layers are ~270 nm (540 nm) thick and the defect layer is ~610 nm (1170 nm) thick.

Reflectivity spectra measurements were performed under a microscope using a (Nicolet SpectraTech NicPlan) Infrared Microscope and Fourier Transform Infrared Spectrometer (Magna 860) with a lens numerical aperture (NA) corresponding to 30 degrees of angular spread. They exhibited a single-mode Fabry–Perot resonant mode at 1.74 and 3.2 μm for the 240 and 460 micron diameter fiber, respectively (Figure 12).

Because of the range of incident angles, the measured quality factor (Q-factor, defined as its spectral position divided by the full width half maximum (FWHM)) was equal to 31. Using the different thicknesses and the real and imaginary part of the refractive index ($n \& k$) of As_2Se_3 and PES carefully measured with a broadband (300 nm to 16 microns) spectroscopic ellipsometer (Sopra GES-5), the reflectivity spectra of these fibers were computed with the TMM by approximating them as a one-dimensional planar stacking. By averaging the calculated spectra over the accessible incident angles

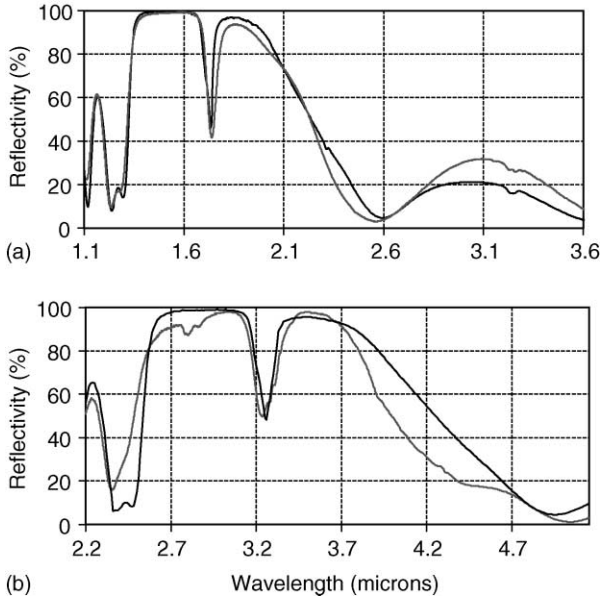


Figure 12 Computed (black lines) and measured (grey lines) reflectivity spectra for the 240 micron (a) and of the 460 micron diameter fibers (b) with Fabry–Perot resonant modes at 1.74 and 3.2 μm, respectively. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

and polarizations, we obtained a very good agreement between the simulation and the measurements.

Simulation of the Opto-Mechanical Behavior of the Fabry–Perot Fibers

We focused our analysis on the elastic regime, which ultimately limits the operational range of these fibers. The fiber's Young's modulus (E) can be approximated by modeling this multilayer structure as independent parallel springs under axial (along the z-axis of the fiber) strain assumed to be equal for all the layers, leading to a Young's modulus of 2.64 GPa (using $E_{PES} = 2.4$ GPa for bulk PES and $E_{As_2Se_3} = 15$ GPa reported for 1.5 μm thick films).

Neglecting the possible strain-induced refractive index variation of the materials, the normalized shift of the Fabry–Perot resonant mode can be related to the applied axial strain by calculating the radial stresses σ_{rr} (resulting from the difference between the Poisson ratios of the materials – $\nu_{PES} = 0.45$ and $\nu_{As_2Se_3} = 0.289$ – and the adhesion condition between the layers) and displacements u_r in each layer under axial strain. Starting from the equilibrium equations:

$$\frac{\partial \sigma_{zz}}{\partial z} = 0 \qquad [4]$$

$$\frac{\partial \sigma_{rr}}{\partial r} + \frac{\sigma_{rr} - \sigma_{\theta\theta}}{r} = 0 \Leftrightarrow \frac{d^2 u_r}{d^2 r} + \frac{1}{r}\frac{du_r}{dr} - \frac{u_r}{r^2} = 0 \qquad [5]$$

and using Hooke's law and Lame's equations, we can derive a general expression for u_r and σ_{rr}, with two unknowns A and B per layer:

$$u_r = \frac{A}{r} + Br \qquad [6]$$

$$\sigma_{rr} = -\frac{2GA}{r^2} + 2(\eta + G)B + \lambda\varepsilon_{zz} \qquad [7]$$

where $G = E/[2(1+\nu)]$ is the shear modulus and $\eta = E\nu/[(1+\nu)(1-2\nu)]$ the Lame modulus. The continuity of the displacement and the radial stress at each interface, plus the boundary conditions (σ_{rr} vanishes at free surfaces) can be expressed as a linear system whose unique solution allows us to relate the radial strain in each layer to the applied axial strain as a linear relation $\varepsilon_{rr} = C\varepsilon_{zz}$, where C can be interpreted as an effective Poisson ratio. Finally, taking into account that the Fabry–Perot resonant mode itself is linearly shifted within the bandgap because of the different effective Poisson ratios of the glass and polymer layers, we obtain a linear relation between the normalized shift of the Fabry–Perot resonant mode and the applied axial strain:

$$\frac{\Delta\lambda}{\lambda} = -0.373\varepsilon_{zz} \qquad [8]$$

All the layers (except the outer protective polymer layer) are under tensile radial stress whose maximum (0.22 MPa under 1% axial strain) is located at the interface between the layers and the polymer core where delamination is most likely to occur.

Mechanical Tuning Experiment and Discussion

Measurements were performed on fibers ~30 cm long, which were fixed at one end with epoxy to a load cell (Transducer Techniques MDB-2.5) while the other end was attached with strong tape to a pole mounted on a stepper rotational stage (Newport PR50) (Figure 13). This end of the fiber was also screwed to the pole to further secure it in place.

The diameter of the pole was equal to 2.1 cm, leading to a normalized shift precision below 0.005%. The uncertainty on the Young's modulus, due to the precision of the load cell, was lower than 30 MPa. All the reflectivity spectra were normalized to a background taken with a flat gold mirror. The measurements were realized as far as possible from the fixed ends of the fiber where edge effects are likely to occur. These edge effects result in a reduction of the length of the fiber that deforms uniformly and consequently increase the real strain far from the edges by a factor of 1.14 and 1.15 for the 240 and the 460 micron diameter fiber, respectively (determined experimentally by measuring the position of two reference points on the fiber) compared to the strain calculated from the rotation of the pole. Moreover, to avoid measuring slight variations in the spectral position of the bandgap resulting from outer diameter variations, typically of the order of 4–5 microns over meters of fibers, the measurements were realized at a fixed reference position on the fiber.

By focusing on the fundamental bandgap of the 240 and the 460 micron diameter fiber, we demonstrated the tuning of the Fabry–Perot resonant mode under increasing axial strain (Figure 14).

A drop in the reflectivity of 13% was observed at 1.71 μm (dash line) for the 240 micron diameter fiber when increasing the applied axial strain from 0.23% (light gray) to 1.07% (dark gray). The normalized-shift versus strain curves (Figure 15) appeared to be linear for both fibers up to approximately 0.9% (dash line) with a slope equal to −0.3859 and −0.3843 for the 240 and the 460 micron diameter fiber, respectively, close to the predicted value (−0.373, eqn [8]).

The normalized shift was equal to −0.347% for 0.9% applied axial strain, which seems to be the limit of the elastic regime and corresponds to small applied

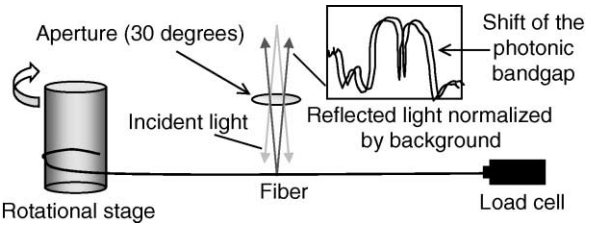


Figure 13 Experimental setup used for mechanical tuning demonstration. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

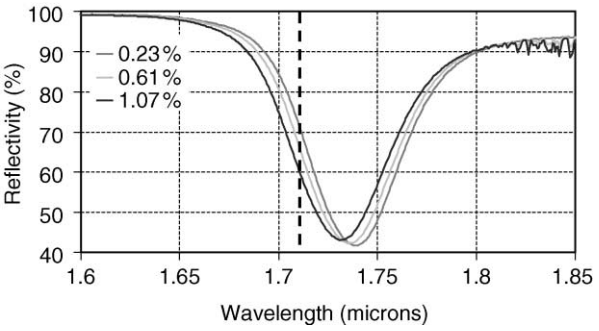


Figure 14 Reflectivity versus wavelength plot showing the shift of the Fabry–Perot resonant mode of the 240 micron diameter fiber for three increasing values of the applied axial strain. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

loads: 76 and 300 g for the 240 and 460 micron diameter fiber, respectively. Under higher strains, the normalized-shift versus strain curves were no longer linear and the measured loads would start decreasing slightly with time, which could be a sign of delamination between the layers and the polymer core, of plastic deformation and/or of relaxation in the layers. The stress–strain curves also exhibit an elastic regime up to approximately 1% strain with a corresponding Young's modulus equal to 2.59 GPa for the 240 μm OD fiber and to 2.39 GPa for the 460 μm OD fiber. These values are slightly lower than the predicted value, possibly due to relaxation effects in the polymer and the glass layers (for As_2Se_3 and PES, $T_g = 175$ and 220 °C, respectively).

Wavelength-Scalable Hollow Optical Fibers with Large Photonic Bandgaps for CO_2 Laser Transmission

Hollow optical transmission fibers offer the potential to circumvent fundamental limitations associated with conventional index guided fibers and thus have been the subject of active research in recent years. Here we report on the materials selection, design, fabrication, and characterization of extended lengths of hollow optical fiber lined with an interior omnidirectional dielectric mirror. These fibers consist of a hollow air core surrounded by multiple alternating submicron-thick layers of a high-refractive-index glass and a low-index polymer, resulting in large infrared photonic bandgaps. These gaps provide strong confinement of optical energy in the hollow fiber core and lead to light guidance in the fundamental and up to fourth-order gaps. We show that the fiber transmission windows can be scaled over a large wavelength range covering at least 0.75 to 10.6 microns. The utility of our approach is further demonstrated by the design and fabrication of tens of meters of hollow photonic bandgap fibers for 10.6 micron radiation transmission. We demonstrate transmission of carbon dioxide (CO_2) laser light with high power-density through more than 4 meters of hollow fiber and measure the losses to be less than 1.0 dB/m at 10.6 microns. This establishes suppression of fiber waveguide losses by orders of magnitude compared to the intrinsic fiber material losses.

Silica optical fibers have been extremely successful in telecommunications applications, and other types of solid-core fibers have been explored at wavelengths where silica is not transparent. However, all fibers that rely on light propagation principally through a solid material have certain fundamental limitations stemming from nonlinear effects, light absorption by electrons or phonons, material dispersion, and Rayleigh scattering that limit maximum optical transmission power and increase attenuation losses. These limitations have, in turn, motivated the study of a fundamentally different light-guiding methodology: the use of hollow waveguides having highly reflecting walls. Light propagation through air in a hollow fiber eliminates or greatly reduces the problems of nonlinearities, thermal lensing, and end-reflections, facilitating high-power laser guidance and other applications which may be impossible using conventional fibers. Hollow metallic or metallo-dielectric waveguides have been studied fairly extensively and found useful practical application, but their performance has been bounded by the notable losses occurring in metallic reflections at visible and infrared (IR) wavelengths, as well as by the limited length and mechanical flexibility of the fabricated waveguides. Hollow all-dielectric fibers, relying on specular or attenuated total reflection, have also been explored, but high transmission losses have prevented their broad application. More recently, all-dielectric fibers,

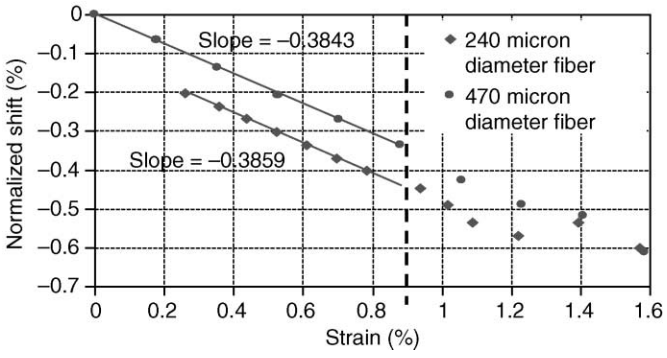


Figure 15 Normalized shift of the Fabry–Perot resonant mode (defined as $\Delta\lambda/\lambda$) versus applied axial strain for the 240 micron (diamonds) and the 460 micron diameter fiber (dots). The curve obtained for the 240 micron diameter fiber is lower because the experimental 0% strain is likely to correspond to a nonzero positive strain, necessary to keep this thinner fiber straight for the measurement (equivalent to a load less than 10 g). Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

consisting of a periodic array of air holes in silica, have been used to guide light through air using narrow photonic bandgaps. Solid-core, index-guiding versions of these silica photonic crystal fibers have also been explored for interesting and important applications, such as very large core single-mode fibers, nonlinear enhancement and broadband supercontinuum generation, polarization maintenance, and dispersion management. However, the air-guiding capabilities of such waveguides thus far remain inferior to transmission through solid silica, due to various factors such as the difficulties in fabricating long, uniform fibers which must have a high volume fraction of air and many air–hole periods, as well as by the large electromagnetic (EM) penetration depths associated with the small photonic bandgaps achievable in these air–silica structures. In our fiber, the hollow core is surrounded by a solid high-refractive-index-contrast multilayer structure leading to large photonic bandgaps and omnidirectional reflectivity. The pertinent theoretical background and recent analyses indicate that such fibers may be able to achieve ultralow losses and other unique transmission properties. The large photonic bandgaps result in very short EM penetration depths within the layer structure, significantly reducing radiation and absorption losses while increasing robustness. Omnidirectional reflectivity is expected to reduce intermode coupling losses.

To achieve high index contrast in the layered portion of the fiber, we combined a chalcogenide glass with a refractive index of ~2.8 As_2Se_3, and a high-performance polymer with a refractive index of ~1.55 PES. We recently demonstrated that these materials could be thermally co-drawn into precisely layered structures without cracking or delamination, even under large temperature excursions. The same polymer was used as a cladding material, resulting in fibers composed of ~98% polymer by volume (not including the hollow core) and thus combine high optical performance with polymeric processability and mechanical flexibility. We fabricated a variety of fibers by depositing a 5–10 micron thick As_2Se_3 layer through thermal evaporation onto a 25–50 micron thick PES film and the subsequent 'rolling' of that coated film into a hollow multilayer tube called a fiber preform. This hollow macroscopic preform was consolidated by heating under vacuum and cladded with a thick outer layer of PES; the layered preform was then placed in an optical fiber draw tower and drawn down into tens or hundreds of meters of fiber having well-controlled submicron layer thicknesses. The nominal positions of the photonic bandgaps were determined by laser monitoring of the fiber OD during the draw process. Typical standard deviations in the fiber OD were ~1% of the OD. The resulting fibers were designed to have large hollow cores, useful in high-energy transmission.

SEM analysis (Figure 16) reveals that the drawn fibers maintain proportionate layer thickness ratios and that the PES and As_2Se_3 films adhere well during rigorous thermal cycling and elongation. Within the multilayer structure shown in Figure 1, the PES layers (gray) have a thickness of 900 nm, and the As_2Se_3 layers (bright) are 270 nm thick (except for the first and last As_2Se_3 layers, which are 135 nm). Broadband fiber transmission spectra were measured with a Fourier transform infrared (FTIR) spectrometer (Nicolet Magna 860), using a parabolic mirror to couple light into the fiber and an external detector. The results of these measurements are shown in the lower panel of Figure 2 for fibers having two different layer structures. For each spectrum, light is guided at the fundamental and high-order photonic bandgaps. Also shown in the upper panel of Figure 2, is the corresponding photonic band diagram for an infinite periodic multilayer structure calculated using the experimental parameters of our fiber (layer thicknesses and indices). Good agreement is found between the positions of the measured transmission peaks and the calculated bandgaps, corroborated with the SEM-measured layer thicknesses, verifying that transmission is dominated by the photonic bandgap mechanism. In order to demonstrate 'wavelength scalability' (i.e., the control of transmission through the fiber's structural parameters) another fiber was produced, having the same cross-section but with thinner layers. We compared the transmission

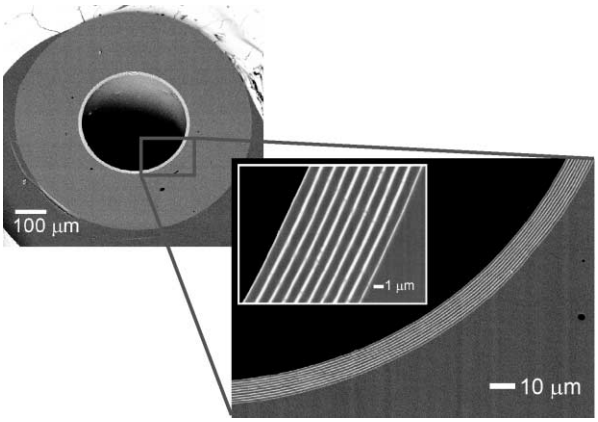


Figure 16 Cross-sectional SEM micrographs at various magnifications of hollow cylindrical multilayer fiber mounted in epoxy. The hollow core appears black, the PES layers and cladding gray, and the As_2Se_3 layers bright white. This fiber has a fundamental photonic bandgap at a wavelength of ~3.55 microns. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

spectra for the original 3.55 micron bandgap fibers to the fiber with the scaled-down layer thicknesses.

Figure 17 shows the shifting of the transmission bands, corresponding to fundamental and high-order photonic bandgaps, from one fiber to the next. The two fibers analyzed in Figure 17 were fabricated from the same fiber preform using different draw-down ratios (fibers with a fundamental bandgap centered near 3.55 microns have an OD of 670 microns; those with a gap at 3.1 microns have an OD of 600 microns). The high-order bandgaps are periodically spaced in frequency, as expected for such a photonic crystal structure.

The wavelength scalability of our fibers was further demonstrated in the fabrication of hollow fibers designed for the transmission of 10.6 micron EM radiation. This not only shows that these structures can be made to guide light at extremely disparate wavelengths, but that specific useful bandgap wavelengths can be accurately targeted during fabrication and fiber drawing. Powerful and efficient CO_2 lasers

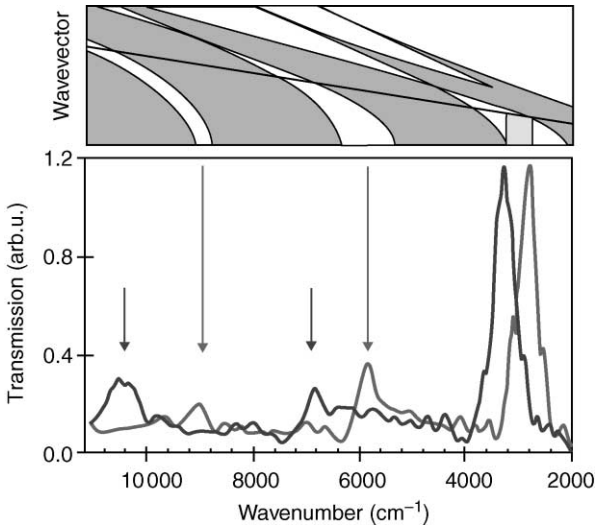


Figure 17 Upper panel: Calculated photonic bandstructure associated with the dielectric mirror lining of the hollow fiber. Modes propagating through air and reflected by the fiber walls lie in the bandgaps (white) and within the light cone defined by the glancing-angle condition (black line). The gray regions represent modes radiating through the mirror. The fundamental bandgap has the widest range-to-mid-range ratio, with a region of omnidirectional reflectivity highlighted in black. Lower panel: Comparison of transmission spectra for two different hollow fibers of ~30 cm length having similar structure but scaled photonic crystal (multilayer) period dimensions. The spectrum in gray is from a fiber with a fundamental photonic bandgap at 3.55 microns; the black spectrum is from a fiber where the corresponding bandgap is at 3.1 microns. High-order bandgaps (indicated by arrows) are periodically spaced in frequency. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

are available that emit at 10.6 microns and are used in such applications as laser surgery and materials processing, but waveguides operating at this wavelength have remained limited in length or loss levels. Using the fabrication techniques outlined above, we produced fibers having hollow core diameters of 700–750 microns and ODs of 1300–1400 microns with a fundamental photonic bandgap spanning the 10–11 micron wavelength regime, centered near 10.6 microns. Figure 5 depicts a typical FTIR transmission spectrum for these fibers, measured using ~30 cm long straight fibers.

In order to quantify the transmission losses in these 10.6 micron bandgap hollow fibers, fiber cutback measurements were performed. This involved the comparison of transmitted intensity through ~4 meters of straight fiber with the intensity of transmission through the same section of fiber cut to shorter lengths (Figure 18 inset). This test was performed on multiple sections of fiber, and the results found to be nearly identical for the different sections tested. The measurements were performed using a 25 watt CO_2 laser (GEM-25, Coherent-DEOS) and high power detectors (Newport 818T-10). The fiber was held straight, fixed at both ends as well as at multiple points in the middle to prevent variations in the input coupling and propagation conditions during fiber cutting. The laser beam was sent through focusing lenses as well as 500 micron diameter pinhole apertures and the input end face of the fiber was coated with a metal film to prevent accidental

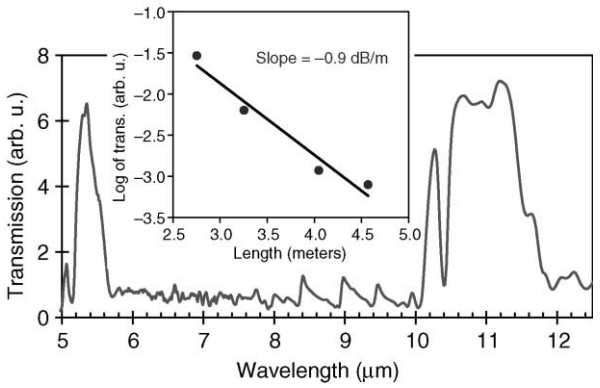


Figure 18 Typical transmission spectrum of hollow fibers designed to transmit CO_2 laser light. The fundamental photonic bandgap is centered near a wavelength of 10.6 microns and the second-order gap is at ~5 microns. Inset: Log of transmitted power (arbitrary units) versus length of fiber (meters). The slope of this graph is the loss in dB/m. The measured fiber has a hollow core diameter of 700 microns. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

laser damage from misalignment. The transmission losses in the fundamental bandgap at 10.6 microns were measured to be 0.95 dB/m, as shown in the inset of Figure 18, with an estimated measurement uncertainty of 0.15 dB/m. These loss measurements are comparable to some of the best reported loss values for other types of waveguides operating at 10.6 microns. A bending analysis for fibers with a bandgap centered at 10.6 microns revealed bending losses below 1.5 dB for 90 degree bends with bending radii from 4–10 cm. We expect that these loss levels could be lowered even further by increasing the number of layers, through optimization of the layer thickness ratios and by creating a cylindrically symmetric multilayer fiber with no inner seam (present here because of the 'rolling' fabrication method). In addition, using a polymer with lower intrinsic losses should greatly improve the transmission characteristics.

One reasonable figure of merit for optical transmission losses through hollow all-dielectric photonic bandgap fibers is to compare the hollow fiber losses to the intrinsic losses of the materials used to make the fiber. As_2Se_3 has been explored as an IR-transmitting material, yet the losses at 10.6 microns reported in the literature are ~10 dB/m for highly purified material, and more typically are greater than 10 dB/m for commercially available materials such as those used in our fabrication. Based on FTIR transmission and spectroscopic ellipsometer measurements that we have performed on PES, the optical losses associated with propagation through solid PES should be greater than 40 000 dB/m at 10.6 microns. This demonstrates that guiding light through air in our hollow bandgap fibers leads to waveguide losses that are orders of magnitude lower than the intrinsic fiber material losses, which has been one of the primary goals of hollow photonic bandgap fiber research. These comparatively low losses are made possible by the very short penetration depths of EM waves in the high refractive index contrast photonic crystal structure, allowing these materials to be used at wavelengths that may have been thought improbable. Another long-standing motivation of infrared fiber research has been the transmission of high-power laser light. As a qualitative demonstration of the potential of these fibers for such applications, both straight and smoothly bent fibers of lengths varying from 0.3–2.5 meters were used to transmit enough CO_2 laser energy to burn holes through paper. The maximum laser power density coupled into our fibers in these trials was approximately 300 W/cm^2, more than sufficient to burn a homogeneous polymer material, including PES. No damage to the fibers was observed when the laser beam was properly coupled into the hollow fiber core. These results indicate the feasibility of using hollow multilayer photonic bandgap fibers as a low-loss wavelength-scalable transmission medium for high-power laser light.

See also

Diffractive Systems: Design and Fabrication of Diffractive Optical Elements. **Fiber and Guided Wave Optics:** Fabrication of Optical Fiber. **Lasers:** Carbon Dioxide Lasers. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Abeles F (1950) Investigations on the propagation of sinusoidal electromagnetic waves in stratified media. Application to thin films. *Annales de Physique* 5: 706.

Benoit G and Fink Y, Spectroscopic Ellipsometry Database, http://mit-pbg.mit.edu/Pages/Ellipsometry.html

Born M and Wolf E (1980) *Principles of Optics*, 6th edn. New York: Pergamon Press, p. 67.

Engeness TD, Ibanescu M, Johnson SG, *et al.* (2003) Dispersion tailoring and compensation by modal interactions in OmniGuide fibers. *Optics Express* 11: 1175–1198. http://www.opticsexpress.org/abstract.cfm?URI = OPEX-11-10-1175

Fink Y, Winn JN, Fan S, *et al.* (1998) A dielectric omnidirectional reflector. *Science* 282: 1679–1682.

Fink Y, Ripin DJ, Fan S, Chen C, Joannopoulos JD and Thomas EL (1999) Guiding optical light in air using an all-dielectric structure. *Journal of Lightwave Technology* 17: 2039–2041.

Hart SD, Maskaly GR, Temelkuran B, Prideaux PH, Joannopoulos JD and Fink Y (2002) External reflection from omnidirectional dielectric mirror fibers. *Science* 296: 510–513.

Ibanescu M, Johnson SG, Soljacic M, *et al.* (2003) Analysis of mode structure in hollow dielectric waveguide fibers. *Physics Review E* 67: 046608-1–8.

Johnson SG, Ibanescu M, Skorobogatiy M, *et al.* (2001) Low-loss asymptotically single-mode propagation in large-core OmniGuide fibers. *Optics Express* 9: 748–779. http://www.opticsexpress.org/abstract.cfm?URI = OPEX-9-13-748

Kuriki K, Shapira O, Hart SD, *et al.* (2004) Hollow multilayer photonic bandgap fibers for NIR applications. *Optics Express* 12: 8.

Soljacic M, Ibanescu M, Johnson SG, Joannopoulos JD and Fink Y (2003) Optical bistability in axially modulated OmniGuide fibers. *Optics Letters* 28: 516–518.

Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

Yeh P, Yariv A and Hong C-H (1977) Electromagnetic propagation in periodic stratified media. 1. General theory. *Journal of the Optical Society of America* 67(4): 423.

Yeh P, Yariv A and Marom E (1978) Theory of Bragg fiber. *Journal of the Optical Society* 68: 1196–1201.

Wave Optical Modeling and Design

F Wyrowski, Friedrich-Schiller University, Jena, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Why do ray optics dominate the literature on optical engineering? The major practical reason may be that historically optical engineering has dealt with systems for image forming. These systems transform divergent spherical waves from points in an object, into spherical waves converging to focal points in the image. Thus, the basic operation of an imaging system performed by a lens is changing the radius of curvature of spherical fields. The propagation of spherical waves is modeled well by geometrical optics, so long as truncation by apertures and other high frequency aberrations are not significant. Moreover, the basic merit function in imaging evaluates the appearance of the image point at its correct location by so-called dot diagrams. Evaluation of this merit function requires only a ray bundle representation of the output field. The combination of a ray representation with geometrical optics propagation is called raytracing, which explains the dominance of raytracing in conventional optical engineering. Because the design of illumination systems is also often based on principles of imaging, raytracing remains a suitable technique in that case too. However, systems that perform general light transformations must apply wave-optical engineering techniques, which constitutes a generalization of conventional optical engineering. In this sense it includes the modeling of image-forming systems as a special case.

Modeling Principles

Modeling in wave-optical engineering addresses the propagation of electromagnetic fields through systems as illustrated in Figure 1. The introduction of operators and operands as mathematical tools simplifies a systematic discussion of the modeling concepts. Formally, an electromagnetic field may be expressed by the operand $\mathbf{f} \in \mathsf{F}$ with the set of fields F. The actual form of $\mathbf{f} \in \mathsf{F}$ depends on the physical approach chosen to model electromagnetic fields. For instance, $\mathbf{f} \in \mathsf{F}$ may stand for a ray bundle, a complex amplitude, or a coherence function. The photon source generates a field which acts as the input field $\mathbf{f}$ into the optical system. The quality of the output field $\mathbf{f}_{\text{out}}$ is evaluated with the help of a merit function $\mathbf{\Omega}$. This function may consist of M components which yields:

$$\mathbf{\Omega}(\mathbf{f}_{\text{out}}) = \{\Omega_1(\mathbf{f}_{\text{out}}), \ldots, \Omega_\mu(\mathbf{f}_{\text{out}}), \ldots, \Omega_M(\mathbf{f}_{\text{out}})\} \quad [1]$$

with $\Omega_\mu : \mathsf{F} \rightarrow \mathsf{R}^+$ and $\mu = 1, \ldots, M$. Each component Ω_μ defines a mathematical way to obtain one positive valued quantity for a given $\mathbf{f}_{\text{out}}$. Examples are the rms-value to evaluate the quality of a wavefront, M^2-value to judge the profile of a laser beam, center of gravity of a light distribution, light transformation efficiency, and the value of an overlap integral to determine the coupling efficiency into a fiber. Any property of the output field can be evaluated by an appropriate definition of a function Ω_μ, as long as the demanded field parameters are accessible for the field model in use. However, it should be mentioned, that it is not clear *a priori*, that all quality criteria Ω_μ can be satisfied simultaneously. How the different Ω_μ are related is dependent on the application.

The propagation of the input field through the system is mathematically summarized by the operator equation:

$$\mathbf{f}_{\text{out}} = S(p_\alpha)\mathbf{f}_{\text{in}} \quad [2]$$

with the operator $S : \mathsf{F} \rightarrow \mathsf{F}$. p_α stands for all parameters p_1, p_α, p_{A} which specify the system, for instance position of elements, refractive index, surface profiles, and layer thickness. Obviously a full electromagnetic model of S does not exist for arbitrary systems. The propagation problem eqn [2] is much too complex for a wave-optical solution. However, for a sequential approach the situation is different. If the propagation through a system is performed sequentially through the elements and the homogenous regions, at least approximate wave-optical methods are available to model the propagation S. In a sequential approach, either the transmitted or reflected field is considered per region

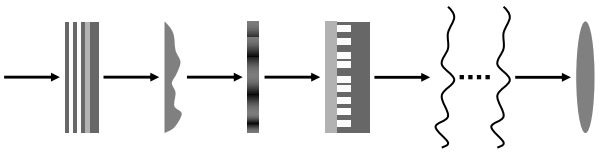


Figure 1 An optical system may include various types of elements to perform the desired optical function. Layered media, free-form surfaces, index-modulated regions, microstructured interfaces, lenses, and other structured media offer the flexibility needed to realize general light transformations. Typically structured media and homogenous regions follow each other.

of the system. Although this is an approximation, it turns out to be very useful in wave-optical engineering. With regard to Figure 2 the sequential approach is mathematically expressed by

$$\mathbf{f}_{\text{out}} = S(p_\alpha)\mathbf{f}_{\text{in}} = \mathcal{P}_J(n_J, \Delta z_J) S_{J-1}(p_{\alpha_{J-1}}) \cdots S_2(p_{\alpha_2}) \times \mathcal{P}_1(n_1, \Delta z_1)\mathbf{f}_{\text{in}} \quad [3]$$

$\mathcal{P}_j(n_j, \Delta z_j)$ expresses the propagation through a homogenous dielectric with refractive index n_j and extent Δz_j. This type of propagation is also called free propagation. Sometimes the term free propagation is restricted to propagation in vacuum; Here it is used more generally. $S_j(p_{\alpha_j})$ models the propagation through an optical component which is specified by the parameters p_{α_j}. Optical engineering needs physical models for the field operand $\mathbf{f}$, the free propagation $\mathcal{P}$, and the propagation S_j through elements. The most fundamental demand in wave-optical engineering is a wave-optical model to express the field $\mathbf{f}$ in homogenous regions. It is common to restrict to harmonic fields. The generalization to more general radiation is often possible by suitable superposition of harmonic fields. This approach assumes linear optical systems as typical in optical engineering. From Maxwell's equations it follows that harmonic fields in homogenous regions are completely specified if two independent field components are determined in one plane z_0 within the region. Therefore, the electromagnetic field operand takes the form:

$$\mathbf{f}(x, y, z_0) = [E_x(x, y, z_0), E_y(x, y, z_0)] \quad [4]$$

with the complex amplitudes E_x and E_y of the x and y components of the electric field vector. Because the knowledge of $\mathbf{f}(x, y, z_0)$ allows the direct calculation of the missing complex amplitudes E_z, H_x, H_y, H_z via Maxwell's equations, the entire electromagnetic harmonic field in a homogenous region is specified by eqn [4].

With respect to the two components of the field operand $\mathbf{f}$ propagation operators S are 2×2-matrices of the form:

$$S = \begin{Bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{Bmatrix} \quad [5]$$

In the case of a diagonal matrix, that is $S_{xy} = S_{yx} = 0$, E_x and E_y propagate independently and are considered decoupled channels of the system. If in addition $S_{xx} = S_{yy}$, both channels behave identically, and it is sufficient to model only one channel. This is often referred to as the scalar model. Of course, also in the scalar model the electromagnetic field still consists of two independent components and maintains its vectorial nature.

Numerical implementation of eqn [3] for the field operand of eqn [4] constitutes the backbone of modeling in wave-optical engineering.

Free Propagation

The propagation $\mathcal{P}$ of harmonic fields $\mathbf{f}$ through homogenous regions of a system has a rigorous electromagnetic solution in form of the spectrum of plane waves integral which yields:

$$\begin{aligned} \mathbf{f}(x,y,z \geq z_0) &= \mathcal{P}_{\text{SPW}} \mathbf{f}(x,y,z_0) \\ &= \begin{pmatrix} \mathcal{P}_{\text{SPW}} & 0 \\ 0 & \mathcal{P}_{\text{SPW}} \end{pmatrix} \begin{pmatrix} E_x(x,y,z_0) \\ E_y(x,y,z_0) \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{F}^{-1}([\mathcal{F}E_x(x,y,z_0)]e^{i(k_x x + k_y y + k_z \Delta z)}) \\ \mathcal{F}^{-1}([\mathcal{F}E_y(x,y,z_0)]e^{i(k_x x + k_y y + k_z \Delta z)}) \end{pmatrix} \end{aligned} \quad [6]$$

with the Fourier transform $\mathcal{F}$ and $\Delta z = z - z_0$. The term $\mathcal{F}E(x,y,z_0)$ is called spectrum of planewaves (SPW). The operator matrix is diagonal and $\mathcal{P}_{\text{SPW}} = \mathcal{P}_{xx} = \mathcal{P}_{yy}$, that means the x- and y-components propagate independent of each other according to the same formula in homogenous dielectrics.

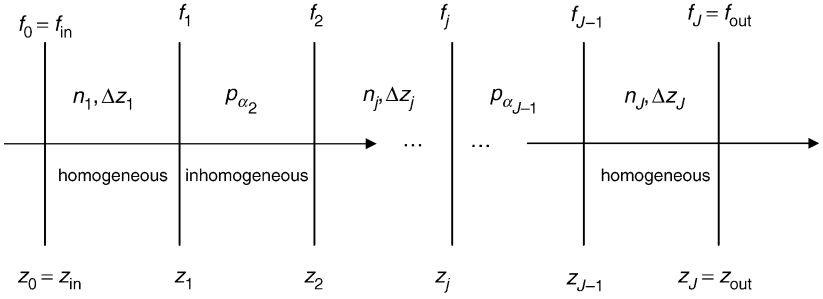


Figure 2 Formally a system may be subdivided into a sequence of homogenous dielectrics like air and inhomogeneous regions which include the elements. We indicate a region by the letter *j*.

In this sense, free propagation is a scalar problem which allows the convenient use of $U(x,y,z)$ instead of two field components for the sake of simpler mathematical expressions.

Equation [6] allows the derivation of other rigorous versions of the free propagation operator $\mathcal{P}_{\mathrm{SPW}}$, for instance Rayleigh's integral formula, which is well-known from diffraction theory. Moreover, eqn [6] leads to important approximate operators for far-field propagation and the propagation of paraxial fields. Propagation of a field into its far field is given by the operator:

$$\begin{aligned} U(r\mathbf{s}) &= \mathcal{P}_{\mathrm{far}} U(x,y,z_0) \\ &= -\mathrm{i}2\pi k s_z [\mathcal{F}U(x,y,z_0)]|_{(ks_x,ks_y)} \frac{\exp(\mathrm{i}kr)}{r} \end{aligned} \quad [7]$$

with the wave number $k = 2\pi n/\lambda$, $r = |\mathbf{r}| = (x^2+y^2+(\Delta z)^2)^{1/2}$, and the direction vector $\mathbf{s} = (s_x, s_y, s_z) = \mathbf{r}/r$. Here, and in what follows, λ denotes the vacuum wavelength. $|_{(ks_x,ks_y)}$ indicates the variables to be inserted into the Fourier transformed field. Though the far-field formula is derived for $r \to \infty$, it is also of great practical value for the propagation of fields generated by nonparaxially emitting sources like laser diodes. Roughly speaking eqn [7] predicts a spherical wave in the far field which is modulated by the spectrum of planewaves of the field in z_0. $\mathcal{P}_{\mathrm{far}}$ propagates from the plane z_0 to the spherical surface with radius r and not to a plane z.

The paraxial approximation of eqn [6] follows if the spectrum of planewaves $\mathcal{F}U(x,y,z_0)$ possesses significant values only for small k_x and k_y, that is if both electric field components propagate approximately along the z-axis. The resulting propagation operator is known as Fresnel's integral formula. In practice, its formulation with the Fourier transform $\mathcal{F}$ is typically more suitable:

$$\begin{aligned} U(x,y,z \geq z_0) &= \mathcal{P}_{\mathrm{paraxial}} U(x,y,z_0) \\ &= \frac{k \exp[\mathrm{i}k\Delta z]}{\mathrm{i}2\pi\Delta z} \exp\left[\frac{\mathrm{i}k(x^2+y^2)}{2\Delta z}\right] \\ &\quad \times \left| \left\{ \mathcal{F}\left(\exp\left[\frac{\mathrm{i}k(x^2+y^2)}{\Delta z}\right] U(x,y,z_0) \right) \right\} \right|_{(x,y)n/\lambda\Delta z} \end{aligned} \quad [8]$$

$|_{(x,y)n/\lambda\Delta z}$ indicates the variables to be inserted after Fourier transform. The far-field approximation of eqn [8] leads to the well-known Fraunhofer diffraction formula.

Equations [6]–[8] provide fundamental operators for the propagation of fields in wave-optical engineering. All integrals are based on Fourier transforms which allows the application of the fast Fourier transform (FFT) in practice. Dependent on the propagation technique, one or two FFTs must be performed. In practice, the sampling of fields is of great concern. Naturally the amount of data should be as small as possible. According to the rigorous propagation operator eqn [6] the (x,y)-coordinate system and therefore the sampling period in z_0 and z are identical, that is, scaling does not occur. As illustrated in Figure 3, that can lead to an enormous amount of data for large propagation distances. For demonstration purpose the field amplitude depicted at the top is chosen. The field window has the size 1 mm × 4 mm. The wavelength λ is 632.8 nm, $n = 1$, and the sampling distance 10 μm. That results in 100 × 400 sampling points. Propagation of this field with the spectrum of planewave operator maintains the size of the field window. The resulting field for a propagation distance $\Delta z = 10$ mm is shown in the middle. Because of diffraction, the field widens. As long as the initial window is large enough to encompass the propagated field, the propagation method works well. At the bottom, the resulting field for $\Delta z = 200$ mm is shown. The field exceeds the size of the window and that leads to numerical errors (aliasing). This can be avoided by embedding the initial field into a window which is large enough with respect to the expected size of the propagated field. As a result, one obtains the

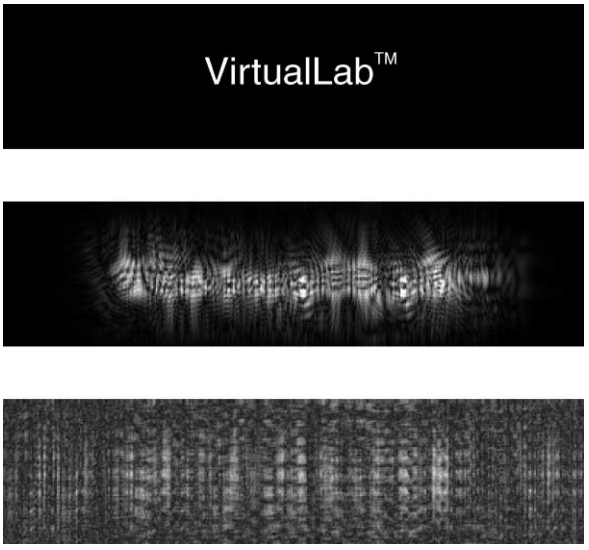


Figure 3 Illustration of numerical properties of the spectrum of plane wave propagation technique. The field amplitude depicted at the top is numerically propagated in free space and the distributions in the middle and the bottom are obtained. The simulations in this and all following figures were performed. The simulations in this and all following figures were performed with the wave-optical engineering software VirtualLab™ from LightTrans GmbH.

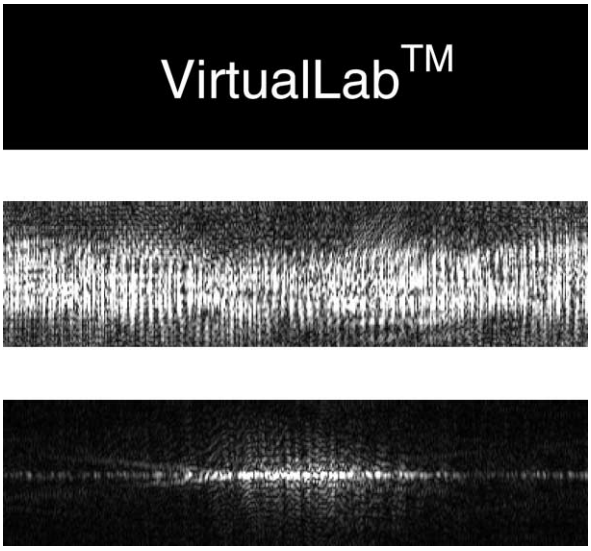


Figure 4 Starting from the same input field and the same distances as in Figure 3 the propagated fields depicted in the middle and the bottom are obtained using the paraxial propagation operator. Because of the inherent scaling the field window sizes are 632.8 μm × 632.8 μm and 12.656 mm × 12.656 mm, respectively. The number of sampling points remains unchanged. Obviously, the field in the middle is too small to encompass the propagated field and numerical errors occur. For the long distance the field size is perfectly matched to the field size obtained by propagation.

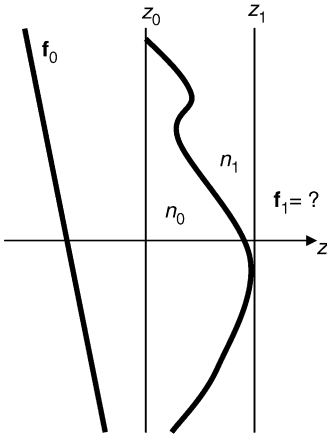


Figure 5 A planewave $\mathbf{f}_0$ propagates through an interface between two homogeneous dielectrics with refractive index n_0 and n_1. The interface is of general shape and therefore a rigorous solution of the propagation problem does not exist. The virtual planes in z_0 and z_1 encompass the interface in a way which minimizes the distance $z_1 - z_0$.

demanded accuracy at cost of increasing data memory requirements and computation time. Thus in practice the SPW propagation operator is particularly suitable for short distances. Fortunately the paraxial propagation operator eqn [6] possesses a complementary behavior. Because of $(x, y)n/\lambda\Delta z$, the (x, y)-coordinate system changes its scaling with the distance. As illustrated in Figure 4, that makes it well suited for large propagation distances but problems occur for small ones. Therefore, all paraxial propagation problems have an efficient numerical solution by a suitable combination of $\mathcal{P}_{\text{paraxial}}$ and $\mathcal{P}_{\text{SPW}}$. In the nonparaxial case the far-field operator eqn [7] often helps to realize a numerically efficient propagation for large distances. However, there are sometimes situations, in which free propagation of nonparaxial fields causes significant numerical problems.

Propagation Through Elements

The propagation of harmonic fields $\mathbf{f}$ through elements is a challenging task. For example, Figure 5 illustrates propagation through smooth or microstructured interfaces between two homogenous media. The field $\mathbf{f}_1(x, y, z_1)$ should be determined if $\mathbf{f}_0(x, y, z_0)$, the shape of the interface, and the refractive indices n_0 and n_1 are known. The reflection mode of operation can be formulated analogously. It is remarkable that this basic problem possesses a rigorous solution only for a few special interface geometries. Of particular importance are the rigorous solutions for plane and periodic interfaces. In the first case, the solution comprizes Snell's law of refraction, the law of reflection, and Fresnel's formulae. Grating theory provides the solution for periodic interfaces. It is possible to use these methods also for nonperiodic interfaces but at the cost of immense computational effort. Thus, they are only practicable for structures with detail no larger than a few wavelengths. For general interfaces, accurate approximate methods are required. One approach is to decompose a general propagation problem into a set of simpler problems that can be solved rigorously. The results are then combined to obtain an approximate solution of the original problem. Two basic concepts for subdividing the general problem are: (i) Decomposition of the incident field into laterally truncated fragments which illuminate elementary fractions of the interface. (ii) Decomposition of the response of the interaction of the incident field with the interface into approximately independent ones related to local interface features. The first technique is referred to as local elementary interface approximation (LEIA) and the second one as local independent response approximation (LIRA). Elementary interfaces of particular importance are plane, periodic, and spherical ones.

The local plane interface approximation (LPIA) is of particular importance in wave-optical engineering. Figure 6 illustrates the basic concept for the example

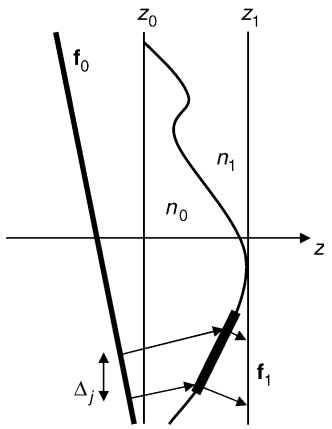


Figure 6 If a lateral fraction of the input field is propagated towards the interface, it is often possible to laterally restrict the major part of the propagated field to a local plane interface. Because of diffraction it is never possible to do that rigorously. Neglecting diffraction in the region between z_0 and z_1 provides a simple model to propagate a section of the incident plane field through the interface. Fresnel's formulae and Snell's law of refraction can then be applied directly to obtain the transmitted plane field, which is propagated to the plane z_1.

of a plane incident field. It has been proven that the approximation of geometrical optics propagation within the regions of the component, here that means between z_0 and z_1, provides results which are surprisingly accurate in numerous situations of practical interest. A full wave-optical version of LPIA is currently under research and development. In general, however, LPIA is applicable if the characteristic lateral extent of the local modulation of the interface is not close to the wavelength. It should be emphasized that LPIA is, as most other approximate propagation techniques, a 2×2 matrix, like eqn [5], with nonvanishing components. That means, for instance, that in general, an E_y-component occurs in z_1 even if $E_y = 0$ in z_0. That is, the two field channels experience crosstalk. A further approximation in geometrical optics LPIA, popular in Fourier optics holography, and diffractive optics is the thin element approximation (TEA). Geometrical optics LPIA leads to TEA if refraction and the local dependence of Fresnel losses at the interface are neglected. Obviously, that is allowed only if all angles involved, including the slope of the interface and the direction of the input field, are small. Therefore, TEA is a paraxial and geometrical optics approximation. Moreover, small local slopes of the interface lead to a restriction to thin elements, which explains the name TEA. It is also referred to as scalar approximation, because the operator matrix becomes diagonal with $S_{\mathrm{TEA}} = S_{xx} = S_{yy}$, that means the two field channels propagate independently according the same formula through the interface. In the thin element only approximation the dependence of the optical path through the element on position (x, y) is considered and mathematically takes the form:

$$\begin{aligned} U(x, y, z_1) &= S_{\mathrm{TEA}}\, U(x, y, z_0) \\ &= c_{\mathrm{Fresnel}}\, U(x, y, z_0) \exp\left[\frac{\mathrm{i}2\pi n_1(z_1 - z_0)}{\lambda}\right] \\ &\quad \times \exp\left[\frac{\mathrm{i}2\pi h(x, y)(n_0 - n_1)}{\lambda}\right] \end{aligned} \qquad [9]$$

with the vacuum wavelength λ. The profile height $h(x, y)$ is expressed as the distance from z_0 to the interface. Propagation through an interface between dielectrics modeled with TEA affects the phase of the incident field proportional to the profile height. The magnitude remains unchanged besides a constant factor c_{Fresnel}, due to Fresnel's equations. If complex valued refractive indices are taken into consideration, absorption results in an amplitude modulation completely analogous to the phase modulation described in eqn [9]. This way it is also possible to address the propagation through apertures and the combination of S_{TEA} with $\mathcal{P}$ leads to the well-known diffraction integrals. In fact, Kirchhoff's boundary conditions of field propagation through a hole in a completely opaque screen can be interpreted as TEA for complex valued refractive indices.

The simple form of the TEA operator allows a straightforward numerical implementation. The field operand and the profile are equidistantly sampled and the resulting field is sampled analogously. That makes the combination of S_{TEA} with numerical implementations of the free propagation operators $\mathcal{P}$ easy. As soon as refraction is considered in geometrical optics LPIA, the equidistant sampling of the output field is destroyed which enforces additional interpolation effort before a free propagation by fast Fourier transforms can be implemented, because fast numerical Fourier transform algorithms require an equidistant sampling.

In wave-optical engineering the combination of geometrical optics LPIA with a wave-optical field model is mandatory. In this context it is instructive to mention, that its combination with a ray-bundle representation of the incident electromagnetic field leads to raytracing through the interface, that is it reduces to the basic technique in conventional optical engineering.

Similar to LPIA it is possible to consider linear gratings as elementary local interface and the local linear grating approximation (LLGA) results. It is of special concern for all types of diffractive elements, for which the interface can locally considered

periodically. An important example is given for diffractive lenses in the outer zones of high numerical aperture lenses. In the center and for low numerical aperture TEA is appropriate for propagating through diffractive lenses.

Design Principles

The design problem in wave-optical engineering can be stated as follows: specify system parameters p_α in a way that ensures for a given input field $\mathbf{f}_{\text{in}}$ an output field $\mathbf{f}_{\text{out}}$ which satisfies the quality criteria defined by the merit functions Ω_μ. In addition, one may demand a system which is most simple, particularly cheap, or various other system properties of concern in practice.

Design constitutes an optimization problem. Restrictions to the system parameters and the merit function sets design constraints as well as design freedoms.

System parameters can be restricted, for instance, by the materials available, the maximum size of a system, and types of element profiles. Within these limits the system parameters are chosen freely during optimization to achieve the optical function.

The merit functions determines which parameters of the output field are important and should be emphasized, for instance the intensity distribution. All output fields, whose parameters satisfy the constraints satisfy the design problem. In most applications, not all field parameters are constrained. For instance, in laser materials processing or illumination applications, the merit function is related to energy quantities and then the phase of the output field can be used as a free parameter distribution. Moreover, the output field is typically not constrained in the entire output plane, but only in so-called signal regions, which introduces freedom of amplitude outside the signal regions. Hence, dependent on the application, there exist field parameters that can be chosen partly or completely free during design.

Successful design strategies use the freedoms to improve the matrix function. It is often not clear a priori, if a solution does exist. Then the field constraints are satisfied as good as possible in terms of the merit function.

In imaging systems, typical system parameters include, for instance, distances, refractive indices, curvature of lenses, and parameters to specify aspherical surfaces. For high-end lens systems, the resulting number of parameters can be large, but they are still small enough to utilize parameter optimization algorithms if the designer starts with a reasonable initial guess. In wave-optical engineering, the number of system parameters can become huge due to the generalized nature of the structures, which typically prevents the use of system parameter optimization techniques. For example, the profile of diffractive elements is typically specified by 10^4–10^6 parameters. Thus a direct optimization of the system parameters, which requires the evaluation of $S(p_\alpha)$ (see eqn [3]) after any change of system parameters p_α, is not realistic.

Instead of optimizing the system parameters directly, which is referred to as design in the structural embodiment of the system, a technique wellknown from Fourier optics and diffractive optics can be generalized for wave-optical engineering. It does not start with structure related element parameters but replaces all or at least some of the operators $S_j(p_{\alpha_j})$ of eqn [3] by mathematical operators $\mathcal{T}_j(t_{\tau_j})$. t_τ denotes the parameters of $\mathcal{T}$. The propagation operator through a system expressed in this functional embodiment has the form:

$$\begin{aligned}\mathbf{f}_{\text{out}} &= S_{\text{functional}}\mathbf{f}_{\text{in}} \\ &= \mathcal{P}_J(n_J, \Delta z_J)\mathcal{T}_{J-1}(t_{\tau_{J-1}})\cdots\mathcal{T}_2(t_{\tau_2}) \\ &\quad \times \mathcal{P}_1(n_1, \Delta z_1)\mathbf{f}_{\text{in}}\end{aligned} \qquad [10]$$

if all propagations through elements are expressed in functional form. A very important example of a functional operator $\mathcal{T}$ is given by a simple multiplication of both field components with identical or different complex functions $t(x, y)$, that is:

$$\mathcal{T}_t(t_\tau)\mathbf{f}_{\text{in}} = \begin{pmatrix} t_{xx}(x, y) & 0 \\ 0 & t_{yy}(x, y) \end{pmatrix}\mathbf{f}_{\text{in}}(x, y) \qquad [11]$$

The values of both $t(x, y)$ form the parameters t_τ of $\mathcal{T}$. The functions $t(x, y)$ are also called ideal transmission functions of elements. Of special interest are phase-only transmission functions $t(x, y) = \exp[i\tau(x, y)]$ (see discussion below). If $t(x, y) = t_{xx}(x, y) = t_{yy}(x, y)$, the design is simplified. Fortunately, in most actual applications this assumption is valid.

Designing in the functional embodiment requires the evaluation of $S_{\text{functional}}$ instead of S of eqn [3], which reduces the numerical complexity enormously. That improves the likelihood for developing efficient design algorithms to optimize the distances Δz_j, the refractive indices n_j, and parameters t_{τ_j}.

The design in the functional embodiment does not deliver the structure parameters of elements but only parameters t_τ of the operator $\mathcal{T}$ which gives information about how a field should be changed at the position z where $\mathcal{T}$ takes effect. Thus, in a next design step, elements must be obtained which cause the required effect. This so-called structure design

step is also generally challenging because it requires the solution of an inverse propagation problem through interfaces. For a further discussion the most important $\mathcal{T}$, that is the transmission type operator of eqn [11], is considered. For the sake of simplicity $t(x, y) = t_{xx}(x, y) = t_{yy}(x, y)$ is assumed. Then, a solution which uses geometrical optics LPIA in combination with phase-only transmissions is known in the form of a recursive structure design strategy. In its paraxial approximation, that is TEA, the solution becomes very simple. For a transmission $t(x, y) = \exp[i\tau(x, y)]$, obtained from a functional design, an element with a surface profile:

$$h(x, y) = \frac{\lambda\tau(x, y)}{2\pi(n_{\text{element}} - n_{\text{free}})} \qquad [12]$$

results from eqn [9]. The refractive index of the element is denoted by n_{element} and the homogenous region behind the element has the index n_{free}, eqn [12] constitutes a direct relationship between the phase values of the transmission function and the profile $h(x, y)$. This structure design technique is accurate for all paraxial situations. However, it is also a reasonable method for nonparaxial structure design since it provides an initial profile for further optimization. Obviously it is applicable for phase-only transmission functions only. That underlines the importance of this type of functional operator for the design in the functional embodiment. Often the fabrication of profiles is done with lithographic techniques which result in surface reliefs with a few discrete height levels, that means the profile $h(x, y)$ is quantized. According to eqn [12] a quantization of $h(x, y)$ directly leads to quantized phase values $\tau(x, y)$. This quantization must be taken into account when designing $\tau(x, y)$. An example should serve to discuss further aspects of structure design according to eqn [9]. A fundamental transmission in various applications is the phase of a spherical wave and its paraxial version, the quadratic phase. An example of a spherical phase $\tau(x, y)$ is shown in Figure 7 where the 2π-modularity of the exponential function is taken into account – as normal for phase data which has been obtained by numerical calculations. Inserting this phase into eqn [9] leads to the profile of a diffractive lens. Because of the use of inverse TEA for structure design it works as demanded for moderate numerical apertures and starts to suffer from aberrations for large numerical apertures because of electromagnetic effects in the very outer zones. It is also possible to unwrap the phase before inserting it into eqn [9]. Then a smooth surface, that is a refractive spherical lens, results. It is well-known that a spherical lens works accurately in the paraxial domain and suffers from aberrations for moderate and large numerical apertures, which are more severe than for the diffractive lens. The use of inverse geometrical optics LPIA as a structure design technique is not restricted to the paraxial domain and therefore it leads to a suitable aspherical surface in order to ensure a spherical phase transmission for larger numerical apertures too. This is illustrated in Figure 8. The lens example shows another advantage of the two-step design method, which consists of the design in the functional embodiment prior to the one in the structural form of the element. In the functional

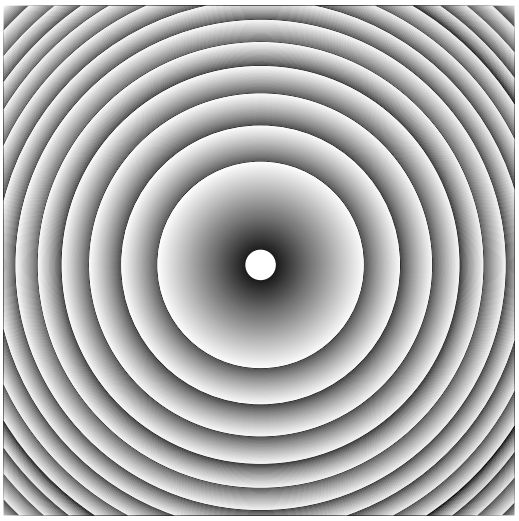


Figure 7 The phase of a spherical phase in the 2π-modulo representation. This phase can be unwrapped which leads to a smooth spherical phase which varies from 0 to various 2π. The noisy edges are due to the use of a cartesian coordinate system.

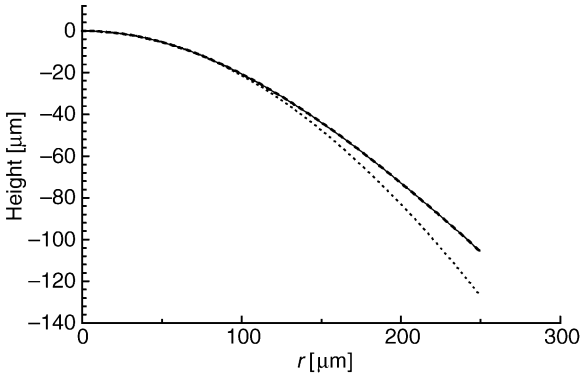


Figure 8 Application of the thin lens approximation for structure design leads to a spherical surface for a spherical phase-only transmission. The corresponding surface is depicted by the dotted line. If geometrical optics LPIA is used for structure design, an aspherical surface is obtained as required. The aspherical surface is depicted by the solid line. The height of the surfaces are shown as function of the radius *r*. Naturally the difference between TEA and geometrical optics LPIA occurs for larger *r* only.

form the transmission is obtained. It is not necessary to decide about the use of a diffractive or a refractive element in an early stage of the design. Moreover, it provides the designer with insight into selecting optical elements to achieve the function. If the phase-only transmission can be unwrapped, a refractive as well as a diffractive element can be used. An optical engineer has the freedom to decide about the suitable choice taking fabrication, size, costs, and other additional constraints into account. However, as soon as the phase-only transmission includes so-called vortices, unwrapping is not possible and diffractive optics is mandatory. An example of a vortex is shown in Figure 9. It turns out that vortices are more likely to appear in wave-optical engineering the less image-like the optical function is (see also last section). It is important to note, that functional design assumes a diagonal operator matrix, that is, it excludes crosstalk of the two field channels. Transferring the system into a structural form may violate this assumption because of vectorial effects in the nonparaxial domain. For instance, geometrical optics LPIA in general does not correspond to a diagonal operator. Therefore, it is necessary to analyze the final system for $\mathbf{f}_{\text{in}}$ – even in the case of linear polarized light – using more accurate propagation techniques than S_{TEA} if it is not ensured that the paraxial approximation is accurate enough.

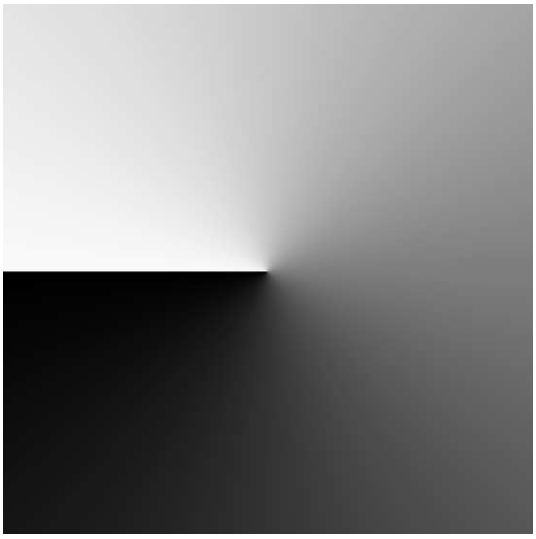


Figure 9 Illustration of a single vortex of charge one. The phase varies from 0 to 2π. Because of the vortex, this phase is not unwrappable, that means a smooth phase function cannot be obtained. The transition in the phase function requires a microstructure in the profile of the element, which generates this phase-only transmission.

The so-called thin lens, or ideal lens, in conventional optical engineering constitutes a special case of the two-step design strategy. In the first design step, distances and lens focal lengths are determined, without taking into account the actual shape of a lens. In the second step lenses, or even lens systems, are inserted according to the position and the focal length specified in the first step. Thus, wave-optical engineering extends our understanding by broadening established methods in optical engineering.

Fundamental Systems

A smart approach to design minimizes the number of elements required to achieve a desired light transformation. Thus, it is logical to try a functional design with a system of the form:

$$\begin{aligned} U_{\text{out}}(x,y) &= S_{\text{functional}}\, U_{\text{in}}(x,y) \\ &= \mathcal{P}(n_{\text{free}}, \Delta z_{\text{out}}) \exp[\mathrm{i}\tau(x,y)] \\ &\quad \times \mathcal{P}(n_{\text{free}}, \Delta z_{\text{in}}) U_{\text{in}}(x,y) \end{aligned} \qquad [13]$$

which is illustrated in Figure 10. One phase-only transmission, properly placed between the input and the output plane, should realize the desired transformation of the function. Scalar terminology is used, because both components are decoupled in the functional embodiment.

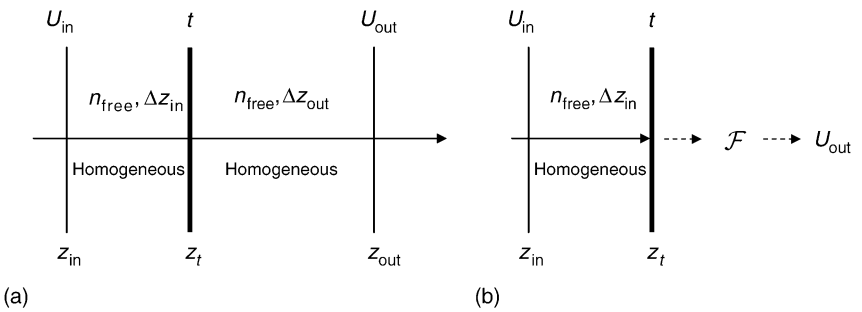


Figure 10 A basic system for light transformation can be described functionally by two propagations through homogeneous dielectrics with refractive index n_{free} and one phase-only transmission function. (b) If the second propagation is replaced by an operation in functional form, here the Fourier transform, the setup shown on the right side results.

If the output field is completely specified by the merit function, for instance: if the complex amplitude of a laser beam should be transformed into another one, the so-called amplitude matching strategy is a very powerful technique to solve the design problem stated by eqn [13]. In amplitude matching one determines the Distances Δz_{in} and Δz_{out} such that the inverse propagated U_{out} and the forward propagated U_{in} possess identical magnitudes. If such distances exist, the transformation can be realized using a phase-only transmission in the plane z_t. The phase simply follows from the ratio of the propagated fields. If no distances exist where amplitude matching occurs, the design problem has no solution and at least one more transmission function must be introduced.

As mentioned above, in numerous applications the output field is not specified completely. In particular, phase is often a free parameter. Then, an even simpler system approach than the one suggested by eqn [13], often leads to reasonable solution. Instead of $\mathcal{P}(n_{\text{free}}, \Delta z_{\text{out}})$, a Fourier transform is assumed, then if,

$$\begin{aligned} U_{\text{out}}(k_x, k_y) &= S_{\text{functional}} U_{\text{in}}(x, y) \\ &= \mathcal{F}\{\exp[\mathrm{i}\tau(x, y)]\mathcal{P}(n_{\text{free}}, \Delta z_{\text{in}}) U_{\text{in}}(x, y)\} \end{aligned} \qquad [14]$$

The output field is expressed in k-coordinates as long as the system is discussed in functional form. For a functional design the distance Δz_{in} is of no concern, however it must be chosen together with the structure design as discussed below. Thus, in functional design, one concentrates only on the design of the transmission function. The iterative Fourier transform algorithm (IFTA), an extremely powerful algorithm, has been developed to accomplish this design. Its principle is illustrated in Figure 11. The algorithm sequentially applies constraints on the transmission and in the Fourier plane. Properly controlled, the algorithm converges to a Fourier pair for which the transmission plane constraint is satisfied completely and the Fourier domain constraint is satisfied as accurately as required by the merit function. The initial distribution is of enormous importance for a successful transmission design. Typically the initial distribution is defined in the Fourier plane as shown in Figure 11. A field $U_{\text{sig}}(k_x, k_y)$ is chosen, which satisfies the merit function. It is called the signal field. Because of design freedoms such as phase, typically there exists a set of signal fields F_{sig}. If phase freedom is available, any field with the desired amplitude and an arbitrary phase is an element of F_{sig} and can be used as initial distribution of the IFTA. Dependent on the application, the choice of the phase is of fundamental concern. In the following we assume

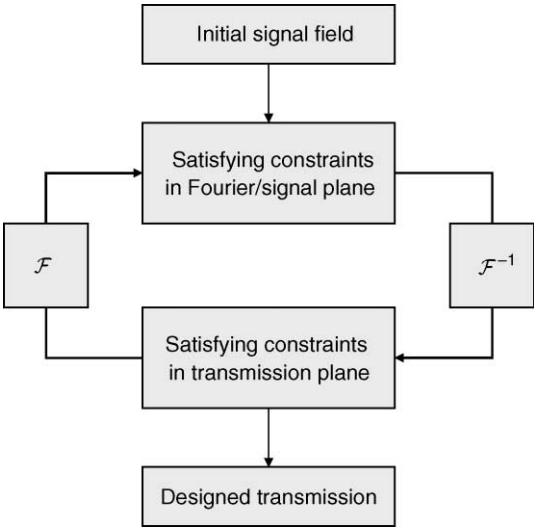


Figure 11 Illustration of the basic principle of the iterative Fourier transform algorithm (IFTA). By a sequential satisfaction of constraints in the signal and the transmission plane, the design in the functional embodiment can be accomplished. If both constraints are satisfied accurate enough, the IFTA is terminated with a final satisfaction of the transmission plane constraints. The depicted IFTA version can be generalized by replacing $\mathcal{F}$ by $\mathcal{P}$ or any S.

phase freedom is available, but IFTA is also useful if this is not the case.

It is helpful to distinguish between two types of signal fields. If amplitudes for only a finite number of discrete k-values, (i.e., angles or directions), need to be determined, the transmission acts as a diffractive beamsplitter. In all other cases more general light transformations are concerned.

The constraints in the transmission plane follow from the structure design. In most cases the structure design is done by TEA according to eqn [12]. Then, phase-only transmissions are required. Typically the fabrication of the corresponding element profile demands a phase quantization in addition. Both constraints have been intensively investigated for the IFTA.

As mentioned above, practically signal fields are always restricted to one or more signal regions of finite extent. For example, a detector has finite extent. Outside of these regions the signal may be unrestricted or only partially restricted. This freedom in amplitude outside the signal regions is fundamental for a successful design of phase-only transmissions. It has been shown, that the only phase-only function whose Fourier transform exists over a finite area is a linear phase. The transform of all other phase-only functions contribute energy outside of the signal regions. In other words, the use of amplitude freedom is mandatory. Energy losses due to stray light outside

the signal region cannot be avoided. As a result, it is not possible to generate the desired response with 100% efficiency. For any given signal field U_{sig} an upper bound $\eta_{bound}(U_{sig})$ of the achievable efficiency $\eta(U_{sig})$ of light transformation with a phase-only transmission has been derived, that is $\eta(U_{sig}) \leq \eta_{bound}(U_{sig})$. The equal sign is only valid for the special case of a linear phase transmission, for which the signal field consists of one k-direction only. Then $\eta_{bound}(U_{sig}) = \eta(U_{sig}) = 1$ is obtained. After structure design for a linear phase-only transmission with TEA, a so-called blazed grating results. If TEA is a sufficient approximation one concludes that a blazed grating has a theoretical efficiency of 100%. For all other signals the theoretical efficiency is smaller than 100%. Moreover, the upper bound is a function of U_{sig} and thus also of its phase. Therefore, in a first step of a transmission design it is recommendable to perform a phase synthesis which maximizes the upper bound. This first design step delivers information about achieveable efficiency and a suitable initial signal field for the IFTA.

If fabrication requires phase quantization, amplitude freedom can be used to separate quantization noise from the signal regions. Unfortunately this reduces the achievable efficiency even more. If the phase to be quantized is homogenously distributed, the upper bound for a phase-only transmission quantized to Q equidistant phase levels is given by:

$$\eta_{bound}(U_{sig}, Q) = \mathrm{sinc}^2\left[\frac{1}{Q}\right]\eta_{bound}(U_{sig}) \qquad [15]$$

For the special case of linear phase, the well-known formula $\eta(Q) = \mathrm{sinc}^2[1/Q]$ follows. The values of $\mathrm{sinc}^2[1/Q]$ and therefore of quantized blazed gratings in the paraxial approximation are 0.405, 0.812, 0.950, 0.987, and 0.997 for $Q = 2$, 4, 8, 16, and 32, respectively. This result does not mean that, for example, a diffractive beamsplitter with a splitting ratio $1 : M_{beams}$ and 16 height levels will provide ~99% efficiency. According to eqn [15], the theoretical values are smaller and depend on M_{beams} because $\eta_{bound}(U_{sig})$ depends via U_{sig} on M_{beams}. If fabrication processes and the inherent imperfections are taken into consideration, the values decrease even more. As a rule of thumb, it is possible to claim that efficiencies larger than 90% are always challenging to obtain. With sophisticated design and fabrication techniques, efficiencies larger than 80% are reasonable.

Up to now, we have discussed only the importance of the signal field phase to efficiency. But the phase of the initial signal field has another extremely important effect on the result of the transmission design. From mapping-type considerations, a smooth initial signal field phase can be constructed in order to realize a light transformation. An example is illustrated in Figure 12. It shows the central part of the phase of the transmission function which transforms a Gaussian input beam into a flat-top line beam. Because of the smooth signal phase also the transmission phase is smooth and therefore unwrappable. This approach is typical for a so-called beam shaping design. It requires a well-defined input beam and high accuracy in adjustment of the beam shaping optics to obtain the output field with good quality. If, instead of a smooth initial phase, a random one is applied, the situation illustrated in Figure 13 (left) is achieved. Here the phase of the transmission looks random, which is typical for beamsplitter and diffuser. Because such phase functions possess numerous vortices, they do not allow an unwrapping. Thus, diffractive elements are the logical result of structure design. An example is shown in Figure 13 (right).

Figure 14 shows how light from a laser diode passes through the diffuser element of Figure 13 (right). The radiation is transformed into an intensity pattern that depicts the 'skyline' of Jena. Although globally the intensity possesses the requested shape, a closer look shows that the intensity distribution is speckled. However, in numerous applications this is not a problem so long as the size of the speckles is well adapted to the resolution of the detector.

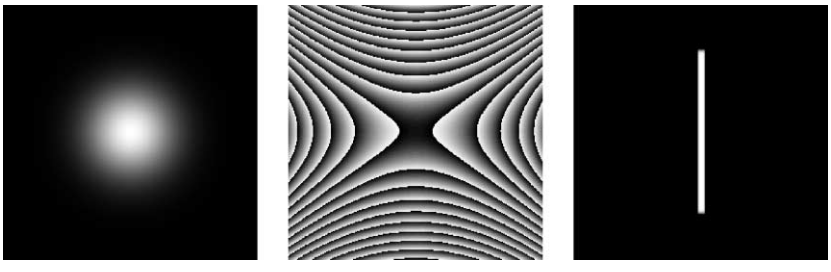


Figure 12 The input beam on the left propagates through the transmission function the central part of the phase of which is shown in the middle and in the signal plane the amplitude shown on the right results. This Gaussian-to-line transformation is a typical beam shaping design task.

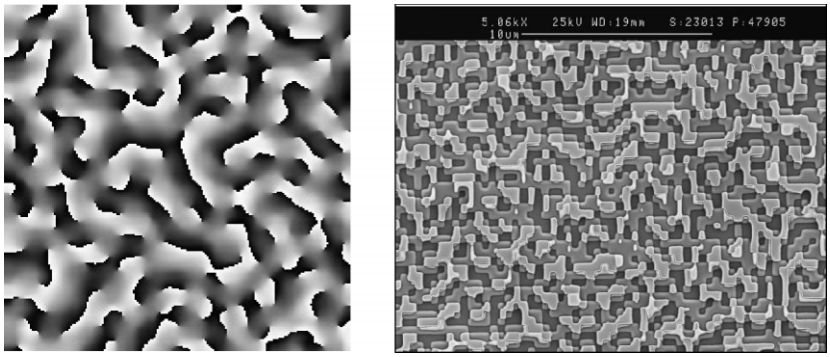


Figure 13 A typical phase distribution of a transmission (left) which results from design with IFTA for a random initial signal field phase. Numerous vortexes prevent its unwrapping. Therefore, structure design leads to diffractive elements. Then, a quantization of the phase is demanded in most cases. An example is shown on the right-handed side. It depicts a portion of a four level diffuser which results in the output field shown in **Figure 14**. The pixel size is 400 nm and the corresponding element was designed and fabricated in a cooperation of LightTrans GmbH and the Institute of Applied Physics at the University of Jena.

This technique is called diffuser-type light transformation. The appearance of the global intensity distribution is robust against variations in the input beam and thus is well-suited also for partial coherent illumination, for example radiation from an excimer laser and LED. In that case, even the appearance of speckles is reduced or vanishes completely. The mapping-type and the diffuser-type approach to choose the initial signal phase constitute two extreme cases of a great variety to control basic aspects of light transformation using different initial phases.

After choosing a suitable initial signal field $U_{\text{sig}}(k_x, k_y)$ the IFTA typically performs well and a transmission function is obtained, which leads to $|U_{\text{out}}(k_x, k_y)|^2 \propto |U_{\text{sig}}(k_x, k_y)|^2$ in the signal region with minimum error. In the next step the structural design must be performed. The structure of the element is calculated via eqn [12].

Besides designing structure to produce the desired transmission function one needs to realize optically the Fourier transformation. Placing the observation plane in the far field behind the element is the most straightforward way to achieve this. However, field Fourier transformation can also be realized by lenses as shown in **Figure 15**. From simple geometric considerations using paraxial approximation one obtains:

$$\alpha = \frac{2\pi n_{\text{free}}}{\lambda f_{\text{e}}} \qquad [16]$$

Each system produces a Fourier transform that is scaled specially by α, where the output field is considered in a dielectric with refractive index n_{free}. The effective focal length f_{e} is equal to the focal length of the lens if the element is placed in front of the lens (see top of **Figure 15**), the distance of the element to the lens is of no concern for the resulting intensity,

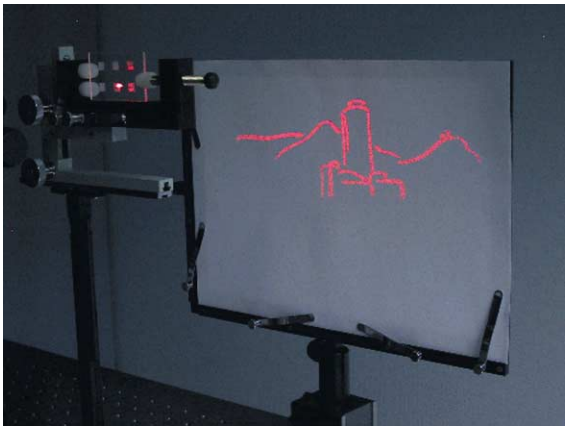


Figure 14 Optical demonstration of the use of a diffuser (see in Figure (right)) transforming a laser diode beam into an intensity displaying the 'skyline' of Jena. No additional optical element is needed! Stray light appears but is laterally separated from the desired signal and thus it is easy to be filtered.

however it is recommended to minimize the distance between element and lens for elements with deflection angles close to the paraxial limit. That avoids light losses due to truncation by the lens aperture. If the focal length itself is chosen as the distance of the element to the lens, a so-called $2f$-setup results, which generate, in addition to the correct magnitude, the correct phase of the Fourier transformation in the focal plane. If an alternative distance to the lens is used, an additional quadratic phase occurs in the focal plane. If the element is placed behind the lens (see middle row of **Figure 15**), f_{e} is the distance between element and focal plane. This setup provides some freedom in adjusting the output scaling. On the other hand, it requires an accurate adjustment of position to obtain a specific scaling. This is not required for the first setup. In order to obtain a specific

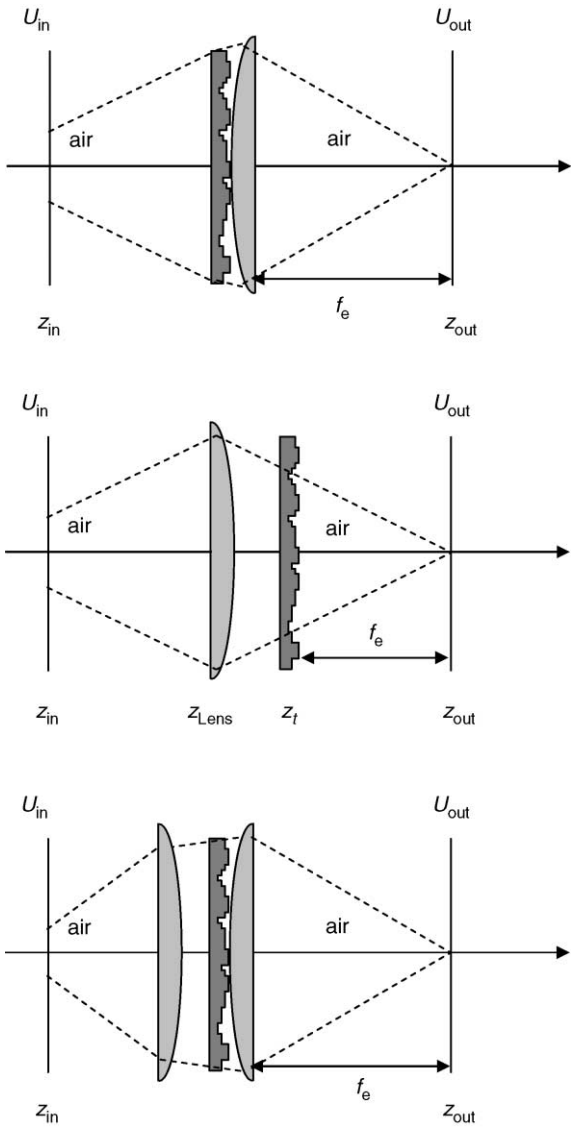


Figure 15 Three setups to realize a Fourier transformation for a beam shaping or diffuser/splitter element, which deflects the light only in a paraxial way. In all cases the output is obtained in the focal plane of the lens.

scaling of the output field $U_{out}(\alpha x, \alpha y)$, the effective focal length and the size of $U_{sig}(k_x, k_y)$ in the k-space must be chosen properly according to eqn [16]. In addition to the size of the output field, its resolution is also of fundamental importance. According to eqn [14] and the convolution theorem the output field is given by

$$U_{out}(k_x, k_y) = \mathcal{F}\ \exp[i\tau(x, y)] * \mathcal{F}\mathcal{P}(\Delta z_{in}) U_{in}(x, y) \quad [17]$$

where $*$ denotes the convolution. For a constant transmission function one would obtain the focus of the input beam. In functional embodiment that is $\mathcal{F}\mathcal{P}(\Delta z_{in}) U_{in}(x, y)$ only. In the structural form, the focus of the beam must be calculated by propagating the input field through the system without the transmission function. The Fourier transformed transmission function, that is $\mathcal{F}\ \exp[i\tau(x, y)]$, is convolved with this beam focus in order to obtain the resulting output field. In beamsplitting that means a replication of the beam. In general the output field can be understood as being composed of numerous replicated beam foci which are overlapping. That makes plausible that the size of the focus of the input beam limits the resolution of the whole output field. The size of the focus is determined by the basic beam focusing lens system and not by the diffractive or refractive element. In conclusion, a highly resolved output field requires a focusing optic with high numerical aperture. For examples, it is not possible to obtain a flat-top beam profile from a Gaussian beam where focus has the same size as the flat-top should have. The steeper the sidewalls of the flat-top profile, the smaller must be the focal spot without the beam shaping element. The choice of a suitable numerical aperture of the lens system must be done before the transmission design is performed because it also determines the extent of the transmission function. The distance Δz_{in} must be chosen in such a way that ensures that the propagated input field illuminates the whole aperture of the system.

If a high numerical aperture on the output side is required to achieve a high resolution, the setup in the middle of Figure 15 should be avoided and the first setup is preferable. Otherwise the element would be illuminated by a highly nonparaxial field which violates the structure design with TEA. With the same argument, but now on the input side, an input field with high divergence should be at least partly collimated to achieve a paraxial illumination of the element. Then, the setup at the bottom of Figure 15 results.

For nonparaxial light transformation, that means the output field is large and hence the transformation requires nonparaxial deflection of the input field; the described transmission and structure design is still useful to obtain an initial design for a further optimization by wave-optical modeling techniques.

Though a Fourier approach to obtain general light transformations seems rather restricted, it covers a broad field of applications. The combination of suitable beam optics with the refractive free-form or diffractive element can be very simple (see optical setup of Figure 14) or very complex. That mainly depends on the resolution and size of the demanded output field as it is well known from imaging systems.

Each design should be finalized with a complete wave-optical analysis, typically including adjustment and fabrication error tolerancing, to achieve a full

understanding of the optical properties of the system. After this final investigation it is ready for fabrication and application providing a further innovation through optics and photonics on the base of wave-optical engineering.

See also

Fourier Optics.

Further Reading

Aagedal H, Schmid M, Beth T, Teiwes S and Wyrowski F (1996) Theory of speckles in diffractive optics and its application to beam shaping. *Journal of Modern Optics* 43(7): 1409–1421.

Aagedal H, Schmid M, Egner S, Müller-Quade J, Beth T and Wyrowski F (1997) Analytical beam shaping with application to laser-diode arrays. *Journal of the Optical Society of America A* 14(7): 1549–1553.

Aagedal H, Wyrowski F and Schmid M (1997) Paraxial beamsplitting and shaping. In: Turunen J and Wyrowski F (eds) *Diffractive Optics for Industrial and Commercial Applications*, chapter 6, pp. 165–188. Berlin: Akademie Verlag.

Born M and Wolf E (1980) *Principles of Optics*, 6th edn. New York: Pergamon Press.

Bryngdahl O (1974) Optical map transformations. *Optics Communications* 10(2): 164–168.

Bühling S, Schimmel H and Wyrowski F (2003) Wave-optical engineering with virtuallab. *Proceedings SPIE*, 5182: 24–33.

Bühling S, Schimmel H and Wyrowski F (2003) Challenges in industrial applications of diffractive beam splitting. *Proceedings SPIE* 5182: 123–131.

Collier RJ, Burckhardt CB and Lin LH (1971) *Optical Holography*. London: Academic Press.

Fienup JR (1980) Iterative method applied to image reconstruction and to computer-generated holograms. *Optical Engineering* 19(3): 297–305.

Gerchberg RW and Saxton WO (1972) A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* 35(2): 237–246.

Goodmann JW (1968) *Introduction to Fourier Optics*. New York: McGraw-Hill.

Goodman JW and Silvestri AM (1970) Some effects of Fourier-domain phase quantization. *IBM Journal of Reserved Development* 14: 478–484.

Krackhardt U, Mait JN and Streibl N (1992) Upper bound on the diffraction efficiency of phase-only fanout elements. *Applied Optics* 31(1): 27–37.

Lajunen H, Tervo J, Turunen J, Vallius T and Wyrowski F (2003) Simulation of light propagation by local spherical interface approximation. *Applied Optics* 34: 6804–6810.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge: Cambridge University Press.

Pfeil AV and Wyrowski F (2000) Wave-optical structure design with the local plane-interface approximation. *Journal of Modern Optics* 47(13): 2335–2350.

Pfeil AV, Wyrowski F, Drauschke A and Aagedal H (2000) Analysis of optical elements with the local plane-interface approximation. *Applied Optics* 39(19): 3304–3313.

Schimmel H, Bühling S and Wyrowski F (2003) Use of ray tracing in wave-optical engineering. *Proceedings SPIE* 5182: 6–14.

Schimmel H and Wyrowski F (2000) Amplitude matching strategy for wave-optical design of monofunctional systems. *Journal of Modern Optics* 47(13): 2295–2321.

Sergienko N, Stamnes JJ, Kettunen V, Kuittinen M, Turunen J, Vahimaa P and Friberg AT (1999) Comparison of electromagnetic and scalar methods for evaluation of diffractive lenses. *Journal of Modern Optics* 46(1): 65–82.

Smith WJ (1997) *Practical Optical System Layout*. New York: McGraw-Hill.

Smith WJ (2000) *Modern Optical Engineering*, 3rd edn. Bellingham: SPIE Press.

Turunen J (1997) Diffraction theory of microrelief gratings. In: Herzig HP (ed.) *Micro-Optics Elements, Systems and Applications*, chapter 2, pp. 31–52. London: Taylor & Francis.

Turunen J and Wyrowski F (1997) *Diffractive Optics for Industrial and Commercial Applications*. Berlin: Akademie Verlag.

Vallius T, Jefimovs K, Kettunen V, Kuittinen M, Laakkonen P and Turunen J (2001) Design of non-paraxial array illuminators by step-transition perturbation approach. *Journal of Modern Optics* 48(12): 1869–1879.

Wyrowski F (1990) Diffractive optical elements: Iterative calculation of quantized, blazed phase structures. *Journal of the Optical Society of America A* 7: 961–969.

Wyrowski F (1991) Digital phase encoded inverse filter for optical pattern recognition. *Applied Optics* 30: 4650–4657.

Wyrowski F (1991) Considerations on convolution and phase factors. *Optics Communications* 81(6): 353–358.

Wyrowski F (1991) Upper bound of the diffraction efficiency of diffractive phase elements. *Optics Letters* 16(24): 1915–1917.

Wyrowski F (1993) Design theory of diffraction elements in the paraxial domain. *Journal of the Optical Society of America A* 10: 1553–1561.

Wyrowski F (2003) Comments on wave-optical engineering. *Proceedings SPIE* 5182: 1–5.

Wyrowski F and Turunen J (2002) Wave-optical engineering. In: Guenther AH (ed.) *International Trends in Applied Optics*, chapter 21, pp. 471–496. Bellingham: SPIE Press.

DISPERSION MANAGEMENT

A E Willner, Y-W Song, J Mcgeehan and Z Pan, University of Southern California, Los Angeles, CA, USA

B Hoanca, University of Alaska Anchorage, Anchorage, AK, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical communications systems have grown explosively in terms of the capacity that can be transmitted over a single optical fiber. This trend has been fueled by two complementary techniques, those being the increase in data-rate-per-channel coupled with the increase in the total number of parallel wavelength channels. However, there are many considerations as to the total number of wavelength-division-multiplexed (WDM) channels that can be accommodated in a system, including cost, information spectral efficiency, nonlinear effects, and component wavelength selectivity.

Dispersion is one of the critical roadblocks to increasing the transmission capacity of optical fiber. The dispersive effect in an optical fiber has several ingredients, including intermodal dispersion in a multimode fiber, waveguide dispersion, material dispersion, and chromatic dispersion. In particular, chromatic dispersion is one of the critical effects in a single mode fiber (SMF), resulting in a temporal spreading of an optical bit as it propagates along the fiber. At data rates ≤ 2.5 Gbit/s, the effects of chromatic dispersion are not particularly troublesome. For data rates ≥ 10 Gbit/s, however, transmission can be tricky and the chromatic dispersion-induced degrading effects must be dealt with in some way, perhaps by compensation. Furthermore, the effects of chromatic dispersion rise quite rapidly as the bit rate increases – when the bit rate increases by a factor of four, the effects of chromatic dispersion increase by a factor of 16! This article will only deal with the management of chromatic dispersion in single mode fiber.

One of the critical limitations of optical fiber communications comes from chromatic dispersion, which results in a pulse broadening as it propagates along the fiber. This occurs as photons of different frequencies (created by the spreading effect of data modulation) travel at different speeds, due to the frequency-dependent refractive index of the fiber core. Compounding the problems cause by chromatic dispersion is the fact that as bit rates rise, chromatic dispersion effects rise quadratically with respect to the increase in the bit rate. One can eliminate these effects using fiber with zero chromatic dispersion, known as dispersion shifted fiber (DSF). However, with zero dispersion, all channels in a wavelength-division-multiplexed (WDM) system travel at the same speed, in-phase, and a number of deleterious nonlinear effects such as cross-phase modulation (XPM) and four-wave mixing (FWM) result. Thus, in WDM systems, some amount of chromatic dispersion is necessary to keep channels out-of-phase, and as such chromatic dispersion compensation is required. Any real fiber link may also suffer from 'dispersion slope' effects, in which a slightly different dispersion value is produced in each WDM channel. This means that while one may be able to compensate one channel exactly, other channels may progressively accumulate increasing amounts of dispersion, which can severely limit the ultimate length of the optical link and the wavelength range that can be used in a WDM system. This article will address the concepts of chromatic dispersion and dispersion slope management followed by some examples highlighting the need for tunability to enable robust optical WDM systems in dynamic environments. Some dispersion monitoring techniques are then discussed and examples given.

Chromatic Dispersion in Optical Fiber Communication Systems

In any medium (other than vacuum) and in any waveguide structure (other than ideal infinite free space), different electromagnetic frequencies travel at different speeds. This is the essence of chromatic dispersion. As the real fiber-optic world is rather distant from the ideal concepts of both vacuum and infinite free space, dispersion will always be a concern when one is dealing with the propagation of electromagnetic radiation through fiber. The velocity in fiber of a single monochromatic wavelength is constant. However, data modulation causes a broadening of the spectrum of even the most monochromatic laser pulse. Thus, all modulated data have a nonzero spectral width which spans several wavelengths, and the different spectral components of modulated data travel at different speeds. In particular, for digital data intensity modulated on an optical carrier, chromatic

dispersion leads to pulse broadening – which in turn leads to chromatic dispersion limiting the maximum data rate that can be transmitted through optical fiber (see Figure 1).

Considering that the chromatic dispersion in optical fibers is due to the frequency-dependent nature of the propagation characteristics, for both the material (the refractive index of glass) and the waveguide structure, the speed of light of a particular wavelength λ will be expressed as follows, using a Taylor series expansion of the value of the refractive index as a function of the wavelength:

$$v(\lambda) = \frac{c_0}{n(\lambda)} = \frac{c_0}{n_0(\lambda_0) + \frac{\partial n}{\partial \lambda}\delta\lambda + \frac{\partial^2 n}{\partial \lambda^2}(\delta\lambda)^2} \qquad [1]$$

Here, c_0 is the speed of light in vacuum, λ_0 is a reference wavelength, and the terms in $\partial n/\partial\lambda$ and $\partial^2 n/\partial\lambda^2$ are associated with the chromatic dispersion and the dispersion slope (i.e., the variation of the chromatic dispersion with wavelength), respectively. Transmission fiber has positive dispersion, i.e. longer wavelengths result in longer propagation delays. The units of chromatic dispersion are picoseconds per nanometer per kilometer, meaning that shorter time pulses, wider frequency spread due to data modulation, and longer fiber lengths will each contribute linearly to temporal dispersion. Higher data rates inherently have both shorter pulses and wider frequency spreads. Therefore, as network speed increases, the impact of chromatic dispersion rises precipitously as the square of the increase in data rate. The quadratic increase with the data rate is a result of two effects, each with a linear contribution. On one hand, a doubling of the data rate makes the spectrum twice as wide, doubling the effect of dispersion. On the other hand, the same doubling of the data rate makes the data pulses only half as long (hence twice as sensitive to dispersion). The combination of a wider signal spectrum and a shorter pulse width is what leads to the overall quadratic impact. Moreover, the data modulation format used can significantly affect the sensitivity of a system to chromatic dispersion. For example, the common nonreturn-to-zero (NRZ) data format, in which the optical power stays high throughout the entire time slot of a '1' bit, is more robust to chromatic dispersion than is the return-to-zero (RZ) format, in which the optical power stays high in only part of the time slot of a '1' bit. This difference is due to the fact that RZ data have a much wider channel frequency spectrum compared to NRZ data, thus incurring more chromatic dispersion. However, in a real WDM system, the RZ format increases the maximum allowable transmission distance by virtue of its reduced duty cycle (compared to the NRZ format), making it less susceptible to fiber nonlinearities as can be seen in Figure 2.

A rule for the maximum distance over which data can be transmitted is to consider a broadening of the pulse equal to the bit period. For a bit period B, a dispersion value D and a spectral width $\Delta\lambda$, the dispersion-limited distance is given by

$$L_D = \frac{1}{D \cdot B \cdot \Delta\lambda} = \frac{1}{D \cdot B \cdot (cB)} \propto \frac{1}{B^2} \qquad [2]$$

(see Figure 3). For example, for single mode fiber, $D = 17$ ps/nm/km, so for 10 Gbit/s data the distance is $L_D = 52$ km. In fact, a more exact calculation

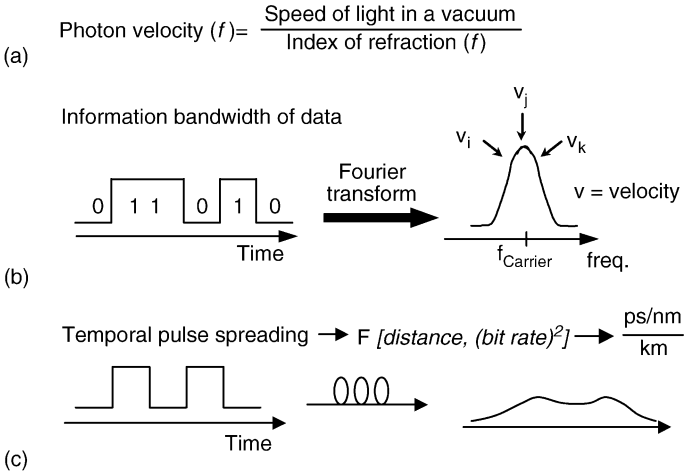


Figure 1 The origin of chromatic dispersion in data transmission. (a) Chromatic dispersion is caused by the frequency-dependent refractive index in fiber. (b) The nonzero spectral width due to data modulation. (c) Dispersion leads to pulse broadening, proportional to the transmission distance and the data rate.

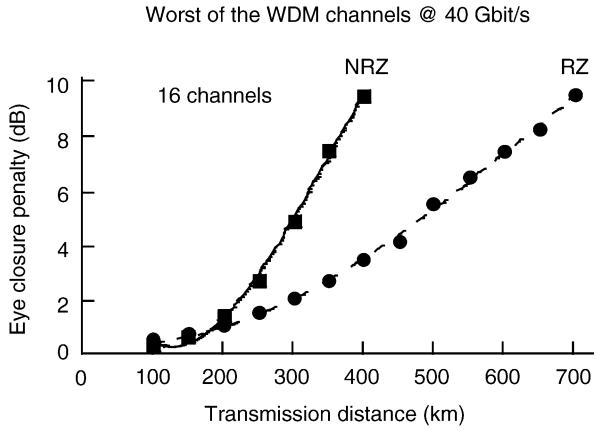


Figure 2 Performances of RZ and NRZ formats in a real fiber transmission link. (Reproduced with permission from Hayee I and Willner AE (1999) NRZ versus RZ in 10–40-Gb/s dispersion-managed WDM transmission systems. *IEEE Photon. Tech. Lett.* 11(8): 991–993.)

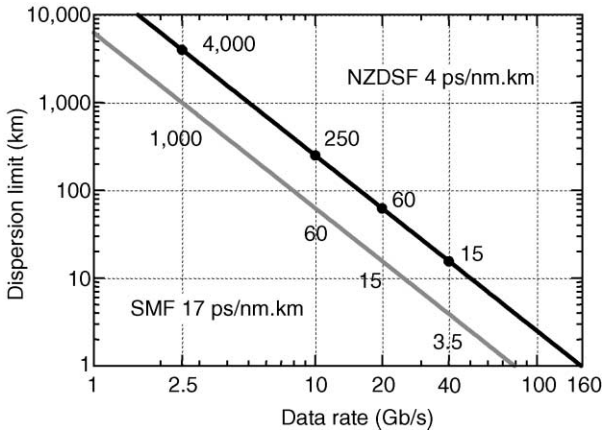


Figure 3 Transmission distance limitations due to uncompensated dispersion in SMF as a function of data rate for intensity modulated optical signals. (Reproduced with permission from Garrett LD (2001) Invited Short Course, *Optical Fiber Communication Conference*.)

shows that for 60 km, the dispersion induced power penalty is less than 1 dB (see Figure 4). The power penalty for uncompensated dispersion rises exponentially with transmission distance, and thus to maintain good signal quality, dispersion compensation is required.

Chromatic Dispersion Management

Optical Nonlinearities as Factors to be Considered in Dispersion Compensation

Even though it is possible to manufacture fiber with zero dispersion, it is not practical to use such fiber for WDM transmission, due to large penalties induced by fiber nonlinearities. Most nonlinear effects originate from the nonlinear refractive index of fiber, which is

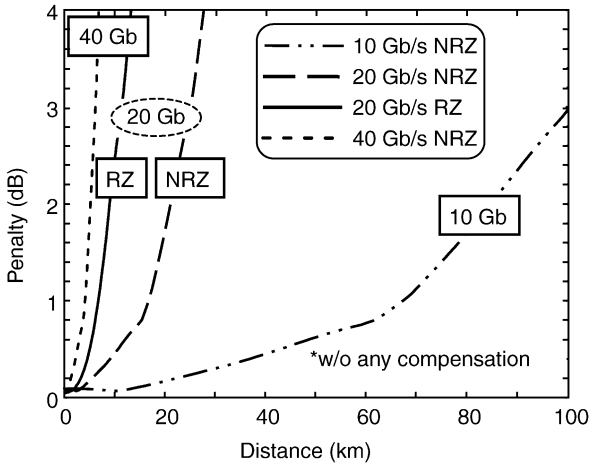


Figure 4 Power penalties due to uncompensated dispersion in single mode fiber (SMF) as a function of distance and data rate. (Reproduced with permission from Garrett LD (2001) Invited Short Course, *Optical Fiber Communication Conference*.)

not only dependent on the frequency of light but also on the intensity (optical power), and is related to the optical power as:

$$\bar{n}(f, P) = n(f) + n_2 \frac{P}{A_{\text{eff}}} \qquad [3]$$

where $n(f)$ is the linear part of the refractive index, P is the optical power inside the fiber, and n_2 is the nonlinear-index coefficient for silica fibers. The typical value of n_2 is 2.6×10^{-20} m^2/W. This number takes into account the averaging of the polarization states of the light as it travels in the fiber. The intensity dependence of the refractive index gives rise to three major nonlinear effects.

Self-phase modulation (SPM)

A million photons 'see' a different glass than does a single photon, and a photon traveling along with many other photons will slow down. SPM occurs because of the varying intensity profile of an optical pulse on a single WDM channel. This intensity profile causes a refractive index profile and, thus, a photon speed differential. The resulting phase change for light propagating in an optical fiber is expressed as:

$$\Phi_{\text{NL}} = \gamma P L_{\text{eff}} \qquad [4]$$

where the quantities γ and L_{eff} are defined as:

$$\gamma = \frac{2\pi n_2}{\lambda A_{\text{eff}}} \quad \text{and} \quad L_{\text{eff}} = \frac{1 - e^{-\alpha L}}{\alpha} \qquad [5]$$

where A_{eff} is the effective mode area of the fiber and α is the fiber attenuation loss. L_{eff} is the effective nonlinear length of the fiber that accounts for fiber loss, and γ is the nonlinear coefficient measured

in rad/km/W. A typical range of values for γ is between 10–30 rad/km/W. Although the nonlinear coefficient is small, the long transmission lengths and high optical powers, that have been made possible by the use of optical amplifiers, can cause a large enough nonlinear phase change to play a significant role in state-of-the-art lightwave systems.

Cross-phase modulation (XPM)

When considering many WDM channels co-propagating in a fiber, photons from channels 2 through N can distort the index profile that is experienced by channel 1. The photons from the other channels 'chirp' the signal frequencies on channel 1, which will interact with fiber chromatic dispersion and cause temporal distortion. This effect is called cross-phase modulation. In a two-channel system, the frequency chirp in channel 1, due to power fluctuation within both channels, is given by

$$\Delta B = \frac{d\Phi_{NL}}{dt} = \gamma L_{eff}\frac{dP_1}{dt} + 2\gamma L_{eff}\frac{dP_2}{dt} \qquad [6]$$

where, dP_1/dt and dP_2/dt are the time derivatives of the pulse powers of channels 1 and 2, respectively. The first term on the right-hand side of the above equation is due to SPM, and the second term is due to XPM. Note that the XPM-induced chirp term is double that of the SPM-induced chirp term. As such, XPM can impose a much greater limitation on WDM systems than can SPM, especially in systems with many WDM channels.

Four-wave-mixing (FWM)

The optical intensity propagating through the fiber is related to the electric field intensity squared. In a WDM system, the total electric field is the sum of the electric fields of each individual channel. When squaring the sum of different fields, products emerge that are beat terms at various sum and difference frequencies to the original signals. Figure 5 depicts that if a WDM channel exists at one of the four-wave-mixing beat-term frequencies, then the beat term will interfere coherently with this other WDM channel and potentially destroy the data.

Dispersion Maps

While zero-dispersion fiber is not a good idea, a large value of the accumulated dispersion at the end of a fiber link is also undesirable. An ideal solution is to have a 'dispersion map,' alternating sections of positive and negative dispersion as can be seen in Figure 6. This is a very powerful concept: at each point along the fiber the dispersion has some nonzero value, eliminating FWM and XPM, but the total dispersion at the end of the fiber link is zero, so that no pulse broadening is induced (Table 1). The most advanced systems require periodic dispersion compensation, as well as pre- and post-compensation (before and after the transmission fiber).

The addition of negative dispersion to a standard fiber link has been traditionally known as 'dispersion compensation,' however, the term 'dispersion management' is more appropriate. SMF has positive dispersion, but some new varieties of nonzero dispersion-shifted fiber (NZDSF) come in both positive and negative dispersion varieties. Some examples are shown in Figure 7. Reverse dispersion fiber is also now available, with a large dispersion

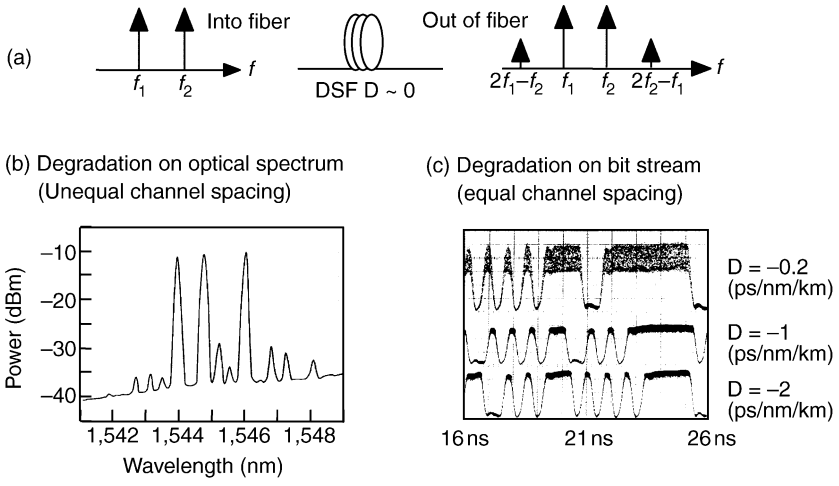


Figure 5 (a) and (b) FWM induces new spectral components via nonlinear mixing of two wavelength signals. (c) The signal degradation due to FWM products falling on a third data channel can be reduced by even small amounts of dispersion. (Reproduced with permission from Tkach RW, Chraplyvy AR, Forghieri F, Gnauck AH and Derosier RM (1995) Four-photon mixing and high-speed WDM systems. *Journal of Photon Technology* 13(5): 841–849.)

comparable to that of SMF, but with the opposite sign. When such flexibility is available in choosing both the magnitude and sign of the dispersion of the fiber in a link, dispersion-managed systems can be fully optimized to the desired dispersion map using a combination of fiber and dispersion compensation devices (see Figure 8).

Dispersion is a linear process, so first-order dispersion maps can be understood as linear systems. However, the effects of nonlinearities cannot be ignored, especially in WDM systems, with many tens of channels, where the launch power may be very high. In particular, in systems deploying dispersion compensating fiber (DCF), the large nonlinear coefficient of the DCF can dramatically affect the dispersion map.

Corrections to Linear Dispersion Maps

Chromatic dispersion is a necessity in WDM systems, to minimize the effects of fiber nonlinearities. A chromatic dispersion value as small as a few ps/nm/km is usually sufficient to make XPM and FWM negligible. To mitigate the effects of nonlinearities but maintain small amounts of chromatic dispersion, NZDSF is commercially available. Due to these nonlinear effects, chromatic dispersion must be managed, rather than eliminated.

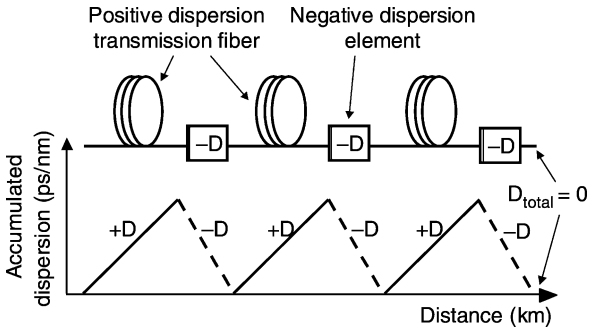


Figure 6 Dispersion map of a basic dispersion managed system. Positive dispersion transmission fiber alternates with negative dispersion compensation elements such that the total dispersion is zero end-to-end.

If a dispersion-management system was perfectly linear, it would be irrelevant whether the dispersion along a path is small or large, as long as the overall dispersion is compensated to zero (end to end). Thus, in a linear system the performance should be similar, regardless of whether the transmission fiber is SMF, and dispersion compensation modules are deployed every 60 km, or the transmission fiber is NZDSF (with approximately a quarter of the dispersion value of SMF) and dispersion compensation modules are deployed every 240 km. In real life, optical nonlinearities are very important, and recent results seem to favor the use of large, SMF-like, dispersion values in the transmission path and correspondingly high dispersion compensation devices. A recent study of performance versus channel spacing showed that the capacity of SMF could be more than four times that of NZDSF. This is because the nonlinear coefficients are much higher in NZDSF than in SMF, and for dense WDM the channel interactions become a limiting factor. A critical conclusion is that not all dispersion compensation maps are created equal: a simple calculation of the dispersion compensation, to cancel the overall dispersion value, does not lead to optimal dispersion map designs.

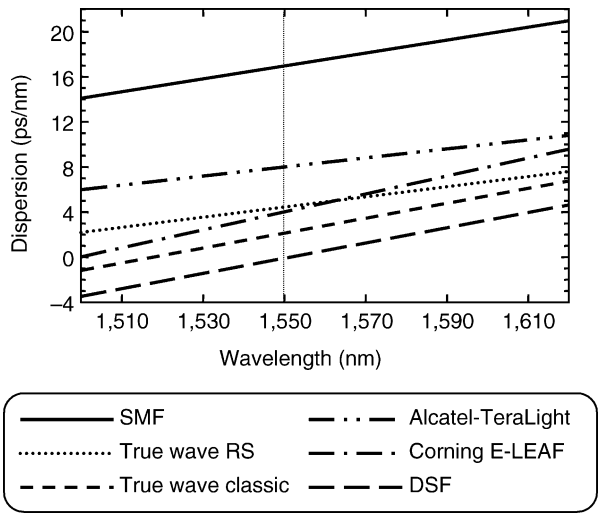


Figure 7 Chromatic dispersion characteristics of various commercially available types of transmission fiber.

Table 1 Commercially available fibers and their characteristics

	Dispersion @ 1,550 nm [ps/nm/km]	*Dispersion slope [ps/nm^2/km]*	*Attenuation [dB]*	*Mode field diameter [μm]*	*PMD [ps/km$^{0.5}$]*
SMF	18	0.08	≤0.25	11	≤0.5
Tera Light	8.0	0.058	≤0.2	9.2	≤0.04
TW-RS	4.4	0.043	≤0.25	8.4	≤0.03
LEAF	4.0	0.085	≤0.25	9.6	≤0.08
Standard DCF	−90	−0.22	≤0.5	5.2	≤0.08

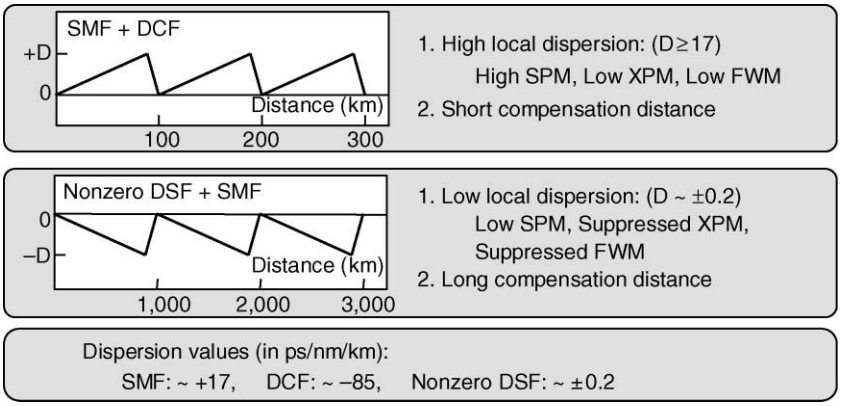


Figure 8 Various dispersion maps for SMF-DCF and NZDSF-SMF.

Additionally, several solutions have been shown to be either resistant to dispersion, or have been shown to rely on dispersion itself for transmission. Such solutions include chirped pulses (where prechirping emphasizes the spectrum of the pulses so that dispersion does not broaden them too much), dispersion assisted transmission (where an initial phase modulation tailored to the transmission distance leads to full-scale amplitude modulation at the receiver end due to the dispersion), and various modulation formats robust to chromatic dispersion and nonlinearities.

Dispersion Management Solutions

Fixed Dispersion Compensation

From a systems point of view, there are several requirements for a dispersion compensating module: low loss, low optical nonlinearity, broadband (or multichannel) operation, small footprint, low weight, low power consumption, and clearly low cost. It is unfortunate that the first dispersion compensation modules, based on DCF only, met two of these requirements: broadband operation and low power consumption. On the other hand, several solutions have emerged that can complement or even replace these first-generation compensators.

Dispersion compensating fiber (DCF)

One of the first dispersion compensation techniques was to deploy specially designed sections of fiber with negative chromatic dispersion. The technology for DCF emerged in the 1980s and has developed dramatically since the advent of optical amplifiers in 1990. DCF is the most widely deployed dispersion compensator, providing broadband operation and stable dispersion characteristics, and the lack of a dynamic, tunable DCF solution has not reduced its popularity.

As can be seen in Figure 9, the core of the average dispersion compensating fiber is much smaller than that of standard SMF, and beams with longer wavelengths experience relatively large changes in mode size (due to the waveguide structure) leading to greater propagation through the cladding of the fiber, where the speed of light is greater than that of the core. This leads to a large negative dispersion value. Additional cladding layers can lead to improved DCF designs that can include negative dispersion slope to counteract the positive dispersion slope of standard SMF.

In spite of its many advantages, DCF has a number of drawbacks. First, it is limited to a fixed compensation value. In addition, DCF has a weakly guiding structure and has a much smaller core cross-section, 19 μm^2, compared to the 85 μm^2 of SMF. This leads to higher nonlinearity, higher splice losses, as well as higher bending losses. Last, the length of DCF required to compensate for SMF dispersion is rather long, about one-fifth of the length of the transmission fiber for which it is compensating. Thus DCF modules induce loss, and are relatively bulky and heavy. The bulk is partly due to the mass of fiber, but also due to the resin used to hold the fiber securely in place. One other contribution to the size of the module is the higher bend loss associated with the refractive index profile of DCF; this limits the radius of the DCF loop to 6–8 inches, compared to the minimum bend radius of 2 inches for SMF.

Traditionally, DCF-based dispersion compensation modules are usually located at amplifier sites. This serves several purposes. First, amplifier sites offer relatively easy access to the fiber, without requiring any digging or unbraiding of the cable. Second, DCF has high loss (usually at least double that of standard SMF), so a gain stage is required

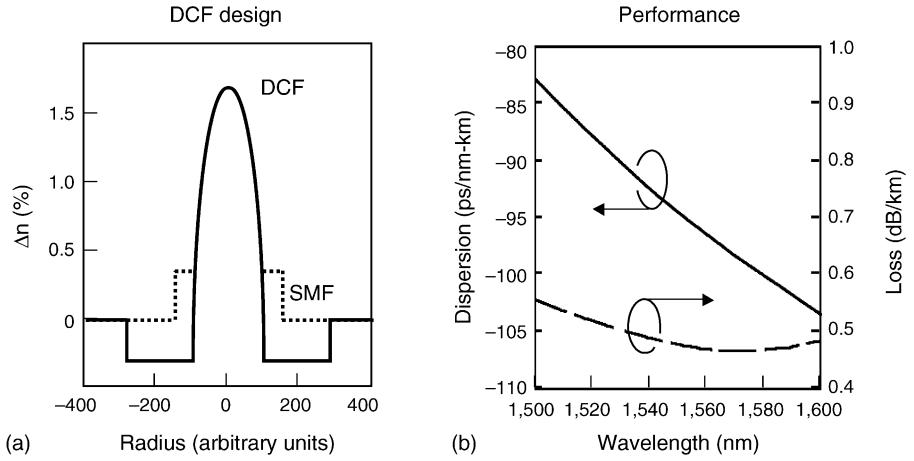


Figure 9 Typical DCF (a) refractive index profile and (b) dispersion and loss as a function of wavelength. Δn is defined as refractive index variation relative to the cladding.

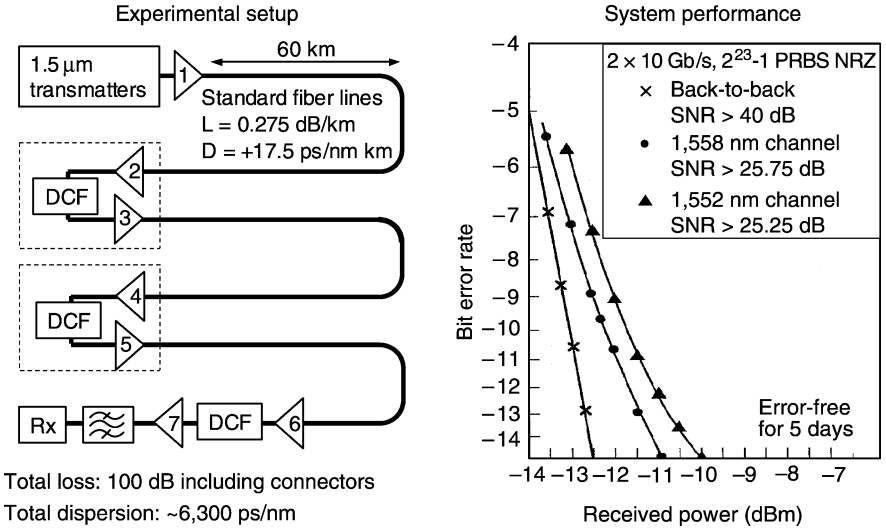


Figure 10 System demonstration of dispersion compensation using DCF. (Reproduced with permission from Park YK, Yeates PD, Delavaux J-MP, *et al.* (1995) A field demonstration of 20-Gb/s capacity transmission over 360 km of installed standard (non-DSF) fiber. *Photon. Technol. Lett.* 7(7): 816–818.)

before the DCF module to avoid excessively low signal levels. DCF has a cross-section four times smaller then SMF, hence a higher nonlinearity, which limits the maximum launch power into a DCF module. The compromise is to place the DCF in the mid-section of a two-section EDFA. This way, the first stage provides pre-DCF gain, but not to a power level that would generate excessive nonlinear effects in the DCF. The second stage amplifies the dispersion compensated signal to a power level suitable for transmission though the fiber link. This launch power level is typically much higher than could be transmitted through DCF without generating large nonlinear effects. Many newer dispersion compensation devices have better performance than DCF, in particular lower loss and lower nonlinearities. For this reason, they may not have to be deployed at the mid-section of an amplifier. Figure 10 shows the real demonstration results using the DCF.

Chirped fiber Bragg gratings

Fiber Bragg gratings have emerged as major components for dispersion compensation because of their low loss, small footprint, and low optical nonlinearity. Bragg gratings are sections of single-mode fiber in which the refractive index of the core is modulated in a periodic fashion, as a function of the spatial coordinate along the length of the fiber. When the spatial periodicity of the modulation matches what is known as a Bragg condition with respect to the wavelength of light propagating through the grating,

the periodic structure acts like a mirror, reflecting the optical radiation that is traveling through the core of the fiber. An optical circulator is traditionally used to separate the reflected output beam from the input beam.

When the periodicity of the grating is varied along its length, the result is a chirped grating which can be used to compensate for chromatic dispersion. The chirp is understood as the rate of change of the spatial frequency as a function of position along the grating. In chirped gratings the Bragg matching condition for different wavelengths occurs at different positions along the grating length. Thus, the roundtrip delay of each wavelength can be tailored by designing the chirp profile appropriately. Figure 11 compares the chirped FBG with uniform FBG. In a data pulse that has been distorted by dispersion, different frequency components arrive with different amounts of relative delay. By tailoring the chirp profile such that the frequency components see a relative delay which is the inverse of the delay of the transmission fiber, the pulse can be compressed back. The dispersion of the grating is the slope of the time delay as a function of wavelength, which is related to the chirp.

The main drawback of Bragg gratings is that the amplitude profile and the phase profile as a function of wavelength have some amount of ripple. Ideally, the amplitude profile of the grating should have a flat (or rounded) top in the passband, and the phase profile should be linear (for linearly chirped gratings) or polynomial (for nonlinearly chirped gratings). The grating ripple is the deviation from the ideal profile shape. Considerable effort has been expended on reducing the ripple. While early gratings were plagued by more than 100 ps of ripple, published results have shown vast improvement to values close to ± 3 ps.

Higher order mode dispersion compensation fiber

One of the challenges of designing standard DCF is that high negative dispersion is hard to achieve unless the cross-section of the fiber is small (which leads to high nonlinearity and high loss). One way to reduce both the loss and the nonlinearity is to use a higher-order mode (HOM) fiber (LP_{11} or LP_{02} near cutoff instead of the LP_{01} mode in the transmission fiber).

Such a device requires a good-quality mode converter between LP_{01} and LP_{02} to interface between the SMF and HOM fiber. HOM fiber has a dispersion per unit length greater than six times that of DCF. Thus, to compensate for a given transmission length in SMF, the length of HOM fiber required is only one sixth the length of DCF. Thus, even though losses and nonlinearity per unit length are larger for HOM fiber than for DCF, they are smaller overall, because of the shorter HOM fiber length. As an added bonus, the dispersion can be tuned slightly by changing the cutoff wavelength of LP_{02} (via temperature tuning). A soon to be released HOM fiber-based commercial dispersion compensation module is not tunable, but can fully compensate for dispersion slope.

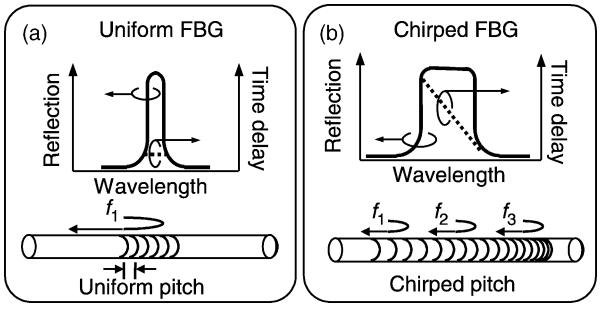


Figure 11 Uniform and chirped FBGs. (a) A grating with uniform pitch has a narrow reflection spectrum and a flat time delay as a function of wavelength. (b) A chirped FBG has a wider bandwidth, a varying time delay and a longer grating length.

Tunable Dispersion Compensation

The need for tunability

In a perfect world, all fiber links would have a known, discrete, and unchanging value of chromatic dispersion. Network operators would then deploy fixed dispersion compensators periodically along every fiber link to exactly match the fiber dispersion. Unfortunately, several vexing issues may necessitate that dispersion compensators are tunability, that they have the ability to adjust the amount of dispersion to match system requirements.

First, there is the most basic business issue of inventory management. Network operators typically do not know the exact length of a deployed fiber link nor its chromatic dispersion value. Moreover, fiber plants periodically undergo upgrades and maintenance, leaving new and nonexact lengths of fiber behind. Therefore, operators would need to keep in stock a large number of different compensator models, and even then the compensation would only be approximate. Second, we must consider the sheer difficulty of 40 Gbit/s signals. The tolerable threshold for accumulated dispersion for a 40 Gbit/s data channel is 16 times smaller than at 10 Gbit/s. If the compensation value does not exactly match the fiber to within a few percent of the required dispersion value, then the communication link will not work. Tunability is considered a key enabler for this bit rate (see Figures 12 and 13). Third, the accumulated dispersion changes slightly with temperature, which begins to be an issue for 40 Gbit/s systems and 10 Gbit/s ultra long-haul systems. In fiber, the zero-dispersion wavelength changes with temperature at a

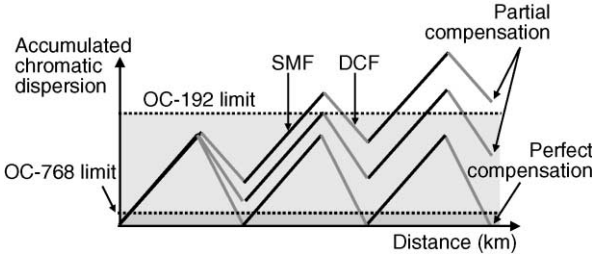


Figure 12 The need for tunability. The tolerance of OC-768 systems to chromatic dispersion is 16 times lower than that of OC-192 systems. Approximate compensation by fixed in-line dispersion compensators for a single channel may lead to rapid accumulation of unacceptable levels of residual chromatic dispersion.

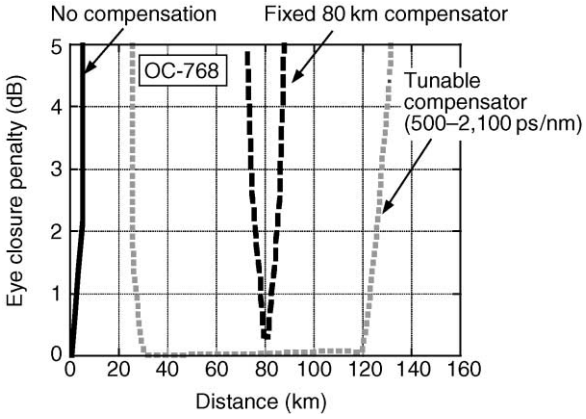


Figure 13 Tunable dispersion compensation at OC-768 (40 Gb/s) is essential for achieving a comfortable range of acceptable transmission distances (80 km for tunable, only ~4 km for fixed compensation).

typical rate of 0.03 nm/°C. It can been shown that a not-uncommon 50 °C variation along a 1,000 km 40 Gbit/s link can produce significant degradation (see Figure 14). Fourth, we are experiencing the dawn of reconfigurable optical networking. In such systems, the network path, and therefore the accumulated fiber dispersion, can change. It is important to note that even if the fiber spans are compensated span-by-span, the pervasive use of compensation at the transmitter and receiver suggests that optimization and tunability based on path will still be needed.

Other issues that increase the need for tunability include: (i) laser and (de)mux wavelength drifts for which a data channel no longer resides on the flat-top portion of a filter, thereby producing a chirp on the signal that interacts with the fiber's chromatic dispersion; (ii) changes in signal power that change both the link's nonlinearity and the optimal system dispersion map; and (iii) small differences that exist in transmitter-induced signal chirp.

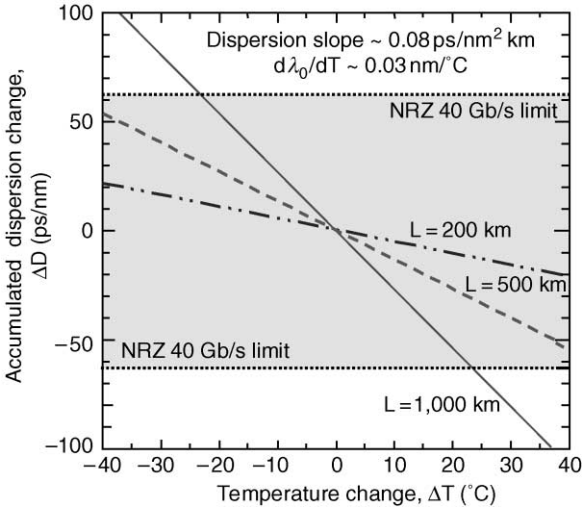


Figure 14 Accumulated dispersion changes as a function of the link length and temperature fluctuation along the fiber link.

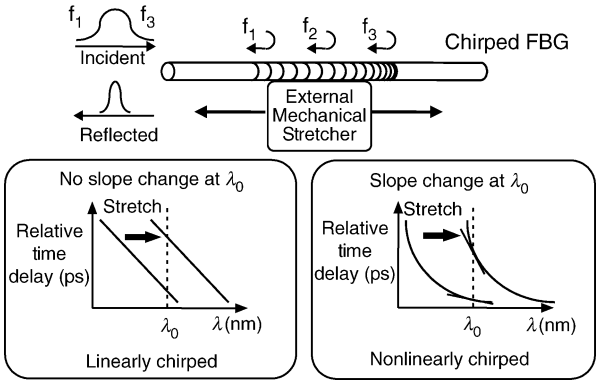


Figure 15 Tuning results for both linearly and nonlinearly chirped FBGs using uniform stretching elements. The slope of the dispersion curve at a given wavelength λ_0 is constant when the linearly chirped grating is stretched, but changes as the nonlinearly chirped grating is stretched.

Approaches to tunable dispersion compensation

A host of techniques for tunable dispersion compensation have been proposed in recent years. Some of these ideas are just interesting research ideas, but several have strong potential to become viable technologies.

Fiber gratings offer the inherent advantages of fiber compatibility, low loss, and low cost. If a FBG has a refractive-index periodicity that varies nonlinearly along the length of the fiber, it will produce a time delay that also varies nonlinearly with wavelength (see Figure 15). Herein lies the key to tunability. When a linearly chirped grating is stretched uniformly by a single mechanical element, the time delay curve is shifted towards longer wavelengths, but the slope of the ps-vs.-nm curve remains constant at all wavelengths within the passband. When a

nonlinearly-chirped grating is stretched, the time delay curve is shifted toward longer wavelengths, but the slope of the ps-vs.-nm curve at a specific channel wavelength changes continuously. Ultimately, tunable dispersion compensators should accommodate multichannel operation. Several WDM channels can be accommodated by a single chirped FBG in one of two ways: fabricating a much longer (i.e., meters-length) grating, or using a sampling function when writing the grating, thereby creating many replicas of transfer function of the FBG in the wavelength domain (see Figure 16).

One free-space-based tunable dispersion compensation device is the virtually imaged phased array (VIPA), based on the dispersion of a Fabry–Perot interferometer. The design requires several lenses, a movable mirror (for tunability), and a glass plate with a thin film layer of tapered reflectivity for good mode matching. Light incident on the glass plate undergoes several reflections inside the plate. As a result, the beam is imaged at several virtual locations, with a spatial distribution that is wavelength-dependent.

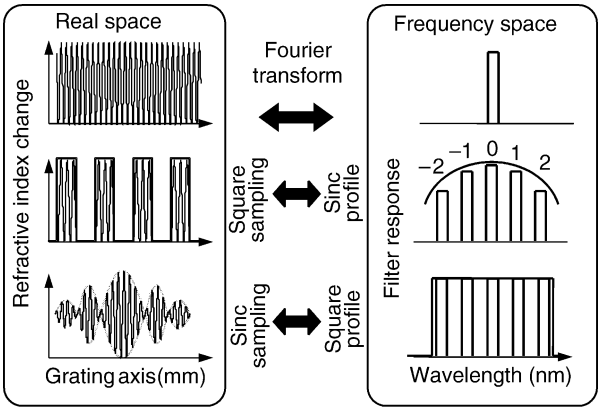


Figure 16 The concept of 'sampled' FBGs, where a superstructure is written on top of the grating that produces a Fourier transform in the frequency domain, leading to multiple grating passbands.

Several devices used for dispersion compensation can be integrated on a chip, using either an optical chip media (semiconductor-based laser or amplifier medium) or an electronic chip. One such technology is the micro-ring resonator, a device that, when used in a structure similar to that of an all-pass filter (see Figure 17), can be used for dispersion compensation on a chip-scale. Although these technologies are not yet ready for deployment as dispersion compensators, they have been used in other applications and have the potential to offer very high performance at low cost.

As the ultimate optical dispersion compensation devices, photonic bandgap fibers (holey fibers) are an interesting class in themselves (see Figure 18). These are fibers with a hollow structure, with holes engineered to achieve a particular functionality. Instead of being drawn from a solid preform, holey fibers are drawn from a group of capillary tubes fused together. This way, the dispersion, dispersion slope, the nonlinear coefficients could in principle all be precisely designed and controlled, up to very small or very large values, well outside the range of those of the solid fiber.

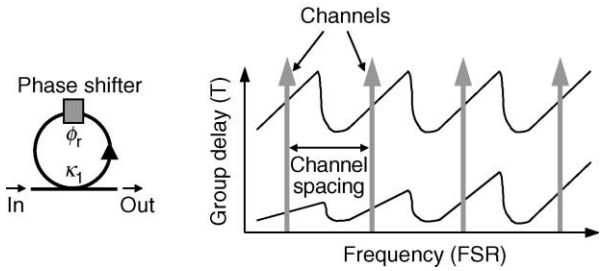


Figure 17 Architecture of an all-pass filter structure for chromatic dispersion and slope compensation. (Reproduced with permission from Madsen CK, Lenz G, Bruce AJ, *et al.* (1999) Integrated all-pass filters for tunable dispersion and dispersion slope compensation. *Photon. Technol. Lett.* 11(12): 1623–1625.)

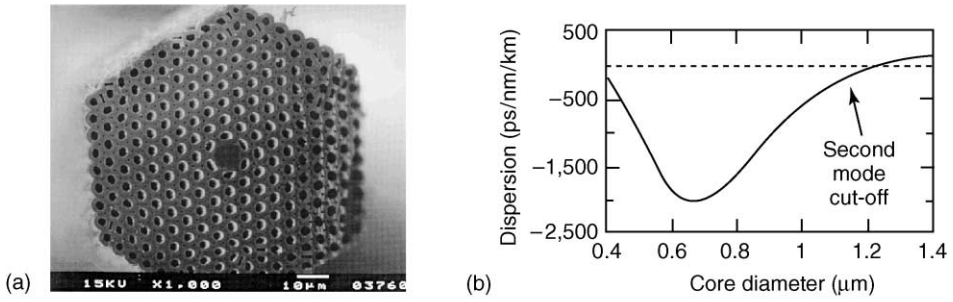


Figure 18 (a) SEM image of a photonic crystal fiber (holey fiber), and (b) net dispersion of the fiber at 1550 nm as a function of the core diameter. (Reproduced with permission from Birk TA, Mogihetsev D, Knight JC and Russell PSt.J (1999) Integrated all-pass filters for tunable dispersion and dispersion slope compensation. *Photon. Technol. Lett.* 11(6): 674–676.)

Dispersion Slope Mismatch

Transmission fiber, especially fiber with dispersion compensation built in, may suffer from a dispersion slope in which a slightly different dispersion value is produced for each WDM channel (see Figure 19). Even though the compensator would be able to cancel the dispersion of the fiber at the design wavelength, there will be residual dispersion left at the other wavelength channels unless the compensator can match the slope of the dispersion curve of the fiber as well. Some solutions for dispersion slope compensation are described in this section.

First, DCF, with negative dispersion slope, is a prime candidate for deployment as a dispersion slope compensator even though it cannot easily be made tunable. By designing the DCF with the same ratio of dispersion to dispersion slope as that of a real fiber link, new types of DCF can be used to compensate for both dispersion and dispersion slope, much like DCF is used for dispersion compensation today. DCF's popularity, wideband functionality, and stable dispersion characteristics make this a particularly attractive solution. Designing the DCF to match the dispersion characteristics of the transmission fiber is the critical engineering challenge in this slope compensation scheme. Second, third-order nonlinearly chirped FBG can act as a tunable dispersion slope compensator. A simple modification of the nonlinearly chirped FBG allows tuning of the compensated dispersion slope value via stretching the grating. The grating is prepared such that the time delay as a function of wavelength has a cubic profile that covers several WDM channels over a continuous bandwidth of many nanometers. Since the resulting dispersion curve is quadratic over the grating bandwidth, the dispersion slope experienced by the WDM channels can be tuned by stretching the grating using a single mechanical element (see Figure 20). Third, combining the VIPA, with either a 3D mirror or diffraction grating, can also provide tunable free-space dispersion slope compensation. Slope tuning is achieved by dynamically controlling the MEMS-based 3D mirror or the diffraction grating. Fourthly, using an FBG with many spaced thin-film heater sections can also enable tunable dispersion slope. Each heater can be individually electrically controlled, allowing the time delay profile of the grating to be dynamically tuned via changing the temperature along the length of the grating. Advanced applications of this technique can allow alterations to the entire time delay profile of the grating, providing a truly flexible dispersion and dispersion slope compensation mechanism.

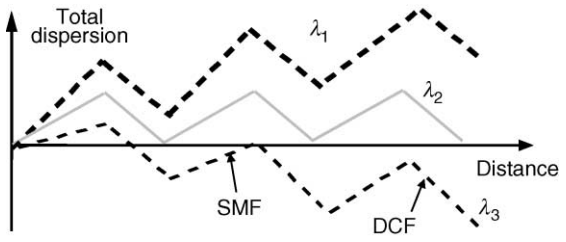


Figure 19 The chromatic dispersion slope mismatch caused by the different slopes of transmission fiber (SMF or NZDSF) and DCF.

Chromatic Dispersion Monitoring

Another important issue related to dispersion management is dispersion monitoring techniques. In a reconfigurable system, it is necessary to reconfigure

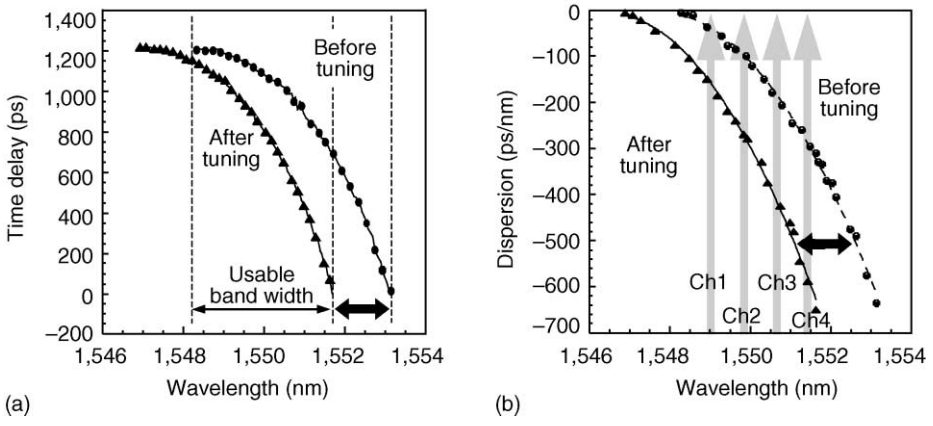


Figure 20 Tunable dispersion slope compensation using a third-order nonlinearly chirped FBG. (a) Cubic time delay curves of the grating, and (b) quadratic dispersion curves showing the change in dispersion in each channel before and after tuning. (Reproduced with permission from Song YW, Motoghian SMR, Starodubov D, *et al.* (2002) Tunable dispersion slope compensation for WDM systems using a non-channelized third-order-chirped FBG. *Optical Fiber Communication Conference* 2002. Paper ThAA4.)

any tunable chromatic dispersion compensation modules on the fly as the network changes. An in-line chromatic dispersion monitor can quickly measure the required dispersion compensation value while data are still being transmitted through the optical link. This is very different from the more traditional chromatic dispersion measurement techniques where dark fiber is used and the measurement is done off-line over many hours (or days).

Chromatic dispersion monitoring is most often done at the receiving end, where the Q-factor of the received data or some other means is employed to assess the accumulated dispersion. Existing techniques that can monitor dispersion in-line, fast, and with relatively low cost include: (i) general performance monitoring using the bit-error rate (BER) or eye opening, but this approach cannot differentiate among different degrading effects; (ii) detecting the intensity modulation induced by phase modulation at the transmitter; (iii) extracting the bit-rate frequency component (clock) from photo-detected data and monitoring its RF power (see Figure 21); (iv) inserting a subcarrier at the transmitter and subsequently monitoring the subcarrier power degradation; (v) extracting several frequency components from the optical data using very narrow-band optical filters and detecting the optical phase; (vi) dithering the optical-carrier frequency at the transmitter and measuring the resultant phase modulation of the clock extracted at receiver with an additional phase-locked loop; and (vii) using an optical filter to select the upper and lower vestigial sideband (VSB) signals in transmitted optical data and determine the relative group delay caused by dispersion (see Figure 22).

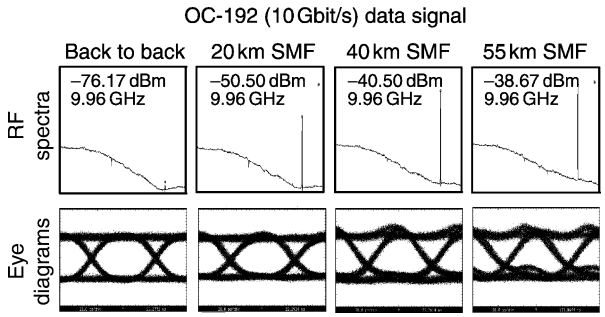
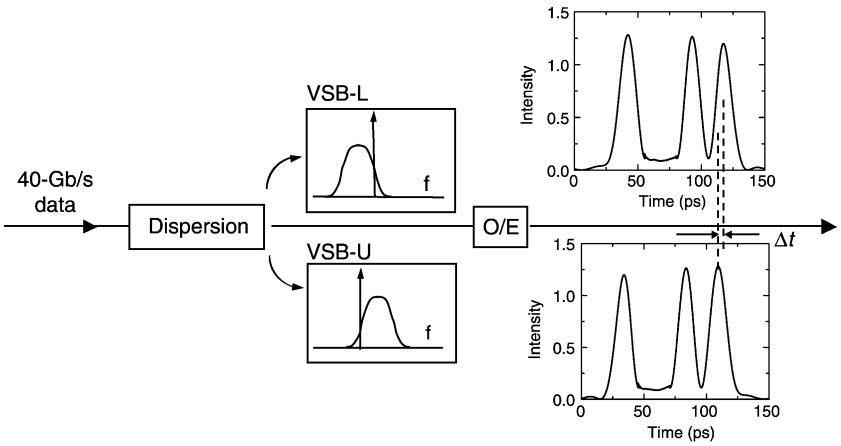


Figure 21 Clock regenerating effect due to chromatic dispersion for NRZ data. As the amount of residual dispersion increases, so does the amount of power at the clock frequency. This power can be used to monitor the amount of uncompensated chromatic dispersion. (Reproduced with permission from Pan Z, Yu Q, Xie Y, *et al.* (2001) Chromatic dispersion monitoring and automated compensation for NRZ and RZ data using clock regeneration and fading without adding signalling. *Optical Fiber Communication Conference 2001.* Paper WH5.)

Conclusion

Chromatic dispersion is a phenomenon with profound implications for optical fiber communications systems. It has negative effects, broadening data pulses, but it also helps reduce the effects of fiber nonlinearities. For this reason, managing dispersion, rather than trying to eliminate it altogether, is the key. Fixed dispersion components are suitable for point-to-point OC-192 systems. Tunable dispersion components are essential for dispersion management in reconfigurable and OC-768 systems. These dispersion compensation elements must have low loss, low nonlinearity, and must be cost effective.



Clock phase shift = $2\pi \times R_b \times$ (Time delay) = f(dispersion)

Figure 22 Chromatic dispersion monitoring using the time delay (Δt) between two VSB signals, which is a function of chromatic dispersion. (Reproduced with permission from Yu Q, Yan L-S, Pan Z and Willner AE (2002) Chromatic dispersion monitor for WDM systems using vestigial-sideband optical filtering. *Optical Fiber Communication Conference 2002.* Paper WE3.)

Although several technologies have emerged that meet some or all of the above requirements, no technology is a clear winner. The trend is towards tunable devices, or even actively self-tunable compensators, and such devices will allow system designers to cope with the shrinking system margins and with the emerging rapidly reconfigurable optical networks.

List of Units and Nomenclature

Chromatic dispersion	The time-domain pulse broadening in an optical fiber caused by the frequency dependence of the refractive index. This results in photons at different frequencies traveling at different speeds [ps/nm/km].
Chromatic dispersion slope	Different wavelengths have different amounts of dispersion values in an optical fiber [ps/nm^2/km].
Conventional single mode fiber (SMF)	Fiber that transmits only a single optical mode by virtue of a very small core diameter relative to the cladding. Provides lowest loss in the 1,550 nm region and has zero dispersion at 1,300 nm.
Cross-phase modulation (XPM)	A nonlinear Kerr effect in which a signal undergoes a nonlinear phase shift induced by a copropagating signal at a different wavelength in a WDM system.
Dispersion compensating fiber (DCF)	Optical fiber that has both a large negative dispersion and dispersion slope around 1,550 nm.
Fiber Bragg gratings (FBGs)	A small section of optical fiber in which there is a periodic change of refractive index along the core of the fiber. An FBG acts as a wavelength-selective mirror, reflecting only specific wavelengths (Bragg wavelengths) and passing others.
Four-wave mixing (FWM)	A nonlinear Kerr effect in which two or more signal wavelengths interact to generate a new wavelength.
Kerr effect	Nonlinear effect in which the refractive index of optical fiber varies as a function of the intensity of light within the fiber core.
Modulation	The process of encoding digital data onto an optical signal so it can be transmitted through an optical network.
Nonlinear effects	Describes the nonlinear responses of a dielectric to intense electromagnetic fields. One such effect is the intensity dependence of the refractive index of an optical fiber (Kerr effect).
Nonzero dispersion-shifted fiber (NZDSF)	Optical fiber that has a small amount of dispersion in the 1,550 nm region in order to reduce deleterious nonlinear effects.
Polarization mode dispersion (PMD)	Dispersion resulting from the fact that different light polarizations within the fiber core will travel at different speeds through the optical fiber [ps/km$^{0.5}$].
Self-phase modulation (SPM)	A nonlinear Kerr effect in which a signal undergoes a self-induced phase shift during propagation in optical fiber.
Wavelength division multiplexing (WDM)	Transmitting many different wavelengths down the same optical fiber at the same time in order to increase the amount of information that can be carried.

See also

Nonlinear Optics, Basics: Four-Wave Mixing. **Photonic Crystals:** Nonlinear Optics in Photonic Crystal Fibers.

Further Reading

Agrawal GP (1997) *Fiber-Optic Communication Systems*. Rochester, NY: John Wiley & Sons.

Agrawal GP (1997) *Nonlinear Fiber Optics*, 2nd edn. Academic Press.

Gowar J (1993) *Optical Communication Systems*, 2nd edn. Prentice Hall.

Kaminow IP and Koch TL (1997) *Optical Fiber Telecommunications IIIA & IIIB*. Academic Press.

Kashyap R (1999) *Fiber Bragg Gratings*. Academic Press.

Kazovsky L, Benedetto S and Willner A (1996) *Optical Fiber Communication Systems*. Artech House.

Willner AE and Hoanca B (2002) Fixed and tunable management of fiber chromatic dispersion. In: Kaminow IP and Li T (eds) *Optical Fiber Telecommunications IV*. Academic Press.

Yariv A and Yeh P (1984) *Optical Waves in Crystals*. John Wiley & Sons.

DISPLAYS

R L Donofrio, Display Device Consultants LLC, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article is a brief review of display technology being used today. It emphasizes some of the optical aspects of displays. The details of the basic construction of the display devices are also discussed to aid in understanding. The displays covered are the CRT, VFD, FED, PDP, LED, OLED, LCD-AMLCD, transmissive and transflective LCDs.

Definition of a Display

A display can be defined as an object which transfers information to the viewer and/or an output unit that gives a visual representation of data.

Optical Characteristics of a Display

A display has (in general) the optical characteristics of luminance, reflectivity, contrast ratio, color, surface gloss, resolution, and image sharpness (sometimes called small area contrast). Measurement methods can be obtained from the EIA (Electronics Industries Alliance), JT-31 (Optical Characteristics of Display Devices), SAE (Society of Automotive Engineers and ISO Standards) and VESA Standards, to name a few.

Luminance

The luminance of a display is its light output and is given in units of candela/square meter or cd/m^2 (in the English system fL). Luminance is measured using a photometer, photometer/colorimeter, or spectroradiometer.

Illuminance

The illuminance is the light incident on the surface being measured. Its units are Lumen per m^2 or Lux (and in the English system Ft.cd). This is measured using illuminance meters.

Reflectivity

The reflectivity for CRTs (cathode ray tubes) is measured with incident light at 45 degrees (in some cases for LCDs, a smaller 30 degree angle is used). The reflected light from the surface is measured with a detector normal to (perpendicular to) the surface. Devices such as 'tube face reflectivity' (TFR) can be used to measure this parameter and it is published in the EIA-JT-31 measurement methods. Other similar methods are used for flat panel displays.

Color

Although color perception is different for different individuals, standard responses to color stimuli have been developed and we can measure the color of different objects and display images. Color can be measured with such equipment as a colorimeter or spectroradiometer. The output data are given in various color systems. Color systems, such as the 1931 CIE (Commission International de I'Eclairage) Color System (x, y), the 1960 CIE (u, v) or 1976 CIE (u', v') color coordinates are all used. All these color systems can be converted to each other by simple mathematical relations and these conversions are usually available to the user in the data output. The choice of the color space/system is dependent on application but the system in most general use for displays is the 1976 CIE (u', v'). This color system has become more widely used because equally perceived color differences in this color system are at almost equal distances in this color space. In reality, the quest for a system with equal distances equating to equal color differences can be achieved through a complicated matrix approach in which the constants used in the matrix vary from point to point in the color space.

Gloss

Gloss is a measure of surface reflectivity under Snell's Law criteria (equal incidence and reflection angles). The measurement angles of 20, 45, 60, and 85 degrees are used to measure the gloss of common surfaces. Display glass surface measurements tend to use the 60 degree gloss angle. The surface of a display can be made rougher to scatter incident light away from the viewer or it can be optically coated to reduce the reflected light to achieve improved viewing. Some times both methods are used. Reduced gloss surfaces also are involved with the reduction in surface glare. The level of increased roughness of the surface affects the user's ability to see detail.

Contrast Ratio

When people talk about contrast they generally mean 'contrast ratio'. The contrast ratio (CR) of an area on a display is a unitless parameter and is defined as the excited luminance (L_e) divided by

the unexcited luminance (L_u). The excited luminance (L_e) equals the glass transmission (T) times some excited internal luminance (L_{ei}) and for simplicity, L_u equals T times an internal zero ambient unexcited luminance light level of (L_{ui}). Factors, such as internal scattered light in LCDs and scattered electrons and scattered light in CRTs, give a nonzero un-excited luminance L_u.

When dealing with the CR, the effect of ambient illumination A contributes to the numerator and the denominator. Thus, when the ambient illumination is increased, it lowers the contrast ratio. In a simple case where the ambient illumination is reflected from the surface of the display whose reflectivity is ρ, the contrast ratio is given below:

$$\mathrm{CR} = (L_e + A\rho)/(L_u + A\rho) \quad [1]$$

For the case of the reflectivity of a phosphor screen (reflectance R) coated panel with a glass transmission (T), the reflectivity to first approximation is

$$\rho = cRT^2 \quad [2]$$

where c is a constant.

So we see that changes in contrast can come about through changing the display's glass panel transmission. For example, if the L_e of the white field of a CRT was 100 fL with a 90% panel glass (along with $R = 0.5$) and had a L_u of 2 fL, then the zero ambient contrast would be 50. However, with an ambient illumination of 50 Ftcd the contrast would be 4.76. Now, when we use a 50% transmission panel glass at the ambient illumination of 50 Ft.cd the CR is increased to approximately 8.39. Equations [1] and [2] show how to calculate the contrast. It should be noted that decreased panel glass transmission or the use of 'dark glass' increases contrast because the reflectivity is a 'T squared' effect and the luminance L is a function 'T'. There are second-order effects involved with the reduction of reflected light and unexcited luminance which make the equations more complicated. These second-order effects will not be covered as they are beyond the basic discussion of this article.

In comparing display CR, one must compare the displays at the same ambient illumination. Additionally, you will find that a number of flat panel display advertising reports will report the CR at zero ambient illumination. As we have shown, these CR values are the highest ones for the display. The CR for emissive displays tends to be constant over a wide range of viewing angles, whereas the contrast ratio for liquid crystal displays (LCDs) is more variable over large-angle viewing. Recent new electrode position developments, such as the 'in-plane switching' technology for active matrix liquid crystal displays (AMLCDs) has gone a long way to reduce this reduced angle-of-view effect in that type of LCD display.

Figure 1 shows how a viewer in a sport utility vehicle (SUV) would see the changes in contrast with viewing angle for two types of twisted nematic LCDs (TN-LCDs). One of the TN LCDs does not

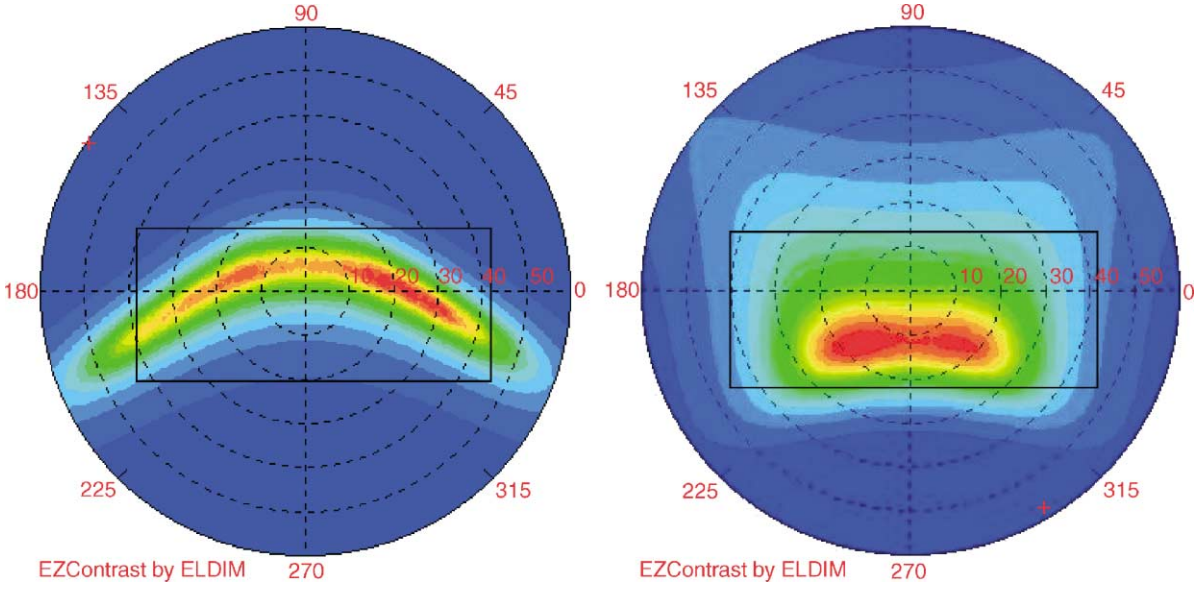


Figure 1 A typical rectangular viewing envelope for a mid-size SUV vehicle superimposed on LCD iso-contrast plots. Left is standard TN LCD. Right, is film-compensated TN LCD. Reproduced from Smith-Gillespie R, *et al.* (2001) Design requirements for automotive entertainment displays. *Proceedings of the 8th Annual Symposium on Vehicle Displays (2001).* Permission for reprint courtesy of the Society for Information Displays.

have the film compensation and the other does. For relative contrast appraisal, the red areas show the highest contrast regions and the blue the lowest. It is obvious that different technologies can give significant changes in viewed contrast.

Phosphor Aging

The fall off of luminance as the phosphor screen is bombarded with the electron beam is called phosphor aging. This effect is due to two main factors: (i) the thermal quenching of the phosphor luminescence (which is reversible); and (ii) a nonreversible aging due to damage to the screen components (such as phosphor, binder, and aluminum coating). It has been found that the aging (when we are using an electron beam) is a function of the number of electrons bombarding the screen. Aging models help us to understand and predict the amount of aging a tube will show after a fixed amount of time. The first aging model in 1951 was on ion bombardment and showed that:

$$I = I_0/(1 + AN) \quad [3]$$

where I_0 is the initial luminance before aging, and I is the final luminance.

In 1961 a similar (one parameter) Model $I = I_0/(1 + CN)$ was used by Pfhanl, where C is the aging constant to half luminance and N is the number of coulombs. In 1981, a two-component model for low current aging was presented, where:

$$I = I_0/(1 + N(B/Q^D)) \quad [4]$$

Here we have two constants B and D, and Q is the cumulated charge in number of coulombs. When $D = 0$ then the equation given above goes to the Pfhanl case.

Later, in 1996 the three-parameter model was described and in 1997 the Gaussian beam aging model was described.

Ambient Illumination and Reflection

To see the information shown on a display, the information must be spatially detected by the human eye over the reflected ambient illumination (such as room and sunlight). The ambient illumination is generally from localized or diffuse sources. As we have seen previously, the effect of diffuse illumination can be reduced through the use of neutral density filters. But, when we are dealing with very high ambient light such as that entering the canopy of a fighter aircraft then, the monochromatic information from the display can be viewed using selective spectral filters. However, for the usual computer monitor, the 'specular reflection' (Snell's Law type of reflection) off the front of the display is due to a localized external light source. This reflection is reduced by methods such as anti-glare (AG) and anti-reflecting (AR) surfaces applied to the front surface of the display. A flat versus curved front surface of the display can also be a glare reduction technique. Surface roughness is used in AG surfaces to scatter a portion of the specular light away from the viewer. Surface roughness can be achieved through surface abrading, surface plastic coating, surface silicate coating, and surface acid etching to name a few methods. AR surfaces use optical interference techniques to reduce the reflected light to the viewer. In some cases, both methods are applied to a display surface. Surface gloss of the display is usually measured with a gloss meter in 60 degree gloss values but 20, 45, and 80 degree angles are also used. Resolution is measured in line pairs/mm or pixels/mm and image sharpness is associated with modulation transfer function (MTF) measurements. We should note here that resolution and image sharpness are not the same thing. High resolution is equated with the high frequency limit of the MTF measurement. Since the reduction in specular reflection through the use of surface roughness can affect the image sharpness, we see that a system of trade-offs can be developed to optimize a display for a given viewing environment. AR coatings improve glare reduction without scattering and thus give higher-resolution capability to the displays but this is usually obtained at a higher price then the AG coatings.

Types of Displays

Displays can be divided into two types; emissive and nonemissive. By emissive we mean that the display surface gives off light. In the nonemissive, the display or a small portion of the display acts as a shutter or spatial filter which also can be reflective in nature. In this nonemissive mode, the display can pass light from a source in back of or in front of the shutter (or both, as we shall see later).

Table of Emissive and Nonemissive Displays

Thus from Table 1 we see that under the title of emissive displays we have the cathode ray tube (CRT), projection CRT, vacuum fluorescent display (VFD), field emissive display (FED), electroluminescent (EL) display, light emitting diode (LED), organic light emitting diode (OLED), and plasma. These devices, which are in the realm of emissive displays, use certain types of emissive materials called

Table 1 Table of emissive and nonemissive displays

Emissive	*Nonemissive*
CRT	LCD and AMLCD
VFD	Bi-Stable LCD
FED	Digital Mirror D
EL – inorganic	LCOS (Liquid Crystal on Si)
LED	Electronic Inks
OLED	
PLASMA	

lumophors (or phosphors). The lumophors are called cathodoluminescent, photoluminescent, and electroluminescent, depending whether the excitation action was due to the interaction of the phosphor with cathode rays, photons, or low-voltage electrons. Cathodoluminescent lumophors are used in cathode ray tubes (CRTs – television tubes, monitor tubes, projection CRTs), (VFDs), and (FEDs).

EL displays can use inorganic lumophors (phosphors) or organic lumophors such as those used in OLEDs. Under photoluminescence displays, we have 'plasma displays' (plasma TV), IR-up conversion, and fluorescent lamps. In plasma displays, the excitation of the plasma in a localized region gives rise to UV photons which strike the phosphor screen in that region. In IR-up conversion, IR photons are absorbed in such a way that two IR photons can excite an electron with twice the energy of one IR photon. This then allows the use of IR lasers to create displays through the excitation of these up-conversion phosphors. In fluorescent lamps, the UV created by the excitation of Hg or Xe, strikes the special lamp phosphors (coated on the inside of the glass envelope), which become luminescent.

Under the heading of shutters we can consider LCDs which consist of AMLCDs, super twisted nematic liquid crystal displays (STN LCDs) and bistable LCDs. Another shutter type is the liquid crystal on silicon (LCOS) display and the movable digital mirror display (DMD), and the digital light processing (DLP) projecting engines. Another spatial display type is both normal and dynamic inks. A dynamic ink can be a 'polarizable' ink display in which a two shade (light/dark) ink sphere element can rotate under the presence of an electric field to show the light or dark side. Although they are very interesting, we will not cover ink displays in this article.

The shutter action is also used in some projection TV sets (or projectors) in which the LCD either transmits or reflects (such as LCOS) the light from an arc lamp. The DMD movable mirrors reflect the digital information to the lens system and the unwanted rays are reflected away from the lens to a black light capture area.

The Cathode Ray Tube (CRT)

The CRT is a display which has been in use for over 100 years (the Braun CRT tube was developed in 1898). The CRT uses a phosphor screen which is bombarded by energetic electrons and emits visible light radiation. The single electron gun CRT was used for years in early monochrome television sets and monochrome computer displays. Monochrome green, amber, and white phosphors have been used in monitor tubes. Some color selectivity can be obtained with the single electron gun using the following three methods; (i) voltage penetration phosphors, (ii) current sensitive phosphors and (iii) beam index methods. In a CRT, the penetration of the electron beam follows Hans Bethe's energy loss equation for electron penetration in solids and thus is a function of the electron energy or CRT voltage.

Hans Bethe's energy loss equation is:

$$dE/dx = -2\Pi e^4 (N_e/E)\ln(\alpha E/I) \qquad [5]$$

where: E = energy of the incident electron; x = depth of material; N_e = density of atomic electrons; I = mean excitation energy; and α = a constant.

In the voltage penetration tube, a cascaded screen is used to give a color change when the electron voltage is increased. For example, if a cascade screen is made from a green phosphor layer placed on top of the red phosphor layer then the electrons strike the green phosphor first. Then with increased voltage, the electrons pass through the top phosphor and excite the bottom red phosphor. In the current sensitive color tube, the phosphor screen is composed of (for example) red and green phosphor whose luminance versus current curves are different. In these cases, we are referring to the linearity of the change in luminescence from the phosphor as a result of a change in electron beam current. The red phosphor (generally a rare earth oxide) is a linear phosphor, whereas the green is nonlinear. Changing the current thus can change the color of the display. In the beam index tube, a UV (or secondary electron emitters have also been used) emitting index layer is put along side a red, green, or blue triad (for example, between the blue and the red phosphor stripes). The scanning electron beam, through the use of a UV sensor (or secondary electron detector) mounted in the rear portion of the tube and the scanning electronics, can determine the location of the red, green, and blue phosphors. The electron beam can then excite only the individual phosphor areas of interest. With this tube, generally, the

electron beam is at a low current to give a small enough cross-section to only excite one color and thus has limited application.

Another application of the monochrome is the projection CRT system. Here, three tubes for red, green, and blue are used to project an image on a viewing screen. The terminology is such that, when the CRTs are in the back of the screen and the viewer in front of the screen, it is called a 'rear' projection set and if the CRTs and the viewer are in front of the screen it is a 'front' projection set. The principle here is that, either using one CRT tube at a time or by a superposition two or three of the CRTs, the color of the displayed image will be within the color gamut defined by the primary colors.

The Color CRT

The more ubiquitous method of obtaining a color display is the device called the shadow mask CRT (and the similar Trinitron CRT). This color CRT is designed in such a manner that through the use of an aperture mask, (seen in **Figure 1**) the electrons from the red electron gun strike the red phosphor and likewise electrons from the green and blue electron guns strike the green and blue phosphors. The combination allows the viewer to see a color image. This color separation action is seen in **Figure 2**. Although **Figure 2** shows a delta tri-color design, there are also in-line systems in which the mask consists of an array of slots which are somewhat rectangular in nature. In this array, many slots can be on a line, or a mask made can be made of a series of wires as used in the Trinitron design.

The disadvantage of the shadow mask technique is that the mask can capture as much as 80% of the electrons from the electron guns. This yields lower screen luminance and lower contrast due to scattered electrons. As the screen size gets larger from 19 inches to 36 inches or more, additional temperature compensation methods must be used to reduce the effects of mask heating. A number of methods to control the shift of the electron beam on the phosphor as the mask heats up have been developed. Initially, axially positioned bimetal springs were used on masks to compensate for the heating effect. Then lighter masks were used with corner lock suspension systems. Another method is using mask materials with lower coefficients of expansion such as INVAR.

Contrast in color CRTs has been improved by first applying a carbon coating between the panel glass and the phosphor screen. This requires additional process steps which form 'windows' in the carbon coating (matrix) through which the excited phosphor is seen. The use of pigmented phosphors also helped to reduce the reflectivity of the phosphor/matrix screen.

The phosphors used in the shadow mask color CRT are the red Y_2O_2S:Eu or, in some cases, the Y_2O_3:Eu phosphor. The green phosphor is ZnS:Cu and the blue phosphor is ZnS:Ag.

The CRT seen in **Figure 3** shows the new high-definition television-type (HDTV) tube with the 16×9 aspect ratio as contrasted with the older 4×3 aspect ratio (screen width $\times$ height).

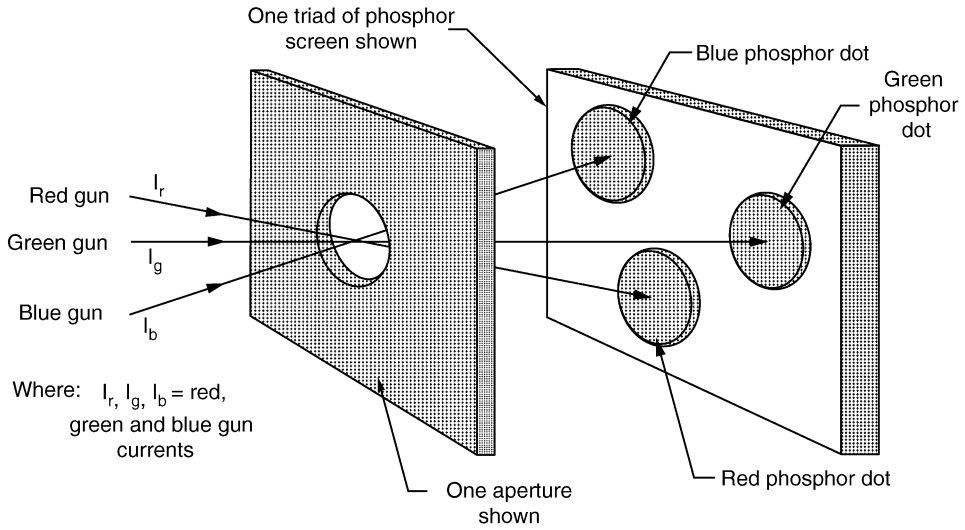


Figure 2 The action of a shadow mask.



Figure 3 A typical HDTV color CRT. Reproduced from Donofrio RL (2002) The future of direct-view CRTs. *SID Information Display Magazine* 6: 10. Permission for reprint courtesy of the Society for Information Display.

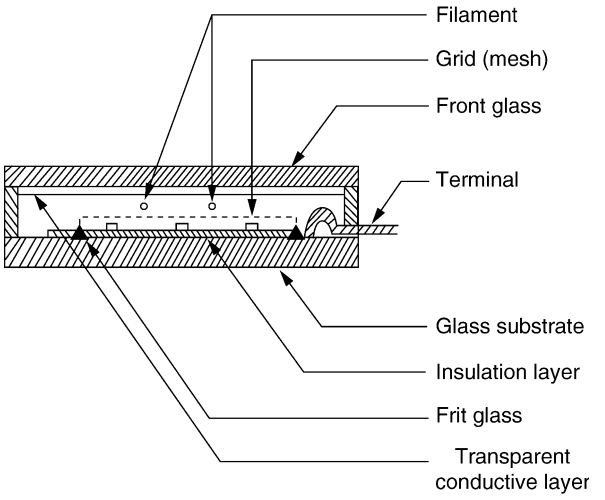


Figure 4 The cross-section of a vacuum fluorescent display assembly.

The Vacuum Fluorescent Display (VFD)

The VFD can be thought of as a low-voltage CRT. Just as in the CRT, electrons from a cathode are controlled by a grid which allows them to be accelerated and strike the phosphor screen anode. However, the electrons in a CRT can be accelerated with a voltage of 30 000 V, whereas the VFD accelerates the electrons with voltages of 12 to 40 V.

In the operation of the VFD, electrons are emitted by the filament. The filament is a very thin tungsten wire coated with a tri-carbonate oxide (barium, strontium, and calcium carbonates) which are the same constituents of the CRT oxide cathode. The filament is held between a filament support and an anchor which applies tension to the filament wire. A current is passed through the filament to achieve a temperature to ~600 °C. At that temperature, thermonic emission occurs. Between the filament and the phosphor elements is a stainless steel mesh grid. By the application of a positive or negative voltage on this grid, electrons can either be cut off or accelerated to the phosphor on the anode. The anode can be a conductor such as graphite, coated with a phosphor layer such as ZnO:Zn (zinc oxide). This is a low-voltage cathodoluminescent phosphor which emits a green color with an emission maxima of about 505 nm. Other phosphors are also used to obtain red, orange, and blue (Figure 4).

In addition to the filament, grid and anode, the VFD is a vacuum device (like the CRT) and contains a getter to assist achieving and maintaining a good vacuum and an internal transparent conductive coating to protect the display from external electrostatic fields. The VFD display is used frequently in the auto industry and aging of the phosphor is very small, achieving 100 000 hrs to half of the initial luminance.

VFD units have been made with phosphor elements on either side of the filament. This gives rise to a 3D effect. Graphic and fixed segment VFDs have been made for multimedia products and 16 × 16 graphics for Kanji and other characters.

Experimentally, 2D VFD arrays have been made similar to the FEDs covered in the next section.

Field Emissive Displays (FEDs)

The FED is similar to the CRT and the VFD in that electrons from a source bombard a phosphor screen. In both the CRT and the VFD thermionic electrons are used. However, in the FED the electrons arise from field emission. Field emission devices create electrons though quantum mechanical tunneling from the emitter surface into the vacuum by the action of a large electrical field. This process is governed by the Fowler–Nordheim equation. This equation takes the form:

$$J = A(\beta V)^2/\Phi \exp(-B\Phi^{3/2}/\beta V) \qquad [6]$$

where, J is the electron current density; V is the applied voltage; Φ is the work function of the emitting tip; A & B are constants; and β is the field enhancement factor.

The emitters can be of the Molybdenum Spindt type and also the polydiamond or diamond-like carbon (DLC) coated emitters. We have the field enhancement with a sharp tip and reduced work function with the DLC. There has been renewed interest in FEDs, with the development of the 'carbon nanotube' as an electron field emitting source. The carbon nanotube is a fullerene material, which is a distinct form of carbon. This 'bucky ball'-like

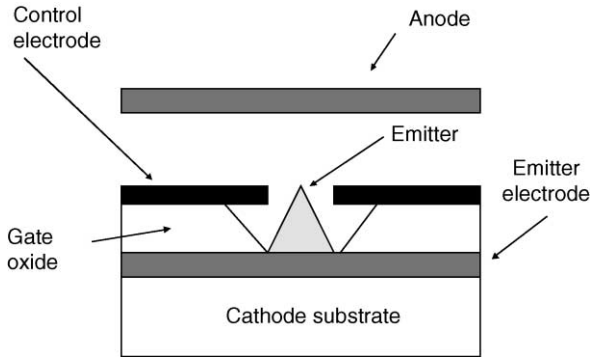


Figure 5 Field emission display cross-section. Reproduced from Green P and Haven D (1998) Fundamentals of emissive displays. *SID Short Course S-3*. Permission for reprint courtesy of the Society for Information Display.

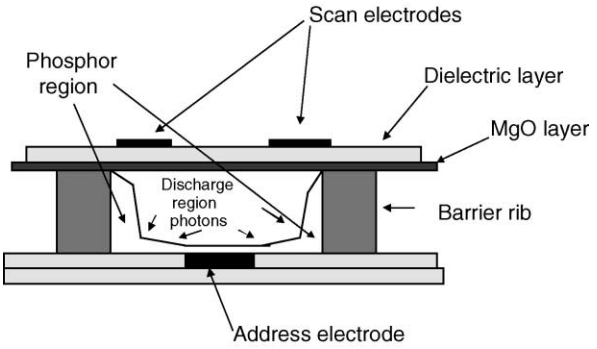


Figure 6 ACPDP partial cross-section view. Reproduced from Green P and Haven D (1998) Fundamentals of emissive displays. *SID Short Course S-3*. Permission for reprint courtesy of the Society for Information Display.

material is cylindrical instead of spherical in nature, has great strength, and high conductivity.

There are two types of FEDs, a low-voltage (LVFED) and a high-voltage (HVFED). The HVFED uses an aluminum film over the phosphor to protect the phosphor screen from aging. The HVFED can use the high voltage (2–10 KV) electrons to bombard the aluminized screen and it can use the more efficient, somewhat standard CRT phosphors ZnS:Cu green, ZnS:Ag Blue, and Y_2O_3:Eu red. The LVFED (12–200 V) cannot use the standard CRT phosphors but uses the low-voltage phosphor such as the ZnO:Zn green phosphor. This does not use an aluminum layer, since about 2 KV is needed to penetrate the Al layer. The commonly used CRT red phosphor Y_2O_2S:Eu was found unsuitable for HVFED use due to release of the sulfur during electron bombardment.

Figure 5 shows that for this basic FED structure, the electrons from the Spindt emitter are controlled by the voltages of the anode, control electrode and emitter electrode.

Plasma Displays

Earlier plasma displays, made using excited Ne gas with the characteristic amber glow required no phosphor and were used in the printing industry. Present plasma displays used for television and information signs are photoluminescent displays. As shown in Figure 6, the excitation of the plasma gas made of elements such as He, Ne, Ar, and Xe give rise to a glow discharge and the creation of vacuum ultraviolet photons that strike the screen and cause it to luminesce. In construction, the display is similar to the FED in that it is a matrix addressed device. It has been described in literature that the similarity goes even further with the need for anode and cathode surfaces, a means of evacuating and a perimeter seal. There are two types of plasma displays, DC and AC. The AC design appears to be the more successful. Typical PDP phosphors for the triad are (Y,Gd)BO_3:Eu for the red, Zn_2SiO_4:Mn for the green, and $BaMgAl_{10}O_{17}$:Eu (called BAM for short) for the blue. Figure 6 below shows the basic structure of an AC plasma display. One phosphor element is shown with barrier ribs and address and scan electrodes. The discharge of the plasma takes place in the region of the phosphor and excites the photoluminescent phosphor, which is shaped (in this example) along the side of the barrier ribs and over the address electrode.

Electroluminescence

Electroluminescence was first seen in the 1930s by Destriau. EL is the generation of (nonthermal) light by use of an electric field. High electric field EL units emit light through the impact of high energy electrons in luminescent centers in a semi-conducting material. The electric field can be alternating current (AC) or direct current (DC). Powder phosphor AC EL displays appear to be one of the successful inorganic EL display methods. Because of their rugged nature, alternating current electroluminescence ACEL displays have been used in trucks, tractors and the M1 Abrams Tank. Thick-film EL panels are also used in some vehicle displays as back lights for LCDs. These light panels are free from mercury and are more environmentally friendly.

Inorganic LED Displays

Light-emitting diodes are being used as backlights for LCD displays, replacing some of the fluorescent light backlighting methods. They are also used in indicator lights. Large displays have been made of LEDs. Using a matrix array row and column connectors,

we can digitally select a given color LED and pattern in an LED array and generate a color image. Similar methods were carried out many years ago using incandescent lights. However, now they are bright and fast reacting. LEDs operate by semiconductor P–N junction rules and the flow of electrons and holes under forward bias operation give rise to junction light emission. Materials such as GaP:Zn,O are used to create a red LED with an emission maxima wavelength of 699 nm.

Red, green, and blue-UV/blue LEDs are now available, with an increasing demand for white LEDs. Two methods are used to produce a white LED. One method is to use a blue LED with the front of the LED coated with a green photoluminescent phosphor. If the phosphor is thin enough, the blue light penetrates the screen and excites the green to give a blue/green white. Another method is to use a blue LED and coat it with a white photoluminescent phosphor.

Organic LEDs

The basic process of light generation in OLEDs is the formation of a P–N junction. Electrons injected from the N-type material and holes from the P-type material recombine at the junction and light is generated. There are presently two competing methods to fabricate an OLED. The first developed by Kodak in 1987 is the small molecule molecular weight material approach. In OLED operation, the electron hole pair recombine to form excitons (on the average one singlet and three triplet excitons are formed for every four electron hole pairs). The light for the Kodak OLED is from the recombination of singlet excitons and the energy transferred to the fluorescent dye. The terminology is such that we have small molecule-OLEDs (sm-OLEDs) and polymer-OLED (PLED). In later developments, the sm-OLED device structure was reported to have three layers. The first is the hole transport layer (hole transport) enhancing layer copper phthalocyanine (CuPc), a layer NPB (60 nm), the light emitting aluminum tris-8-hydroxyquinolate Alq layer (75 nm), and electron transport layer. The electron transport layer is in contact with the three layer structure. The top electrode is a low work function material, such as MgAg (75–200 nm thick) called an electron injection layer and topped with a transparent indium tin oxide (ITO) layer for transparent OLEDS. All of these coatings are applied in a vacuum chamber.

In the phosphorescent material method, the light emission is from the triplet excitons through the use of phosphorescent dyes and the polymer-based devices. The common base material is a polyphenylene-vinylene PPV and other polymers such as polyfluorene have been used. The construction of the phosphorescent OLED is similar in that the ITO is also used as a starting substrate but the hole transport layer is applied by spin casting from solution. Spin casting is also used for the electron transfer material. After drying (removal of solvents) a calcium cathode material is applied. A contrast of 3000/1 has been achieved (for a military type application) using a green OLED with an AR coated circular polarizer and at an ambient light level of 500 lux, and a 6/1 contrast ratio has been achieved with an ambient illumination of 50 000 lux and an output of 370 cd/m^2. Additionally, OLEDs can be made on flexible (F) substrates and are called FOLEDs. The use of a transparent (T) cathode gives rise to a top emission OLED or a TOLED.

The first OLED display used in auto products was developed by Pioneer and used in a radio display and the Ford Motor Company has shown a White OLED display in one of their 'Concept Cars'. This display allows them to have a reconfigurable instrument

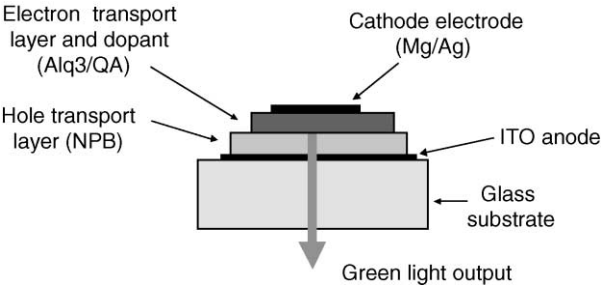


Figure 7 OLED partial cross-section – Kodak type. Reproduced from Green P and Haven D (1998) Fundamentals of emissive displays. *SID Short Course S-3*. Permission for reprint courtesy of the Society for Information Display.



Figure 8 Philips poled display in an auto mirror. Reproduced with permission of Philips Mobile Display Systems.

panel (IP) and is made up of a number of 5 inch × 8 inch displays across the instrument panel (Figure 7).

The POLED as seen in Figure 8 is now being used by Philips Mobile Systems in a rear view car mirror, as seen below. It is the display from a video camera looking at the rear of a car to see any obstacles to backing up safely.

Thin Film EL (TFEL) Displays

A simple type of TFEL display is the color by white structure shown below. In this case, the color is generated using a white phosphor and red, green, and blue optical filters. To make this ACEL display function properly, insulators are used on either side of the phosphor. In Figure 9, the ITO conducting stripes (depicted) horizontally on the top and vertically on the bottom, give the means of addressing each color pixel. The white phosphor, for example, is a ZnS:Mn, SrS:Ce blend which gives a white electroluminescence.

Other phosphors such as ZnS:Mn can be used but its yellow luminescence requires the use of a green and red filter and only two metal electrodes on the bottom to give a two-color display. Better red green color separation can be achieved through the use of a $Zn_{1-x}Mg_xS$:Mn and ZnS:Mn phosphor mix. New EL blue phosphor developments have allowed larger color gamuts to be achieved.

One of the important characteristics of the alternating current thin-film electroluminescence (ACTFEL) is that it has an operating temperature range from −40 °C to 85 °C which is key for automotive use. It has a viewing angle of 160 degrees with a life of more than 50 000 hours. Active matrix (active addressed) AMEL structures allow fast

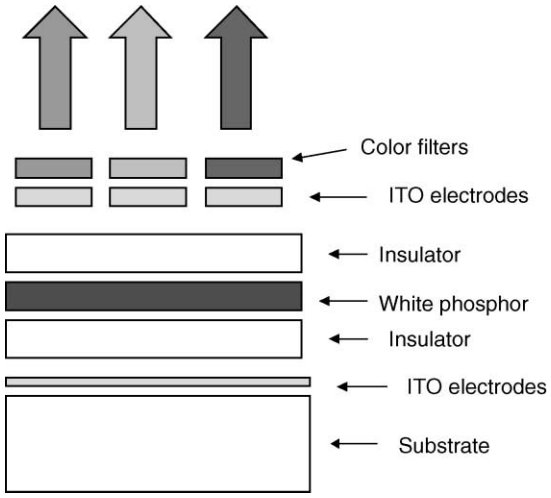


Figure 9 Color thin film electroluminescent display (TFEL) cross-section. Reproduced from Dakkala A (1998) New EL displays for vehicular applications. In: Donofrio RL and Musa S (eds) *Flat Panel and Vehicle Display '98.* Permission for reprint courtesy of the Society for Information Display.

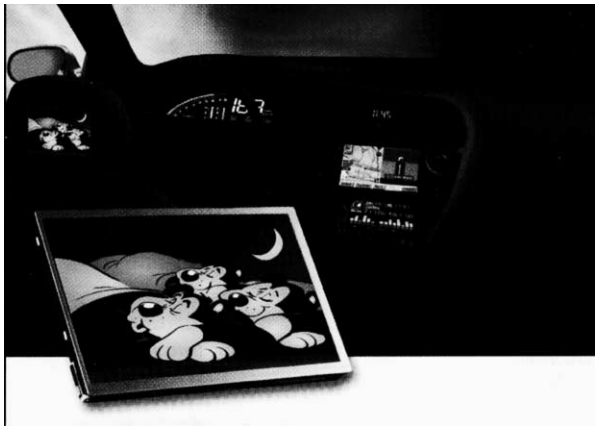


Figure 10 AMLCDs used in automobiles. Reproduced from Donofrio RL (2003) The road show. *SID Information Display Magazine* 7: 26. Permission for reprint courtesy of the Society for Information Display.

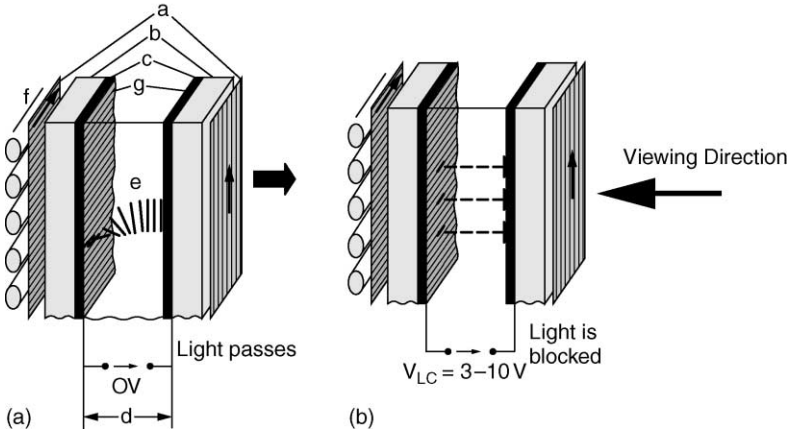


Figure 11 The operation of a simple liquid crystal display. Reproduced from Lueder E (1997) Fundamentals of passive and active-addressed liquid crystal displays. *SID Short Course S-1* (1997). Permission for reprint courtesy of the Society for Information Display.

response and moving graphic displays. These have found application in 'heads up displays' (HUD) for the military and medical industries and in consumer goods such as portable computers, pagers, cell phones, and 3D gaming.

The Active Matrix LCD

The active matrix liquid crystal display (AMLCD) device consists of a series of liquid crystal cells or windows whose characteristics can be controlled by an applied voltage to pass or restrict the passage of light through the windows via the external polarizers and the polarizing action of the LCD cells. These displays are used in laptop and tabletop computers, TVs, and for automotive displays (for driving information and entertainment). Figure 10 shows some present applications for AMLCDs in the automobile.

When the simple LCD is used without a transistor matrix, it is termed a passive display rather than an active one. Figure 11 below shows how a simple LCD cell operates.

The action shown in this figure is the twisted alignment nature of the liquid crystal material. When the polarizer on the left has its high transmission axis horizontal and the polarizer on the right of this cell has perpendicular alignment then the light from the backlight (f) is passed through the crossed polarizers. However, when voltage is applied to the cell, the liquid crystal is no longer twisted and the crossed polarizers prevent the light from going though the cell to the viewer. If the polarizers were not crossed but the polarization transmission axis was in the same direction in both polarizers, then the opposite effect would occur.

In order to show a color image, the LCD is designed to have three or more cells clustered per color pixel. The following (Figure 12) is the transmission profiles of each of these red, green, and blue filters.

The backlight used is a fluorescent lamp with a white phosphor (made up of red, green, and blue photoluminescent phosphors). Figure 13 shows what would happen if only the blue phosphor was used and also shows the superposition of the mercury lines that penetrate the phosphor and contribute to the final spectral energy distribution.

Reflective LCD

For applications where there is suitable external light to view the display, no internal light source is used.

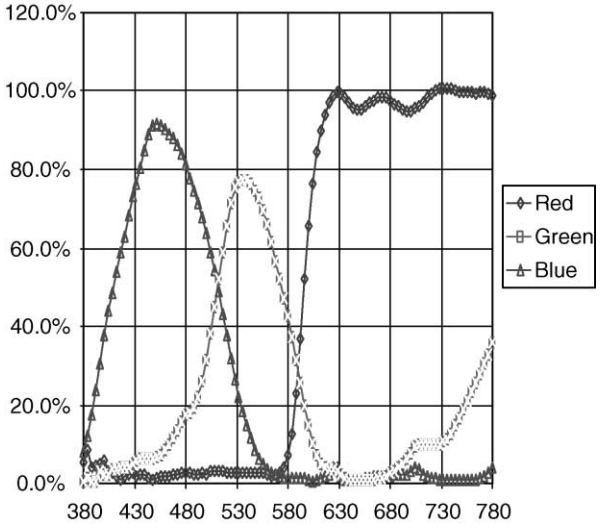


Figure 12 Typical color LCD filter transmission characteristics. Reproduced from Donofrio RL and Abileah A (1998) LCD backlighting system – phosphors and colors. In: Donofrio RL and Musa S (eds) *Flat Panel and Vehicle Display '98, p. 123*. Permission for reprint courtesy of the Society for Information Display.

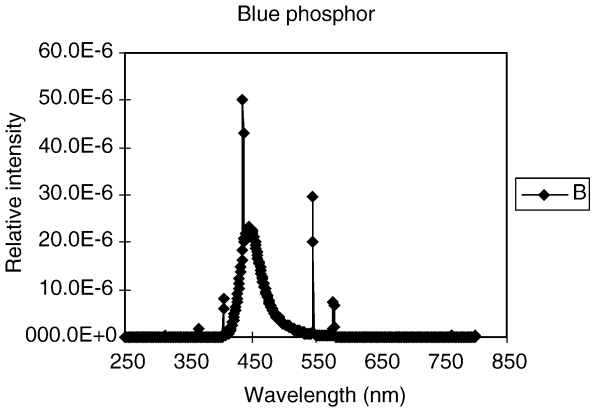


Figure 13 The relative luminance spectrum of the LCD. Backlight blue phosphor showing mercury lines. Reproduced from Donofrio RL and Abileah A (1998) LCD backlighting system – phosphors and colors. In: Donofrio RL and Musa S (eds) *Flat Panel and Vehicle Display '98, p. 123*. Permission for reprint courtesy of the Society for Information Display.

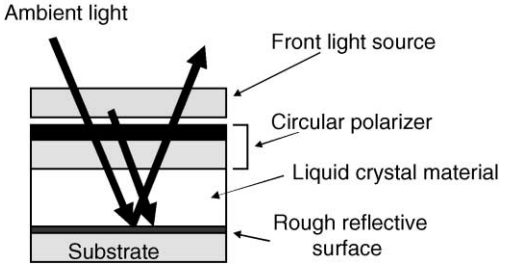


Figure 14 The action and construction of a reflective LCD. Reproduced with permission from Watanabe R and Tomita O (2003) Active-Matrix LCDs for Mobile Telephones in Japan. *SID Information Display Magazine* 7(3): 30–35. Permission for reprint courtesy of the Society for Information Display.

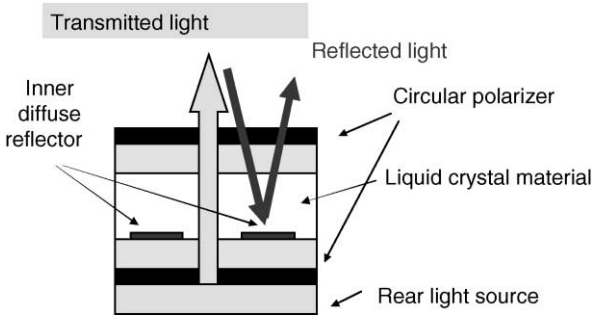


Figure 15 The action and cross-section of a transflective LCD. Reproduced with permission from Watanabe R and Tomita O (2003) Active-Matrix LCDs for Mobile Telephones in Japan. *SID Information Display Magazine* 7(3): 30–35. Permission for reprint courtesy of the Society for Information Display.

This reflective LCD consists of a polarizer and retarder plates stacked on the front side of the cell. The cell with spacers contains the liquid crystal material and on the bottom of the cell is located a reflecting surface. It has been found that a dimpled reflecting surface makes a diffuse reflecting surface (Figure 14).

Transflective LCD

The importance of the transflective device is that it has better contrast under high ambient illumination than the standard transmission back lighted LCD and fair contrast for low ambient illumination.

The construction of this device consists of front and rear circular polarizers, back light, reflective elements, and novel filters. The back light is in the rear of the display. However, like the reflective LCD, internal reflectors are used on each pixel. The filters are designed spatially to be more transmissive for the reflective light than for the back light. Additionally the cell gap for the reflective light is smaller then the transmitted light region because the reflected light passes the cell twice whereas the transmitted light passes the cell once. This type of device seen in Figure 15, is finding uses in the auto industry and where the LCD is needed for both low and high ambient light levels.

Conclusions

In this brief review of displays we have tried to cover many different types of displays. The reader will in time find that there are many other variations on the displays discussed. The object is to give the reader an up-to-date review of the topics involved with displays and the various types of display in use today. By its nature it is a broad approach.

See also

Color and the World. **Instrumentation:** Photometry. **Light Emitting Diodes**. **Materials for Nonlinear Optics:** Liquid Crystals for NLO.

Further Reading

Donofrio RL (1981) Low Current Density Aging. *Proceedings of the Electro-Chemical Society*, Abstract #152, May 11, 1981.

Donofrio RL (2003) The road show. *Information Display* 7(3): 26–29.

Donofrio RL and Abileah A (1998) LCD backlighting. *Flat Panel and Vehicle Displays '98* : 119–124. Metropolitan Detroit Chapter of the SID.

Donofrio RL, Musa S and Wise K (eds) (1999) Vehicle Displays and Microsensors '99. *Proceedings of the 6th Strategic Forum and Technical Symposium*, Sept 22–23, 1999.

Flat Panel Display Measurements Standard, Version 1.0 VESA (Video Electronics Standards Association) May 15, 1998.

Forrest S (2002) *Fundamentals of Organic Light-Emitting Devices*. Short Course S-2 SID.

Green P and Haven D (1998) *Fundamentals of Emissive Displays*, Short Course S-3, SID.

Keller P (1991) *The Cathode-Ray Tube, Technology, History, and Applications*. New York: Palisades Press.

Leverence HW (1968) *An Introduction to Luminescence of Solids*. Dover Publications.

MacDonald W and Lowe A (1997) *Display Systems – Design and Applications*. SID Series in Display Technology, Wiley and Sons Ltd.

Sherr S (1993) *Electronic Displays*, 2nd edn. John Wiley.

Shionoya S and Yen WM (1998) *Phosphor Handbook*. CRC Press.

Stupp EH and Brennesholtz MS (1999) *Projection Displays*. SID Series in Display Technology, Wiley and Sons Ltd.

Watanabe R and Tomita O (2003) Active-Matrix LCDs for Mobile Telephones in Japan. *SID Information Display Magazine* 7(3): 30–35.

Web site atis.org/tg2k/. American National Standard TI.523-2001 'Def of Display Device'.

Wu ST and Yang DK (2001) *Reflective Liquid Crystal Displays*. SID Series in Displays Technology, Wiley and Sons Ltd.

E

ELECTROMAGNETICALLY INDUCED TRANSPARENCY

J P Marangos, Imperial College of Science, Technology and Medicine, London, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In this article we examine how the process of electromagnetically induced transparency (EIT) can be used to increase the efficiency of non-linear optical frequency mixing schemes. We illustrate the basic ideas by concentrating upon the enhancement of resonant four-wave mixing schemes in atomic gases. To start we introduce the quantum interference phenomenon that gives rise to EIT. The calculation of EIT effects in a three-level atomic medium using a density matrix approach is presented. Next we examine the modified susceptibilities that result and that can be used to describe nonlinear mixing, and show how large enhancements in nonlinear conversion efficiency can result. Specific examples are briefly discussed along with a further discussion of the novel aspects of refractive index and pulse propagation modifications that result from EIT. The potential benefits to nonlinear optics of electromagnetically induced transparency are discussed. In an article of this nature it is not possible to credit all of the important contributions to this field, but we include a bibliography for further reading indicating some of the seminal contributions.

Electromagnetically induced transparency is the term applied to the process by which an otherwise opaque medium can be rendered transparent through laser-induced quantum mechanical interference processes. The linear susceptibility is substantially modified by EIT. Absorption is cancelled due to destructive interference between excitation pathways. The dispersion of the medium is likewise modified such that where there is no absorption the refractive index is unity and there can be a large value of normal dispersion leading to low values of group velocity. In contrast to the linear susceptibility the nonlinear susceptibility involving these same resonant laser fields will undergo constructive rather than destructive interference. This leads to a strong enhancement in nonlinear frequency mixing. For a normally transparent medium the nonlinear optical couplings will be small unless large-amplitude fields are applied. The most important consequence of EIT, in contrast to the usual case when a medium is transparent, is that the dispersion is not vanishing and the nonlinear couplings can be very large.

To understand how this comes about we must consider the system illustrated in Figure 1a where because of the resonant condition satisfied for the applied fields the atom can be simplified to just three-levels. The important parameters in this model system are that state $|2\rangle$ is metastable and has a dipole-allowed coupling to state $|3\rangle$. The coupling between states $|1\rangle$ and $|3\rangle$ is also dipole-allowed

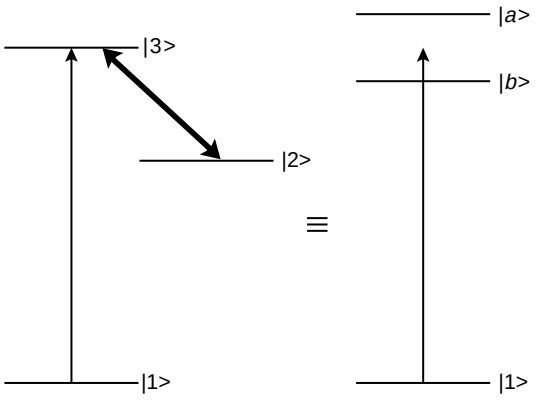


Figure 1 A three-level atomic system coupled to laser fields in the lambda configuration is shown on the left-hand side. In the limit of a strong coupling field this is equivalent to the dressed states $|a\rangle$ and $|b\rangle$ coupled to the ground state by the probe field alone as shown on the right-hand side.

and as a consequence state |3> can decay radiatively to either |1> or |2>. Let a coupling field be applied resonantly to the |2> − |3> transition, which may be cw or pulsed but should in the latter case be transform-limited. We define the Rabi coupling as $\Omega_{23} = \frac{\mu_{23}E}{\hbar}$ where E is the laser electric field amplitude and μ_{23} is the dipole moment of the transition. The Rabi frequency needs to be larger than the inhomogeneous broadening of the sample. A second field, the probe, typically of much lower intensity, is then applied in resonance with the |1> − |3> transition. If the condition $\Omega_{23} \gg \Omega_{13}$ is satisfied then it is convenient to replace the bare atomic states |2> and |3> with the dressed states (see Figure 1b):

$$|a> = \frac{1}{\sqrt{2}}[|2> + |3>] \qquad [1a]$$

$$|b> = \frac{1}{\sqrt{2}}[|2> - |3>] \qquad [1b]$$

Transitions from state |1> induced by the probe field to this pair of near-resonant dressed states are subject to exact cancellation at resonance if |2> is perfectly metastable. This is because the only contribution to the excitation amplitude comes from the matrix elements $<1|\mathbf{r}\cdot\mathbf{E}|3>$ as the |1> − |2> transition is dipole forbidden. The contributions from the equally detuned states |a> and |b> thus enter into the overall amplitude with opposite signs and equal magnitude as can be seen by inspection of eqns [1a] and [1b]. This leads to cancellation of the absorption amplitude. This type of absorption cancellation is well known and closely related to the so-called *dark states*.

Theoretical Treatment of EIT in a Three-Level Medium

It was realized by several workers in the 1970s that laser-induced interference effects could lead to a cancellation of absorption at certain frequencies. To gain a more quantitative understanding of the effects of the coupling field upon the optical properties of a dense ensemble of three-level atoms we require a treatment that computes the optical susceptibility of the medium. A treatment originally carried out by Harris *et al.* for a Λ scheme similar to that illustrated in Figure 1 was the first to derive the modified susceptibilities that will be discussed below. In that treatment the state amplitudes in the three-level atom were solved in the steady-state limit and from these the linear and nonlinear susceptibilities (see below) are then derived. In what follows we adopt a density matrix treatment as employed by various workers. This readily allows the inclusion of dephasing as well as population decay terms. The critical parameters in this system are the strengths of the fields (described in terms of the Rabi couplings), the detuning of the fields from resonance $\Delta\omega_{13} = \omega_{13} - \omega_P$ and $\Delta\omega_{23} = \omega_{23} - \omega_C$ (see Figure 2), the radiative decays from |3> to |1> and |2>, γ_c and γ_d and from |2> − |1> γ_a (although the latter is anticipated to be smaller). Extension to other configurations of the three states, such as a V or ladder scheme is implicit within this general treatment.

Figure 2 illustrates the system considered. This treatment will also address the nonlinear response in a four-wave mixing scheme completed by the two-photon resonant coupling applied between state |1> and |2>.

We write the interaction Hamiltonian as:

$$H = H_0 + V \qquad [2]$$

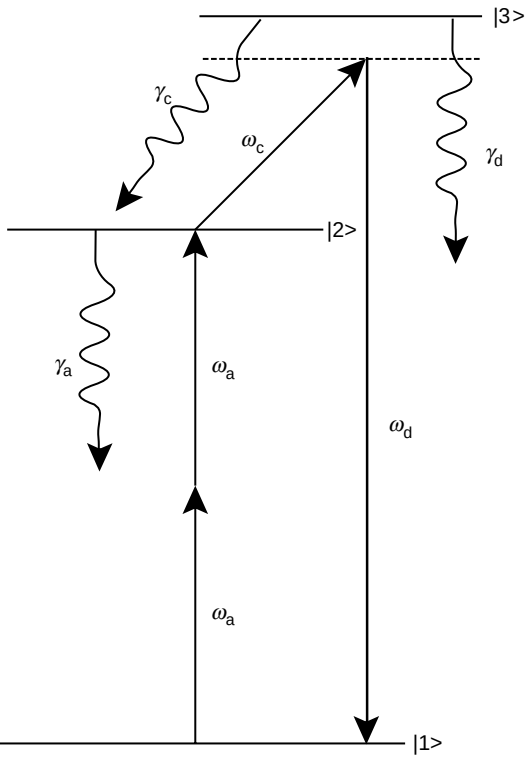


Figure 2 A four-wave mixing scheme incorporating the three-level lambda system in which electromagnetically induced transparency has been created for the generated field ω_d by the coupling field ω_c. The decay rates γ_i from the three atomic levels are also shown. For a full explanation of this system see text.

where H_0 is the unperturbed Hamiltonian of the system and is written as

$$H_0 = \hbar\omega_1|1\rangle\langle 1| + \hbar\omega_2|2\rangle\langle 2| + \hbar\omega_3|3\rangle\langle 3| \quad [3]$$

and V is the interaction Hamiltonian and can be expressed

$$\begin{aligned} V = {} & \hbar\Omega_a e^{-i\omega_a t}|2\rangle\langle 1| + \hbar\Omega_c e^{-i\omega_c t}|2\rangle\langle 3| \\ & + \hbar\Omega_d e^{-i\omega_d t}|3\rangle\langle 1| + c.c \end{aligned} \quad [4]$$

Note that the Rabi frequencies in the equation can be described as $\hbar\Omega_{ij} = \mu_{ij}|E(\omega_{ij})|$, where μ_{ij} is the dipole moment of the transition between states |i> and |j>, the Rabi frequency Ω_a is a two-photon Rabi frequency that characterizes the coupling between the laser field *a* and the atom for this two-photon transition. We have assumed for simplicity that $\omega_a = \omega_b = \omega_{12}/2$.

Assuming all the fields lie close to the resonance, the rotating wave approximation can be applied to the interaction picture and the interaction Hamiltonian V^I is given as

$$\begin{aligned} V^I = {} & \hbar\Omega_a e^{i\Delta_a t}|2\rangle\langle 1| + \hbar\Omega_c e^{i\Delta_c t}|2\rangle\langle 3| \\ & + \hbar\Omega_d e^{i\Delta_d t}|3\rangle\langle 1| + c.c \end{aligned} \quad [5]$$

where Δ_a, Δ_c and Δ_d refer the detunings of the fields and can be written as:

$$\begin{aligned} \Delta_a &= \omega_{12} = 2\omega_a \\ \Delta_c &= \omega_{32} - \omega_c \\ \Delta_d &= \omega_{13} - \omega_d \end{aligned} \quad [6]$$

For the evaluation of the density matrix with this interaction V^I, the Schrödinger equation can be restated in terms of the density matrix components. This form is called the Liouville equation and can be written as:

$$\hbar\frac{\partial}{\partial t}\varrho_{ij}(t) = -i\sum_k H_{ik}(t)\varrho_{ij}(t) + i\sum_k \varrho_{ij}(t)H_{ik}(t) + \Gamma_{ij} \quad [7]$$

where Γ_{ij} represents phenomenologically added decay terms (i.e. spontaneous decays, collisional broadening, etc.). This formalism leads to a set of nine differential equations for nine different density matrix elements that describe the three-level system.

To remove the optical frequency oscillations, a coordinate transform is needed and to incorporate the relevant oscillatory detuning terms into the off-diagonal elements we make the substitution:

$$\begin{aligned} \tilde{\varrho}_{12} &= \varrho_{12}e^{-i\Delta_a t} \\ \tilde{\varrho}_{23} &= \varrho_{23}e^{-i\Delta_c t} \\ \tilde{\varrho}_{31} &= \varrho_{31}e^{-i\Delta_d t} \end{aligned} \quad [8]$$

This operation conveniently eliminates the time dependencies in the rate equations and the equations of motion for the density matrix are given by:

$$\begin{aligned} \frac{\partial}{\partial t}\varrho_{11} &= i\frac{1}{2}\Omega_a\tilde{\varrho}_{12} + i\frac{1}{2}\Omega_d\tilde{\varrho}_{13} - i\frac{1}{2}\Omega_a^*\tilde{\varrho}_{21} \\ &\quad - i\frac{1}{2}\Omega_d^*\tilde{\varrho}_{31} + \Gamma_{11} \\ \frac{\partial}{\partial t}\varrho_{22} &= i\frac{1}{2}\Omega_c^*\tilde{\varrho}_{23} + i\frac{1}{2}\Omega_a^*\tilde{\varrho}_{21} - i\frac{1}{2}\Omega_c\tilde{\varrho}_{32} \\ &\quad - i\frac{1}{2}\Omega_a\tilde{\varrho}_{12} + \Gamma_{22} \\ \frac{\partial}{\partial t}\varrho_{33} &= i\frac{1}{2}\Omega_c^*\tilde{\varrho}_{23} - i\frac{1}{2}\Omega_d^*\tilde{\varrho}_{31} + i\frac{1}{2}\Omega_d\tilde{\varrho}_{13} \\ &\quad - i\frac{1}{2}\Omega_c\tilde{\varrho}_{32} + \Gamma_{33} \\ \frac{\partial}{\partial t}\varrho_{23} &= i\frac{1}{2}\Omega_c\tilde{\varrho}_{22} - i\frac{1}{2}\Omega_c\tilde{\varrho}_{33} - i\Delta_a\tilde{\varrho}_{23} \\ &\quad + i\frac{1}{2}\Omega_d\tilde{\varrho}_{21} - i\frac{1}{2}\Omega_a\tilde{\varrho}_{13} + \Gamma_{23} \\ \frac{\partial}{\partial t}\varrho_{21} &= i\Omega_a\tilde{\varrho}_{22} + i\Omega_a\tilde{\varrho}_{33} + i\Omega_d\tilde{\varrho}_{23} - i\Delta_c\tilde{\varrho}_{21} \\ &\quad - i\frac{1}{2}\Omega_c\tilde{\varrho}_{31} - i\frac{1}{2}\Omega_a\tilde{\varrho}_{31} + \Gamma_{21} \\ \frac{\partial}{\partial t}\varrho_{31} &= i\frac{1}{2}\Omega_d\tilde{\varrho}_{22} + i\Omega_d\tilde{\varrho}_{33} + i\Omega_d\tilde{\varrho}_{23} - i\Delta_c\tilde{\varrho}_{21} \\ &\quad - i\frac{1}{2}\Omega_c\tilde{\varrho}_{31} - i\frac{1}{2}\Omega_a\tilde{\varrho}_{31} + \Gamma_{31} \end{aligned} \quad [9]$$

Using the set of equations above and the relation $\tilde{\varrho}_{ij} = \tilde{\varrho}_{ij}^*$ we obtain equations for $\tilde{\varrho}_{12}$, $\tilde{\varrho}_{32}$, and $\tilde{\varrho}_{13}$. For the incoherent population relaxation the decays can be written:

$$\begin{aligned} \Gamma_{11} &= \gamma_a\varrho_{22} + \gamma_d\varrho_{33} \\ \Gamma_{22} &= -\gamma_a\varrho_{22} + \gamma_c\varrho_{33} \\ \Gamma_{33} &= -(\gamma_c + \gamma_d)\varrho_{33} \end{aligned} \quad [10]$$

and for the coherence damping:

$$\begin{aligned} \Gamma_{21} &= -\left\{\frac{1}{2}[\gamma_a + \gamma_c] + \gamma_{21}^{col}\right\}\varrho_{21} \\ \Gamma_{23} &= -\left\{\frac{1}{2}[\gamma_a + \gamma_c + \gamma_d] + \gamma_{23}^{col}\right\}\varrho_{23} \\ \Gamma_{31} &= -\left\{\frac{1}{2}[\gamma_d + \gamma_c] + \gamma_{31}^{col}\right\}\varrho_{31} \end{aligned} \quad [11]$$

where γ_{ij}^{col} represents collisional dephasing terms which may be present.

This system of equations can be solved by various analytical or numerical methods to give the individual density matrix elements. Analytical solutions are possible if we can assume for instance that $\Omega_c \gg \Omega_a$, Ω_d and that there are continuous-wave fields so a steady-state treatment is valid. For pulsed fields or in the case where the generated field may be significant numerical solutions are in general required.

We are specifically interested in the optical response to a probe field at ω_d (close to resonance with the $|1> - |3>$ transition) that is governed by the magnitude of the coherence ϱ_{13}. We find ϱ_{13} making the assumption that only the coupling field is strong, i.e., that $\Omega_c \gg \Omega_a, \Omega_d$ holds. From this quantity the macroscopic polarization is obtained from which the susceptibility can be computed. The macroscopic polarization P at the transition frequency ω_{13} can be related to the microscopic coherence ϱ_{13} via the expression:

$$P_{13} = N\mu_{13}\varrho_{13} \qquad [12]$$

where N is the number of equivalent atoms in the ground state within the medium, and μ_{13} is the dipole matrix element associated with the transition. In this way real and imaginary parts of the linear susceptibility χ at frequency ω can be directly related to ϱ_{13} via the macroscopic polarization since the latter can be defined as:

$$P_{13}(\omega) = \varepsilon_0 \chi(\omega) E \qquad [13]$$

where E is the electric field amplitude at frequency ω_d. The linear susceptibility (real and imaginary parts) is given by the following expressions:

$$\mathrm{Re}\chi_D^{(1)}(-\omega_d,\omega_d) = \frac{|\mu_{13}|^2 N}{\varepsilon_0 \hbar} \times \left[\frac{-4\Delta_{21}(|\Omega_2|^2 - 4\Delta_{21}\Delta_{32}) + 4\Delta_{31}\gamma_a^2}{4\Delta_{31}\Delta_{21} - \gamma_d\gamma_a - |\Omega_2|^2)^2 + 4(\gamma_d\Delta_{21} + \gamma_a\Delta_{31})^2} \right] \qquad [14]$$

$$\mathrm{Im}\chi_D^{(1)}(-\omega_d,\omega_d) = \frac{|\mu_{13}|^2 N}{\varepsilon_0 \hbar} \times \left[\frac{8\Delta_{21}^2\gamma_d + 2\gamma_a(|\Omega_2|^2 + \gamma_d\gamma_a)}{(4\Delta_{31}\Delta_{21} - \gamma_d\gamma_a - |\Omega_2|^2)^2 + 4(\gamma_d\Delta_{21} + \gamma_a\Delta_{31})^2} \right] \qquad [15]$$

We now consider the additional two-photon resonant field ω_a. This leads to a four-wave mixing process that generates a field at ω_d (i.e., the probe frequency). The nonlinear susceptibility that describes the coupling of the fields in a four-wave mixing process is given by the expression:

$$\chi_D^{(3)}(-\omega_d,\omega_a,\omega_a,\omega_c) = \frac{\mu_{23}\mu_{13}N}{6\varepsilon_0\hbar^3((\Delta_{21} - j\gamma_a/2)(\Delta_{31} - j\gamma_d/2) - (|\Omega_c|^2/4))^*} \times \sum_i \mu_{i1}\mu_{i3}\left(\frac{1}{\omega_i - \omega_a} + \frac{1}{\omega_i - \omega_b}\right)$$

Nonlinear Optical Processes

The dependence of the susceptibilities upon the detuning is plotted in Figure 3. In this plot the effects of inhomogeneous (Doppler) broadening are

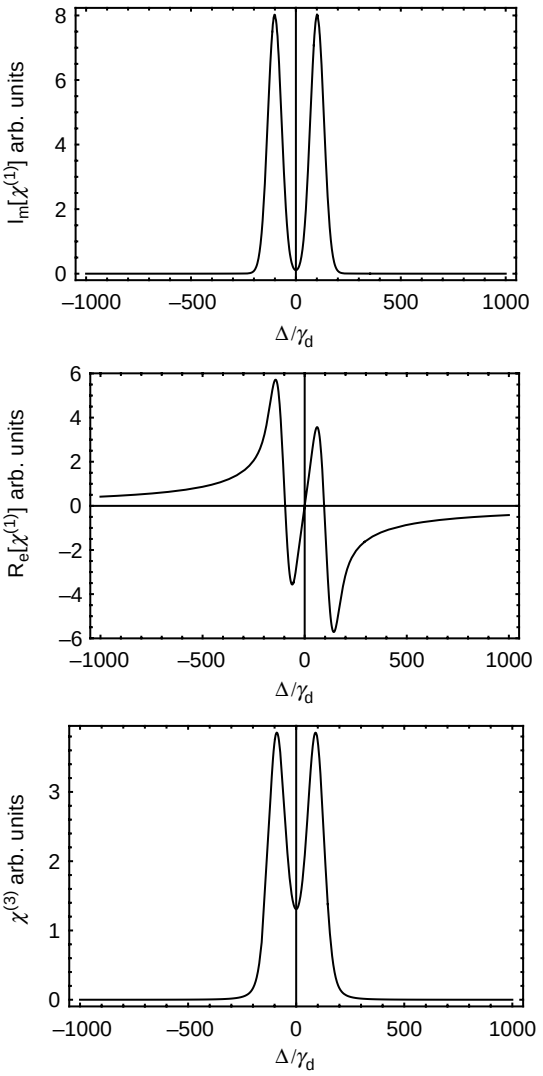


Figure 3 Electromagnetically induced transparency is shown in the case of significant Doppler broadening. The Doppler averaged values of the imaginary $\mathrm{Im}[\chi^{(1)}]$ and real $\mathrm{Re}[\chi^{(1)}]$ parts of the linear susceptibility and the nonlinear susceptibility $\chi^{(3)}$ are plotted in the three frames. The Doppler width is taken to be $20\gamma_d$, $\Omega_c = 100\gamma_d$ and $\gamma_d = 50\gamma_a$, $\gamma_c = 0$.

also incorporated by computing the susceptibilities over the inhomogeneous profile (see below for more details).

By inspection of eqn [14] we see that the absorptive loss at the minimum of $\mathrm{Im}[\chi^{(1)}]$ varies as γ_a/Ω_c^2. This dependence is a consequence of the interference discussed above. In the absence of interference (i.e., just simply Autler–Townes splitting) the minimum loss would vary as γ_d/Ω_c^2. Since $|3> - |1>$ is an allowed decay channel (in contrast to $|2> - |1>$) it follows that $\gamma_b \gg \gamma_a$ and so the absorption is much less as a consequence of EIT. $\mathrm{Re}[\chi^{(1)}]$ in eqn [15] and Figure 3 is also modified significantly. The resonant value of refractive index is equal to the vacuum value where the absorption is a minimum, the dispersion is normal in this region with a gradient determined by the strength of the coupling laser, a point we will return to shortly. For an unmodified system the refractive index is also unity at resonance but in that case there is high absorption and steep anomalous absorption. Reduced group velocities result from the steep normal dispersion that accompanies EIT. Inspection of the expression [16] and Figure 3 shows that $\chi^{(3)}$ is also modified by the coupling field. The nonlinear susceptibility depends upon $1/\Omega_c^2$ as is expected for a laser dressed system, however in this case there is not destructive but rather constructive interference, between the field-split components. This result is of great importance: it ensures that the absorption can be minimized at frequencies where the nonlinear absorption remains large.

As a consequence of constructive interference the nonlinear susceptibility remains resonantly enhanced whilst the linear susceptibility vanishes or becomes very small at resonance due to destructive interference. Large nonlinearity accompanying vanishing absorption (transparency) of course match conditions for efficient frequency mixing as a large atomic density can then be used. Moreover the dispersion (controlling the refractive index) also vanishes at resonance; this leads to perfect phase matching (i.e., zero wavevector mismatch between the fields) in the limit of a simple three-level system. As a result of these features a large enhancement of the conversion efficiency in this type of scheme can be achieved.

To compute the generated field strength Maxwell's equations must be solved using these expression for the susceptibility to describe the absorption, refraction, and nonlinear coupling in the medium. We will treat this within the slowly varying envelope approximation since the fields are cw or nanosecond pulses. To be specific we assume that the susceptibilities are time independent, i.e., that we are in the steady-state (cw) limit. We make the assumptions also that there is no pump field absorption and that we have plane waves. Under these assumptions the generated field amplitude A_d is given in terms of the other field amplitudes A_i by:

$$\frac{\partial}{\partial z}A_\mathrm{d} = \mathrm{i}\frac{\omega_\mathrm{d}}{4c}\chi^{(3)}A_\mathrm{a}^2A_\mathrm{c}\mathrm{e}^{-\mathrm{i}\Delta k_\mathrm{d}z} - \frac{\omega_\mathrm{d}}{2c}\mathrm{Im}[\chi^{(1)}]A_\mathrm{d} + \mathrm{i}\frac{\omega_\mathrm{d}}{2c}\mathrm{Re}[\chi^{(1)}]A_\mathrm{d} \qquad [16]$$

where the wavevector mismatch is given by:

$$\Delta k_\mathrm{d} = k_\mathrm{d} + k_\mathrm{c} - 2k_\mathrm{a} \qquad [17]$$

The wavevector mismatch will be zero on resonance for the three-level atom considered in this treatment. In fact the contribution to the refraction from all the other atomic levels must be taken into account whilst computing Δk_d and it is implicit that these make a finite contribution to the wavevector mismatch.

We can solve this first-order differential equation with the boundary condition $A_\mathrm{d}(z=0)=0$ to find the generated intensity $I(\omega_\mathrm{d})$ after a length z:

$$I(\omega_\mathrm{d}) = \frac{3n\omega_\mathrm{d}^2}{8Z_0c^2}|\chi^{(3)}|^2|A_\mathrm{a}|^4|A_\mathrm{c}|^2 \times \frac{\left\{1+\mathrm{e}^{-\frac{\omega_\mathrm{d}}{c}\mathrm{Im}[\chi^{(1)}]z} - 2\mathrm{e}^{-\frac{\omega_\mathrm{d}}{c}\mathrm{Im}[\chi^{(1)}]z}\cos\left(\Delta k + \frac{\omega_\mathrm{d}}{2c}\mathrm{Re}[\chi^{(1)}]\right)z\right\}}{\frac{\omega_\mathrm{d}^2}{4c^2}\mathrm{Im}[\chi^{(1)}]^2 + \left\{\Delta k + \frac{\omega_\mathrm{d}}{2c}\mathrm{Re}[\chi^{(1)}]\right\}^2}$$

where Z_0 is the impedance of free space. This expression is quantitatively correct for the case of the assumptions made. More generally the qualitative predictions and general scaling remain valid in the limit where the pulse duration is significantly longer than the time required to traverse the medium. Note that both real and imaginary parts of $\chi^{(1)}$ and $\chi^{(3)}$ play an important role, as we would expect for resonant frequency mixing. The influence of that part of the medium refraction which is due to other levels is contained in the terms with Δk. In the case of a completely transparent medium this becomes a major limit to the conversion efficiency.

Propagation and Wave-Mixing in a Doppler Broadened Medium

Doppler shifts arising from the Maxwellian velocity distribution of the atoms in the medium lead to a corresponding distribution in the detunings for

the various members of the atomic ensemble. The response of the medium, as characterized by the susceptibilities, must therefore include the Doppler effect by performing a weighted sum over possible detunings. The weighting is determined by the Gaussian form of the Maxwellian velocity distribution. From this operation the effective values of the susceptibilities at a given frequency are obtained, and these quantities can be used to calculate the generated field. This step is of considerable practical importance as in most up-conversion schemes it is not possible to achieve Doppler-free geometries and the use of laser cooled atoms, although in principle possible, limits the atomic density that can be employed. The interference effects persist in the dressed profiles providing the coupling laser Rabi frequency is comparable to or larger than the inhomogeneous width. This is because the Doppler profile follows a Gaussian distribution which falls off much faster in the wings of the profile than the Lorentzian profile due to the natural broadening.

In the case considered with weak probe field, excited state populations and coherences remain small. The two-photon transition need not be strongly driven (i.e., a small two-photon Rabi frequency can be used) but a strong coupling laser is required. The coupling laser must be intense enough that its Rabi frequency is comparable to or exceeds the inhomogenous widths in the system (i.e., Doppler width), and a laser intensity of above 1 MW cm^{-2} is required for a typical transition. This is trivially achieved even for unfocused pulsed lasers, but does present a serious limit to cw lasers unless a specific Doppler-free configuration is employed. The latter is not normally suitable for a frequency up-conversion scheme if a significant up-conversion factor is required, e.g., to the vacuum ultraviolet (VUV); however recent experiments report significant progress in cw frequency up-conversion using EIT and likewise a number of other possibilities, e.g., laser-cooled atoms and standing-wave fields, have been proposed.

A transform-limited single-mode laser pulse is essential for the coupling laser field since a multimode field will cause an additional dephasing effect on the coherence, resulting in a deterioration of the quality of the interference. In contrast, whilst it is advantageous for the field driving the two-photon transition to be single mode (in order to achieve optimal temporal and spectral overlap with the EIT hole induced by the dressing laser), this is not essential since this field does not need to drive the coherence responsible for interference.

When a pulsed laser field is used additional issues must be considered. The group velocity is modified for pulses propagating in the EIT large reductions, e.g., by factors down to $<c/100$, in the group velocity have been observed. Another consideration beyond that found in the simple steady-state case is that the medium can only become transparent if the pulse contains enough energy to dress all the atoms in the interaction volume. The minimum pulse energy to prepare a transparency is:

$$E_{\text{preparation}} = \frac{f_{13}}{f_{23}} NL\hbar\omega \qquad [18]$$

where f_{ij} are the oscillator strengths of the transitions and NL the product of the density and the length. Essentially the number of photons in the pulse must exceed the number of atoms in the interaction volume to ensure all atoms are in the appropriate dressed state. This puts additional constraints on the laser pulse parameters.

Up-conversion to the UV and vacuum UV has been enhanced by EIT in a number of experiments. Only pulsed fields have so far been up-converted to the VUV with EIT enhancement. The requirements on a minimum value of $\Omega_c > \Delta_{\text{Doppler}}$ constrains the conversion efficiency that can be achieved because the $1/\Omega_c^2$ factor in eqn [17] ultimately leads to diminished values of $\chi^{(3)}$. The use of gases of higher atomic weight at low temperatures is therefore highly desirable in any experiment utilizing EIT for enhancement of four-wave mixing to the VUV. Conversion efficiencies, defined in terms of the pulse energies by E_d/E_a or E_d/E_c of a few percent have been achieved using the EIT enhancement technique. It is typically most beneficial to maximize the conversion efficiency defined by the first ratio since ω_a is normally in the UV and is the lower energy of the two applied pulses.

Nonlinear Optics with a Pair of Strong Coupling Fields in Raman Resonance

An important extension of the EIT concept occurs when two strong fields are applied in Raman resonance between a pair of states in a three-level system. Considering the system illustrated in Figure 1 we can imagine that both applied fields are now strong. Under appropriate adiabatic conditions the system evolves to produce the maximum possible value for the coherence $\varrho_{12} = 0.5$. Adiabatic evolution into the maximally coherent state is achieved by adjusting either the Raman detuning or the pulse sequence (counter to-intuitive order). The pair of fields may also be in single-photon resonance with a

third level, in which case the EIT-like elimination of absorption will be important. This situation is equivalent to the formation of a darkstate, since neither of the two strong fields is absorbed by the medium. For sufficiently strong fields the single-photon condition need not be satisfied and a maximum coherence will still be achieved.

An additional field applied to the medium can participate in sum- or difference-frequency mixing with the two Raman resonant fields. The importance of the large value of coherence is that it is the source polarization that drives the new fields generated in the frequency mixing process. Complete conversion can occur over a short distance that greatly alleviates the constraints usually set by phase-matching in nonlinear optics. Recently near unity conversion efficiencies to the far-UV were reported in an atomic lead system where maximum coherence had been created. In a molecular medium large coherence between vibrational or rotational levels has also been achieved using adiabatic pulse pairs. Efficient multi-order Raman sideband generation has been observed to occur. This latter observation may lead the way to synthesizing very short duration light pulses since the broadband Raman sideband spectrum has been proved to be phase-locked.

Pulse Propagation and Nonlinear Optics for Weak CW Fields

In a Doppler-free medium a new regime can be accessed. This is shown in Figure 4 where the possibility now arises of extremely narrow transparency dips since very small values of Ω_c are now sufficient to induce EIT. The widths of these features are typically subnatural and are therefore accompanied by very steep normal dispersion, which corresponds to a much reduced group velocity. The ultraslow propagation of pulses is one of the most dramatic manifestations of EIT in this regime. Nonlinear susceptibilities are now very large as there is constructive interference controlling the value and the splitting of the two states is negligible compared to their natural width. Nonlinear optics at very low light levels, i.e., at the few-photon limit, is possible in this regime.

Propagation of pulses is significantly modified in the presence of EIT. Figure 4 shows the changes to $\mathrm{Re}[\chi^{(1)}]$ that arise. Within the transparency dip there exists a region of steep normal dispersion. In the vicinity of resonance this is almost linear and it becomes reasonable to consider the leading term only that describes the group velocity. An analysis of the

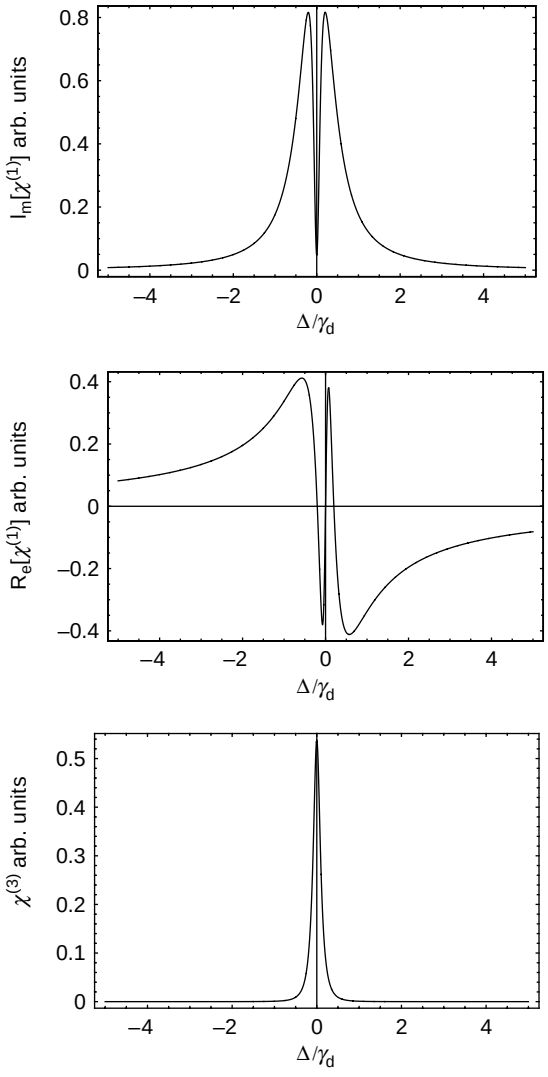


Figure 4 Electromagnetically induced transparency is shown in the case where there is no significant Doppler broadening. The values of the imaginary $\mathrm{Im}[\chi^{(1)}]$ and real $\mathrm{Re}[\chi^{(1)}]$ parts of the linear susceptibility and the nonlinear susceptibility $\chi^{(3)}$ are plotted in the three frames. We take $\Omega_c = \gamma_d/5$ and $\gamma_d = 50\gamma_a$, $\gamma_c = 0$.

refractive changes has been provided by Harris who expanded the susceptibilities (both real and imaginary parts) of the dressed atom around the resonance frequency to determine various terms in $\mathrm{Re}[\chi^{(1)}]$. The first term of the series (zero order) $\mathrm{Re}[\chi^{(1)}](\omega_{13}) = 0$ corresponds to the vanishing dispersion at resonance. The next term $\partial[\mathrm{Re}\chi^{(1)}](\omega)/\partial\omega$ gives the slope of the dispersion curve; at ω_{13} this takes the value:

$$\frac{\partial \mathrm{Re}\chi^{(1)}(\omega_{13})}{\partial\omega} = \frac{|\mu_{13}|^2 N 4(\Omega_c^2 - \gamma_a^2)}{2\pi\varepsilon_0(\Omega_c^2 + \gamma_a\gamma_d)} \qquad [19]$$

The slope of the dispersion profile leads to a reduced group velocity v_g:

$$\frac{1}{v_g} = \frac{1}{c} + \frac{\pi}{\lambda}\frac{\partial \chi^{(1)}}{\partial \omega} \qquad [20]$$

From the expression for $\partial\chi/\partial\omega$ we see that this slope is steepest (and so v_g mimimum) in the case where $\Omega_c \gg \Gamma_2$ and $\Omega_c^2 \gg \Gamma_2\,\Gamma_3$ but is still small compared to Γ_3 (i.e., $\Omega_C < \Gamma_3$) and hence $\partial\chi/\partial\omega \propto 1/\Omega_c^2$. In the limit of small Ω_c the following expression for v_g therefore holds:

$$v_g = \frac{\hbar c \varepsilon_0 [\Omega_c]^2}{2\omega_d |\mu_{13}|^2 N} \qquad [21]$$

Extremely low group velocities, down to a few meters per second, are achieved in this limit using excitation of the hyperfine split ground states in either laser cooled atomic samples or Doppler-free configurations in finite temperature samples. Recently similar light slowing has been observed in solids. Storage of the optical pulse within the medium has also been achieved by adiabatically switching off the coupling field and thus trapping the optical excitation as an excitation within the hyperfine ground states for which the storage time can be very long (>1 ms) due to very low dephasing rates. Since the storage scenario should be valid even for single photons this process has attracted considerable attention recently as a means to enable quantum information storage.

Extremely low values of Ω_c are sufficient to induce complete transparency (albeit in a very narrow dip) and at this location the nonlinear response is resonantly enhanced. Very high efficiency nonlinear frequency mixing and phase conjugation at low light levels have been reported under these conditions. It is expected that high-efficiency nonlinear optical processes will persist to ultralow intensities (the few photon level) in an EIT medium of this type.

One example of highly enhanced nonlinear interactions is the predicted large values of the Kerr type nonlinearity (nonlinear refractive index). The origin of this can be seen by considering the steep dispersion profile in the region of the transparency dip in Figure 4. Imagine that we apply an additional field ω_f, perhaps a very weak one, at a frequency close to resonance between state |2> and a fourth level |4>. The ac Stark shift caused by this new field to the other three level, although small, will have a dramatic effect upon the value of the refractive index because of the extreme steepness of the dispersion profile. Essentially even a very small shift of the resonant wavelength causes a large change in the refractive index for a field applied close to this frequency. It is predicted that strong cross-phase modulations will be created in this process between the fields ω_f and ω_d, even in the quantum limit for these fields. This is predicted to lead to improved schemes for quantum nondemolition measurements of photons through the measurement of the phase-shifts they induce (through cross-phase modulation) on another quantum field. This type of measurement has direct application in quantum information processing.

See also

Nonlinear Optics, Applications: Phase Matching; Raman Lasers. **Scattering:** Raman Scattering.

Further Reading

Arimondo E (1996) Coherent population trapping in laser spectroscopy. *Progress in Optics* 35: 257–354.

Harris SE (1997) Electromagnetically induced transparency. *Physics Today* 50: 36–42.

Harris SE and Hau LV (1999) Non-linear optics at low light levels. *Physical Review Letters* 82: 4611–4614.

Harris SE, Field JE and Imamoglu A (1990) Nonlinear optical processes using electromagnetically induced transparency. *Physical Review Letters* 64: 1107–1110.

Hau LV, Harris SE, Dutton Z and Behroozi CH (1999) Light speed reduction to 17 metres per second in an ultra-cold atomic gas. *Nature* 397: 594–598.

Marangos JP (2001) Electromagnetically induced transparency. In: Bass M, *et al.* (ed.) *Handbook of Optics*, vol. IV, ch. 23. New York: McGraw-Hill.

Merriam AJ, Sharpe SJ, Xia H, *et al.* (1999) Efficient gas-phase generation of vacuum ultra-violet radiation. *Optics Letters* 24: 625–627.

Schmidt H and Imamoglu A (1996) Giant Kerr nonlinearities obtained by electromagnetically induced transparency. *Optics Letters* 21: 1936–1938.

Scully MO (1991) Enhancement of the index of refraction via quantum coherence. *Physical Review Letters* 67: 1855–1858.

Scully MO and Suhail Zubairy M (1997) Quantum Optics. Cambridge, UK: Cambridge University Press.

Sokolov AV, Walker DR, Yavuz DD, *et al.* (2001) Femtosecond light source for phase-controlled multi-photon ionisation. *Physical Review Letters* 87: 033402-1.

Zhang GZ, Hakuta K and Stoicheff BP (1993) Nonlinear optical generation using electromagnetically induced transparency in hydrogen. *Physical Review Letters* 71: 3099–3102.

ENVIRONMENTAL MEASUREMENTS

Contents

Doppler Lidar
Hyperspectral Remote Sensing of Land and the Atmosphere
Laser Detection of Atmospheric Gases
Optical Transmission and Scatter of the Atmosphere

Doppler Lidar

R M Hardesty, National Oceanic and Atmospheric Administration, Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Lidars developed for the measurement of atmospheric winds make use of the Doppler effect, in which radiation scattered or emitted from a moving object is shifted in frequency as a result of the movement of the particles. Atmospheric Doppler lidars irradiate a volume of atmosphere with a pulse of very narrowband, laser-produced radiation, then detect the change in frequency of the radiation backscattered from atmospheric aerosol particles or molecules present in the volume. Although Doppler lidars are conceptually similar to Doppler weather radars, lidar wavelengths are 3–4 orders of magnitude shorter than radar wavelengths, producing some important differences in propagation and scattering characteristics. Radiation at wavelengths in the optical region of the spectrum is efficiently scattered by aerosols and molecules, therefore Doppler lidars, unlike commonly used Doppler weather radars, do not require the presence of hydrometeors or insects to obtain useful results. Also, optical radiation can be tightly collimated, virtually eliminating effects of ground clutter and enabling lidar probing of small volumes to within a few meters of terrain or structures.

The primary disadvantage of lidar techniques is the severe attenuation of the optical radiation by cloud water droplets and fog. Doppler lidars do not typically probe beyond the edge of most atmospheric clouds – the one exception being tenuous ice clouds such as cirrus, which often are characterized by low optical extinction and high backscatter, making them excellent lidar targets. Doppler lidars also do not have the extended range, which can exceed 100 km, available from contemporary meteorological radars when scatterers are present.

In clear air, however, the characteristics and capabilities of Doppler lidar are well suited for obtaining detailed wind and turbulence observations for a wide variety of applications. Lidar beams can be easily scanned to characterize motions within very confined three-dimensional spaces such as shallow atmospheric boundary layers, narrow canyons, and turbulent structures. The relative compactness of Doppler lidars makes deployment on aircraft or other moving platforms extremely viable, and a Doppler lidar deployed on a satellite has been proposed as a way to obtain global measurements of atmospheric wind fields. By scanning a lidar beam from an orbiting satellite and analyzing the returns backscattered from clouds, aerosols, and molecules, a satellite-based instrument could provide important wind information for numerical forecast models.

Doppler Lidar Fundamentals

Backscattered Light

As laser light propagates through the atmosphere, some of the incident energy is scattered by the atmospheric molecules and constituents through which the light passes. In lidar applications, the light backscattered (scattered directly back toward the source) is collected and analyzed. The backscattering properties of an atmospheric particle depend on its refractive index, shape, and size. In the special case when the particle is much smaller than the wavelength of the incident radiation, as is the case for laser radiation scattered by atmospheric molecules, the scattering process is characterized as Rayleigh scattering. Because, in the Rayleigh scattering regime the energy backscattered by a particle increases proportionally as the inverse of the fourth power of the wavelength, Doppler lidar systems designed for molecular scatter operate at short wavelengths. Molecular Doppler lidar systems typically operate in the visible or ultraviolet spectral regions.

Aerosol particles, the other component of the atmosphere that scatters laser light, produce Mie scattering, which applies when the diameter of the scatterers is not orders of magnitude smaller than the incident wavelength. Aerosol particles present in the atmosphere include dust, soot, smoke, and pollen, as well as liquid water and ice. Although for Mie scattering the relationship between the power backscattered by an ensemble of aerosol particles and the incident wavelength is not simply characterized, most studies have shown that the energy backscattered increases roughly proportional to the first or second power of the inverse of the incident wavelength, depending on the characteristics of the scattering aerosol particles. In a polluted environment, such as in the vicinity of urban areas, an abundance of large particles often results in a roughly linear relationship between the inverse of the incident wavelength and the backscattered energy, while in more pristine environments, such as the free troposphere, the inverse wavelength/backscatter relationship can approach or exceed a square-law relationship.

The primary objective in Doppler lidar is to measure the Doppler frequency shift of the scattered radiation introduced by the movements of the scattering particles. Figure 1 shows a typical spectrum of the radiation collected at the lidar receiver for a volume of atmosphere irradiated by a monochromatic laser pulse. Molecular scattering produces the broadband distribution in Figure 1, where the broadening results from the Doppler shifts of the radiation backscattered from molecules moving at their thermal velocities. The width of the molecular velocity distribution in the atmosphere ranges from ~ 320 to 350 m s^{-1}, scaling as the square root of the temperature. In the center of the spectrum is a much narrower peak resulting from scattering of the light by aerosol particles. Because the thermal velocity of the much larger aerosol particles is very low, the width of the distribution of the aerosol return is determined by the range of velocities of particles moved about by small-scale turbulence within the scattering volume. This is typically on the order of a few m s^{-1}. Also shown in Figure 1 is an additional broadband distribution, due to scattered solar radiation collected at the receiver. If the laser source is not monochromatic, the backscattered signal spectrum is additionally broadened, with the resulting spectrum being the convolution of the spectrum shown in Figure 1 with the spectrum of the laser pulse. As seen in the figure, the entire spectrum is Doppler-shifted in frequency, relative to the frequency of the laser pulse. The object for a Doppler lidar system is to measure this shift, given by $\delta f = 2v_{\text{rad}}/\lambda$, where v_{rad} is the component of the mean velocity of the particles in the direction of propagation of the lidar pulse and λ is the laser wavelength.

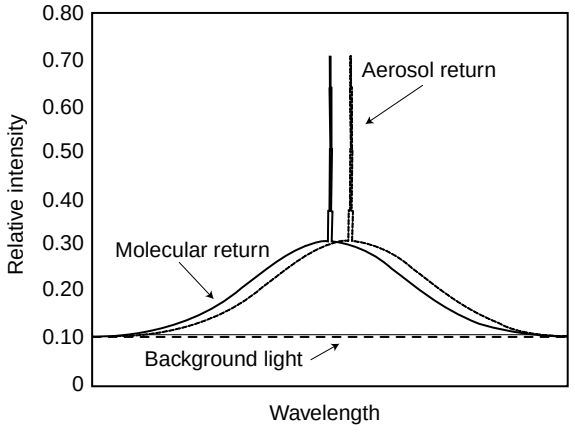


Figure 1 Distribution of wind velocities in a lidar return, showing Doppler shift resulting from wind-induced translation of the scatterers. The ratio of aerosol to molecular signals increases with increasing wavelength.

Elements of a Doppler Lidar

Doppler lidar systems can be designed primarily to measure winds from aerosol-scattered radiation, molecule-scattered radiation, or both. The type of system places specific requirements on the primary components that comprise a Doppler lidar system. A Doppler lidar consists of a laser transmitter to produce pulses of energy that irradiate the atmospheric volume of interest, a receiver which collects the backscattered photons and estimates the energy and Doppler shift of the return, and a beam pointing mechanism for directing the laser beam and receiver field of view together in various directions to probe different atmospheric volumes. Independent of whether the primary scatterers are molecules or aerosol particles, the system design criteria for a Doppler system are driven by a fundamental relationship between the error in the estimate of mean frequency shift δf_1, the bandwidth of the return f_2 (proportional to the distribution of velocities), and the number of incident backscattered photons detected N as

$$\delta f_1 \propto f_2/N^{1/2} \qquad [1]$$

Thus, the precision of the velocity estimate is improved by increasing the number of detected signal photons and/or decreasing the bandwidth of the backscattered signal. It is obvious from the equation that a significantly greater number of photons are required to achieve the same precision in a Doppler measurement from a molecular backscattered signal,

characterized by higher bandwidth f_2, compared to the number required to compute velocities from aerosol returns. The improved measurement precision gained from a narrow bandwidth return also indicates that the laser transmitter in a Doppler lidar system should be characterized by narrow spectral width (considerably narrower than the spectral broadening caused by the distribution of velocities), and to increase the number of photons detected, maximum transmit energy.

Narrowband, single-frequency pulses are obtained by employing a low-power, precisely frequency-stabilized master oscillator laser, whose radiation is either used to seed a stabilized, higher-energy laser cavity (injection seeding) or amplified in one or more optical amplifiers (master-oscillator, power amplifier, (MOPA)) to produce frequency-stabilized pulses.

The lidar receiver gathers the backscattered photons and extracts the wind velocity as a function of the range to the scattering volume. This requires a telescope, to gather and focus the scattered radiation, and a receiver to detect the scattered radiation and determine the Doppler shift. Frequency analysis in a Doppler lidar receiver is carried out using one of two techniques: coherent detection (also known as heterodyne detection) or direct detection (alternately labeled incoherent detection). The techniques differ fundamentally. In a heterodyne receiver, frequency analysis is carried out on a digitized time series created by mixing laser radiation with the backscattered radiation, while in direct detection lidar an interferometer optically analyzes the backscattered radiation to produce a light pattern which contains the information on the frequency content.

Coherent (Heterodyne) Doppler Lidar

Description

Coherent or heterodyne lidar is implemented by optically mixing the backscattered laser light with radiation from a stable, continuous-wave, local oscillator (LO) laser whose frequency is precisely controlled to be a known frequency offset, typically on the order of tens of MHz, from that of the laser transmitter (Figure 2). The mixing process at the face of an optical detector generates, after high pass filtering, an electrical signal with amplitude proportional to amplitude of the backscattered electromagnetic field and frequency equal to the difference between the backscattered field frequency and the LO laser field frequency. This signal is sampled, and then digitally processed to estimate the range-dependent mean frequency shift of the backscattered signal, from which the radial wind component can be

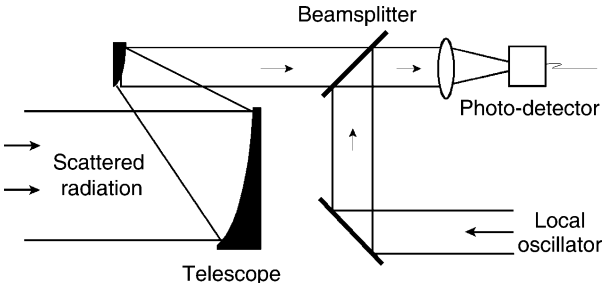


Figure 2 Schematic of coherent detection of backscattered radiation.

derived. A property of coherent lidar is that, because of constructive and destructive interference of the returns from individual scatterers, single pulse returns are characterized by random fluctuations in the amplitude and phase of the detected time series and Fourier spectrum. Consequently, averaging (sometimes referred to as accumulation) over multiple pulses usually is necessary to produce a sufficiently precise estimate of the signal mean frequency estimate. Because the phase of the detected time series is random within each pulse, averaging must be carried out in the power spectrum or autocorrelation function domain, rather than on the detected time series.

The optical mixing process in coherent Doppler lidar provides both benefits and design challenges. Because receiver efficiencies (equivalent fraction of incident photons entering the receiver that are converted to electrons) are quite high, wind velocities can be estimated from very weak backscattered signals. Also, signal processing is performed on a time series derived from the mixed signal, enabling the use of electrical filters to produce a narrow receiver bandwidth and effectively eliminating broadband solar background light as a major source of noise. As a result, unlike many lidars coherent Doppler lidar performance is not degraded under daylight conditions.

The primary limitation of coherent detection results from the added photon noise from the local oscillator radiation (which is usually much stronger than the backscattered radiation). The effect of local oscillator photon noise is to define a system noise level below which very weak signals cannot be practically extracted, even with substantial multiple-pulse averaging. In addition, because efficient mixing requires phase coherence of the backscattered signal field across the detector, coherent lidar performance at longer ranges is degraded by strong optical turbulence along the path of the laser pulse. These adverse effects of turbulence increase for shorter wavelength systems.

Although measurements at wavelengths near 1.06 μm have been demonstrated, coherent Doppler lidar systems used for regular atmospheric probing operate in the eye-safe, infrared portion of the spectrum at wavelengths longer than 1.5 μm. Currently, the two most common system wavelengths for coherent lidar wind systems are in atmospheric window spectral regions around 2 and 10.6 μm. Early coherent Doppler lidar measurements, beginning in the 1970s, employed CO_2 laser transmitters and local oscillators operating at wavelengths near 10.6 μm. Pulsed CO_2 laser transmitters with as much as 10 J of energy have since been demonstrated, and systems with 1 J lasers have probed the atmosphere to ranges of 30 km or more. In the late 1980s, solid state laser transmitters operating near 2 μm wavelengths were introduced into coherent lidar wind-measuring systems. The compact size and potential reliability advantages of solid-state transmitters, in which the transmitter laser is optically pumped by an array of laser diodes, provide advantages over larger CO_2 laser technology. Also, because for a given measurement accuracy the range-resolution obtainable is proportional to wavelength, 2 μm instruments have an enhanced capability to probe small-scale features. However, although development of higher energy, single-frequency coherent lidar sources operating at 2 μm region, as well as at 1.5 μm, is currently an active research area, solid state lasers with pulse energies greater than several tens of mJ have yet to be incorporated into lidar systems for atmospheric probing.

Applications of Coherent Doppler Lidar

Coherent lidars have been used to measure winds for a variety of applications, and from an assortment of platforms, such as ships and aircraft. Since these lidars operate in the infrared where aerosol scattering dominates molecular scattering, they require aerosol particles to be present at some level to obtain usable returns. Although clouds also provide excellent lidar targets, most of the more useful applications of coherent lidars involve probing the atmospheric boundary layer or lower troposphere where aerosol content is highest. Because of the capability to scan the narrow lidar beam directly adjacent to terrain, a unique application of lidar probing is the measurement of wind structure and evolution in complex terrain such as mountains and valleys. Over the past two decades, Doppler lidar studies have been used, for example, to study structure of damaging windstorms on the downwind side of mountain ranges, advection of pollution by drainage flows from valleys, and formation of mountain leeside turbulence as a potential hazard to landing aircraft. The interactions of wind flows with complex terrain can produce dangerous conditions for transportation, especially aircraft operations. Figure 3 shows a strong mountain wave measured by a Doppler lidar near Colorado Springs, during an investigation of the effects of downslope winds and turbulence on aircraft operations at the Colorado Springs airport. The mountains are on the right of the figure, where the strong wave with wavelength of about 6 km can be seen. Note the reversal in winds near the location where the wave descends to the surface. The Colorado Springs study was prompted by the crash of a passenger jet while landing at the airport during a period when the winds were down the slope of the Rocky Mountains just to the west of the airport.

A Doppler lidar was recently deployed in an operational mode for wind shear detection at the new Hong Kong International Airport. Because winds flowing over mountainous terrain just south of the

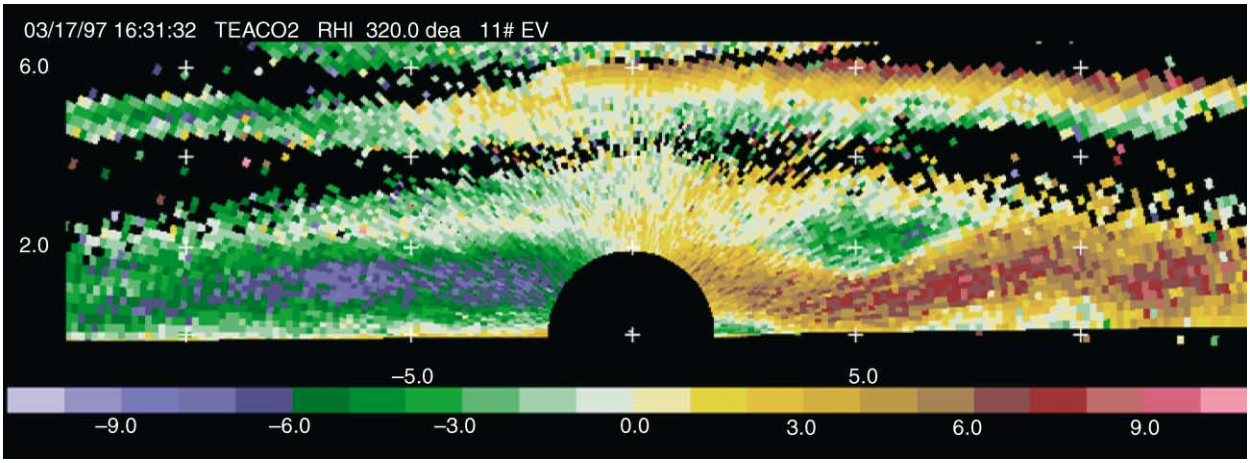


Figure 3 Vertical scan of the winds in a mountain wave measured by a 10.6 μm coherent Doppler lidar near Colorado Springs, Colorado. Courtesy L. Darby, NOAA.

airport can produce mountain waves and channeled flows, wind shear is frequently encountered during landing and takeoff operations. Figure 4 shows a gust front situated just to the west of the airport, characterized by a sharp wind change (which would produce a corresponding change in airspeed) approaching 13 m s^{-1} over a distance of less than one kilometer. Providing a capability to detect such events near the airport and warn pilots during approach or preparation for takeoff was the primary reason for deployment of the lidar. Within the first year of operation the Hong Kong lidar was credited with improving wind shear detection rates and providing more timely warnings for pilots of the potentially hazardous conditions. As of August 2003 this lidar had logged more than 10 000 hours of operation.

Because coherent Doppler lidars are well matched to applications associated with probing small-scale, turbulent phenomena, they have also been deployed at airports to detect and track wing tip vortices generated by arriving or departing aircraft on nearby parallel runways. In the future, a network of ground-based lidars could provide information on vortex location and advection speed as well as wind shear, leading to a potential decrease in congestion at major airports. Also, compact lidar systems looking directly ahead of research aircraft have shown the capability to detect wind changes associated with potentially hazardous clear-air turbulence. In the future, a Doppler lidar installed on the commercial aircraft fleet could potentially look ahead and provide a warning to passengers to fasten seat belts before severe turbulence is encountered.

The high resolution obtainable in a scanning lidar enables visualization of finely structured wind and turbulence layers. Figure 5 shows an image of turbulence associated with a nocturnal low-level jet just 50 m above the surface obtained at night over flat

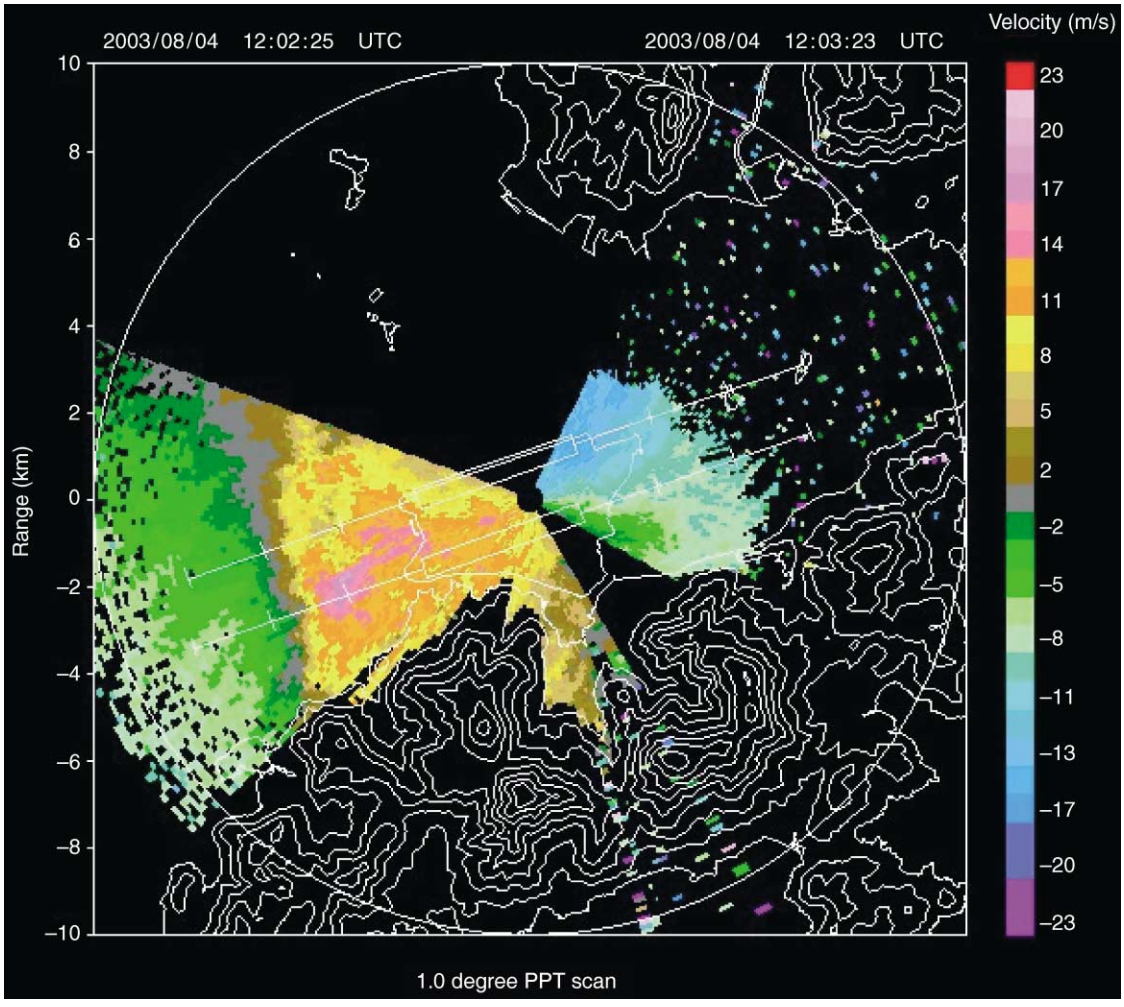


Figure 4 Horizontal depiction of the radial component of the wind field measured at Hong Kong Airport by a 2.02 μm Doppler lidar during a gust front passage, showing strong horizontal shear just west of the airport. Courtesy C. M. Shun, Hong Kong Observatory.

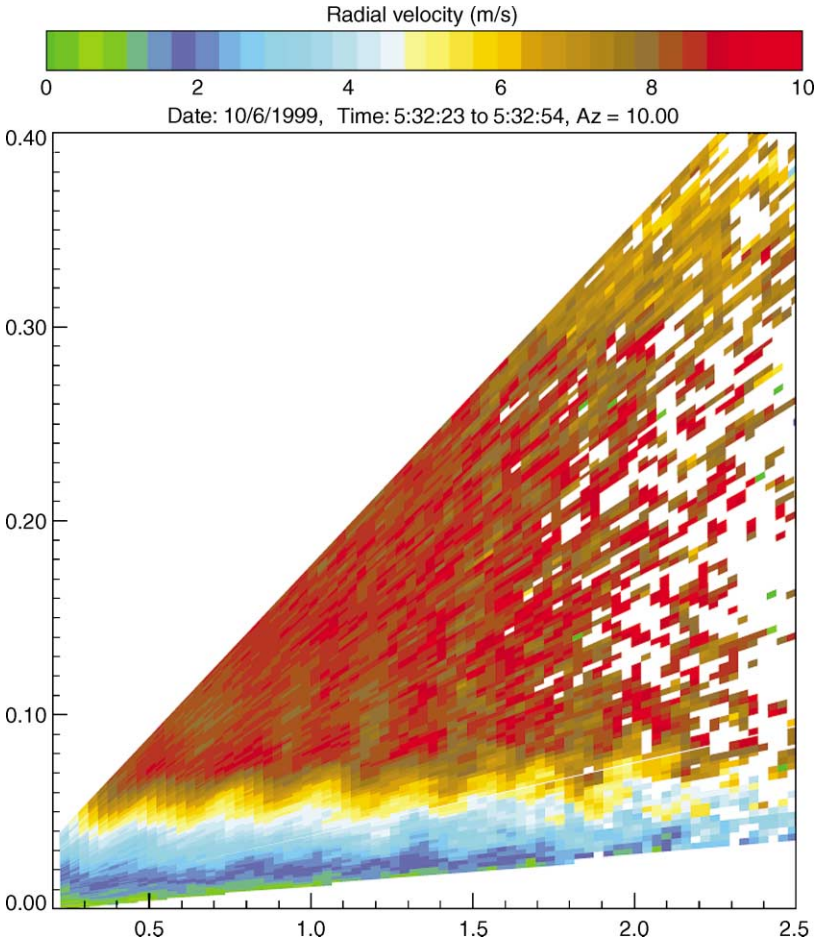


Figure 5 Turbulence structure within a nocturnal stable layer at a height of 50 m. Note the stretching of the vertical scale. Courtesy R. Banta, NOAA.

terrain in Kansas. Although a low-level jet was present on just about every evening during the experiment, similar images obtained at different times by the Doppler lidar illustrated markedly different characteristics, such as a wide variation in the observed mechanical turbulence along the interface. Such observations enable researchers to improve parameterizations of turbulence in models and better understand the conditions associated with vertical turbulent transport and mixing of atmospheric constituents such as pollutants.

By deploying Doppler lidars on moving platforms such as ships and aircraft, the spatial coverage of the measurements can be greatly increased. Although removal of platform motion and changes in orientation are not trivial, aircraft-mounted lidar can map out such features as the low-level jet and the structure of boundary layer convective plumes. Figure 6 shows horizontal and vertical motions associated with convective plumes measured along an approximately 60 km path over the southern Great Plains of the United States. From such measurements estimates of turbulence intensity, integral scale, and other boundary layer characteristics can be computed.

Direct Detection Doppler Lidar

Description

Direct detection or incoherent Doppler lidar has received significant attention in recent years as an alternative to coherent lidar for atmospheric wind measurements. In contrast to coherent lidar, in which an electrical signal is processed to estimate Doppler shift, an optical interferometer, usually a Fabry–Perot etalon, serves as the principal element in a direct detection lidar receiver for determining the frequency shift of the backscattered radiation.

One implementation of a direct-detection Doppler lidar receiver is the 'fringe imaging' technique. In this design, the interferometer acts as a spectrum analyzer. The backscatter radiation is directed through a

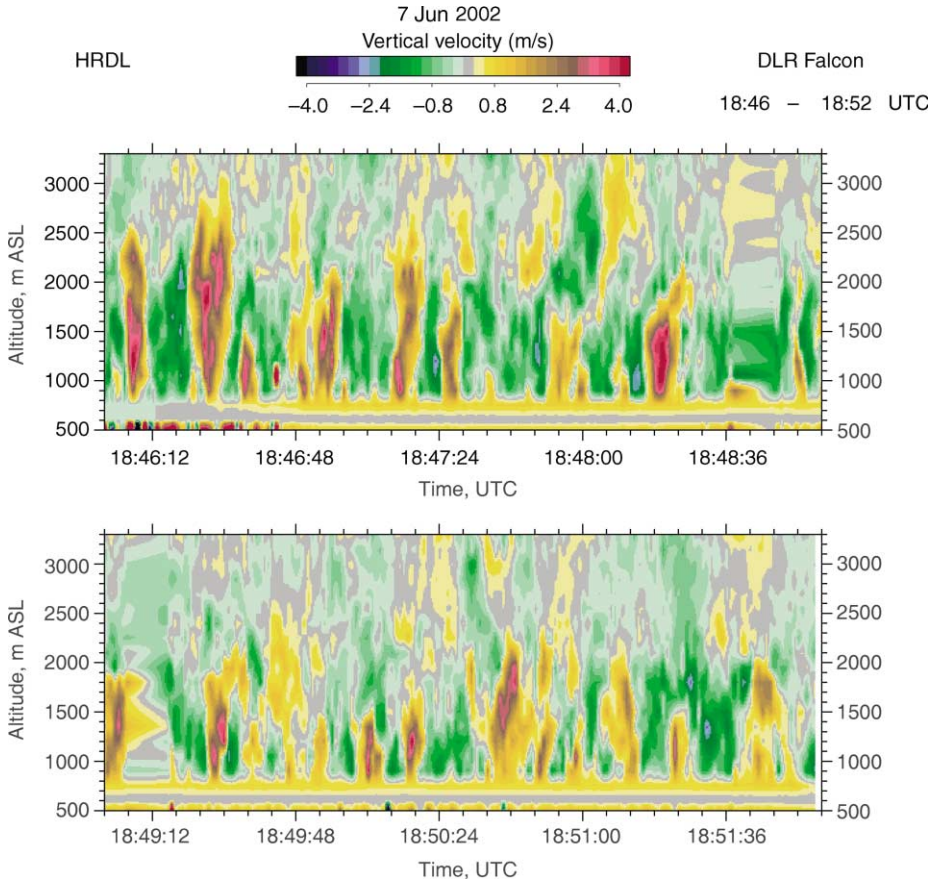


Figure 6 Vertical motion in convective plumes measured by a vertically pointing airborne coherent Doppler lidar. Total horizontal extent of the measurements is approximately 60 km, reddish colors correspond to upward motions.

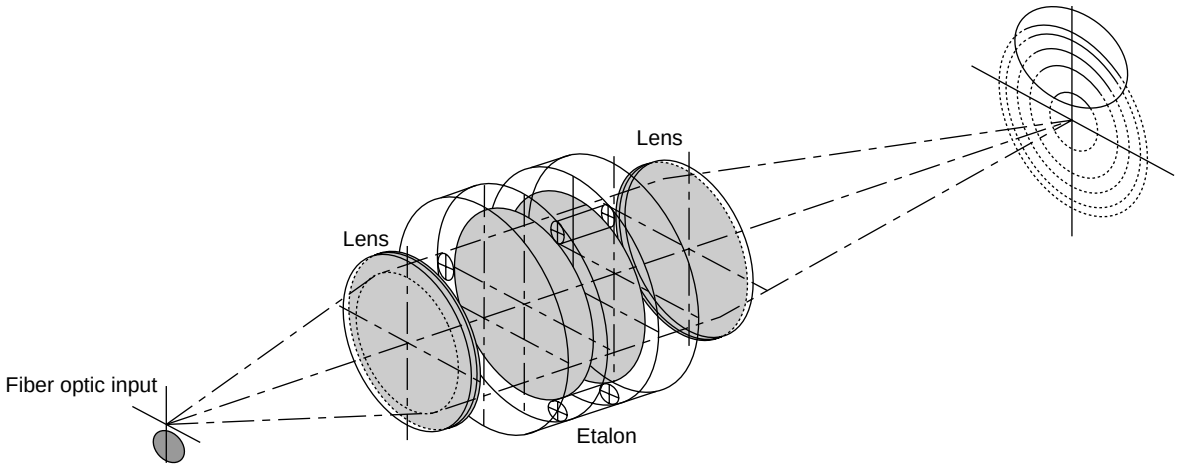


Figure 7 Schematic of Fabry–Perot etalon in a direct detection, fringe-imaging lidar receiver. Courtesy P. Hays, Michigan Aerospace Corporation.

Fabry–Perot interferometer, which produces a ring pattern in the focal plane (Figure 7). The spectral content information of the incident radiation is contained in the radial distribution of the light. Each ring corresponds to an order of the interferometer and is equivalent to a representation of the backscattered signal frequency spectrum. As the mean frequency of the backscattered radiation changes, the rings move inward or outward from the center. To extract a spectrum of the backscattered light, one or

more of the rings are imaged onto a two-dimensional detector, and the resulting pattern analyzed. The rings can either be detected using a multi-element ring detector, or can be converted to a linear pattern with a circle-to-line converter optic and then imaged on to a charge-coupled device array device.

An alternate direct detection receiver configuration, generally called the double edge technique, is to employ two interferometers as bandpass filters, with the center wavelength of each filter set above and below the laser transmitter wavelength, as shown in Figure 8. The incoming radiation is split between the two interferometers, and the wavelength shift is computed by examining the ratio of the radiation transmitted by each interferometer. Both the double edge and fringe-imaging techniques have demonstrated wind measurements to heights well into the stratosphere. The major challenge associated with the double edge receiver is optimizing the instrument when both aerosol and molecular scattered radiation are present, since in general the change in transmission as a function of velocity is different for the aerosol and molecular signals. Double edge receivers optimized for both aerosol and molecular returns place the bandpass of the etalon filters at the precise wavelength where the change in transmission with change in Doppler shift is the same for both aerosol and molecular returns.

For both types of direct detection receivers described above, much of the radiation incident on the interferometer is reflected out of the system, reducing the overall efficiency of the receiver. Recently, designs that incorporate fiber optics to collect a portion of the reflected radiation and 'recycle' it back into the etalon have been demonstrated as a method to improve the transmission efficiency of the etalon in a fringe imaging receiver.

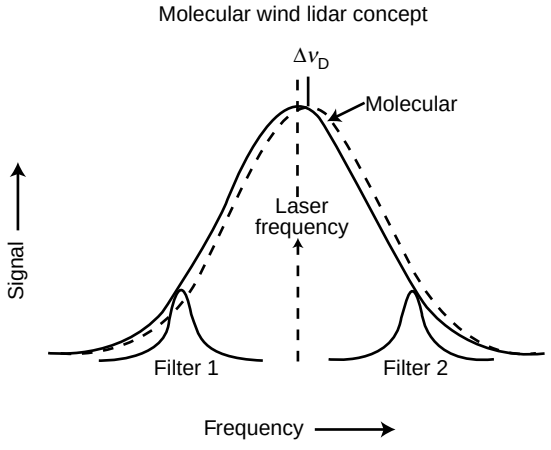


Figure 8 Spectrum of molecular lidar return showing placement of bandpass filters for a dual channel (double edge) direct detection receiver. Courtesy B. Gentry, NASA.

Doppler Wind Measurements Based on Molecular Scatter

One of the primary advantages of direct detection Doppler lidar is its capability for measurements based on scatter from atmospheric molecules. Measurement of Doppler shifts from molecular scattered radiation is challenging because of the large Doppler-broadened bandwidth of the return. Because one wants to measure a mean wind velocity with a precision of a few m s^{-1} or better, in the order of 10^5 photons are required. Some combination of multiple pulse averaging, powerful lasers, and large receiver optics is required to obtain these high photon counts from backscattered returns.

Molecular-scatter wind measurements have been demonstrated in the visible spectral region at 532 nm wavelength as well as at 355 nm in the near ultraviolet. The ultraviolet region has the dual advantages of enhanced molecular scatter and less restrictive laser eye-safety restrictions. Figure 9 shows the time series of a wind profile measured in the troposphere from both aerosols and clouds using a molecular-scatter, ground-based 355-nm wavelength Doppler fringe-imaging lidar. The figure shows measurements from receiver channels optimized for the wideband molecular signal and the narrowband aerosol return. For this measurement, backscattered photons were collected by a 0.5 m receiver aperture, averaged for 1 minute, and processed. In the absence of clouds, direct detection Doppler lidars have measured wind profiles continuously from the surface to beyond 15 km height. Figure 10 shows that the estimated wind error for the same 355 nm lidar is less than 1 m s^{-1} to about 10 km, and less than 4 m s^{-1} at 15 km height.

Heterodyne and Direct-Detection Doppler Trade-Offs

Lively debates within the lidar community have occurred over the past decade regarding the relative merits of heterodyne versus direct detection Doppler lidars. To a large extent, the instruments are complementary. Generally, heterodyne instruments are much more sensitive when significant aerosols are present. Coherent lidar processing techniques have been developed that can produce accurate wind measurement rates using only a few lidar pulses with as few as 100 detected photons, such that several wind observations per second can be obtained for nominal pulse rates. This inherent sensitivity has led to numerous applications in which a lidar beam has been scanned rapidly over a large volume to obtain time-varying, three-dimensional wind measurements.

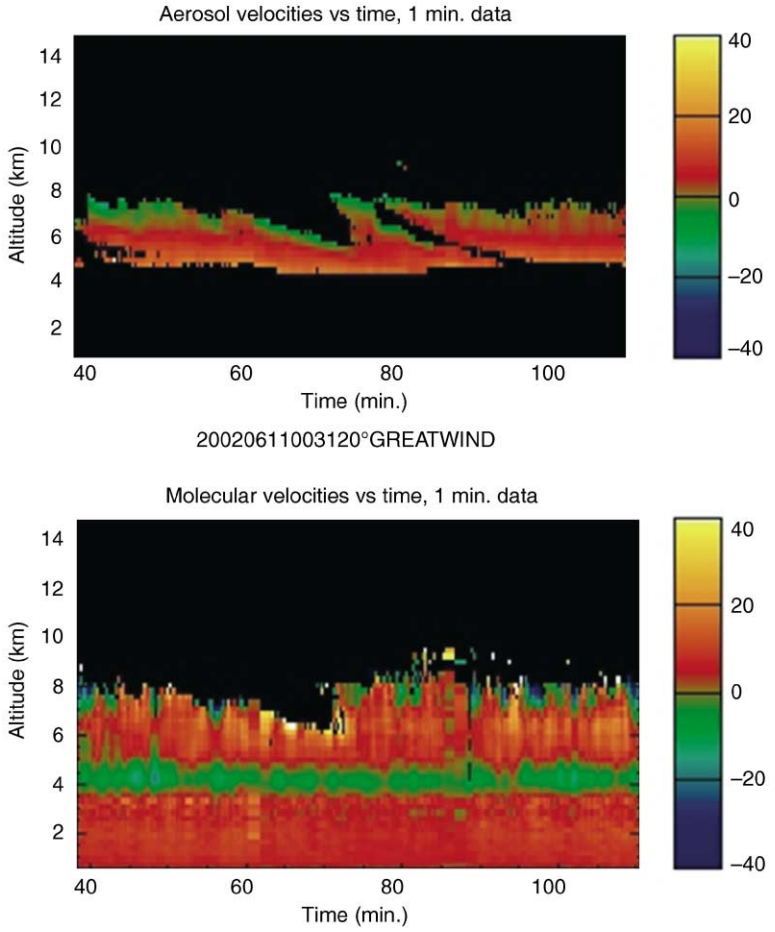


Figure 9 Time series of radial wind speed profiles measured by a 355 nm fringe-imaging direct detection lidar aerosol channel (top) and molecular channel (bottom) at Maura Loa Observatory, HI. Change in wind speed and blocking of the return by clouds is clearly seen. Courtesy C. Nardell, Michigan Aerospace Corp.

The primary advantage of direct detection instruments is their demonstrated capability to measure winds from molecular-backscattered returns in the middle and upper troposphere. In very pristine air, direct detection offers the only method for long-range wind measurements, even though significant averaging may be required. Direct-detection lidars have the additional advantage of being unaffected by atmospheric refractive turbulence. Unlike heterodyne lidars, which require a very pure laser pulse and a diffraction-limited receiver field of view matched to and precisely aligned with the transmitted beam, direct detection systems can have a wider bandwidth transmitter and a receiver field of view several times diffraction-limited. In direct detection lidar design, the field of view is usually constrained by the need to limit background light during daytime operation. A major design challenge for direct detection instruments is holding the Fabry–Perot etalon plate spacing stable over a range of temperatures and in high vibration environments.

Global Wind Measurements

A satellite-based Doppler lidar has frequently been proposed as a way to measure wind fields over most of the Earth. At present, winds are the one major meteorological variable not well-measured from orbiting platforms. Measurement of winds is especially important over regions of the Earth that are not currently well sampled, such as over Northern Hemisphere oceans, as well as over most of the tropics and Southern Hemisphere. Wind profile information is currently obtained from radiosondes and by tracking cloud and water vapor inhomogeneities using satellite imagers. Doppler lidar wind measurements would greatly augment the current data set by providing wind estimates throughout the troposphere under clear conditions, and highly height-resolved observations down to cloud tops when cloud decks are present. Observing system simulation experiments conducted in recent years indicate that satellite-based lidar global wind

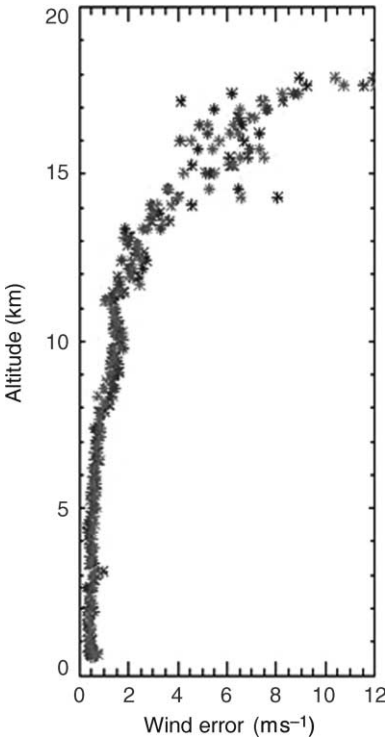


Figure 10 Estimated error in the radial velocity estimate versus altitude for a 355 nm direct detection, fringe imaging Doppler lidar system at Mauna Loa observatory on Hawaii. The lidar provides data for assimilation into local mesoscale forecast models. Courtesy C. Nardell, Michigan Aerospace Corp.

measurements could lead to a significant improvement in long-term forecast skill, provided the wind fields can be observed with sufficient accuracy and spatial resolution.

In a Doppler lidar wind mission, a satellite carrying a lidar system would orbit the Earth in a nearly polar orbit. The pulsed laser beam would be scanned conically about the nadir to obtain different components of the wind velocity. The scanning could be either continuous or 'stop and stare'. After sufficient returns are averaged at a given pointing angle to produce an acceptable estimate, the radial component of the velocity would be computed and assimilated directly into numerical analysis and forecast models.

Doppler lidar measurement of winds from space is theoretically feasible but technologically difficult. Depending on the orbital height, the scattering volume is anywhere from 450 to ~850 km from the satellite, which challenges the sensitivity of current system types. Because weight and power consumption are critical parameters for space systems, telescope diameter and laser power cannot be easily increased to obtain the necessary sensitivity. Similarly, the ability to average returns from multiple pulses is also limited by time limitations. Because a satellite moves at about 7 km s^{-1}, in order to obtain measurements over a horizontal distance of 300 km (the resolution of the radiosonde network) only about 45 seconds are available to make enough observations from multiple look angles to obtain a useful measurement. It should also be noted that, as a result of the high orbital velocity of the satellite, precise knowledge of beam pointing is extremely critical for measurements from satellites. For a lidar operating at a nadir angle of 45 degrees, an error in the knowledge of pointing angle of just 1 mrad results in an error of about 5 ms^{-1} in the measured radial component of the wind.

Despite the challenge of employing a satellite-based Doppler lidar, efforts are continuing to develop the appropriate technology and to assess the impact of the observations. The European Space Agency is planning a Doppler wind lidar demonstration mission for the late 2000s that would incorporate a nonscanning, direct-detection instrument with both aerosol and molecular channels. Doppler lidar technology research and numerical simulations aimed at satellite-based wind sensing is ongoing at several research centers within the United States, Europe, and Japan. One option currently being studied is a 'hybrid' lidar system that would combine a direct detection lidar for measurement of winds in the clear free troposphere with a low-energy coherent system that would obtain its observations from clouds and the aerosol-rich boundary. Although a hybrid instrument reduces some requirements on system power and aperture size, combining the two techniques on a single platform will likely require innovative engineering and new approaches to beam combining and scanning.

See also

Environmental Measurements: Optical Transmission and Scatter of the Atmosphere. **Scattering:** Scattering Theory.

Further Reading

Baker WE, Emmitt GD, Robertson F, *et al.* (1998) Lidar-measured wind from space: a key component for weather and climate prediction. *Bulletin of the American Meteorological Society* 79: 581–599.

Banta RM, Olivier LD and Levinson DH (1993) Evolution of the Monterey Bay sea breeze layer as observed by pulsed Doppler lidar. *Journal of Atmospheric Science* 50 (24): 3959–3982.

Banta RM, Olivier LD, Gudiksen PH and Lange R (1996) Implications of small-scale flow features to modeling dispersion over complex terrain. *Journal of Applied Meteorology* 35: 330–342.

Banta RM, Darby LS, Kaufmann P, Levinson DH and Zhu C-J (1999) Wind flow patterns in the Grand Canyon as revealed by Doppler lidar. *Journal of Applied Meteorology* 38: 1069–1083.

Frehlich R (1996) Coherent Doppler lidar measurements of winds in the weak signal regime, 1997. *Applied Optics* 36: 3491–3499.

Gentry B, Chen H and Li SX (2000) Wind measurements with a 355 nm molecular Doppler lidar. *Optics Letters* 25: 1231–1233.

Grund CJ, Banta RM, George JL, *et al.* (2001) High-resolution Doppler lidar for boundary-layer and cloud research. *Journal of Atmosphere and Oceanic Technology* 18: 376–393.

Huffaker RM and Hardesty RM (1996) Remote sensing of atmospheric wind velocities using solid state and CO_2 coherent laser systems. *Proceedings of the IEEE* 84: 181–204.

Menzies RT and Hardesty RM (1989) Coherent Doppler lidar for measurements of wind fields. *Proceedings of the IEEE* 77: 449–462.

Post MJ and Cupp RE (1990) Optimizing a pulsed Doppler lidar. *Applied Optics* 29: 4145–4158.

Rees D and McDermid IS (1990) Doppler lidar atmospheric wind sensor: reevaluation of a 355 nm incoherent Doppler lidar. *Applied Optics* 29: 4133–4144.

Rothermel J, Cutten DR, Hardesty RM, *et al.* (1998) The multi-center airborne coherent lidar atmospheric wind sensor. *Bulletin of the American Meteorological Society* 79: 581–599.

Skinner WR and Hays PB (1994) Incoherent Doppler lidar for measurement of atmospheric winds. *Proceedings of SPIE* 2216: 383–394.

Souprayen C, Garnier A and Hertzog A (1999) Rayleigh–Mie Doppler wind lidar for atmospheric measurements. II: Mie scattering effect, theory, and calibration. *Applied Optics* 38: 2422–2431.

Souprayen C, Garnier A, Hertzog A, Hauchecorne A and Porteneuve J (1999) Rayleigh–Mie Doppler wind lidar for atmospheric measurements. I: Instrumental setup, validation, first climatological results. *Applied Optics* 38: 2410–2421.

Hyperspectral Remote Sensing of Land and the Atmosphere

W H Farrand, Space Science Institute, Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Hyperspectral remote sensing is a true marriage of imaging technology with spectroscopy. Hyperspectral remote sensing systems (also known as imaging spectrometer systems) fully sample the optical wavelength range of interest, whether it be the reflected solar range (0.35 to 2.5 μm) or the range of thermal emission from the Earth's surface (3.0 to 14 μm). A hyperspectral sensor views the Earth's surface in a series of contiguous and spectrally narrow image bands. Figure 1 presents the concept of hyperspectral remote sensing, wherein each spatial element of an image has an associated full resolution spectrum. The calibrated reflectance or emittance spectra collected by a hyperspectral system are meant to be directly comparable with that of materials measured in the laboratory. With such high spectral resolution, it thus becomes possible to do reflectance or emittance spectroscopy of the Earth's surface from an overhead perspective.

Underlying Principles

Optical remote sensing is done over wavelength intervals, or windows, in which the atmosphere is largely transparent. For the reflective solar portion of the spectrum, the atmosphere is mostly transparent from approximately 0.35 to 2.5 μm. Water, CO_2, and other gases have absorptions of varying strength in this range. This window is subdivided into the range of human vision (the visible) and the region of slightly longer wavelengths known as the near infrared; together they are the visible/near infrared or VNIR and extend from approximately 0.35 to 1.0 μm. The region from approximately 1.0 to 2.5 μm is known as the short wavelength infrared or SWIR. In the range of emitted terrestrial radiation there are two window regions. The first extends from approximately 3 to 5 μm and is known as the medium wavelength infrared or MWIR, and the second, the long wavelength infrared or LWIR, extends from approximately 8 to 14 μm.

The light that is intercepted by the entrance aperture of a sensor is known by the quantity of radiance which is measured in units of microwatts per square centimeter per nanometer per unit of solid angle. Physical quantities related to the material properties of Earth surface materials are reflectance and emittance. Both properties are the result of ratios and so are unitless. One definition of reflectance is that it is the ratio of radiance reflected from a material, divided by the radiance reflected from an identically illuminated perfectly diffuse reflector. Likewise, a definition for emittance is the ratio of radiance emitted from a material, divided by the ratio

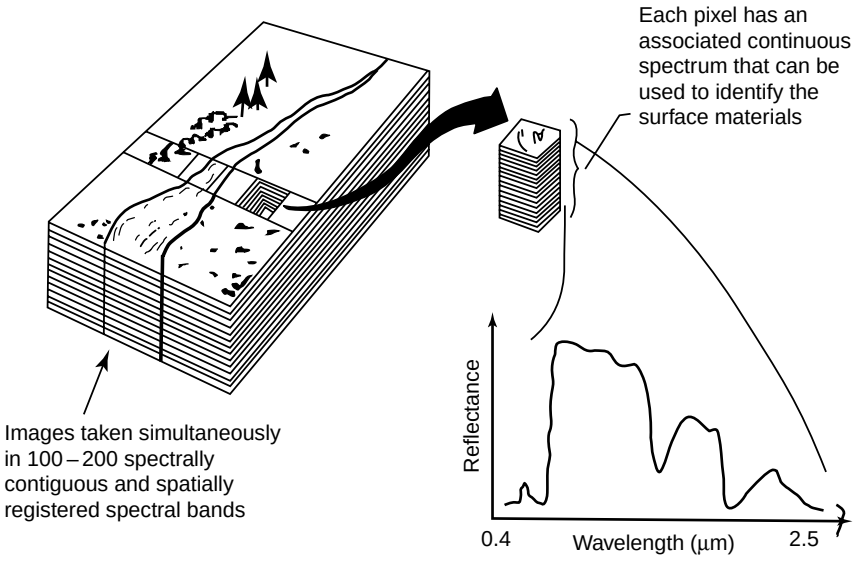


Figure 1 Hyperspectral remote sensing concept.

of radiance emitted from a perfectly emitting material of the same temperature.

Materials covering the Earth's surface, or gases in the atmosphere, can be identified in hyperspectral data on the basis of absorption features, or bands, in the spectra recorded by the sensor. Absorption bands result from the preferential absorption of energy in some wavelength interval. The principal types of processes by which light can be absorbed include, in order of decreasing energy of the process, electronic charge transfers, electronic crystal field effects, and molecular vibrations. The former two processes are observed in minerals and manmade materials that contain transition group metal cations, most usually, iron. Charge transfer absorptions are the result of cation-to-anion or cation-to-cation electron transfers. Crystal field absorptions are explainable by the quantum mechanical theory of atomic structure, wherein an atom's electrons are contained in orbital shells. The transition group metals have incompletely filled *d* orbital shells. When a transition metal cation, such as iron, is surrounded by anions, and potentially selects other cations, an electric field, known as the crystal field, exists. Crystal field absorptions occur when radiant energy causes that cation's orbital shell energy levels to be split by interaction with the crystal field. The reflectance spectra of different iron-bearing minerals are unique, because the iron cation has a unique positioning with respect to the anions and other cations that compose the mineral.

A material's component molecules have bonds to other molecules and as a result of interaction with radiant energy, these bonds can stretch or bend. These molecular vibrations, and overtones of those vibrations, are less energetic processes than the electronic absorption processes described above and so the resulting absorption features occur at longer wavelengths. Fundamental vibrational features occur in the MWIR and LWIR. Overtones and combination overtones of these vibrations are manifested as weaker absorption features in the SWIR. These overtone features in the SWIR include absorptions diagnostic of carbonate and certain clay minerals. Absorption features caused by vibrational overtones of C–H, O–H, and N–H stretches are also manifested in the SWIR and these absorption features are characteristic of important vegetative biochemical components such as cellulose, lignin, starch, and glucose. An example of the effect of the changes in reflected energy recorded at a sensor, due to the absorption of incident energy by green vegetation and other surface materials, is provided in **Figure 2**.

Laboratory spectrometers have been used for many years to analyze the diagnostic absorptions of materials caused by the above effects. More recently, spectrometers were mounted on telescopes to determine the composition of the Moon and the other solid bodies in our solar system. Technology progressed to where profiling spectrometers (instruments that measure a successive line of points on the surface) could be mounted on aircraft. The next logical step was the construction of imaging spectrometers (hyperspectral sensors) that measured spectra in two spatial dimensions. In fact, the data produced by a hyperspectral sensor are often thought of as an image cube (**Figure 3**) because they consist of three dimensions: two spatial and one spectral.

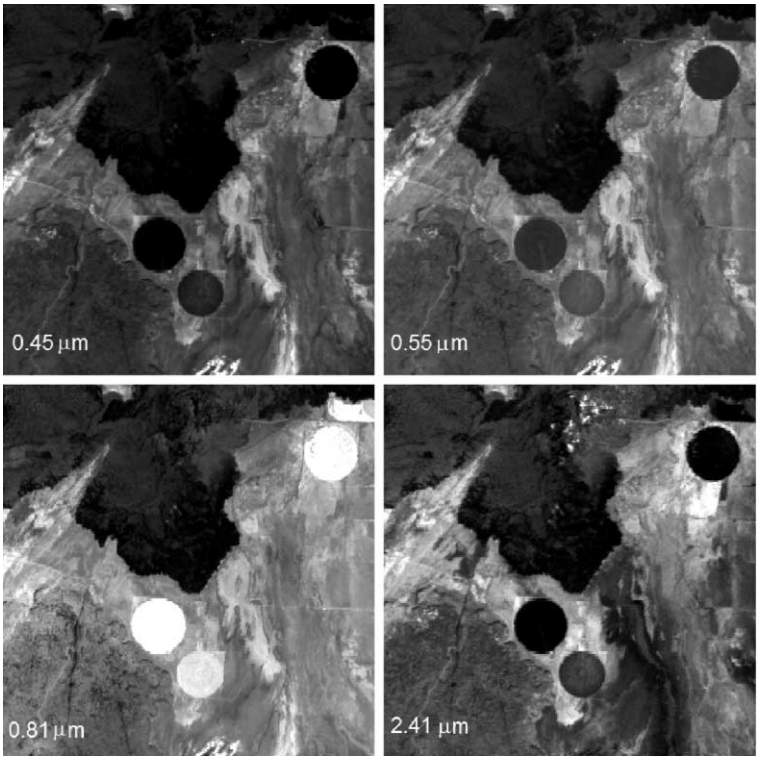


Figure 2 Portion of a scene from the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) from Utah that includes a basalt flow, circularly irrigated grasses sparsely vegetated plains, and damp alluvial sediments. AVIRIS channels centered at 0.45, 0.55, 0.81, and 2.41 μm are shown. Green vegetation is brightest in the 0.81 μm band which samples the peak of the near infrared plateau (see **Figure 7**) just beyond the 'red edge'. The green vegetation and damp sediments are darkest in the 2.41 μm band where water has a very low reflectance (note also **Figure 8**). The bright patches in the basalt flow in the 2.41 μm image are occurrences of oxidized red cinders which have a high reflectance in the SWIR.

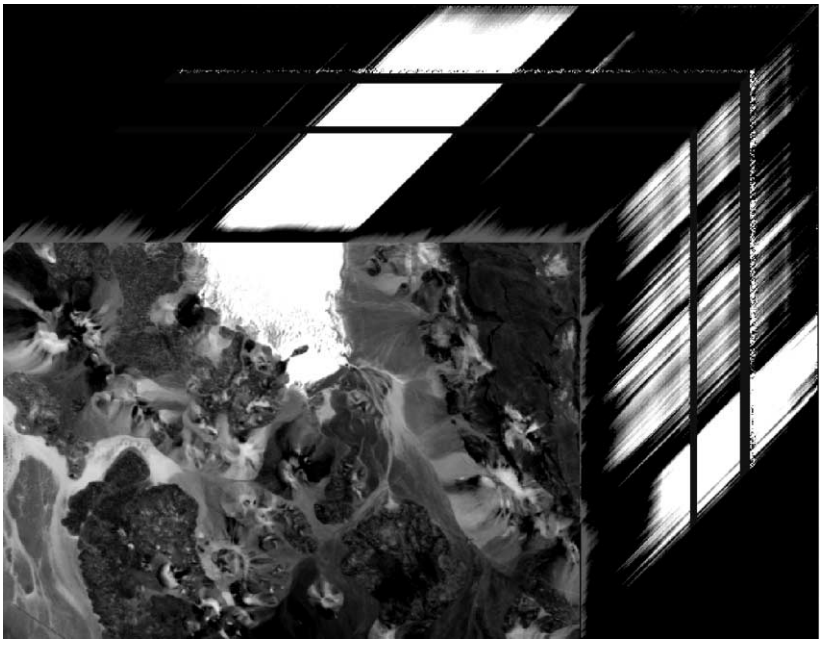


Figure 3 Visualization of a hyperspectral image cube. A 1.7 μm image band from the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over the Lunar Crater Volcanic Field, Nevada is shown with each of the 224 rightmost columns and 224 topmost lines arrayed behind it. In the lower right-hand corner, the low reflectance of ferric oxide-rich cinders in the blue and green show up as dark in the stacking of columns with the high reflectance in the infrared showing up as bright.

Advances in hyperspectral sensor technology have been accompanied by advances in associated technologies, such as computing power and the evolution of data processing algorithms.

Data Processing Approaches

An early impediment to the widespread use of hyperspectral data was the sheer volume of the data sets. For instance, a single scene from NASA's Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) is 614 samples by 512 lines by 224 bands. With its short integer storage format, such a scene takes up 140.8 Mbytes. However, since the dawn of the hyperspectral era, in the 1980s, computing power has expanded immensely and computing tasks, which once seemed prohibitive, are now relatively effortless. Also, numerous processing algorithms have been developed which are especially well suited for use with hyperspectral data.

Spectral Matching

In one class of algorithms, the spectrum associated with each spatial sample is compared against one or a group of model spectra and some similarity metric is applied. The model spectrum can be measured by a laboratory or field portable spectrometer of pure materials or can be a single pixel spectrum, or average of pixel spectra, covering a known occurrence of the material(s) to be mapped. A widely used similarity metric is the Spectral Angle Mapper (SAM). SAM is the angle, θ, obtained from the dot product of the image pixel spectrum, t, and the library spectrum, r, as expressed by

$$\theta = \cos^{-1}\left(\frac{t * r}{\|t\| * \|r\|}\right) \qquad [1]$$

SAM is insensitive to brightness variations and determines similarity based solely on spectral shape. Lower θ values mean more similar spectra.

In a related approach, matching is done, not on the entire spectrum, but rather on characteristic absorption features of the materials of interest. In Figure 4, the geometry of an absorption band is plotted. To map a given mineral, with a specific absorption feature, a high resolution laboratory spectrum is resampled to the spectral resolution of the hyperspectral sensor. The spectrum is subsampled to match the specific absorption feature (i.e., only bands from the short wavelength shoulder of the absorption to the long wavelength shoulder are included). A straight line continuum (calculated, based on a line between the two-band shoulders) is divided out from both the laboratory and the sensor spectra. Contrast between the library and the sensor spectra is mitigated through the use of an additive constant. This constant is incorporated into a set of equations which are solved through the use of standard least squares. The result of such a band-fitting algorithm is two data numbers per spatial sample of the hyperspectral scene. First the band depth (or, alternatively, band area) of the feature is calculated, and second, a goodness-of-fit parameter is calculated. These two parameters are most often combined in a band-depth times goodness-of-fit image. More sophisticated versions of the band-mapping algorithm target multiple absorption features and are, essentially, expert systems.

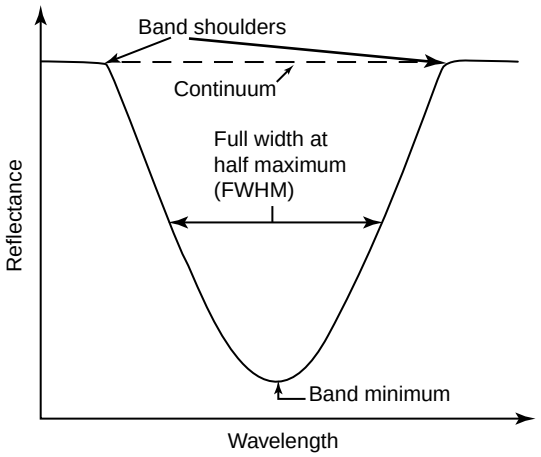


Figure 4 Parameters that describe an absorption band.

Spectral Mixture Analysis

A fundamentally different methodology for analyzing hyperspectral data sets is to model the measured spectrum of each spatial sample as a linear combination of endmember spectra. This methodology is based on the fact that the ground area corresponding to any given spatial sample likely will be covered by more than one material, with the consequence that the measured reflectance or emittance spectrum is a mixed pixel spectrum. The objective of linear SMA is to try to determine the fractional abundance of the component materials, or endmembers. The basis of linear spectral unmixing is to model the response of each spatial element as a linear combination of endmember spectra. The basis equation for linear SMA is

$$r(x, y) = \alpha\mathbf{M} + \varepsilon \qquad [2]$$

where $r(x, y)$ = the relative reflectance spectrum for the pixel at position (x, y), α = the vector of endmember abundances, $\mathbf{M}$ = the matrix of endmember spectra, and ε = the vector of residuals between the modeled and the measured reflectances. Application

of SMA results in a series of fraction images for each endmember wherein the data numbers range, ideally, between 0 and 1. Fraction image pixels with a digital number (DN) of 0 are taken to be devoid of the endmember material. Fraction image pixels with a DN of 1.0 are taken to be completely covered with the endmember material.

Techniques related to SMA have been developed which also map the abundance of a target material on a per pixel basis. These include implementations highly similar to SMA such as Orthogonal Subspace Projection (OSP) wherein, as with SMA, all endmembers must be determined a priori. Other approaches such as Constrained Energy Minimization (sometimes referred to as 'Matched Filter') do not require a priori knowledge of the endmembers. Instead, only the reflectance or emittance spectrum of the target need be known and the undesired spectral background is estimated, and ultimately compensated for, using the principal eigenvectors of the sample correlation or covariance matrix of the hyperspectral scene.

Expert Systems and Artificial Neural Networks

Another means of extracting information from hyperspectral data sets is to use computer approaches that in some way mimic the human thought process. This can take the form of an expert system which applies a set of rules or tests as each pixel spectrum is successively analyzed. For example, if a spectrum has an absorption at 2.2 μm it could tentatively be classified as being caused by a clay mineral. Additionally, if the absorption is a doublet, the specific clay mineral is likely a kaolinite. In practice, expert system codes are lengthy and complex although excellent results can be demonstrated from their use, albeit at the expense of processing time. Another approach in this vein is the use of an artificial neural network (ANN) for data processing. The use of ANNs is motivated by their power in pattern recognition. ANN architectures are well suited for a parallel processing approach and thus have the potential for rapid data processing.

Hyperspectral Remote Sensing of the Land

Geology

Early efforts with airborne hyperspectral sensors focused on geologic remote sensing. This was, at least in part, a consequence of the mineralogic diversity of the Earth's surface and the fact that many minerals have absorption features which are unique and diagnostic of the mineral's identity. The expectation, which was soon borne out by results from these early sensors, was that, given exposure of the surface, mineralogic maps as detailed as the spatial resolution of the sensor could be derived from hyperspectral data. As some of the processing techniques discussed above have become more readily available, it has become possible to produce mineralogic maps from hyperspectral data far more rapidly and cost-effectively than geologic maps produced by standard means (e.g., by walking over the ground and manually noting which lithologies are underfoot).

The propensity for transition metals, most especially iron, to absorb energy through charge transfer and crystal field effects, is noted above. Fe–O and Fe^{2+}–Fe^{3+} charge transfers cause a profound absorption in the reflectance spectrum of Fe-bearing minerals shortwards of 0.4 μm. The wing of this absorption causes the low blue and green reflectance of Fe-bearing minerals. Crystal field bands cause an absorption feature in the 0.9 to 1.0 μm region. The ability to discriminate subtle differences among Fe-bearing minerals in hyperspectral data has proven extremely valuable in applications such as mineral exploration, volcanology, and in the characterization of abandoned mine lands.

In the SWIR, absorptions caused by vibrational overtones of molecular bonds within minerals. These include absorptions in the 2.2 and 2.3 μm region which are characteristic of many Al- and Mg-bearing clay minerals. Absorptions in the SWIR are narrower in width than the Fe-generated crystal field bands discussed previously. In fact, the requirement to efficiently resolve these vibrational overtone absorption bands (which have full width at half maximum (FWHM) bandwidths of 20 to 40 nm) helped to drive the selection of the nominal 10 nm bandwidth of early and most current hyperspectral sensors. Certain clay minerals can be indicators of hydrothermal activity associated with economic mineral deposits; thus, the SWIR is an important spectral region for mineral exploration. Clay minerals, by definition, have planar molecular structures that are prone to failure if subjected to shearing stresses. The ability to uniquely identify and map these minerals, using hyperspectral data, has thus been used to good effect to identify areas on volcanoes and other hydrothermally altered terrains that could be subject to landslides.

Minerals with adsorbed or molecularly bound OH^- and/or water have vibrational overtone absorptions near 1.45 and 1.9 μm, although these features are masked in remotely sensed data by prominent atmospheric water vapor bands at 1.38 and 1.88 μm. The reflectance of water and OH-bearing minerals decrease to near zero at wavelengths longwards of

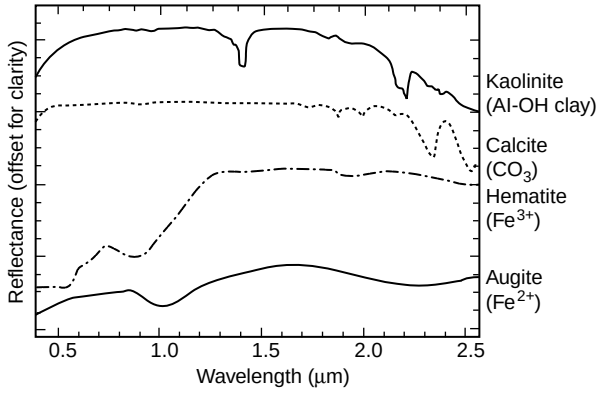


Figure 5 Reflectance spectra of representative minerals. Kaolinite represents clay minerals with its diagnostic 2.2 μm band caused by an overtone vibration of the Al–OH bond. The 2.335 μm band of calcite is caused by an overtone vibration of the carbonate molecule. The features shortwards of 1.0 μm in the hematite spectrum are caused by crystal field and charge transfer absorptions due to Fe^{3+}. The bands near 1.0 and longwards of 2.0 μm in the augite spectrum are the result of Fe^{2+} crystal field bands.

2.5 μm, due to the presence of water and OH fundamental vibrations near 3.0 μm. Minerals bearing the carbonate (CO_3^{-2}) molecule are abundant on the Earth's surface due to the ubiquitous occurrences of limestone and dolomite. Calcite and dolomite have overtone absorptions centered at 2.335 and 2.315 μm, respectively. A good test for the spectral resolution and spectral calibration of a hyperspectral sensor is its ability to successfully discriminate calcite from dolomite on the basis of the aforementioned absorption features. Figure 5 shows representative spectra of carbonate, clay, and ferric and ferrous iron-bearing minerals.

Absorptions resulting from fundamental molecular absorptions are manifested in the LWIR. These absorptions are of great geologic interest because they include absorptions resulting from vibrations of the Si–O bond and silicate minerals form the bulk of the Earth's crust. The wavelength at which the Si–O stretching feature occurs is dictated by the level of polymerization (or molecule-to-molecule bonding) of the silicate mineral. Highly polymerized framework silicate minerals, such as quartz and feldspar, have a shorter wavelength absorption than do silicate minerals, such as olivine which are composed of disconnected SiO_4 molecules. In Figure 6, laboratory emission spectra of highly and poorly polymerized silicate minerals are shown.

Vegetation and the Environment

While different minerals are, by definition, generally composed of diverse component molecules, different species of vegetation represent variations of the same general biochemical constituents (e.g., chlorophyll, proteins, lignin, cellulose, sugar, starch, etc.). In the thermal IR, vegetation has a generally flat emittance, thus it is in the reflective solar spectrum that most vegetative remote sensing is performed. The major features in the reflectance spectrum of green vegetation are shown in Figure 7. In the visible, the major absorption features are caused by the presence of chlorophyll. Chlorophyll has strong absorptions in the blue and the red, leaving a reflectance maximum (the green peak) at 0.55 μm. In the NIR, scattering in the leaf structure causes high reflectance leaving a profound absorption edge (the red edge) between

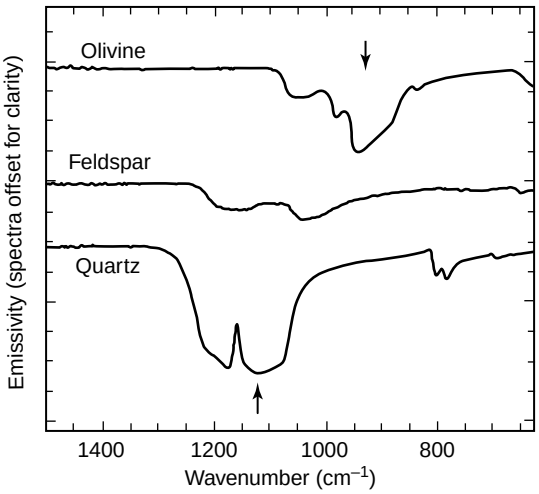


Figure 6 Emissivity spectra of SiO_4-bearing minerals showing the shift in band minimum to higher wave number with increasing polymerization.

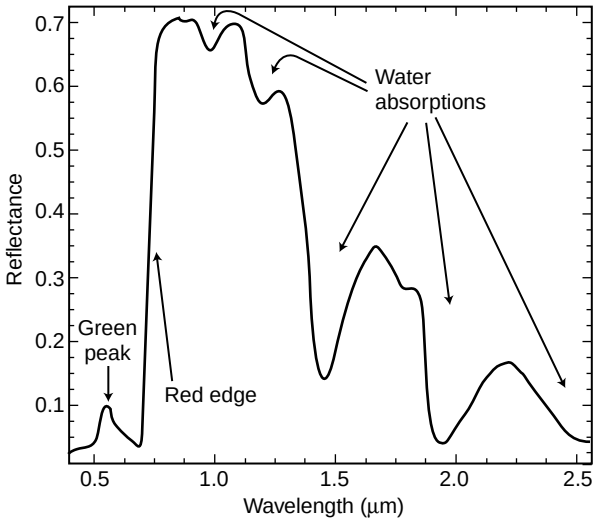


Figure 7 Major spectral features in the reflectance spectrum of green vegetation.

0.7 and 0.78 μm. Past the red edge and into the SWIR, the spectrum of green vegetation is dominated by water in the leaf, with leaf water absorptions occurring at 0.97, 1.19, 1.45, 1.93, and 2.50 μm. Studies of vegetation, using multispectral systems, have made use of a number of broadband indices. Hyperspectral systems allow for the discrimination of individual absorption features and subtle spectral shape differences in vegetation spectra which are unresolvable using broadband multispectral systems. For example, the absorptions caused by chlorophyll can be used for studies of chlorophyll content in vegetative canopies. The unique interaction of chlorophyll with leaf structures can provide a morphology to the green peak that is unique to a given species and thus is mappable using a spectral feature fitting approach.

A desired goal in the study of vegetation and ecosystem processes is to be able to monitor the chemistry of forest canopies. The foliar biochemical constituents making up forest canopies have associated absorption features in the SWIR, that result from vibrational overtones and combinations of C–O, O–H, C–H, and N–H molecular bonds. However the absorptions from these biochemical components overlap so that the chemical abundance of any one plant component cannot be directly related to any one absorption feature. An even more serious challenge is introduced by the strong influence of water in the SWIR in green vegetation spectra. Water constitutes 40 to 80% of the weight of leaves. However, the unique characteristics of high-quality hyperspectral data (e.g., excellent radiometric and spectral calibration, high signal-to-noise ratio, high spectral resolution) have been used to detect even these subtle features imprinted on the stronger water features.

The amount of water present in plant leaves is, in itself, a valuable piece of information. Hyperspectral data can be used to determine the equivalent water thickness present in vegetative canopies. Conversely, the amount of dry plant litter and/or loose wood can be estimated using spectral mixture analysis and related techniques. Taken together, all these data: vegetation species maps, determinations of leaf water content and the relative fractions of live vegetation versus litter, can be used to characterize woodlands for forest fire potential. The ability to map different species of vegetation has also proved useful in the field of agriculture, for assessing the impact of invasive weed species on farm and ranch land. Being able to assess the health of crops, in terms of leaf water content, is also important for agriculture.

Another subtle vegetative spectral feature, that provides information on plant vitality, is the aforementioned red edge. Shifts in the position of the red edge have been linked to vegetation stress. Studies have shown that vegetation stress and consequent shifts in the position of the red edge can be caused by a number of factors, including insufficient water intake or the intake of pernicious trace metals. Vegetation growing over mineral deposits can display red edge shifts and this can be used to assist in mineral exploration efforts.

Another application area, in which hyperspectral remote sensing has found great utility, is in the analysis of snow-covered areas. Over the reflective solar spectrum, the reflectance of snow varies from values near zero in the SWIR to values near one in the blue. Differences in the grain size of snow result in differences in reflectance as is illustrated in Figure 8. The spectral variability of snow makes it easily distinguishable from other Earth surface materials.

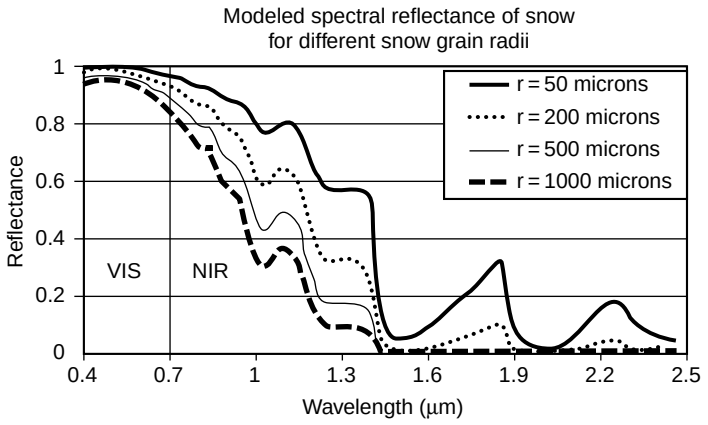


Figure 8 Differences in snow reflectance spectra resulting from differences in snow grain size. Courtesy of Dr Anne Nolin of Oregon State University.

Linear spectral unmixing, using hyperspectral data, has been shown to successfully discriminate between snow, vegetation, rock, and clouds and to accurately map snow cover fraction in mixed pixels. Accurate maps of snow cover, types of snow, and fractions of liquid water admixed with snow grains, are required for forecasting snowmelt runoff and stream discharge in watersheds dominated by snow cover.

Hyperspectral Remote Sensing of the Atmosphere

The physical quantity most often sought in land remote sensing studies is surface reflectance. However, the quantity recorded by a hyperspectral sensor is radiance at the entrance aperture of the sensor. In order to obtain reflectance, the background energy level of the Sun and/or Earth must be removed and the scattering and absorbing effects of the atmosphere must be compensated for. In the VNIR through SWIR, there are seven atmospheric constituents with significant absorption features: water vapor, carbon dioxide, ozone, nitrous oxide, carbon monoxide, methane, and oxygen. Molecular scattering (commonly called Rayleigh scattering) is strong in the blue but decreases rapidly with increasing wavelength. Above 1 μm, its effect is negligible. Scattering caused by atmospheric aerosols, or Mie scattering, also is more prominent at shorter wavelengths and decreases with increasing wavelength, but the dropoff of effects from Mie scattering is not as profound as that for Rayleigh scattering. Consequently, effects from aerosol scattering can persist into the SWIR.

There are three primary categories of methods for removing the effects of atmosphere and solar insolation from hyperspectral imagery, in order to derive surface reflectance. Methods of atmospheric corrections can be considered as being either an image-based, empirical, or model-based approach. An image-based, or 'in-scene' approach uses only data measured by the instrument. Empirical methods make use of the remotely sensed data in combination with field measurements of reflectance, to solve a simplified equation of at-sensor radiance such as eqn [3].

$$L_s = A\rho + B \qquad [3]$$

where L_s is the at-sensor radiance, ρ is the reflectance of the surface, and A and B are quantities that incorporate, respectively, all multiplicative and additive contributions to the at-sensor radiance. All the quantities expressed in eqn [3] can be considered as varying as a function of wavelength, λ. Approximations for A and B from eqn [3] constitute a set of gains and offsets derived from the empirical approach. In some empirical approaches, atmospheric path radiance is ignored and only a multiplicative correction is applied.

Model-based approaches seek to model what the at-sensor radiance should be on a pixel-by-pixel basis, including the contribution of the atmosphere. The at-sensor radiance, L_s, at any wavelength, λ, can be expressed as

$$L_s = \frac{1}{\pi}(E\rho + M_T)\tau_\theta + L_u \qquad [4]$$

where E is the irradiance at the surface of the Earth, ρ is the reflectance of the Earth's surface, M_T is the spectral radiant exitance of the surface at temperature, T, τ_θ is the transmissivity of the atmosphere at zenith angle θ, and L_u is the spectral upwelling path radiance of the atmosphere. The ability to solve eqn [4] and do atmospheric correction on a pixel-by-pixel basis is appealing, in that it negates the shortcomings of an empirical approach where a correction based on calibration targets in one part of a scene might not be appropriate for another part of the scene, due to differences in atmospheric pathlength or simple atmospheric heterogeneity. Model-based approaches are also able to take advantage of the greater spectral dimensionality of hyperspectral data sets for the derivation of atmospheric properties (e.g., the amount of column water vapor, CO_2 band depths, etc.) directly from the data.

The main atmospheric component, affecting imagery in the reflective solar portion of the spectrum, is water vapor. Model-based atmospheric correction techniques determine the column water vapor for each pixel in the scene, based on the depth of the 940 and/or 1130 nm atmospheric water bands. Thus for each pixel, an appropriate amount of water can be removed. The maps of atmospheric water vapor distribution are themselves of interest to atmospheric scientists. Absorption features caused by well-mixed gases, such as the 760 nm O_2 band, can be used by model-based programs, to produce images of approximate scene topography. More advanced model-based approaches also solve for the scattering effects of aerosols on the hyperspectral data.

As noted above, atmospheric constituents, including industrial effluents, can be detected and spatially mapped by hyperspectral sensors. The fundamental vibrational absorption of gases of interest occurs in the MWIR and the LWIR. Overtones of these gaseous molecular vibrations occur in the VNIR to SWIR, but are generally too weak to be detected. The ability to detect anthropomorphically produced gases depends on a number of factors, including the

temperature difference between the gas cloud and the background atmosphere, the concentration of gas within the cloud, and the size of the cloud (e.g., the path length of light through the cloud).

See also

Imaging: Infrared Imaging. **Instrumentation:** Spectrometers.

Further Reading

Clark RN (1999) Spectroscopy of rocks and minerals, and principles of spectroscopy. In: Rencz AN (ed.) *Remote Sensing for the Earth Sciences*, pp. 3–57. New York: John Wiley and Sons.

Farrand WH and Harsanyi JC (1997) Mapping the distribution of mine tailings in the Coeur d'Alene River Valley, Idaho through the use of a Constrained Energy Minimization technique. *Remote Sensing Environment* 59: 64–76.

Gao B-C and Goetz AFH (1995) Retrieval of equivalent water thickness and information related to biochemical components of vegetation canopies from AVIRIS data. *Remote Sensing Environment* 52: 155–162.

Gao B-C, Heidebrecht KH and Goetz AFH (1993) Derivation of scaled surface reflectances from AVIRIS data. *Remote Sensing Environment* 44: 165–178.

Goetz AFH (1989) Spectral remote sensing in geology. In: Asrar G (ed.) *Theory and Applications of Optical Remote Sensing*, pp. 491–526. New York: John Wiley & Sons.

Goetz AFH, Vane G, Solomon JE and Rock BN (1985) Imaging spectrometry for Earth remote sensing. *Science* 228: 1147–1153.

Hapke B (1993) *Theory of Reflectance and Emittance Spectroscopy*. Cambridge, UK: Cambridge University Press.

Kirkland LE, Herr KC and Salisbury JW (2001) Thermal infrared spectral band detection limits for unidentified surface materials. *Applied Optics* 40: 4852–4862.

Martin ME, Newman SD, Aber JD and Congalton RG (1998) Determining forest species composition using high spectral resolution remote sensing data. *Remote Sensing Environment* 65: 249–254.

Painter TH, Roberts DA, Green RO and Dozier J (1998) The effect of grain size on spectral mixture analysis of snow-covered area with AVIRIS data. *Remote Sensing Environment* 65: 320–332.

Pieters CM and Englert PAJ (eds) *Remote Geochemical Analysis: Elemental and Mineralogical Composition*. Cambridge, UK: Cambridge University Press.

Pinzon JE, Ustin SL, Castaneda CM and Smith MO (1997) Investigation of leaf biochemistry by hierarchical foreground/background analysis. *IEEE Transactions on Geoscience and Remote Sensing* 36: 1–15.

Roberts DA, Green RO and Adams JB (1997) Temporal and spatial patterns in vegetation and atmospheric properties using AVIRIS. *Remote Sensing Environment* 62: 223–240.

Wessman CA, Aber JD and Peterson DL (1989) An evaluation of imaging spectrometry for estimating forest canopy chemistry. *International Journal of Remote Sensing* 10: 1293–1316.

Laser Detection of Atmospheric Gases

E V Browell, W B Grant and S Ismail, NASA Langley Research Center, Hampton, VA, USA

Published by Elsevier Ltd.

Introduction

Lidar (light detection and ranging) systems are able to measure profiles of atmospheric aerosols, clouds, and gases by transmitting a pulsed laser beam into the atmosphere and collecting the backscattered radiation from aerosols and molecules in the atmosphere with a receiver located near the transmitter. The differential absorption lidar (DIAL) approach is the most widely used technique for measuring a variety of gases. In this approach, two closely spaced laser wavelengths are used: one which is absorbed by the gas of interest, and the other which is only weakly, or not at all, absorbed by the gas. A differential with respect to range and wavelength is calculated, to determine the average gas concentration along any segment of the lidar path, using the familiar Beer–Lambert law for an absorbing medium.

The DIAL equation can be expressed in its simple form as

$$N = \left(1/\left(2(R_2 - R_1)(\sigma_{on} - \sigma_{off})\right)\right) \times \ln\left(P_{off}(R_2)P_{on}(R_1)\right)/\left(P_{off}(R_1)P_{on}(R_2)\right) \quad [1]$$

where N is the average gas concentration, R is the range, σ_{on} and σ_{off} are the absorption cross sections at the on- and off-line wavelengths, and $P_{on}(R)$ and $P_{off}(R)$ are the powers received at the on- and off-line wavelengths. However, since it is performed in the atmosphere and not a laboratory, great care must be taken to ensure that the data are analyzed properly to minimize random and systematic errors. Random errors arise from noise in the backscattered signal, the solar background signal, and the inherent

detector noise, and this type of error can be reduced by signal averaging. Systematic errors arise from uncompensated instrument and atmospheric effects and must be carefully considered in system design and operation and data processing.

Another approach, the Raman lidar approach, uses inelastic scattering (scattered/emitted light has different wavelength than the illuminating light) from gases where the wavelength shift corresponds to vibrational or rotational energy levels of the molecules. Any illuminating laser wavelength can be used, but since the Raman scattering cross-section varies as λ^{-4}, where λ is the wavelength, the shorter laser wavelengths, such as in the near UV spectral region, are preferred. While even shorter wavelengths in the solar blind region below ~300 nm would permit operation in daytime, the atmospheric attenuation, due to Rayleigh scattering and UV molecular absorption, limits the measurement range. High-power lasers are used due to the low Raman scattering cross-sections. The Raman lidar measurements have to be carefully calibrated against a profile measured using another approach, such as is done for water vapor (H_2O) measurements with the launch of hygrometers on radiosondes, since the lidar system constants and atmospheric extinctions are not usually well known or modeled. In order to obtain mixing ratios of H_2O with respect to atmospheric density, the H_2O signals are ratioed to the nitrogen Raman signals. However, care must be exercised in processing these data due to the spectral dependences of atmospheric extinction.

Surface-Based Lidar Systems

Surface-based lidar systems measure the temporal evolution of atmospheric profiles of aerosols and gases. The first lidar systems used to remotely measure atmospheric gases were Raman lidar systems. The Raman lidar systems were thought to be very promising since one high-power laser could be used to measure a variety of gases. They are easier to develop and operate than DIAL systems, because the laser does not have to be tuned to a particular gas absorption feature. However, the Raman scattering cross-section is low, and the weak, inelastically scattered signal is easily contaminated by daylight background radiation, resulting in greatly reduced performance during daytime. Raman lidar systems have been developed primarily for measuring H_2O and retrieving atmospheric temperature.

Surface-based UV DIAL systems, that are part of the Network for the Detection of Stratospheric Change (NDSC) have made important contributions to the understanding of stratospheric O_3. There are DIAL systems at many locations around the Earth, with sites in Ny-Ålesund, Spitzbergen (78.9°N, 11.9°E), Observatoire Haute Provence, France (43.9°N, 5.7°E), Table Mountain, California (34.4°N, 118.2°W), Mauna Loa, Hawaii (19.5°N, 155.6°W), and Lauder, New Zealand (45.0°S, 169.7°E). First established in the 1980s, these DIAL systems are strategically located so that O_3 in different latitude bands can be monitored for signs of change. In addition, they can provide some profiles for comparison with space-based O_3 measuring instruments. The parameters of a typical ground-based UV DIAL system used in the NDSC are given in Table 1.

Airborne DIAL Systems

Airborne lidar systems expand the range of atmospheric studies beyond those possible by surface-based lidar systems, by virtue of being located in aircraft that can be flown to high altitudes and to remote locations. Thus, they permit measurements at locations inaccessible to surface-based lidar systems. In addition, they can make measurements of large atmospheric regions in times that are short compared with atmospheric motions, so that the large-scale patterns are discernible. Another advantage of airborne lidar operation is that lidar systems perform well in the nadir (down) direction since the atmospheric density and aerosol loading generally increases with decreasing altitude towards the surface, which helps to compensate for the R^{-2} decrease in the lidar

Table 1 Parameters for the Jet Propulsion Laboratory's Mauna Loa DIAL systems for stratospheric O_3 measurements

	XeCl	Nd:YAG (3rd Harmon.)
Wavelength (nm)	308 (on)	355 (off)
Pulse energy (mJ)	300	150
Pulse repetition frequency (Hz)	200	100
Area (m^2)	0.79	
Optical efficiency (%)	~40	
Wavelengths–Rayleigh (nm)	308	355
N_2 Raman (nm) (for aerosol correction)	332	387
Measurement range* (km)	15–55	
Vertical resolution (km)	3 at bottom, 1 at O_3 peak, 8–10 at top	
Measurement averaging time (hours)	1.5	
Measurement accuracy	<5% at peak, 10–15% at 15 km, >40% at 45 km	

*Chopper added to block beam until it reaches 15 km in order to avoid near field signal effects.

signal with range. For the zenith direction, the advantage is that the airborne lidar system is closer to the region being measured.

Field Measurement Programs

Global O_3 Measurements

The first airborne DIAL system, which was developed by the NASA Langley Research Center (LaRC), was flown for O_3 and aerosol investigations in conjunction with the Environmental Protection Agency's Persistent Elevated Pollution Episodes (PEPE) field experiment, conducted over the east coast of the US in the summer of 1980. This initial system has evolved into the advanced UV DIAL system that will be described in the next section. Airborne O_3 DIAL systems have also been developed and used in field measurement programs by several other groups in the United States, Germany, and France.

The parameters of the current NASA LaRC airborne UV DIAL system are given in Table 2. The on-line and off-line UV wavelengths of 288.2 and 299.6 nm are used for DIAL O_3 measurements during tropospheric missions, and 301 and 310 nm are used for stratospheric missions. This system also transmits 1,064 and 600 nm beams for aerosol and cloud measurements. The time delay between the on- and off-wavelength pulses is 400 μs, which is sufficient time for the return at the first set of wavelengths to end, but short enough that the same region of the atmosphere is sampled. This system has a demonstrated absolute accuracy for O_3 measurements of better than 10% or 2 ppbv (parts per billion by volume), whichever is larger, and a measurement precision of 5% or 1 ppbv with a vertical resolution of 300 m and an averaging time of 5 minutes (about 70 km horizontal resolution at typical DC-8 ground speeds).

The NASA LaRC airborne UV DIAL systems have made significant contributions to the understanding of both tropospheric and stratospheric O_3, aerosols, and clouds. These systems have been used in 18 international and 3 national field experiments over the past 24 years, and during these field experiments, measurements were made over, or near, all of the oceans and continents of the world. A few examples of the scientific contributions made by these airborne UV DIAL systems are given in Table 3.

The NASA LaRC airborne UV DIAL system has been used extensively in NASA's Global Tropospheric Experiment (GTE) program which was started in the early 1980s, had as its primary mission the study of tropospheric chemistry in remote regions of the Earth, in part to study and gain a better understanding of atmospheric chemistry in the unperturbed atmosphere. A related goal was to document the Earth's atmosphere in a number of places during seasons when anthropogenic influences are largely absent, then return to these areas later to document the changes. The field missions have included campaigns in Africa, Alaska, the Amazon Basin,

Table 2 Parameters of the NASA LaRC airborne UV DIAL system

L [illegible]: Nd:YAG-pumped dye lasers, frequency doubled into the UV			
Pulse repetition frequency (Hz)	30		
Pulse length (ns)	8–12		
Pulse energy (mJ) at 1.06 μm	250–300		
Pulse energy (mJ) at 600 nm	50–70		
UV pulse energy (mJ)			
For troposphere at 288/300 nm	20		
For stratosphere at 301/310 nm	20		
Dimensions (l × w × h) (cm)	594 × 102 × 109		
Mass (kg)	1735		
Power requirement (kW)	30		
		Wavelength region (nm)	
R [illegible]	289–311	572–622	1064
Area (m^2)	0.086	0.086	0.864
Receiver optical efficiency (%)	30	40	30
Detector quantum efficiency (%)	26 (PMT)	8 (PMT)	40 (APD)
Field-of-view (mrad)	≤1.5	≤1.5	≤1.5
[illegible]			
Measurement range (km)	up to 10–15 (nadir and zenith)		
Vertical resolution (m)	300–1500, depending on range		
Horizontal resolution (km)	≤70		
Measurement accuracy	≤10% or 2 ppbv, whichever is greater		

Table 3 Examples of significant contributions of airborne O_3 DIAL systems to the understanding of tropospheric and stratospheric O_3 (see also **Figures 1–5**)

[illegible]:	
	Air mass characterizations in a number of remote regions
	Continental pollution plume characterizations (see **Figures 1** and **2**)
	Study of biomass burn plumes and the effect of biomass burning on tropospheric O_3 production (see **Figure 3**)
	Case study of warm conveyor belt transport from the tropics to the Arctic
	Study of stratospheric intrusions and tropopause fold events (see **Figure 4**)
	Observation of the decay of a cutoff low
	Power plant plume studies
[illegible]:	
	Contributions to the chemical explanation of behavior of Antarctic O_3
	Quantification of O_3 depletion in the Arctic (see **Figure 5**)
	Polar stratospheric clouds – particle characterizations
	Intercomparison with surface-based, airborne and space-based instruments
	Quantification of O_3 reduction in tropical stratosphere after the June 1991 eruption of Mount Pinatubo and characterization of the tropical stratospheric reservoir edge
	Cross-vortex boundary transport

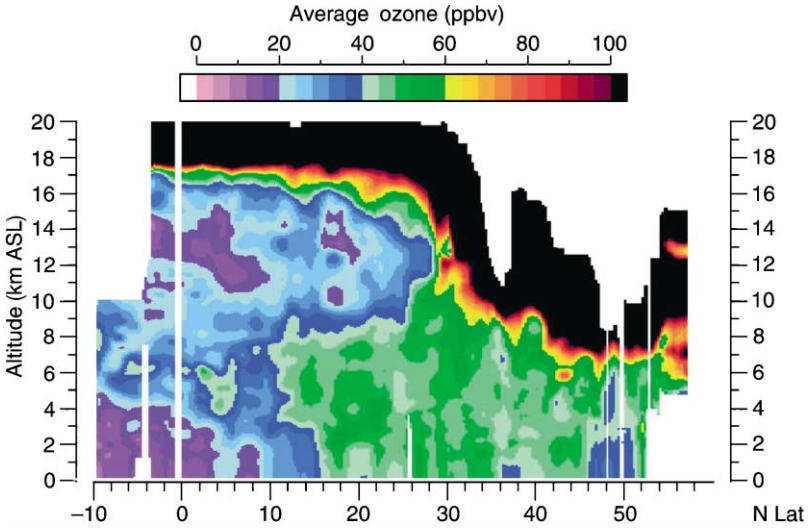


Figure 1 Latitudinal distribution of ozone over the western Pacific Ocean obtained during the second Pacific Exploratory Mission (PEM West B) in 1994.

Canada, and North America, and over the Pacific Ocean from Antarctica to Alaska, with concentrated measurements off the east coast of Asia and in the tropics. One of the bigger surprises of the two decades of field missions, was the discovery in the mid-1990s, that the tropical and South Pacific Ocean had very high tropospheric O_3 concentrations in plumes from biomass burning in Africa and South America during the austral spring. There were few indications from surface-based or space-based measurements that there were extensive biomass burn plumes in the area, primarily since the plumes were largely devoid of aerosols due to being stripped out during cloud convective lofting. Once over the ocean, horizontal transport appears to proceed relatively unimpeded unless a storm system is encountered.

The NASA LaRC airborne UV DIAL system has also been flown in all the major stratospheric O_3 campaigns starting in 1987 with the Airborne Antarctic Ozone Experiment (AAOE) to determine the cause of the Antarctic ozone hole. The UV DIAL system documented the O_3 loss across the ozone hole region. Later, when attention was turned to the Arctic and the possibility of an ozone hole there, the system was used to produce an estimate of O_3 loss during the winter season as well as to better characterize the polar stratospheric cloud (PSC) particles. The UV DIAL system was also used to study O_3 loss in the tropical stratospheric reservoir following the eruption of Mount Pinatubo in June 1991. The loss was spotted by the Microwave Limb Sounder (MLS) on the Upper Atmospheric Research Satellite (UARS), but the MLS

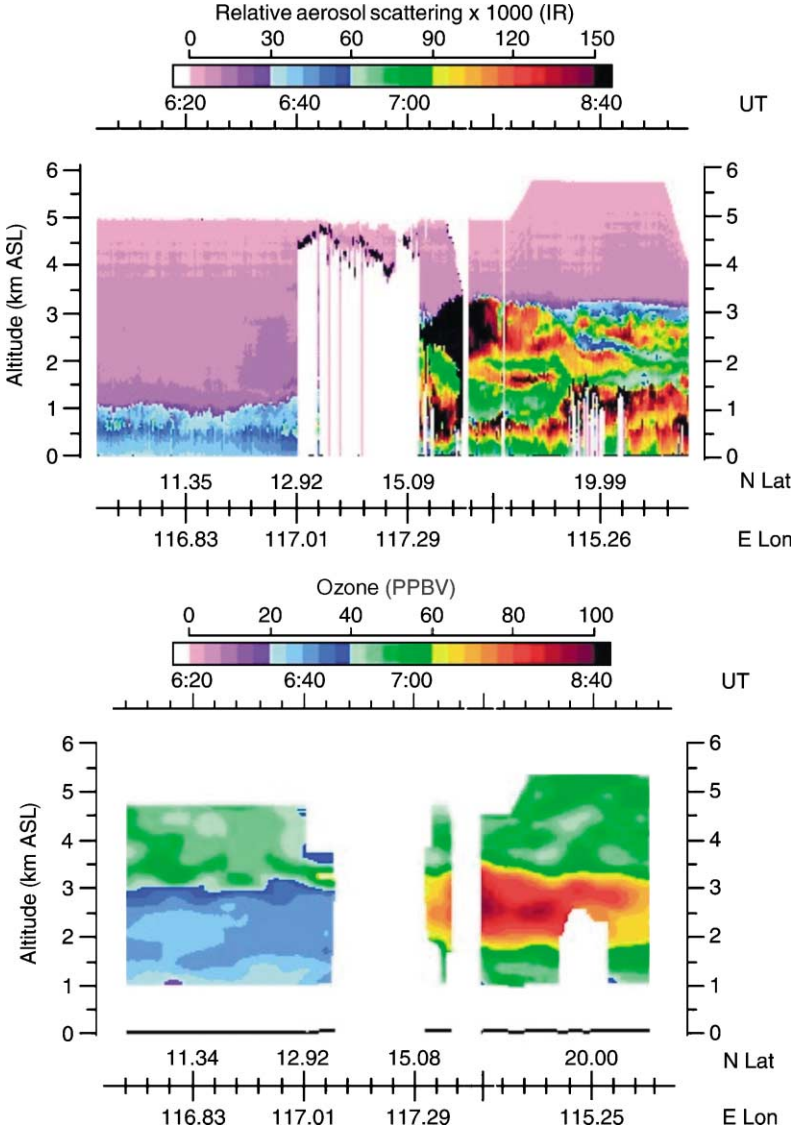


Figure 2 Pollution outflow from China over the South China Sea (right side of figure) with clean tropical air on south side of a front (left side of figure), observed during PEM West B in 1994.

was unable to study the loss in detail due to its low vertical resolution (5 km) compared to the small-scale (2–3 km) features of the O_3 loss. Other traditional space-based O_3 measuring instruments also had difficulty during this period, due to the high aerosol loading in the stratosphere following the eruption.

Space-Based O_3 DIAL System

In order to obtain nearly continuous, global distributions of O_3 in the troposphere, a space-based O_3 DIAL system is needed. A number of key issues could be addressed by a space-based O_3 DIAL system including: the global distribution of photochemical O_3 production/destruction and transport in the troposphere; location of the tropopause; and stratospheric O_3 depletion and dynamics. High-resolution airborne O_3 DIAL and other aircraft measurements show that to study tropospheric processes associated with biomass burning, transport of anthropogenic pollutants, tropospheric O_3 chemistry and dynamics, and stratosphere–troposphere exchange, a vertical profiling capability from space with a resolution of 2–3 km is needed, and this capability cannot currently be achieved using passive remote sensing satellite instruments. An example of the type of latitudinal O_3 cross-section that could be provided by a space-based O_3 DIAL system is shown in Figure 1. This figure shows many different aspects of O_3 loss and production; vertical and horizontal transport; and stratosphere–troposphere exchange that occurs from the tropics to high latitudes.

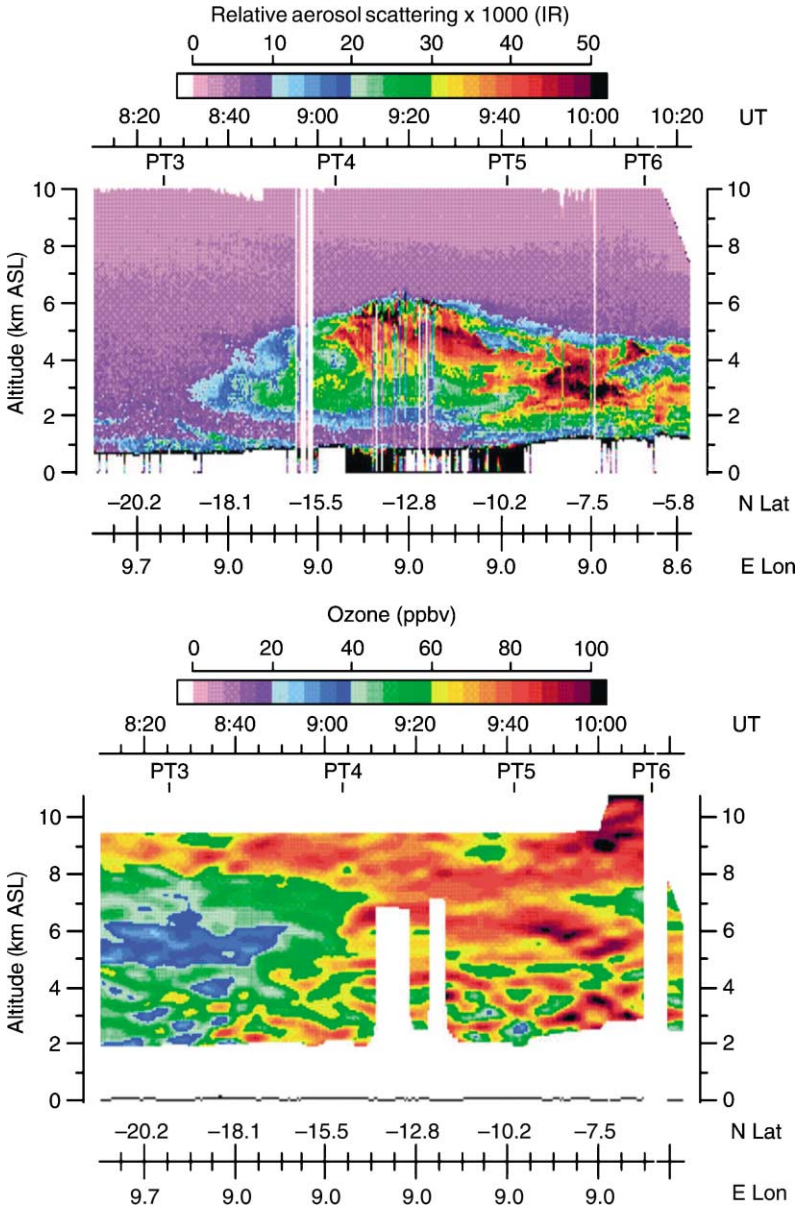


Figure 3 Biomass burning plume over the Atlantic Ocean arising from biomass burning in the central part of western Africa, observed during the TRACE-A mission in 1992.

This type of data would be available from just one pass from a space-based O_3 DIAL system. A space-based O_3 DIAL system optimized for tropospheric O_3 measurements (see system description in Table 4a,b) would also permit high-resolution O_3 measurements in the stratosphere (1 km vertical, 100 km horizontal), along with high-resolution aerosol measurements (100 m vertical, 10 km horizontal). In addition, these DIAL measurements will be useful in assisting in the interpretation of passive remote sensing measurements and in helping to improve their data processing algorithms.

Global H_2O Measurements

H_2O and O_3 are important to the formation of OH in the troposphere, and OH is at the center of most of the chemical reactions in the lower atmosphere. In addition H_2O is an excellent tracer of vertical and horizontal transport of air masses in the troposphere, and it can be used as a tracer of stratosphere–troposphere exchange. Increased aerosol sizes, due to high relative humidities, can also affect heterogeneous chemical processes and radiation budgets in the boundary layer and in cloud layers. Knowledge of H_2O is important to weather forecasting and climate

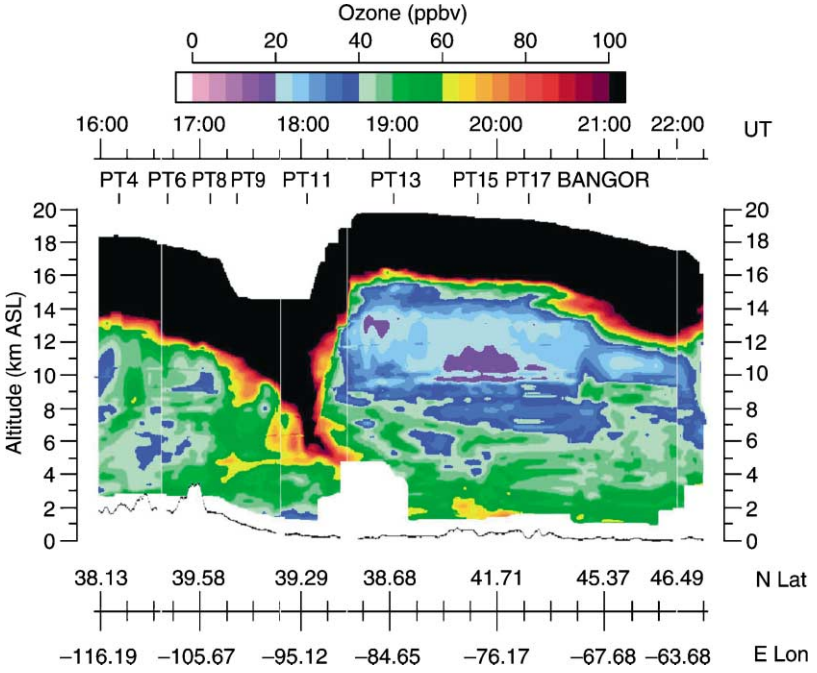


Figure 4 Ozone distribution observed on flight across US during the SASS (Subsonic Assessment) Ozone and Nitrogen Experiment (SONEX) in 1997. A stratospheric intrusion is clearly evident on left side of figure, and low ozone air from tropics transported to mid-latitudes can be seen in the upper troposphere on the right.

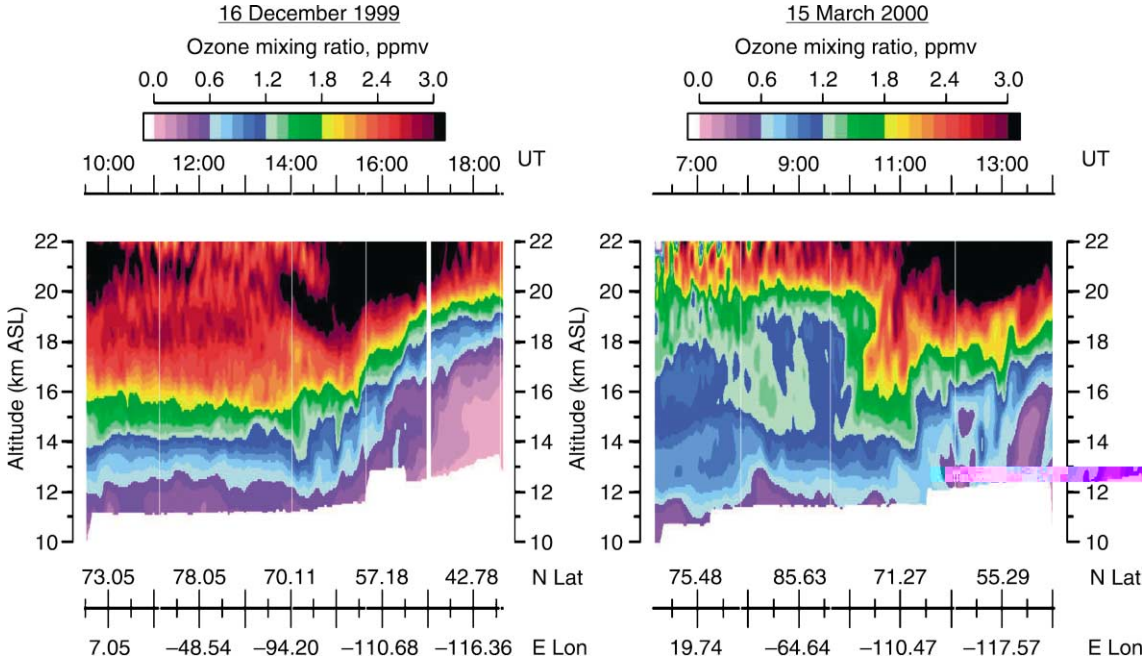


Figure 5 Ozone cross-sections in the stratosphere measured in the winter of 1999/2000 during the SOLVE mission. The change in ozone number density in the Arctic polar vortex due to chemical loss during the winter is clearly evident at latitudes north of 72°N.

predictions. Thus, H_2O distributions can be used in several different ways to better understand chemical, transport, radiation and meteorological processes in the global troposphere.

H_2O Raman Lidar Systems

The first Raman lidar measurements of H_2O were made in the late 1960s, but not much progress in using the Raman approach for H_2O was made until a

system using an abandoned searchlight mirror was employed by the NASA Goddard Space Flight Center in the mid-1980s, to show that if the signal collector (telescope) was large enough, useful Raman measurements could be made to distances of several kilometers. Until recently, Raman lidar measurements of H_2O were largely limited to night-time, due to the high sunlight background interference. Measurements are now made during daytime, using very narrow telescope fields of view and very narrowband filters in the receiver. A good example of a Raman lidar system used to measure H_2O is that at the Department of Energy's Atmospheric Radiation Measurement – Cloud and Radiation Test Bed (ARM-CART) site in Oklahoma. The site is equipped with a variety of instruments aimed at studying the radiation properties of the atmosphere. The system parameters are given in Table 4a,b.

Raman lidar systems have been used for important measurements of H_2O from a number of ground-based locations. Some of the important early work dealt with the passage of cold and warm fronts. The arrival of the wedge-shaped cold front, pushing the warm air up, was one of these studies. Raman lidar have also been located at the ARM-CART site in Kansas and Oklahoma, where they could both provide a climatology of H_2O as well as provide correlative measurements of H_2O for validation of space-based instruments.

Table 4a Parameters for the US Department of Energy's surface-based Raman lidar system for measurements of H_2O

L	Nd:YAG Laser
Wavelength (nm)	355
Pulse energy (mJ)	400
Pulse repetition frequency (Hz)	30
R	
Area (m^2)	1.1
Wavelengths (nm)	
Water vapor	407
Nitrogen	387
Measurement range (km)	
Night-time	near surface to 12
Daytime	near surface to 3
Range resolution (m)	39 at low altitudes, 300 >9 km
Measurement accuracy (%)	
Night-time	<10 <7 km (1 min avg) 10–30 at high altitudes (10–30 min avg)
Daytime	10 <1 km; 5–15 for 1–3 km (10 min avg)

Table 4b Parameters for the NASA Goddard Space Flight Center's Scanning Raman lidar system for measurements of H_2O

	D /*N*	*N*
L	Nd:YAG Laser	XeF
Wavelength (nm)	355	351
Pulse energy (mJ)	300	30–60
Pulse repetition frequency (Hz)	30	400
R		
Area (m^2)	1.8	1.8
Wavelengths (nm)		
Water vapor	407	403
Nitrogen	387	382
Measurement range (km)		
Night-time	near surface to 12	
Daytime	near surface to 4	
Range resolution (m)	7.5 at low altitudes, 300 >9 km	
Measurement accuracy (%)		
Night-time	<10 <5 km (10 sec and 7.5 m range resolution) <10 <7 km (1 min avg) 10–30 at high altitudes (10–30 min avg)	
Daytime	<10 <4 km (5 min avg)	

H_2O DIAL Systems

H_2O was first measured with the DIAL approach using a temperature-tuned ruby laser lidar system in the mid-1960s. The first aircraft-based H_2O DIAL system was developed at NASA LaRC, and was flown in 1982, as an initial step towards the development of a space-based H_2O DIAL system. This system was based on Nd:YAG-pumped dye laser technology, and it was used in the first airborne H_2O DIAL atmospheric investigation, which was a study of the marine boundary layer over the Gulf Stream. This laser was later replaced with a flashlamp-pumped solid-state alexandrite laser, which had high spectral purity, i.e., little out-of-band radiation, a requirement since water vapor lines are narrow, and this system was used to make accurate H_2O profile measurements across the lower troposphere.

A third H_2O DIAL system, called LASE (Lidar Atmospheric Sensing Experiment) was developed as a prototype for a space-based H_2O DIAL system, and it was completed in 1995. This was the first fully autonomously operating DIAL system. LASE uses a Ti:sapphire laser that is pumped by a double-pulsed, frequency-doubled Nd:YAG to produce laser pulses in the 815 nm absorption band of H_2O (see Table 5). The wavelength of the Ti:sapphire laser is controlled by injection seeding with a diode laser that is frequency locked to a H_2O line using an

Table 5 Parameters of LASE H_2O DIAL system

L	Ti:sapphire
Wavelength (nm)	813–818
Pulse energy (mJ)	100
Pulse-pair repetition frequency	5 (on- and off-line pulses separated by 300 μs)
Linewidth (pm)	<0.25
Stability (pm)	<0.35
Spectral purity (%)	>99
Beam divergence (mrad)	<0.6
Pulse width (ns)	35
R	
Area (m^2)	0.11
Receiver optical efficiency (%)	50 (night), 35 (day)
Avalanch Photodiode (APD) detector quantum efficiency (%)	80
Field-of-view (mrad)	1.0
Noise equivalence power ($W\ Hz^{-0.5}$)	2×10^{-14}
Excess noise factor	3
Measurement range (altitude) (km)	15
Range resolution (m)	300–500
Measurement accuracy (%)	5

Table 6 Examples of significant contributions of airborne H_2O DIAL systems to the understanding of H_2O distributions (see also **Figures 6–8**)

NA A L H₂O DIAL LA E
Study of marine boundary layer over Gulf Stream
Observation of H_2O transport at a land/sea edge
Study of large-scale H_2O distributions across troposphere (see **Figure 6**)
Correlative and remote measurements
Observations of boundary layer development (see **Figure 7**)
Cirrus cloud measurements
Hurricane studies (see **Figure 8**)
Relative humidity effects on aerosol sizes
Ice supersaturation in the upper troposphere
Studies of stratospheric intrusions
H_2O distributions over remote Pacific Ocean
O H₂O DIAL
Boundary layer humidity fluxes
Lower-stratospheric H_2O studies

absorption cell. Each pulse pair consists of an on-line and off-line wavelength for the H_2O DIAL measurements. To cover the large dynamic range of H_2O concentrations in the troposphere (over 3 orders of magnitude), up to three line pair combinations are needed. LASE uses a novel approach of operating from more than one position on a strongly absorbing H_2O line. In this approach, the laser is electronically tuned at the line center, side of the line, and near the wing of the line to achieve the required absorption cross-section pairs (on and off). LASE has demonstrated measurements of H_2O concentrations across the entire troposphere using this 'side-line' approach. The accuracy of LASE H_2O profile measurements was determined to be better than 6% or 0.01 g/kg, whichever is larger, over the full dynamic range of H_2O concentrations in the troposphere. LASE has participated in over eight major field experiments since 1995. See **Table 6** for a listing of topics studied using airborne H_2O DIAL systems (**Figures 6–8**).

Space-Based H_2O DIAL System

The technology for a space-based H_2O DIAL system is rapidly maturing in the areas of: high-efficiency, high-energy, high-spectral-purity, long-life lasers with tunability in the 815- and 940-nm regions; low-weight, large-area, high-throughput, high-background-rejection receivers; and high-quantum-efficiency, low-noise, photon-counting detectors. With the expected advancements in lidar technologies leading to a 1-J/pulse capability, a space-based H_2O DIAL system could be flown on a long-duration space mission this decade.

Space-based DIAL measurements can provide a global H_2O profiling capability, which when combined with passive remote sensing with limited vertical resolution, can lead to 3-dimensional measurements of global H_2O distributions. High vertical resolution H_2O (≤1 km), aerosol (≤100 m), and cloud top (≤50 m) measurements from the lidar along the satellite ground-track, can be combined with the horizontally contiguous data from nadir passive sounders to generate more complete high-resolution H_2O, aerosol, and cloud fields for use in the various studies indicated above. In addition, the combination of active and passive measurements can provide significant synergistic benefits leading to improved temperature and relative humidity measurements. There is also strong synergism with aerosol and cloud imaging instruments and with future passive instruments that are being planned or proposed for missions addressing atmospheric chemistry, radiation, hydrology, natural hazards, and meteorology.

Tunable Laser Systems for Point Monitoring

Tunable diode laser (TDL) systems are also used to measure atmospheric gases on a global scale from aircraft and balloons. TDLs are small lasers that emit extremely narrowband radiation and can be tuned in the near IR spectral region using a combination of temperature and current. TDLs can be built into very sensitive and compact systems and located on the surface or flown on aircraft or balloons and used to measure such species as CO, HCl, CH_4, and oxides of nitrogen such as N_2O and NO_2. They derive their

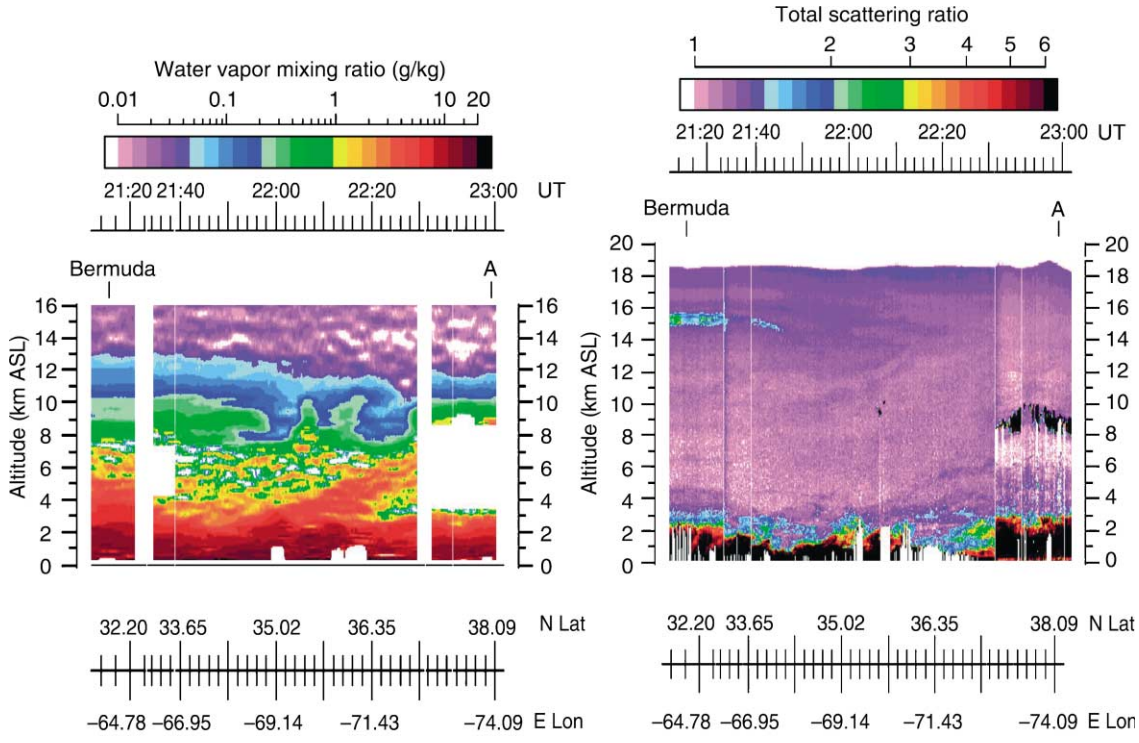


Figure 6 LASE measurements of water vapor (left) and aerosols and clouds (right) across the troposphere on an ER-2 flight from Bermuda to Wallops during the Tropospheric Aerosol Radiative Forcing Experiment (TARFOX) conducted in 1996.

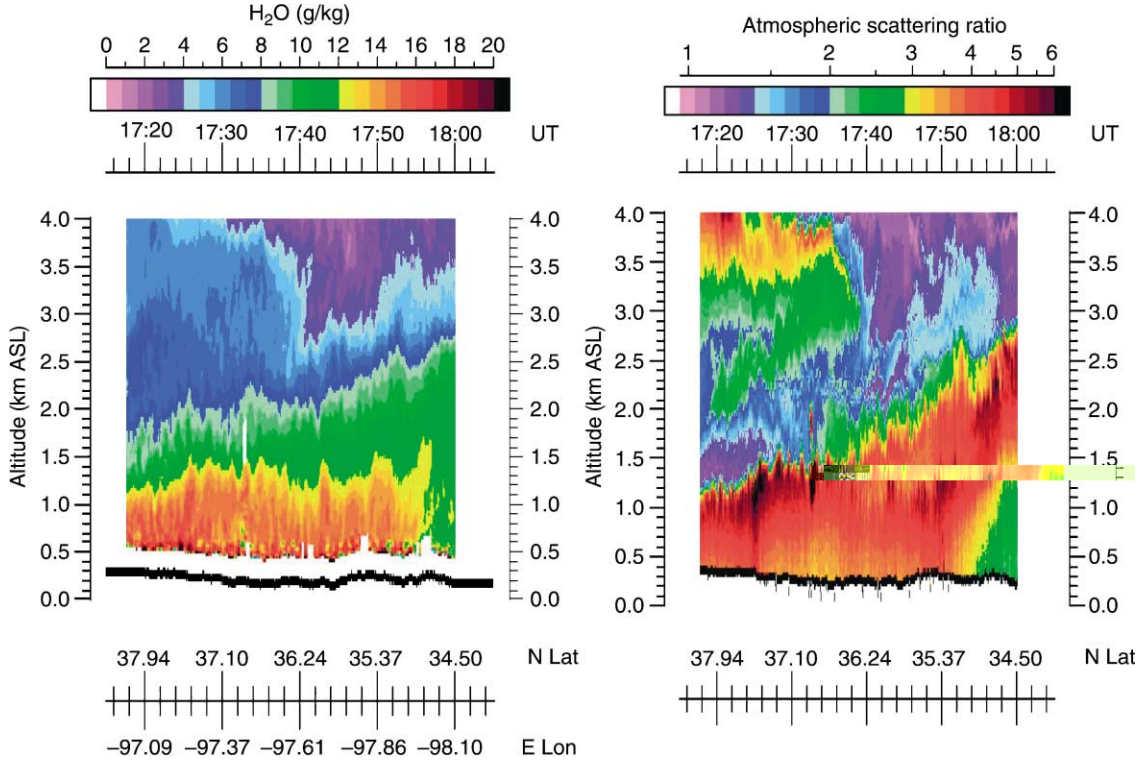


Figure 7 Water vapor and aerosol cross-section obtained on a flight across a cold front during the Southern Great Plains (SGP) field experiment conducted over Oklahoma in 1997.

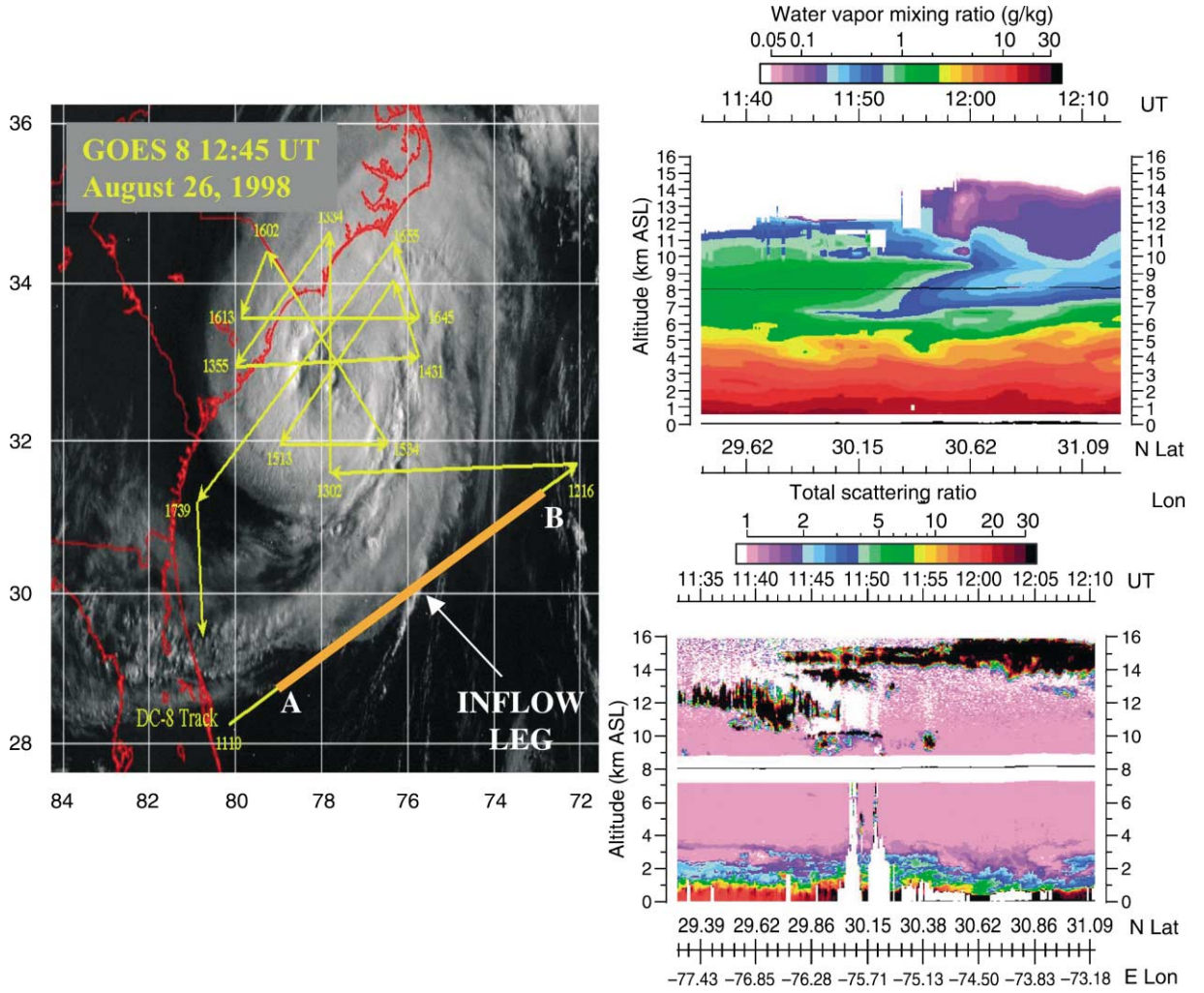


Figure 8 Measurements of water vapor, aerosols, and clouds in the inflow region of Hurricane Bonnie during 1998 Convection and Moisture Experiment (CAMEX-3). A rain band can be clearly seen at the middle of Leg-AB on the satellite and LASE cross-sections.

high sensitivity to the fact that the laser frequency is modulated at a high frequency, permitting a small spectral region to be scanned rapidly. The second- or fourth-harmonic of the scan frequency is used in the data acquisition, effectively eliminating much of the low-frequency noise due to mechanical vibrations and laser power fluctuations. In addition, multipass cells are employed, thereby generating long paths for the absorption measurements.

TDL systems have been used in a number of surface-based measurement programs. One system was mounted on a ship doing a latitudinal survey in the Atlantic Ocean and monitored NO_2, formaldehyde (HCHO), and H_2O_2. Another TDL system was located at the Mauna Loa Observatory and was used to monitor HCHO and H_2O_2 during a photochemistry experiment, finding much lower concentrations of both gases than models had predicted. The TDL system used to measure HCHO has been used in a number of additional ground-based measurement programs and has also been flown on an aircraft in a couple of tropospheric missions.

One TDL system, called DACOM (Differential Absorption CO Measurement), has been used in a large number of NASA GTE missions. It makes measurements of CO in the 4.7 μm spectral region and CH_4 in the 3.3 or 7.6 μm spectral region. This instrument is able to make measurements at a 1 Hz rate and with a precision of 0.5–2.0%, depending on the CO value, and an accuracy of ±2%. DACOM has been very useful in determining global distributions of CO due to the wide-ranging nature of the GTE missions. It has also contributed to

the characterization and understanding of CO in air masses encountered during the GTE missions (Table 7).

A pair of TDL instruments, the Airborne Tunable Laser Absorption Spectrometer (ATLAS) and the Aircraft (ER-2) Laser Infrared Absorption Spectrometer (ALIAS), have been flown on several NASA missions to explore polar O_3 chemistry and atmospheric transport. The ATLAS instrument measures N_2O, and ALIAS is a 4-channel spectrometer that measures a large variety of gases, including HCl, N_2O, CH_4, NO_2, and CO and is currently configured to measure water isotopes across the tropopause.

A new class of lasers, tunable quantum-cascade (QC) lasers, are being added to the list of those available for *in situ* gas measurement systems. Using a cryogenically cooled QC laser during a series of 20 aircraft flights beginning in September 1999 and extending through March 2000, measurements were made of CH_4 and N_2O up to ~20 km in the stratosphere over North America, Scandinavia, and Russia, on the NASA ER-2. Compared with its companion lead salt diode lasers, that were also flown on these flights, the single-mode QC laser, cooled to 82 K, produced higher output power (10 mW), narrower laser linewidth (17 MHz), increased measurement precision (a factor of 3), and better spectral stability (~0.1 cm^{-1} K). The sensitivity of the QC laser channel was estimated to correspond to a minimum-detectable mixing ratio of approximately 2 ppbv of CH_4.

Laser Long-Path Measurements

Laser systems can also be used in long-path measurements of gases. The most important of such programs entailed the measurement of hydroxyl radical (OH) in the Rocky Mountains west of

Table 7 Examples of significant contributions of airborne tunable diode laser systems to the understanding of atmospheric chemistry and transport

B	
	The first measurements of the suite NO_2, NO, O_3, and the NO_2 photolysis rate to test NO_x (NO_2 + NO) photochemistry
	The first stratospheric measurements of NO_x over a full diurnal cycle to test N_2O_5 chemistry
	measurements of NO_2 and HNO_3 over the 20–35 km region to assess the effect of Mt. Pinatubo aerosol on heterogeneous atmospheric chemistry
	Measurements of HNO_3 and HCl near 30 km
	Measurements of CH_4, HNO_3, and N_2O for validation of several satellite instruments
	Intrusions from midlatitude stratosphere to tropical stratospheric reservoir
A	
C	
	Measurement of CO from biomass burning from Asia, Africa, Canada, Central America, and South America
	Detection of thin layers of CO that were transported thousands of miles in the upper troposphere
	Observed very high levels of urban pollution in plumes off the Asian continent
	Emission indices for many gases have been calculated with respect to CO
M	
	Determined CH_4 flux over the Arctic tundra, which led to rethinking of the significance of tundra regions as a global source of CH_4
	Found that biomass burning is a significant source of global CH_4
	Found strong enhancements of CH_4 associated with urban plumes
A	
R	
	Extreme denitrification observed in Antarctic winter vortex from NO_y:N_2O correlation study
	Observed very low N_2O in Antarctic winter vortex, which helped refute the theory that the O_3 hole is caused by dynamics
	Contributed to the study of transport out of the lower stratospheric Arctic vortex by Rossby wave breaking
	Measurement of concentrations of gases involved in polar stratospheric O_3 destruction and production
	Activation of chlorine in the presence of sulfate aerosols
M	
	Measurements in aircraft exhaust plumes in the lower stratosphere, especially of reactive nitrogen species and CO
	Vertical profiles of CO in the troposphere and lower stratosphere
	Determination of the hydrochloric acid and the chlorine budget of the lower stratosphere
	Measurement of NO_2 for testing atmospheric photochemical models
	Trends in HCl/Cl_y in the stratosphere $<$21 km, 1992–1998
	Gas concentration measurements for comparison with a balloon-borne Fourier transform spectrometer observing the Sun
	Near-IR TDL laser hygrometers (ER-2, WB57, DC-8) for measuring H_2O and total water in the lower stratosphere and upper troposphere
	Remnants of Arctic winter vortex detected many months after breakup
	Tropical entrainment time scales inferred from stratospheric N_2O and CH_4 observations

Boulder, Colorado. OH was measured using a XeCl excimer laser operating near 308 nm and transmitting a beam to a retroreflector 10.3 km away. The measurements were quite difficult to conduct, primarily since OH abundance is very low (10^5–10^7 cm^{-3}), yielding very low absorption ($\sim$0.02% for an abundance of 10^5 cm^{-3}) over the 20.6 km path. In addition, the excimer laser could generate OH from ambient H_2O and O_3, unless the laser energy density was kept low. To help ensure good measurements, a white light source was also employed during the measurements to monitor H_2O, O_3, and other gases. The measurements were eventually very successful and led to new values for OH abundances in the atmosphere.

Laser-Induced Fluorescence (LIF)

The laser-induced fluorescence approach has been used to measure several molecular and ionic species *in situ*. The LIF approach has been used to measure OH and HO_2 (HO_x) on the ground and on aircraft platforms. One airborne LIF system uses a diode-pumped Nd:YAG-pumped, frequency-doubled dye laser to generate the required energy near 308 nm. The laser beam is sent into a White cell where it can make 32–36 passes through the gas in the cell to increase the LIF signal strength. NO is used to convert HO_2 to OH. The detection limit in 1 minute of about 2–3 ppqv (10^{-15}) above 5 km altitude, which translates into a concentration of about 4×10^4 molec/cm^3 at 5 km and 2×10^4 molec/cm^3 at 10 km altitude. One of the interesting findings from such measurements is that HO_x concentrations are up to 5 times larger than model predictions based on NO_x concentrations, suggesting that NO_x emissions from aircraft could have a greater impact on O_3 production than originally thought.

NO is detected using the LIF technique in a two-photon approach: electrons are pumped from the ground state using 226 nm radiation and from that state to an excited state using 1.06 μm radiation. The 226 nm radiation is generated by frequency doubling a dye laser to 287 nm and then mixing that with 1.1 μm radiation derived from H_2 Raman shifting of frequency-mixed radiation from a dye laser and a Nd:YAG laser. From the excited level, 187–201 nm radiation is emitted. In order to measure NO_2, it is first converted to the photofragment NO via pumping at 353 nm from a XeF excimer laser. One of the interesting findings from airborne measurements in the South Pacific is that there appeared to be a large missing source for NO_x in the upper troposphere.

Summary

DIAL and Raman lidar systems have played important roles in studying the distribution of gases such as O_3 and H_2O on local, regional, and global scales, while TDL systems have played corresponding roles for such gases as CO, HCl, and oxides of nitrogen. It is anticipated that these approaches will continue to yield valuable information on these and other gases, as new and more capable systems are developed. Within the next decade, it is expected that a DIAL system will be placed in orbit to make truly global measurements of O_3, H_2O, and/or carbon dioxide.

See also

Imaging: Lidar. **Scattering:** Raman Scattering.

Further Reading

Browell EV (1994) Remote sensing of trace gases from satellites and aircraft. In: Calvert J (ed.) *Chemistry of the Atmosphere: The Impact on Global Change*, pp. 121–134. Cambridge, MA: Blackwell Scientific Publications.

Browell EV, Ismail S and Grant WB (1998) Differential absorption lidar (DIAL) measurements from air and space. *Applied Physics. B* 67: 399–410.

Dabas A, Loth C and Pelon J (eds) (2001) *Advances in Laser Remote Sensing*. Selected Papers presented at the 20th International Laser Radar Conference (ILRC), Vichy, France, 10–14 July 2000, pp. 357–360. Palaiseau Cedex, France: A. Edition d'Ecole Polytechnique.

Goldsmith JEM, Bisson SE, Ferrare RA, *et al.* (1994) Raman lidar profiling of atmospheric water vapor: simultaneous measurements with two collocated systems. *Bulletin American Meteorological Society* 75: 975–982.

Goldsmith JEM, Blair FH, Bisson SE and Turner DD (1998) Turn-key Raman lidar for profiling atmospheric water vapor, clouds, and aerosols. *Applied Optics* 37: 4979–4990.

Grant WB (1995) Lidar for atmospheric and hydrospheric studies. In: Duarte F (ed.) *Tunable Laser Applications*, pp. 213–305. New York: Marcel Dekker Inc.

Grant WB, Browell EV, Menzies RT, Sassen K and She C-Y (eds) (1997) *Laser Applications in Remote Sensing*. SPIE Milestones series, 690 pp, 86 papers.

Grisar R, Booettner H, Tacke M and Restelli G (1992) *Monitoring of Gaseous Pollutants by Tunable Diode Lasers*. Proceedings of the International Symposium held in Freiburg, Germany, 17–18 October 1991, 372 pp. Dordrecht, The Netherlands: Kluwer Academic Publishers.

McDermid IS, Walsh TD, Deslis A and White M (1995) Optical systems design for a stratospheric lidar system. *Applied Optics* 34: 6201–6210.

Mégie G (1988) Laser measurements of atmospheric trace constituents. In: Measures RM (ed.) *Laser Remote Chemical Analysis*, pp. 333–408. New York: J Wiley & Sons.

Podolske JR and Loewenstein M (1993) Airborne tunable diode laser spectrometer for trace gas measurements in the lower stratosphere. *Applied Optics* 32: 5324–5330.

Pougatchev NS, Sachse GW, Fuelberg HE, *et al.* (1999) Pacific Exploratory Mission-Tropics carbon monoxide measurements in historical context. *Journal of Geophysics Research* 104: 26,195–26,207.

Sachse GW, Collins Jr JE, Hill GF, *et al.* (1991) Airborne tunable diode laser sensor for high precision concentration and flux measurements of carbon dioxide and methane. *Proceedings of the SPIE International Society for Optical Engineering* 1443: 145–156.

Svanberg S (1994) Differential absorption lidar (DIAL). In: Segrist MW (ed.) *Air Monitoring Spectroscopic Techniques*, pp. 85–161. New York: J Wiley & Sons.

Webster CR, *et al.* (1994) Aircraft (ER-2) laser infrared absorption spectrometer (ALIAS) for in situ measurements of HCl, N_2O, CH_4, NO_2, and HNO_3. *Applied Optics* 33: 454–472.

Whiteman DN, Melfi SH and Ferrare RA (1992) Raman lidar system for the measurement of water vapor and aerosols in the Earth's atmosphere. *Applied Optics* 31: 3068–3082.

Zanzottera E (1990) Differential absorption lidar techniques in the determination of trace pollutants and physical parameters of the atmosphere. *Critical Reviews in Analytical Chemistry* 21: 279–319.

Optical Transmission and Scatter of the Atmosphere

S M Adler-Golden and A Berk,
Spectral Sciences, Inc., Burlington, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The ability to understand and model radiative transfer (RT) processes in the atmosphere is critical for remote sensing, environmental characterization, and many other areas of scientific and practical interest. At the Earth's surface, the bulk of this radiation, which originates from the sun, is found in the ultraviolet to infrared range between around 0.3 and 4 μm. This article describes the most significant RT processes for these wavelengths, which are absorption (light attenuation along the line of sight LOS) and elastic scattering (redirection of the light). Transmittance, T, is defined as one minus the fractional extinction (absorption plus scattering). At longer wavelengths (in the mid- and long-wave infrared) the major light source is thermal emission.

A few other light sources are mentioned here. The moon is the major source of visible light at night. Forest fires can be a significant source of mid-wave infrared radiation. Manmade light sources include continuum sources such as incandescent lamps and spectrally narrow sources such as fluorescent lamps and lasers. In general, spectrally narrow sources need to have a different RT treatment than continuum sources due to the abundance of narrow spectral absorption lines in the atmosphere, as is discussed below.

The challenge of atmospheric RT modeling is essentially to solve the following equation that describes monochromatic light propagation along the LOS direction:

$$\left(\frac{1}{k}\right)\frac{\mathrm{d}I}{\mathrm{d}u} = -I + J \qquad [1]$$

where I is the LOS radiance (watts per unit area per unit wavelength per steradian), k is the extinction coefficient for the absorbing and scattering species (per unit concentration per unit length), u is the material column density (in units of concentration times length), and J is the radiance source function. I is a sum of direct (i.e., from the sun) and diffusely scattered components. The source function represents the diffuse light scattered into the LOS, and is the angular integral over all directions Ω_i of the product of the incoming radiance, $I(\Omega_i)$, and the scattering phase function, $p(\Omega_o, \Omega_i)$:

$$J(\Omega_o) = \int_{\Omega_i} p(\Omega_o, \Omega_i) I(\Omega_i) \mathrm{d}\Omega_i \qquad [2]$$

The scattering phase function describes the probability density for incoming light from direction Ω_i scattering out at angle Ω_o, and is a function of the difference (scattering) angle θ.

The direct radiance component, I', is described by eqn [1] with the source function omitted. Integrating along the LOS leads to the well-known Beer's Law

equation for transmittance:

$$T = I'/I'_0 = \exp(-ku) \qquad [3]$$

where I'_0 is the direct radiance at the boundary of the LOS. The quantity $ku = \ln(1/T)$ is known as the optical depth.

Atmospheric Constituents

The atmosphere has a large number of constituents, including numerous gaseous species and suspended liquid and solid particulates. Their contributions to extinction are depicted in Figure 1. The largest category is gases, of which the most important are water vapor, carbon dioxide, and ozone. Of these, water vapor is the most variable and carbon dioxide the least, although the CO_2 concentration is gradually increasing. In atmospheric RT models, carbon dioxide is frequently taken to have a fixed and altitude-independent concentration, along with other 'uniformly mixed gases' (UMGs). The concentration profiles of the three major gas species are very different. Water vapor is located mainly in the lowest 2 km of the atmosphere. Ozone has a fairly flat profile from the ground through the stratosphere (~30 km). The UMGs decline exponentially with altitude, with a scale height of around 8 km.

Gases

Gas molecules both scatter and absorb light. Rayleigh scattering by gases scales inversely with the fourth power of the wavelength, and is responsible for the sky's blue color. For typical atmospheric conditions, the optical depth for Rayleigh extinction is approximately $0.009/\lambda^4$ per air mass (λ is in μm and air mass is defined by the vertical column from ground to space). The Rayleigh phase function has a $(1 + \cos^2\theta)$ dependence.

Absorption by gases may consist of a smooth spectral continuum (such as the ultraviolet and visible electronic transitions of ozone) or of discrete spectral lines, which are primarily rotation lines of molecular vibrational bands. In the lower atmosphere (below around 25 km altitude), the spectral shape of these lines is determined by collisional broadening and described by the normalized Lorentz line shape formula:

$$f_{\text{Lorentz}}(\nu) = \frac{\alpha_c/\pi}{\alpha_c^2 + (\nu - \nu_0)^2} \qquad [4]$$

Here ν is the wavenumber (in cm^{-1}) [$\nu = (10\,000\ \mu\text{m}/\text{cm})/\lambda$], ν_0 is the molecular line transition frequency and α_c is the collision-broadened half-width (in cm^{-1}), which is proportional to pressure. The constant of proportionality, known as the pressure-broadening parameter, has a typical value on the order of 0.06 $\text{cm}^{-1}\,\text{atm}^{-1}$ at ambient temperature. The extinction coefficient $k(\nu)$ is the product of $f_{\text{Lorentz}}(\nu)$ and the integrated line strength S, which is commonly in units of $\text{atm}^{-1}\,\text{cm}^{-2}$.

At higher altitudes in the atmosphere the pressure is reduced sufficiently that Doppler broadening

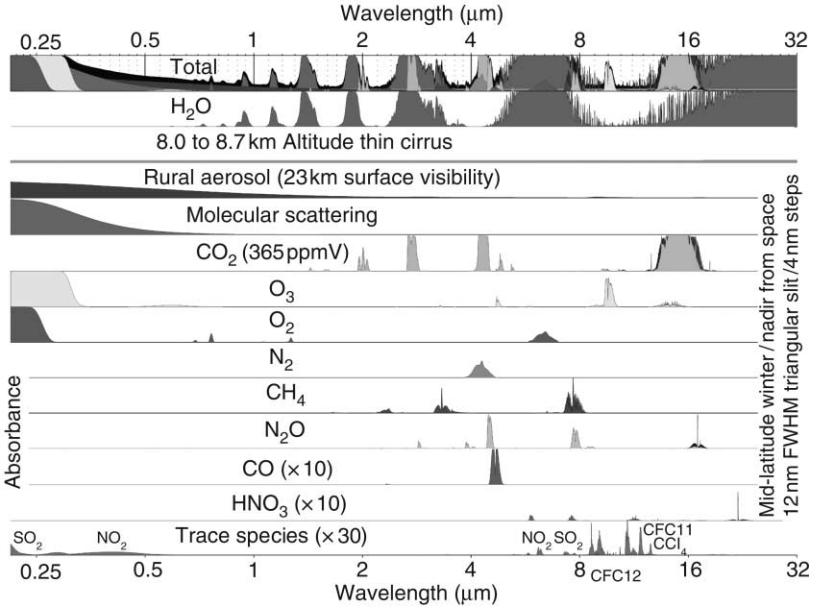


Figure 1 Spectral absorbance (1 – transmittance) for the primary sources of atmospheric extinction. The 12 nm resolution data were generated by MODTRAN for a vertical path from space with a mid-latitude winter model atmosphere.

becomes competitive with collisional broadening, and the Lorentz formula becomes inaccurate. The general lineshape for combined collisional and Doppler broadening is the Voigt lineshape, which is proportional to the real part of the complex error (probability) function, w:

$$f_{\text{Voigt}}(\nu) = \frac{1}{\alpha_{\text{d}}\sqrt{\pi}}\text{Re}\left[w\left(\frac{\nu-\nu_0}{\alpha_{\text{d}}} + \frac{\alpha_{\text{c}}}{\alpha_{\text{d}}}\text{i}\right)\right] \qquad [5]$$

Here, α_{d} is the Doppler 1/e half-width.

Comprehensive spectral databases have been compiled of the transition frequencies, strengths, and pressure-broadened half-widths for atmospherically important molecules throughout the electromagnetic spectrum. Perhaps the most notable of these databases is HITRAN, which was developed by the US Air Force Research Laboratory and is currently maintained at the Harvard-Smithsonian Center for Astrophysics in Cambridge, MA.

Liquids and Solids

The larger particulates in the atmosphere (greater than a few μm in radius) typically belong to clouds. Low-altitude clouds consist of nearly spherical water droplets, while high-altitude cirrus clouds are mainly a collection of ice crystals. Other large particulates include sand dust. Their light scattering is close to the geometric limit at visible and ultraviolet wavelengths. This means that the extinction is nearly wavelength-independent, and the scattering phase function and single-scattering albedo may be reasonably modeled with ray-tracing techniques that account for the detailed size and shape distributions of the particles. However, Mie scattering theory is typically used to calculate cloud optical properties because it is exact for spherical particles of any size.

The smaller particulates in the atmosphere belong to aerosols, which are very fine liquid particles, and dusts, which are solids such as minerals and soot. These particulates are concentrated mainly in the lower 2 km or so of the atmosphere; however, they are also present at higher altitudes in smaller concentrations. Their optical properties are typically modeled using Mie theory. The wavelength dependence of the scattering is approximately inversely proportional to a low power of the wavelength, typically between 1 and 2, as befits particulates intermediate in size between molecular and geometric-limit. The scattering phase functions have a strong forward-scattering peak; values of the asymmetry parameter g (the average value of $\cos\theta$) typically range from 0.6 to 0.8 at solar wavelengths.

Solution Methods

Geometry

Atmospheric properties are primarily a function of altitude, which determines pressure, temperature and species concentration profiles. Accordingly, most RT methods define a stratified atmosphere. The most accurate treatments of transmission and scattering account for the spherical shape of the layers and refraction; however, most RT models use a plane-parallel approximation for at least some computations, such as multiple scattering.

Spectral Resolution

Optical instruments have finite, and frequently broad, wavelength responses. Nevertheless, modeling their signals requires accounting for the variation of absorption on an extremely fine wavelength scale, smaller than the widths of the molecular lines.

'Exact' monochromatic methods

The most accurate RT solution method involves explicitly solving the RT problem for a very large number of monochromatic wavelengths. This line-by-line method is used in a number of RT models, such as FASCODE. It allows Beer's law to be applied to combine transmittances from multiple LOS segments, and provides an unambiguous definition of the optical parameters. It is suitable for use with spectrally structured light sources, such as lasers. The one major drawback of this method is that it is computationally intensive, and therefore may not be practical for problems where large wavelength ranges, multiple LOS views and multiple atmospheric conditions need to be treated.

To alleviate the computational burden of monochromatic calculations, some approximate methods have been developed that model RT in finite spectral intervals, as described below.

Statistical band models

The band model method represents spectral lines in a narrow interval, $\Delta\nu$, statistically using such parameters as the total line strength, the mean pressure-broadening parameter, and the effective number of lines in the interval. An example of a popular band model-based RT algorithm is MODTRAN, which is described below.

A key to the success of band models is the availability of approximate analytical formulas for the integrated absorption for an individual molecular transition of strength S, known as the single-line total

equivalent width, W_{sl}:

$$W_{sl} = \int_{-\infty}^{\infty} [1 - \exp(-Suf_{\nu})]\mathrm{d}\nu \qquad [6]$$

In the optically thin (small absorption) limit, W_{sl} is proportional to the molecular species column density, while in the optically thick (large absorption) limit it scales as the square root of the column density. A further assumption made by MODTRAN is that the line centers are randomly located within the interval, i.e., spectrally uncorrelated. With this assumption, the net transmittance can be expressed as the product of the transmittances for each individual line, whether the line belongs to the same molecular species or a different species.

A particular challenge in band models is treatment of the inhomogeneous path problem – that is, the variations in path properties along a LOS and their effect on the statistical line parameters, which arise primarily from differences in pressure and hence line width. The Curtis–Godson path averaging method provides a reasonable way to define 'equivalent' homogeneous path parameters for the band model. Another challenge is to define an effective extinction optical depth for each layer in order to solve the RT problem with scattering. One option in MODTRAN is to define it by computing the cumulative transmittance through successive layers in a vertical path.

Correlated-*k* model

Another well-known approximate RT algorithm for spectral intervals is the correlated-*k* method. This method starts with an 'exact' line-by-line calculation of extinction coefficients (k's) within the interval on a fine spectral grid, from which a *k*-distribution (vs. cumulative probability) is computed. A database of *k* values and probabilities summing to 1 is built from these *k*-distributions for a grid of atmospheric pressures and temperatures, and for all species contributing to the spectral interval. Inhomogeneous paths are handled by recognizing that the size order of the k's is virtually independent of pressure and typically only weakly dependent on temperature. LOS transmittances, LOS radiances, and fluxes are calculated by interpolating the database over temperature and pressure to define the k's for each LOS segment, solving the monochromatic RT equation at each fixed distribution location, and finally integrating over the distribution. The correlated-*k* method has been found to be quite accurate for atmospheric paths containing a single molecular species; however, corrections must be applied for spectral intervals containing multiple species.

Scattering Methods

When the diffuse light field is of interest, scattering methods are used to calculate the source function *J* of eqns [1] and [2].

Single scattering

If scattering is weak, the approximation may be made that the solar radiation scatters only once. Thus the integral over the scattering phase function, eqn [2], is straightforwardly calculated using the direct radiance component, which is given by eqn [3]. The neglect of multiple scattering (i.e., the diffuse contribution to the source function) means that the diffuse radiance is underestimated; however, single scattering is sufficiently accurate for some atmospheric problems in clear weather and at infrared wavelengths.

Multiple scattering

A number of different methods have been developed to solve the multiple scattering problem. Two-stream methods, which are the simplest and fastest, resolve the radiance into upward and downward directions. These methods generally produce reasonably accurate values of hemispherically averaged radiance, which are also referred to as horizontal fluxes or irradiances.

A much more accurate approach to the multiple scattering problem is the method of discrete ordinates. It involves expansion of the radiation field, the phase function and the surface reflectance as a series of spherical harmonics, leading to a system of linear integral equations. Evaluation of the integrals by Gaussian quadrature leads to a solvable system of linear differential equations. An important approximation called delta-M speeds up convergence of the discrete ordinates method, especially when scattering phase functions are strongly forward peaked, by representing the phase function as the sum of a forward direction δ-function and a remainder term. For most scattering problems, the solution is converged upon with a modest number ($\sim$8 to 16) of quadrature points (streams).

A very different type of multiple scattering technique, called the Monte Carlo method, is based on randomly sampling a large number of computer-simulated 'photons' as they travel through the atmosphere and are absorbed and scattered. The basic idea here is that sensor radiance can be expressed as a multiple path integral over the local source terms, and Monte Carlo methods solve integrals by sampling the integrand. The major advantage of this method is that it is flexible enough to allow for all of the complexity of a realistic atmosphere, often neglected by other methods. The major drawback is its

computational burden, as a large number of photons is required for reasonable convergence; the Gaussian error in the calculation declines with the square root of the number of photons. The convergence problem is moderated to a large extent by using the physics of the problem being solved to bias the selection of photon paths toward those trajectories which contribute most; mathematically, this is equivalent to requiring that the integrand be sampled most often where its contributions are most significant.

An Example Atmospheric RT Model: MODTRAN

MODTRAN, developed collaboratively by the Air Force Research Laboratory and Spectral Sciences, Inc., is the most widely used atmospheric radiation transport model. It defines the atmosphere using stratified layering and computes transmittances, radiances, and fluxes using a moderate spectral resolution band model with IR through UV coverage. The width of the standard spectral interval, or bin, in MODTRAN is 1 cm^{-1}. At this resolution, spectral correlation among extinction sources is well characterized as random. Thus, the total transmittance from absorption and scattering of atmospheric particulates and molecular gases is computed as the product of the individual components.

Rayleigh, aerosol, and cloud extinction are all spectrally slowly varying and well represented by Beer's Law absorption and scattering coefficients on a 1 cm^{-1} grid. Calculation of molecular absorption is more complex because of the inherent spectral structure and the large number of molecular transitions contributing to individual spectral bins. As illustrated in Figure 2, MODTRAN partitions the spectral bin molecular attenuation into 3 components:

- *Line center absorption* from molecular transitions centered within the spectral bin;
- *Line tail absorption* from the tails of molecular lines centered outside of the spectral bin but within 25 cm^{-1}; and
- H_2O and CO_2 *continuum absorption* from distant (>25 cm^{-1}) lines.

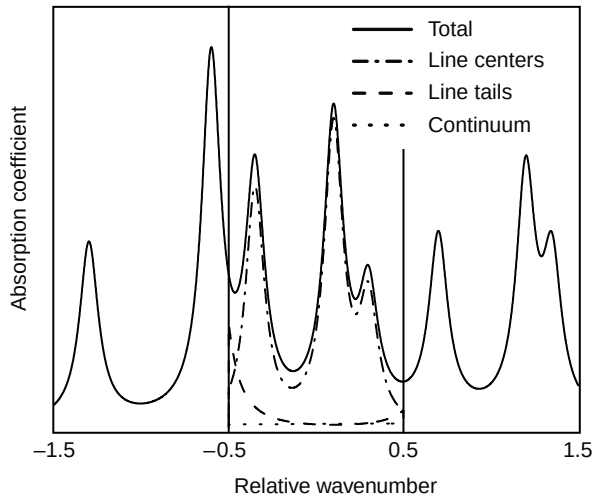


Figure 2 Components of molecular absorption. The line center, line tail and continuum contributions to the total absorption are illustrated for the central 1 cm^{-1} spectral bin.

Within the terrestrial atmosphere, only H_2O and CO_2 have sufficient concentrations and line densities to warrant inclusion of continuum contributions. These absorption features are relatively flat and accurately modeled using 5 cm^{-1} spectral resolution Beer's Law absorption coefficients.

Spectral bin contributions from neighboring line tails drop off, often rapidly, from their spectral bin edge values, but the spectral curves are typically simple, containing at most a single local minimum. For MODTRAN, these spectral contributions are pre-computed for a grid of temperature and pressure values, and fit with Padé approximants, specifically the ratio of quadratic polynomials in wavenumber. These fits are extremely accurate and enable line tail contributions to be computed on an arbitrarily fine grid. MODTRAN generally computes this absorption at a resolution equal to one-quarter the spectral bin width, i.e., 0.25 cm^{-1} for the 1.0 cm^{-1} band model.

The most basic *ansatz* of the MODTRAN band model is the stipulation that molecular line center absorption can be approximated by the absorption of n identical Voigt lines randomly distributed within the band model spectral interval, $\Delta\nu$. Early in the development of RT theory, Plass derived the expression for the transmittance from these n randomly distributed lines:

$$T = \left(1 - \frac{W'_{sl}}{\Delta\nu}\right)^n \qquad [7]$$

MODTRAN's evolution has resulted in a fine tuning of the methodology used to define both the effective line number n and the in-band single-line equivalent width W'_{sl}. The effective line number is initially estimated from a relationship developed by Goody, in which lines are weighted according the to square root of their strength, but MODTRAN combines nearly degenerate transitions into single lines because these multiplets violate the random distribution assumption. The initial effective line number values are refined to insure a match with higher resolution transmittance predictions degraded to the band model resolution. The in-band equivalent

width is computed for an off-centered Voigt line of strength S. Lorentz and Doppler half-widths are determined as strength-weighted averages. The off-center distance is fixed to insure that the weak-line Lorentz equivalent width exactly equals the random line center value.

MODTRAN scattering calculations are optimally performed using the DISORT discrete ordinates algorithm developed by Stamnes and co-workers. Methods for computing multiple scattering such as DISORT require additive optical depths, i.e., Beer's Law transmittances. Since the in-band molecular transmittances of MODTRAN do not satisfy Beer's Law, MODTRAN includes a correlated-k algorithm option. The basic band model *ansatz* is re-invoked to efficiently determine k-distributions; tables of k-data are pre-computed as a function of Lorentz and Doppler half-widths and effective line number, assuming spectral intervals contain n randomly distributed identical molecular lines. Thus, MODTRAN k-distributions are statistical, only dependent on band model parameter values, not on the exact distribution of absorption coefficients in each spectral interval.

Applications

Among the many applications of atmospheric transmission and scattering calculations, we briefly describe two complementary ones in the area of remote sensing, which illustrate many of the RT features discussed earlier as well as current optical technologies and problems of interest.

Earth Surface Viewing

The first example is Earth surface viewing from aircraft or spacecraft with spectral imaging sensors. These include hyperspectral sensors, such as AVIRIS, which have typically a hundred or more contiguous spectral channels, and multispectral sensors, such as Landsat, which typically have between three and a few tens of channels. These instruments are frequently used to characterize the surface terrain, materials and properties for such applications as mineral prospecting, environmental monitoring, precision agriculture, and military uses. In addition, they are sensitive to properties of the atmosphere such as aerosol optical depth and column water vapor. Indeed, in order to characterize the surface spectral reflectance, it is necessary to characterize and remove the extinction and scattering effects of the atmosphere.

Figure 3 also shows an example of data collected by the AVIRIS sensor at ~3 km altitude over thick vegetation. The apparent reflectance spectrum is the observed radiance divided by the Top-of-Atmosphere horizontal solar flux. Atmospheric absorption by water vapor, oxygen, and carbon dioxide is evident, as well as Rayleigh and aerosol scattering. Figure 3 shows the surface reflectance spectrum inferred by modeling and then removing these atmospheric effects (this process is known as atmospheric removal, compensation, or correction). It has the characteristic smooth shape expected of vegetation, with strong chlorophyll absorption in the visible and water bands at longer wavelengths. A detailed analysis of such a spectrum may yield information on the vegetation type and its area coverage and health.

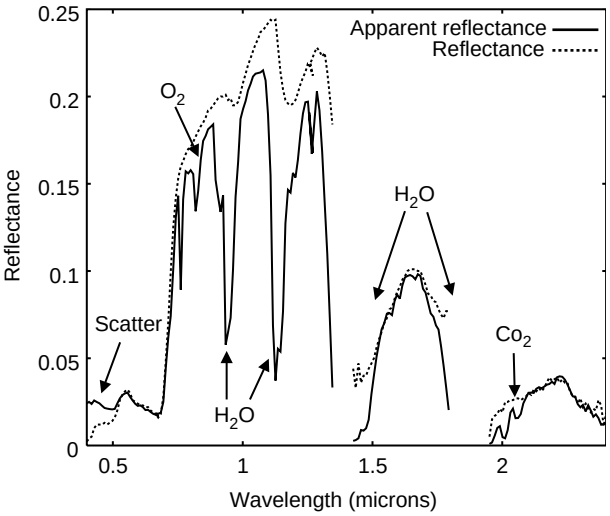


Figure 3 Vegetation spectrum viewed from 3 km altitude before (apparent reflectance) and after (reflectance) removal of atmospheric scatter and absorption.

Sun and Sky Viewing

The second remote sensing example is sun and sky viewing from the Earth's surface with a spectral radiometer, which can yield information on the aerosol content and optical properties as well as estimates of column concentrations of water vapor, ozone, and other gases. Figure 4 shows data from a Yankee Environmental Systems, Inc. multi-filter rotating shadow-band radiometer, which measures both 'direct flux' (the direct solar flux divided by the cosine of the zenith angle) and diffuse (sky) flux in narrow wavelength bands. The plot of ln(direct signal) versus the air mass ratio is called a Langley plot, and is linear for most of the bands, illustrating Beer's Law. The extinction coefficients (slopes) vary with wavelength consistent with a combination of Mie and Rayleigh scattering. The water-absorbing 940 nm band has the lowest values and a curved

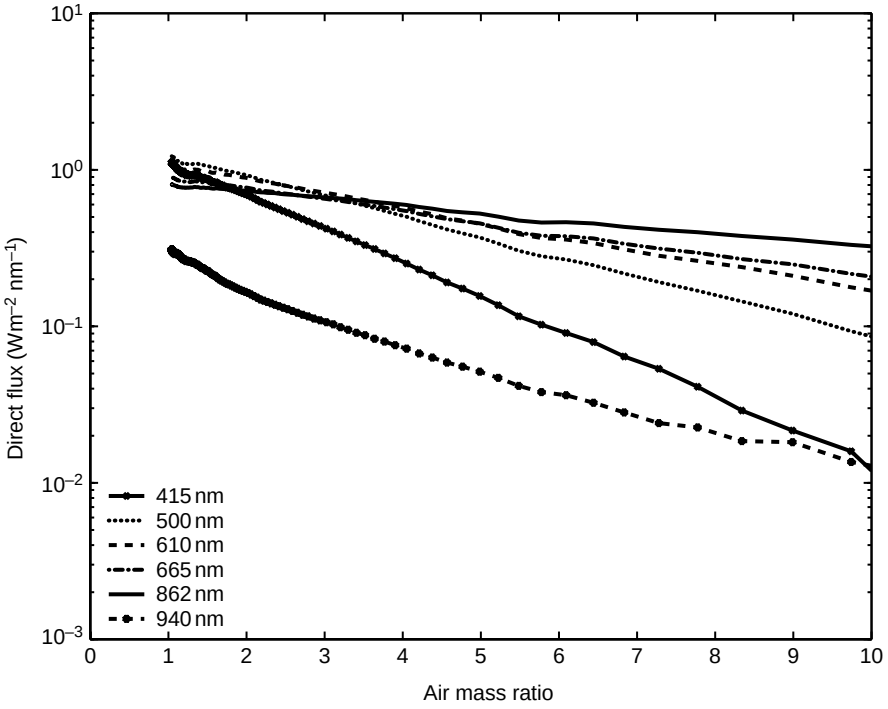


Figure 4 Direct solar flux versus air mass at the surface measured by a Yankee Environmental Systems, Inc. multi-filter rotating shadow-band radiometer at different wavelengths.

plot, in accordance with the square-root dependence of the equivalent width for optically thick lines. The diffuse fluxes, which arise from aerosol and Rayleigh scattering, have a very different dependence on air mass than the direct flux. In particular, the diffuse contributions often increase with air mass for high sun conditions.

The ratio of the diffuse and direct fluxes is related to the total single-scattering albedo, which is defined as the ratio of the total scattering coefficient to the extinction coefficient. Results from a number of measurements showing lower than expected diffuse-to-direct ratios have suggested the presence of black carbon or some other continuum absorber in the atmosphere, which would have a significant impact on radiative energy balance at the Earth's surface and in the atmospheric boundary layer.

List of Units and Nomenclature

Line shape function [cm]	f_ν
Lorentz line shape function [cm]	$f_{\text{Lorentz}}(\nu)$
Voigt line shape function [cm]	$f_{\text{Voigt}}(\nu)$
Scattering asymmetry parameter	g
Direct plus diffuse radiance (W cm^{-1} sr^{-1} or W cm^{-2} μm^{-1} sr^{-1})	I
Direct radiance (W cm^{-1} sr^{-1} or W cm^{-2} μm^{-1} sr^{-1})	I'
Direct radiance at the LOS boundary (W cm^{-1} sr^{-1} or W cm^{-2} μm^{-1} sr^{-1})	I'_0
Source function (W cm^{-1} sr^{-1} or W cm^{-2} μm^{-1} sr^{-1})	J
Extinction coefficient (cm^{-1} atm^{-1})	k
Effective number of lines in bin	n
Scattering phase function (sr^{-1})	p
Line strength (cm^{-2} atm^{-1})	S
Transmittance	T
Column density (atm cm)	u
Complex error (probability) function	w
Single-line total equivalent width (cm^{-1})	W_{sl}
Single-line in-band equivalent width (cm^{-1})	W'_{sl}
Collision-broadened half-width at half maximum (cm^{-1})	α_{c}
Doppler half-width at 1/e (cm^{-1})	α_{d}
Spectral interval (bin) width (cm^{-1})	$\Delta\nu$
Scattering angle (radian)	θ
Wavelength (μm)	λ
Wavenumber (cm^{-1})	ν
Molecular line transition frequency	ν_0
Direction of incoming light (sr)	Ω_{i}
Direction of outgoing light (sr)	Ω_{o}

See also

Environmental Measurements: Laser Detection of Atmospheric Gases. **Instrumentation:** Spectrometers. **Scattering:** Scattering from Surfaces and Thin Films.

Further Reading

Berk A, Anderson GP, Acharya PK, Bernstein LS, Shettle EP, Adler-Golden SM, Lee J and Muratov L (2004) MODTRANS User's Manual. Airforce Research Laboratory, Hanscom AFB, MA, in press.

Chandrasekhar S (1960) *Radiative Transfer*. New York: Dover Publications, Inc., Originally published by Oxford University Press, London (1950).

Fenn RW, Clough SA, Gallery WO, Good RE, Kneizys FX, Mill JD, Rothman LS, Shettle EP and Volz FE (1985) Optical and infrared properties of the atmosphere. In: Jursa AS (ed.) *Handbook of Geophysics and the Space Environment*, chap. 18, US Air Force Geophysics Laboratory, Hanscom AFB, MA, AFGL-TR-85-0315, available from NTIS ADA 167000.

Goody RM and Yung YL (1989) *Atmospheric Radiation, Theoretical Basis*, 2nd edn. New York: Oxford University Press.

HITRAN Special Issue, J. Quant. Spectrosc. Rad. Transfer, vol. 82, Nos. 1–4, 15 November–15 December (2003).

Houghton J (2001) *The Physics of Atmospheres*, 3rd edn, Cambridge, UK: Cambridge University Press.

Killinger DK, Churnside JH and Rothman LS (1994) Atmospheric optics. In *Handbook of Optics*, 2nd edn, 2-Volume Set, chap. 44. New York: McGraw-Hill.

Liou KN (1997) *An Introduction to Atmospheric Radiation*, 2nd edn. London: Academic Press.

Smith FG (ed.) (1993) *The Infrared and Electro-Optical Systems Handbook*, vol. 2, *Atmospheric Propagation of Radiation*. Ann Arbor: Environmental Research Institute of Michigan.

Thomas GE and Stamnes K (1999) *Radiative Transfer in the Atmosphere and Ocean*. Cambridge, UK: Cambridge University Press.

Van de Hulst HC (1981) *Light Scattering by Small Particles*. New York: Dover Publications, Inc., Originally published in 1957, New York: John Wiley & Sons, Inc.

F

FIBER AND GUIDED WAVE OPTICS

Contents

Overview
Dispersion
Fabrication of Optical Fiber
Light Propagation
Measuring Fiber Characteristics
Nonlinear Effects (Basics)
Nonlinear Optics
Optical Fiber Cables
Passive Optical Components

Overview

A Mickelson, University of Colorado, Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

It is said that some in ancient Greece knew that light could be guided inside slabs of transparent material. From this purported little-known curiosity in the ancient world, guided wave optics has grown to be the technology of the physical level of the systems which transfer most of the information about the world; the waveguide of the worldwide telecommunications network is the optical fiber. The optical fiber itself is but a passive waveguide, and guided wave optics is the technology which includes all of the passive and active components which are necessary to prepare optical signals for transmission, regenerate optical signals during transmission, route optical signals through systems and code onto optical carriers, and decode the information from optical carriers, to more conventional forms. In this article, some introduction to this rather encompassing topic will be given.

This article will be separated into four parts. In the first section, discussion will be given to fiber optics, that is, the properties of the light guided in optical waveguides which allow the light to be guided in distinctly nonrectilinear paths over terrestrial distances. The second section will then turn to the components which can be used along with the fiber optical waveguides in order to form useful systems. These components include sources and detectors as well as optical amplifiers. In the third section, we will discuss the telecommunications network which has arisen due to the availability of fiber optics and fiber optic compatible components. The closing section will discuss integrated optics, the field of endeavor which has such great promise to form the future of optical technology.

Fiber Optics

Fiber optics is a term which generally refers to a technology in which light (actually infrared, visible, or ultraviolet radiation) is transmitted through the transparent cores of small threads of composite material. These threads, or fibers as they are called, when surrounded by a cladding material and coated with polymer for environmental protection (see Figure 1) can be coiled like conventional wire and

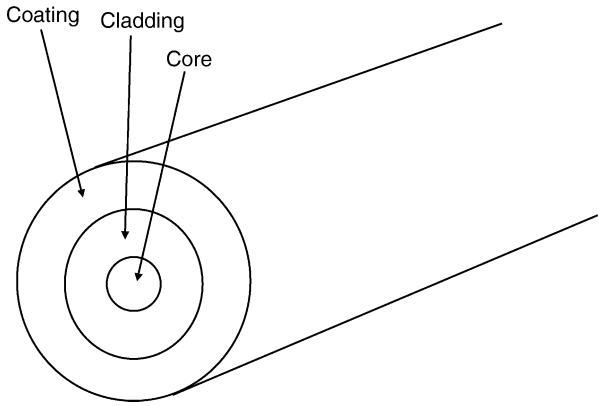


Figure 1 A depiction of the structure of an optical fiber. The innermost cylinder is the core region in which the majority of the light is confined. The next concentric region is the cladding region which is still made of a pure material but one of a lower index of refraction than the innermost core region, such that the light of the lightwave decays exponentially with extension into this region. The outermost region is a coating which protects the fused silica of the core and cladding from environmental contaminants such as water. For telecommunications fibers, the core is comprised of fused silica doped with germanium and typically has an index of refraction of 1.45, whereas the cladding differs from this index by only about 1%. The coating is often a polyimide plastic which has a higher index (perhaps 1.6) but no light guided light should see the cladding-coating interface.

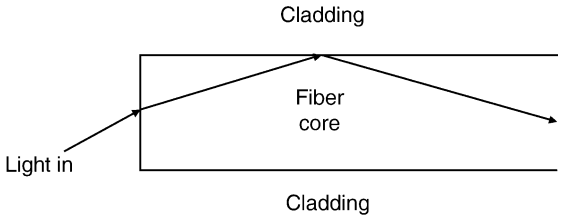


Figure 2 Schematic depiction of the geometrical optics interpretation of the coupling of light into an optical fiber. Here a ray that is incident on the center of the fiber core at an upward angle is first refracted at the fused silica–air interface into the fused silica core. The ray is then totally internally reflected at the core cladding interface such that the ray power remains in the core. The ray picture of light coupling is applicable to single rays coupling to a multimode fiber. When there are multiple rays, interference between these rays must be taken into account which is difficult to do in the ray picture. When there is only one (diffraction limited) mode in the fiber, the interference between a congruence of (spatially coherent) incident rays is necessary to describe the coupling. This description is more easily effected by a quasi-monochromatic mode picture of the propagation.

when cabled can resemble (a very lightweight flexible version of) conventional transmission wire. Although most of the optical fiber in use today is fabricated by a process of gas phase chemical deposition of fused silica doped with various other trace chemicals, fiber can be made from a number of different material systems and in a number of different configurations for use with various types of sources. In what follows in this opening section, we will limit discussion to the basic properties of the light guided by the fiber and leave more technological discussion to following sections.

There are two complementary mathematical descriptions of the propagation of light in an optical waveguide. In the ray description, light incident on a fiber endface is considered to be made up of a bundle of rays. In a uniform homogeneous medium, each ray is like an arrow that exhibits rectilinear propagation from its source to its next interface with a dissimilar material. These rays satisfy Snell's laws of reflection and refraction at interfaces between materials with dissimilar optical properties that they encounter along their propagation path. That is, at an interface, a fraction of the light is reflected backwards at an angle equal to the incident angle and a portion of the light is transmitted in a direction which is more directed toward the normal to the interface when the index increases across the boundary and is directed more away from the normal when the index decreases. At the input endface of a waveguide, a portion of the energy guided by each ray is refracted due to the change in refractive index at the guide surface and then exhibits a more interesting path within the fiber. In a step index optical fiber, where the index of refraction is uniformly higher in the fiber core than in a surrounding cladding, the rays will propagate along straight paths until encountering the core cladding interface. Guided rays (see Figure 2) are totally internally reflected back into the fiber core to again be totally internally reflected at the next core cladding interface and so on. Radiating rays (see Figure 3) will be only partially reflected at the core cladding interface and will rapidly die out in propagating down the fiber when it is taken into account that typical distances between successive encounters with the boundary may be sub-millimeter and total propagation distances may be many kilometers. In graded index fibers, the refractive index within the fiber core varies continuously from a maximum somewhere within the core to a minimum at which the core ends and attaches continuously to a cladding. The ray paths within such fibers are curved and the guided rays are characterized by the fact that once in the fiber they never encounter the cladding. Radiating rays encounter the cladding and are refracted out of the fiber. This description of fiber propagation is quite simple and pleasing but does not take into account that each ray actually is carrying a clock that remembers how long it has been following its given path. When two rays come together, they can

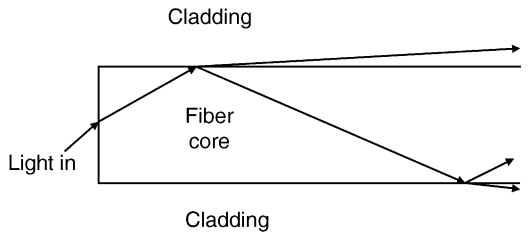


Figure 3 Schematic depiction of the existence of a cut-off angle for coupling into an optical fiber. A ray incident from air on the center of the fused silica fiber core is refracted at the air fused silica interface and then split when it comes to the core cladding interface within the fiber. Were the refracted angle of this ray zero (along the core cladding interface) we would say the ray is totally internally reflected and no power escapes. As it is, there is refracted power into the cladding and the ray will rapidly attenuate as it propagates along the fiber. Although most telecommunications fiber in this day and age is graded index (the rays will curve rather than travel along straight paths), a typical repeat period in such a fiber is a millimeter, whereas propagation distances may be 100 km. With a loss of even only 0.1% at each reflection form the cladding, the light light at cut-off will be severely attenuated in one meter, no loss, one kilometer.

either add or subtract, depending on the reading on their respective clocks. When the light is nearly monochromatic, the clocks' readings are simply a measure of the phase of the ray when it was radiated from its source and the interference between rays can be quite strong. This interference leads to conditions which only allow certain ray paths to be propagated. When the light has a randomly varying phase in both space and time (and is therefore polychromatic), the interference all but disappears. But the condition that allows coupling to a specific ray path (mode) also is obliterated by the phase randomness, and coupling to a single mode guide, for example, becomes inefficient.

When we need to preserve phase relations in order to preserve information while it propagates along a fiber path, we need to use single modes excited by nearly transform limited sources, that is sources whose time variation exhibits well-defined (nonrandom) phase variation. As a monochromatic wave carries no information, we will require a source which can be modulated. When modulation can be impressed on a carrier without loss of phase information of either information or carrier, information can be propagated at longer distances than when the frequency spectrum of the carrier is broadened by noise generated during the modulation. We refer to such a source that can be modulated without broadening as a coherent or nearly coherent source. We cannot easily adapt the ray picture to the description of propagation of such coherent radiation in an optical fiber. Instead we must resort to using the time harmonic form of Maxwell's equations, the equations which describe electromagnetic phenomena, whereas rays can be described by a simplification of the time-dependent form of these equations. In the time harmonic approach, we assume that the source is monochromatic and then solve for a set of modes of the fiber at the assumed frequency of the source. These modes have a well-defined phase progression as a function of carrier frequency and possess a given shape in the plane transverse to the direction of propagation. Information can be included in the propagation by assuming the quasi-monochromatic variation of the source, that is, one assumes that the source retains the phase relations between the various modulated frequency components even while its amplitude and phase is being varied externally. When one assumes that time harmonically (monochromatic) varying fields are propagating along the fiber axis, one obtains solutions of Maxwell's equations in the form of a summation of modes. These are guided modes, ones that propagate down the fiber axis without attenuation. There are also radiation modes that never couple into a propagating mode in the fiber. These modes are analogous to the types of rays we see in the ray description of fiber propagation. There are also another set of modes which are called evanescent modes which show up at discontinuities in the fiber or at junctions between the fiber and other components. In the modal picture of fiber propagation, each of these modes is given an independent complex coefficient (that is, a coefficient with both amplitude and phase). These coefficients are determined first by any sources in the problem and then must be recalculated at each discontinuity plane along the propagation path in the fiber. When the source(s) in the problem is not transform limited, the phases of the coefficients become smeared out and the coupling problems at discontinuities take on radically different solutions.

In some limit, these randomized solutions must appear as the ray solutions. Optical fibers are generally characterized by their numerical aperture (NA), as well as the number which characterizes its transverse dimension. The numerical aperture is essentially the sine of the maximum angle into which the guide will radiate into free space. When the guide is multimoded, then the transverse dimension is generally given as a diameter. When a guide is single-moded, the transverse dimension is generally given as a spotsize, that is, by a measure of the size of a unity magnification image of the fiber endface. Multimode guides can be excited by even poorly coherent sources so long as they radiate into the NA or capture angle at the input of the guide. Single-mode fibers require an excitation which matches the shape and size of their fundamental mode.

Active Fiber Compatible Components

Our discussion of propagation in fiber optic waveguides would be a rather sterile one if there were not a number of fiber compatible components that can generate, amplify, and detect light streams available to us in order to construct fiber optic systems. In this section, we will try to discuss some of the theory of operation of active optical components in order to help elucidate some of the general characteristics of such components and the conditions that inclusion of such components in a fiber system impose on the form that the system must take.

In a passive component such as an optical fiber, we need only consider the characteristics of the light that is guided. Operation of active components requires that the optical fields interact with a medium in such a manner that energy can be transferred from the field through an excitation of the medium to the controlling electrical stream and/or vice versa. Generally one wants the exchange to go only one way, that is from field to electrical stream or from electrical stream to field. In a detector this is not difficult to achieve, whereas in lasers and amplifiers it is quite hard to eliminate the back reaction of the field on the device. Whereas a guiding medium can be considered as a passive homogeneous continuum, the active medium has internal degrees of freedom of its own as well as a granularity associated with the distribution of the microscopic active elements. At least, we must consider the active medium as having an active index of refraction with a mind (set of differential equations anyway) of its own. If one also wants to consider noise characteristics, one needs to consider the lattice of active elements which convert the energy. The wavelength of operation of that active medium is determined by the energy spacing between the upper and lower levels of a transition which is determined by the microscopic structure of these grains, or quanta, that make up the medium. In a semiconductor, we consider these active elements to be so uniformly distributed and ideally spaced that we can consider the quanta (electron hole pairs) to be delocalized but numerous and labeled by their momentum vectors. In atomic media such as the rare earth doped optical fibers which serve as optical amplifiers, we must consider the individual atoms as the players. In the case of the semiconductor, a current, flowing in an external circuit, controls the population of the upper (electronic) and lower (hole) levels of each given momentum state within the semiconductor. When the momentum states are highly populated with electrons and holes, there is a flow of energy to the field (at the transition wavelength) and when the states are depopulated, the field will tend to be absorbed and populate the momentum states by giving up its energy to the medium. The situation is similar with the atomic medium except that the pump is generally optical (rather than electrical) and at a different wavelength than the wavelength of the field to be amplified. That is to say, one needs to use a minimum of three levels of the localized atoms in the medium in order to carry out an amplification scheme.

The composition of a semiconductor determines its bandgap, that is, the minimum energy difference between the electron and hole states of a given momentum value. A source, be it a light emitting diode (LED) or laser diode (see **Figure 4**) will emit light at a wavelength which corresponds to an energy slightly above the minimum gap energy (wavelength equals the velocity of light times Planck's constant divided by energy) whereas a detector can detect almost any energy above the bandedge. Only semiconductors with bandgaps which exhibit a minimum energy at a zero momentum transfer can be made to emit light strongly. Silicon does not exhibit a direct gap and although it can be used as a detector, it cannot be used as a source material. The silicon laser is and has been the 'Holy Grail' of electronics because of the ubiquity of silicon electronics. To even believe that a 'Holy Grail' exists requires a leap of faith. Weak luminescence has been

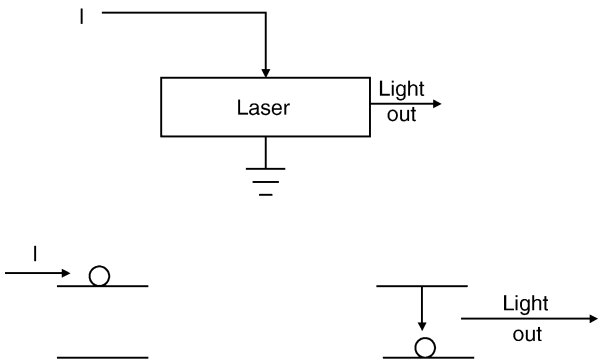


Figure 4 A schematic depiction of the workings of a semiconductor laser light source. The source is fabricated as a diode, and normal operation of this diode is in forward bias, that is, the p-side of the junction is biased positively with respect to ground which is attached to the n-side of the junction. With the p-side positively biased, current should freely flow through the junction. This is not quite true as there is a heterojunction region between the p- and n-regions in which electrons from the n-side may recombine with holes from the p-side. This recombination gives off a photon which is radiated. In light emitting diodes (LEDs), the photon simply leaves the light source as a spontaneously emitted photon. In a laser diode, the heterojunction layer serves as a waveguide and the endfaces of the laser as mirrors to provide feedback and allow laser operation in which the majority of the photons generated are generated in stimulated processes.

observed in certain silicon structures. This luminescence has been used for mid-infrared sources (where thermal effects mask strong luminescence) and in concert with rare earth dopants to make visible light emitting diodes. The silicon laser is still only a vision.

Source materials are, therefore, all made of compound semiconductors. Detectors are in essence the inverse of sources (see Figure 5) but also must detect at wavelengths above their gap energy. For primarily historical reasons, as will be further discussed below, telecommunications employs essentially three windows of wavelengths for different applications, wavelengths centered about 0.85 microns in the first window, wavelengths about 1.3 microns in the second window, and wavelengths about 1.55 microns in the third window. Materials made of layers of different compositions of AlGaAs and mounted on GaAs substrates are used in the first band, while the other bands require the increased degree of freedom allowed by the quaternary alloys of InGaAsP mounted on InP substrates. Other material mixes are possible, these above-mentioned materials are the most common. Rare earth ions exhibit many different transitions. The most useful ones have proven to be the transitions of Er in the third telecommunications window, the one that covers the spectral region to either side of 1.55 microns. Although both Nd and Pr can be made to amplify near the 1.3 micron window, amplifiers made of these materials have not proven to be especially practical.

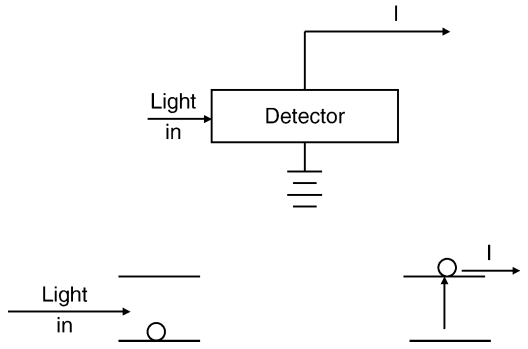


Figure 5 A schematic depiction of the workings of a semiconductor detector. As with a semiconductor source, the detector is fabricated to operate as a diode, that is a p-n junction, but in the case of the detector this diode is to be operated only in reverse bias, that is, with the n-material being biased positively with respect to the ground which is attached to the p-material's contact. In this mode of operation, no current should flow through the junction as there are no free carriers in the material, they have all been drawn out through the contacts. A photon of sufficiently high energy can change this situation by forming an electron hole pair in the material. The electron and hole are thereafter drawn to the n- and p-contacts, respectively, causing a detectable current to flow in the external circuit.

Telecommunications Technology

The physical transmission layer of the worldwide telecommunications network has come to be dominated by fiber optic technology, in particular, by wavelength division multiplexed or WDM single-mode fiber optics operating in the 1.55 micron telecommunications window. In the following, we will first discuss how this transformation to optical communications took place and then go on to discuss some of the specifics of the technology.

Already by the middle of the 1960s, it was clear that changes were going to have to take place in order that the exponential growth of the telephone system in the United States, as well as in Europe, could continue. The telephone system then, pretty much as now, consisted of a hierarchy of tree structures connected by progressively longer lines. A local office is used to connect a number of lines emanating in a tree structure to local users. The local offices are connected by trunk lines which emanate from a toll office. The lines from there on up the hierarchy are long distance ones which are termed long lines and can be regional and longer. The most pressing problem in the late 1960s was congestion in the so-called trunk lines which connect the local switching offices. The congestion was naturally most severe in urban areas. These trunk lines were generally one kilometer in length at that time. The problem was that there was no more space in the ducts that housed these lines. A solution was to use time division multiplexing or TDM to increase the traffic that could be carried by each of the lines already buried in the conduit. The problem was that the twisted pair lines employed would smear out the edges of the time varying bit streams carrying the information at the higher bitrates (aggregated rates due to the multiplexing) due to the inherent dependence of signal propagation velocity on frequency known as dispersion. After discussion and even development of a number of possible technologies, in 1975 a demonstration of a fiber optic system which employed multimode semiconductor lasers feeding multimode optical fibers all operating at 0.85 micron wavelength, proved to be the most viable model for trunk line replacement. The technology was successful and already in 1980 advances in single-mode laser and single-mode fiber technology operating at the 1.3 micron wavelength had made the inclusion of fiber into the long lines viable as well. For roughly the decade from 1985 onward, single-mode fiber systems dominated long line replacement for terrestrial as well as transoceanic links. The erbium doped fiber amplifier which operated in the 1.55 micron wavelength third telecommunication window had proven

itself to be viable for extending repeater periods by around 1990. Development of the InGaAsP quaternary semiconductor system allowed for reliable lasers to be manufactured for this window as development of strained lattice laser technology in the InGaAs system allowed for efficient pump lasers for the fiber amplifiers. As the optical amplifiers could amplify across the wavelength band that could be occupied by many aggregated channels of TDM signals, the move of the long line systems to the third telecommunications window was accompanied by the adoption of wavelength division multiplexing or WDM. That is, electronics becomes more expensive as the analog bandwidth it can handle increases. Cost-effective digital rates are now limited to a few Gb/s, and analog rates to perhaps 20 GHz. Optical center frequencies are in the order of 2×10^{14} Hz and purely optical filters can be made to cleanly separate information signals spaced as closely as 100 GHz apart on optical carriers. An optical carrier can then be made to carry hundreds of channels of 10 Gb/s separated by 100 GHz and this signal can be propagated thousands of kilometers through the terrestrial fiber optics network. Such is today's Worldwide Web.

Integrated Optics as a Future of Guided Wave Optics

As electronics has been pushed to ever greater levels of integration, its power has increased and its price has dropped. A goal implicit in much of the development of planar waveguide technology has been that optics could become an integrated technology in the same sense as electronics. This has not occurred and probably will not occur. This is not to say that integrated optical processing circuits are not being developed and that they will be not employed in the future. They most definitely will.

Integrated optics was coined as a term when it was used in 1969 to name a new program at Bell Laboratories. This program was aimed at investigation of all the technologies that were possible with which to fabricate optical integrated circuits. The original effort was in no way driven by fiber optics as the Bell System only reluctantly moved to a fiber optic network solution in 1975. The original integrated optics efforts were based on such technologies as the inorganic crystal technologies that allowed for large second-order optical nonlinearities, a necessity in order that a crystal also exhibits a large electro-optic coefficient. An electro-optic coefficient allows one to change the index of refraction of a crystal by applying a low frequency or DC electromagnetic field across the crystal.

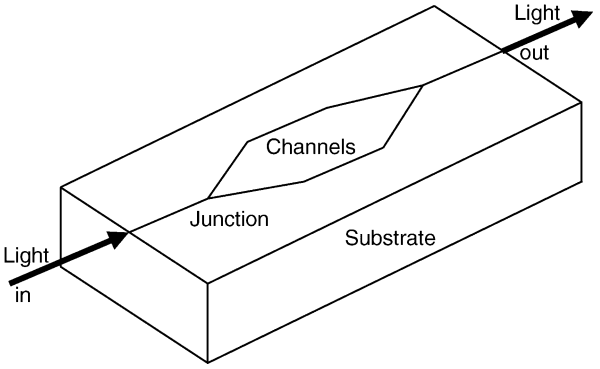


Figure 6 A schematic depiction of an integrated optical device which is often used as a high speed modulator. The device is an integrated version of a Mach–Zehnder interferometer. The basic idea behind its operation is that spatially and temporally coherent light is input into a single mode optical channel. That channel is then split into two channels by a Y-junction. Although not depicted in the figure, the two arms should not have completely equivalent propagation paths. That is to say, that if the light in one of those paths propagates a slightly longer distance (as measured in terms of phase fronts of the wave) then one cannot recombine the power from the two arms. That is, in the second Y-junction, a certain amount of the light which we are trying to combine from the two arms will be radiated out from the junction. In fact, if the light were temporally incoherent (phase fronts not well defined), exactly half of the light would be radiated form the junction. This is a statement of the brightness theorem which was thought to be a law of propagation before there were coherent sources of light. In a high speed modulator, the substrate is an electro-optic crystal and electrodes are placed over the two arms which apply the electrical signal to be impressed on the optical carrier to the channels.

Lithium niobate technology is a technology that has lasted to the present as an optical modulator technology (see **Figure 6**) and an optical switch technology (see **Figure 7**), as well as a technology for parametric wavelength conversion. Unfortunately, although low loss passive waveguide can be fabricated in lithium niobate, the crystal is not amenable to any degree of monolithic integration. Glass was also investigated as an integration technology from the early days of integrated optics, but the lack of second-order nonlinearity in glass strictly limited its applicability. Attention for a period turned to monolithic integration in the semiconductor materials which were well progressing as for use as lasers, detectors, and integrated circuits. Semiconductor crystals, however, are too pure to allow for low loss light propagation which requires the material defects to be so numerous that optical wavelengths cannot sample them. Passive waveguides in semiconductors incur huge losses that can only be mitigated by almost constant optical amplification which, unfortunately, cannot track the rapid optical field variations necessary for information transfer. Early on, attention turned to silicon

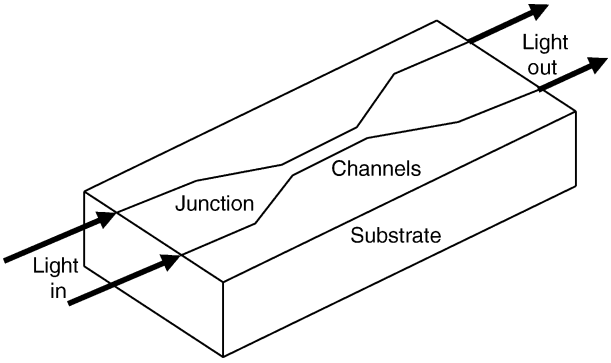


Figure 7 A schematic depiction of an integrated optical device which is often used as an optical switch. This device is interferometric but really has no free space optics counterpart as the Mach–Zhender interferometer does. In this interferometer, spatially and temporally coherent light is input into a single mode input channel. The two single optical mode input channels, which are initially so far apart from each other in terms of their mode sizes that they are effectively uncoupled, are then brought ever closer together until they are strongly coupled and their modes are no longer confined to their separate waveguides but are shared between the two channels. As the channels are symmetric, but the initial excitation, if from a single waveguide alone, is not, the pattern must evolve with propagation. Were both input channels excited, symmetry would still require a symmetry of their phases. If the phases were either equal to each other or completely out of phase, then there would be no evolution because we would have excited coupler modes. If the coupled region is the proper length, light input to one channel will emerge from the other after the channels are pulled apart. If the substrate is an electro-optic crystal and electrodes are placed over the channels, application of a voltage can affect the evolution of the interference in such a manner so as to 'decouple' the channels and switch the light back to the original channel.

optical bench technology, a technology which has presently spawned micro-electro-mechanical-system (MEMs) technology and still persists as an expensive vehicle to hybrid electro-optical integration. In silicon bench, individual components are micro-mounted with solder or silicon pop-up pieces on a silicon wafer which serves as a micro-optical bench. The technology is expensive and hybrid but pervasive in the highest end of cost classes of optical systems.

The tremendous success of the fiber optical network spawned a movement in which groups tried to achieve every possible system functionality in fibers themselves. A notable success has been the optical fiber amplifier. Significantly less successful in the fiber optic network has been the fiber Bragg grating. Attempts to carry out WDM functionalities in all fiber configuration leads to unwieldy configurations requiring large amounts of fiber and significant propagation delays, although, Bragg grating sensors have proven to be somewhat useful in sensing applications. WDM functions have been implemented in a large number of different hybrid configurations, involving various types of glasses, fused silica on silicon and simple hybrids of birefringent crystals.

The telecommunications network is not the driver that it once was. That this is so is perhaps the strongest driving force yet for integrated optics. The driver now is new applications that must be implemented in the most efficient manner. The most efficient manner is quite generally the one in which there is a maximum degree of integration. In the long run, polymer has always been the winning technology for optics of any kind, due to the flexibility of the technology and the drop in cost that accompanies mass production. Indeed, polymer integrated optics progresses. There are also a number of researchers involved in investigating strongly guiding optics, so-called photonic crystal optics. That is, in order to achieve low loss as well as low cost, the fiber and integrated optic waveguides up to the present have used small variations in optical properties of material to achieve guidance at the cost of having structures that are many wavelengths in size. In radio frequency (RF) and microwave technology, for example, guides are tiny fractions of a wavelength. This can lead to an impedance matching problem. RF and microwave impedance matching costs (versus circuit impedance matching costs) are generally quite high, precluding mass application. That radio frequency systems, such as cell phone transceivers link to the rest of the world by antenna, allows that the overall circuit dimension can be kept less than a single wavelength and impedance matching can be foregone in this so-called circuit limit. It is usually hard to keep an overall microwave system to less than a wavelength in overall extent except in the cell phone case or in a microwave oven. Optical miniaturization will require high index contrast guides, and will require optical impedance matching. There are few complete optical systems which will comprise less than an optical wavelength in overall extent. But then this so-called photonic crystal technology will likely be polymer compatible and there may be ways to find significant cost reduction. The future of integrated optics is unclear but bright.

See also

Fiber and Guided Wave Optics: Dispersion; Fabrication of Optical Fiber; Light Propagation; Measuring Fiber Characteristics; Nonlinear Effects (Basics); Nonlinear Optics; Optical Fiber Cables; Passive Optical Components. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Agrawal GP (1995) *Fiber-Optic Communication Systems*. New York: Wiley.
Bette, DeMarchis G and Iannone (1995) *Coherent Communications*. New York: Wiley.
Haken H (1985) *Light*. New York: Elsevier North-Holland.
Hecht E (1998) *Optics*. Reading, MA: Addison-Wesley.
Klein MV and Furtak TE (1986) *Optics*. New York: Wiley.
Marcuse D (1982) *Light Transmission Optics*. New York: Van Nostrand Reinhold.
Marcuse D (1991) *Theory of the Dielectic Optical Waveguide*. Boston, MA: Academic Press.
Mickelson AR, Lee Y-C and Basavanhally NR (1997) *Optoelectronic Packaging*. New York: Wiley.
Saleh BAA and Teich MC (1991) *Fundamentals of Photonics*. New York: Wiley.
Yariv A (1997) *Optical Electronics in Modern Communications*. New York: Van Nostrand Reinhold.

Dispersion

L Thévenaz, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Dispersion designates the property of a medium to propagate the different spectral components of a wave with a different velocity. It may originate from the natural dependence of the refractive index of a dense material to the wavelength of light, and one of the most evident and magnificent manifestations of dispersion is the appearance of a rainbow in the sky. In this case dispersion gives rise to a different refraction angle for the different spectral components of the white light, resulting in an angular dispersion of the sunlight spectrum and explicating the etymology of the term dispersion. Despite this spectacular effect, dispersion is mostly seen as an impairment in many applications, in particular for optical signal transmission. In this case, dispersion causes a rearrangement of the signal spectral components in the time domain, resulting in temporal spreading of the information and eventually in severe distortion.

There is a trend to designate all phenomena resulting in a temporal spreading of information as a type of dispersion, namely the polarization dispersion in optical fibers that should be properly named as polarization mode delay (PMD). In this article chromatic dispersion will be solely addressed, that designates the dependence of the propagation velocity on wavelength and that corresponds to the strict etymology of the term.

Effect of Dispersion on an Optical Signal

An optical signal may always be considered as a sum of monochromatic waves through a normal Fourier expansion. Each of these Fourier components propagates with a different phase velocity if the optical medium is dispersive, since the refractive index n depends on the optical frequency ν. This phenomenon is linear and causal and gives rise to a signal distortion that may be properly described by a transfer function in the frequency domain.

Let $n(\nu)$ be the frequency-dependent refractive index of the propagation medium. When the optical wave propagates in a waveguiding structure the effective wavenumber is properly described by a propagation constant β, that reads:

$$\beta(\nu) = \frac{2\pi\nu}{c_0} n(\nu) \qquad [1]$$

where c_0 is the vacuum light velocity. The propagation constant corresponds to an eigenvalue of the wave equation in the guiding structure and normally takes a different value for each solution or propagation mode. For free-space propagation this constant is simply equal to the wavenumber of the corresponding plane wave.

The complex electrical field $E(z,t)$ of a signal propagating in the z direction may be properly described by the following expression:

$$E(z,t) = A(z,t)\, \mathrm{e}^{\mathrm{i}(2\pi\nu_0 t - \beta_0 z)}$$
$$\text{with} \quad \nu_0\text{: the central optical frequency}$$
$$\text{and} \quad \beta_0 = \beta(\nu_0) \qquad [2]$$

where $A(z,t)$ represents the complex envelope of the signal, supposed to be slowly varying compared to the carrier term oscillating with the frequency ν_0. Consequently, the signal will spread over a narrow band around the central frequency ν_0 and the propagation constant β can be conveniently approximated by a limited expansion to the second order:

$$\beta(\nu) = \beta_0 + \left.\frac{\mathrm{d}\beta}{\mathrm{d}\nu}\right|_{\nu=\nu_0} (\nu - \nu_0) + \left.\frac{\mathrm{d}^2\beta}{\mathrm{d}\nu^2}\right|_{\nu=\nu_0} (\nu - \nu_0)^2 \qquad [3]$$

For a known signal $A(0,t)$ at the input of the propagation medium, the problem consists in

determining the signal envelope $A(z,t)$ after propagation over a distance z. The linearity and the causality of the system make possible a description using a transfer function $H_z(\nu)$ such as:

$$\tilde{A}(z,\nu) = H_z(\nu)\,\tilde{A}(0,\nu) \qquad [4]$$

where $\tilde{A}(z,\nu)$ is the Fourier transform of $A(z,t)$.

To make the transfer function $H_z(\nu)$ explicit, let us assume that the signal corresponds to an arbitrary harmonic function:

$$A(0,t) = A_0\,\mathrm{e}^{\mathrm{i}2\pi ft} \qquad [5]$$

Since this function is arbitrary and the signal may always be expanded as a sum of harmonic functions through a Fourier expansion, there is no loss of generality. The envelope identified as a harmonic function actually corresponds to a monochromatic wave of optical frequency $\nu = \nu_0 + f$ as defined by eqn [2]. Such a monochromatic wave will experience the following phase shift through propagation:

$$\begin{aligned} E(z,t) &= A_0\,\mathrm{e}^{\mathrm{i}[2\pi(\nu_0+f)t-\beta(\nu_0+f)z]} \\ &= A_0\,\mathrm{e}^{\mathrm{i}2\pi ft}\mathrm{e}^{\mathrm{i}[2\pi\nu_0 t-\beta(\nu_0+f)z]} \end{aligned} \qquad [6]$$

On the other hand, an equivalent expression may be found using the linear system described by eqns [2] and [4]:

$$\begin{aligned} E(z,t) &= A(z,t)\,\mathrm{e}^{\mathrm{i}(2\pi\nu_0 t-\beta_0 z)} \\ &= A_z\,\mathrm{e}^{\mathrm{i}2\pi ft}\,\mathrm{e}^{\mathrm{i}(2\pi\nu_0 t-\beta_0 z)} \end{aligned} \qquad [7]$$

Since eqns [6] and [7] must represent the same quantities and using the definition in eqn [4], a simple comparison shows that the transfer function must take the following form:

$$H_z(\nu) = \mathrm{e}^{-\mathrm{i}[\beta(\nu)z-\beta_0 z]} \qquad [8]$$

The transfer function takes a more analytical form using the approximation in eqn [3]:

$$H_z(\nu) = \mathrm{e}^{-\mathrm{i}2\pi(\nu-\nu_0)\tau_\mathrm{D}}\,\mathrm{e}^{-\mathrm{i}\pi D_\nu(\nu-\nu_0)^2 z} \qquad [9]$$

In the transfer function interpretation the first term represents a delay term. It means that the signal is delayed after propagation by the quantity:

$$\tau_\mathrm{D} = \frac{1}{2\pi}\frac{\mathrm{d}\beta}{\mathrm{d}\nu_0}z = \frac{z}{V_\mathrm{g}} \qquad [10]$$

where V_g represents the signal group velocity. This term therefore brings no distortion for the signal and thus states that the signal is replicated at the distance z with a delay τ_D.

The second term is the distortion term which is similar in form to a diffusion process and normally results in time spreading of the signal. In the case of a light pulse it will gradually broaden while propagating along the fiber, like a hot spot on a plate gradually spreading as a result of heat diffusion. The effect of this distortion is proportional to the distance z and to the coefficient D_ν, named group velocity dispersion (GVD):

$$D_\nu = \frac{1}{2\pi}\frac{\mathrm{d}^2\beta}{\mathrm{d}\nu^2} = \frac{\mathrm{d}}{\mathrm{d}\nu}\left(\frac{1}{V_\mathrm{g}}\right) \qquad [11]$$

It is important to point out that the GVD may be either positive (normal) or negative (anomalous) and the distortion term in the transfer function in eqn [9] may be exactly cancelled by propagating in a medium with D_ν of opposite sign. It means that the distortion resulting from chromatic dispersion is reversible and this is widely used in optical links through the insertion of dispersion compensators. These are elements made of specially designed fibers or fiber Bragg gratings showing an enhanced GVD coefficient, with a sign opposite to the GVD in the fiber.

From the transfer function in eqn [9] it is possible to calculate the impulse response of the dispersive medium:

$$h_z(t) = \frac{1}{\sqrt{i|D_\nu|z}}\mathrm{e}^{\mathrm{i}\pi\frac{(t-\tau_\mathrm{D})^2}{D_\nu z}} \qquad [12]$$

so that the distortion of the signal may be calculated by a simple convolution of the impulse response with the signal envelope in the time domain.

The effect of dispersion on the signal can be more easily interpreted by evaluating the dispersive propagation of a Gaussian pulse. In this particular case the calculation of the resulting envelope can be carried out analytically. If the signal envelope takes the following Gaussian distribution at the origin:

$$A(0,t) = A_0\,\mathrm{e}^{-\frac{t^2}{\tau_0^2}} \qquad [13]$$

with τ_0 the $1/e$ half-width of the pulse, the envelope at distance z is obtained by convoluting the initial envelope with the impulse response $h_z(t)$:

$$\begin{aligned} A(z,t) &= h_z(t) \otimes A(0,t) \\ &= A_0\sqrt{\frac{\mathrm{i}z_0}{z+\mathrm{i}z_0}}\mathrm{e}^{\mathrm{i}\pi\frac{(t-\tau_\mathrm{D})^2}{D_\nu(z+\mathrm{i}z_0)}} \end{aligned} \qquad [14]$$

where

$$z_0 = -\frac{\pi\tau_0^2}{D_\nu} \qquad [15]$$

represents the typical dispersion length, that is the distance necessary to make the dispersion effect noticeable.

The actual pulse spreading resulting from dispersion can be evaluated by calculating the intensity of the envelope at distance z:

$$|A(z,t)|^2 = A_0 \frac{\tau_0}{\tau(z)} e^{-\frac{2(t-\tau_D)^2}{\tau^2(z)}} \qquad [16]$$

that is still a Gaussian distribution centered about the propagation delay time τ_D, with $1/e^2$ half-width:

$$\tau(z) = \tau_0\sqrt{1+(z/z_0)^2} \qquad [17]$$

The variation of the pulse width $\tau(z)$ is presented in Figure 1 and clearly shows that the pulse spreading starts to be nonnegligible from the distance $z = z_0$. This gives a physical interpretation for the dispersion length z_0. It must be pointed out that there is a direct formal similarity between the pulse broadening of a Gaussian pulse in a dispersive medium and the spreading of a free-space Gaussian beam as a result of diffraction. Asymptotically for distances $z \gg z_0$, the pulsewidth increases linearly:

$$\tau(z) \simeq |D_\nu| \frac{z}{\pi\tau_0} \qquad [18]$$

It must be pointed out that the width increases proportionally to the dispersion D_ν, but also inversely proportionally to the initial width τ_0. This results from the larger spectral width corresponding to a narrower pulsewidth, giving rise to a stronger dispersive effect.

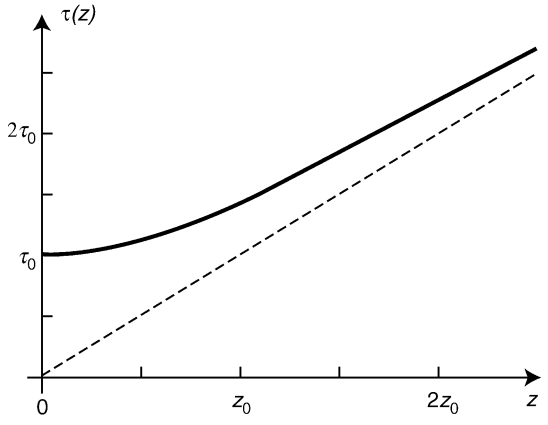


Figure 1 Variation of the $1/e^2$ half-width of a Gaussian pulse, showing the pulse spreading effect of dispersion. The dashed line shows the asymptotic linear spreading for large propagation distance.

The chromatic dispersion does not modify the spectrum of the transmitted light, as any linear effect. This can be straightforwardly demonstrated by evaluating the intensity spectrum of the signal envelope at any distance z, using the eqns [4] and [8]:

$$|\tilde{A}(z,\nu)|^2 = |H_z(\nu)\tilde{A}(0,\nu)|^2 = |e^{-i[\beta(\nu)z-\beta_0 z]}|^2|\tilde{A}(0,\nu)|^2 = |\tilde{A}(0,\nu)|^2 \qquad [19]$$

It means that the pulse characteristics in the time and frequency domains are no longer Fourier-transform limited, since after the broadening due to dispersion, the spectrum should normally spread over a narrower spectral width. This feature results from a re-arrangement of the spectral components within the pulse. This can be highlighted by evaluating the distribution of instantaneous frequency through the pulse. The instantaneous frequency ω_i is defined as the time-derivative of the wave phase factor $\phi(t)$ and is uniformly equal to the optical carrier pulsation $\omega_i = 2\pi\nu_0$ for the initial pulse, as can be deduced from the phase factor at $z = 0$ by combining eqns [2] and [13]. This constant instantaneous frequency means that all spectral components are uniformly present within the pulse at the origin.

After propagation through the dispersive medium the phase factor $\phi(t)$ can be evaluated by combining eqns [2] and [14] and evaluating the argument of the resulting expression. The instantaneous frequency ω_i is obtained after a simple time derivative of $\phi(t)$ and reads:

$$\omega_i(t) = \omega_0 + \frac{2\pi z}{D_\nu(z^2+z_0^2)}(t-\tau_D) \qquad [20]$$

For $z > 0$ the instantaneous frequency varies linearly over the pulse, giving rise to a frequency chirp. The frequency components are re-arranged in the pulse, so that the lower frequency components are in the leading edge of the pulse for a normal dispersion $D_\nu > 0$ and the higher frequencies in the trailing edge. For an anomalous dispersion $D_\nu < 0$, the arrangement is opposite, as can be seen in Figure 2. The effect of this frequency chirp can be visualized in Figure 3, showing that the effect of dispersion is equivalent to a frequency modulation over the pulse. The chirp is maximal at position z_0 and the pulsewidth takes its minimal value τ_0 when the chirp is zero. It is evident with this description that the pulse broadening may be entirely compensated through propagation in a medium of opposite

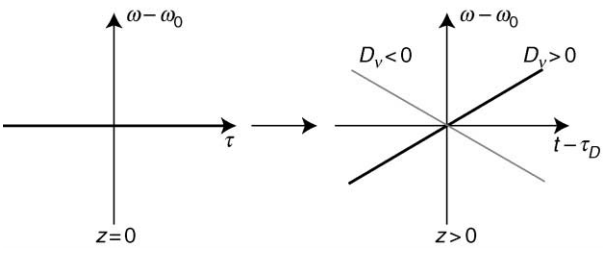


Figure 2 Distribution of the instantaneous frequency through a Gaussian pulse. At the origin the distribution is uniform (left) and the dispersion induces a frequency chirp that depends on the sign of the GVD coefficient D_ν.

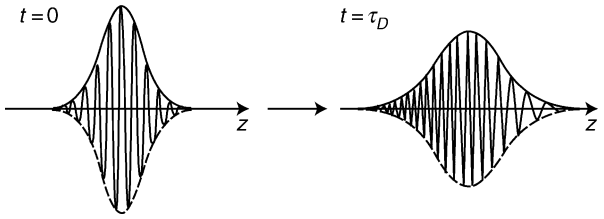


Figure 3 The dispersion results in a pulse broadening together with a frequency chirp, here for a normal GVD, that can be seen like a frequency modulation.

group velocity dispersion; this is equivalent to reversing the time direction in Figure 3. Moreover a pre-chirped pulse can be compressed to its Fourier-transform limited value after propagation in a medium with the proper dispersion sign. This feature is widely used for pulse compression after pre-chirping through propagation in a medium subject to optical Kerr effect.

The above description of the propagation of a Gaussian pulse implicitly states that the light source is perfectly coherent. In the early stages of optical communications it was not at all the case, since most sources were either light emitting diodes or multi-mode lasers. In this case the spectral extent of the signal was much larger than actually required for the strict need of modulation. Each spectral component may thus be considered as independently propagating the signal and the total optical wave can be identified to light merged from discrete sources emitting simultaneously the same signal at a different optical frequency. The group velocity dispersion will cause a delay $\delta\tau$ between different spectral components separated by a frequency interval $\delta\nu$ that can be simply evaluated by a first-order approximation:

$$\delta\tau = \frac{\mathrm{d}\tau_\mathrm{D}}{\mathrm{d}\nu}\delta\nu = \frac{\mathrm{d}}{\mathrm{d}\nu}\left(\frac{z}{V_\mathrm{g}}\right)\delta\nu = D_\nu z\delta\nu \qquad [21]$$

where eqns [10] and [11] have been used. The delay $\delta\tau$ is thus proportional to the GVD coefficient D_ν, to the propagation distance z and the frequency separation $\delta\nu$. This description can be extended to a continuous frequency distribution with a spectral width σ_ν, resulting to the following temporal broadening σ_τ:

$$\sigma_\tau = |D_\nu|\sigma_\nu z \qquad [22]$$

Traditionally the spectral characteristics of a source are given in units of wavelength and the GVD coefficient is expressed in optical fibers accordingly. Following the same description as above, the temporal broadening σ_τ, for a spectral width σ_λ in wavelength units, reads:

$$\sigma_\tau = |D_\lambda|\sigma_\lambda z \qquad [23]$$

Since equal spectral widths must give equal broadening the value of the GVD in units of wavelength can be deduced from the natural definition in frequency units:

$$D_\lambda = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{1}{V_\mathrm{g}}\right) = \frac{\mathrm{d}}{\mathrm{d}\nu}\left(\frac{1}{V_\mathrm{g}}\right)\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} = -\frac{c_0}{\lambda^2}D_\nu \qquad [24]$$

It must be pointed out that the coefficient D_λ takes a sign opposite to D_ν; in other words, a normal dispersion corresponds to a negative GVD coefficient D_λ. It is usually expressed in units of picoseconds of temporal broadening, per nanometer of spectral width and per kilometer of propagating distance, or ps/nm km. For example, a pulse showing a spectral width of 1 nm propagating through a 100 km fiber having a dispersion D_λ of $+10$ ps/nm km, will experience, according to eqn [23], a pulse broadening σ_τ of $10 \times 1 \times 100 = 1000$ ps or 1 ns.

Material Group Velocity Dispersion

Any dense material shows a variation of its index of refraction n as a function of the optical frequency ν. This natural property is called material dispersion and is the dominant contribution in weakly guiding structures such as standard optical fibers. This natural dependence results from the noninstantaneous response of the medium to the presence of the electric field of the optical wave. In other words, the polarization field $P(t)$ corresponding to the material response will vary with some delay or inertia to the change of the incident electric field $E(t)$. This delay between cause and effect generates a memory-type response of the medium that may be described using a time-dependent medium susceptibility $\chi(t)$. The relation between medium polarization at time t and incident field results from the weighted superposition

of the effects of $E(t')$ at all previous times $t' < t$. This takes the form of the following convolution:

$$P(t) = \varepsilon_0 \int_{-\infty}^{+\infty} \chi(t - t')E(t')\mathrm{d}t' \qquad [25]$$

Through application of a simple Fourier transform this relation reads in the frequency domain as:

$$P(\nu) = \varepsilon_0 \chi(\nu)E(\nu) \qquad [26]$$

showing clearly that the noninstantaneous response of the medium results in a frequency-dependent refractive index using the standard relationship with the susceptibility χ:

$$n(\nu) = \sqrt{1 + \chi(\nu)} \quad \text{where} \quad \chi(\nu) = \mathrm{FT}\{\chi(t)\} \qquad [27]$$

This means that a beautiful natural phenomenon such as a rainbow, originates on a microscopic scale from the sluggishness of the medium molecules to react to the presence of light. For signal propagation, it results in a distortion of the signal and in most cases in a pulse spreading, but the microscopic causes are in essence identical. This noninstantaneous response is tightly related to the molecules' vibrations that also give rise to light absorption. For this reason it is convenient to express the propagation in an absorptive medium by adding an imaginary part to the susceptibility $\chi(\nu)$:

$$\chi(\nu) = \chi'(\nu) + \mathrm{i}\chi''(\nu) \qquad [28]$$

so that the refractive index $n(\nu)$ and the absorption coefficient $\alpha(\nu)$ reads in a weakly absorbing medium:

$$n(\nu) = \sqrt{1 + \chi'(\nu)}$$
$$\alpha(\nu) = -\frac{2\pi\chi''(\nu)}{\lambda n(\nu)} \qquad [29]$$

Since the response of the medium, given by the time-dependent susceptibility $\chi(t)$ in eqn [25], is real and causal, the real and imaginary part of the susceptibility in eqn [28] are not entirely independent and are related by the famous Kramers–Kronig relations:

$$\chi'(\nu) = \frac{2}{\pi}\int_0^{\infty} \frac{s\chi''(\nu)}{s^2 - \nu^2}\mathrm{d}s$$
$$\chi''(\nu) = \frac{2}{\pi}\int_0^{\infty} \frac{\nu\chi'(\nu)}{\nu^2 - s^2}\mathrm{d}s \qquad [30]$$

Absorption and dispersion act in an interdependent way on the propagating optical wave and knowing either the absorption or the dispersion spectrum is theoretically sufficient to determine the entire optical response of the medium. The interdependence between absorption and dispersion is typically illustrated in Figure 4. The natural tendency is to observe a growing refractive index for increasing frequencies in low absorption regions. In this case, the dispersion is called normal and this is the most observed situation in transparency regions of a material, which obviously offer the largest interest for optical propagation. In an absorption line the tendency is opposite, a diminishing index for increasing wavelength, and such a response is called anomalous dispersion. The dispersion considered here is the phase velocity dispersion, represented by the slope of $n(\nu)$, that must not be mistaken with the group velocity dispersion (GVD) that only matters for signal distortion. The difference between these two quantities is clarified below.

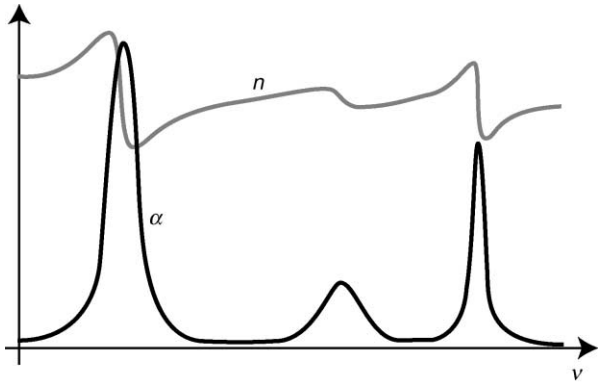


Figure 4 Typical optical response of a transparent medium, showing the spectral interdependence between the absorption coefficient α and the index of refraction n.

To demonstrate that a frequency-dependent refractive index $n(\nu)$ gives rise to a group velocity dispersion and thus a signal distortion we use eqns [1] and [10], so the group velocity V_g can be expressed:

$$V_g = \frac{c_0}{N} \quad \text{with} \quad N = n + \nu\frac{\mathrm{d}n}{\mathrm{d}\nu} = n - \lambda\frac{\mathrm{d}n}{\mathrm{d}\lambda} \qquad [31]$$

N is called group velocity index and differs from the phase refractive index n only if n shows a spectral dependence. In a region of normal dispersion ($\mathrm{d}n/\mathrm{d}\nu > 0$, the group index is larger than the phase index and this is the situation observed in the great majority of transparent materials. From eqn [31] and using eqn [24] the GVD coefficient D_λ can be expressed as a function of the refractive index $n(\lambda)$:

$$D_\lambda = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{1}{V_g}\right) = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{N}{c_0}\right) = -\frac{\lambda}{c_0}\frac{\mathrm{d}^2 n}{\mathrm{d}\lambda^2} \qquad [32]$$

The GVD coefficient is proportional to the second derivative of the refractive index n with respect to

wavelength and is therefore minimal close to a point of inflexion of $n(\lambda)$. As can be seen in Figure 4 such a point of inflexion is always present where absorption is minimal, at the largest spectral distance from two absorption lines. It means that the conditions of low-group velocity dispersion and high transparency are normally fulfilled in the same spectral region of an optical dielectric material. In this region the phase velocity dispersion is normal at any wavelength, but the group velocity is first normal for shorter wavelengths, then is zero at a definite wavelength corresponding to the point of inflexion of $n(\lambda)$, and finally becomes anomalous for longer wavelengths. The situation in which normal phase dispersion and anomalous group dispersion are observed simultaneously is in no way exceptional.

In pure silica the zero GVD wavelength is at 1273 nm, but is subject to be moderately shifted to larger wavelengths in optical fibers, as a result of the presence of doping species to raise the index in the fiber guiding core. This shift normally never exceeds 10 nm using standard dopings; larger shifts are observed resulting from waveguide dispersion and this aspect will be addressed in the next section. The zero GVD wavelength does not strictly correspond to the minimum attenuation in silica fibers, because the dominant source of loss is Rayleigh scattering in this spectral region and not molecular absorption. This scattering results from fluctuations of the medium density as observed in any amorphous materials such as vitreous silica and is therefore a collective effect of many molecules that does not impinge on the microscopic susceptibility $\chi(t)$. It has therefore no influence on the material dispersion characteristics and this explains the reason for the minimal attenuation wavelength at 1550 nm, mismatching and quite distant from the zero material GVD at 1273 nm.

Material GVD in amorphous SiO_2 can be accurately described using a three-term Sellmeier expansion of the refractive index:

$$n(\lambda) = \sqrt{1 + \sum_{j=1}^{3} \frac{C_j \lambda^2}{\lambda^2 - \lambda_j^2}} \qquad [33]$$

and performing twice the wavelength derivative. The coefficients C_j and λ_j are found in most reference handbooks and result in the GVD spectrum shown in Figure 5. Such a spectrum explains the absence of interest for propagation in the visible region through optical fibers, the dispersion being very important in this spectral region. It also explains the large development of optical fibers in the 1300 nm region as a consequence of the minimal material dispersion there. It must be pointed out that it is quite easy to set

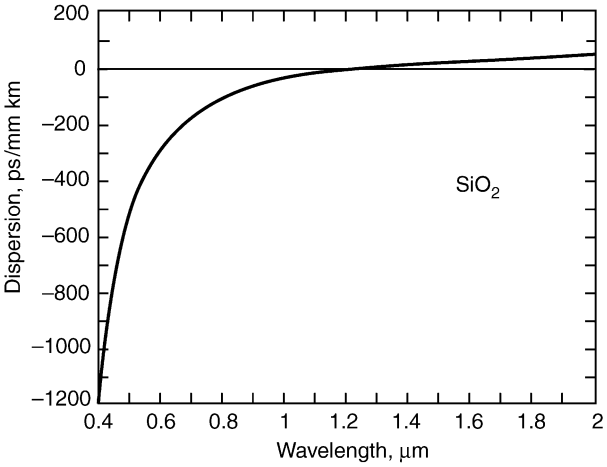


Figure 5 Material dispersion of pure silica. The visible region (0.4–0.7 μm) shows a strong normal GVD that decreases when moving into the infrared and eventually vanishes at 1273 nm. In the minimum attenuation window (1550 nm) the material GVD is anomalous.

up propagation in an anomalous GVD regime in optical fibers, since this regime is observed in the lowest attenuation spectral region. Anomalous dispersion makes possible interesting propagation features when combined with a third-order nonlinearity as the optical Kerr effect, namely soliton propagation and efficient spectral broadening through modulation instability.

Waveguide Group Velocity Dispersion

Solutions of the wave equation in an optical dielectric waveguide such as an optical fiber are discrete and limited. These solutions, called modes, are characterized by an unchanged field distribution along the waveguides and by a uniform propagation constant β over the wavefront. This last feature is particularly important if one recall that the field extends over regions presenting different refractive indices in a dielectric waveguide. For a given mode, the propagation constant β defines an effective refractive index n_{eff} for the propagation by similarity to eqn [1]:

$$\beta = \frac{2\pi\nu}{c_0} n_{\text{eff}} \qquad [34]$$

The value of this effective refractive index n_{eff} is always bound by the value of the core index n_1 and of the cladding index n_2, so that $n_2 < n_{\text{eff}} < n_1$. For a given mode, the propagation constant β, and so the effective refractive index n_{eff}, only depend on a quantity called normalized frequency V that essentially scales the light frequency to the waveguide

optical parameters:

$$V = \frac{2\pi}{\lambda} a \sqrt{n_1^2 - n_2^2} \qquad [35]$$

where a is the core radius. Figure 6 shows the dependence of the propagation constant β of the fundamental mode LP_{01} on the normalized frequency V. The variation is nonnegligible in the single-mode region and gives rise to a chromatic dispersion, since V depends on the wavelength λ, even in the fictitious case of dispersion-free refractive indices in the core and the cladding materials. This type of chromatic dispersion is called waveguide dispersion.

To find an expression for the waveguide dispersion in function of the guiding properties, let us define another normalized parameter, the normalized phase constant b, such as:

$$\beta = \frac{2\pi}{\lambda} \sqrt{n_2^2 + b(n_1^2 - n_2^2)} \qquad [36]$$

The parameter b takes values in the interval $0 < b < 1$, is equal to 0 when $n_{\text{eff}} = n_2$ at the mode cutoff, and is equal to 1 when $n_{\text{eff}} = n_1$. This latter situation is never observed and is only asymptotic for very large normalized frequencies V. Solving the wave equation provides the dispersion relation $b(V)$ and it is important to point out that this relation between normalized quantities depends only on the shape of the refractive index profile. Step-index, triangular or multiple-clad index profiles will result in different $b(V)$ relations, independently of the actual values of the refractive indices n_1 and n_2 and of the core radius a.

From the definitions in eqns [10] and [35] and in the fictitious case of an absence of material dispersion, the propagation delay per unit length reads:

$$\frac{1}{V_g} = \frac{1}{2\pi}\frac{d\beta}{d\nu} = \frac{1}{2\pi}\frac{d\beta}{dV}\frac{dV}{d\nu} = \frac{\lambda}{2\pi} V \frac{d\beta}{dV} \qquad [37]$$

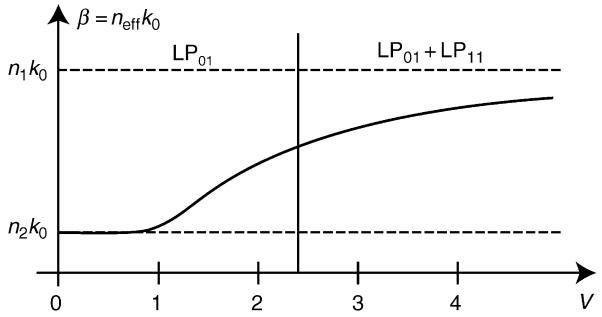


Figure 6 Propagation constant β as a function of the normalized frequency V in a step-index optical fiber. The single mode region is in the range $0 < V < 2.405$.

so that the waveguide group velocity dispersion can be expressed using the relation in eqn [24]:

$$D_\lambda^w = \frac{d}{d\lambda}\left(\frac{1}{V_g}\right) = \frac{d}{d\nu}\left(\frac{1}{V_g}\right)\frac{d\nu}{d\lambda}$$

$$= -\frac{\nu}{\lambda}\frac{d}{d\nu}\left(\frac{1}{V_g}\right) = -\frac{1}{2\pi c_0} V^2 \frac{d^2\beta}{dV^2} \qquad [38]$$

The calculation of the combined effect of material and waveguide dispersions results in very long expressions in which it is difficult to highlight the relative effect of each contribution. Nevertheless, by making the assumption of weak guidance:

$$\frac{n_1 - n_2}{n_2} \simeq \frac{N_1 - N_2}{N_2} \ll 1 \qquad [39]$$

where N_1 and N_2 are the group indices in the core and the cladding, respectively, defined in eqn [31], the complete expression can be drastically simplified to obtain for the delay per unit length:

$$\frac{1}{V_g} = \frac{1}{c_0}\left[N_2 + (N_1 - N_2)\frac{d(bV)}{dV}\right] \qquad [40]$$

and for the total dispersion:

$$D_\lambda = D_2 + (D_1 - D_2)\frac{d(bV)}{dV} - (N_1 - N_2)\frac{N_2}{n_1}\frac{1}{\lambda c_0} V \frac{d^2(bV)}{dV^2} \qquad [41]$$

where

$$D_1 = -\frac{\lambda}{c_0}\frac{d^2 n_1}{d\lambda^2} \quad \text{and} \quad D_2 = -\frac{\lambda}{c_0}\frac{d^2 n_2}{d\lambda^2} \qquad [42]$$

are the material GVD in the core and the cladding, respectively.

The first two terms in eqn [41] represent the contribution of material dispersion weighted by the relative importance of core and cladding materials for the propagating mode. In optical fibers, the difference between D_1 and D_2 is small, so that this contribution can be often well approximated by D_2, independently of any guiding effects.

The last term represents the waveguide dispersion and is scaled by 2 factors:

- *The core-cladding index difference* $n_1 - n_2 \simeq N_1 - N_2$. The waveguide dispersion will be significantly enhanced by increasing the index difference between core and cladding.
- *The shape factor* $V(d^2(bV)/dV^2)$. This factor uniquely depends on the shape of the refractive index profile and may substantially modify the

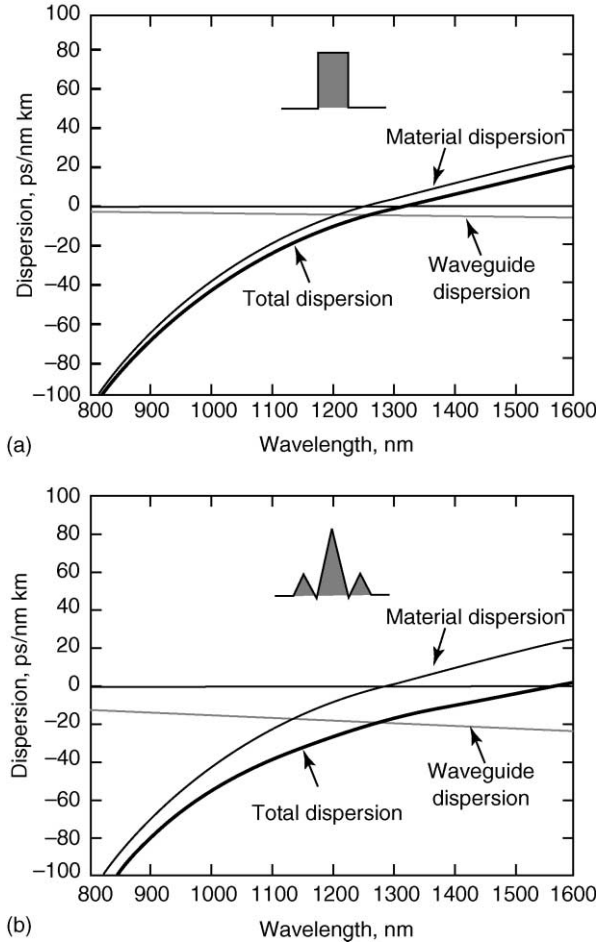


Figure 7 Material, waveguide and total group velocity dispersions for: (a) step-index fiber; (b) triangular profile fiber. The waveguide dispersion can be significantly enhanced by changing the shape of the index profile, making possible a shift of the zero GVD to the minimum attenuation window.

spectral dependence of the waveguide dispersion, making possible a great variety of dispersion characteristics.

Due to this degree of freedom brought by waveguide dispersion it is possible to shift the zero GVD wavelength to the region of minimal attenuation at 1550 nm in silica optical fibers. Figure 7a shows the total group velocity dispersion of a step-index core fiber, together with the separate contributions of material and waveguide GVD. In this case, the material GVD is clearly the dominating contribution, while the small waveguide GVD results in a shift of the zero GVD wavelength from 1273 nm to 1310 nm. A larger shift could be obtained by increasing the core-cladding index difference, but this also gives rise to an increased attenuation from the doping and no real benefit can be expected from such a modification. In Figure 7b, the core shows a triangular index profile to enhance the shape factor in eqn [41], so that the contribution of waveguide GVD is significantly increased with no impairing attenuation due to excessive doping. This makes it possible to realize the ideal situation of an optical fiber showing a zero GVD at the wavelength of minimum attenuation. These dispersion-shifted fibers (DSF) have now become successful in modern telecommunication networks.

Nevertheless, the absence of dispersion favors the efficiency of nonlinear effects and several classes of fibers are now proposed, showing a small but nonzero GVD at 1550 nm, with positive or negative sign, nonzero DSF (NZDSF). By interleaving fibers with positive and negative GVDs it is possible to propagate along the optical link in a dispersive medium and thus minimize the impact of nonlinearities, while maintaining the overall GVD of the link close to zero and canceling any pulse spreading accordingly.

List of Units and Nomenclature

Chromatic dispersion [ps nm^{-1} km^{-1}]	D_λ
Electric field [V m^{-1}]	E
Group index	N
Group velocity [m s^{-1}]	V_g
Group Velocity Dispersion GVD [s^2 m^{-1}]	D_ν
Linear attenuation coefficient [m^{-1}]	α
Medium susceptibility	χ
Normalized frequency	V
Optical frequency [s^{-1}]	ν
Propagation constant [m^{-1}]	β
Propagation delay [s]	τ_D
Polarization density field [A s m^{-2}]	P
Refractive index	n
Vacuum light velocity [m s^{-1}]	c_0
Wavelength [m]	λ

See also

Dispersion Management. **Fiber and Guided Wave Optics:** Overview. **Fourier Optics**. **Nonlinear Optics, Basics:** Kramers–Kronig Relations in Nonlinear Optics.

Further Reading

Ainslie BJ and Day CR (1986) A review of single-mode fibers with modified dispersion characteristics. *IEEE Journal of Lightwave Technology* LT-4(8): 967–979.

Bass M (ed.) (1994) *Handbook of Optics*, sponsored by the Optical Society of America, 2nd edn, vol. II, chap. 10 & 33. New York: McGraw-Hill.

Buck JA (1995) In: Goodman JW (ed.) *Fundamentals of Optical Fibers*, chap. 1, 5, 6. New York: Wiley Series in Pure and Applied Optics.

Gloge D (1971) Dispersion in weakly guiding fibers. *Applied Optics* 10: 2442–2445.
Jeunhomme LB (1989) *Single Mode Fiber Optics: Principles and Applications*. Optical Engineering Series, No. 23. New York: Marcel Dekker.
Marcuse D (1974) *Theory of Dielectric Optical Waveguides*. Series on Quantum Electronics – Principles and Applications, chap. 2. New York: Academic Press.
Marcuse D (1980) Pulse distortion in single-mode fibers. *Applied Optics* 19: 1653–1660.
Marcuse D (1981) Pulse distortion in single-mode fibers. *Applied Optics* 20: 2962–2974.
Marcuse D (1981) Pulse distortion in single-mode fibers. *Applied Optics* 20: 3573–3579.
Mazurin OV, Streltsina MV and Shvaiko-Shvaikovskaya TP (1983) *Silica Glass and Binary Silicate Glasses (Handbook of Glass Data; Part A)* Series on Physical Sciences Data. Amsterdam: Elseviervol. 15, pp. 63–75.
Murata H (1988) *Handbook of Optical Fibers and Cables*. Optical Engineering Series, No. 15, chap. 2. New York: Marcel Dekker.
Saleh BEA and Teich MC (1991) In: Goodman JW (ed.) *Fundamentals of Photonics*, chap. 5, 8 & 22. New York: Wiley Series in Pure and Applied Optics.

Fabrication of Optical Fiber

D Hewak, University of Southampton, Southampton, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The drawing of optical fibers from silica preforms has, over a short period of time, progressed from the laboratory to become a manufacturing process capable of producing millions of kilometers of telecommunications fiber a year. Modern optical fiber fabrication processes produce low-cost fiber of excellent quality, with transmission losses close to their intrinsic loss limit. Today, fiber with transmission losses of 0.2 dB per kilometer of fiber are routinely drawn through a two-stage process that has been refined since the 1970s.

Although fibers of glass have been fabricated and used for hundreds of years, it was not until 1966 that serious interest in the use of optical fibers for communication emerged. At this time, it was estimated that the optical transmission loss in bulk glass could be as low as 20 dB km^{-1} if impurities were sufficiently reduced, a level at which practical applications were possible. At this time, no adequate fabrication techniques were available to synthesize glass of high purity, and fiber-drawing methods were crude.

Over the next five years, efforts worldwide addressed the fabrication of low-loss fiber. In 1970, a fiber with a loss of 20 dB km^{-1} was achieved. The fiber consisted of a titania doped core and pure silica cladding. This result generated much excitement and a number of laboratories worldwide actively began researching optical fiber. New fabrication techniques were introduced, and by 1986, fiber loss had been reduced close to the theoretical limit.

All telecommunications fiber that is fabricated today is made of silica glass, the most suitable material for low-loss fibers. Early fiber research studied multicomponent glasses, which are perhaps more familiar optical materials; however, low-loss fiber, could not be realized, partly due to the lack of a suitable fabrication method. Today other glasses, in particular the fluorides and sulfides, continue to be developed for speciality fiber applications, but silica fiber dominates in most applications.

Silica is a glass of simple chemical structure containing only two elements, silicon and oxygen. It has a softening temperature of about 2,000°C at which it can be stretched, i.e. drawn into fiber. An optical fiber consists of a high purity silica glass core, doped with suitable oxide materials to raise its refractive index (Figure 1). This core, typically on the order of 2–10 microns in diameter, is surrounded by silica glass of lower refractive index. This cladding layer extends the diameter to typically 125 microns. Finally a protective coating covers the entire structure. It is the phenomenon of total internal reflection at the core cladding interface that confines light to the core and allows it to be guided. The basic requirements of an optical fiber are as follows:

1. The material used to form the core of the fiber must have a higher refractive index than the cladding material, to ensure the fiber is a guiding structure.

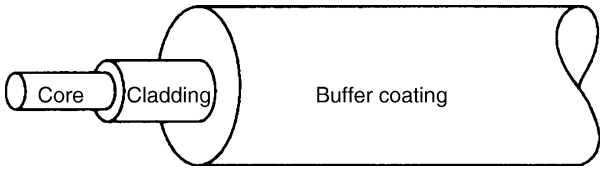


Figure 1 Structure of an optical fiber.

2. The materials used must be low loss, providing transmission with no absorption or scattering of light.
3. The materials used must have suitable thermal and mechanical properties to allow them to be drawn down in diameter into a fiber.

Silica (SiO_2) can be made into a glass relatively easily. It does not easily crystallize, which means that scattering from unwanted crystalline centers within the glass is negligible. This has been a key factor in the achievement of a low-loss fiber. Silica glass has high transparency in the visible and near-infrared wavelength regions and its refractive index can be easily modified. It is stable and inert, providing excellent chemical and mechanical durability. Moreover, the purification of the raw materials used to synthesize silica glass is quite straightforward.

In the first stage of achieving an optical fiber, silica glass is synthesized by one of three main chemical vapor processes. All use silicon tetrachloride ($SiCl_4$) as the main precursor, with various dopants to modify the properties of the glass. The precursors are reacted with oxygen to form the desired oxides. The end result is a high purity solid glass rod with the internal core and cladding structure of the desired fiber.

In the second stage, the rod, or preform as it is known, is heated to its softening temperature and stretched to diameters of the order of 125 microns. Tens to hundreds of kilometers of fiber are produced from a single preform, which is drawn continuously, with minimal diameter fluctuations. During the drawing process one or more protective coatings are applied, yielding long lengths of strong, low-loss fiber, ready for immediate application.

Preform Fabrication

The key step in preparing a low-loss optical fiber is to develop a technique for completely eliminating transition metal and OH ion contamination during the synthesis of the silica glass. The three methods most commonly used to fabricate a glass optical fiber preform are: the modified chemical vapor deposition process (MCVD); the outside vapor deposition process (OVD); and the vapor-axial deposition process (VAD).

In a typical vapor phase reaction, halide precursors undergo a high temperature oxidation or hydrolysis to form the desired oxides. The completed chemical reaction for the formation of silica glass is, for oxidation:

$$SiCl_4 + O_2 \rightarrow SiO_2 + 2Cl_2 \qquad [1]$$

For hydrolysis, which occurs when the deposition occurs in a hydrogen-containing flame, the reaction is:

$$SiCl_4 + 2H_2O \rightarrow SiO_2 + 4HCl \qquad [2]$$

These processes produce fine glass particles, spherical in shape with a size of the order of 1 nm. These glass particles, known as soot, are then deposited and subsequently sintered into a bulk transparent glass. The key to low-loss fiber is the difference in vapor pressures of the desired halides and the transition metal halides that cause significant absorption loss at the wavelengths of interest. The carrier gas picks up a pure vapor of, for example, $SiCl_4$, and any impurities are left behind.

Part of the process requires the formation of the desired core/cladding structure in the glass. In all cases, silica-based glass is produced with variations in refractive index produced by the incorporation of dopants. Typical dopants used are germania (GeO_2), titania (TiO_2), alumina (Al_2O_3), and phosphorous pentoxide (P_2O_5) for increasing the refractive index, and boron oxide (B_2O_2) and fluorine (F) for decreasing it. These dopants also allow other properties to be controlled, such as the thermal expansion of the glass and its softening temperatures. In addition, other materials, such as the rare earth elements, have also been used to fabricate active fibers that are used to produce optical fiber amplifiers and lasers.

MCVD Process

Optical fibers were first produced by the MCVD method in 1974, a breakthrough that completely solved the technical problems of low-loss fiber fabrication (Figure 2).

As shown schematically in Figure 3, the halide precursors are carried in the vapor phase by oxygen carrier gas into a pure silica substrate tube. An oxy-hydrogen burner traverses the length of the tube, which it heats externally. The tube is heated to temperatures of about 1,400°C which then oxidizes the halide vapor materials. The deposition temperature is sufficiently high to form a soot made of glassy particles which are deposited on the inside wall of the substrate tube but low enough to prevent the softened silica substrate tube from collapsing. The process usually takes place on a horizontal glass-working lathe.

During the deposition process, the repeated traversing of the burner forms multiple layers of soot. Changes to the precursors entering the tube and thus the resulting glass composition are introduced for layers which will form the cladding and then the core. The MCVD method allows germania to be doped into



Figure 2 Fabrication of an optical fiber preform by the MCVD method.

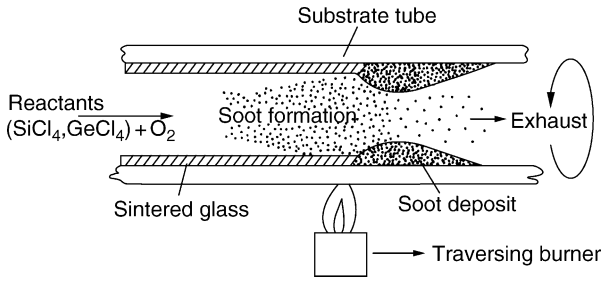


Figure 3 Schematic of preform fabrication by the MCVD method.

the silica glass and the precise control of the refractive index profile of the preform. When deposition is complete, the burner temperature is increased and the hollow, multilayered structure is collapsed to a solid rod. A characteristic of fiber formed by this process is a refractive index dip in the center of the core of the fiber. In addition, the deposition takes place in a closed system, which dramatically reduces contamination by OH^- ions and maintains low levels of other impurities. The high temperature allows high deposition rates compared to traditional chemical vapor deposition (CVD) and large performs, built up from hundreds of layers can be produced.

The MCVD method is still widely used today though it has some limitations, particularly on the preform size that can be achieved and thus the manufacturing cost. Diameter of the final preform is determined by the size of the initial silica substrate tube, which, due to the high purity required, accounts for a significant portion of the cost of the preform. A typical preform fabricated by MCVD yields about 5 km of fiber. Its success has spurred improvements to the process, in particular to address the fiber yield from a single preform.

OVD Process

The outside vapor deposition (OVD) methods, also known as outside vapor phase oxidation (OVPO), synthesize the fine glass particles within a burner flame. The precursors, oxygen, and fuel for the burner, are introduced directly into the flame. The soot is deposited onto a target rod that rotates and traverses in front of the burner. As in MCVD, the preform is built up layer by layer, though now initially by depositing the core glass and then building up the cladding layers over this. After the deposition process is complete, the preform is removed from the target rod and is collapsed and sintered into a transparent glass preform. The center hole remains, but disappears during the fiber drawing process. This technique has advantages in both size of preform which can be obtained and the fact that a high-quality silica substrate tube is no longer required. These two advantages combine to make a more economical process. From a single preform, several hundred kilometers of fiber can be produced.

VAD Process

The most recent refinement to the fabrication process was developed again to aid the mass production of high-quality fibers. In the VAD process, both core and cladding glasses are deposited simultaneously. Like OVD, the soot is synthesized and deposited by flame hydrolysis, as shown in Figure 4, the precursors are blown from a burner, oxidized, and deposited onto a silica target rod.

Burners for the VAD process consist of a series of concentrate nozzles. The first delivers an inert carrier and the main precursors $SiCl_4$, the second delivers an inert carrier glass and the dopants, the third delivers hydrogen fuel, and the fourth delivers oxygen. Gas flows are up to a liter per minute and deposition rates can be very high. With the VAD process, both core and cladding glasses can be deposited simultaneously. The main advantage is that this is a continuous process, as the soot which forms the core and cladding glasses are deposited axially onto the end of the rotating silica rod, it is slowly drawn upwards into a furnace which sinters and consolidates the soot into a transparent glass. The upward motion is such that the end at which deposition is occurring remains in a fixed position and essentially the preform is grown from this base.

The advantages of the VAD process are a preform without a central dip and, most importantly, the mass production associated with a continuous process. The glass quality produced is uniform and the resulting fibers have excellent reproducibility and low loss.

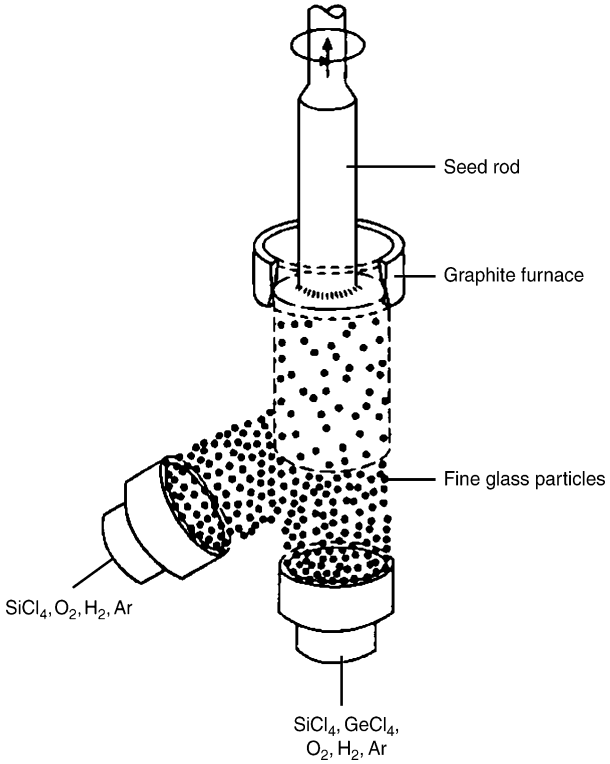


Figure 4 Schematic of preform fabrication by the VAD method.



Figure 5 A commercial scale optical fiber drawing tower.

Other Methods of Preform Fabrication

There are other methods of preform fabrication, though these are not generally used for today's silica glass based fiber. Some methods, such as fabrication of silica through sol–gel chemistry could not provide the large low-loss preforms obtained by chemical vapor deposition techniques. Other methods, such as direct casting of molten glass into rods, or formation of rods and tubes by extrusion, are better suited to and find application in other glass chemistries and speciality fibers.

Fiber Drawing

Once a preform has been made, the second step is to draw the preform down in diameter on a fiber drawing tower. The basic components and configuration of the drawing tower have remained unchanged for many years, although furnace design has become more sophisticated and processes like coating have become automated. In addition, fiber drawing towers have increased in height to allow faster pulling speeds (Figure 5).

Essentially, the fiber drawing process takes place as follows. The preform is held in a chuck which is mounted on a precision feed assembly that lowers the preform into the drawing furnace at a speed which matches the volume of preform entering the furnace to the volume of fiber leaving the bottom of the furnace. Within the furnace, the fiber is drawn from the molten zone of glass down the tower to a capstan, which controls the fiber diameter, and then onto a drum. During the drawing process, immediately below the furnace, is a noncontact device that measures the diameter of the fiber. This information is fed back to the capstan which speeds up to reduce the diameter or slows down to increase the fiber diameter. In this way, a constant diameter fiber is produced. One or more coatings are also applied in-line to maintain the pristine surface quality as the fiber leaves the furnace and thus to maximize the fiber strength.

The schematic drawing in Figure 6 shows a fiber drawing tower and its key components. Draw towers are commercially available, ranging in height from 3 m to greater than 20 m, with the tower height increasing as the draw speed increases.

Typical drawing speeds are on the order of $1\ m\ s^{-1}$. The increased height is needed to allow the fiber to cool sufficiently before entering the coating applicator, although forced air-cooling of the fiber is commonplace in a manufacturing environment.

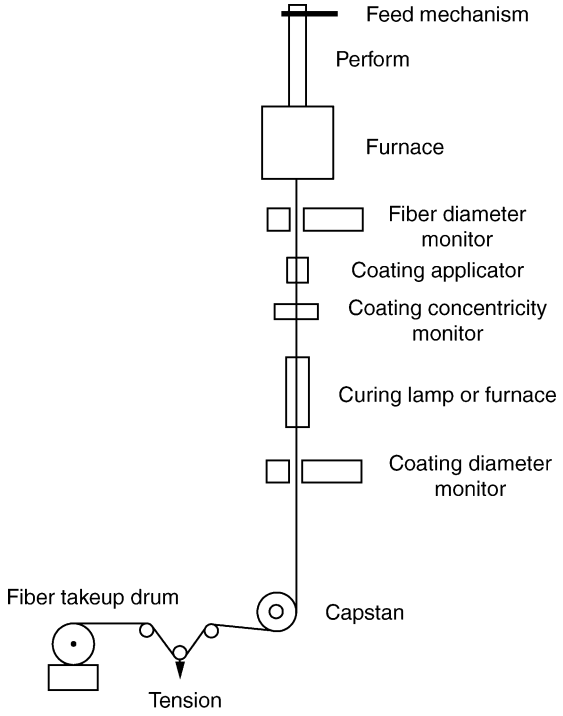


Figure 6 Schematic diagram of an optical fiber drawing tower and its components.

Furnaces

A resistance furnace is perhaps the most common and economical heating source used in a fiber drawing tower. A cylindrical furnace, with graphite or tungsten elements, provides heating to the preform by blackbody radiation. These elements must be surrounded by an inert gas, such as nitrogen or argon, to prevent oxidation. Gas flow and control is critical to prevent variations in the diameter of the fiber. In addition, the high temperature of the elements may lead to contamination of the preform surface that can then weaken the fiber.

An induction furnace provides precise clean heating through radio frequency (RF) energy that is inductively coupled to a zirconia susceptor ring. This type of furnace is cleaner than the graphite resistance type and is capable of continuous running for several months. It also has the advantage that it does not require a protective inert atmosphere and consequently the preform is drawn in a turbulence-free environment. These advantages result in high-strength fibers with good diameter control.

Diameter Measurement

A number of noncontact methods of diameter measurement can be applied to measure fiber diameter. These include laser scanning, interferometry, and light scattering techniques. To obtain precise control of the fiber diameter, the deviation between the desired diameter and the measured diameter is fed into the diameter control system. In order to cope with high drawing speeds, sampling rates as high as 1,000 times per second are used. The diameter control is strongly affected by the gas flow in the drawing furnace and is less affected by the furnace temperature variation. The furnace gas flow can be used to achieve suppression of fast diameter fluctuations. This is used in combination with drawing speed control to achieve suppression of both fast and slow diameter fluctuations. Current manufacturing processes are capable of producing several hundreds of kilometers of fiber with diameter variations of ± 1 micron.

There are two major sources of fiber diameter fluctuations: short-term fluctuations caused by temperature fluctuations in the furnace; and long-term fluctuations caused by variations in the outer diameter of the preform. Careful control of the furnace temperature, the length of the hot zone, and the flow of gas minimize the short-term fluctuations. Through optimization of these parameters, diameter errors of less than ± 0.5 microns can be realized. The long-term fluctuations in diameter are controlled by the feedback mechanism between the diameter measurement and the capstan.

Fiber Coating

A fiber coating is primarily used to preserve the strength of a newly drawn fiber and therefore must be applied immediately after the fiber leaves the furnace. The fiber coating apparatus is typically located below the diameter measurement gauge at a distance determined by the speed of fiber drawing, the tower height, and whether there is external cooling of the fiber. Coating is usually one of the limiting factors in the speed of fiber drawing. To minimize any damage or contamination of the pristine fiber leaving the furnace, this portion of the tower, from furnace to the application of the first coating, is enclosed in a clean, filtered-air chamber.

A wide variety of coating materials has been applied; however, conventional, commercial fiber generally relies on a UV curable polymer as the primary coating. Coating thickness is typically 50–100 microns. Subsequent coatings can then be applied for specific purposes. A dual coating is often used with an inner primary coating that is soft and an outer, secondary coating which is hard. This ratio of low to high elastic modulus can minimize stress on the fiber and reduce bending loss.

Alternative coatings include hermetic coatings of a low melting temperature metal, ceramics, or amorphous carbon. These can be applied in-line before the polymeric coating. Metallic coatings are applied by passing the fiber through a molten metal while ceramic or amorphous coatings utilize an in-line chemical vapor deposition reactor.

Types of Fiber

The majority of silica fiber drawn today is single mode. This structure consists of a core whose diameter is chosen such that, with a given refractive index difference between the core and cladding, only a single guide mode propagates at the wavelength of interest. With a discrete index difference between core and cladding, it is often referred to as a step index fiber. In typical telecommunications fiber, single mode operation is obtained with core diameters of 2–10 microns with a standard outer diameter of 125 microns.

Multimode fiber has core diameters considerably larger, typically 50, 62.5, 85, and 110 microns, again with a cladded diameter of 125 microns. Multimode fibers are often graded index, that is the refractive index is a maximum in the center of the fiber and smoothly decreases radially until the lower cladding index is reached. Multimode fibers find use in non-telecommunication applications, for example optical fiber sensing and medicine.

Single mode fibers, which are capable of maintaining a linear polarization input to the fiber, are known as polarization preserving fibers. The structure of these fibers provides a birefringence that removes the degeneracy of the two possible polarization modes. This birefringence is a small difference in the effective refractive index of the two polarization modes that can be guided and it is achieved in one of two ways. Common methods for the realization of this birefringence are an elliptical core in the fiber, or through stress rods, which modify the refractive index in one orientation. These fiber structures are shown in Figure 7.

While the loss minimum of silica-based fiber is near 1.55 microns, step index single-mode fiber offers zero dispersion close to 1.3 micron wavelengths and dispersion at the loss minimum is considerable. A modification of the structure of the fiber, and in particular a segmented refractive index profile in the core, can be used to shift this dispersion minimum to 1.55 microns. This fiber, illustrated in Figure 7 is known as dispersion shifted fiber. Similarly fibers, with a relatively low dispersion over a wide wavelength range, known as dispersion flattened fibers, can be obtained by the use of multiple cladding layers.

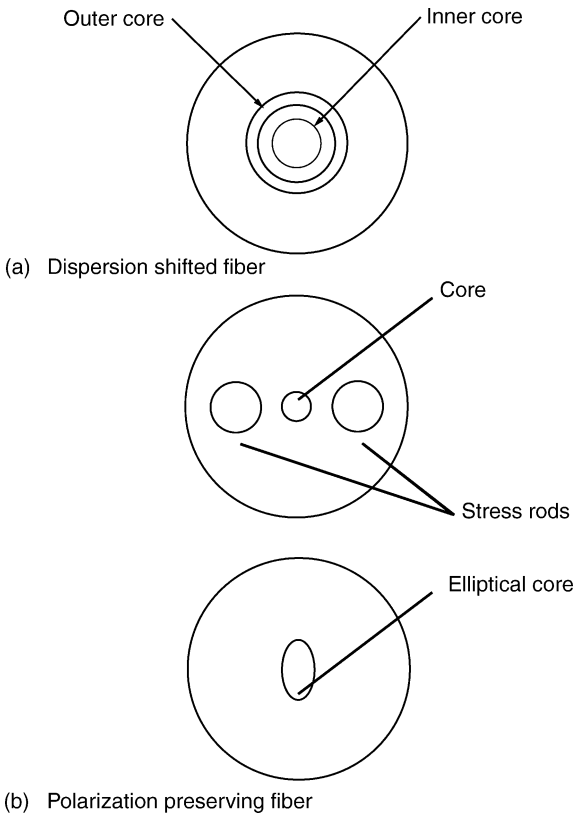


Figure 7 Structure of (a) dispersion shifted fiber and (b) two methods of achieving polarization preserving fiber.

List of Units and Nomenclature

Dopants	Elements or compounds added, usually in small amounts, to a glass composition to modify its properties.
Fiber drawing	The process of heating and thus softening an optical fiber preform and then drawing out a thin thread of glass.
Fiber loss	The transmission loss of light as it propagates through a fiber, usually measured in dB of loss per unit length of fiber. Loss can occur through the absorption of light in the core or scattering of light out of the core.
Glass	An amorphous solid formed by cooling from the liquid state to a rigid solid with no long range structure.
Modified chemical vapor deposition (MCVD)	A process for the fabrication of an optical preform where gases flow into the inside of a rotating tube, are heated and react

	to form particles of glass which are deposited onto the wall of the glass tube. After deposition, the glass particles are consolidated into a solid preform.
Outside vapor deposition (OVD)	A process for the fabrication of an optical preform where glass soot particles are formed in an oxy-hydrogen flame and deposited on a rotating rod. After deposition, the glass particles are consolidated into a solid preform.
Preform	A fiber blank, a bulk glass rod consisting of a core and cladding glass composite which is drawn into fiber.
Refractive index	A characteristic property of glass, which is defined by the speed of light in the material relative to the speed of light in a vacuum.
Silica	A transparent glass formed from silicon dioxide.
Vapor-axial deposition	A process similar to OVD, where the core and cladding layers are deposited simultaneously.

See also

Detection: Fiber Sensors. **Fiber and Guided Wave Optics:** Dispersion; Light Propagation; Measuring Fiber Characteristics; Nonlinear Effects (Basics); Optical Fiber Cables. **Fiber Gratings**. **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Historical Development; Wavelength Division Multiplexing. **Optical Materials:** Optical Glasses.

Further Reading

Agrawal GP (1992) *Fiber-Optic Communication Systems*. New York: John Wiley & Sons.

Bass M and Van Stryland EV (eds) (2001) *Fiber Optics Handbook: Fiber, Devices, and Systems for Optical Communications*. New York: McGraw-Hill Professional Publishing.

Fujiura K, Kanamori T and Sudo S (1997) Fiber materials and fabrications. In: Sudo S (ed.) *Optical Fiber Amplifiers*, ch. 4, pp. 193–404. Boston, MA: Artech House.

Goff DR and Hansen KS (2002) *Fiber Optic Reference Guide: A Practical Guide to Communications Technology*, 3rd edn. Oxford, UK: Butterworth-Heinemann UK.

Hewak D (ed.) (1998) *Glass and Rare Earth Doped Glasses for Optical Fibres*. EMIS Datareview Series, INSPEC. London: The Institution of the Electrical Engineers.

Keck DB (1981) Optical fibre waveguides. In: Barnoski MK (ed.) *Fundamentals of Optical Fiber Communications*. New York: Academic Press.

Li T (ed.) (1985) *Optical Fiber Communications: Fiber Fabrication*. New York: Academic Press.

Personick SD (1985) *Fiber Optics: Technology and Applications*. New York: Plenum Press.

Schultz PC (1979) In: Bendow B and Mitra SS (eds) *Fiber Optics*. New York: Plenum Press.

Light Propagation

F G Omenetto, Los Alamos National Laboratory, Los Alamos, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Optical fibers have become a mainstay of our modern way of life. From phone calls to the internet, their presence is ubiquitous. The unmatched ability that optical fibers have to transmit light is an amazing asset that, besides being strongly part of our present, is poised to shape the future.

The foundation of the unique ability of optical fibers to transmit light hinges on the well-known physical concept of total internal reflection. This phenomenon is described by considering the behavior of light traveling from one material (A) to a different material (B). When light traverses the interface of the two materials with a specific angle, its direction of propagation is altered with respect to the normal to the surface, according to the well-known Snell law which relates the incident angle (onto the interface between A and B) to the refracted angle (away from the interface between A and B). This is written as $n_A \sin i = n_B \sin r$, where i is the angle of incidence, r the angle of refraction, and n_A and n_B are the refractive indices of the materials A and B. Under appropriate conditions, a critical value of the incidence angle (i_{CR}) exists for which the light does not propagate into B and is totally reflected back into A. This value is given by $i_{CR} = \arcsin(n_A/n_B)$. It follows that total internal reflection occurs when light is traveling from a medium with a higher index of

refraction into a medium with a lower index of refraction. As an approximation, this is the basic idea behind light propagation in fibers.

Optical fibers are glass strands that include a concentric core with a cladding wrapped around it. The index of refraction of the core is slightly higher than the index of refraction of the cladding, thereby creating the necessary conditions for the confinement of light within them. Such index mismatches between the core and the cladding are determined by using different materials for each, by doping the glass matrix before pulling the fiber to create an index gradient, by introducing macroscopic defects such as air-holes in the structure of the fiber cladding (such as in the recently developed photonic crystal fibers), or by using materials other than glass, such as polymers or plastic.

The real physical picture of light guiding in fibers is more complex than the conceptual description stated above. The light traveling in the fiber is, more precisely, an electromagnetic wave with frequencies that lie in the optical range, and the optical fiber itself constitutes an electromagnetic waveguide. As such, the waveguide supports guided electromagnetic modes which can take on a multitude of shapes (i.e., of energy distribution profiles), depending on the core/cladding index of refraction profiles. These profiles are often complex and determine the physical boundary conditions that influence the field distribution within the fiber. The nature of the fiber modes has been treated extensively in the literature where an in-depth description of the physical solutions can be found.

There are many distinguishing features that identify different optical fiber types. Perhaps the most important one, in relation to the electromagnetic modes that the fiber geometry supports, is the distinction that is made between single-mode fibers and multimode fibers. As the name implies, a single-mode fiber is an optical fiber whose index of refraction profile between core and cladding is designed in such a way that there is only one electromagnetic field distribution that the fiber can support and transmit. That is to say, light is propagated very uniformly across the fiber. A very common commercially used fiber of this kind is the Corning SMF-28 fiber, which is employed in a variety of telecommunication applications. Multimode fibers have larger cores and are easier to couple light into, yet the fact that many modes are supported implies that there is less control over the behavior of light during its propagation, since it is more difficult to control each individual mode.

The condition for single-mode propagation in optical fibers is determined during the design and manufacture of the fiber by its geometrical parameters and is specific to a certain operation wavelength. A quantity V is defined as $V = (2\pi/\lambda) \times a(n_{\text{core}}^2 - n_{\text{cladding}}^2)^{1/2}$, where a is the core radius and λ is the wavelength of light. For example, a step-index fiber (i.e., a fiber where the transition from the index of refraction of the core to the index of refraction in the cladding is abrupt), supports a single electromagnetic mode if $V < 2.405$.

Light propagation in optical fibers, however, is affected by numerous other factors besides the electromagnetic modes that the fiber geometry allows.

First, the ability to efficiently transmit light in optical fiber is dependent on the optical transparency of the medium that the fiber is made of. The optical transparency of the medium is a function of the wavelength (λ) of light that needs to be transmitted through the fiber. Optical attenuation is particularly high for shorter wavelengths and the transmission losses for ultraviolet wavelengths become considerably higher than in the near infrared. The latter region, due to this favorable feature, is the preferred region of operation for optical fiber based telecommunication applications. This ultimate physical limit is determined by the process of light being scattered off the atoms that constitute the glass, a process called Rayleigh scattering. The losses due to Rayleigh scattering are inversely proportional to the fourth power of the wavelength (losses $\propto \lambda^{-4}$), and therefore become quite considerable as λ is decreased.

The ability of optical fibers to transmit light over a certain set of wavelengths, is quantified by their optical attenuation. This parameter is defined as the ratio of the output power versus the input power and is usually measured in units of decibels (dB), i.e., attenuation (dB) $= 10 \log_{10}(P_{\text{out}}/P_{\text{in}})$.

The values that are found in the literature relate optical attenuation to distance and express the attenuation per unit length (such as dB/km). Single-mode fibers used in telecommunications are usually manufactured with doped silica glasses designed to work in the near infrared, in a wavelength range between 1.3 and 1.6 μm ($1\ \mu\text{m} = 10^{-6}$ m) which provides the lowest levels of Rayleigh scattering and is located before the physical infrared absorption of silica (~1.65 μm) kicks in. Such fibers, which have been developed and refined for decades, have attenuation losses of less that 0.2 dB/km at a wavelength of 1.55 μm. Many other materials have been used to meet specific optical transmission needs. While silica remains by far the most popular material used for optical fiber manufacturing, other glasses are better suited for different portions of the spectrum: quartz glass can be particularly transmissive in the ultraviolet whereas plastic is sometimes

conveniently used for short distance transmission in the visible.

The ability to collect light at the input of the fiber represents another crucial and defining parameter. Returning to total internal reflection, such an effect can only be achieved by sending light onto an interface at a specific angle. In fibers, the light needs to be coupled into a circular core and reach the core cladding interface at that specific geometrical angle. The measure of the ability to couple light in the fiber is referred to as 'coupling efficiency'. The core dimension defines the geometry of the coupling process which, in turn, determines the efficiency of the propagation of light inside the fiber. In order to be coupled efficiently and therefore guided within the fiber, the light has to be launched within a range of acceptable angles. Such a range is usually defined or derivable from the numerical aperture (NA) of the fiber which is approximately $\mathrm{NA} = (n_{\mathrm{core}}^2 - n_{\mathrm{cladding}}^2)^{1/2}$, or can be alternatively described as a function of the critical angle for total internal reflection within the fiber by $\mathrm{NA} = n_{\mathrm{core}} \sin i_{\mathrm{CR}}$. Also, light has to be focused into the core of the fiber and therefore the interplay between core size, focusability, and NA (i.e., acceptance angle) of the fiber influence the coupling efficiency into the fiber (which in turn affects transmission). Typical telecommunication fibers have a 10 μm core (a NA of ~0.2) and have very high coupling efficiencies (of the order of 70–80%). Specialized single mode fibers (such as photonic crystal fibers) can have core diameters down to a few microns, significantly lower coupling efficiencies, and high numerical apertures. Multimode fibers, in contrast, are easy to couple into given their large core size.

Other factors that contribute to losses during propagation are due to the physical path that the actual optical fiber follows. A typical optical fiber strand is rarely laid along a straight line. These types of losses are called bending losses and can be intuitively understood by thinking that when the fiber is straight, the light meets its total reflection condition at the core-cladding interface, but when the fiber bends, the interface is altered and varies the incidence conditions of light at the interface. If the bend is severe, light is no longer totally reflected and escapes from the fiber.

Another important issue that influences propagation arises when the light that travels through the fiber is sent in bursts or in pulses of light. Pulsed light is very important in fiber transmission because the light pulse is the carrier of optical information, most often representing a binary digit of data. The ability to transmit pulses reliably is at the heart of modern day telecommunication systems. The features that affect the transmission of pulses through optical fibers depend, again, largely on the structure and constituents of the fiber. One of the most important effects on pulsed light is dispersion. This phenomenon causes the broadening in time of the light pulses during their travel through the fiber. This can be easily explained by a Fourier analogy by thinking that a short pulse in time (such as the 10–20 picosecond pulse duration that is used for optical communications, $1\ \text{picosecond} = 10^{-12}$ seconds) has a large frequency content (i.e., is composed of a variety of 'colors'). Since the speed of light through bulk media, and therefore through the fiber constituent, depends on wavelength, different 'colors' will travel at different speeds, arriving at the end of the fiber slightly delayed with respect to one another. This causes a net pulse broadening and is dependent on the properties of the fiber that are depicted in the dispersion curve of the fiber. This type of dispersion is called chromatic dispersion and is dependent on the distance the pulse travels in the fiber. A measure of the pulse broadening can be calculated by multiplying the dispersion value for the fiber (measured in ps/km) at the wavelength of operation times and the distance in kilometers that the pulse travels.

There are other types of dispersion that affect pulses traveling through fibers: light with different polarization travels at a different speed, thereby causing an analogous problem to chromatic distortion. In multimode fibers, different modes travel with different properties. Dispersion issues are extremely important in the design of the next-generation telecommunication systems. As a push is made toward higher transmission capacity, pulses become shorter and these issues have to be dealt with carefully. Specifically, transmission of light in optical fibers becomes more complex because the interaction between the light and the fiber material constituents becomes nonlinear. As pulses become shorter, their energy is confined in a smaller temporal interval, making their peak power (which is equal to energy/time) increase. If the peak power becomes sufficiently high, the atoms that form the glass of which the fiber is made are excited in a nonlinear fashion adding effects that need to be addressed with caution. Energetic pulses at a certain wavelength change their shape during propagation and can become severely distorted or change wavelength (i.e., undergo a frequency conversion process), thereby destroying the information that they were meant to carry.

Several physical manifestations of optical nonlinearity affect pulsed light propagation, giving rise to new losses but also providing new avenues for efficient transmission.

Perhaps the most successful example of the positive outcome of nonlinear effects in fibers is the optical soliton. The latter is an extremely stable propagating pulse that is transmitted through the fiber undistorted, since it is a result of a peculiar balance between the linear distortion and nonlinear distortion that cancel each other by operating in a specific region of dispersion of the optical fiber (the anomalous dispersion region).

Nonlinear effects in fibers are an extremely rich area of study which, besides several drawbacks, carries immense opportunities if the nonlinearities can be controlled and exploited.

See also

Dispersion Management. **Nonlinear Optics, Applications:** Pulse Compression via Nonlinear Optics. **Solitons:** Optical Fiber Solitons, Physical Origin and Properties.

Further Reading

Hecht J (2001) *Understanding Fiber Optics*. New York: Prentice Hall.

Snyder AW and Love JD (1983) *Optical Waveguide Theory*. Boston: Kluwer Academic Press.

Agrawal GP (2001) *Nonlinear Fiber Optics*. San Diego: Academic Press.

Measuring Fiber Characteristics

A Girard, EXFO, Quebec, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The optical fiber is divided in two types: multimode and singlemode. Each type is used in different applications and wavelength ranges and is consequently characterized differently. Furthermore, the corresponding test methods also vary.

The optical fiber characteristics may be divided into four categories:

- The optical characteristics (transmission related);
- The dimensional characteristics;
- The mechanical characteristics; and
- The environmental characteristics.

These categories will be reviewed in the following sections, together with their corresponding test methods.

Fiber Optical Characteristics and Corresponding Tests Methods

The following sections will describe the following optical characteristics:

- attenuation;
- macrobending sensitivity;
- microbending sensitivity;
- cut-off wavelength;
- multimode fiber bandwidth;
- differential mode delay for multimode fibers;
- chromatic dispersion;
- polarization mode dispersion;
- polarization crosstalk; and
- nonlinear effects.

Attenuation

The spectral attenuation of an optical fiber follows exponential power decay from the power level at a cross-section 1 to the power level at cross-section 2, over a fiber length L, as follows:

$$P_2(\lambda) = P_1(\lambda) \cdot e^{-\gamma(\lambda)L} \qquad [1]$$

$P_1(\lambda)$ is the optical power transmitted through the fiber core cross-section 1, expressed in mW; $P_2(\lambda)$ is the optical power transmitted through the fiber core cross-section 2 away from cross-section 1, expressed in mW; $\gamma(\lambda)$ is the spectral attenuation coefficient in linear units, expressed in km^{-1}; and L is the fiber length expressed in km.

Attenuation may be characterized at one or more specific wavelengths or as a function of wavelength. In the later case, attenuation is referred to spectral attenuation. Figure 1 illustrates such power decay.

Equation [1] may be expressed in relative units as follows:

$$\log_{10} P_2 = (\log_{10} P_1) \cdot - \gamma L \cdot \log_{10} e \qquad [2]$$

P is expressed in dBm units using the following definition.

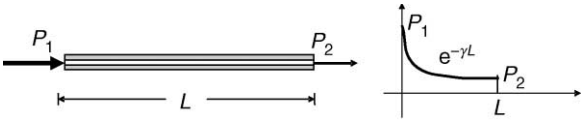


Figure 1 Power decay in an optical fiber due to attenuation.

The power in dBm is equal to 10 times the logarithm in base 10 of the power in mW; or

$$0\ \text{dBm} = 10\log_{10}(1\ \text{mW}) \quad [3]$$

Then

$$\gamma(\lambda)[\text{km}^{-1}] = [(10\log_{10} P_1) - (10\log_{10} P_2)] \div L\log_{10} e \quad [4]$$

$$\alpha(\lambda)[\text{dB/km}] = [P_1(\text{dBm}) - P_2(\text{dBm})]/L \quad [5]$$

A new relative attenuation coefficient $\alpha(\lambda)$ with units of dB/km has been introduced, which is related to the linear coefficient $\gamma(\lambda)$ with units of km^{-1} as follows:

$$\alpha(\lambda) = (\log_{10} e)\cdot\gamma(\lambda) \approx 4{,}343\,\gamma(\lambda) \quad [6]$$

$\alpha(\lambda)$ is the spectral attenuation coefficient of a fiber and is illustrated in Figure 2.

Test methods for attenuation

The attenuation may be measured by the following methods:

- cut-back method;
- backscattering method; and
- insertion loss method.

Cut-back method. The cut-back method is a direct application of the definition in which the power levels P_1 and P_2 are measured at two points of the fiber change of input conditions. P_2 is the power emerging from the far end of the fiber and P_1 is the power emerging from a point near the input after cutting the fiber.

The output power P_2 is recorded from the fiber under test (FUT) placed in the measurement setup. Keeping the launching conditions fixed, the FUT is cut to the cut-back length, for example 2 m from the launching point. The FUT attenuation, between the points where P_1 and P_2 have been measured, is calculated using P_1 and P_2, from the definition equations provided above.

Backscattering method. The attenuation coefficient of a singlemode fiber is characterized using bidirectional backscattering measurements. This method is also used for:

- attenuation uniformity;
- optical continuity;
- physical discontinuities;
- splice losses; and
- fiber length.

An optical time domain reflectometer (OTDR) is used for performing such characterization. Adjustment of laser pulsewidth and power is used to obtain a compromise between resolution (a shorter pulsewidth provides a better resolution but at lower power) and dynamic range/fiber length (higher power provides better dynamic range but with longer pulsewidth). An example of such an instrument is shown in Figure 3.

The measurement is applicable to various FUT configurations (e.g., cabled fiber in production or deployed in the field, fiber on a spool, etc.). Two unidirectional backscattering curves are obtained, one from each end of the fiber (Figure 4).

Each backscattering curve is recorded on a logarithmic scale, avoiding the parts at the two ends of the curves, due to parasitic reflections.

The FUT length is found from the time interval between the two ends of the backscattering loss

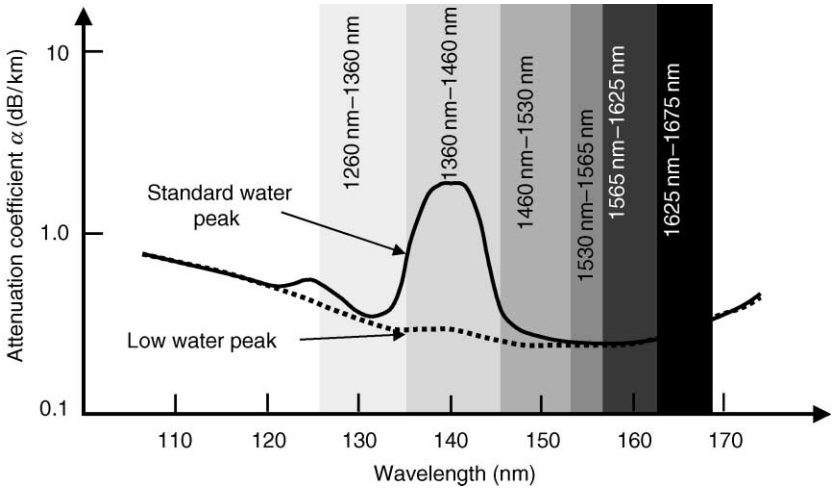


Figure 2 Typical spectral attenuation of a singlemode fiber.

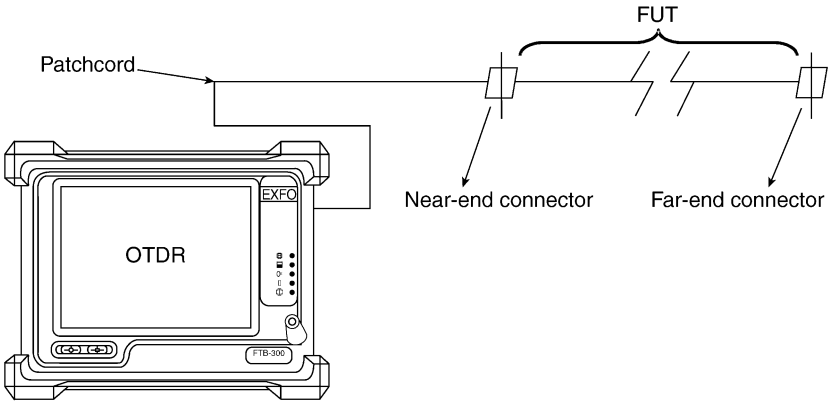


Figure 3 The backscattering method (OTDR).

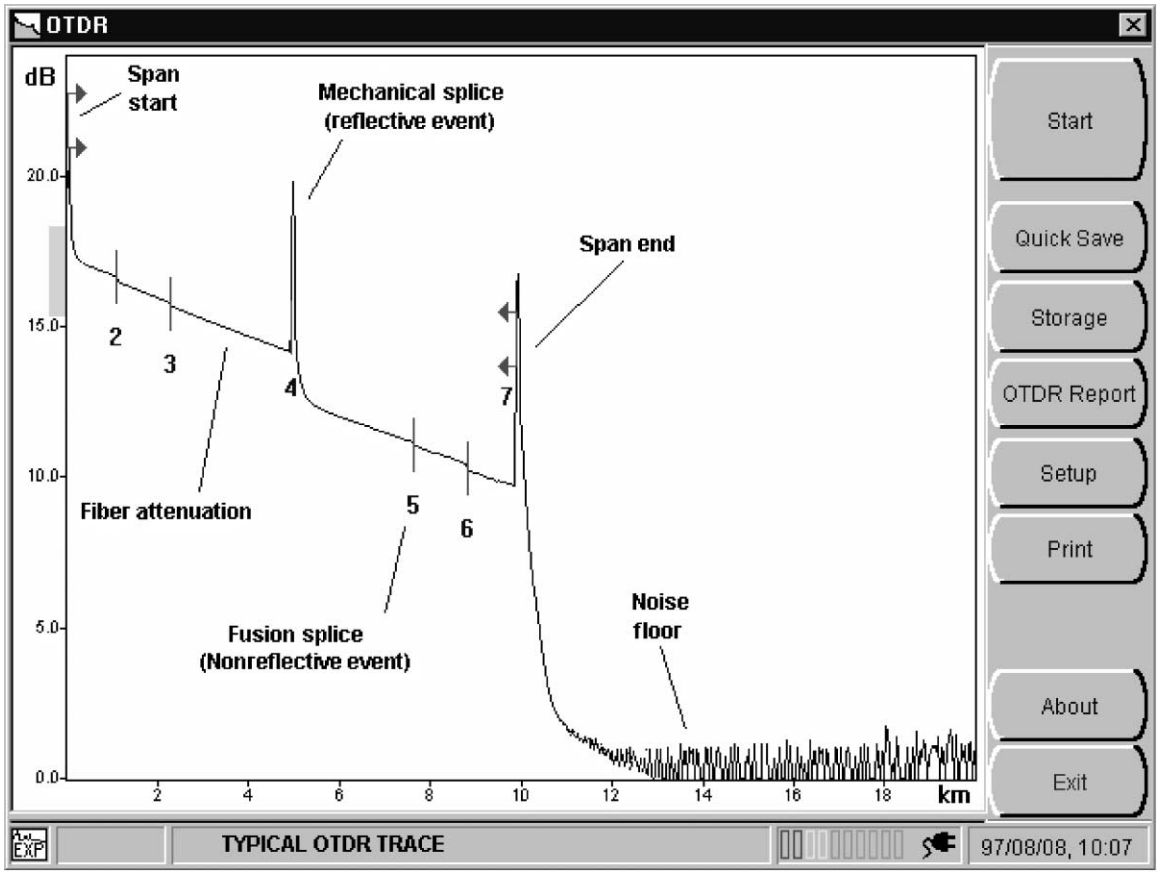


Figure 4 Unidirectional OTDR backscattering loss measurement.

curve together with the FUT group index of refraction, n_g, as:

$$L_f = c_{fs} \cdot \frac{\Delta T_f}{n_g} \quad [7]$$

c_{fs} is the velocity of light in free space.

The bidirectional backscattering curve is obtained using two unidirectional backscattering curves and calculating the average loss between them. The end-to-end FUT attenuation coefficient is obtained from the difference between two losses divided by the difference of their corresponding distances.

Insertion loss method. The insertion loss method consists of the measurement of the power loss due to

the FUT insertion between a launching and a receiving system, previously interconnected (reference condition). The powers P_1 and P_2 are evaluated in a less straightforward way than in the cut-back method. Therefore, this method is not intended for use in manufacturing.

The insertion loss technique is less accurate than the cut-back, but has the advantage of being non-destructive for the FUT. Therefore, it is particularly suitable in the field.

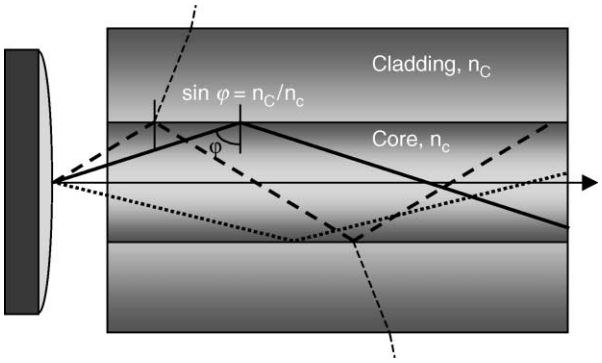


Figure 5 Total internal reflection and macrobending effect on the light rays.

Macrobending Sensitivity

Macrobending sensitivity is the property by which there is a certain amount of light leaking (loss) into the cladding when the fiber is bent and the bending angle is such that the condition of total internal reflection is no longer met at the core-cladding interface. Figure 5 illustrates such a case.

Macrobending sensitivity is a direct function of the wavelength: the longer the wavelength and/or the smaller the bending diameter, the more loss the fiber experiences. It is recognized that 1625 nm is a wavelength that is very sensitive to macrobending (see Figure 6).

The macrobending loss is measured by the power monitoring method (OTDR, especially for field assessment, see Figure 6) or the cut-back method.

Microbending Sensitivity

Microbending is a fiber property by which the core-cladding concentricity randomly changed along

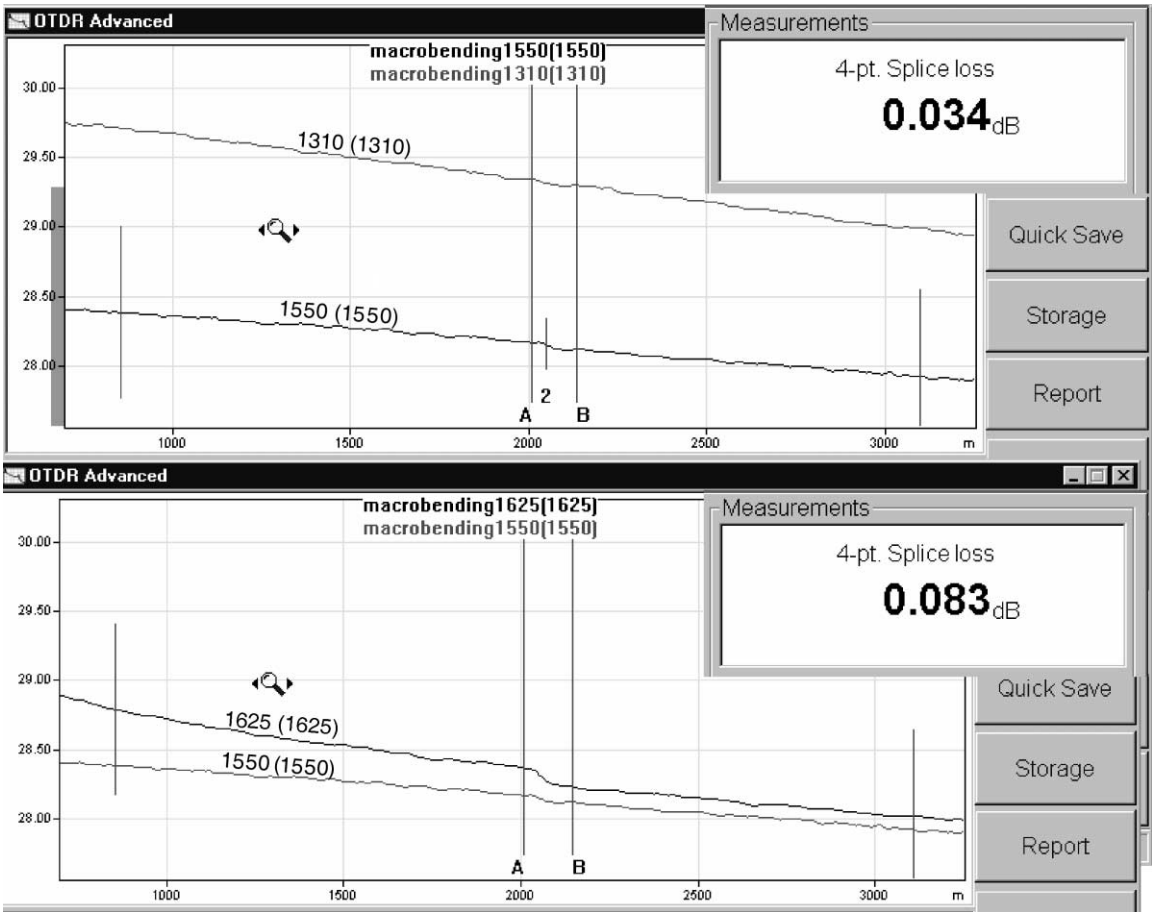


Figure 6 Macrobending sensitivity as a function of wavelength.

the fiber length. It causes the core to wobble inside the cladding and along the fiber length.

Four methods are available for characterizing microbending sensitivity in optical fibers:

- expandable drum for singlemode fibers and optical fiber ribbons over a wide range of applied linear pressure or loads;
- fixed diameter drum for step-index multimode, singlemode, and ribbon fibers for a fixed linear pressure;
- wire mesh and applied loads for step-index multimode and singlemode fibers over a wide range of applied linear pressure or loads; and
- 'basketweave' wrap on a fixed diameter drum for singlemode fibers.

The results from the four methods can only be compared qualitatively. The test is nonroutine for general evaluation of optical fiber.

Cut-off Wavelength

The cut-off wavelength is the shortest wavelength at which a single mode can propagate in a singlemode fiber. This parameter can be computed from the fiber refractive index profile (RIP). At wavelengths below the cut-off wavelength, several modes propagate and the fiber is no longer singlemode, but multimode.

In optical fibers, the change from multimode to singlemode behavior does not occur at a specific wavelength, but rather over a smooth transition as a function of wavelengths. Consequently, from a fiber-optic network standpoint, the actual threshold wavelength for singlemode performance is more critical. Thus an effective cut-off wavelength is described below.

The cut-off wavelength is defined as the wavelength greater than the ratio between the total power, in the higher-order modes and the fundamental mode, which has decreased to less than 0.1 dB. According to this definition, the second-order mode LP_{11} undergoes 19.3 dB more attenuation than the fundamental LP_{01} mode when the modes are equally excited.

Because the cut-off wavelength depends on the fiber length, bends, and strain, it is defined on the basis of the following three cases:

- fiber cut-off wavelength;
- cable cut-off wavelength; and
- jumper cable cut-off wavelength.

Fiber cut-off wavelength: Fiber cut-off wavelength λ_{fc} is defined for uncabled primary-coated fiber and is measured over 2 m with one loop of 140 mm radius loosely constrained, with the rest of the fiber kept essentially straight. The presence of a primary coating on the fiber usually will not affect λ_{fc}. However, the presence of a secondary coating may result in λ_{fc} being significantly shorter than that of the primary coated fiber.

Cable cut-off wavelength: Cable cut-off wavelength is measured prior to installation on a substantially straight 22 m cable length prepared by exposing 1 m of primary-coated fiber at both ends, the exposed ends each incorporating a 40 mm radius loop. Alternatively, this parameter may be measured on 22 m primary-coated uncabled fiber in the same configuration as for the λ_{fc} measurement.

Jumper cable cut-off wavelength: Jumper cable cut-off wavelength is measured over 2 m with one loop of 76 mm radius, or equivalent (e.g., split mandrel), with the rest of the jumper cable kept essentially straight.

Multimode Fiber Bandwidth for Multimode Fibers

The −3 dB bandwidth of a multimode optical fiber (or modal bandwidth) is defined as the lowest frequency where the magnitude of the baseband frequency response in optical power has decreased by 3 dB relative to the power at zero frequency. Modal bandwidth is also called intermodal dispersion as it takes into account the dispersion between the modes of propagation of the transmitted signal into the multimode fiber.

Various methods of reporting the results are available, but the results are typically expressed in terms of the −3 dB (optical power) frequency. Figure 7 illustrates modal bandwidth.

The bandwidth or pulse broadening may be normalized to a unit length, such as GHz · km, or ns/km.

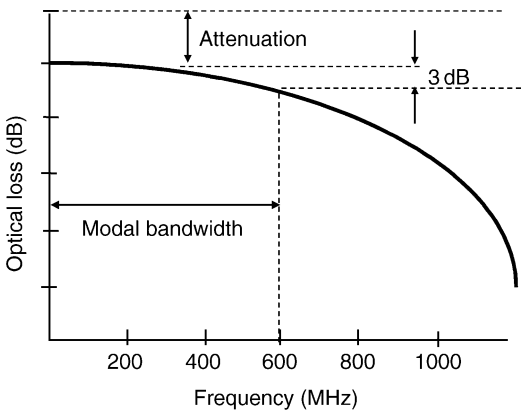


Figure 7 Determination of modal bandwidth.

Two methods are available for determining transmission capacity of multimode fibers:

- the frequency domain measurement method in which the baseband frequency response is directly measured in the frequency domain by determining the fiber response to a sinusoidally modulated light source; or
- the optical time domain measurement method (pulse distortion) in which the baseband response is measured by observing the broadening of a narrow pulse of light.

Differential Mode Delay for Multimode Fibers

Differential mode delay (DMD) characterizes the modal structure of a graded-index glass-core multimode fiber. DMD is useful for assessing the bandwidth performance of a fiber when used with short-pulse, narrow spectral-width laser sources.

The output from a singlemode probe fiber excites the multimode FUT at the test wavelength. The probe spot is scanned across the FUT endface, and the optical pulse delay is determined at specified radial offset positions between an inner and an outer limit. DMD is the difference in optical pulse delay time between the FUT fastest and slowest modes excited for all radial offset positions between and including the inner and the outer limits.

The related critical issues influencing DMD are the temporal width of the optical pulse, jitter in the timing, the finite bandwidth of the optical detector, and the mode broadening due to the source spectral width and the FUT chromatic dispersion.

The test method is commonly used in production and research facilities, but is not easily accomplished in the field. DMD may be a good predictor of the source launching conditions. DMD may be normalized to a unit length, such as ps/m.

Chromatic Dispersion

Chromatic dispersion in a singlemode fiber is a combination of material dispersion and waveguide dispersion (see Figure 8), and it contributes to pulse broadening and distortion in a digital signal.

Material dispersion is produced by the dopants used in glass and is important in all fiber types. Waveguide dispersion is produced by the wavelength dependence of the index of refraction and is critical in singlemode fibers only.

From the point of view of the transmitter, this is due to two causes:

- The presence of wavelengths in the source optical spectrum. Each wavelength has a different phase delay and group delay (different group velocities) along the fiber, because they travel under different index of refraction (or phase) as the index varies as a function of wavelengths, as shown in Figure 9.
- The other cause is the modulation of the source, which itself has two effects:
  - As bit-rates increase, the spectral width of the modulated signal increases and can become comparable to or exceed the spectral width of the source.
  - Chirp occurs when the source wavelength spectrum varies during the pulse. By convention, positive chirp at the transmitter occurs when the spectrum during the rise/fall of the pulse shifts towards shorter/longer wavelengths respectively. For a positive fiber dispersion coefficient, longer wavelengths are delayed relative to shorter wavelengths. Hence if the sign of the product of chirp and dispersion is positive, the two processes combine to produce pulse broadening. If the product is negative, pulse compression can occur over an initial fiber length until the pulse reaches a minimum width and then broadens again with increasing dispersion.

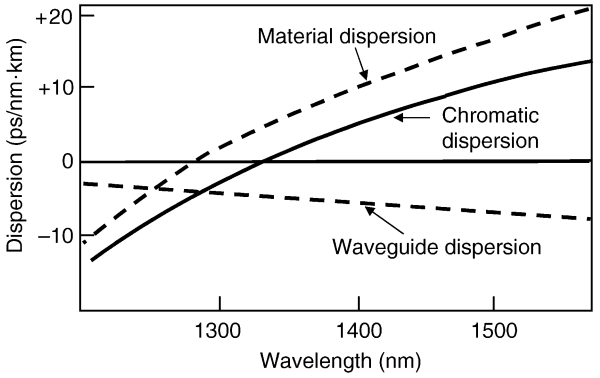


Figure 8 Contribution of the material and waveguide dispersions to the chromatic dispersion.

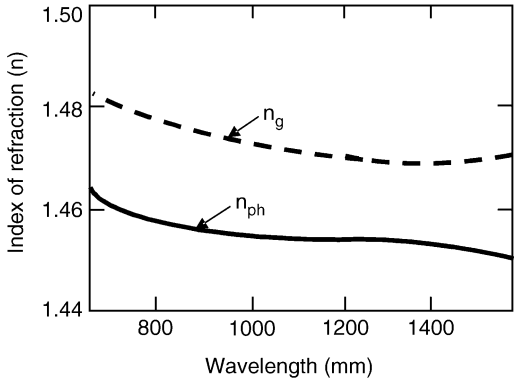


Figure 9 Difference between the phase and the group index of refraction.

The electric field propagating into the FUT may be simply described as follows:

$$E(t,z) = E_0 \sin(\omega t - \beta z) \quad [8]$$

$\omega = 2\pi c/\lambda$ [rad/s] is the angular frequency; $\beta = kn = (\omega/c)n$ is the effective index; as β has units of m^{-1} it is often referred to as wavenumber or even sometimes propagation constant; k is the propagation constant.

The group delay τ_g is then given by:

$$d\beta/d\omega = \beta_1 = \tau_g \quad [9]$$

An example of the group delay spectral distribution is shown in Figure 10. Assuming n is not complex; β_1 is the first-order derivative of β.

The group velocity v_g is given by

$$v_g = (d\beta/d\omega)^{-1} = -(\lambda^2/2\pi c)(d\beta/d\lambda)^{-1} \quad [10]$$

The FUT input–output delay is given by

$$\tau_d = L/v_g \quad [11]$$

L is the FUT length.

The group index of refraction n_g is given by

$$n_g = c/v_g = n - \lambda(dn/d\lambda) \quad [12]$$

The dispersion parameter or dispersion coefficient D (ps/nm · km) is given by

$$D = -(\omega/\lambda)(d\tau_g/d\omega) = -(2\pi c/\lambda^2)(d^2\beta/d\omega^2)$$

$$= -(\lambda/c)(d^2n/d\lambda^2) \quad [13]$$

$$d^2\beta/d\omega^2 = \beta_2 \quad [14]$$

An example of the spectral distribution of D obtained from the group delay is shown in Figure 10.

β_2 (ps^2/km) is the group velocity dispersion parameter, so D may be related to β_2, as follows:

$$D = -(\omega/\lambda)\beta_2 \quad [15]$$

When β_2 is positive then D is negative and vice-versa. The region where β_2 is positive is called normal dispersion while the negative-β_2 region is called anomalous dispersion.

At λ_0, β_1 is minimum, and $\beta_2 = 0$, then $D = 0$.

The dispersion slope S (ps/nm^2 · km), also called the differential dispersion parameter or second-order dispersion, is given by

$$S = dD/d\lambda = (\omega/\lambda)\beta_3 + (2\omega/\lambda^2)\beta_2 \quad [16]$$

$$\beta_3 = d\beta_2/d\omega = d^3\beta/d\omega^3$$

At λ_0, β_1 is minimum, $\beta_2 = 0$, then $D = 0$; but S is not zero and depends on β_3. An example of the spectral distribution of S and S_0 is illustrated in Figure 10.

Overall, the general expression of β is given by

$$\beta(\omega) = \beta_0 + (\omega - \omega_0)\beta_1 + (1/2)(\omega - \omega_0)^2\beta_2$$

$$+ (1/12)(\omega - \omega_0)^3\beta_3 + \cdots \quad [17]$$

Figure 11 illustrates the difference between dispersion unshifted fiber (ITU-T Rec. G.652), dispersion shifted fiber (ITU-T Rec. G.653) and nonzero dispersion shifted fiber (ITU-T Rec. G.655).

Test methods for the determination of chromatic dispersion

All methods measure the group delay at a specific wavelength over a range and use agreed fitting functions to evaluate λ_0 and S_0.

In the phase shift method, the group delay is measured in the frequency domain, by detecting, recording, and processing the phase shift of a sinusoidal modulating signal between a reference, a secondary fiber path, and the channel signal.

Setup variances exist and some do not require the secondary reference path. For instance, by using a reference optical filter at the FUT output it is possible to completely decouple the source from the phase-meter. With such an approach, chromatic dispersion may now be measured in the field over very long links using optical amplifiers (see Figure 12).

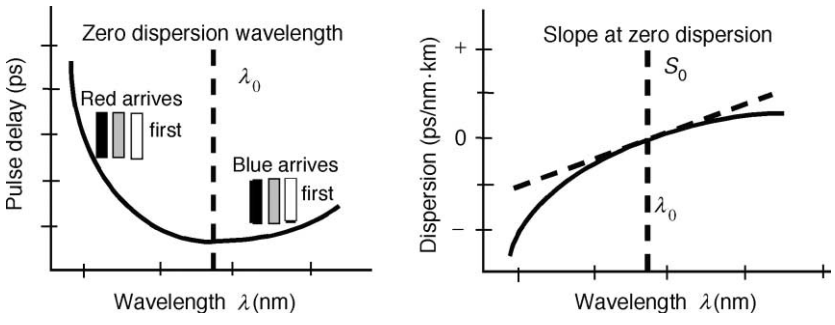


Figure 10 Relation between the pulse (group) delay and the (chromatic) dispersion.

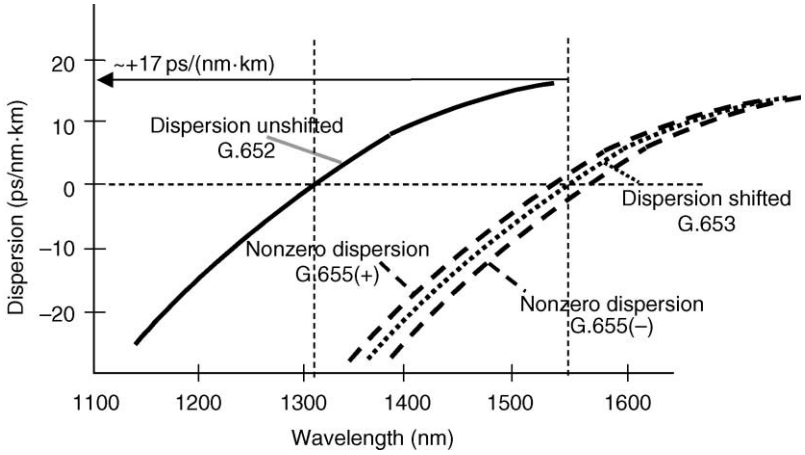


Figure 11 Chromatic dispersion for various types of fiber.

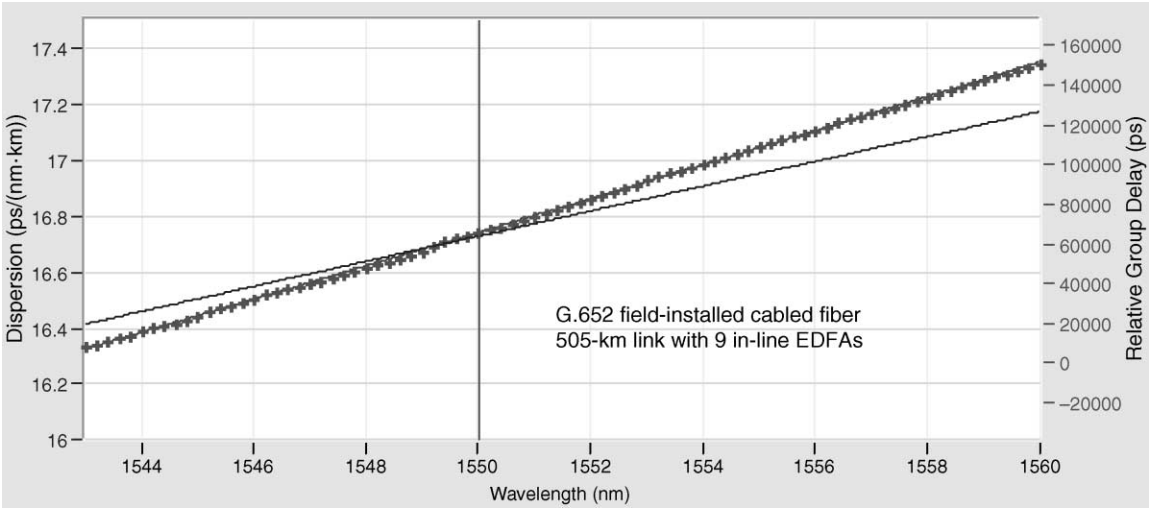


Figure 12 Test results for field-related chromatic dispersion using the phase-shift technique.

In the differential phase shift method, two detection systems are used together with two wavelength sources at the same time. In this case the chromatic dispersion may be determined directly from the two group delays. This technique usually offers faster and more reliable results but costs much more than the phase-shift technique which is usually preferred.

In the interferometric method, the group delay between the FUT and a reference path is measured by a Mach–Zehnder interferometer. The reference delay line may be an air path or a singlemode fiber standard reference material (SRM). The method can be used to determine the following characteristics:

- longitudinal chromatic dispersion homogeneity; and
- effect of overall or local influences, such as temperature changes and macrobending losses.

In the pulse delay method, the group delay is measured in the time domain, by detecting, recording, and processing the delay experienced by pulses at various wavelengths.

Polarization Mode Dispersion

Polarization mode dispersion (PMD) causes an optical pulse to spread in the time domain and may impair the performance of a telecommunications system. The effect can be related to differential phase and group velocities and corresponding arrival time of different polarization components of the pulse signal. For a sufficiently narrowband source, the effect can be related to a differential group delay (DGD), $\Delta\tau$, between a pair of orthogonally polarized principal states of polarization (PSP) at a given wavelength or optical frequency (see **Figure 13a**).

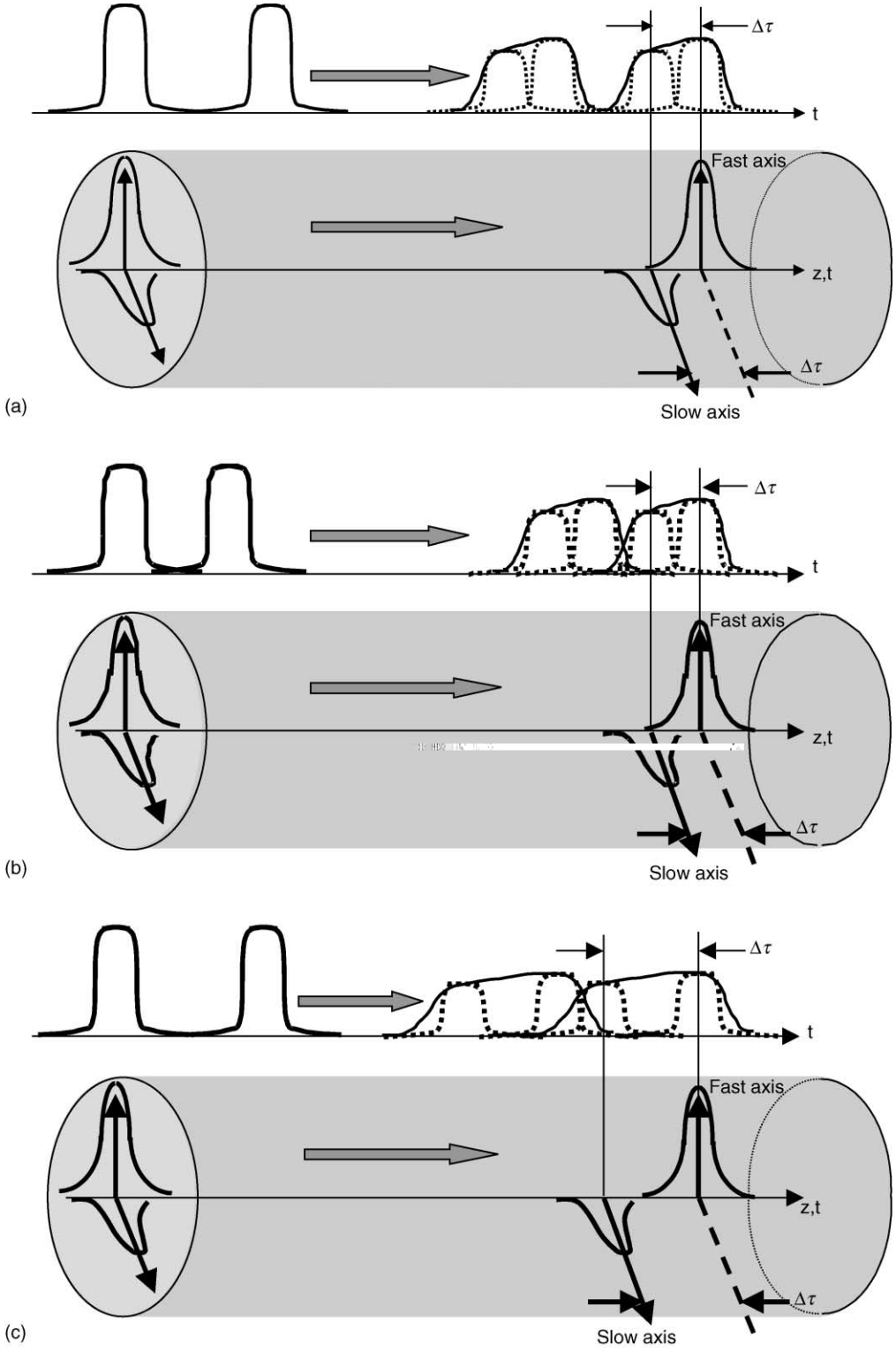


Figure 13 PMD effect on the pulse broadening and its possible pulse impairment. (a) The pulse is spread by the DGD $\Delta\tau$ but the bit rate is too low to create an impairment. (b) The pulse is spread by the DGD $\Delta\tau$ but the bit rate is high enough to create an impairment. (c) The DGD $\Delta\tau$ is large enough even at low bit rate to make the pulse spreading and creating impairment.

In an ideal circular symmetric fiber, the two PSPs propagate with the same velocity. However:

- a real fiber is not perfectly circular;
- the core is not perfectly concentric with the cladding;
- the core may be subjected to microbending;
- the core may present localized clusters of dopants; and
- the environmental conditions may stress the deployed cable and affect the fiber.

Each time the fiber undergoes local stresses and consequently birefringence. These asymmetry characteristics vary randomly along the fiber and in time, lead to a statistical behavior of PMD.

For a deployed cabled fiber at a given time and optical frequency, there always exist two PSPs such that the pulse spreading due to PMD vanishes, if only one PSP is excited. On the contrary, the maximum pulse spread due to PMD occurs when both PSPs are equally excited, and is related to the difference in their group delays, the DGD associated with the two PSPs. For broadband transmission, the DGD statistically varies as a function of wavelengths or frequencies and result in an output pulse that is spread out in the time domain (see **Figures 13a–c**). In this case, the spreading can be related to the RMS (root mean square) of DGD values $\langle \Delta\tau^2 \rangle^{1/2}$. However, if a known distribution such as the Maxwell distribution may be fit to the DGD distribution probability, then a mean (or average) value of the DGD $\langle \Delta\tau \rangle$ may be correlated to the RMS value and used as a system performance predictor in particular with a maximum value of the DGD distribution associated to a low probability of occurrence. This maximum DGD may then be used to define the quality of service that would tolerate values lower than this maximum DGD.

Test methods for polarization mode dispersion

Three methods are generically used for measuring PMD. Other methods or analyses may exist but they are generally not standardized or are limited in their applications.

- Stokes parameter evaluation (SPE)
  - Jones matrix eigenanalysis (JME)
  - Poincaré sphere analysis (PSA)
- Fixed analyzer (FA)
  - Extrema counting (EC)
  - Fourier transform (FT)
- Interferometry (INTY)
  - Traditional analysis (TINTY)
  - General analysis (GINTY).

All methods use a linearly polarized source at the FUT input and are suitable for laboratory measurements of factory lengths of fiber and cable. However, the interferometric method is the only method appropriate for measurements of cabled fiber that may be moving or vibrating such as is found in the field.

Stokes parameter evaluation. SPE determines PMD by measuring a response to a change of narrowband light (from a tuneable light source with broadband detector – JME, or a broadband source with a filtered detector such as an interferometer – PSA) across a wavelength range. The Stokes vector of the output light is measured for each wavelength. The change of these Stokes vectors with angular optical frequency (wavelength), ω and with the change in input SOP (state of polarization), yields the DGD as a function of wavelength.

For both JME and PSA analyses, three distinct and known linear SOPs (orthogonal on the Poincaré sphere) must be launched for each wavelength. **Figure 14** illustrates the test setup and examples of test results.

The JME and PSA method are mathematically equivalent.

Fixed analyzer. FA determines PMD by measuring a response to a change of narrowband light across a wavelength range. For each SOP, the change in output power that is filtered through a fixed polarization analyzer, relative to the power detected without the analyzer, is measured as a function of wavelength. **Figure 15** illustrates a test setup and examples of test results.

The resulting measured function can be analyzed in one of two ways:

- by counting the number of peaks and valleys (EC) of the curve and application of a formula. This analysis is considered as a frequency domain approach; and
- by taking the Fourier transform (FT) of the measured function. This FT is equivalent to the pulse spreading obtained by TINTY.

Interferometry. INTY uses a broadband light source and an interferometer. The fringe pattern containing the source auto-correlation, together with the PMD related cross-correlation of the emerging electromagnetic field, is determined by the interference pattern of the output light, i.e., the interferogram. The PMD determination for the wavelength range associated with the source spectrum is based on the envelope of the fringe

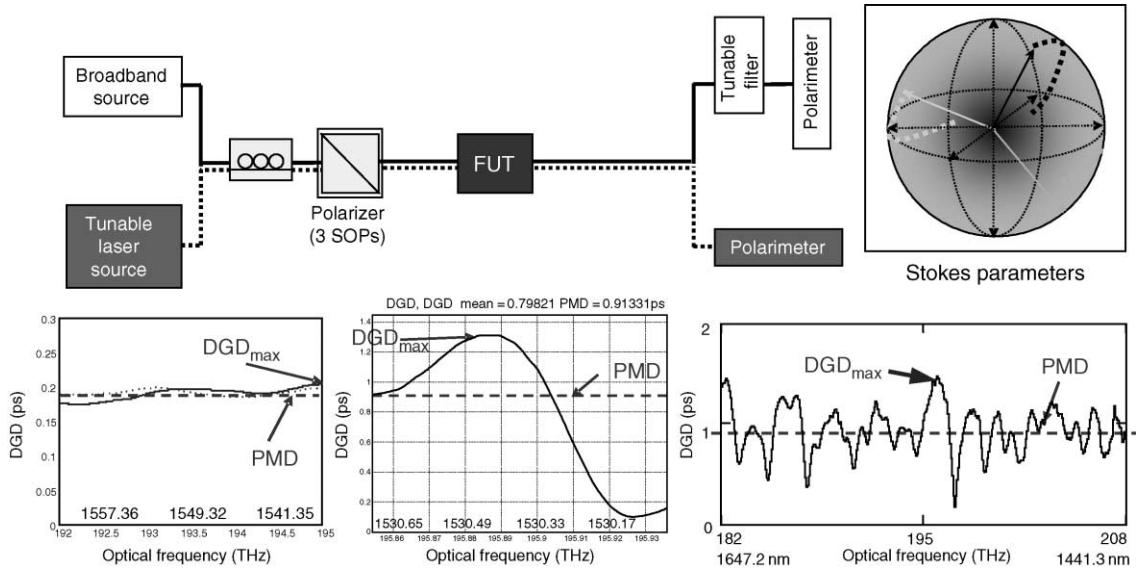


Figure 14 PMD by Stokes parameter evaluation method.

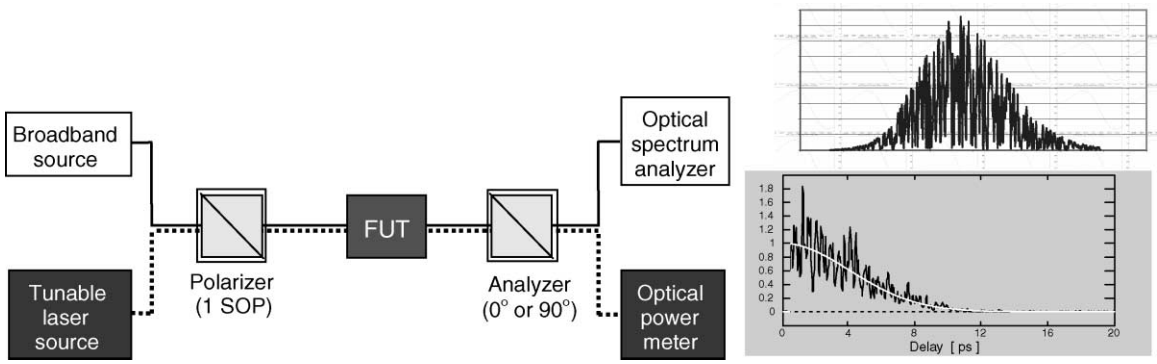


Figure 15 PMD by fixed analyzer method.

pattern interferogram. Two analyses are available to obtain the PMD:

- TINTY uses a set of specific operating conditions for its successful applications and a basic setup; and
- GINTY uses no limiting operating conditions, but in addition to the same basic setup, also using a modified setup compared to TINTY.

Figure 16 illustrates the test setup for both approaches and examples of test results.

Polarization Crosstalk

Polarization crosstalk is a characteristic of energy mixing/transfer/coupling between the two PSPs in a PMF (polarization maintaining fiber) when their isolation is imperfect. It is the measure of the strength of mode coupling or output power ratio between the PSPs within a PMF.

A PMF is an optical fiber capable of transmitting, under external perturbations, such as bending or lateral pressures, both HE_{11}^{x} and HE_{11}^{y} polarization modes whose electric field vector directions are orthogonally to each other and which have different propagation constants β_x and β_y.

Two methods are available for measuring the polarization crosstalk of PMF:

- *power ratio method*, which uses the maximum and minimum values of output power at a specified wavelength, and is applicable to fibers and connectors jointed to a PMF, and to two or more PMFs joined in series; and
- *in-line method*, which uses an analysis of the Poincaré sphere, and is applicable to single or cascaded sections of PMF, and to PMF interconnected with optical devices.

Nonlinear Effects

When the power of the transmission signal is increased to achieve longer span lengths at high bit

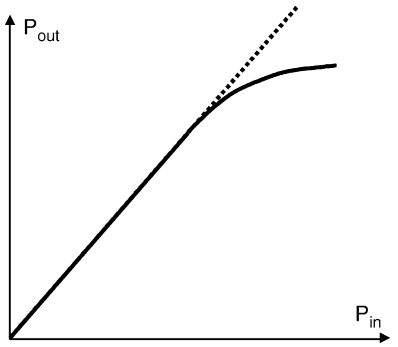


Figure 17 Power output-to-power input relationship for production of nonlinear effects.

expressed as follows:

$$n = n_0 + n_2 I \qquad [18]$$

n is the nonlinearity dependent index; n_0 is the linear part of the index; n_2 is the nonlinear index, also called the Kerr nonlinear index (2.2 to 3.4 $\times 10^{-16}$ cm^2/W); and I is the optical intensity inside the fiber.

The field propagation at a distance L into the fiber is described by the following equation:

$$E_{\text{out}}(L) = E_{\text{in}}(0)\exp[-a/2 + \mathrm{i}\beta + \gamma P(L,t)/2]L \qquad [19]$$

$a/2$ is the attenuation; $\mathrm{i}\beta$ is the phase of the wave; and $\gamma P(L,t)/2$ is the nonlinearity term;

$$\gamma = 2\pi n_2/\lambda A_{\text{eff}} \qquad [20]$$

γ is the nonlinearity coefficient and may be a complex number; A_{eff} is the fiber core effective area; $P(L,t)$ is the total power; and λ is the signal wavelength and t the time variable.

The nonlinear coefficient is defined as n_2/A_{eff}. This coefficient plays a critical role in the fiber and is closely related to system performance degradation due to nonlinearities when very high power is used.

Methods for measuring the nonlinear coefficient. Two methods are available for measuring the nonlinear coefficient:

- Continuous-wave dual-frequency (CWDF); and
- Pulsed single-frequency (PSF).

In the CWDF method, light from two wavelengths is injected into the fiber. At higher power, the light from the two wavelengths beat due to the nonlinearity and produce an output spectrum that is spread. The relationship of the power level to a particular spreading is used to calculate the nonlinear coefficient.

In the PSF method, the pulsed light from a single wavelength is injected into the fiber. Very short pulses (<1 ns) and their input peak power must be measured and related to the nonlinear spreading of the output spectrum.

Stimulated Brillouin scattering

In an intensity modulated system using a source with a narrow linewidth, significant optical power is transferred from the forward-propagating signal to a backward-propagating signal when the SBS power threshold is exceeded. At that point periodic regions of index of refraction produce a grating traveling at speed of sound away from the source. The grating reflects backward part of the incident light. The reflected sound waves (acoustic phonons) scatter light back to the source. Phase matching (or momentum conservation) dictates that the scattered light preferentially travels in the backward direction. The scattered light will be Doppler-shifted (downshifted or Brillouin-shifted) by approximately 11 GHz (at 1550 nm, for G.652 fiber). The scattered light has a very narrow spectrum (very coherent) and very close to the carrier signal and may be very detrimental.

Stimulated Raman scattering

SRS is an interaction between light and the fiber molecular vibrations as adjacent atoms vibrate in opposite directions (an 'optical phonon'). Some of the energy of the main carrier (optical pump wave) is transferred to the molecules, thereby further increasing the amplitude of their vibrations. If the vibrations become large enough, a threshold is reached at which the local index of refraction changes. These local changes then scatter light in all directions similar to Rayleigh scattering. However, unlike Rayleigh scattering, the wavelength of the Raman scattered light is shifted to longer wavelengths by an amount that corresponds to the molecular vibration frequencies and the Raman signal spreads over a large spectrum.

Self-phase modulation

SPM is the effect that a powerful pulse has on its own phase, considering that in eqn [18], $I(t)$ varies in time:

- $I(t) \rightarrow n(t) = n_0 + n_2 I(t) \rightarrow$ modulates the phase $\beta(t)$ of the pulse; and
- $\mathrm{d}I/\mathrm{d}t \rightarrow \mathrm{d}n/\mathrm{d}t \rightarrow \mathrm{d}\beta/\mathrm{d}t(\text{chirp}) \rightarrow$ broadening in the frequency domain $\rightarrow$ broadening in the time domain.

$I(t)$ peaks at the center of the pulse (peak power) and consequently increases the index of refraction. A higher index causes the wavelengths in the center of the pulse to accumulate phase more quickly than at the wings of the pulse:

- this causes wavelength stretching (shift to longer wavelengths) at pulse leading edge (risetime); and
- this causes wavelength compression (shift to shorter wavelengths) at pulse trailing edge (falltime).

The pulse will then broaden with negative (normal) dispersion and shorten with positive (anomalous) dispersion. SPM may then be used for dispersion compensation considering that self-phase modulation imposes $C > 0$ (positive chirp). It can cancel the dispersion if properly manage as a function of the sign of the dispersion. SPM is one of the most critical nonlinear effects for the propagation of soliton or very short pulses over very long distance.

Cross-phase modulation

XPM is the effect that a powerful pulse has on the phase of an adjacent pulse from another WDM system channel traveling in phase or at slightly the same group velocity. It concerns spectral interference between two WDM channels:

- the increasing $I(t)$ at the leading edge of the interfering pulse shifts the other pulse to longer wavelength; and
- decreasing $I(t)$ at the trailing edge of the interfering pulse shifts the other pulse to shorter wavelengths.

This produces spectral broadening, which dispersion converts to temporal broadening depending on the sign of the dispersion. XPM effect is similar to SPM except it depends on the channel count.

Four-wave (four-photon) mixing

FWM is the by-product production effect from two or more WDM channels. For two channels $I(t)$ modulates the phase of each signal (ω_1 and ω_2). An intensity modulation appears at the beat frequency $\omega_1 - \omega_2$.

Two sideband frequencies are created in a similar way as harmonics generation. New wavelengths are created in a number equal to $N^2(N-1)/2$, where $N =$ number of original wavelengths.

Fiber Dimension Characteristics and Corresponding Tests Methods

Table 1 provides a list of the various fiber dimensional characteristics and their corresponding test methods.

Table 1 Fiber dimensional characteristics

Attribute	*Measured parameter*
Fiber geometry	Core/cladding diameter Core/cladding noncircularity Core-cladding concentricity error
Numerical aperture	
Mode field diameter	
Coating geometry	
Length	

Fiber Geometry Characteristics

The fiber geometry is related to the core and cladding characteristics.

Core

The core center is the center of a circle which best fits the points at a constant level in the near-field intensity profile emitted from the central region of the fiber, using wavelengths above and/or below the cut-off wavelength.

The RIP can be measured by refracted near field (RNF) or transverse interferometry techniques and transmitted near field (TNF).

The core concentricity error is the distance between the core center and the cladding center. This definition applies very well for multimode fibers. The distance between the center of the near field profile and the center of the cladding is also used for singlemode fibers.

The mode field diameter (MFD) represents a measure of the transverse electromagnetic field intensity of the mode in a fiber cross-section and it is defined from the far-field intensity distribution.

The MF is the singlemode field distribution of the LP_{01} mode, giving rise to a spatial intensity distribution in the fiber.

The MF concentricity error is the distance between the MF center and the cladding center.

The core noncircularity is a measure of the core ellipticity. This parameter is one of the causes for creating birefringence in the fiber and consequently PMD.

Cladding

The cladding is the outermost region of constant refractive index in the fiber cross-section.

The cladding center is the center of a circle best fitting the outer limit (boundary) of the cladding.

The cladding diameter is the diameter of the circle defining the cladding center.

The cladding noncircularity is a measure of the difference between the diameters of the two circles defined by the cladding tolerance field divided by the nominal cladding diameter.

Coating

The primary coating is one or more layers of protective material applied to the cladding during or after the drawing process to protect the cladding surface (e.g., a 250 μm protective coating). The secondary coating is one or more layers of protective material applied over the primary coating in order to give additional protection or to provide a particular structure.

Measurement of the fiber geometrical attributes
The fiber geometry is measured by the following methods:

- TNF;
- RNF;
- Side-view technique/transverse interference;
- TNF image technique; and
- Mechanical diameter.

Test instrumentation may incorporate two or more methods such as the one shown in Figure 18.

Transmitted near-field technique. The cladding diameter, core concentricity error, and cladding noncircularity are determined from the near-field intensity distribution. Figure 19 provides a series of examples of test results from TNF measurements.

Refracted near-field technique. The RIP across the entire fiber (core and cladding) can be directly obtained from the RNF measurement, as shown in Figure 20.

The geometrical characteristics of the fiber can be obtained from the refractive index distribution using suitable algorithms:

- core/cladding diameter;
- core/cladding concentricity error;
- core/cladding noncircularity;
- maximum numerical aperture (NA); and
- index and relative index of refraction difference.

Figure 21 illustrates the core geometry.

Side-view technique/transverse interference. The side-view method is applied to singlemode fibers to determine the core concentricity error, cladding diameter and cladding noncircularity by measuring the intensity distribution of light that is refracted inside the fiber. The method is based on an interference microscope focused on the side view of an FUT illuminated perpendicular to the FUT axis. The fringe pattern is used to determine the RIP.

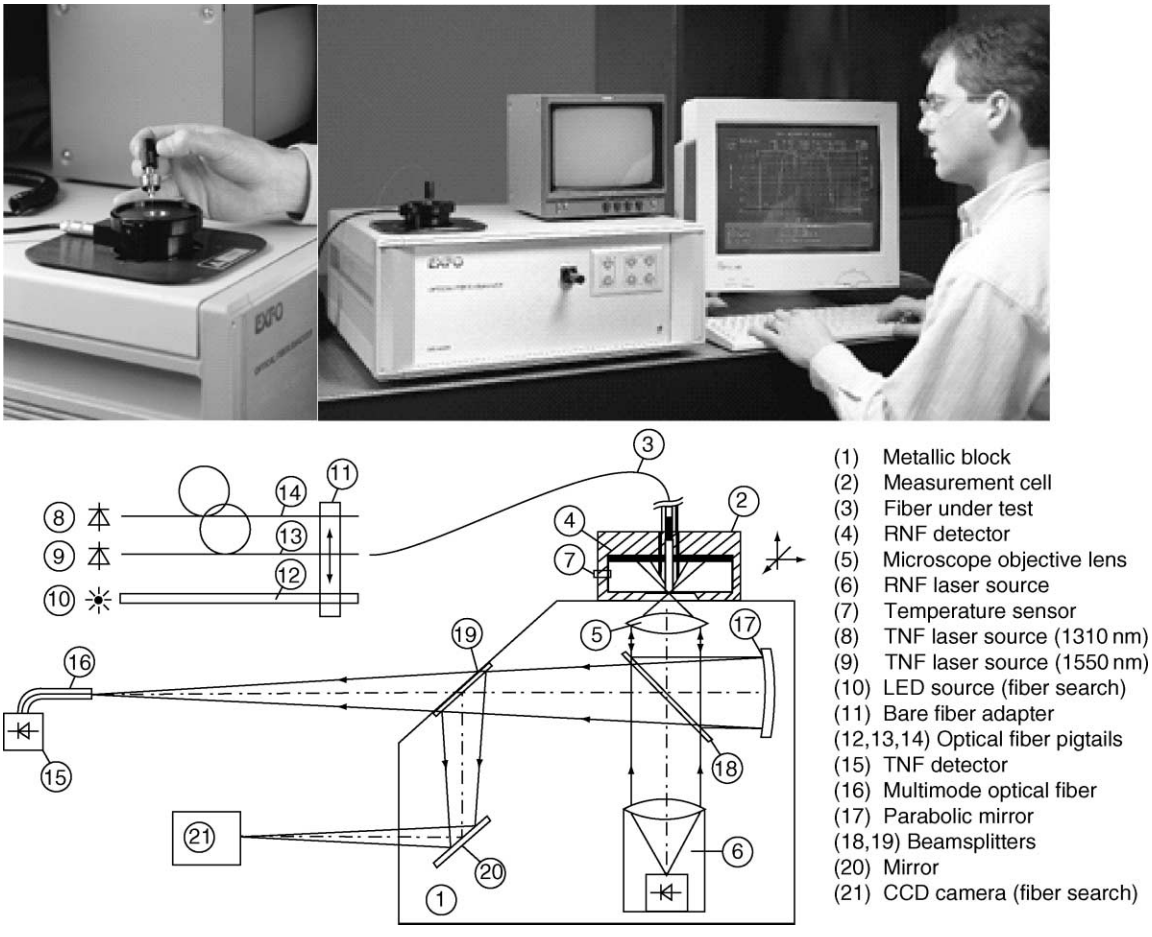


Figure 18 RNF/TNF combined instrumentation.

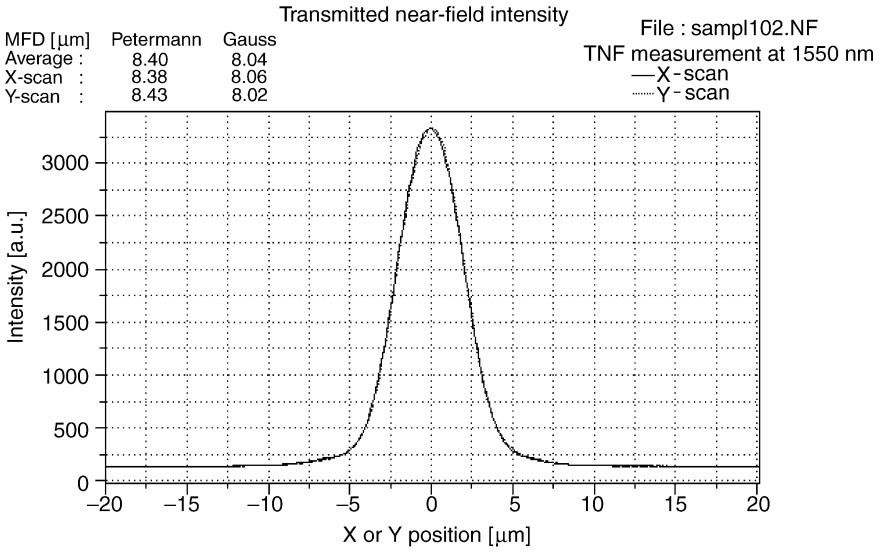


Figure 19 MFD measurement by TNF.

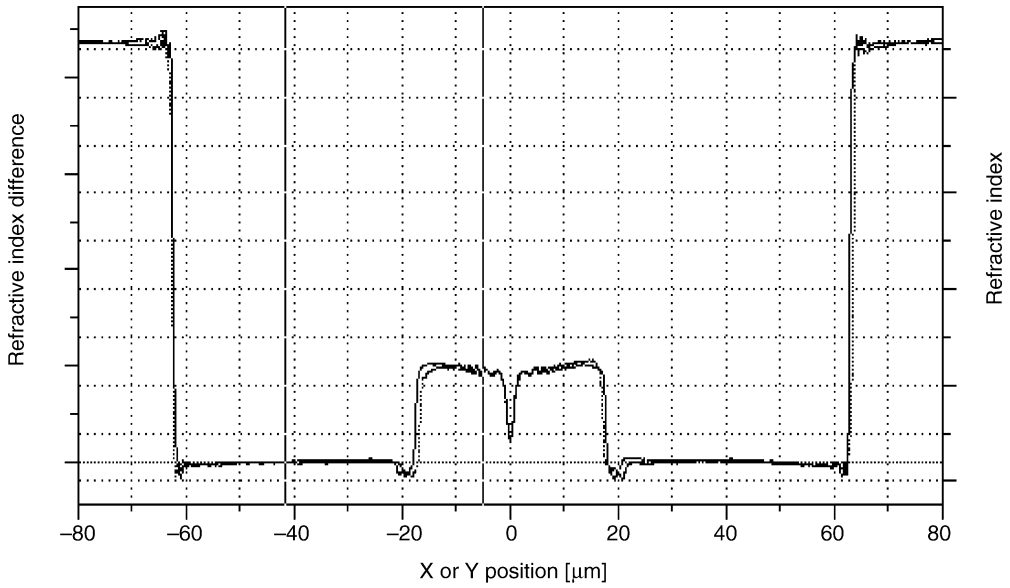


Figure 20 RIP from RNF measurement.

TNF image technique. The TNF image technique, also called near-field light distribution, is used for the measurement of the geometrical characteristics of singlemode fibers. The measurement is based on analysis of magnified images at the FUT output. Two subsets of the method are available:

- grey-scale technique which performs an x–y near-field scan using a video system; and
- Single near-field scanning technique performing a one-dimensional scan.

Mechanical diameter. This is a precision mechanical diameter measurement technique used to accurately determine the cladding diameter of silica fibers. The technique uses an electronic micrometer such as based on a double-pass Michelson interferometer. The technique is used for providing calibrated fibers to the industry as SRM.

Numerical Aperture

The NA is an important attribute for multimode fibers in order to predict their launching efficiency, joint loss at splices and micro/macrobending characteristics.

A method is available for the measurement of the angular radiant intensity distribution (far-field)

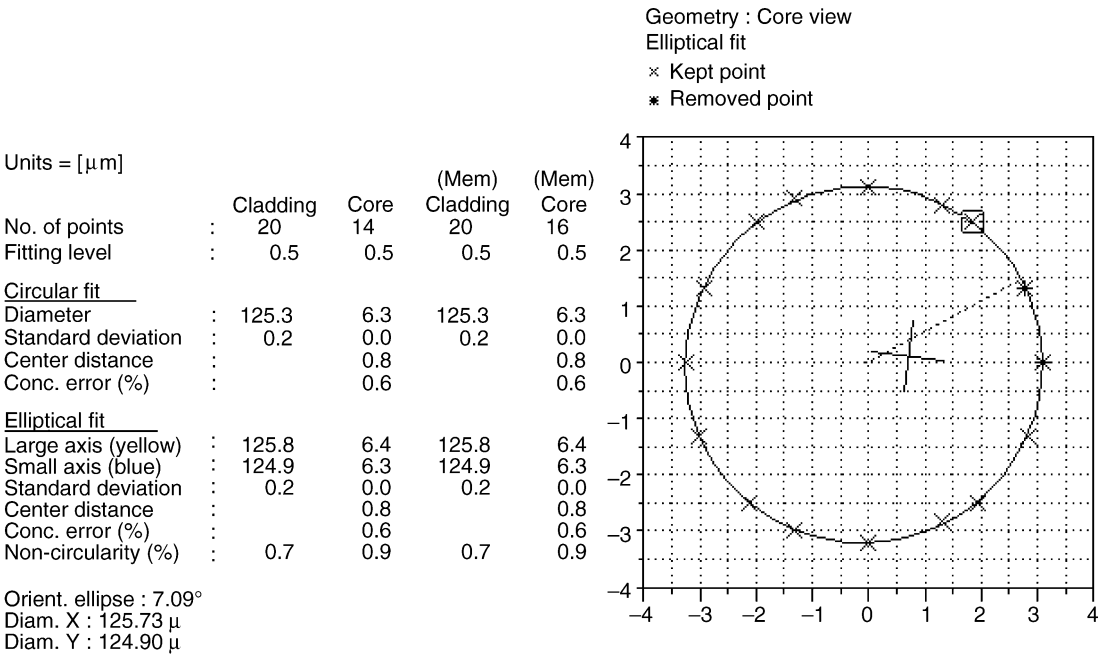


Figure 21 Core geometry by RNF measurement.

distribution or the RIP at the output of an FUT. NA can be determined from the analysis of the test results.

Mode Field Diameter

The definition of the MFD is given in the section describing the core center above. Four measurement methods are available:

- direct far-field scan determines MFD from the far-field intensity distribution;
- variable aperture in far field determines MFD from the complementary aperture transmission function $a(x)$, $x = d \tan \theta$ being the aperture radius and d the distance between the aperture and the FUT:

$$\mathrm{MFD} = \frac{\lambda}{\pi d}\left[\int_0^\infty a(x)\frac{x}{(x^2+d^2)^2}\,\mathrm{d}x\right]^{-1/2} \qquad [21]$$

Equation [21] is valid for small-θ approximation.

- near-field scan determines MFD from the near-field intensity distribution I_{NF}, r being the radial coordinate:

$$\mathrm{MFD} = 2\left[2\frac{\int_0^\infty rI_{\mathrm{NF}}(r)\,\mathrm{d}r}{\int_0^\infty r\left[\dfrac{\mathrm{d}I^{1/2}(r)}{\mathrm{d}r}\right]^2\mathrm{d}r}\right]^{1/2} \qquad [22]$$

Equation [22] is valid for small-θ approximation; and

- bidirectional backscattering uses an OTDR and bidirectional measurements to determine MFD by comparing the FUT results with a reference fiber.

Effective Area

A_{eff} is a critical nonlinearity parameter and is defined as follows:

$$A_{\mathrm{eff}} = \frac{2\pi[\int_0^\infty I(r)r\,\mathrm{d}r\,]^2}{\int_0^\infty I(r)^2 r\,\mathrm{d}r} \qquad [23]$$

$I(r)$ is the field intensity distribution of the fiber fundamental mode at radius r. The integration in the equation is carried out over the entire fiber cross-section. For a Gaussian approximation:

$$I(r) = \exp - 2\left(\frac{2r}{\mathrm{MFD}}\right)^2 \qquad [24]$$

which yields:

$$A_{\mathrm{eff}} = \frac{\pi}{4}\mathrm{MFD}^2 \qquad [25]$$

Three methods are available for the measurement of A_{eff}:

- direct far-field;
- variable aperture in far-field; and
- near-field.

Mechanical Measurement and Test Methods

Table 2 describes the various mechanical characteristics and their corresponding test methods.

Proof Stressing

The proof stress level is the value of tensile stress or strain applied to a full fiber length over a short period of time. The method for fiber proof stressing is the longitudinal tension which describes procedures for applying tensile loads to a length of fiber. The fiber stress is calculated from the applied tension. The tensile load is applied over short periods of time but not too short in order for the fiber to experience proof stress.

Residual Stress

Residual stress is the stress built up by the thermal expansion difference between core and cladding during the fiber drawing process or splicing. Methods are available for measuring residual stress based on polarization effects. A light beam produced by a rotating polarizer propagates to the x-axis direction while the beam polarization is in the $y-z$ plane and the fiber longitudinal axis in the z-axis. The light experiences different phase shift in the y- and z-axis due to the FUT birefringence. The photoelastic effect gives the relationship between this phase change and residual stresses.

Stress Corrosion Susceptibility

The stress corrosion susceptibility is related to the dependence of crack growth on applied stress. It depends on the environmental conditions and static and dynamic values may be observed.

Environmental Characteristics

Table 3 lists the fiber characteristics related to the effect of the environment.

Table 2 Fiber mechanical characteristics

Attribute
Proof stress
Residual stress
Stress corrosion susceptibility
Tensile strength
Stripability
Fiber curl

Table 3 Fiber environmental characteristics

Attribute
Hydrogen aging
Nuclear gamma irradiation
Damp heat
Dry heat
Temperature cycling
Water immersion

Hydrogen Aging for Low-Water-Peak Single-Mode Fiber

Hydrogen aging on low-water peak fibers, such as G.652.C, is based on a test performed at 1.0 atmosphere of hydrogen pressure at room temperature over a period of one month. Other proportional combinations are possible.

Nuclear Gamma Irradiation

Nuclear radiation is considered on the basis of a steady state response of optical fibers and cables exposed to gamma radiation and to determine the level of related radiation-induced attenuation produced in singlemode or multimode cabled or uncabled fibers.

The fiber attenuation generally increases when exposed to gamma radiation. This is primarily due to the trapping of radiolytic electrons and holes at defect sites in the glass (i.e., the formation of 'color centers'). Two regimes are considered:

- the low dose rate suitable for estimating the effect of environmental background radiation; and
- the high dose rate suitable for estimating the effect of adverse nuclear environments.

The effects of environmental background radiation are measured by the attenuation (cut-back method). The effects of adverse nuclear environments are tested by power monitoring before, during and after FUT exposure.

See also

Fiber and Guided Wave Optics: Nonlinear Effects (Basics). **Imaging:** Interferometric Imaging. **Interferometry:** Gravity Wave Detection; Overview. **Polarization:** Introduction. **Scattering:** Scattering Phenomena in Optical Fibers.

Further Reading

Agrawal GP (1997) *Fiber-Optic Communication Systems*, 2nd edn. New York: John Wiley & Sons.

Agrawal GP (2001) *Nonlinear Fiber Optics*, 3rd edn. San Diego, CA: Academic Press.
Girard A (2000) *Guide to WDM Technology and Testing*. Quebec: EXFO, pp. 194.
Hecht J (1999) *Understanding Fiber Optics*, 3rd edn. Upper Saddle River, NJ: Prentice Hall.
Masson B and Girard A (2004) *FTTx PON Guide, Testing Passive Optical Networks*. Quebec: EXFO, pp. 56.
Miller JL and Friedman E (2003) *Optical Communications Rules of Thumb*. New York: McGraw-Hill, pp. 428.
Neumann E-G (1988) *Single-Mode Fibers, Fundamentals*. Berlin: Springer-Verlag.

Nonlinear Effects (Basics)

G Millot and P Tchofo-Dinda, Université de Bourgogne, Dijon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Many physical systems in various areas such as condensed matter or plasma physics, biological sciences, or optics, give rise to localized large-amplitude excitations having a relatively long lifetime. Such excitations lead to a host of phenomena referred to as nonlinear phenomena. Of the many disciplines of physics, the optics field is probably the one in which practical applications of nonlinear phenomena have been the most fruitful, in particular, since the discovery of the laser in 1960. This discovery has thus led to the advent of a new branch in optics, referred to as *nonlinear optics*. The applications of nonlinear phenomena in optics include the design of various kinds of laser sources, optical amplifiers, light converters, light-wave communication systems for data transmission purposes, to name a few.

In this article, we present an overview of some basic principles of nonlinear phenomena that result from the interaction of light waves with dielectric waveguides such as optical fibers. These nonlinear phenomena can be broadly divided into two main categories, namely, parametric effects and scattering phenomena. Parametric interactions arise whenever the state of the dielectric matter is left unchanged by the interaction, whereas scattering phenomena imply transitions between energy levels in the medium. More fundamentally, parametric interactions originate from the electron motion under the electric field of a light wave, whereas scattering phenomena originate from the motion of heavy ions (or molecules).

Linear and Nonlinear Signatures

The macroscopic properties of a physical system can be obtained by analyzing the response of the system under an external excitation. For example, consider at time t the response of a system, such as an amplifier, to an input signal $E_1 = A\sin(\omega t)$. In the low-amplitude limit of the output signal, the response R_1 of the system is proportional to the excitation:

$$R_1 = \alpha_1 E_1 \qquad [1]$$

where α_1 is a constant. This type of behavior corresponds to the so-called linear response. In general, a physical system executes a linear response when the superposition of two (or more) input signals E_1 and E_2 yields a response which is a superposition of the output signals, as schematically represented in Figure 1:

$$R = \alpha_1 R_1 + \alpha_2 R_2 \qquad [2]$$

Now, in almost all real physical systems, if the amplitude of an excitation E_1 becomes sufficiently large, distortions will occur in the output signals. In other words, the response of the system will no longer be proportional to the excitation, and consequently, the law of superposition of states will no longer be observed. In this case, the response of the system may take the following form:

$$R = \alpha_1 E_1 + \alpha_2 E_1^2 + \alpha_3 E_1^3 + \ldots \qquad [3]$$

which involves not only a signal at the input frequency ω, but also signals at frequencies 2ω, 3ω, and so on. Thus, harmonics of the input signal are generated. This behavior, called *nonlinear response*, is at the origin of a host of important phenomena in many

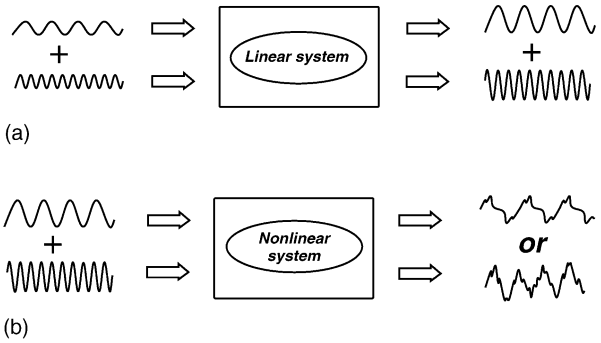


Figure 1 Schematic representation of linear and nonlinear responses of a system, to two input signals.

branches of sciences, such as in condensed matter physics, biological sciences, or in optics.

Physical Origin of Optical Nonlinearity

Optical nonlinearity originates fundamentally from the action of the electric field of a light wave on the charged particles of a dielectric waveguide. In contrast to conductors where charges can move throughout the material, dielectric media consist of *bound* charges (ions, electrons) that can execute only relatively limited displacements around their equilibrium positions. The electric field of an incoming light wave will cause the positive charges to move in the polarization direction of the electric field whereas negative charges will move in the opposite direction. In other words, the electric field will cause the charged particles to become dipoles, as schematically represented in Figure 2.

Then each dipole will vibrate under the influence of the incoming light, thus becoming a source of radiation. The global light radiated by all the dipoles represents the scattered light. When the charge displacements are proportional to the excitation (incoming light), i.e., in the low-amplitude limit of scattered radiation, the output light vibrates at the same frequency as that of the excitation. This process corresponds to Rayleigh scattering. On the other hand, if the intensity of the excitation is sufficiently large to induce displacements that are not negligible with respect to atomic distances, then the charge displacements will no longer be proportional to the excitation. In other words, the response of the medium, which is no longer proportional to the excitation, becomes nonlinear. In this case, the scattered waves will be generated not only at the excitation frequency (the Kerr effect), say ω, but also at frequencies that differ from ω (e.g., 2ω, 3ω). On the other hand, it is worth noting that when all induced dipoles vibrate coherently (that is, their relative phase does not vary randomly), their individual radiation may, under certain conditions, interfere constructively and lead to a global field of high intensity.

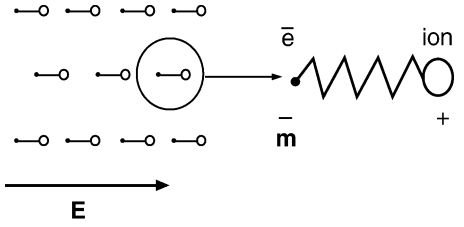


Figure 2 Electric dipoles in a dielectric medium under an external electric field.

The condition of constructive interference is the phase-matching condition.

In practice, the macroscopic response of a dielectric is given by the polarization, which corresponds to the total amount of dipole moment per unit volume of the dielectric. As the mass of an ion is much larger than that of an electron, the amplitude of the ion motion is generally negligible with respect to that of the electrons. As a consequence, the electron motion generally leads to the dominant contribution in the macroscopic properties of the medium. The behavior of an electron under an optical electric field is similar to that of a particle embedded in an anharmonic potential. A very simple model (called the Lorentz model) that provides a deep insight into the dielectric response consists of an electron of mass m and charge $-e$ connected to an ion by an elastic spring (see Figure 2). Under the electric field $E(t)$, the electron executes a displacement $x(t)$ with respect to its equilibrium position, which is governed by the following equation:

$$\frac{d^2x}{dt^2} + 2\lambda\frac{dx}{dt} + \omega_0^2 x + (a^{(2)}x^2 + a^{(3)}x^3 + \cdots) = -\frac{e}{m}E(t) \qquad [4]$$

where $a^{(2)}$, $a^{(3)}$, and so on, are constant parameters, ω_0 is the resonance angular frequency of the electron, and λ is the damping coefficient resulting from the dipolar radiation. When the amplitude of the electric field is sufficiently large, then the restoring force on the electrons becomes a nonlinear function of x; hence the presence of terms such as $a^{(2)}x^2$, $a^{(3)}x^3$, and so on, in eqn [4]. In this situation, the macroscopic response of the dielectric is the polarization

$$P = -\sum ex(\omega, 2\omega, 3\omega, \ldots) \qquad [5]$$

where $x(\omega, 2\omega, 3\omega, \ldots)$ is the solution of eqn [4] in the frequency domain, and the summation extends over all the dipole moments per unit volume. In terms of the electric field E the polarization may be written as

$$P = \varepsilon_0(\chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \cdots) \qquad [6]$$

where $\chi^{(1)}$, $\chi^{(2)}$, $\chi^{(3)}$, and so on, represent the susceptibility coefficients. Figure 3 (top left) illustrates schematically the polarization as a function of the electric field.

In particular, one can clearly observe that when the amplitude of the incoming electric field is sufficiently small (bottom left in Figure 3), the polarization (top right) is proportional to the electric field (bottom left), thus implying that the electric dipole radiates a wave having the same frequency as that of the incoming light (bottom right). On the other hand, Figure 4

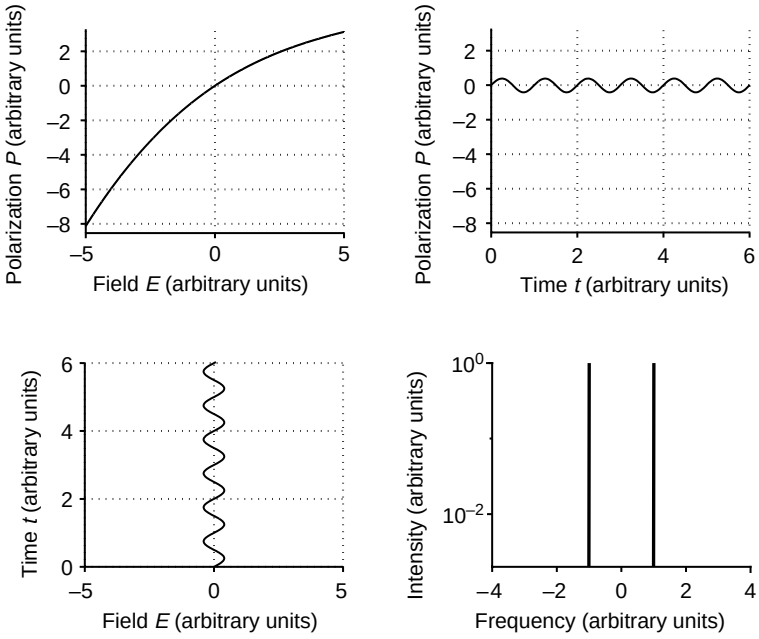


Figure 3 Polarization induced by an electric field of small amplitude. Nonlinear dependence of the polarization as a function of field amplitude (top left) and time dependence of the input electric field (bottom left). Time dependence of the induced polarization (top right) and corresponding intensity spectrum (bottom right).

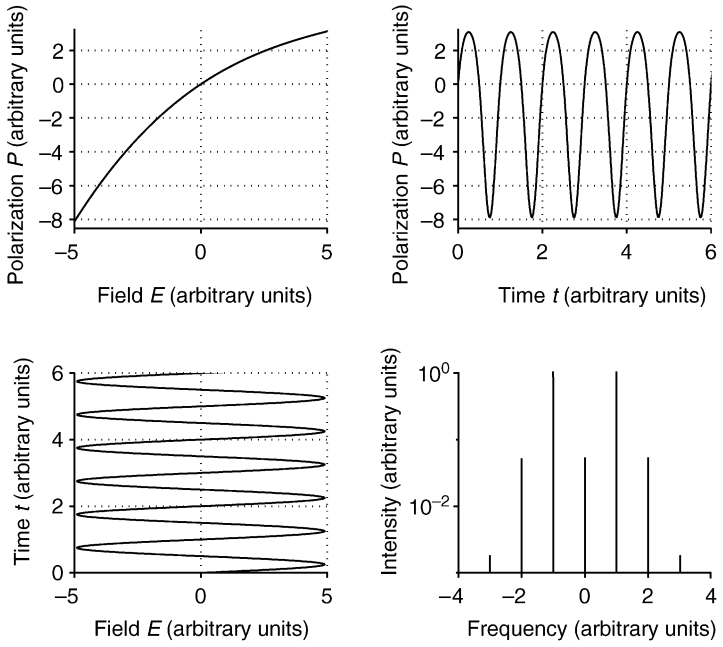


Figure 4 Polarization induced by an incoming electric field of large amplitude.

shows that for an electric field of large amplitude, the polarization is no longer proportional to the electric field, leading to radiation at harmonic frequencies (see Figure 4, bottom right).

This nonlinear behavior leads to a host of important phenomena in optical fibers, which are useful for many optical systems but detrimental for other systems.

Parametric Phenomena in Optical Fibers

In anisotropic materials, the leading nonlinear term in the polarization, i.e., the $\chi^{(2)}$ term, leads to phenomena such as the harmonic generation or optical rectification. This $\chi^{(2)}$ term vanishes in homogeneous isotropic materials such as cylindrical

optical fibers, and there the leading nonlinear term becomes the $\chi^{(3)}$ term. Thus, most of outstanding nonlinear phenomena in optical fibers originate from the third-order nonlinear susceptibility $\chi^{(3)}$. Some of those phenomena are described below.

Optical Kerr Effect

The optical Kerr effect is probably the most important nonlinear effect in optical fibers. This effect induces an intensity dependence of the refractive index, which leads to a vast wealth of fascinating phenomena such as self-phase modulation (SPM), cross-phase modulation (CPM), four-wave mixing (FWM), modulational instability (MI) or optical solitons. The Kerr effect can be conveniently described in the frequency domain through a direct analysis of the polarization, which takes the following form:

$$P_{\mathrm{NL}}(\omega) = \frac{3}{4}\varepsilon_0\chi^{(3)}(\omega)|E(\omega)|^2E(\omega) \qquad [7]$$

The constant 3/4 comes from the symmetry properties of the tensor $\chi^{(3)}$. Setting $P_{\mathrm{NL}}(\omega) = \varepsilon_0\varepsilon_{\mathrm{NL}}E(\omega)$, where $\varepsilon_{\mathrm{NL}} = \frac{3}{4}\chi^{(3)}|E|^2$ is the nonlinear contribution to the dielectric constant, the total polarization takes the form

$$P(\omega) = P_{\mathrm{L}} + P_{\mathrm{NL}} = \varepsilon_0[\chi^{(1)}(\omega) + \varepsilon_{\mathrm{NL}}]E(\omega) \qquad [8]$$

As eqn [8] shows, the refractive index n, at a given frequency ω, is given by

$$n^2 = 1 + \chi^{(1)} + \varepsilon_{\mathrm{NL}} = (n_0 + \Delta n_{\mathrm{NL}})^2 \qquad [9]$$

with $n_0^2 = 1 + \chi^{(1)}$. In practice $\Delta n_{\mathrm{NL}} \ll n_0$, and then, the refractive index is given by

$$n(\omega, |E|^2) = n_0(\omega) + n_2^e|E|^2 \qquad [10]$$

where n_2^e is the nonlinear refractive index defined by $n_2^e = 3\chi^{(3)}/(8n_0)$. The linear polarization P_{L} is responsible for the frequency dependence of the refractive index, whereas the nonlinear polarization P_{NL} causes an intensity dependence of the refractive index, which is referred to as the optical Kerr effect. Knowing that the wave intensity I is given by $I = \alpha|E|^2$, with $\alpha = \frac{1}{2}\varepsilon_0 c n_0$, the refractive index can be then rewritten as

$$n(\omega, I) = n_0(\omega) + n_2 I \qquad [11]$$

with $n_2 = n_2^e/\alpha = 2n_2^e/(\varepsilon_0 c n_0)$. For fused silica fibers one has typically: $n_2 = 2.66 \times 10^{-20}\ \mathrm{m^2\ W^{-1}}$. For example, an intensity of $I = 1\ \mathrm{GW\ cm^{-2}}$ leads to $\Delta n_{\mathrm{NL}} = 2.66 \times 10^{-7}$, which is much smaller than $n_0 \approx 1.45$.

Four-Wave Mixing

The four-wave mixing (FWM) process is a third-order nonlinear effect in which four waves interact through an energy exchange process. Let us consider two intense waves, $E_1(\omega_1)$ and $E_2(\omega_2)$, with $\omega_2 > \omega_1$, called pump waves, propagating in an optical fiber. Hereafter we consider the simplest case when waves propagate with the same polarization. In this situation the total electric field is given by

$$\begin{aligned} E_{\mathrm{tot}}(\mathbf{r}, t) &= E_1 + E_2 \\ &= A_1(\omega_1)\exp[\mathrm{i}(\mathbf{k}_1\cdot\mathbf{r} - \omega_1 t)] \\ &\quad + A_2(\omega_2)\exp[\mathrm{i}(\mathbf{k}_2\cdot\mathbf{r} - \omega_2 t)] \end{aligned} \qquad [12]$$

where $\mathbf{k}_1$ and $\mathbf{k}_2$ are the wavevectors of the fields E_1 and E_2, respectively. Equation [7], which gives the nonlinear polarization induced by a single monochromatic wave, remains valid provided that the frequency spacing between the two waves is relatively small, i.e., $|\Delta\omega| = |\omega_2 - \omega_1| \ll \omega_0 = (\omega_1 + \omega_2)/2$. In this context, eqn [7] leads to

$$P_{\mathrm{NL}} \approx \frac{3}{4}\varepsilon_0\chi^{(3)}(\omega_0)|E_{\mathrm{tot}}|^2E_{\mathrm{tot}} \qquad [13]$$

Substituting eqn [12] in eqn [13] yields

$$\begin{aligned} P_{\mathrm{NL}} = 2n_0n_2\varepsilon_0\Big[&\left(|E_1|^2 + 2|E_2|^2\right)E_1 \\ &+ \left(|E_2|^2 + 2|E_1|^2\right)E_2 + E_1^2E_2^* + E_1^*E_2^2\Big] \end{aligned} \qquad [14]$$

The term $|E_1|^2E_1$ in the right-hand side of eqn [14] represents the self-induced Kerr effect on the wave ω_1. The term $|E_2|^2E_1$, which corresponds to a modification of the refractive index seen by the wave ω_1, due to the presence of the wave ω_2, represents the cross-Kerr effect. Thus, the refractive index at frequency ω_1 depends simultaneously on the intensities of the two pumps, i.e., $n = n(\omega_1, |E_1|^2, |E_2|^2)$. Similarly the third and fourth terms in the right-hand side of eqn [14] illustrate the self-induced and cross-Kerr effects on the wave ω_2, respectively. Note that the self-induced Kerr effect is responsible for a self-phase modulation, which induces a spectral broadening of a pulse propagating through the optical fiber, whereas the cross-Kerr effect leads to a cross-phase modulation that induces a spectral broadening of one wave in the presence of a second wave. The two last terms in the right-hand side of eqn [14] correspond to the generation of new waves at frequencies $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$, respectively. The wave with the smallest frequency $2\omega_1 - \omega_2 = \omega_{\mathrm{s}}$ is called the Stokes wave, whereas the wave with the highest frequency $2\omega_2 - \omega_1 = \omega_{\mathrm{as}}$ is called the anti-Stokes wave. These two newly generated waves interact

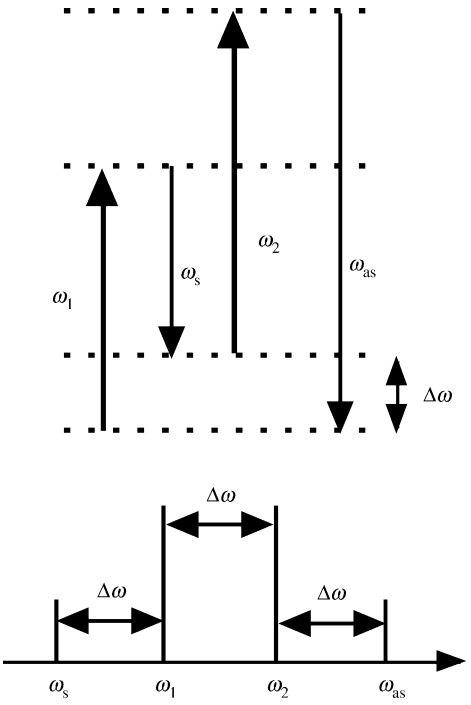


Figure 5 Schematic diagram representing a non-degenerate four-wave mixing process (top) and the corresponding frequency spectrum (bottom).

with the two pumps through an energy exchange process. This interaction is commonly referred to as a four-wave mixing process.

From a quantum mechanical point of view, four-wave mixing corresponds to a process in which two photons with frequencies ω_1 and ω_2 are annihilated with simultaneous creation of two photons at frequencies ω_s and ω_{as}, respectively. The new waves are generated at frequencies ω_s and ω_{as} such that

$$\omega_1 + \omega_2 = \omega_s + \omega_{as} \qquad [15]$$

which states that the total energy is conserved during the interaction. But the condition for this FWM process to occur is that the total momentum is conserved, that is,

$$\Delta\mathbf{k} = \mathbf{k}_s + \mathbf{k}_{as} - \mathbf{k}_1 - \mathbf{k}_2 = 0 \qquad [16]$$

or equivalently,

$$n(\omega_s)\omega_s + n(\omega_{as})\omega_{as} - n(\omega_1)\omega_1 - n(\omega_2)\omega_2 = 0 \quad [17]$$

Equation [16] is known as the phase-matching condition. Figure 5 displays a schematic diagram of a four-wave mixing process with the corresponding frequency spectrum.

FWM with different pump frequencies $\omega_1 \neq \omega_2$ is called 'nondegenerate FWM', whereas the case $\omega_1 = \omega_2$ is referred to as 'degenerate FWM', or more simply, as three-wave mixing.

Conclusion

Optical fibers constitute a key device for many areas of optical sciences, and in particular for ultrafast optical communications. The nonlinear phenomena that arise through the nonlinear refractive index change (induced mainly by the Kerr effect) have been largely investigated these last two decades for applications such as all-optical wavelength conversion, parametric amplification, generation of new optical frequencies or ultrahigh repetition rate pulse trains. However, in many other optical systems such as optical communication systems, these nonlinear phenomena become harmful effects, and in this case control processes are being developed in order to suppress or at least reduce their impact in the system.

List of Units and Nomenclature

Nonlinear refractive index n_2 [$m^2\ W^{-1}$]
Optical intensity I [$GW\ cm^{-2} = 10^9\ W\ cm^{-2}$]

CPM: cross phase modulation
FWM: four wave mixing
MI: modulational instability
SPM: self-phase modulation

See also

Fiber and Guided Wave Optics: Light Propagation; Measuring Fiber Characteristics. **Lasers:** Optical Fiber lasers. **Optical Amplifiers:** Semiconductor Optical Amplifiers. **Optical Communication Systems:** Wavelength Division Multiplexing. **Scattering:** Scattering Theory. **Solitons:** Soliton Communication Systems.

Further Reading

Agrawal GP (2001) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Bloembergen N (1977) *Nonlinear Optics*. Reading, MA: Benjamin.

Boyd RW (1992) *Nonlinear Optics*. San Diego, CA: Academic Press.

Butcher PN and Cotter DN (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Dianov EM, Mamyshev PV, Prokhorov AM and Serkin VN (1989) *Nonlinear Effects in Optical Fibres*. Switzerland: Harwood Academic Publishers.

Hellwarth R (1977) Third-order optical susceptibilities of liquids and s\olids. *Progress in Quantum Electronics* 5: 1–68.

Pocholle JP, Papuchon M, Raffy J and Desurvire E (1990) Non linearities and optical amplification in single mode fibres. *Revue Technique Thomson-CSF* 22: 187–268.

Shen YR (1984) *Principles of Nonlinear Optics*. New York: Wiley.

Nonlinear Optics

K Thyagarajan and A K Ghatak, Indian Institute of Technology, New Delhi, India

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The invention of the laser provided us with a light source capable of generating extremely large optical power densities (several MW/m^2). At such large power densities, matter behaves in a nonlinear fashion and we come across new optical phenomena such as second-harmonic generation (SHG), sum and difference frequency generation, intensity dependent refractive index, mixing of various frequencies, etc. In SHG, an incident light beam at frequency ω interacts with the medium and generates a new light wave at frequency 2ω. In sum and difference frequency generation, two incident beams at frequencies ω_1 and ω_2 mix with each other producing sum $(\omega_1 + \omega_2)$ and difference $(\omega_1 - \omega_2)$ frequencies at the output. Higher-order nonlinear effects, such as self-phase modulation, four-wave mixing, etc. can also be routinely observed today. The field of nonlinear optics dealing with such nonlinear interactions is gaining importance due to numerous demonstrated applications in many diverse areas such as optical fiber communications, all-optical signal processing, realization of novel sources of optical radiation, etc.

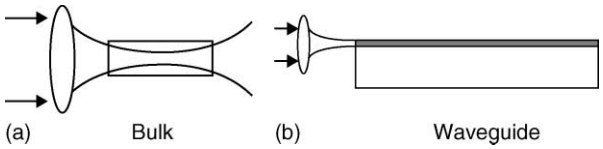


Figure 1 (a) In bulk media, tighter focusing produces larger intensities, but over shorter interaction lengths. (b) In optical waveguides diffraction effects are balanced by waveguiding and the interaction lengths are much larger.

Nonlinear optical interactions become prominent when the optical power densities are high and interaction takes place over long lengths. The usual method to increase optical intensity is to focus the light beam using a lens system. For a given optical power, the tighter the focusing, the larger will be the intensity for a given optical power; however, greater will be the divergence of the beam. Thus tighter focusing produces larger intensities, but over shorter interaction lengths (see Figure 1a). Optical waveguides, in which the light beam is confined to a small cross-sectional area, are currently being explored for realizing efficient nonlinear devices. In contrast to bulk media, in waveguides, diffraction effects are balanced by waveguiding and the beam can have small cross-sectional areas over much longer interaction lengths (see Figure 1b). A simple optical waveguide consists of a high index dielectric medium surrounded by a lower index dielectric medium so that light waves can be trapped in the high index region by the phenomenon of total internal reflection. Figure 2 shows a planar waveguide, a channel waveguide and an optical fiber. In the planar waveguide a film of refractive index n_f is deposited/diffused on a substrate of refractive index n_s and has a cover of refractive index n_c (with n_s, $n_c < n_f$). The waveguide has typical cross-sectional dimensions of a few micrometers. Unlike planar waveguides, in which guidance takes place only in one dimension, in channel waveguides the diffused region in a substrate is surrounded on all sides by lower index media. Optical fibers are structures with cylindrical symmetry and have a central cylindrical core of doped silica surrounded by a concentric cylindrical cladding of pure silica which has a slightly lower refractive index. Light guidance in all these waveguides takes place through the phenomenon of total internal reflection.

In contrast to bulk interactions requiring beam focusing, in the case of optical waveguides, the beam can be made to have extremely small cross-sectional

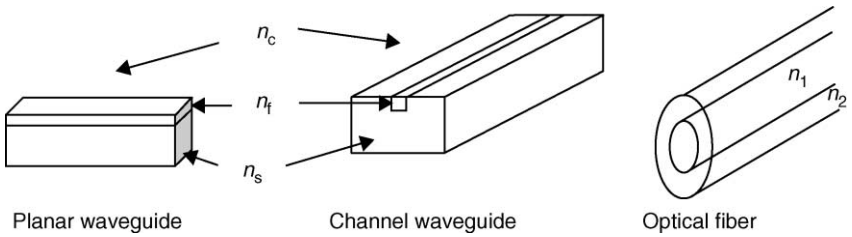


Figure 2 A planar waveguide, a channel waveguide and an optical fiber.

areas ($\sim$25 μm^2) over very long interaction lengths ($\sim$20 mm in integrated optical waveguides and tens of thousands of kilometers in the case of optical fibers). This leads to very much increased nonlinear interaction efficiencies even at moderate powers (approximately a few tens of mW).

In the following sections, we will discuss some of the important nonlinear interactions that are being studied with potential applications to various branches of science and engineering. Obviously it is impossible to cover all aspects of nonlinear optics – several books have been written in this area (see Further Reading section at the end of this article) – what we will do is to discuss the physics of some of the important nonlinear effects.

Apart from intensity and length of interaction, one of the most important requirements for many efficient nonlinear interactions is the requirement of phase matching. The concept of phase matching can be easily understood from the point of view of SHG. In SHG, the incident wave at frequency ω generates a nonlinear polarization at frequency 2ω and this nonlinear polarization is responsible for the generation of the wave at 2ω. Now, the phase velocity of the nonlinear polarization wave at 2ω is the same as the phase velocity of the electromagnetic wave at frequency ω, which is usually different from the phase velocity of the electromagnetic wave at frequency 2ω; this happens due to wavelength dispersion in the medium. When the two phase velocities are unequal, the polarization wave at 2ω (which is the source) and the electromagnetic wave at 2ω pass in and out of phase with each other as they propagate through the medium. Due to this, the energy flowing in from ω to 2ω cannot add constructively and the efficiency of second-harmonic generation is limited. If the phase velocities of the waves at ω and 2ω are matched then the polarization wave and the wave at 2ω remain in phase leading to drastically increased efficiencies. This condition is referred to as phase matching and plays a very important role in most nonlinear interactions.

Nonlinear Polarization

In a linear medium, the electric polarization P is assumed to be a linear function of the electric field E:

$$P = \varepsilon_0 \chi E \qquad [1]$$

where, for simplicity, a scalar relation has been written. The quantity χ is termed as linear dielectric susceptibility. At high optical intensities (which corresponds to high electric fields), all media behave in a nonlinear fashion. Thus eqn [1] is modified to

$$P = \varepsilon_0(\chi E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \ldots) \qquad [2]$$

where $\chi^{(2)}, \chi^{(3)}, \ldots$ are higher order susceptibilities giving rise to the nonlinear terms. The second term on the right-hand side is responsible for SHG, sum and difference frequency generation, parametric interactions, etc. while the third term is responsible for third-harmonic generation, intensity dependent refractive index, self-phase modulation, four-wave mixing, etc. For media possessing an inversion symmetry, $\chi^{(2)}$ is zero and there is no second-order nonlinear effect. Thus silica optical fibers, which form the heart of today's communication networks, do not possess the second-order nonlinearity.

We will first discuss second-harmonic generation and parametric amplification which arises due to the second-order nonlinear term and then go on to effects due to third-order nonlinear interaction.

Second-Harmonic Generation in Crystals

The first demonstration of SHG was made in 1961 by focusing a 3 kW ruby laser pulse ($\lambda_0 = 6943$ Å) on a quartz crystal. An incident beam from a ruby laser (red color) after passing through a crystal of KDP, gets partly converted into blue light which is the second harmonic. Ever since, SHG has been one of the most widely studied nonlinear interactions.

We consider a plane wave of frequency ω propagating along the z-direction through a medium and consider the generation of the second-harmonic frequency 2ω as the beam propagates through the medium. Now, the field at ω generates a polarization at 2ω, which acts as a source for the generation of an electromagnetic wave at 2ω. Corresponding to the frequencies ω and 2ω, the electric fields are assumed to be given by

$$E^{(\omega)} = \frac{1}{2}(E_1(z)e^{i(\omega t - k_1 z)} + \text{c.c.}) \qquad [3]$$

and

$$E^{(2\omega)} = \frac{1}{2}(E_2(z)e^{i(2\omega t - k_2 z)} + \text{c.c.}) \qquad [4]$$

respectively; c.c. stands for the complex conjugate of the preceding quantities. The quantities:

$$k_1 = \omega\sqrt{(\varepsilon_1 \mu_0)} = (\omega/c)n^{\omega} \qquad [5]$$

and

$$k_2 = 2\omega\sqrt{(\varepsilon_2 \mu_0)} = (2\omega/c)n^{2\omega} \qquad [6]$$

represent the propagation vectors at ω and 2ω, respectively; ε_1 and ε_2 represent the dielectric permittivities at ω and 2ω, and n^{ω} and $n^{2\omega}$ represent

the corresponding wave refractive indices. It should be noted that the amplitudes E_1 and E_2 are assumed to be z dependent – this is because at $z = 0$ (where the beam is incident on the medium) the amplitude E_2 is zero and this would develop as the beam propagates through the medium. We will now develop an approximate theory for the generation of the second harmonic.

We start with the wave equation:

$$\nabla^2 E - \varepsilon\mu_0 \frac{\partial^2 E}{\partial t^2} = \mu_0 \frac{\partial^2 P_{\mathrm{NL}}}{\partial t^2} \qquad [7]$$

where P_{NL} is the nonlinear polarization. An incident wave at frequency ω generates a polarization at 2ω, which acts as a source for the generation of an electromagnetic wave at 2ω.

In order to consider SHG, we write the wave equation corresponding to 2ω with the nonlinear polarization at 2ω given by

$$P_{\mathrm{NL}}^{(2\omega)} = \frac{1}{2}(\tilde{P}_{\mathrm{NL}}^{(2\omega)} \, \mathrm{e}^{2\mathrm{i}(\omega t - k_1 z)} + \mathrm{c.c.}) \qquad [8]$$

where

$$\tilde{P}_{\mathrm{NL}}^{(2\omega)} = dE_1 E_1 \qquad [9]$$

and

$$d = \frac{\varepsilon_0 \chi^{(2)}}{2} \qquad [10]$$

represents the effective nonlinear coefficient and depends on the nonlinear material, the polarization states of the fundamental and the second harmonic and also on the propagation direction. Simple manipulations give us, for the rate of change of amplitude of the second harmonic wave:

$$\frac{\mathrm{d}E_2}{\mathrm{d}z} = -\frac{i\mu_0 \, dc\omega}{n^{2\omega}} E_1^2(z) \mathrm{e}^{\mathrm{i}(\Delta k)z} \qquad [11]$$

where

$$\Delta k = k_2 - 2k_1 = (2\omega/c)(n^{2\omega} - n^{\omega}) \qquad [12]$$

and we have assumed:

$$\mathrm{d}^2 E_2/\mathrm{d}z^2 \ll k_2(\mathrm{d}E_2/\mathrm{d}z) \qquad [13]$$

In order to solve eqn [11] we neglect depletion of the fundamental field, i.e., $E_1(z)$ is almost a constant and the quantity E_1^2 on the right-hand side can be assumed to be independent of z. If we now integrate eqn [11], we obtain:

$$E_2(z) = -\frac{i\mu_0 \, \mathrm{d}c\omega}{n^{2\omega}} z E_1^2 \mathrm{e}^{\mathrm{i}\sigma} \frac{\sin\sigma}{\sigma} \qquad [14]$$

where

$$\sigma = \frac{1}{2}(\Delta k)z = (\omega/c)(n^{2\omega} - n^{\omega})z \qquad [15]$$

Now, the powers associated with the beams corresponding to ω and 2ω are given by:

$$P_1 = \frac{n^{\omega}}{2c\mu_0} S|E_1|^2, \qquad P_2 = \frac{n^{2\omega}}{2c\mu_0} S|E_2|^2 \qquad [16]$$

where S represents the cross-sectional area of the beams. Substituting for $|E_2|^2$ from eqn [16], we get after some elementary simplifications:

$$\eta = \frac{P_2}{P_1} = \frac{2c^3\mu_0^3 d^2 \omega^2}{(n^{\omega})^2 n^{2\omega}} z^2 \frac{P_1}{S} \left(\frac{\sin\sigma}{\sigma}\right)^2 \qquad [17]$$

when η represents the SHG efficiency. Note that the efficiency of SHG increases if P_1/S, the intensity of the fundamental wave increases. Also η increases if the nonlinear coefficient d increases (obviously) and if the frequency increases. However, for a given power P_1, the conversion efficiency increases if the area of the beam decreases – thus a focused beam will have a greater SHG efficiency. The most important factor is $(\sin\sigma/\sigma)^2$ which is a sharply peaked function around $\sigma = 0$, attaining a maximum value of unity at $\sigma = 0$. Thus for maximum SHG efficiency:

$$\sigma = 0 \Rightarrow n^{2\omega} = n^{\omega} \qquad [18]$$

i.e., the refractive index at 2ω must be equal to the refractive index at ω – this is known as the phase matching condition.

The phase matching condition can be pictorially represented by a vector diagram (see Figure 3). In Figure 3, $\mathbf{k}_1\omega$ and $\mathbf{k}_2\omega$ represent the wave vectors of the fundamental and the second harmonic respectively. To achieve reasonably effective SHG, it is very important to satisfy the phase-matching condition. Efficiencies under nonphase matched operation can be orders of magnitude lower than under phase-matched operation.

We see from eqn [17] that the smallest z for which η is maximum is:

$$z = L_{\mathrm{c}} = \frac{\pi}{\Delta k} = \frac{\pi c}{2\omega(n^{2\omega} - n^{\omega})} \qquad [19]$$

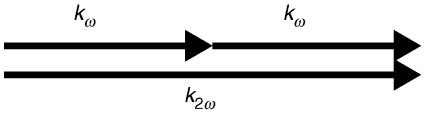


Figure 3 Vector diagram representing the phase matching condition for SHG.

where we have used eqn [15] for Δk. The length L_c is called the phase coherence length and represents the maximum crystal length up to which the second-harmonic power increases. Thus, if the length of the crystal is less than L_c, the second-harmonic power increases almost quadratically with z. For $z > L_c$, the second-harmonic power begins to reduce again.

In general, because of dispersion, it is very difficult to satisfy the phase-matching condition. However, in a birefringent medium, for example with $n_o > n_e$, it may be possible to find a direction along which the refractive index of the *o*-wave for ω equals the refractive index of the *e*-wave for 2ω (see Figure 4). For media with $n_e > n_o$, the direction would correspond to that along which the refractive index of the *e*-wave for ω equals the refractive index for the *o*-wave for 2ω. This can readily be understood by considering a specific example. We consider the SHG in KDP corresponding to the incident ruby laser wavelength ($\lambda_0 = 0.6943\ \mu m$, $\omega = 2.7150 \times 10^{15}$ Hz). For this frequency:

$$\left\{ \begin{array}{ll} n_o^{\omega} = 1.50502, & n_e^{\omega} = 1.46532 \\ n_o^{2\omega} = 1.53269, & n_e^{2\omega} = 1.48711 \end{array} \right\} \qquad [20]$$

The refractive index variation for the *e*-wave is given by:

$$n_e(\theta) = \left(\frac{\sin^2\theta}{n_e^2} + \frac{\cos^2\theta}{n_o^2} \right)^{-\frac{1}{2}} \qquad [21]$$

where θ is the angle that the wave propagation direction makes with the optic axis. Now, as can be seen from the above equations:

$$n_e < n_e(\theta) < n_o \qquad [22]$$

Since n_o at 0.6943 μm lies between n_e and n_o at 0.34715 μm, there will always exist an angle θ_m along which $n_e^{2\omega}(\theta_m) = n_o^{\omega}$. We can solve for θ_m and obtain:

$$\cos\theta_m = \left[\frac{(n_o^{\omega})^2 - (n_e^{2\omega})^2}{(n_o^{2\omega})^2 - (n_e^{2\omega})^2} \right]^{1/2} \frac{n_o^{2\omega}}{n_o^{\omega}} \qquad [23]$$

For the values given by eqn [20], we find $\theta_m = 50.5°$.

Quasi Phase Matching (QPM)

As mentioned earlier, phase matching is extremely important for any efficient nonlinear interaction. Recently the technique of quasi phase matching (QPM) has become a convenient way to achieve phase matching at any desired wavelength in any material. QPM relies on the fact that the phase mismatch in the two interacting beams can be compensated by periodically readjusting the phase of interaction through periodically modulating the nonlinear characteristics of the medium at a spatial frequency equal to the wavevector mismatch of the two interacting waves. Thus in SHG, when the nonlinear polarization at 2ω and the electromagnetic wave at 2ω have an accumulated phase difference of π, then the sign of the nonlinear coefficient is reversed so that the energy flowing from the polarization to the wave can add constructively with the existing energy (see Figure 5). Thus, by properly choosing the period of the spatial modulation of the nonlinear coefficient, one could achieve phase matching. This scheme, QPM, is being very widely studied for application to nonlinear interactions in bulk and in waveguides.

In a ferroelectric material such as lithium niobate, the signs of the nonlinear coefficients are linked to the direction of the spontaneous polarization. Thus a

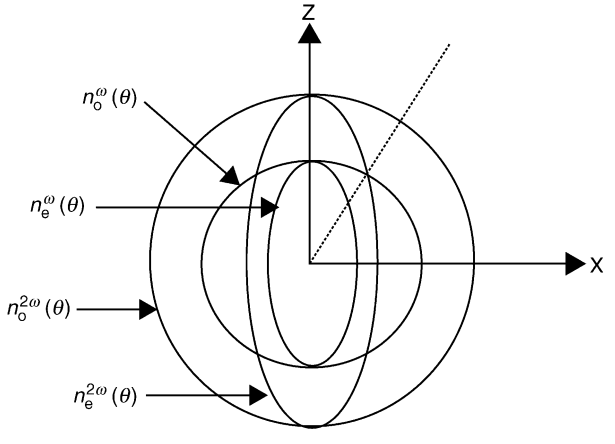


Figure 4 In a birefringent medium, for example with $n_o > n_e$, it may be possible to find a direction along which the refractive index of the *o*-wave for ω equals the refractive index of the *e*-wave for 2ω.

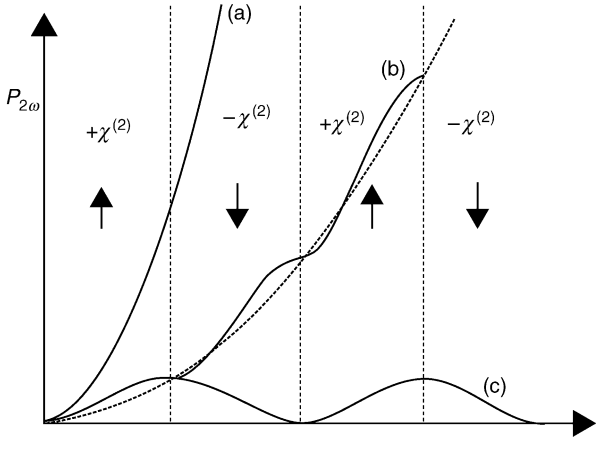


Figure 5 By reversing the sign of the nonlinear coefficient after every coherence length, the energy in the second harmonic can be made to grow. (a) Perfect phase matching; (b) Quasi phase matching; (c) Nonphase matched.

periodic reversal of the domains of the crystal can be used for achieving QPM (see Figure 6). This is the currently used technique to obtain high-efficiency SHG and other nonlinear interactions in $LiNbO_3$, $LiTaO_3$, and KTP. The most popular technique today, to achieve periodic domain reversal in $LiNbO_3$, is the technique of electric field poling. In this method, a high electric field pulse is applied to properly oriented lithium niobate crystal using lithographically defined electrode patterns to produce a permanent periodic domain reversed pattern. Such a periodically domain reversed $LiNbO_3$ crystal with the periodically reversed domains going through the entire depth of the crystal is also referred to as PPLN (periodically poled lithium niobate, pronounced 'piplin') and is now commercially available.

In order to analyze SHG in a periodically poled material, let us assume that the nonlinear coefficient d varies sinusoidally with a period Λ. In such a case we have:

$$d = d_0 \sin(Kz) \qquad [24]$$

where d_0 is the amplitude of modulation of the nonlinear coefficient and $K(= 2\pi/\Lambda)$ represents the spatial frequency of the periodic modulation. For easier understanding we are assuming the modulation to be sinusoidal; in general, the modulation will be periodic but not sinusoidal. Any periodic modulation can be written as a superposition of sinusoidal and cosinusoidal variations. Thus our discussion is valid for one of the Fourier components of the variation.

By using eqn [24], eqn [11] for the variation of the amplitude of the second harmonic becomes:

$$\frac{dE_2}{dz} = -\frac{i\mu_0 d_0 c\omega}{n^{2\omega}} E_1^2(z) e^{i(\Delta k)z} \sin(Kz) \qquad [25]$$

which can be written as:

$$\frac{dE_2}{dz} = -\frac{\mu_0 d_0 c\omega}{2n^{2\omega}} E_1^2(z) e^{i(\Delta k)z} [e^{iKz} - e^{-iKz}] \qquad [26]$$

Using similar arguments as earlier, it can be shown that if $\Delta k - K = 0$, then only the second term within the brackets in eqn [26] contributes to the growth of the second harmonic and similarly if $\Delta k + K = 0$, then only the first term within the brackets contributes to the growth of the second harmonic. The first condition implies that:

$$2\frac{2\pi}{\lambda_0}(n^{2\omega} - n^{\omega}) - \frac{2\pi}{\Lambda} = 0 \qquad [27]$$

where $\Lambda = 2\pi/K$ represents the spatial period of the modulation of the nonlinear coefficient, and λ_0 is the wavelength of the fundamental. Thus the modulation period Λ required for QPM SHG is:

$$\Lambda = \frac{\lambda_0}{2(n^{2\omega} - n^{\omega})} = 2L_c \qquad [28]$$

Figure 7 shows the vector diagram corresponding to QPM SHG. The phase mismatch between the fundamental and second harmonic is compensated by the spatial frequency vector of the periodic variation of d.

In the case of waveguides, n^{ω} and $n^{2\omega}$ would represent the effective indices of the modes at the fundamental and the second-harmonic frequency. It may be noted that the index difference between the fundamental and the second harmonic is typically 0.1 and thus the required spatial periods for a fundamental wavelength of 800 nm is ~ 4 μm.

The main advantage of QPM is that it can be used at any wavelength within the transparency window of the material; only the period needs to be correctly chosen for a specific fundamental wavelength. One can also choose appropriate polarization to make use of the largest nonlinear optical coefficients. Another advantage is the possibility of varying the domain reversal period (chirp) to achieve specific interaction characteristics such as increased bandwidth, etc.

In general, the spatial variation of the nonlinear grating is not sinusoidal. In this case the efficiency of interaction would be determined by the Fourier component of the spatial variation at the spatial frequency corresponding to the period given by eqn [28]. Also, since the required periods are very small, it is possible to use a higher spatial period of modulation and use one of the Fourier components for the nonlinear interaction process. Thus, in the case of periodic reversal of the nonlinear coefficient

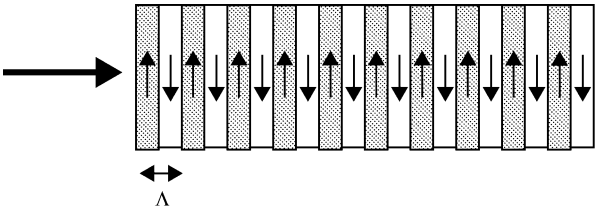


Figure 6 QPM can be achieved by a periodic reversal of the domains of a ferroelectric material. Arrows represent the direction of the spontaneous polarization of the crystal.

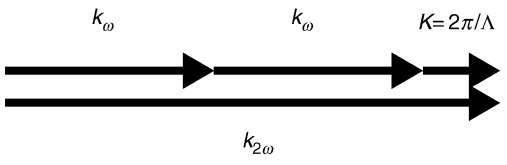


Figure 7 Vector diagram corresponding to QPM-SHG.

with a spatial period given by:

$$\Lambda_g = m\frac{\lambda_0}{2(n^{2\omega} - n^{\omega})}; \qquad m = 1, 3, 5, \ldots \qquad [29]$$

which happens to be the m^{th} harmonic of the fundamental spatial frequency required for QPM, the corresponding nonlinear coefficient that would be responsible for SHG would be the Fourier amplitude at that spatial frequency. This can be taken into account by defining an effective nonlinear coefficient:

$$d_{QPM} = \frac{2d_0}{m\pi} \qquad [30]$$

Of course the largest effective nonlinear coefficient is achieved by using the fundamental frequency with $m = 1$. Higher spatial periods are easier to fabricate but would lead to reduced nonlinear efficiencies.

As compared to bulk devices, in the case of waveguides, the interacting waves are propagating in the form of modes having specific intensity distributions in the transverse plane. Because of this, the nonlinear interaction also depends on the overlap between the periodically inverted nonlinear medium, the fields of the fundamental and second-harmonic waves. Thus if $E_\omega(x, y)$ and $E_{2\omega}(x, y)$ represent the electric field distributions of the waveguide modes at the fundamental and second-harmonic frequency, then the efficiency of SHG depends on the following overlap integral:

$$I_{ovl} = \int\int d_0(x, y)E_\omega^2(x, y)E_{2\omega}(x, y)\,dx\,dy \qquad [31]$$

where $d_0(x, y)$ represents the transverse variation of the nonlinear coefficient of the waveguide. Thus optimization of SHG efficiency in the case of waveguide interactions has to take account of the overlap integral.

Since QPM relies on periodic phase matching, it is highly wavelength dependent. Thus the required period Λ is different for different fundamental frequencies. Any deviation in the frequency of the fundamental would lead to a reduction in the efficiency due to deviation from QPM condition. Thus the pump laser needs to be highly stabilized at a frequency corresponding to the fabricated period.

Apart from SHG, QPM has been used for other three-wave interaction processes such as difference frequency generation, parametric amplification, etc. Among the many materials that have been studied for SHG using QPM, the important ones are lithium niobate, lithium tantalite, and potassium titanyl phosphate (KTP). Many techniques have been developed for periodic domain reversal to achieve a periodic variation in the nonlinear coefficient. This includes electric field poling, electron bombardment, thermal diffusion treatment, etc.

Third-Order Nonlinear Effects

In the earlier sections, we discussed effects arising out of second-order nonlinearity, i.e., the term proportional to E^2 in the nonlinear polarization. This nonlinear term is found only in media not possessing an inversion symmetry. Thus amorphous materials or crystals possessing inversion symmetry do not exhibit second-order effects. The lowest-order nonlinear effects in such a medium is of an order three wherein the nonlinear polarization is proportional to E^3.

Self-phase modulation (SPM), cross-phase modulation (XPM) and four-wave mixing (FWM) represent some of the very important consequences of third-order nonlinearity. These effects have become all the more important as they play a significant and important role in wavelength division multiplexed optical fiber communication systems.

Self-Phase Modulation (SPM)

Consider the propagation of a plane light wave at frequency ω through a medium having $\chi^{(3)}$ nonlinearity. The polarization generated in the medium is given by

$$P = \varepsilon_0\chi E + \varepsilon_0\chi^{(3)}E^3 \qquad [32]$$

If we consider a single frequency wave with an electric field given by

$$E = E_0\cos(\omega t - kz) \qquad [33]$$

then

$$P = \varepsilon_0\chi E_0\cos(\omega t - kz) + \varepsilon_0\chi^{(3)}E_0^3\cos^3(\omega t - kz) \qquad [34]$$

Expanding $\cos^3\theta$ in terms of $\cos\theta$ and $\cos 3\theta$, we obtain the following expression for the polarization at frequency ω:

$$P = \varepsilon_0\left(\chi + \frac{3}{4}\chi^{(3)}E_0^2\right)E_0\cos(\omega t - kz) \qquad [35]$$

For a plane wave given by eqn [33], the intensity is

$$I = \frac{1}{2}c\varepsilon_0 n_0 E_0^2 \qquad [36]$$

where n_0 is the refractive index of the medium at low intensity. Then

$$P = \varepsilon_0\left(\chi + \frac{3}{2}\frac{\chi^{(3)}}{c\varepsilon_0 n_0}I\right)E \qquad [37]$$

The polarization P and electric field are related through the following equation:

$$P = \varepsilon_0(n^2 - 1)E \qquad [38]$$

where n is the refractive index of the medium. Comparing eqns [37] and [38], we get

$$n^2 = n_0^2 + \frac{3}{2}\frac{\chi^{(3)}}{c\varepsilon_0 n_0} I \qquad [39]$$

where

$$n_0^2 = 1 + \chi \qquad [40]$$

Since the last term in the eqn [39] is usually very small, we get

$$n = n_0 + n_2 I \qquad [41]$$

where

$$n_2 = \frac{3}{4}\frac{\chi^{(3)}}{c\varepsilon_0 n_0^2} \qquad [42]$$

is the nonlinear coefficient.

For fused silica $n_0 \approx 1.47$, $n_2 \approx 3.2 \times 10^{-20}$ m^2/W and if we consider power of 100 mW having a cross-sectional area of 100 μm^2, the resultant intensity is 10^9 W/m^2 and the corresponding change in refractive index is

$$\Delta n \approx n_2 I \approx 3.2 \times 10^{-11} \qquad [43]$$

Although this is very small, when the beam propagates over an optical fiber over long distances (a few hundred to a few thousand kilometers), the accumulated nonlinear effects can be significant.

In the case of an optical fiber, the light beam propagates as a mode having a specific transverse electric field distribution and thus the intensity is not constant across the cross-section. In such a case, it is convenient to express the nonlinear effect in terms of the power carried by the mode (rather than in terms of intensity). If the linear propagation constant of the mode is represented by β, then in the presence of nonlinearity, the effective propagation constant is given by

$$\beta_{\mathrm{NL}} = \beta + \frac{k_0 n_2}{A_{\mathrm{eff}}} P \qquad [44]$$

where $k_0 = 2\pi/\lambda_0$, P is the power carried by the mode. The quantity A_{eff} represents the effective transverse cross-sectional area of the mode and is defined by

$$A_{\mathrm{eff}} = \frac{\left[\int_0^{\infty}\int_0^{2\pi} \psi^2(r) r \, \mathrm{d}r \, \mathrm{d}\phi\right]^2}{\int_0^{\infty}\int_0^{2\pi} \psi^4(r) r \, \mathrm{d}r \, \mathrm{d}\phi} \qquad [45]$$

where $\psi(r)$ represents the transverse mode field distribution of the mode. For example, under the Gaussian approximation:

$$\psi(r) = \psi_0 \, \mathrm{e}^{-r^2/w_0^2} \qquad [46]$$

where ψ_0 is a constant and $2w_0$ represents the mode field diameter (MFD), we get:

$$A_{\mathrm{eff}} = \pi w_0^2 \qquad [47]$$

It is usual to express the nonlinear characteristic of an optical fiber by the coefficient given by

$$\gamma = \frac{k_0 n_2}{A_{\mathrm{eff}}} \qquad [48]$$

Thus, for the same input power and same wavelength, smaller values of A_{eff} lead to greater nonlinear effects in the fiber. Typically:

$$A_{\mathrm{eff}} \sim 50\text{–}80 \ \mu\mathrm{m}^2 \quad \text{and} \quad \gamma \approx 2.4 \ \mathrm{W}^{-1}\,\mathrm{km}^{-1} \ \text{to} \ 1.5 \ \mathrm{W}^{-1}\,\mathrm{km}^{-1} \qquad [49]$$

When a light beam propagates through an optical fiber, the power decreases because of attenuation. Thus, the corresponding nonlinear effects also reduce. Indeed, the phase change suffered by a beam in propagating from 0 to L is given by

$$\phi = \int_0^L \beta_{\mathrm{NL}} \, \mathrm{d}z = \beta L + \gamma \int_0^L P \, \mathrm{d}z \qquad [50]$$

If α represents the attenuation coefficient, then

$$P(z) = P_0 \, \mathrm{e}^{-\alpha z} \qquad [51]$$

and we get

$$\phi = \beta L + \gamma P_0 L_{\mathrm{eff}} \qquad [52]$$

where

$$L_{\mathrm{eff}} = \frac{1 - \mathrm{e}^{-\alpha L}}{\alpha} \qquad [53]$$

is referred to as the effective length. For $\alpha L \gg 1$, $L_{\mathrm{eff}} \approx 1/\alpha$ and for $\alpha L \ll 1$, $L_{\mathrm{eff}} \approx L$.

The effective length represents the length of the fiber over which most of the nonlinear effects has accumulated. For a loss coefficient of 0.20 dB/km, $L_{\mathrm{eff}} \approx 21$ km.

If we consider a fiber length much longer than L_{eff}, then to have reduced impact of SPM, we must have

$$\gamma P_0 L_{\mathrm{eff}} \ll 1 \qquad [54]$$

or

$$P_0 \ll \frac{1}{\gamma L_{\mathrm{eff}}} \approx \frac{\alpha}{\gamma} \qquad [55]$$

For $\alpha = 4.6 \times 10^{-2}$ km^{-1} (which corresponds to an attenuation of 0.2 dB/km) and $\gamma = 2.4$ W^{-1} km^{-1}, we get

$$P_0 \ll 19 \ \mathrm{mW} \qquad [56]$$

Propagation of a Pulse

When an optical pulse propagates through a medium, it suffers from the following effects:

(i) attenuation;
(ii) dispersion; and
(iii) nonlinearity.

Attenuation refers to the reduction in the pulse energy due to various mechanisms, such as scattering, absorption, etc. Dispersion is caused by the fact that a light pulse consists of various frequency components and each frequency component travels at a different group velocity. Dispersion causes the temporal width of the pulse to change; in most cases it results in an increase in pulse width, however, in some cases the temporal width could also decrease. Dispersion is accompanied by chirping, the variation of the instantaneous frequency of the pulse within the pulse duration. Since both attenuation and dispersion cause a change in the temporal variation of the optical power, they closely interact with nonlinearity in deciding the pulse evolution as it propagates through the medium.

Let $E(x, y, z, t)$ represent the electric field variation of an optical pulse. It is usual to express E in the following way:

$$E(x, y, z, t) = \frac{1}{2}[A(z, t)\psi(x, y)\mathrm{e}^{\mathrm{i}(\omega_0 t - \beta_0 z)} + \text{c.c.}] \quad [57]$$

where $A(z, t)$ represents the slowly varying complex envelope of the pulse, $\psi(x, y)$ represents the transverse electric field distribution, ω_0 represents center frequency, and β_0 represents the propagation constant at ω_0.

In the presence of attenuation, second-order dispersion and third-order nonlinearity, the complex envelope $A(z, t)$ can be shown to satisfy the following equation:

$$\frac{\partial A}{\partial z} = -\frac{\alpha}{2}A - \beta_1\frac{\partial A}{\partial t} + i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial t^2} - i\gamma|A|^2 A \quad [58]$$

Here

$$\beta_1 = \left|\frac{\mathrm{d}\beta}{\mathrm{d}\omega}\right|_{\omega=\omega_0} = \frac{1}{v_\mathrm{g}} \quad [59]$$

represents the inverse of the group velocity of the pulse, and

$$\beta_2 = \left|\frac{\mathrm{d}^2\beta}{\mathrm{d}\omega^2}\right|_{\omega=\omega_0} = -\frac{\lambda_0^2}{2\pi c}D \quad [60]$$

where D represents the group velocity dispersion (measured in ps/km nm).

The various terms on the RHS of the eqn [58] represent the following:

I term: attenuation
II term: group velocity term
III term: second-order dispersion
IV term: nonlinear term

If we change to a moving frame defined by coordinates $T = t - \beta_1 z$, eqn [58] becomes

$$\frac{\partial A}{\partial z} = -\frac{\alpha}{2}A + i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial T^2} - i\gamma|A|^2 A \quad [61]$$

If we neglect the attenuation term, we obtain the following equation which is also referred to as the nonlinear Schrödinger equation:

$$\frac{\partial A}{\partial z} = i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial T^2} - i\gamma|A|^2 A \quad [62]$$

The above equation has a solution given by

$$A(z, t) = A_0 \operatorname{sech} \sigma T\, \mathrm{e}^{-\mathrm{i}gz} \quad [63]$$

with

$$A_0^2 = -\frac{\beta_2}{\gamma}\sigma^2, \qquad g = -\frac{\sigma^2}{2}\beta_2 \quad [64]$$

Equation [63] represents an envelope soliton and has the property that it propagates undispersed through the medium. The full width at half maximum (FWHM) of the pulse envelope will be given by $\tau_\mathrm{f} = 2\tau_0$, where

$$\operatorname{sech}^2 \sigma\tau_0 = \frac{1}{2} \quad [65]$$

which gives the FWHM τ_f:

$$\tau_\mathrm{f} = 2\tau_0 = \frac{2}{\sigma}\ln(1 + \sqrt{2}) \approx \frac{1.7627}{\sigma} \quad [66]$$

The peak power of the pulse is:

$$P_0 = |A_0|^2 = \frac{|\beta_2|}{\gamma}\sigma^2 \quad [67]$$

Replacing σ by τ_f, we obtain

$$P_0\tau_\mathrm{f}^2 \approx \frac{\lambda_0^2}{2cy}D \quad [68]$$

where we have used eqn [60]. The above equation gives the required peak for a given τ_f for the formation of a soliton pulse.

As an example, we have $\tau_\mathrm{f} = 10$ ps, $\gamma = 2.4\ \mathrm{W}^{-1}\ \mathrm{km}^{-1}$, $\lambda_0 = 1.55\ \mu\mathrm{m}$, $D = 2$ ps/km nm and the required peak power will be $P_0 = 33$ mW.

Soliton pulses are being extensively studied for application to long distance optical fiber

communication. In actual systems, the pulses have to be optically amplified at regular intervals to compensate for the loss suffered by the pulses. The amplification could be carried out using erbium doped fiber amplifiers (EDFAs) or fiber Raman amplifiers.

Spectral Broadening due to SPM

In the presence of only nonlinearity, eqn [61] becomes

$$\frac{\mathrm{d}A}{\mathrm{d}z} = -i\gamma|A|^2A \qquad [69]$$

whose solution is given by

$$A(z,t) = A(z=0,t)\mathrm{e}^{-\mathrm{i}\gamma Pz} \qquad [70]$$

where $P = |A|^2$ is the power in the pulse. If P is a function of time, then time dependent phase term at $z = L$ becomes

$$\mathrm{e}^{\mathrm{i}\phi(t)} = \mathrm{e}^{\mathrm{i}[\omega_0 t - \gamma P(t)L]} \qquad [71]$$

We can define an instantaneous frequency as

$$\omega(t) = \frac{\mathrm{d}\phi}{\mathrm{d}t} = \omega_0 - \gamma L\frac{\mathrm{d}P}{\mathrm{d}t} \qquad [72]$$

for a Gaussian pulse:

$$P = P_0\,\mathrm{e}^{-2T^2/\tau_0^2} \qquad [73]$$

giving

$$\omega(t) = \omega_0 + \frac{4\gamma LTP_0\,\mathrm{e}^{-2T^2/\tau_0^2}}{\tau_0^2} \qquad [74]$$

Thus the instantaneous frequency within the pulse changes with time, leading to chirping of the pulse (see Figure 8). Note that since the pulsewidth has not changed, but the pulse is chirped, the frequency spectrum of the pulse has increased. Thus SPM leads to the generation of new frequencies. By Fourier transform theory, an increased spectral width implies that the pulse can now be compressed in the temporal domain by passing it through a medium with the proper sign of dispersion. This is indeed one of the standard techniques to realize ultrashort femtosecond optical pulses.

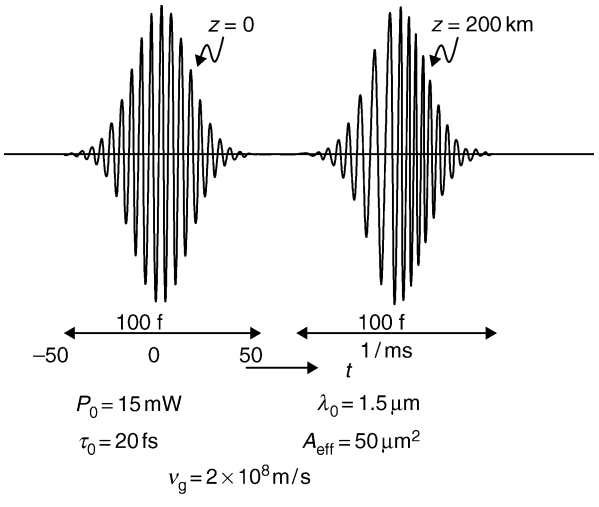


Figure 8 Due to self phase modulation, the instantaneous frequency within the pulse changes with time leading to chirping of the pulse.

Cross Phase Modulation (XPM)

Like SPM, cross-phase modulation also arises due to the intensity dependence of refractive index, leading to spectral broadening. Unlike SPM, in the case of XPM, intensity variations of a light beam at a particular frequency modulate the phase of light beam at another frequency. If the signals at both frequencies are pulses, then due to difference in group velocities of the pulses, there is a walk off between the two pulses, i.e., if they start together, they will separate as they propagate through the medium. Nonlinear interaction takes place as long as they physically overlap in the medium. Smaller the dispersion, smaller will be the difference in group velocities (assuming closely spaced wavelengths) and the longer they will overlap. This would lead to stronger XPM effects. At the same time, if two pulses pass through each other, then since one pulse will interact with both the leading and the trailing edge of the other pulse, XPM effects will be nil provided there is no attenuation. In the presence of attenuation in the medium, the pulse will still get modified due to XPM. A similar situation can occur when the two interacting pulses are passing through an optical amplifier.

To study XPM, we assume simultaneous propagation of two waves at two different frequencies through the medium. If ω_1 and ω_2 represent the two frequencies, then one obtains for the variation of the amplitude A_1 of the frequency ω_1:

$$\frac{\mathrm{d}A_1}{\mathrm{d}z} = -i\gamma(\tilde{P}_1 + 2\tilde{P}_2)A_1 \qquad [75]$$

where $\tilde{P}_1$ and $\tilde{P}_2$ represent the powers at frequencies ω_1 and ω_2, respectively. The first term in eqn [75] represents SPM while the second term corresponds to XPM. If the powers are assumed to attenuate at the same rate, i.e.:

$$\tilde{P}_1 = P_1\,\mathrm{e}^{-\alpha z}, \qquad \tilde{P}_2 = P_2\,\mathrm{e}^{-\alpha z} \qquad [76]$$

then the solution of eqn [75] is

$$A_1(L) = A_1(0)\,\mathrm{e}^{-\mathrm{i}\gamma(P_1+2P_2)L_{\mathrm{eff}}} \qquad [77]$$

where, as before, L_{eff} represents the effective length of the medium. When we are studying the effect of

power at ω_2 on the light beam at frequency ω_1, we will refer to the wave at frequency ω_2 as pump, and the wave at frequency ω_1 as probe or signal. From eqn [77] it is apparent that the phase of signal at frequency ω_1 is modified by the power at another frequency. This is referred to as XPM. Note also that XPM is twice as effective as SPM.

Similar to the case of SPM, we can now write for the instantaneous frequency in the presence of XPM as eqn [72]:

$$\omega(t) = \omega_0 - 2\gamma L_{\text{eff}} \frac{dP_2}{dt} \qquad [78]$$

Hence the part of the signal that is influenced by the leading edge of the pump will be down-shifted in frequency (since in the leading edge $dP_2/dt > 0$) and the part overlapping with the trailing edge will be up-shifted in frequency (since $dP_2/dt < 0$). This leads to a frequency chirping of the signal pulse just as in the case of SPM.

If the probe and pump beams are pulses, then XPM can lead to induced frequency shifts depending on whether the probe pulse interacts only with the leading edge or trailing edge or both, as they both propagate through the medium. Let us consider a case when the group velocity of pump pulse is greater than that of the probe pulse. Thus, if both pulses enter the medium together, then since the pump pulse travels faster, the probe pulse will interact only with the trailing edge of the pump. Since in this case dP_2/dt is negative, the probe pulse suffers a blue-induced frequency shift. Similarly if the pulses enter at different instants but completely overlap at the end of the medium, then $dP_2/dt > 0$ and the probe pulse would suffer a red-induced frequency shift. Indeed, if the two pulses start separately and walk through each other, then there is no induced shift due to cancellation of shifts induced by leading and trailing edges of the pump. Figure 9 shows measured induced frequency shift of 532 nm probe pulse as a function of the delay between this pulse and a pump pulse at 1064 nm, when they propagate through a 1 m long optical fiber.

When pulses of light at two different wavelengths propagate through an optical fiber, due to different group velocities of the pulses, they pass through each other, resulting in what could be termed as a collision. In the linear case, the pulses will pass through without affecting each other, but when intensity levels are high, XPM induces phase shifts in both pulses. We can define a parameter termed walk off length L_{wo} which is the length of the fiber required for the interacting pulses to walk off relative to each other. The walk off length is given by

$$L_{\text{wo}} = \frac{\Delta\tau}{D\Delta\lambda} \qquad [79]$$

where D represents the dispersion coefficient, and $\Delta\lambda$ represents the wavelength separation between the interacting pulses. For return to zero (RZ) pulses, $\Delta\tau$ represents the pulse duration while for nonreturn to zero (NRZ) pulses, $\Delta\tau$ represents the rise term or fall time of the pulse. Closely spaced channels will thus interact over longer fiber lengths, thus leading to greater XPM effects. Larger dispersion coefficients will reduce L_{wo} and thus the effects of XPM. Since the medium is attenuating, the power carried by the pulses decreases as they propagate and thus leading to reduced XPM effect. The characteristic length for attenuation is the effective length L_{eff}, defined by eqn [53]. If $L_{\text{wo}} \ll L_{\text{eff}}$, then over the length of interaction of the pulses, the intensity levels do not change appreciably and the magnitude of the XPM-induced effects will be proportional to the wavelength spacing $\Delta\lambda$. For small $\Delta\lambda$'s, $L_{\text{wo}} \gg L_{\text{eff}}$ and the interaction length is now determined by the fiber losses (rather than by walk off) and the XPM-induced effects become almost independent of $\Delta\lambda$. Indeed, if we consider XPM effects between a continuous wave (cw) probe beam and a sinusoidally industry modulated pump beam, then the amplitude of the XPM-induced

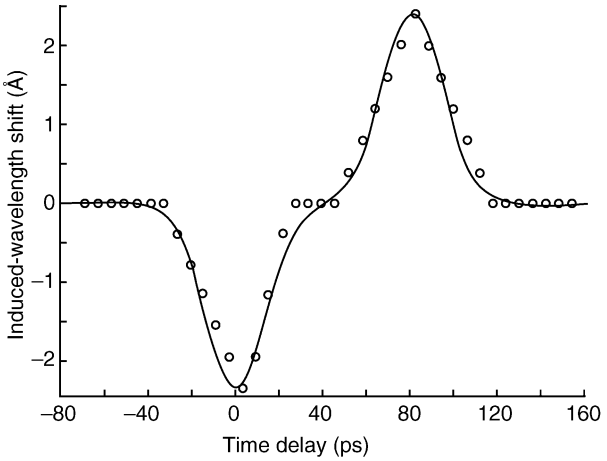


Figure 9 Measured induced frequency shift of 532 nm probe pulse as a function of the delay between this pulse and a pump pulse at 1064 nm, when they propagate through a 1 m long optical fiber. (Reproduced from Baldeck PL, Alfano RR and Agrawal GP (1998) Induced frequency shift of copropagating ultrafast optical pulses. *Applied Physics Letters* 52: 1939–1941 with permission from the American Institute of Physics.)

phase shift ($\Delta\Phi_{pr}$) in the probe beam is given by:

$$\begin{aligned} \Delta\Phi_{pr} &\approx 2\gamma P_{2m} L_{eff} \quad \text{for} \quad L_{wo} \gg L_{eff} \\ \Delta\Phi_{pr} &\approx 2\gamma P_{2m} L_{wo} \quad \text{for} \quad L_{wo} \ll L_{eff} \end{aligned} \qquad [80]$$

Here P_{2m} is the amplitude of the sinusoidal power modulation of the pump beam.

XPM-induced intensity interference can be studied by simultaneously propagating an intensity modulated pump signal and a cw probe signal at a different wavelength. The intensity modulated signal will induce phase modulation on the cw probe signal and the dispersion of the medium will convert the phase modulation to intensity modulation of the probe. Thus, the magnitude of the intensity fluctuation of the probe signal serves as an estimate of the XPM induced interference. Figure 10 shows the intensity fluctuations on a probe signal at 1550 nm, induced by a modulated pump for a channel separation of 0.6 nm. Figure 11 shows the variation of the RMS value of probe intensity modulation with the wavelength separation between the intensity modulated signal and the probe. The experiment has been performed over four amplified spans of 80 km of standard single mode fiber (SMF) and nonzero dispersion shifted fiber (NZDSF). The large dispersion in SMF has been compensated using dispersion compensating chirped gratings. The probe modulation, in the case of SMF, decreases approximately linearly with $1/\Delta\lambda$ for all $\Delta\lambda$s; the modulation is independent of $\Delta\lambda$. In contrast the NZDSF shows a constant modulation for $\Delta\lambda \leq 1$ nm.

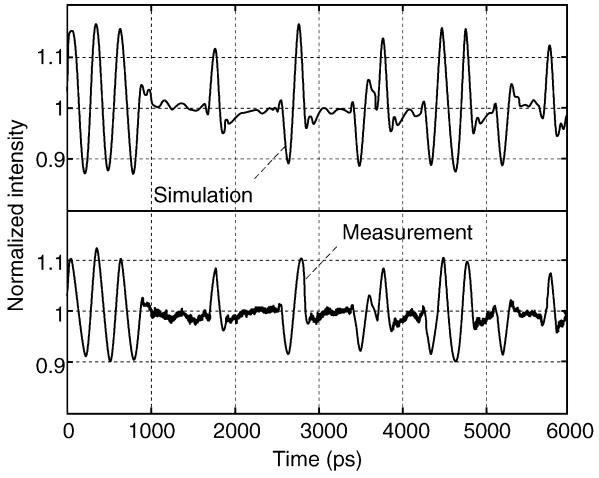


Figure 10 Intensity fluctuations induced by cross phase modulation on a probe signal at 1550 nm by a modulated pump for a channel separation of 0.6 nm. (Reproduced with permission from Rapp L (1997) Experimental investigation of signal distortions induced by cross phase modulation combined with dispersion. *IEEE Photon Technical Letters* 9: 1592–1595, © 2004 IEEE.)

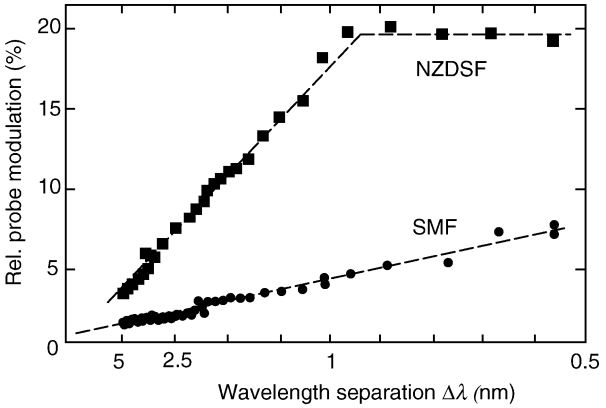


Figure 11 Variation of the RMS value of probe intensity modulation with the wavelength separation between the intensity modulated signal and the probe. (Reproduced with permission from Shtaif M, Eiselt M and Garret LD (2000) Cross-phase modulation distortion measurements in multispan WDM systems. *IEEE Photon Technical Letters* 12: 88–90, © 2004 IEEE.)

This is consistent with the earlier discussion in terms of L_{wo} and L_{eff}.

Four-Wave Mixing (FWM)

Four-wave mixing (FWM) is a nonlinear interaction that occurs in the presence of multiple wavelengths in a medium, leading to the generation of new frequencies. Thus, if light waves at three different frequencies ω_2, ω_3, ω_4 are launched simultaneously into a medium, the same nonlinear polarization that led to intensity dependent refractive index, leads to nonlinear polarization component at a frequency:

$$\omega_1 = \omega_3 + \omega_4 - \omega_2 \qquad [81]$$

This nonlinear polarization, under certain conditions, leads to the generation of electromagnetic waves at ω_1. This process is referred to as four-wave mixing due to the interaction between four different frequencies. Degenerate four-wave mixing corresponds to the case when two of the input waves have the same frequency. Thus, if $\omega_3 = \omega_4$, then inputting waves at ω_2 and ω_3 leads to the generation of waves at the frequency

$$\omega_1 = 2\omega_3 - \omega_2 \qquad [82]$$

During FWM process, there are four different frequencies present at any point in the medium. If we write the electric field of the waves as

$$E_i = \frac{1}{2}[A_i(z)\psi_i(x,y)e^{i(\omega_i t - \beta_i z)} + \text{c.c.}], \quad i = 1,2,3,4 \qquad [83]$$

where, as before, $A_i(z)$ represents the amplitude of the wave, $\psi_i(x, y)$ the transverse field distribution, and β_i the propagation constant of the wave. The total electric field is given by

$$E = E_1 + E_2 + E_3 + E_4 \qquad [84]$$

Substituting for the total electric field in the equation for nonlinear polarization, the term with frequency ω_1 comes out as

$$P_{\mathrm{NL}}^{(\omega_1)} = \frac{1}{2}[P_{\mathrm{NL}}^{(\omega_1)}\, \mathrm{e}^{\mathrm{i}(\omega_1 t - \beta_1 z)} + \mathrm{c.c.}] \qquad [85]$$

where

$$P_{\mathrm{NL}}^{(\omega_1)} = \frac{3\varepsilon_0}{2}\chi^{(3)} A_2^* A_3 A_4 \psi_2 \psi_3 \psi_4\, \mathrm{e}^{-\mathrm{i}\Delta\beta z} \qquad [86]$$

and

$$\Delta\beta = \beta_3 + \beta_4 - \beta_2 - \beta_1 \qquad [87]$$

In writing eqn [86], we have only considered the FWM term, neglecting the SPM and XPM terms.

Substituting the expression for $P_{\mathrm{NL}}^{(\omega_1)}$ in the wave equation for ω_1 and making the slowly varying approximation (in a manner similar to that employed in the case of SPM and XPM), we obtain the following equation for $A_1(z)$:

$$\frac{\mathrm{d}A_1}{\mathrm{d}z} = -2i\gamma A_2^* A_3 A_4\, \mathrm{e}^{-\mathrm{i}\Delta\beta z} \qquad [88]$$

where γ is defined by eqn [48] with $k_0 = \omega/c$, ω representing the average frequency of the four interacting waves, and A_{eff} is the average effective area of the modes.

Assuming all waves to have the same attenuation coefficient α and neglecting depletion of waves at frequencies ω_2, ω_3 and ω_4, due to nonlinear conversion we obtain for the power in the frequency ω_1 as

$$P_1(L) = 4\gamma^2 P_2 P_3 P_4 L_{\mathrm{eff}}^2 \eta\, \mathrm{e}^{-\alpha L} \qquad [89]$$

where

$$\eta = \frac{\alpha^2}{\alpha^2 + \Delta\beta^2}\left[1 + \frac{4\mathrm{e}^{-\alpha L}\sin^2\dfrac{\Delta\beta L}{2}}{(1 - \mathrm{e}^{-\alpha L})^2}\right] \qquad [90]$$

and L_{eff} is the effective length (see eqn [53]). Maximum four-wave mixing takes place when $\Delta\beta = 0$, since in such a case $\eta = 1$. Now:

$$\Delta\beta = \beta(\omega_3) + \beta(\omega_4) - \beta(\omega_2) - \beta(\omega_1) \qquad [91]$$

Since the frequencies are usually close to each other, we can make a Taylor series expansion about any frequency, say ω_2. In such a case, we obtain

$$\Delta\beta = (\omega_3 - \omega_2)(\omega_3 - \omega_1)\left|\frac{\mathrm{d}^2\beta}{\mathrm{d}\omega^2}\right|_{\omega=\omega_2} \qquad [92]$$

In optical fiber communication systems, the channels are usually equally spaced. Thus, we assume the frequencies to be given by

$$\omega_4 = \omega_2 + \Delta\omega, \qquad \omega_3 = \omega_2 - 2\Delta\omega \text{ and} \qquad \omega_1 = \omega_2 - \Delta\omega \qquad [93]$$

Using these frequencies and eqn [60], eqn [92] gives us:

$$\Delta\beta = -\frac{4\pi D\lambda^2}{c}(\Delta\nu)^2 \qquad [94]$$

where $\Delta\omega = 2\pi\Delta\nu$. Thus maximum FWM takes place when $D = 0$. This is the main problem in using wavelength division multiplexing (WDM) in dispersion shifted fibers which are characterized by zero dispersion at the operating wavelength of 1550 nm, as FWM will then lead to crosstalk among various channels. FWM efficiency can be reduced by using fiber with nonzero dispersion. This has led to the development of nonzero dispersion shifted fiber (NZDSF) which have a finite nonzero dispersion of about ± 2 ps/km nm at the operating wavelength.

From eqn [94], we notice that for a given dispersion coefficient D, FWM efficiency will reduce as $\Delta\nu$ increases.

In order to get a numerical appreciation, we consider a case with $D = 0$, i.e., $\Delta\beta = 0$. For such a case $\eta = 1$. If all channels were launched with equal power P_{in} then:

$$P_1(L) = 4\gamma^2 P_{\mathrm{in}}^3 L_{\mathrm{eff}}^2\, \mathrm{e}^{-\alpha L} \qquad [95]$$

Thus the ratio of power generated at ω_1, due to FWM and that existing at the same frequency, is

$$\frac{P_{\mathrm{g}}}{P_{\mathrm{out}}} = \frac{P_1(L)}{P_{\mathrm{in}}\,\mathrm{e}^{-\alpha L}} = 4\gamma^2 P_{\mathrm{in}}^2 L_{\mathrm{eff}}^2 \qquad [96]$$

Typical values are $L_{\mathrm{eff}} = 20$ km, $\gamma = 2.4\ \mathrm{W}^{-1}\ \mathrm{km}^{-1}$. Thus:

$$\frac{P_{\mathrm{g}}}{P_{\mathrm{out}}} \approx 0.01 P_{\mathrm{in}}^2(\mathrm{mW}^2) \qquad [97]$$

Figure 12 shows the output spectrum measured at the output of a 25 km-long dispersion shifted fiber ($D = -0.2$ ps/km nm) when three 3 mW wavelengths are launched simultaneously. Notice the generation of many new frequencies by FWM. Figure 13 shows the

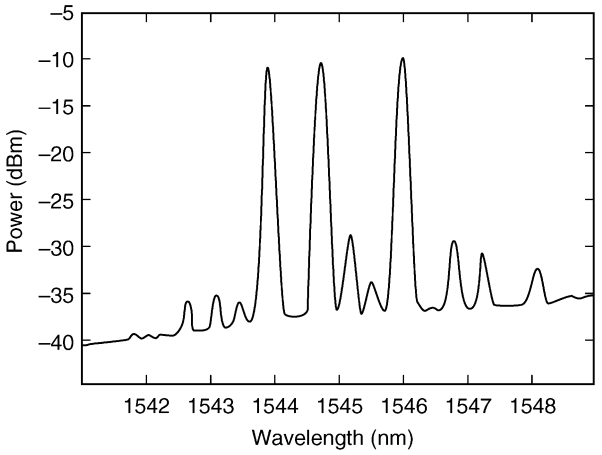


Figure 12 Generation of new frequencies because of FWM when waves at three frequencies are incident in the fiber. (Reproduced with permission from Tkach RW, Chraplyvy AR, Forghiari F, Gnanck AH and Derosier RM (1995) Four-photon mixing and high speed WDM systems. *Journal of lightwave Technology* 13: 841–849, © 2004 IEEE.)

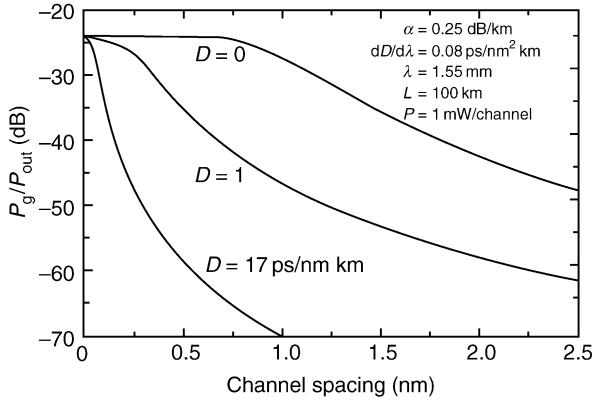


Figure 13 Ratio of generated power to the output as a function of channel spacing $\Delta\lambda$ for different dispersion coefficients. (Reproduced with permission from Tkach RW, Chraplyvy AR, Forghiari F, Gnanck AH and Derosier RM (1995) Four-photon mixing and high speed WDM systems. *Journal of lightwave Technology* 13: 841–849, © 2004 IEEE.)

ratio of generated power to the output as a function of channel spacing $\Delta\lambda$ for different dispersion coefficients. It can be seen that by choosing a nonzero value of dispersion, the four-wave mixing efficiency can be reduced. Larger the dispersion coefficient, smaller can be the channel spacing for the same crosstalk.

Since dispersion leads to increased bit error rates in fiber optic communication systems, it is important to have low dispersion. On the other hand, lower dispersion leads to crosstalk due to FWM. This problem can be resolved by noting that FWM depends on the local dispersion value in the fiber, while the pulse spreading at the end of a link depends on the overall dispersion in the fiber link. If one chooses a link made up of positive and negative dispersion coefficients, then by an appropriate choice of the lengths of the positive and negative dispersion fibers, it would be possible to achieve a zero total link dispersion while at the same time maintaining a large local dispersion. This is referred to as dispersion management in fiber optic systems.

Although FWM leads to crosstalk among different wavelength channels in an optical fiber communication system, it can be used for various optical processing functions such as wavelength conversion, high-speed time division multiplexing, pulse compression, etc. For such applications, there is a concerted worldwide effort to develop highly nonlinear fibers with much smaller mode areas and higher nonlinear coefficients. Some of the very novel fibers that have been developed recently include holey fibers, photonic bandgap fibers, or photonic crystal fibers which are very interesting since they possess extremely small mode effective areas ($\sim 2.5\ \mu m^2$ at 1550 nm) and can be designed to have zero dispersion even in the visible region of the spectrum. This is expected to revolutionize nonlinear fiber optics by providing new geometries to achieve highly efficient nonlinear optical processing at lower powers.

Supercontinuum Generation

Supercontinuum (SC) generation is the phenomenon in which a nearly continuous spectrally broadened output is produced through nonlinear effects on high peak power picosecond and subpicosecond pulses. Such broadened spectra find applications in spectroscopy, wavelength characterization of optical components such as a broadband source from which many wavelength channels can be obtained by slicing the spectrum.

Supercontinuum generation in an optical fiber is a very convenient technique since it provides a very broad bandwidth (>200 nm), the intensity levels can be maintained high over long interaction lengths by choosing small mode areas, and the dispersion profile of the fiber can be appropriately designed by varying the transverse refractive index profile of the fiber.

The spectral broadening that takes place in the fiber is attributed to a combination of various third-order effects such as SPM, XPM, FWM, and Raman scattering. Since dispersion plays a significant role in the temporal evolution of the pulse, different dispersion profiles have been used in the literature to achieve broadband SC.

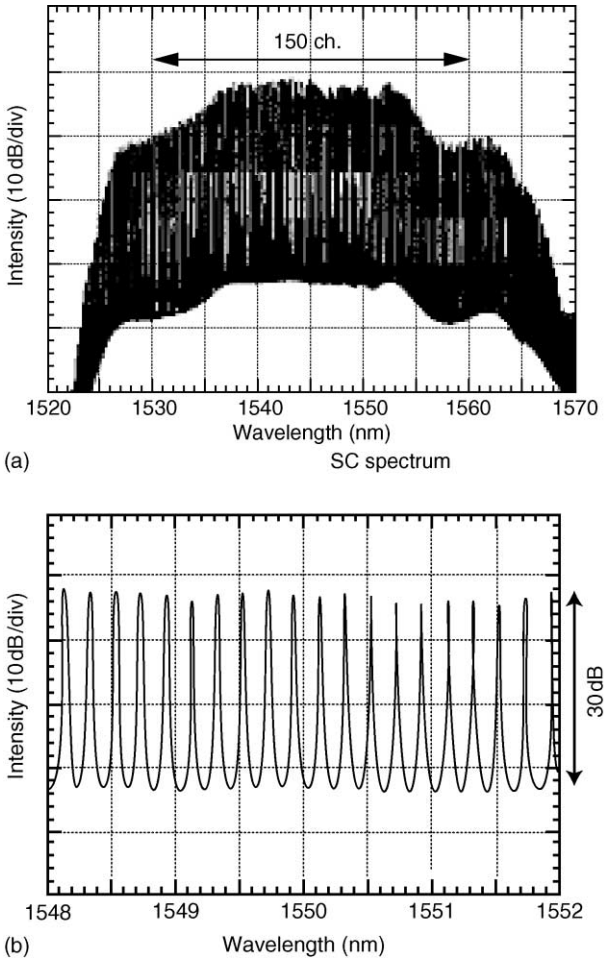


Figure 14 Super Continuum spectrum generated by passing 25 GHz optical pulse train at 1544 nm generated by a mode locked laser diode and amplified by an erbium doped fiber. (Reproduced from Yamada E, Takara H, Ohara T, Satc K, Morioka T, Jinguji K, Itoh M and Ishi M (2001) A high SNR, 150 Ch. Supercontinuum CW optical source with high SNR and precise 25 GHz spacing for 10 Gbit/s DWDM systems. *Electronics Letters* 37: 304–306 with permission from IEEE.)

Some studies have used dispersion decreasing fibers, dispersion flattened fibers, while others have used a constant anamolous dispersion fiber followed by a normal dispersion fiber.

Figure 14 shows the SC spectrum generated by passing 25 GHz optical pulse train at 1544 nm generated by a mode locked laser diode and amplified by an erbium doped fiber. The fiber used for SC generation is a polarization maintaining dispersion flattened fiber. The output is a broad spectrum containing more than 150 spectral components at a spacing of 25 GHz with a flat top spectrum over 18 nm. The output optical powers range from −9 dBm to +3 dBm. Such sources are being investigated as attractive solutions for dense WDM (DWDM) optical fiber communication systems.

Conclusions

The advent of lasers provided us with an intense source of light and led to the birth of nonlinear optics. Nonlinear optical phenomena exhibit a rich variety of effects and coupled with optical waveguides and fibers, such effects can be observed even at moderate optical powers. With the realization of compact femtosecond lasers, highly nonlinear holey optical fibers, quasi phase matching techniques, etc, the field of nonlinear optics is expected to grow further and lead to commercial exploitation such as in compact blue lasers, soliton laser sources and multiwavelength sources and for fiber optic communication.

List of Units and Nomenclature

Attenuation	The decrease in optical power of a propagating mode usually expressed in dB/km
Birefringent medium	A medium characterized by two different modes of propagation characterized by two different polarization states of a plane light wave along any direction
Cross-phase modulation (XPM)	The phase modulation of a propagating light wave due to the nonlinear change in refractive index of the medium brought about by another propagating light wave
Effective length (L_{eff})	The length over which most of the nonlinear effect is accumulated in a propagating light beam
Four-wave mixing (FWM)	The mixing of four propagating light waves due to intensity dependent refractive index of the medium. This mixing leads to the generation of new frequencies from a set of two or three input light beams
Mode effective area (A_{eff})	The cross sectional area of the mode which determines the nonlinear effects on the propagating mode. This is different from the core area
Nonlinear polarization (P_{NL})	When the response of the medium to an applied electric field is not proportional to the applied electric field, then this results in a nonlinear polarization
Optical waveguides	Devices which are capable of guiding optical beams by overcoming diffraction effects using

	the phenomenon of total internal reflection
Phase coherence length	The minimum crystal length upto which the second harmonic power generated by the nonlinear interaction increases
Phase matching	Condition under with the nonlinear polarization generating an electromagnetic wave propagates at the same phase velocity as the generated electromagnetic wave
Pulse dispersion	The broadening of a light pulse as it propagates in an optical fiber
Quasi-phase matching (QPM)	Periodic modulation of the nonlinear coefficient to achieve efficient nonlinear interaction
Second-harmonic generation (SHG)	The generation of a wave of frequency 2ω from an incident wave of frequency ω through nonlinear interaction in a medium
Self-phase modulation (SPM)	The modulation of the phase of a propagating light wave due to the nonlinear change in refractive index of the medium brought about by itself
Supercontinuum generation (SC)	The phenomenon in which a nearly contiunuous spectrally broadband output is produced through nonlinear effects on high peak power picosecond and sub-picosecond pulses
Walk off length (L_{wo})	Length of the fiber required for two interacting pulses at different wavelengths to walk off relative to each other

See also

Nonlinear Optics, Applications: Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation; $\chi^{(3)}$–Third-Harmonic Generation; Four-Wave Mixing. **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Agrawal GP (1989) *Nonlinear Fiber Optics*. Boston, MA: Academic Press.

Armstrong JA, Bloembergen N, Ducuing J and Pershan PS (1962) Interactions between light waves in a nonlinear dielectric. *Physics Review* 127: 1918.

Baldeck PL, Alfano RR and Agrawal GP (1998) Induced frequency shift of copropagating ultrafast optical pulses. *Applied Physics Letters* 52: 1939–1941.

Baldi P (1992) *Generation de fluorescence parametrique guidee sur niobate et tantalite de lithium polarizes periodiquement. Etudes preliminaire d'un oscillateur parmetrique optique integree*. Doctoral Thesis of the University of Nice, France.

Chiang TK, Kagi N, Marhic ME and Kazovsky LG (1996) Cross phase modulation in fiber links with multiple optical amplifiers and dispersion compensators. *Journal of Lightwave Technology* 14: 249–259.

Franken PA, Hill AE, Peter CW and Weinreich G (1961) Generation of optical harmonics. *Physics Review Letters* 7: 118.

Ghatak AK and Thyagarajan K (1987) *Optical Electronics*. Cambridge, UK: Cambridge University Press.

Ghatak AK and Thyagarajan K (1998) *Introduction to Fiber Optics*. Cambridge, UK: Cambridge University Press.

Lim EJ, Fejer MM and Byer RL (1989) Second-harmonic generation green light in periodically poled planar lithium niobate waveguide. *Electronic Letters* 25: 174.

Lim EJ, Fejer MM, Byer RL and Kozovsky WJ (1989) *Electronic Letters* 25: 731–732.

Myers LE and Bosenberg WR (1997) Periodically poled lithium niobate and quasi phase matcher parametric oscillators. *IEEE Journal of Quantum Electronics* 33: 1663–1672.

Newell AC and Moloney JV (1992) *Nonlinear Optics*. Redwood City, CA: Addison Wesley Publishing Co.

Rapp L (1997) Experimental investigation of signal distortions induced by cross phase modulation combined with dispersion. *IEEE Photon Technical Letters* 9: 1592–1595.

Shen YR (1989) *Principles of Nonlinear Optics*. New York: Wiley.

Shtaif M, Eiselt M and Garret LD (2000) Cross-phase modulation distortion measurements in multispan WDM systems. *IEEE Photon Technical Letters* 12: 88–90.

Tkach RW, Chraplyvy AR, Forghiari F, Gnanck AH and Derosier RM (1995) Four-photon mixing and high speed WDM systems. *Journal of Lightwave Technology* 13: 841–849.

Wang QZ, Lim QD, Lim DH and Alfano R (1994) High resolution spectra of self phase modulation in optical fibers. *Journal of Optical Society of America* B-11: 1084.

Webjorn J, Siala S, Nan DW, Waarts RG and Lang RJ (1997) Visible laser sources based frequency doubling in nonlinear waveguides. *IEEE Journal of Quantum Electronics* 33: 1673–1686.

Yamada M, Nada N, Saitoh M and Watanabe K (1993) *Applied Physics Letters* 62: 435–436.

Yamada E, Takara H, Ohara T, Sato K, Morioka T, Jinguji K, Itoh M and Ishi M (2001) A high SNR, 150 Ch. Supercontinuum CW optical source with high SNR and precise 25 GHz spacing for 10 Gbit/s DWDM systems. *Electronics Letters* 37: 304–306.

Yariv A (1997) *Optical Electronics in Modern Communications*, 5th ed. New York: Oxford University Press.

Optical Fiber Cables

G Galliano, Telecom Italia Lab, Torino, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In order to be put to practical use in the telecommunication network, optical fibers must be protected from environmental and mechanical stresses which can change their transmission characteristics and reliability.

The fibers, during manufacturing, are individually protected with a thin plastic layer (extruded during the drawing) to preserve directly their characteristics from damage due to the environment and handling. This coating surrounding the fiber cladding (with an external diameter of 125 μm) is known as the primary coating, which is obtained with a double layer of UV resin for a maximum external diameter of about 250 μm.

Primary coated fiber, at the end of the manufacturing process, is subjected to a dynamic traction test (proof-test) to monitor its mechanical strength. This test, carried out on all fibers manufactured, consists of the application of a well-defined strain (from 0.5% to 1%) for a fixed time (normally 1 s) and thus permits the exclusion of, from successive cabling, those fibers with poor mechanical characteristics.

At the end of the manufacturing process the optical fibers are not directly usable. It is necessary to surround the single fiber (or groups of fibers) with a structure which provides protection from mechanical and chemical hazards and allows for stresses that may be applied during cable manufacture, installation, and service.

Suitable protection is provided using fiber conditioning within the cable (secondary coating or loose cabling), while entire protection is obtained with some cable elements, such as strength members, core assembling, filling, and outer protections. Moreover, the cable structure and its design is heavily influenced by the application (e.g., number of fibers), installation (laying and splicing), and environmental conditions of service (underground, aerial, or submarine).

In this article, the main characteristics of optical cables (with both multimode or single mode fibers) are described and analyzed, together with suitable solutions.

Secondary Fiber Coating

The fiber, before cabling, can be protected with a secondary coating (a tight jacketing) to preserve it during cable stranding. However, fiber with only a primary coating may be cabled with a suitable cable protection.

In the first case the fiber can be directly stranded to form a cable core; in the second case, a single group of fibers must be inserted into a loose structure (tubes or grooves) with a proper extra-length. These two solutions are not always distinguishable; in fact, tight fibers inserted into loose tubes or grooves are often used.

The implementation of the secondary coating and the choice of proper materials requires particular attention, in order to avoid microbending effects on the fiber, which can cause degradation of the transmission characteristics. Microbending is caused by the pressure of the fiber on a microrough surface, or by the fiber buckling, due to the contraction of the structure (e.g., secondary coating) containing it. Microbending causes an increase in fiber attenuation at long wavelengths (1300 and 1550 nm) for both single and multimode fibers, but is particularly emphasized in multimode fibers.

Tight Jacket

In a tight protection, the primary coating fiber is surrounded by a second plastic jacket in contact with the fiber. This jacket mainly protects the fiber from lateral stresses that can cause the microbending.

The fiber can be individually protected with a single or double layer of plastic material or in a ribbon structure. In Figure 1 four different types of tight jacket are shown. In Figure 1a and b the fibers are singularly protected with a single or double layer up to an external diameter of 0.9 mm. The second type is normally preferred because the external hard plastic gives a good fiber protection and the inner soft plastic avoids microbending effects on the fiber.

In a ribbon structure, from 4 to 16 primary coated fibers are laid close and parallel and covered with plastic. Two types of ribbon are normally classified 'encapsulated' (Figure 1c) or 'edge bonded' (Figure 1d). In the first case, the fibers are assembled with a common secondary coating (one or two layers), in the second case, an adhesive is used to stick together the fibers with only the primary coating.

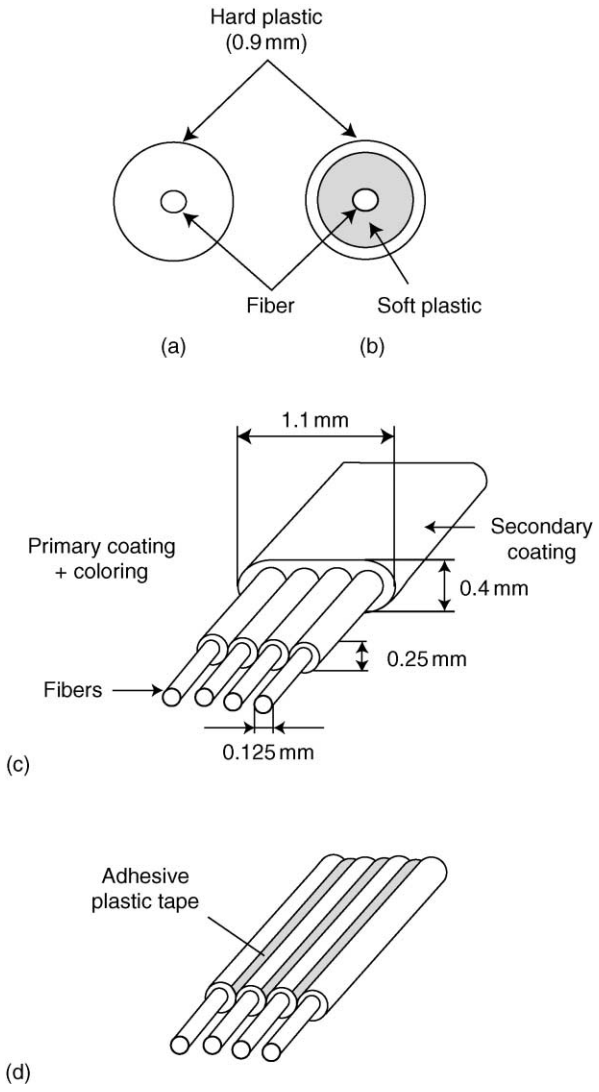


Figure 1 Different types of tight coatings.

Care should be taken in applying tight protection, in particular for a ribbon structure, to avoid an attenuation increase due to the action of the jacket on the fiber (in particular for the thermal contraction of the coating materials). The ribbon structure is designed to permit the simultaneous junction of multiple fibers; for this reason the ribbon geometry must be precision controlled.

Loose Jacket

In a loose protection the fiber can be inserted into a cable with only the primary coating. A thin layer of colored plastic may be added onto the primary coating to identify the fiber in the cable. The fibers are arranged in a tube or in a groove on a plastic core (slotted core). The diameter of the tube or the dimension of the core must be far larger than the diameter of the fiber so that the fiber is completely free inside the structure. The fibers can be inserted into a tube or into a slotted core, singularly or in groups. The space inside the tube around the fibers can be empty or filled with a suitable jelly. Some examples of loose structures are shown in Figure 2. In the tube structure, the external diameter of the jacket depends on the number of the fibers: for a single fiber up to 2 mm (Figure 2a), and for a multifiber tube (Figure 2b) up to 3 mm (for 10 fibers).

Examples of slotted core structures are shown in Figure 2c and d, respectively, for one fiber and for a group of fibers, the dimensions of grooves depending on the number of fibers.

Plastic materials with a high Young modulus are normally used for the tube that is extruded on the single or on the group of fibers. Instead, for the slotted core, plastic materials with good extrusion are used to obtain a precise profile of the grooves.

The loose structure guarantees freedom to the fibers in a defined interval of elongation and compression. Out of this interval, microbending effects can arise which increase fiber attenuation. Axial compression generally occurs when the cable is subjected to low temperatures and is generated from the thermal contraction of the cable materials that is different from that of the fibers. Axial elongation normally occurs when the cable is pulled or by the effects of high temperature.

For a correct design of a cable with a loose structure, the extra length of the fiber with respect to the tube or groove dimension and the stranding pitch must be evaluated, starting from the range of axial compression and elongation. In fact, care must be taken to prevent the fiber from being strained or forced against the tube or groove walls in the designed range.

For the tube solution, the extra length of the fibers is controlled during the tube extrusion; for the slotted core solution, the length is controlled during cabling, giving a controlled tension to the slotted core. The loose structure is often used for the cabling of fibers with tight jacketing, joining the advantages of the tight protection and loose cable. This solution is typical for ribbon cables.

Fiber Identification

To identify the fibers in a cable, a proper coloration with defined color codes must be given to the single fibers. In the tight protection, the tight buffer is directly colored during extrusion. In the loose protection, a thin colored layer is added on the primary coating of the fiber; tubes or slotted core may

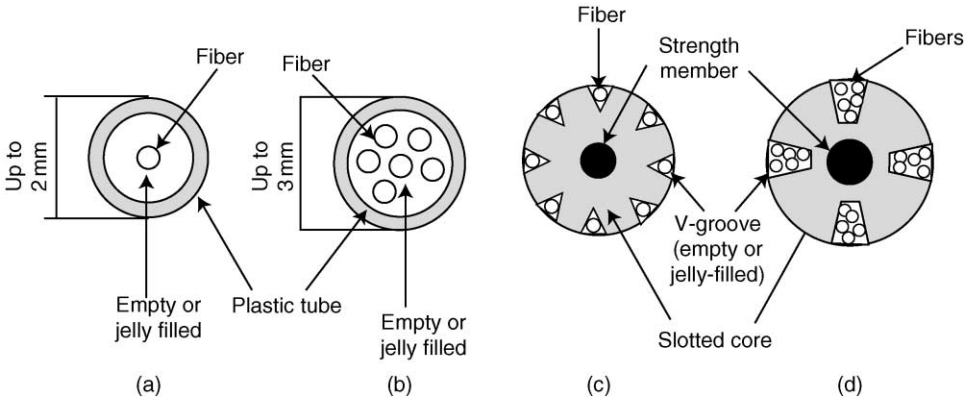


Figure 2 Different types of loose coatings.

be also colored to identify groups of fibers in high potentiality cables.

Cable Components

The fundamental components of an optical cable are: the cable core with the optical fibers, the strength member (one or more) which provide the mechanical strength of the cable, sheaths which provide the external protection, and filling materials.

Cable Core

The optical fibers with the secondary coating (tight or loose) are rejoined together in a cable core. For tight fibers or loose tubes, the cable core is obtained by stranding the fibers or the tubes around central elements that normally act as strength members too (see below).

The stranding pitch must be long enough to avoid excessive bending of the fibers and short enough to guarantee stress apportionment among the fibers in cable bends and maximum elongation of the cable without fiber strain. Normally, a helical stranding is used, but often SO stranding, that consists in an inversion of stranding direction every three of more turns, is employed. The second stranding type makes the cabling easier but may reduce its performance.

The unit core or the slotted core are made of plastic and normally incorporate a central strength member.

Strength Member

The strength member may be defined as an element that provides the mechanical strength of the cable, to withstand elongation and contraction during the installation and to ensure thermal stability. Fiber elongation, when the cable is pulled or the fiber compressed due to low temperatures, depends greatly

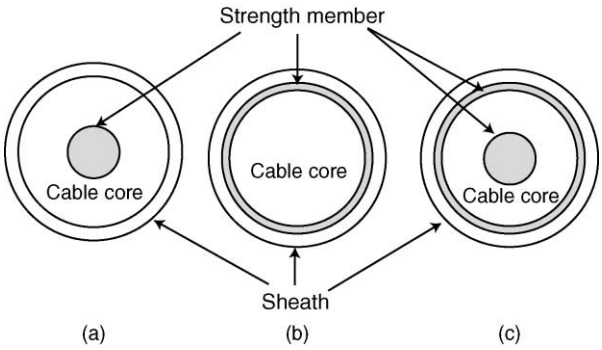


Figure 3 Different locations of the strength member in an optical cable.

on the characteristics of the strength member inserted in the cable.

As the maximum elongation values suggested for the fibers are in the range from 20 to 30% of the proof-test value, the maximum elongation allowed for a tight cable is ~0.1% instead of ~0.3% for a loose cable. To warrant these performances, a material with a high Young modulus is normally used as a strength member. Also, the strength member must be light (to avoid excessive cable weight) and flexible: these properties make it easier when laying the cable in ducts.

The strength member may be metallic or non-metallic and may be inserted in the core (Figure 3a) or in the outer part of the cable (Figure 3b), normally under the external sheath. Often two strength members are placed in an optical cable: the first one in the core and the second one under the sheath (Figure 3c).

Metallic strength members, generally one or more wires of steel, are inserted in the center or in the outer part of the cable. The steel, for its low thermal expansion coefficient and high Young modulus, behaves as a good strength member, particularly

when it is arranged in the central core. Steel is a nonexpensive material, but needs protection against corrosion and electrical-induced voltages, such as by lightning or current flow in the ground.

If cables with high flexibility are requested, steel is not suitable and aramid fiber yarn or glass-fiber reinforced plastic (G-FRP) are used, particularly when the strength member is placed in the outer part of the cable. Glass fiber elements are generally used as a compressive strength member in the central cable core to avoid contraction at low temperatures, but they may be used as strength elements when the traction strength is low.

Aramid yarns are not active as resisting compression elements but are normally arranged in the outer part of the optical cable. Aramid yarns have a high Young modulus (as with steel) and low specific weight: in commercial form they consist of a large number of filaments gathered into a layer. Normally two layers stranded in opposite directions are used to avoid cable torsion during the installation.

Metal-free cables are protected by G-FRP and aramid yarns. This solution results in optical cables which are insensitive to electrical interferences.

Sheaths

The functions of the sheaths are to protect an optical cable during its life, in particular from mechanical and environmental stresses. During the installation or when the cable is in service, it may be subjected to mechanical stresses, for example, presence of humidity or water and chemical agents. The sheaths act as a barrier to avoid degradations of the fibers in the internal cable core.

The plastic sheaths (one or more, depending on the cable type and structure) are extruded over the core. This extrusion is a delicate operation in cable manufacturing because the cable core may be subjected to tension, while during the cooling the thermal compression of the plastic produces a stress on all the cable elements. Examples of cable sheaths are shown in Figure 4.

Different plastic materials, such as polyvinylchloride (PVC), polyethylene, and polyurethane are normally used. PVC has excellent mechanical properties, flexibility, and is flame retardant but is permeable to humidity. On the other hand, high-density polyethylene has a permeability lower than PVC, with good mechanical properties and low coefficient of friction, but it is more flammable than PVC and less flexible. Finally, polyurethane, for its softness and high flexibility, is often used for inner sheaths. When low toxicity is requested (e.g., in indoor cables), halogen-free materials are employed.

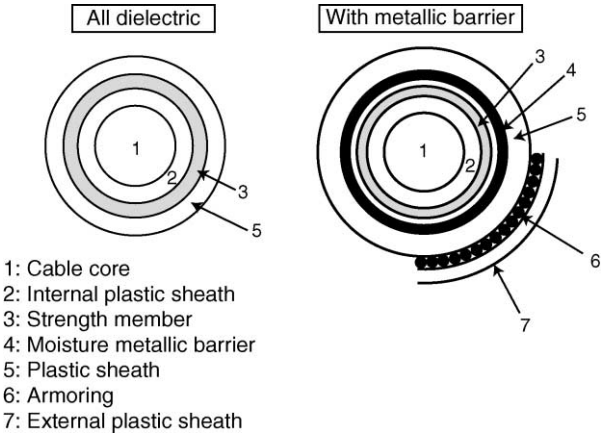


Figure 4 Examples of optical cable sheaths.

Often a metallic barrier is used under the external sheath to prevent moisture entering the cable core. The type most widely used is an aluminum ribbon (of a few tenths of a millimeter) bonded into the external polyethylene sheath, during the sheath extrusion. This structure is normally indicated as a LAP (laminated aluminum polyethylene) sheath.

When the cable is directly buried, a metallic armor (corrugated steel tape or armoring wires) is generally inserted. The metallic armor protects the cable core against radial stresses but it is also used as protection against rodents and insects.

Filling Compounds and Other Components

Fillings are used in optical cables to avoid the presence of moisture and water propagation in the cable core should a failure occur in the cable sheath. Suitable jelly compounds, with constant viscosity at a wide range of temperatures and chemical stabilities, are generally inserted in the cable core and, sometimes, in the outer protection layer. In loose structures, the jelly is in direct contact with the fibers: the compound must be compatible with them and guarantee the fiber freedom. Special compounds may be used to adsorb and permanently fix any hydrogen present in the cable core. The presence of hydrogen gas, that may be generated from the internal materials of the cable (mainly metals), is harmful to the fibers because it causes permanent attenuation increases and thus damage to them.

With the generic terminology 'other components', some other components used in the optical cables are included. They are often similar to those used in conventional cables. The most important are: insulated conductors, plastic fillers, and cushion layers and tapes around the fiber core.

Insulated copper conductors may be incorporated in the cable core, for example, to carry power supply

or as service channels, or to detect the presence of moisture in the cable (in this case, the insulation is missing or partial). The conductors may cause problems due to the voltages induced by power lines or lighting.

Plastic fillers are often used to complete the cable core geometry. Cushion layers are used to protect the cable core against radial compression. Normally, they are based on plastic materials wound around the core. Tapes are wound around the cable core to hold the assemblies together and secondly to provide a heat barrier during the sheath extrusion. Mylar tapes are usually employed.

Optical Cable Structures, Types, and Applications

Cable Structures

The structure of an optical cable is strictly dependent on the construction method and the application type that determines the design and the grade of external protection. The main core (or inner) structures of an optical cable can be classified as: stranded structures (tight and loose); slotted core cable; or ribbon cable. In this section, a few examples of cable structures are presented. Below, separate sections are devoted to submarine cables and to special application cables, for their particular structures and features.

In the stranded tight structure, a low number of tight or loose fibers is stranded around a strength member to form a cable unit. Some of this cable unit may be joined to constitute a cable with a medium/high number of fibers (Figure 5a). Alternatively the fibers may be arranged in a multilayer structure (Figure 5b).

The main advantages of a tight stranded structure are the small dimensions of the cable core, even when a large number of fibers are involved, and also the facility of handling. Care must be taken in the design of the strength member: in fact, the fibers are immediately subjected to tension when the cable is pulled into the ducts or compressed when subjected to low temperatures.

In the stranded loose structures, two categories of cables are defined, one fiber per tube and a group of fibers (up to 20) per tube (Figure 5c). As already mentioned, in the loose structures, the fibers are free inside the tube and for this reason the cable is less critical than the tight cable. The external dimensions of loose cables are larger than tight cables, but the solution with multiform tubes allows for cables with a high number of fibers in relatively small dimensions.

Slotted core cables (Figure 2c,d) contain fibers with only the primary coating and may be divided, as in the stranded loose structure, into one fiber per groove or into a group of fibers per groove. The cable design and the choice of a suitable plastic material for the slotted core joined to a proper strength member, make it possible to obtain cables with fiber stability at a wide range of temperature. As in the loose stranded solutions, the slotted core solution is adopted to obtain cables with a high number of fibers in small dimensions.

In the ribbon solution, the fibers (from 4 to 16) are placed in a linear ribbon array. Many ribbons are stacked together to constitute the optical unit. The ribbon cable normally has a loose structure (loose tube or slotted core). Three structures are shown in Figure 6. In the first example (classical AT & T cable), 12 ribbons of 12 fibers are directly stacked and stranded in a tube (Figure 6a). In the second and third examples, ribbons are inserted into tubes or grooves (Figure 6b,c). Several tubes or slotted cores can be assembled together to give a cable with very high fiber density. The ribbon approach may be efficient when a large number of fibers must be

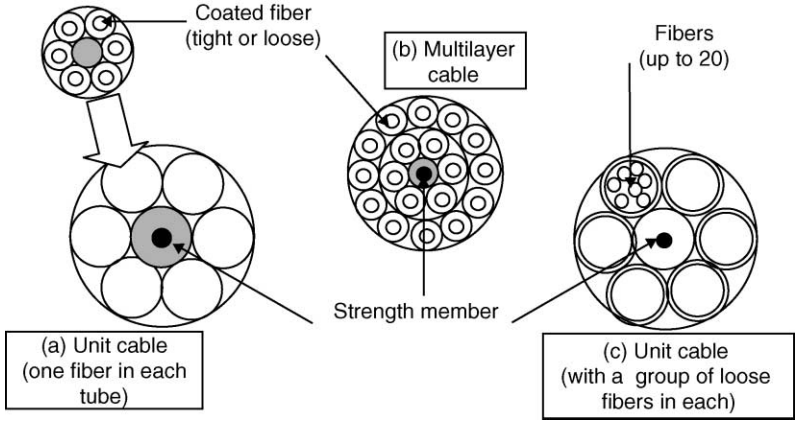


Figure 5 Structures of stranded core cables.

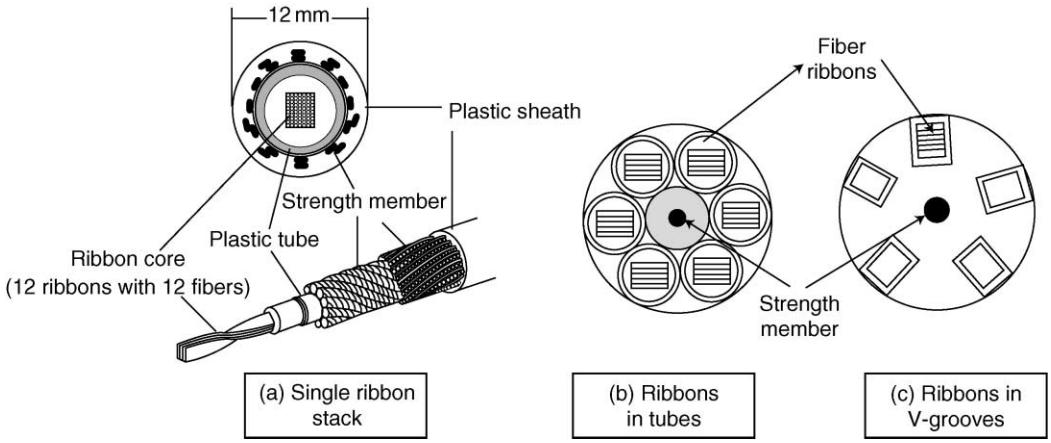


Figure 6 Structures of ribbon cables.

simultaneously handled, spliced, and terminated at connectors.

Types and Applications

The choice of an optical cable is heavily established by many factors, but mainly on the type of installation. On this basis, some different typology of cables may be identified: aerial cables, duct cables, direct buried cables, indoor cables, underwater cables, etc.

Aerial cables can be clasped onto a metal strand or can be self supporting. In the first case, the cable is not subjected to a particular tension but requires good thermal and mechanical performances. This solution is chosen when the cable must withstand strong ice and wind and when long distances between the poles are required. All the structures mentioned above may be adopted. In the second case, the cable is normally exposed to high mechanical stresses during its life and high tensile strength must be guaranted to maintain the fiber elongation at a safety level. Generally, loose structures are preferred for their greater fiber freedom. Often, in self-supporting cables, the strength member is external to the cable core. An example is shown in Figure 7.

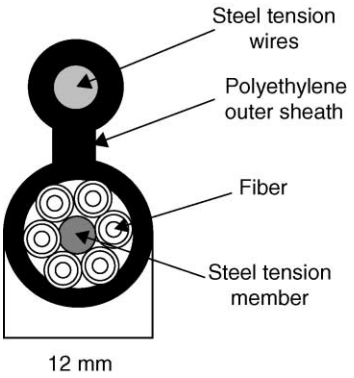


Figure 7 Self-supporting aerial cable.

Cables pulled in ducts must be resistant to pulling and torsion forces, and light and flexible to permit the installation of long sections. The cable must also protect the fibers against water and moisture that may be present in ducts or manholes. Usually the cable is filled with a jelly compound, a metallic barrier is normally used and, often, an armoring is added when protection against rodents is necessary. All the structures described above may be employed in duct cables.

The characteristics of direct buried cables are very similar to those of duct cables, but additional protections, such as metallic armoring, are required to avoid the risk of damages from digging and other earthmoving work.

Cables for indoor applications normally require a smaller number of fibers. Their main characteristics are: small dimensions, flexibility, small curvature radius, and ease in handling and jointing. For indoor applications, the sheath must not permit flame propagation and must have a low emission of toxic gases and dark smoke. Tight cables are preferably used for indoor applications due to their compactness and small dimensions.

Underwater cables are employed in the crossing of rivers, in lakes and lagoons, and for shallow water applications in general. Such cables must have stringent requirements such as resistance to moisture and water penetration and propagation, pulling resistance (during the installation, and recovery in event of failure), and high resistance to static pressure and core structure compatible with land cables (when it is part of a ground link). In addition, due to the presence of metallic structures, attention should be paid to hydrogen effects.

Submarine Cables

Submarine cable is manufactured for laying in deep-sea conditions. In designing a submarine cable, it is necessary to provide high reliability that the mechanical and transmission characteristics of the optical fibers will be stable over a long period of time. Different cables are used for submarine applications, but generally follow the basic scheme of Figure 8.

The central core (with a few fibers – up to 12 – in a slotted core or in a tight structure) is surrounded by a double layer of steel wires that act as strength members against tensile action and water pressure. Metal pipes at the inner or outer sides of the strength member, act as a water barrier and as power supply conductors (for the supply of the undersea regenerators).

The outer polyethylene insulating sheath is the ultimate protection for the ordinary deep-sea cable. For cables requiring special protection, steel wire armoring (anti-attack from fishes), and further polyethylene are added.

The cable must have stringent mechanical requirements, such as resistance to traction, torsion, crushing, impact, and to shark attack. The cable must be suitable for installation using standard cable-laying ships. Furthermore the cable materials must have low content of hydrogen and emissions. If necessary, a hydrogen absorber may be included in the cable core.

Special Cables

This category comprises cables not specifically used in ordinary telecommunication networks, but cables used for specific applications and according to specified requirements. Cables for military use are normally employed for temporary plant restoration. They must be light but crush resistant and with good mechanical characteristics. Cables for mobile applications (robots, elevators, aircrafts, submarines, etc.) generally require high flexibility and tensile strength. Special cables may be installed in particular environments, where they must be insensitive, for instance, to nuclear radiation, chemical agents, and high temperatures.

Optical cables can be installed in power lines, together with the power conductors, or into the ground wires of high-voltage overhead power lines. In this last type of cable, the central stranded steel wire, making up the ground wire, is replaced with a metal tube containing the optical unit (normally a small groups of optical fibers inserted in a tube or in a slotted core).

Blown Fiber

This technology consists of the installation of a single fiber or of small fiber bundles (or cables) into pre-installed tubes or ducts using a flux of compressed air. Initially, cables with empty tubes are installed and successively, at the time of need, the tubes can be filled by blowing fibers (singly, or in bundles) or cables. The installation time is normally fast; some compressors can blow tens of meters in a minute. The typical maximum blowing distance is about 1 km for horizontal planes, but it is reduced

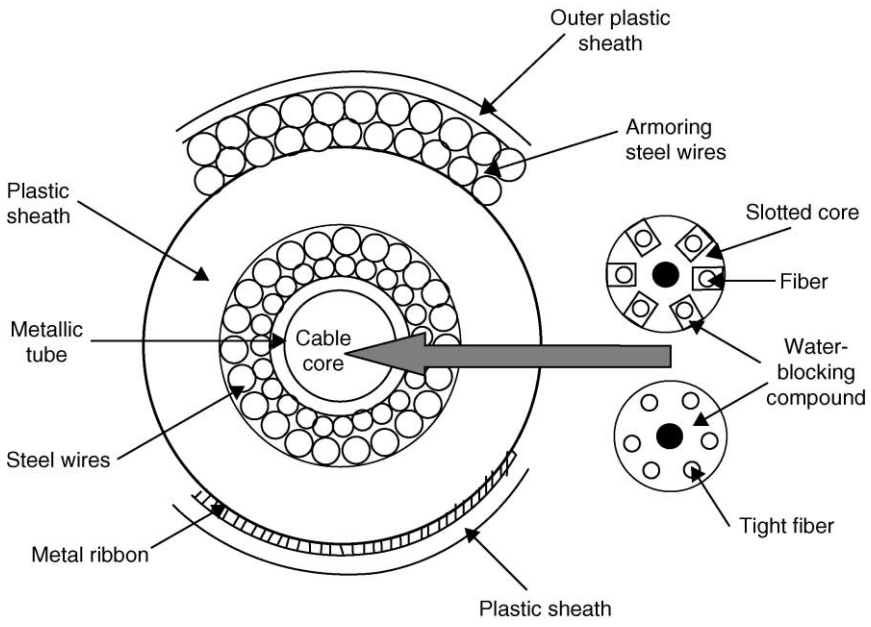


Figure 8 Basic structure of a submarine cable.

along winding roads. The advantage of this technique is to produce variable and flexible structures, increasing the number of fibers to fit the network need. Besides, substitutions or change of fibers can be quickly carried out without substitution of the cable.

See also

Fiber and Guided Wave Optics: Dispersion. **Light Emitting Diodes**.

Further Reading

Griffioen W (1989) The installation of conventional fiber-optic cables in conduits using the viscous flow of air. *Journal of Lightwave Technology* 7(2): 297–302.

Hughes H (1997) *Telecommunications Cables. Design, Manufacture and Installation*. : John Wiley and Sons.

Keiser G (1986) *Optical Fiber Communications*. McGraw-Hill.

Murata H (1988) *Handbook of Optical Fibers and Cables*. Marcel Dekker.

Technical Staff of CSELT (1990) *Fiber Optic Communications Handbook Second Edition*. : TAB Books.

Passive Optical Components

D Suino, Telecom Italia Lab, Torino, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Modern telecommunications networks are based on fiber optics as signal transporters. In these networks the signals travel in the form of light confined within the fibers. As light power travels through the fibers from the source apparatus to the receiver, a number of components are employed to manage the optical signal. All other components that are not used to manage the signal are called 'passive optical components'. They serve several functions, such as to join two different fibers to distribute the signal onto more optical branches to limit the optical power to select particular wavelengths, and so on.

These components can be located anywhere in the network, depending on their function and on the network architecture. However, their main location is near the apparatus in which it is necessary to manage the signal before it is sent through the network or at the junction in the network where it is sent on to the receiver.

The main types of passive components that are found in a network can be grouped as follows:

- Joint devices:
  - Optical connectors;
  - Mechanical splices.
- Branching devices:
  - Nonwavelength-selective branching devices;
  - WDM.
- Attenuators;
- Filters;
- Isolators;
- Circulators.

Optical fibers are also divided into several groups (multimode 50/125, multimode 62.5/125, singlemode, dispersion shifted, NZD, and so on). There are also components of different typology that match up with each particular fiber. However, the general concept in terms of functionality, technology, and characteristic parameters, are the same among passive optical components of the same type (connectors, branching devices, or attenuators) but of different groups (single-mode, multimode, etc.).

Optical Joint Devices

The function of an optical joint is to perform a junction between two fibers, giving optical continuity to the line. Two different classes of joint can be considered: the connectors and the splices. The joints made by using connectors, are flexible points in the network, which means they can be opened and reconnected several times in order, for example, to reconfigure the network plan. However, a splice is a fixed joint and usually cannot be opened and re-mated. There are two type of splice: mechanical splices, holding joints to fibers by glue or mechanical crimp, and fusion splices fusing the two fiber heads with a suitable fusion splicer.

General Considerations

Considering that fibers can be described as pipes, in the core of which flows light, continuity between two fibers means that the two cores must be aligned in a stable and accurate way. Obviously this alignment must be carried out with minimum power loss in the junction. The main factors that

affect the power loss in a joint point between two fibers are:

- x = Lateral offset;
- z = Longitudinal offset;
- θ = Angular misalignment; and
- w_T/w_R = Mode field diameter ratio (in the single mode fiber) or numerical aperture ratio (in the multimode fiber).

These geometrical mismatching between the two fibers is represented in Figure 1.

The function that describes the joint attenuation is different for single-mode and multimode connectors, but the parameters that contribute to it are the same. The general formulation for these functions are complicated and, in particular, for the multimode connector it involves an integral on the shape distribution of the light in the fiber core. These functions become simpler in particular conditions: for example, if we consider a single-mode joint without a gap between the two fibers (that is the more usual condition in the modern connectors) the z parameter is nil, and in this case the function that describes the attenuation of the connector is

$$A(\mathrm{dB}) = -10\log\left\{\left(\frac{2w_T w_R}{w_T^2 + w_R^2}\right)^2 \times \exp\left[-2\frac{(x\lambda)^2 + (\pi n w_T w_R \sin(\vartheta))^2}{\lambda^2(w_T^2 + w_R^2)}\right]\right\} \quad [1]$$

where n is the fiber core refractive index and λ is the wavelength of the light.

Figure 2 shows the trend of eqn [1], where the parameter ranges in the graphs are typical for the most common products on the market. It is evident that the main contribution to attenuation arises from the lateral offset.

For connectors of good quality, both for single-mode and multimode fibers, the attenuation values are in the range of a few tenths of dB (typically less than 0.3 dB). Another important factor for a joint, in particular for connectors used in networks for high speed or analog signals, is the return loss (RL). This is a parameter that gives a measure of the quantity of light back-reflected onto the optical discontinuity of the connector, due to the Fresnel effect. The RL is defined as the ratio in dB between the incident light power (P_{inc}) and the reflected light power (P_{ref}):

$$\mathrm{RL} = -10\log\left(\frac{P_{ref}}{P_{inc}}\right) \quad [2]$$

Fresnel reflection arises when the light passes between two media with different refractive indices. In the case of connectors in which the two fibers are not in contact, the light passes from the silica of the fiber

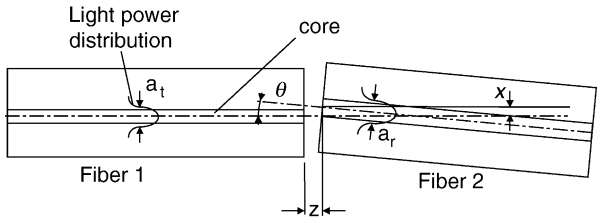


Figure 1 Main parameter that affects the joint power loss.

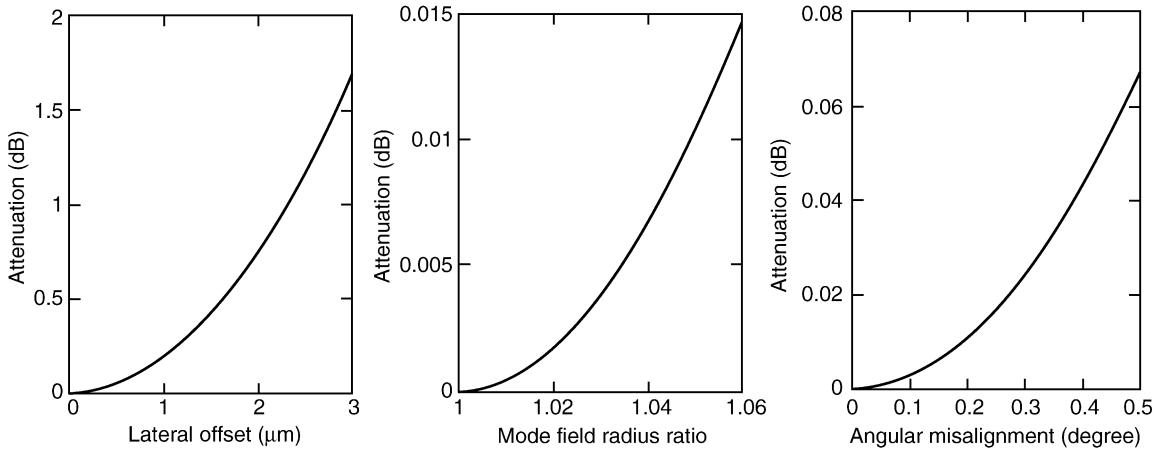


Figure 2 Single mode connector attenuation as a function of the geometrical parameters. Each curve is plotted keeping the other parameter constant.

core to the air in the gap. This causes the reflection of 4% of the incident light that, as RL, is about 14 dB.

To improve the RL performances of an optical joint, it is necessary to reduce the refractive index difference. This can be achieved either by index matching material between the two fibers that reduces the differences between the refractive index of the fiber and the gap, or by performing a physical contact between the fiber heads. The first technique is typically used in mechanical splices but, due to the problem related with the pollution contamination of the index matching, it is not a good solution for connectors that are to be opened and reconnected several times.

The physical contact solution is the most frequently adopted for these connectors. Physical contact is obtained by polishing the ferrule and the fiber end-faces in a convex shape, with the fiber core positioned in the apex (Figure 3). With this technique, the typical values of return loss for a single mode connector are in the range from 35 dB to 55 dB, depending on the polishing grade. The residual back-reflected light is generated in the thin layer of the fiber end surface, because of slight changes in the refractive index due to the polishing process.

Higher return loss values (small quantity of reflected light) are achieved with the angled physical contact (APC) technique. This is obtained by polishing the convex ferrule end face angled with respect to the fiber axis. In this way, the light is not back reflected into the fiber core, but into the cladding, and is thus eliminated (Figure 4). In APC connectors, typical return loss values are greater than 60 dB.

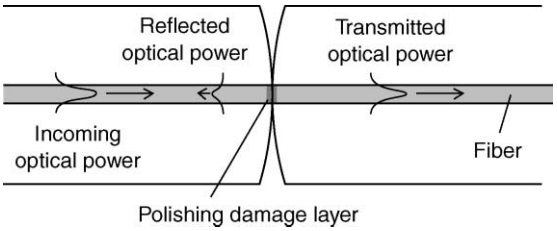


Figure 3 Back reflection effects in PC connectors.

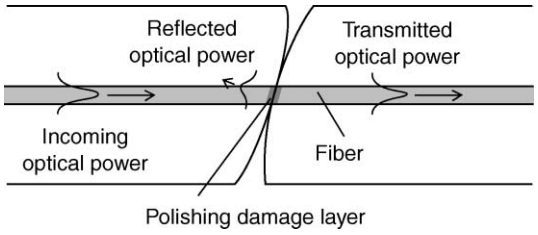


Figure 4 Back reflection effects in APC connectors.

Optical Connectors

On the market are several types of optical connectors. It is possible to divide them into three main groups:

- Connectors with direct alignment of bare fiber;
- Connectors with alignment by means of a ferrule; and
- Multifiber connectors.

Connectors with direct alignment of bare fiber

In this type of connector the bare fibers are aligned on an external structure that is usually a V-groove (Figure 5). Two configurations can be performed: plug and socket or plug–adapter–plug. In the first case, the fiber in the socket is fixed and the one in the plug aligns to it. In the plug–adapter–plug configuration, two identical plugs containing the bare fibers are locked onto an adapter in which the fibers are placed on the aligned structure. These connectors are cheaper than the connectors with a ferrule, (described below), but this is unreliable. In fact, the bare fibers are delicate and brittle and they can be affected by the external mechanical or environmental stresses.

Connectors with alignment by means ferrule

In this type of connector, the fibers are fixed in a secondary alignment structure, usually of cylindrical shape, called a ferrule. The fiber alignment is obtained by inserting the two ferrules containing the fibers into a sleeve. Usually this type of connector is in the plug–adapter–plug configuration (Figure 6). With this technique, the precision of the alignment is moved from the V-groove precision to the dimensional tolerance of the ferrule. Typical tolerance values on the parameters for a cylindrical ferrule with a diameter of 2.5 mm, are in the range of a tenth of a micrometer, and must be verified on the diameter, eccentricity between the hole (through which the fiber is fixed) and cylindricity of the external surface.

Multifiber connectors

In some types of optical cable, the fibers are organized in multiple structures, in which many fibers are glued together to form a ribbon. The ribbon can have

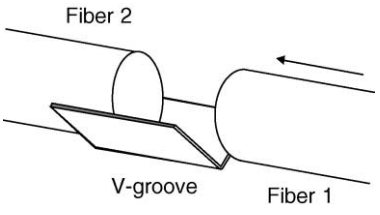


Figure 5 Scheme of the bare fibers alignment on a V-groove.

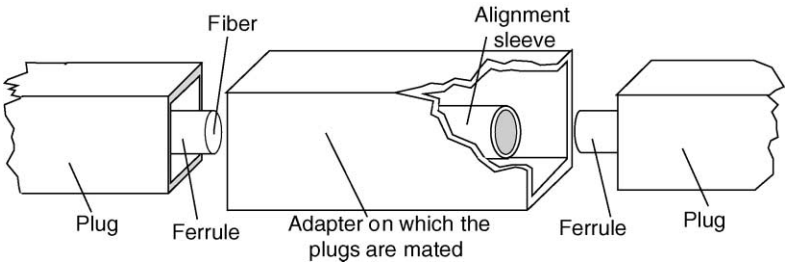


Figure 6 Scheme of a plug–adapter–plug connector with alignment by ferrules and sleeve.

2, 4, 8, 12, or 16 fibers, for example. When these cables are used, it is convenient to maintain the ribbon fiber structure in the connectors (until the point in which it is necessary to use the fibers singularly). Multifiber connectors are typically based on a secondary alignment on a plastic ferrule of rectangular shape in which the fibers maintain the same geometry that they have in the ribbon. The rectangular ferrules of the two plugs are normally aligned by means of two metallic pins placed in two hole in line with the fiber. In Figure 7, a scheme of a multifiber connector is shown.

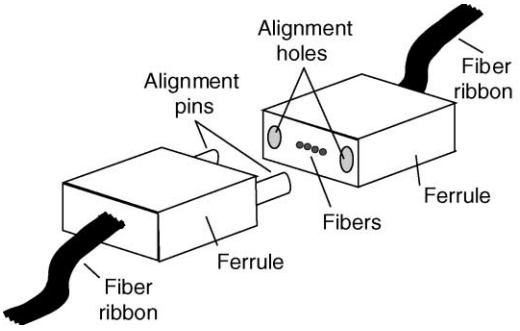


Figure 7 Multifiber connector.

Mechanical Splices

Mechanical splices are yet another way to join together two fibers. All the main considerations for the connectors are true for the mechanical splices. The difference is that the mechanical splice is a fixed joint and, normally, it cannot be opened and re-mated. Mechanical splices, both for single fiber and for multifiber structures, are the most usually manufactured. In a mechanical splice, the fibers are aligned directly in a V-groove and fixed by glue or a mechanical holder.

These components are cheaper than fusion splices but their optical, mechanical, and environmental functions are lower than the fusion splices. Figure 8, shows a mechanical splice for multifibers in which the fibers are crimped on a V-groove array. Usually, to avoid high optical reflection in the junction point, this type of joint is filled with index matching material.

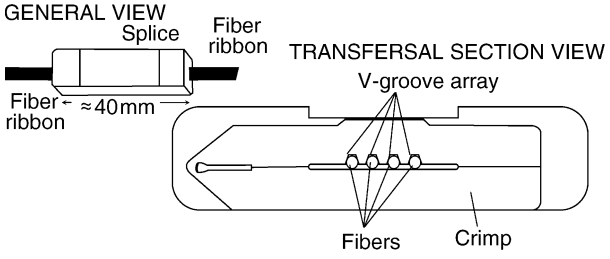


Figure 8 Example of a multifiber mechanical splice.

Branching Devices

Branching devices are passive components devoted to distributing the optical power from an input fiber to two or more fibers. These components can be classified in two broad classes, on the basis of the wavelength dependence. The nonwavelength-selective branching devices, called also couplers or splitters, share the input power to two or more fibers independently from the light wavelength. The components that perform the light distribution on more fibers on the basis of the different wavelengths of the input light are called a WDM (wavelength division (de)multiplexer).

Branching devices have three or more ports, bare fiber or connectors, for the input and/or output of optical power, and share optical power among these ports. A generic branching device can be represented as a multiport device having N input ports and M output ports, as shown in the Figure 9.

The optical characteristics of a branching device can be represented by a square transfer matrix of $n \times n$ coefficients, where n is the total number of the component ports. Each coefficient t_{ij} in the matrix is the fractional optical power transferred among designated ports, that is the ratio of the optical power P_{ij} transferred out of port j with respect to input

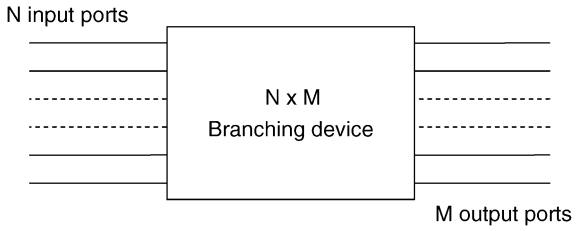


Figure 9 Scheme of a generic N × M branching device.

power P_i into port i:

$$T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & & & \\ \vdots & & t_{ij} & \vdots \\ & & \ddots & \\ t_{n1} & & & t_{nn} \end{bmatrix} \quad [3]$$

where:

$$t_{ij} = \frac{P_{ij}}{P_i} \quad [4]$$

So, according to the relationship between two ports, the coefficients represent several parameters. For example, if the port i is an input port and j is an output port, t_{ij} is the transferred power signal through the component, while if they exist as two isolated ports the coefficients represent the crosstalk between the two ports.

The three main techniques to obtain this type of component are:

(i) All fiber components;
(ii) Micro-optics; and
(iii) Integrated optics.

Nonwavelength-Selective Branching Devices

A nonwavelength-selective branching device is a component having three or more ports, which shares an input signal among the output ports in a predetermined fashion. Usually these components are completely reversible and so they also work as a combiner of signals from more input ports onto a single output port.

There are several types of couplers for different applications:

- Y-coupler with one input port and two output ports;
- X-coupler with two input ports and two output ports; and

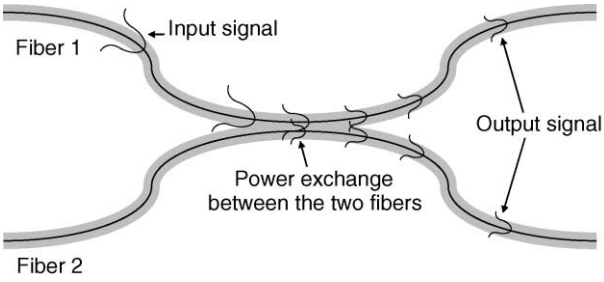


Figure 10 Scheme of a fusion 2 × 2 coupler.

- Star coupler with more than two ports in input and/or in output.

Moreover the couplers can be symmetrical, and in this case they divide the input power light onto n equal output flows, or asymmetrical devices in which the quantity of power light in the output ports is shared in a predefined nonuniform way.

The techniques used to carry out this type of device are mainly the fusion technique or planar technique. The first one is based on the phenomenon of evanescent wave coupling: when the core of two optical guides (as fibers) are located in close proximity, a power exchange from a guide to another is possible. This configuration can be obtained by stretching the fibers under controlled fusion (see Figure 10).

The planar technique, that is the simpler integrated optic application, is based on the optical guides with appropriate shapes being placed directly onto a silicon wafer by means of lithographic processes (Figure 11).

The fusion technique has the advantage of being 'all in fiber' so the component is ready to be inserted into the optical networks by means of usual joint techniques. But for a high number of ports, the coupling ratio among the branches cannot be controlled in a precise way and the cost of the fusion device is proportional to the number of the branches. On the contrary, a device obtained with the planar technology must be coupled with fibers normally used and this operation may be difficult due to the different geometry of the planar guide (rectangular) and the fiber (circular). On the other hand, this technology allows a high precision in the optical characteristics of the component and the cost of the device is independent of the branch number.

WDM

A wavelength-selective branching device, usually called WDM (Wavelength Depending Multiplexer),

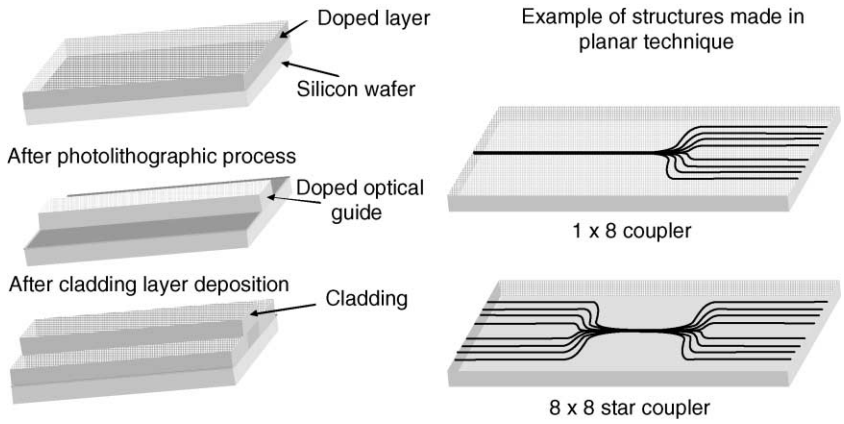


Figure 11 Scheme of the process to obtain a planar coupler and an example of a device.

is a component having three or more ports which shares an input signal among the output ports in a predetermined fashion, depending on the wavelength, so that at least two different wavelength ranges are nominally transferred between two different couples of ports.

WDM devices may be divided in three categories on the basis of the bandwidth of the channels (the spacing between two adjacent wavelength discriminated by the device):

(i) DWDM (Dense WDM) device operates for channel spacing equal or less than 1000 GHz (about 6–8 nm in the range of typical optical wavelength used in telecommunications systems, 1310 or 1550 nm);
(ii) CWDM (Coarse WDM) device operates for channel spacing less than 50 nm and greater than 1000 GHz; and
(iii) WWDM (Wide WDM) device operates for channel spacing equal or greater than 50 nm.

These components are used to combine, at the input side of an optical link in only one fiber, more optical signals with different wavelengths and separate them at the receiving end. This technology allows to send along a fiber, more communication channels, thus increasing the total bit rate transmitted. These types of components can be obtained using filters to select single a wavelength (as a diffractive grating that resolves the light into its monochromatic components) or using the phenomena of evanescent field coupling occurring between two adjacent fiber cores (as described for coupler devices), controlling in a precise way the coupling zone; this is, in fact, dependent on the wavelength (Figure 12).

Using wavelength filters, as they select a singular wavelength, it is necessary to perform a cascade filter structure for separating more than two wavelengths.

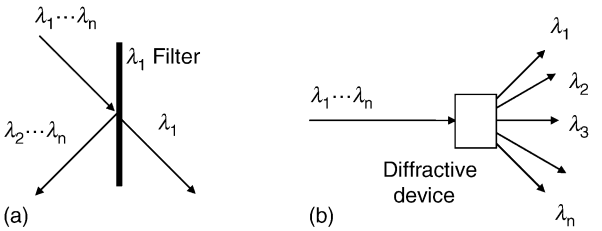


Figure 12 (a) Scheme of the filter technique and (b) diffractive device technique.

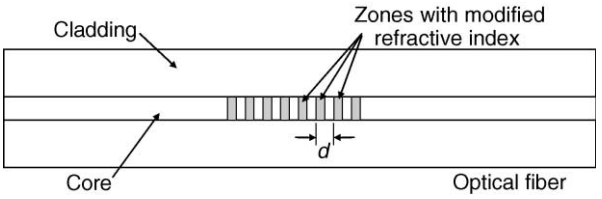


Figure 13 Scheme of FBG.

Micro-optic technologies are generally used for multimode fiber due to the critical collimation problem. For the single-mode fiber the evanescent field coupling technique is applied directly to fiber or to planar optical guides.

Another efficient technique to create WDM devices is with a fiber Bragg grating (FBG) wavelength filter directly applied to the fiber. FBGs consist in periodic modulation (see Figure 13) of the refractive index along the core of the fiber, and they act as reflection filters.

During the fabrication process (see Figure 14) the core of the optical fiber is exposed to UV radiation, and this exposure increases the refractive index of the core in defined zones, with a periodic shape obtained by the interference fringe created by crossing two coherent beams.

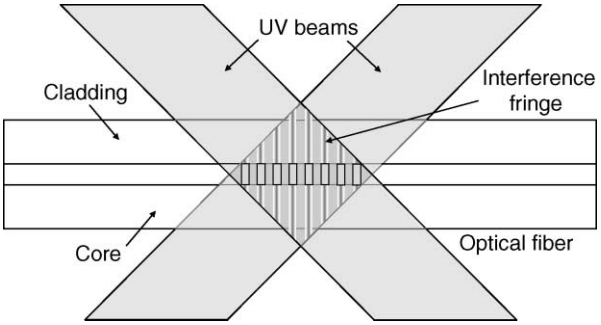


Figure 14 Scheme of FBG production.

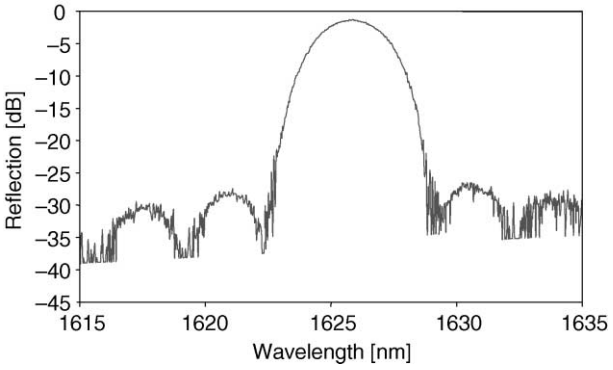


Figure 15 Transmission characteristic of FBG.

Figure 15 shows a representative spectral curve of a FBG filter centered at 1625 nm. The central wavelength, the width, and the value of reflection depend on the grating period (d) and on the deposited UV energy in the fiber core.

Other Passive Optical Components

In an optical network, besides the devices described above, there are a number of other passive optical components that perform particular functions on the optical signal; the main ones are:

Attenuators These are devices used to attenuate the optical power, for example, in front of a receiver when the optical line is short and the power of the received signal is still too high for the detector, or when it is necessary to equalize the signals arriving from different optical lines. It exists both in tunable and fixed versions. The first ones allow adjustment of the attenuation value, however, the second ones have a fixed and predefined value. The fixed optical attenuators can be obtained by inserting a filter or stressing a fiber in a controlled and permanent way. The tunable attenuators are based on variable optical parameters as, for example, the distance or the lateral offset in an optical connection. They exist both as connect lengths of fiber (attenuated optical jumpers) or as compact components similar to a double connector (plug stile attenuator).

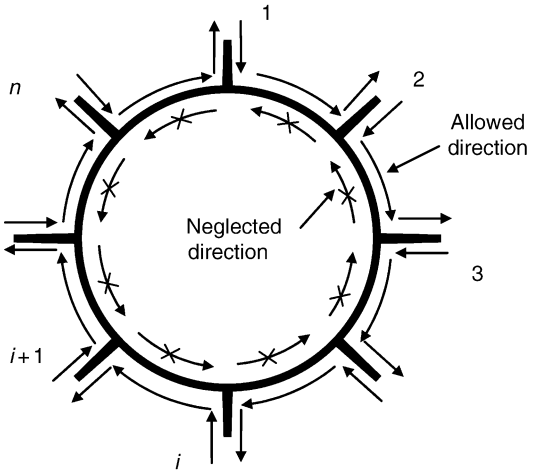


Figure 16 Principle scheme of an optical circulator.

Filters These are components with two ports employed to select or to filter some fixed wavelengths. They can be divided into the following categories:

- short-wave pass (only wavelengths lower than or equal to a specified value are passed);
- long-wave pass (only wavelengths greater than or equal to a specified value are passed);
- band-pass (only an optical window is allowed); and
- notch (only an optical window is inhibited).

It is also possible to have a combination of the above categories. The optical filtering can be obtained by inserting one or more filtering devices (as interferential films) into a fiber or creating a filter directly on the fiber by the FBG described above.

Isolators These are nonreciprocal two-port optical device in which the light flows in the forward direction, and not in the reverse one. The isolators are used to suppress

backward reflections along an optical fiber transmission line, while having minimum insertion loss in the forward direction. Fiber optic isolators are commonly used to avoid back reflections into laser diodes and optical amplifiers, which can make the laser and amplifiers oscillate unstably, and cause noise in the fiber optic transmission system.

Circulators These are passive components having three or more ports numbered sequentially $(1, 2, \ldots, n)$ through which optical power is transmitted nonreciprocally from port 1 to port 2,..., from port (i) to port $(i+1), \ldots$ and from port n to port 1 (Figure 16).

See also

Diffraction: Fresnel Diffraction. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Adam MJ (1981) *An Introduction to Optical Waveguides*. New York: John Wiley & Sons.

Blonder GE, *et al.* (1989) Glass waveguides on silicon for hybrid optical packaging. *IEEE Journal of Lightwave Technology* 7: 1530.

Digonnet M and Shaw HJ (1983) Wavelength multiplexing in single-mode fiber couplers. *Applied Optics* 22: 484.

Jeunhomme LB (1983) *Single-Mode Fiber Optics – Principles and Applications*. New York: Marcel Dekker.

Kashima N (1995) *Passive Optical Components for Optical Fiber Transmission*. Norwood, MA: Artech House.

Kashima N (1993) *Optical Transmission for the Subscriber Loop*. Norwood, MA: Artech House.

Miller CM (1986) *Optical Fiber Splices and Connectors – Theory and Methods*. New York: Marcel Dekker.

SC86B IEC (1999) *IEC 60874-1: Connectors for Optical Fibres and Cables – Part 1: Generic Specification*, 4th edn. Geneva: International Electrotechnical Commission.

SC86B IEC (1999) *IEC 61073-1: Optical Fibres and Cables – Mechanical Splices and Fusion Splice Protectors – Part 1: Generic Specification*, 3rd edn. Geneva: International Electrotechnical Commission.

SC86B IEC (1999) *IEC 60869-1: Fibre Optic Attenuators – Part 1: Generic Specification*, 3rd edn. Geneva: International Electrotechnical Commission.

SC86B IEC (2000) *IEC 60875-1: Non-wavelength Selective Fibre Optic Branching Devices – Part 1:Generic Specification*, 4th edn. Geneva: International Electrotechnical Commission.

SC86B IEC (2000) *IEC 62077-1: Fibre Optic Circulators – Part 1: Generic Specification*, 1st edn. Geneva: International Electrotechnical Commission.

SC86B IEC (2000) *IEC 61977-1 Ed. 1.0: Fibre Optic Filters – Generic Specification*, 1st edn. Geneva: International Electrotechnical Commission.

SC86B IEC (2000) *IEC 61202-1: Fibre Optic Isolators – Part 1: Generic Specification*, 1st edn. Geneva: International Electrotechnical Commission.

Technical Staff of CSELT (1990) *Fiber Optic Communications Handbook*, 2nd edn, TAB Books. New York: McGraw-Hill.

Yokohama I, *et al.* (1988) Novel mass-fabrication of fiber couplers using arrayed fiber ribbons. *Electronic Letters* 24: 1147.

FIBER GRATINGS

P S Westbrook, OFS Laboratories, Murray Hill, NJ, USA

B J Eggleton, University of Sydney, Sydney, NSW, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Fiber gratings are periodic index variations inscribed into the core of an optical fiber, and are important devices for manipulating fiber guided light. Since their first fabrication in the late 1970s, fiber grating use and manufacture has increased significantly and they are now employed commercially as optical filters, reflectors and taps in telecommunications systems, and as strain and temperature sensors in various industries. We review optical fibers and the fundamentals of fiber grating characteristics as well as fabrication. Finally we consider several of the major industrial and scientific applications of fiber gratings.

Fiber Modes

A fiber grating couples light between the modes of an optical fiber, and therefore a knowledge of these modes is essential to understanding the characteristics of fiber gratings. In what follows, we emphasize single-mode fibers used near 1550 nm because these are very important in most practical applications that

use fiber gratings. A single-mode fiber can be viewed as a thin cylinder of silica with a central raised index region of small enough diameter so that only a single mode is guided through total internal reflection at the core boundary. The electric field (*E*-field) of the fundamental mode propagating in the core is described by a characteristic transverse *E*-field profile and a propagation constant k:

$$E = E(r, \varphi)\mathrm{e}^{-\mathrm{i}(\omega t - kz)} \qquad k = 2\pi n_{\mathrm{eff}}/\lambda \qquad [1]$$

Here r and φ are the radial and azimuthal directions, z is measured longitudinally along the fiber, ω is the radial frequency, and λ is the vacuum wavelength, and the *E*-field points in the same direction everywhere. As the second equation indicates, the fiber mode propagates in the fiber as though it has an effective refractive index n_{eff}.

The silica cladding acts to isolate the core mode from the outside, hence the core mode penetrates very little into the cladding. However, if the outer cladding surface allows total internal reflection, the cladding may also guide light. Such 'cladding modes' are also important in understanding fiber gratings, since the grating couples the core mode with the cladding modes. As shown in Figure 1, the cladding modes extend outside the core into the cladding region, and their effective index is therefore always less than that of the core mode. We also note that in the case when the outer coating of the fiber does not support total internal reflection, 'leaky' cladding modes result. In the limit that there is no reflection at the cladding boundary a continuum of 'radiation modes' results. Although these modes are not guided, they are exploited in a variety of applications and can be carefully tailored by a suitably designed fiber grating. Note that the radiation and cladding modes exist regardless of the presence of a grating; a grating simply couples power from the core mode to the cladding/radiation modes; the grating is typically only a small perturbation on the index profile of the waveguide and the corresponding mode fields.

Fiber Grating Theory

Fiber gratings are longitudinal periodic variations in the refractive index (or, more generally, the dielectric function) of the core and/or cladding of an optical fiber. The scattering from any grating may be understood simply by considering the scattering off each successive period and adding these contributions while taking the phase of the *E*-field into account. Figure 2 illustrates this analysis in the case of a fiber Bragg reflector. When the reflections from each period of the grating add constructively, a strong back-reflection results. This is known as Bragg reflection. The condition for Bragg reflection is given by:

$$\lambda_{\mathrm{Bragg}} 2n_{\mathrm{eff}} = \Lambda_{\mathrm{grating}} \qquad [2]$$

Here λ_{Bragg} is the vacuum wavelength of the light (equivalent to its frequency, $\omega = 2\pi c/\lambda$), n_{eff} is the effective index of the mode, and $\Lambda_{\mathrm{grating}}$ is the period of the grating refractive index modulations. This Bragg reflection is evident in a narrow dip in the

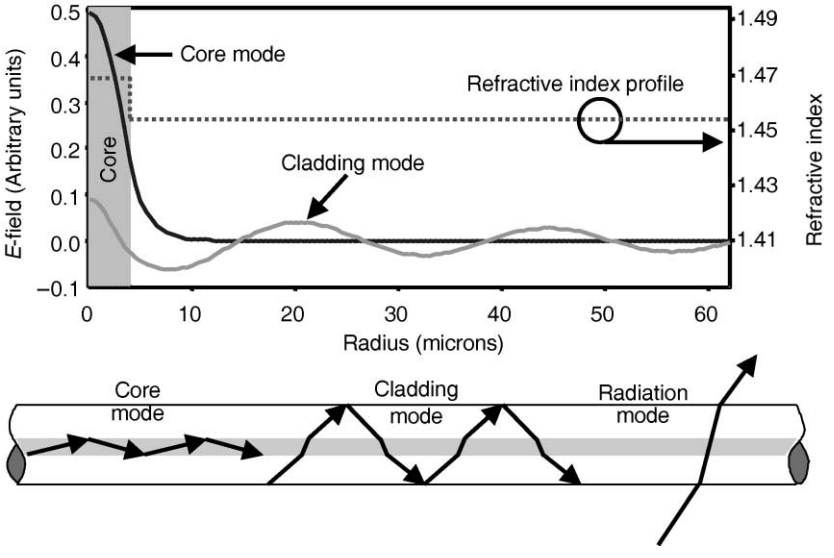


Figure 1 Fiber modes. Refractive index profile of a step index, single-mode fiber (dashed line). Radial profiles of mode *E*-fields of the fundamental (LP01) core mode, guided within the raised index region defined by the core, and one of the many higher-order cladding modes guided by the outer, air–silica interface. Also shown is a fiber indicating the ray paths for core, cladding, and radiation modes.

transmission spectrum of the grating (see Figure 3a). The ability of a fiber grating to selectively reflect the core light at one wavelength is one of its most important practical properties since it can be used as a narrowband filter or reflector. As a fiber component, the fiber grating has the advantage of compactness and provides in-line wavelength filtering.

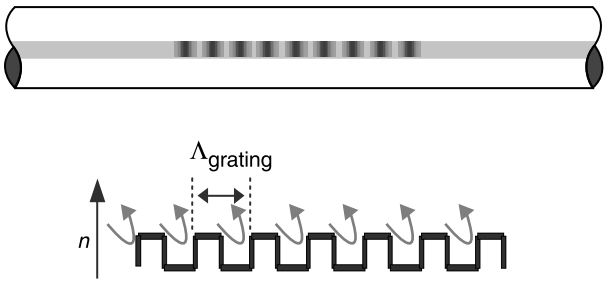


Figure 2 Bragg reflection. Each period of the grating reflects a small fraction of the incident light. When the period of the grating is adjusted so that these reflections add coherently, a strong Bragg reflection results and a narrow dip is observed in the transmission spectrum of the fiber (see Figure 3a). As shown in Figures 3 and 4, Bragg reflection can also occur into cladding modes.

The condition for Bragg reflection may be expressed more generally as a phase-matching condition. Phase matching refers to the condition when the phases of all the grating-scattered waves are the same and there is constructive interference. More specifically, this condition expresses the relationship between the spatial frequencies of the incident light, the scattered light, and the grating modulation, necessary for constructive interference in the scattered light. For a grating, the phase matching condition may be expressed in terms of the grating wavevector, $K_{grating} = 2\pi/\Lambda_{grating}$ and of the two mode propagation constants, $k_{incident}$ and $k_{scattered}$ (defined in eqn [1])

$$k_{incident} - K_{grating} = k_{scattered} \qquad [3]$$

Using phase matching, we can understand the various transmission spectra that result from fiber

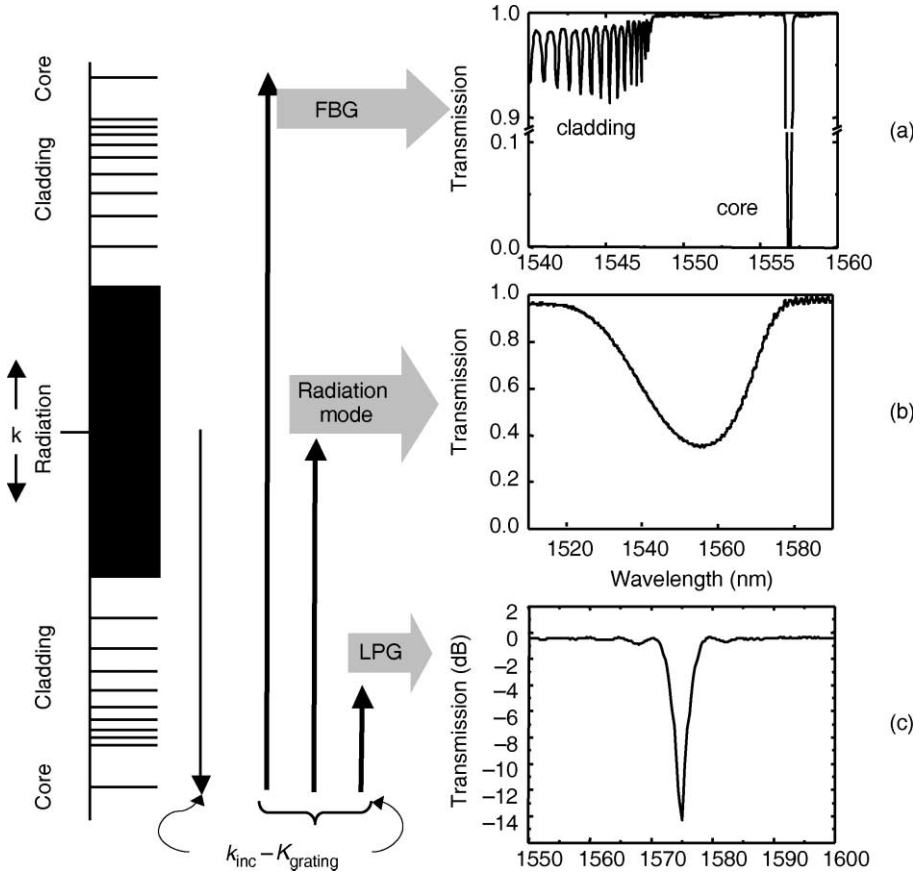


Figure 3 Phase matching. Fiber grating resonances occur when the phase matching condition is satisfied: $k_{inc} - K_{grating} = k_{scat}$. Propagation constants ($k = 2\pi n_{eff}/\lambda$) of the fiber modes are indicated as hash marks on the vertical line. The region indicates the regime of unguided radiation modes. The fiber grating is represented by its longitudinal spatial frequency (wavevector) $K_{grating} = 2\pi/\Lambda_{grating}$, where $\Lambda_{grating}$ is the period of the grating. Assuming an incident forward-going core mode, the grating wavevector then phase matches to (a) counter-propagating core and cladding modes (fiber Bragg grating, FBG), (b) continuum of radiation modes, or (c) a copropagating cladding mode (long-period grating, LPG).

gratings. Figure 3 shows the transmission characteristic observed for core guided light in a single-mode fiber containing three different types of gratings: Bragg reflector, radiation mode coupler, and copropagating mode coupler. Each grating has a different type of phase matching condition. Figure 3 also depicts the phase matching condition in eqn [3] graphically. The vertical line contains a mark representing the propagation constants, k, of modes of a single mode fiber (both positive and negative for the two propagation directions) and a solid black region representing the radiation mode continuum. Bragg reflection into backward-propagating core and cladding modes occurs for short grating periods ($K_{\text{grating}} \sim 2k_{\text{core}}$). Such gratings are known as fiber Bragg gratings (FBGs) and a typical transmission spectrum with a strong Bragg reflection and several cladding mode resonances is shown in Figure 3a (a typical Bragg reflection spectrum is shown in Figure 8 below; the FBG grating period was ~500 nm). Core-cladding mode reflection results in loss resonances for wavelengths shorter than the core–core Bragg reflection. When the grating period is very long (in general, $K_{\text{grating}} < k_{\text{core}}(n_{\text{cladding}} - 1)/n_{\text{cladding}}$; typically Λ_{grating} is ~1000 times that of an FBG), phase matching to copropagating cladding modes occurs. Such gratings are known as long-period gratings (LPGs). A typical LPG spectrum is shown in Figure 3c. The LPG resonance corresponds to coupling between the core mode and a single copropagating cladding mode. For wavevectors in between these two regimes, coupling to the continuum of radiation modes is possible. An example of such a grating spectrum is shown in Figure 3b. We note that fiber cladding modes can be either well guided (resulting in the sharp grating resonances of Figure 3a) or poorly guided (resulting in a broad continuum of Figure 3b) depending on the fiber surface. The radiation mode approximation is valid in the limit that the cladding mode is very poorly guided (i.e., has high loss) within the grating region.

While phase matching yields the resonance wavelengths of a fiber grating, the strengths of these resonances must be derived from a full solution of Maxwell's equations. The most common approximation used in these calculations is known as coupled mode theory (CMT). In CMT, the longitudinal refractive index variation is assumed to be a small perturbation to the waveguide modes, which results only in a change in the propagation constant of the modes but does not change the transverse dependence of the mode E-fields (see eqn [1]). CMT then assumes that the E-field amplitude of each mode varies slowly (compared to wavelength) along the grating. A pair of coupled equations relating the two mode amplitudes as a function of position along the grating can then be derived, and these allow for computation of the grating reflection and transmission. The coupling between two modes is determined by an overlap integral κ_{12} between the product of the E-fields of the two modes (E_1 and E_2) and the periodic refractive index variation:

$$\kappa_{12} \propto \int\int_{\text{Area}} E_1(r,\phi)E_2(r,\phi)\delta n(r,\phi)\,\mathrm{d}A \qquad [4]$$

where $\delta n(r,\phi)$ is derived from the refractive index modulation:

$$\delta n(r,\phi,z) = \delta n(r,\phi)\cos(K_{\text{grating}}z) \qquad [5]$$

This overlap integral is useful because it allows for a rough comparison of the strength of resonances for different modes without the need to obtain a full CMT solution. In the simple case of Bragg reflection of the core mode in a standard single-mode fiber (see Figure 1):

$$\kappa = \frac{\pi\delta n\eta}{\lambda}, \qquad [6]$$

where $\eta \approx 0.8$ is the overlap integral of the core mode intensity with the core.

To understand fiber grating spectra we first discuss a 'uniform' Bragg grating, of length L. Such a grating has uniform index modulation along its length and is shown in Figure 4a. Depending on the length and refractive index modulation of the grating, it may be classified as either weak or strong. The division between these two regimes may be understood by considering the light scattered by each successive period of the grating. The fraction of light scattered from each period can be approximated by $\delta n/2n$ using the Snell's law reflectance formula. The boundary between a strong and a weak grating occurs when the grating is either long enough or strong enough that the sum of the scattering from all periods is of order 1. This division may also be expressed in terms of κ and L: $\kappa L \sim 1$.

In the weak limit, the incident light propagates through the grating with very little scattering. In this case, both the LPG and FBG resonances have transmission spectra that are approximately related to the Fourier transform of the grating index profile. (For LPGs, this may not hold if modal dispersion is large; see the discussion below.) In the case of a uniform grating profile, this results in a $1 - \text{sinc}^2$ wavelength dependence of the transmission spectrum and a spectral width $\delta\lambda/\lambda \sim \Lambda_{\text{grating}}/L = 1/N_{\text{periods}}$ as shown in Figure 4a. An important feature of the uniform grating is the side lobes. The side lobes result from the sudden 'turn on' of the grating. In the FBG,

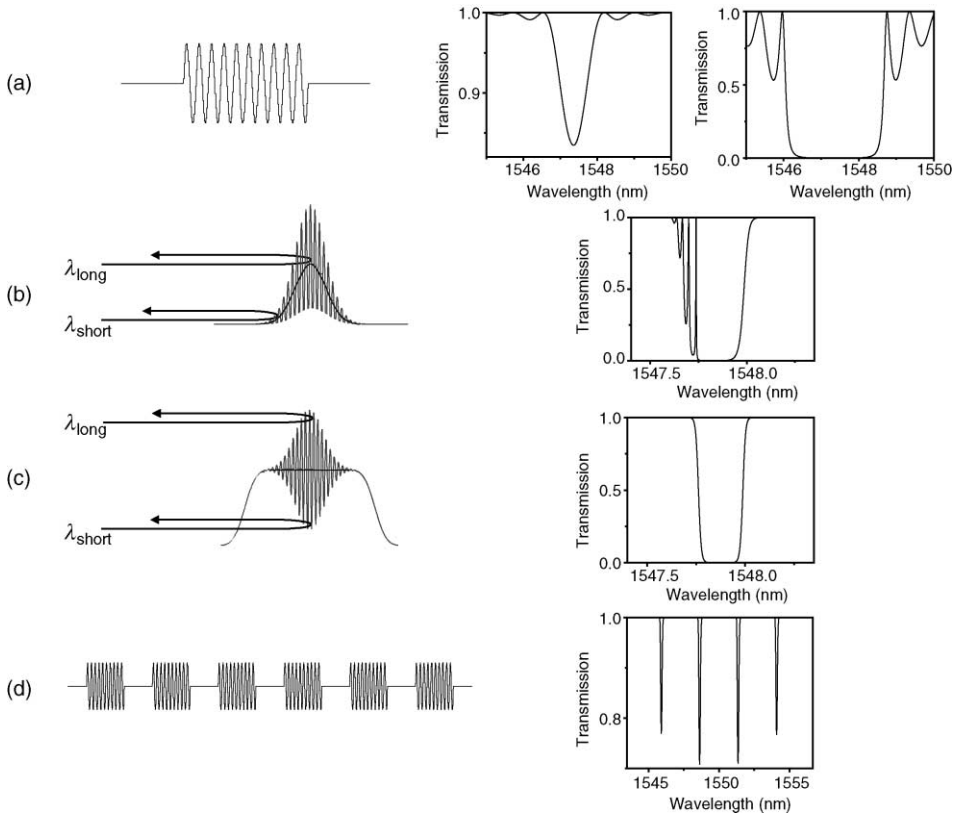


Figure 4 Fiber Bragg grating refractive index profiles and spectra. (a) A uniform profile with transmission spectra for both strong and weak index modulations. (b) An 'unapodized' Gaussian grating profile showing Fabry–Perot-like structure within the grating resonance. (c) A '100% dc apodized' Gaussian grating profile showing a symmetric transmission spectrum. (d) A superstructure grating exhibiting four resonance peaks.

they can be understood as resonances of the effective Fabry–Perot cavity defined by the sharp grating boundaries.

In the strong limit, a substantial fraction of light is scattered out of the incident mode, and the FBG and LPG transmission spectra become different. As shown in Figure 4a, the FBG spectrum widens to a width determined only by δn: $\delta\lambda/\lambda = \delta n/n$. The incident light penetrates only a short distance ($\sim \lambda/\delta n$) into the grating before being completely reflected. The strong LPG, on the other hand, becomes 'overcoupled' as the second mode is reconverted into the incident mode. The result is a higher transmission for the incident core mode (provided that the scattered mode propagates without loss).

While uniform gratings are easy to model and manufacture, most practical fiber gratings have a nonuniform profile. In a nonuniform grating profile, both the refractive index modulation amplitude, $\delta n_{AC}(z)$, and the average effective index value, $\delta n_{DC}(z)$, may vary along the fiber grating length:

$$\delta n(z) = \delta n_{DC}(z) + \delta n_{AC}(z)\cos(K_{grating}z) \qquad [7]$$

By controlling the ac and dc index change, a greatly improved filter may be fabricated. Figure 4b shows the index profile and the grating transmission spectrum of a fiber Bragg grating with an 'unapodized' Gaussian profile. Note that due to the smooth increase in modulation amplitude, the sidebands of the uniform grating have been eliminated. However, because the dc index (and hence Bragg condition) is not constant within the grating, sharp resonant structure is observed within the grating resonance. This structure can be understood as Fabry–Perot resonances in the cavity formed by the nonuniform Bragg condition within the grating. Figure 4c shows a grating with a '100% dc apodized' Gaussian profile. This grating has a constant dc refractive index within the grating and therefore the resonance is smooth and symmetric. Most useful bandpass filters require a profile close to 100% apodization. Although not shown in Figure 4, a smoothly varying index profile also eliminates the sidelobes in LPG resonances.

Other important types of nonuniform fiber gratings include chirped fiber gratings and superstructure fiber gratings. In a chirped fiber grating the grating period

slowly changes along the length of the grating. As discussed below, chirped fiber gratings may be used as chromatic dispersion compensators. They may also be used as broadband reflectors. LPGs may also be chirped to produce a broadband response. Superstructure gratings have oscillations in the period and/or amplitude of the refractive index modulations. The resulting grating can have several resonance peaks (see **Figure 4d**). The properties of nonuniform gratings may be computed using CMT and also understood intuitively using 'band-diagrams' and an effective index model of the fiber grating response.

Grating spectra showing radiation mode coupling are unlike FBGs or LPGs because there is no conversion of the radiation modes back to the incident core mode. The sharp resonances smooth out into a broad continuum. The grating transmission spectrum depends less on the refractive index profile of the grating than on the overlap of the core mode and grating with the radiation modes. Radiation mode coupling is greatly affected by tilting the grating planes with respect to the axis of the fiber. As discussed below, such tilted gratings may be designed as loss filters and optical fiber taps. Grating tilt has also been applied to decrease core mode back-reflection.

LPG resonances (and to some degree FBGs) may also exhibit strong dependence on the modal dispersion (or variation of the propagation constant, k, with wavelength). This wavelength dependence can dominate the wavelength dependence arising from CMT, making the resonance more or less narrow than predicted by CMT without modal dispersion. In an extreme case, very broadband LPG resonances (40 nm) may be produced between the core mode and a specially designed higher-order mode whose propagation constant varies such that it is resonant with the core mode over a large bandwidth.

Fiber Grating Fabrication

Photosensitivity

The key discovery leading to the study and practical application of fiber gratings was the phenomenon of UV-induced photosensitivity in the conventional germanosilicate optical fibers used in telecommunications systems. UV irradiation of the germanosilicate core region can produce refractive index changes in the range 0.01 to 0.0001. Such refractive index changes are large enough to produce useful filters and reflectors. The refractive index change is most efficiently produced by irradiation close to 242 nm or around 193 nm, where germanosilicate glass has strong absorptions. Numerous other dopants have been shown to be photosensitive at wavelengths ranging from 484 nm to 155 nm. These include P, Ce, Pb, Sn, N, Sn-Ge, Sb, and Ge-B. Other dopants, such as Al and F, show very little photosensitivity.

A second major discovery enabling practical manufacture of fiber gratings was the effect of 'hydrogen loading'. The fiber is placed in a pressurized hydrogen atmosphere until molecular hydrogen has diffused into the core region to a level of a few mole percent (a level similar to that of the photosensitive dopants). UV irradiation of the resulting fiber produces index changes an order of magnitude larger than in the unloaded fiber. Hydrogen loading allows gratings to be easily imprinted in conventional single-mode fibers, whose Ge content is too low to allow for strong gratings. The increase in photosensitivity is true for most photosensitive dopants. Index changes in excess of 10^{-2} have been produced with hydrogen loading. In order to avoid increased OH absorption in the telecommunications band (1520–1630 nm), deuterium (D_2) is normally used in place of hydrogen.

The origins of UV photosensitivity are not fully understood; however, at least two mechanisms are agreed to contribute to the index change. Firstly, as is evident from **Figure 5**, the UV absorption spectrum is modified during UV exposure. This bleaching corresponds to a change in refractive index in the infrared which may be computed using the Kramers–Kronig relations. In particular, the absorption band at

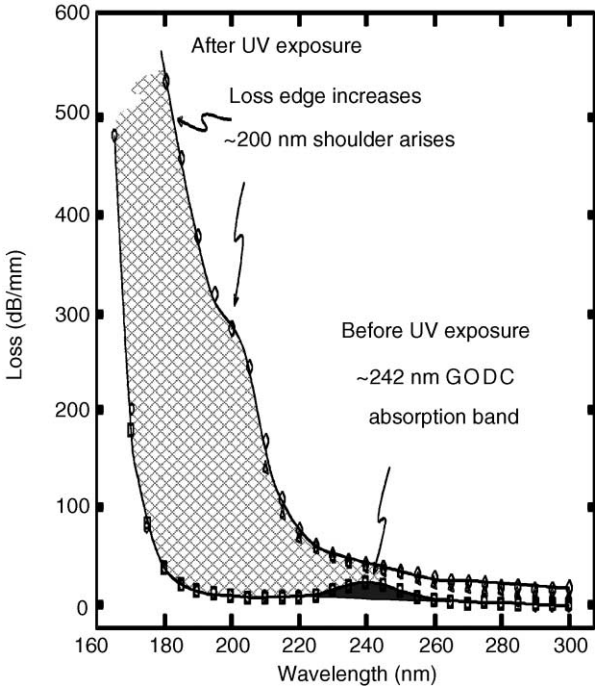


Figure 5 UV absorption spectrum of Ge-doped silica before and after irradiation with 242 nm light, showing bleaching of the Ge-oxygen deficient defect feature at 242 nm.

242 nm resulting from germanium oxygen deficient centers is bleached, indicating that these defects are modified during exposure. The infrared index change results both from this bleaching and from the change in the UV absorption edge at ~190 nm. Secondly, absorption of UV causes the silica matrix to contract. This compaction and the resulting stress distribution also change the refractive index.

Both mechanisms are enhanced by the presence of hydrogen in the silica lattice. The hydrogen reacts with oxygen, thus increasing the number of UV-absorbing germanium oxygen-deficient centers. As a result, in hydrogen loaded fibers, more of the Ge atoms participate in the index change, explaining the large increase in photosensitivity. The number of oxygen-deficient defects may also be increased during fiber manufacture by collapsing the fiber preform in a low-oxygen atmosphere. Such fibers show increased photosensitivity without the need for hydrogen loading. Co-doping a Ge-doped fiber with boron has also been shown to increase the level of photosensitivity of a fiber for a given level of index change in the core.

An important aspect of UV-induced index changes is that they require thermal annealing to ensure long-term stability. Sufficient exposure to elevated temperatures can completely erase the refractive index change in some Ge-doped gratings. Gratings can be completely erased by several hours of heating to greater than 500 °C. Thermal stability has been described by an activation energy model which assumes a distribution of energies for the defects giving rise to the refractive index change. Therefore, at a given anneal temperature a fixed fraction of the index change will quickly decay, leaving only stable defects that easily survive at lower temperatures. By annealing at several temperatures for a fixed time, a relationship between temperature and time required for a given refractive index decay may then be derived and used in accelerated aging tests of fiber gratings. For 20-year grating reliability of ~1% decrease in UV-induced refractive index change, a typical Ge-doped grating requires annealing at 140 °C for 12 hours.

Grating Inscription

The first method used in UV grating inscription is known as 'internal writing'. In this method light propagating in the fiber causes a standing wave to form within the core, and UV photosensitivity from a two-photon process results in permanent grating formation. Both LPG and FBG resonances have been fabricated using this method. The technique is inflexible, however, because the grating parameters are not easily controlled.

Another important advance that made possible the widespread use of fiber gratings is the 'transverse side-writing' technique. In this method, a UV interference pattern is projected through the side of the fiber onto the core region. The first demonstration of this technique employed a free space interferometer to produce the UV fringes. A significant improvement on this technique employs a zero-order nulled phase mask to generate the interference pattern. The phase mask acts to split the incident beam, and these two then form the interference pattern. The method is shown in Figure 6. The advantage of the phase mask

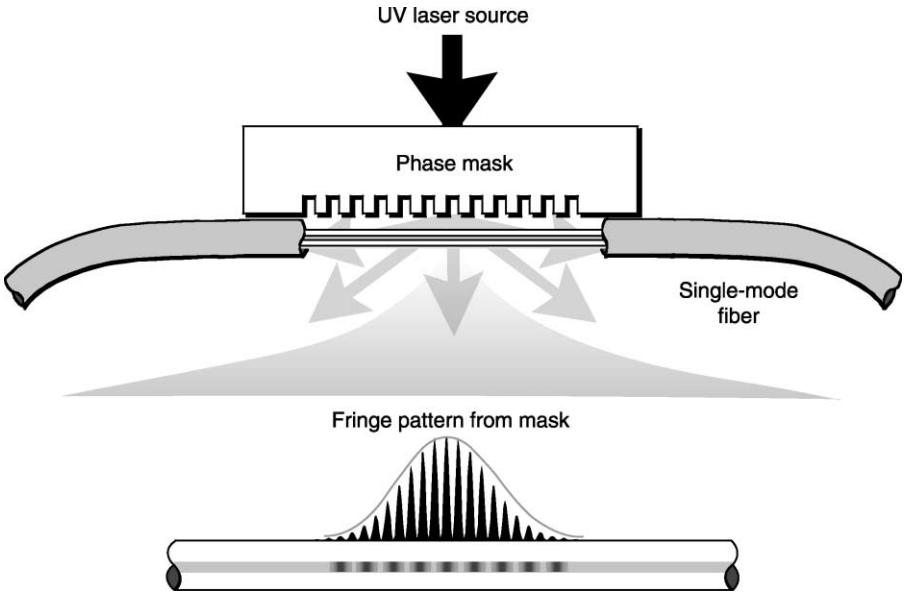


Figure 6 Phase mask fabrication of a fiber grating.

technique is that the fringes are very stable, since the UV beams propagate only a short distance. Another important technique for writing fiber gratings is the 'point-by point' writing method. In a point-by-point technique the UV interference pattern is modulated and translated along a fiber in such a way that a long grating may be formed. Although complex and sensitive, such methods can yield gratings as long as a few meters and also allow for the fabrication of complex grating profiles. Point-by-point methods have been used to fabricate broadband chirped dispersion compensating gratings and Bragg reflectors with very low dispersion.

LPGs may also be fabricated with UV irradiation. An amplitude mask is typically used. Requirements for beam coherence are greatly reduced because of the long period (typically a few hundred microns). Other techniques may also be used to fabricate LPGs. The fiber may be periodically deformed by heating or with periodic microbends. Dynamic LPGs may also be produced by propagating acoustic waves along the fiber.

Applications

Fiber gratings are key enabling technologies in a number of important applications. In this section we review these, highlighting the unique functionalities that can be provided by fiber gratings.

Fiber Grating Strain and Temperature Sensitivity

Fiber gratings are sensitive to both strain and temperature and may therefore be used as sensors. Both LPG and FBG resonances change with strain and temperature. LPGs show greater but more complicated sensitivity and typically FBGs are used in such applications. Strain sensitivity arises from the fact that gratings are extended in length. Because fibers are very thin (diameter $\sim 125\ \mu m$), significant extension is possible. A smaller strain sensitivity arises from the stress optic effect. Typical sensitivity at 1550 nm is ~ 1 pm/1 $\mu\varepsilon$ in FBGs. Temperature sensitivity arises from temperature dependence of the refractive index and gives rise to a change order 0.011 nm/°C of temperature change of the FBG. FBG strain sensors have been applied to map stress distributions and can be employed in a distributed manner with many Bragg reflectors in one fiber being examined with one detector.

These sensitivities may also be used to make fiber grating filters tunable. Temperature changes are more stable but typically give 1–2 nm of change in an FBG resonance. Strain tuning is more difficult to package commercially but can give up to 4 nm in extension and more than 10 nm in compression. Passive thermal stabilization of grating resonances may also be achieved by designing packages in which strain and temperature variations cancel each other.

Fiber Grating Stabilized Diode Lasers and Fiber Lasers

One of the first important commercial applications of fiber gratings was in stabilization of fiber coupled diode lasers. Normally these lasers operate on several cavity modes making the output unstable. If feedback is provided in the form of a narrowband reflection then the laser output will be only in this narrow bandwidth. The fiber grating is implemented in the fiber pigtail of the diode laser and typically provides a few percent reflectivity over a bandwidth of ~ 1 nm. The application is depicted schematically in Figure 7.

More generally, fiber gratings have been used to realize fiber laser cavities and fiber distributed feedback lasers. One important example is the cascaded Raman resonator fiber laser, in which several FBGs recycle Raman pump light to produce high-power, fiber coupled light at any design wavelength.

Optical Add/Drop Multiplexers

Another important commercial FBG application is the optical add/drop multiplexer (see Figure 8). These are key components in wavelength division multiplexed (WDM) networks. WDM networks transmit large amounts of data through a single fiber by modulating several optical carrier wavelengths and sending all the wavelengths through a single fiber. The optical add/drop multiplexer allows individual wavelengths to be added or dropped from a given fiber.

Fiber Bragg gratings are useful in this application because of the very narrow resonances that are achievable (typically a few tenths of a nm at 1550 nm). For wavelengths close to the Bragg wavelength an FBG strongly reflects. The reflectivity and bandwidth are determined by the depth of the index modulation and grating length, as discussed earlier. These degrees of freedom make fiber gratings ideal filters to be used in wavelength division multiplexed lightwave networks, in which multiple channels are transmitted along long lengths of optical fiber. Figure 8 illustrates schematically the principle of

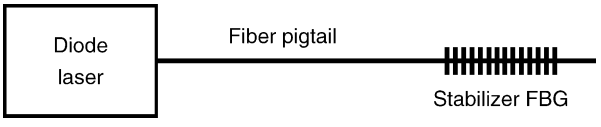


Figure 7 Fiber Bragg grating stabilized diode laser. The fiber grating reflector in the laser pigtail provides a small level of feedback to stabilize the laser output power and wavelength.

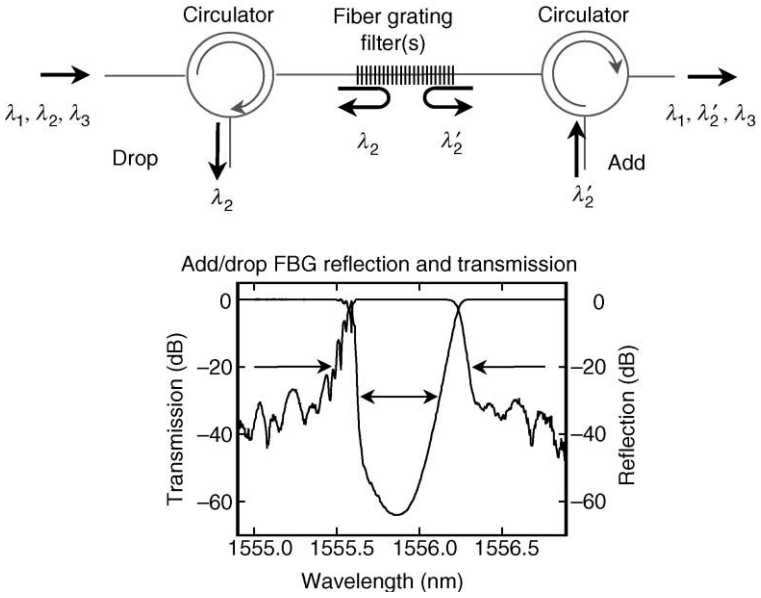


Figure 8 Top: Add/drop node employing fiber grating reflector. Bottom: characteristic reflection and transmission spectrum of an add/drop fiber grating filter.

a fiber grating based add/drop multiplexer. The grating is located between two three-port optical circulators. (A circulator is a bulk optic element that allows transmission only from port n to port $n+1$.) Different channels are incident upon the grating from the left and only one particular channel undergoes reflection at the grating. Note that by carefully designing the grating properties the reflectivity in the adjacent channel can be made to be negligible. This one channel is reflected and re-routed by the three-port circulator. The remainder of the channels are transmitted through the grating. The grating also enables the 'add' functionality. In this case light originating from a different part of the optical network enters the link again through an optical circulator and is reflected off the fiber grating.

Gain-Flattening Loss Filters

A primary strength of fiber gratings is the diversity of filter shapes that is achievable through modification of the grating profile. Gain-flatteners are an important application requiring such filter design. Gain-flattening filters are important in WDM telecommunications systems in order to smooth the wavelength dependence of fiber amplifiers (such as Er or Raman amplifiers). Both LPGs and FBGs may be employed as gain flattening filters. The desired loss spectrum is translated into a corresponding grating profile. LPGs have the advantage of low return loss, but the LPG resonance wavelengths and strengths are very sensitive to the fiber geometry, making the computation of the grating profile less accurate. FBGs may also be used in both tilted and untilted form; however they typically have back-reflections that require careful fiber grating growth or an isolator to suppress back-reflections. Figure 9 shows an example of a gain equalizing loss filter using LPGs.

Dispersion Compensator

Chromatic dispersion can limit the transmission distances and capacities of long-haul fiber links. Chromatic dispersion arises from the variation in propagation velocity with wavelength. An initially sharp pulse is thus dispersed and begins to overlap with adjacent pulses resulting in degraded signal quality. A key technology for long-haul fiber links is therefore 'dispersion compensation'. The most common solution has been 'dispersion compensating fiber' (DCF), which is simply fiber in which the chromatic dispersion has been engineered to be the exact opposite of the dispersion in the fiber link. Limitations of DCF include nonlinearities, higher-order dispersion, and lack of tunability. Fiber gratings offer an alternative approach that overcomes the problems.

Fiber Bragg gratings can provide strong chromatic dispersion when operated in reflection with a three-port circulator (see Figure 10). In this case the grating period is made to vary along the length of the grating. This means that different wavelengths in a pulse reflect at different positions in the grating. This gives rise to wavelength-dependent group delay and therefore chromatic dispersion. The advantage of the fiber grating compared to DCF is that the device can be made to be very compact, with potentially lower

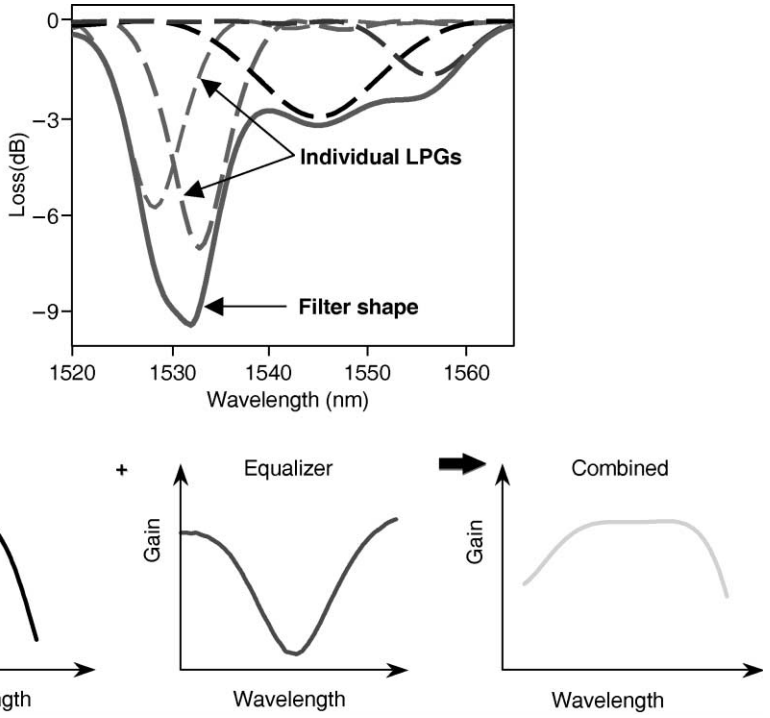


Figure 9 Long-period grating gain equalizer. Several individual gratings (dashed line) are combined to produce the overall shape (solid line).

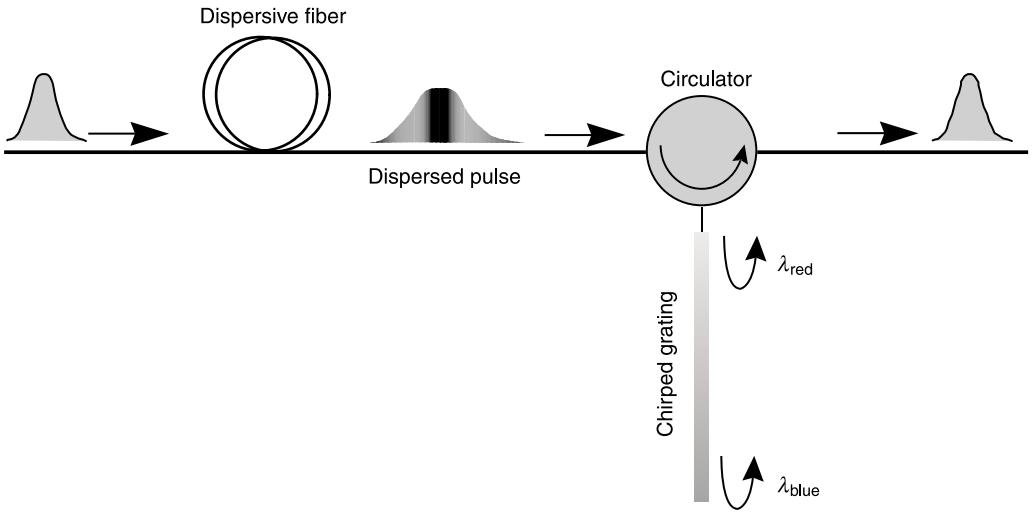


Figure 10 Chirped fiber grating dispersion compensator.

insertion loss, reduced optical nonlinearity, and controllable optical characteristics. In particular fiber gratings offer the possibility of engineering the chromatic dispersion to allow for compensation of dispersion slope and other degradations such as polarization mode dispersion.

Optical Monitors

Fiber gratings may also be employed as compact optical taps coupling core guided light to the radiation modes propagating outside the fiber (see Figure 3b). When a conventional Bragg grating is tilted with respect to the fiber axis it can scatter light out of the fiber in a highly directional manner. The angle of scatter is also wavelength dependent because of the phase matching condition discussed above. This wavelength dependence may form the basis of a compact wavelength monitor. The fiber grating is used to simultaneously couple light out of the fiber and separate the wavelengths. A simple

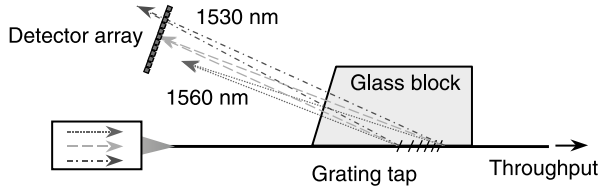


Figure 11 Fiber grating wavelength monitor. The fiber grating acts to scatter light out of fiber and simultaneously disperse the light.

detector array can then measure a spectrum of all the channels in a typical WDM system (see Figure 11).

Fiber grating taps may also scatter light in a polarization sensitive manner. This may be achieved by tilting the fiber grating by 45° with respect to the fiber axis. The resulting broadband taps may be employed in compact polarization monitoring schemes in which detectors are placed on the outside of the fiber to detect light at specific polarizations.

Advanced Topic in Fiber Gratings

Fiber Gratings in Air–Silica Microstructured Optical Fibers

Air–silica microstructure fibers (ASMFs), also known as 'holey' fibers, are fibers with air regions running along their length and allow for more complicated modal properties than conventional silica-doped fiber. It has been shown that fiber gratings can be UV written into photosensitive regions of various types of microstructure fibers. ASMF gratings open new possibilities for fiber grating devices. In addition, they are useful in studying the guidance properties of ASMFs, because they can selectively excite higher-order modes of the fiber. The phase matching condition may be used to derive the effective indices of the higher-order modes and mode profiles may be recorded by imaging a cleaved surface. Such measurements agree well with standard beam propagation computations. Figure 12 shows an example of a mode field image generated with a FBG.

One important application of ASMF gratings is a widely tunable LPG filter. Such filters employ a 'hybrid waveguide' formed by infusing a polymer into the air regions. Because the polymer index is much more sensitive to temperature than silica, the guidance properties of the cladding modes may be tuned. Thus, widely tunable LPG filters may be formed. Moreover, because the cladding modes are confined within an 'inner cladding' such filters are insensitive to external index, allowing for instance,

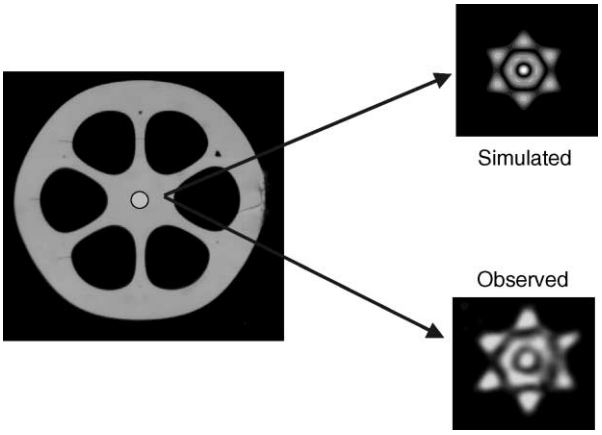


Figure 12 Cross-section of an air–silica microstructure fiber along with measured and simulated modes guided by the air regions.

the use of on-fiber, thin film heaters for power-efficient temperature tuning.

Nonlinear Optics in Fiber Gratings

The linear properties of fiber gratings are well understood and as explained earlier lead to a number of interesting phenomena and applications. When the peak intensity of the optical field is very high a new class of nonlinear effects can occur. These optical nonlinearities arise from the intrinsic nonlinearity of the material system in which the grating is inscribed. In the case of optical fibers, the dominant nonlinearity is third-order nonlinearity which manifests itself in the refractive index of the material depending on the peak intensity of the optical field: $n = n_0 + n_2 I$. This nonlinearity gives rise to a number of dramatic nonlinear effects in fiber gratings, including all-optical switching, bistability, multistability, soliton generation, pulse compression, and wavelength conversion.

To understand these effects we consider an intense optical field incident upon a fiber grating with a central wavelength tuned inside the stopband of the grating where the reflectivity is high. At low intensities the pulse is reflected and only a small amount of light leaks through the grating. As the intensity of the incident optical field is increased the optical nonlinearity becomes significant and results in shifting of the local Bragg wavelength of the grating. This in turn can lead to an increase in transmission of light through the grating and bistable operation.

Theoretical studies of the continuous wave nonlinear properties of periodic media show that such media can exhibit switching as well as bistability and multistability. This prediction has been confirmed

experimentally in a variety of geometries, such as in semiconductors, and in colloidal crystals. Experiments studying nonlinear effects in gratings can also be performed in optical fiber gratings. In these experiments short optical pulses are typically used in order to obtain the very high intensities. This has several key advantages. The first of these is that long, high-quality Bragg gratings can be permanently imprinted into the core of optical fibers by side illumination with interfering ultraviolet beams (see above). This allows for very long interaction lengths (compared to picosecond pulse lengths) and enables nonlinear propagation phenomena to be studied in detail. These include a number of propagation phenomena, such as the formation and propagation of Bragg solitons, modulational instability, and pulse compression. The second advantage is that although the nonlinearity is weak, it has a response time of the order of a few femtoseconds, and thus acts essentially instantaneously on the scale of all but the shortest optical pulses. Finally, the ability to engineer the fiber grating offers a number of unique geometries and potential applications of nonlinear fiber gratings.

A key idea in the interpretation of nonlinear optics in fiber Bragg gratings is the Bragg soliton. Briefly, in these solitons the enormous dispersion provided by the fiber Bragg grating is balanced by the fiber nonlinearity. This is analogous to the optical soliton that has been demonstrated in standard single-mode fiber where the dispersion is predominantly associated with the intrinsic material properties of silica. Since the dispersion of a Bragg grating can be five orders of magnitude larger than that of unprocessed fiber, Bragg solitons can be observed in gratings of only a few centimeters in length, thus rivaling the compactness of other soliton geometries, such as spatial solitons. Of course this implies that the required optical intensities to generate Bragg solitons are correspondingly higher than in uniform media.

Figure 13 shows an example of experimental results when the peak intensity of a pulse incident upon a fiber grating was approximately 10 GW/cm^2 and the incident pulse is tuned to the anomalous dispersion regime of the grating. In particular, Figure 13 shows the transmitted intensity versus time for different values of detuning. By varying the strain on the grating, we can change the detuning of the input pulse, and thus tune to different points on the dispersion curve. Note also from Figure 13 that far from the photonic bandgap edge, where the grating-induced dispersion is negligible, the pulse is unaffected and propagates through the grating at the speed of light. Closer to the edge of the photonic bandgap the pulse is substantially compressed due to the formation of a Bragg soliton. In addition to this compression, the transmitted pulse is substantially delayed in time as a result of the reduced group velocity of the pulse propagating through the grating.

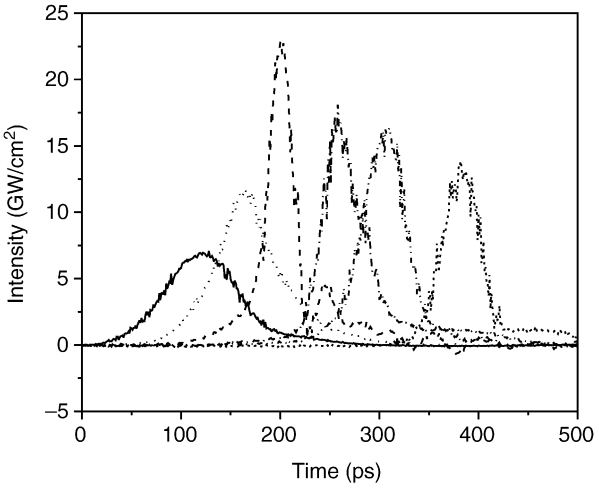


Figure 13 Pulse compression associated with nonlinear pulse propagation through a fiber Bragg grating. Compression is associated with the formation of a Bragg soliton when the incident pulse is tuned just outside of the photonic bandgap formed by the grating.

List of Units and Nomenclature

cw	Continuous wave
FBG	Fiber Bragg grating
LPG	Long-period grating

See also

Detection: Fiber Sensors. **Fiber and Guided Wave Optics:** Fabrication of Optical Fiber; Light Propagation; Measuring Fiber Characteristics; Nonlinear Optics; Optical Fiber Cables. **Lasers:** Optical Fiber Lasers. **Optical Amplifiers:** Optical Amplifiers in Long-Haul Transmission Systems. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Basic Concepts. **Optical Materials:** Optical Glasses. **Photon Picture of Light**.

Further Reading

Adams MJ (1981) *An Introduction to Optical Waveguides*. Chichester, UK: Wiley.

Agrawal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Eggleton BJ, Westbrook PS, Windeler RS, Spalter S and Strasser TA (1999) Grating resonances in air-silica microstructured optical fibers. *Optical Letters* 24: 1460–1462.

Eggleton BJ, Lenz G and Litchinitser NM (2000) Optical pulse compression schemes that use nonlinear Bragg gratings. *Fiber Integrated Optics* 19: 383–421.

Espindola RP, Windeler RS, Abramov AA, *et al.* (1999) External refractive index insensitive air-clad long period fiber grating. *Electronic Letters* 35: 327–328.
Hill KO, Fujii Y, Johnson DC and Kawasaki BS (1978) Photosensitivity in optical fiber waveguides: applications to reflection filter fabrication. *Applied Physics Letters* 32: 647–649.
IEEE Journal of Lightwave Technology (1997) Special issue on fiber gratings, photosensitivity and poling 15(8).
Kashyap R (1999) *Fiber Bragg Gratings*. New York: Academic Press.
Kogelnik H (1990) Theory of optical waveguides. In: Tamir T (ed.) *Guided-Wave Optoelectronics*. Berlin: Springer-Verlag.
Lemaire PJ, Atkins RM, Mizrahi V and Reed WA (1993) High-pressure H_2 loading as a technique for achieving ultrahigh UV photosensitivity and thermal sensitivity in GeO_2 doped optical fibers. *Electronic Letters* 29: 1191–1193.
Marcuse D (1974) *Theory of Dielectric Optical Waveguides*. New York: Academic Press.
Meltz G, Morey WW and Glenn WH (1989) Formation of Bragg gratings in optical fibres by transverse holographic method. *Optical Letters* 14: 823–825.
Othonos A and Kalli K (1999) *Fiber Bragg Gratings: Fundamentals and Applications in Telecommunications and Sensing*. Norwood, MA: Artech House.
Polladian L (1993) Graphical and WKB analysis of nonuniform Bragg gratings. *Physical Review E* 48: 4758–4767.
Westbrook PS, Eggleton BJ, Windeler RS, *et al.* (2001) Cladding mode loss in hybrid polymer-silica microstructured optical fiber gratings. *IEEE Photonic Technology Letters* 12: 495–497.

FOURIER OPTICS

S Jutamulia, University of Northern California, Petaluma, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Coherent optical information processing is almost entirely based on the Fourier-transform property of a lens. A Fourier-transform lens is actually an ordinary lens. If the input transparency is placed in the front focal plane of the lens and illuminated with coherent collimated light (planewave), the amplitude function in the back focal plane of the lens will be the Fourier transform of the input transparency as shown in Figure 1. A coherent optical processor usually consists of two lenses, as shown in Figure 2, thus it performs two Fourier transforms successively. The first lens transforms the input function from space domain into frequency domain, and the second lens transforms back the frequency function from frequency domain to the output function in space domain.

Fourier Transform Property of Lens

Recently, the Fourier-transform property of lenses has been explained, simply using the concept of linear system and geometrical optics. In contrast, the Fourier-transform property of lenses is traditionally analyzed, based on the Fresnel–Kirchhoff diffraction theory involving quadratic phase terms (eqns [2] and [3]). In this article, we adopt the simple geometrical theory based on geometrical optics.

We start our discussion by looking at Figure 1. Consider that a coherent point source is at the center of the front focal plane of the lens, which is denoted by plane (x, y). The point source generates a collimated beam parallel to the optical axis. The amplitude function at the back focal plane of the lens, which is denoted by plane (u, v), can be described as unity. Consequently, we can express mathematically:

$$f(x, y) = \delta(x, y) \qquad [1]$$

$$F(u, v) = 1 \qquad [2]$$

where $f(x, y)$ is the amplitude function in the front focal plane, and $F(u, v)$ is the amplitude function in the back focal plane.

If the point source shifts a distance a along the y-axis, the amplitude function in the front focal plane is

$$f(x, y) = \delta(x, y + a) \qquad [3]$$

The off-axis point source generates a tilted collimated beam. If we place a screen in the plane (u, v), we will see uniform intensity across the screen similar to that which would be seen when the source is centered on the axis. Mathematically, it means:

$$|F(u, v)|^2 = 1 \qquad [4]$$

However, the phase of the amplitude function is not uniform, since the collimated beam is not parallel to

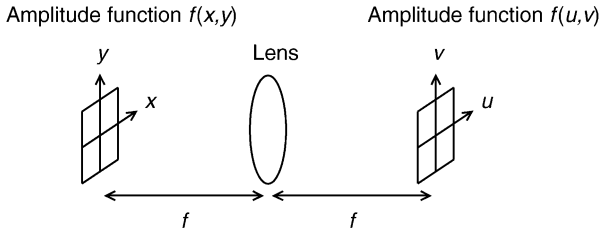


Figure 1 A lens Fourier-transforms amplitude function $f(x, y)$ in the front focal plane to amplitude function $F(u, v)$ in the back focal plane.

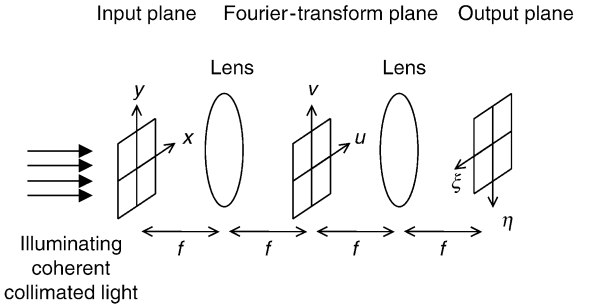


Figure 2 A pair of lenses form 4-f coherent optical processor.

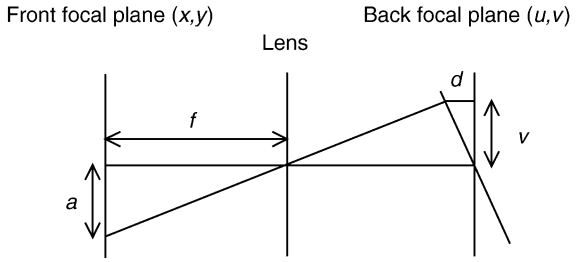


Figure 3 Calculation of phase function in the back focal plane (u, v) based on geometrical optics.

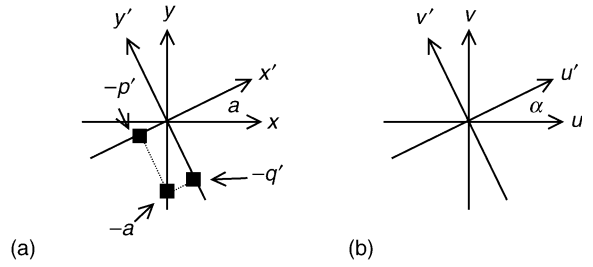


Figure 4 (a) Front focal plane coordinate system (x, y) is rotated by α. (b) Similarly, back focal plane coordinate system (u, v) is rotated by α.

the optical axis. In other words, plane (u, v) is not parallel to the wavefront.

If the phase at the center of plane (u, v) is considered zero, the phase at any point at plane (u, v) can be calculated referring to Figure 3. Using simple geometry, we know that

$$d(u, v) = \frac{a}{f} v \qquad [5]$$

where $d(u, v)$ is the optical path difference. Correspondingly, the phase is

$$\phi(u, v) = \frac{2\pi}{\lambda f} av \qquad [6]$$

where λ is the wavelength of light.

In complex representation, the amplitude function is expressed as

$$F(u, v) = A(u, v) \exp[i\phi(u, v)] \qquad [7]$$

where $A(u, v)$ and $\phi(u, v)$ are magnitude and phase of the amplitude function, respectively. From eqn [4], we know that

$$A(u, v) = |F(u, v)| = 1 \qquad [8]$$

Substituting eqns [8] and [6] into [7] yields

$$F(u, v) = \exp\left(i \frac{2\pi}{\lambda f} av\right) \qquad [9]$$

Equation [9] is the amplitude function in the back focal plane of the lens, when the amplitude function in the front focal plane of the lens is given by eqn [3].

Consider now that coordinate (x,y) is rotated by angle α to new coordinates (x',y') as shown in Figure 4a. The amplitude function in the front focal plane now becomes

$$f(x', y') = \delta(x' + p', y' + q') \qquad [10]$$

where

$$p' = a \sin \alpha \qquad [11a]$$

$$q' = a \cos \alpha \qquad [11b]$$

Correspondingly, coordinate (u, v) is also rotated by the same angle α to new coordinates (u', v') as shown in Figure 4b. The old coordinates u and v can be described in terms of new coordinates u' and v' as follow:

$$u = u' \cos \alpha - v' \sin \alpha \qquad [12a]$$

$$v = u' \sin \alpha + v' \cos \alpha \qquad [12b]$$

Substituting eqn [12b] into eqn [6] yields

$$\phi(u', v') = \frac{2\pi}{\lambda f} a(u' \sin \alpha + v' \cos \alpha) \qquad [13]$$

Further, substituting eqns [11a] and [11b] into [13] yields

$$\phi(u', v') = \frac{2\pi}{\lambda f}(p'u' + q'v') \qquad [14]$$

Thus, the amplitude function in the back focal plane of lens is

$$F(u', v') = \exp\left[i\frac{2\pi}{\lambda f}(p'u' + q'v') \right] \qquad [15]$$

For convenience, we now change parameters x' and y' to x and y, u' and v' to u and v, p' and q' to $-p$ and $-q$. Equations [10] and [15] become

$$f(x, y) = \delta(x - p, y - q) \qquad [16]$$

$$F(u, v) = \exp\left[-i\frac{2\pi}{\lambda f}(pu + qv) \right] \qquad [17]$$

We may simply consider that $f(x, y)$ is an input function and $F(u, v)$ is an output function of a system, since they are both physically the same amplitude functions. If there are two mutually coherent point sources at $(0, 0)$ and $(0, -a)$ with different amplitudes k and l in the front focal plane, the input function is

$$f(x, y) = k\delta(x, y) + l\delta(x, y + a) \qquad [18]$$

These two point sources will generate two collimated beams. Since they are mutually coherent, two collimated beams will interfere, and the resultant amplitude function in the back focal plane is simply the superposition of two beams, which is

$$F(u, v) = k + l \exp\left(i\frac{2\pi}{\lambda f} av \right) \qquad [19]$$

Thus, the lens is a linear system when the light is coherent. Furthermore, if input function $f(x, y)$ is a group of mutually coherent point sources with different amplitudes (i.e., an optical transparency illuminated with a planewave), it can be expressed as follows:

$$f(x, y) = \int\int_{-\infty}^{\infty} f(p, q)\delta(x - p, y - q)\mathrm{d}p\ \mathrm{d}q \qquad [20]$$

Since the lens is a linear system, output function will be the same superposition as follows:

$$F(u,v) = \int\int_{-\infty}^{\infty} f(p,q)\exp\left[-i\frac{2\pi}{\lambda f}(pu+qv) \right]\mathrm{d}p\,\mathrm{d}q \qquad [21]$$

By again changing parameters p and q to x and y, respectively, eqn [21] becomes

$$F(u,v) = \int\int_{-\infty}^{\infty} f(x,y)\exp\left[-i\frac{2\pi}{\lambda f}(xu+yv) \right]\mathrm{d}x\,\mathrm{d}y \qquad [22]$$

which is the optical Fourier transform of the input function $f(x,y)$.

4-f Coherent Optical Processor

If the amplitude function in the front focal plane of a lens is denoted by $f(x, y)$, then the amplitude function in the back focal plane is simply $F(u, v)$ given in eqn [22], which is the Fourier transform of $f(x, y)$. Mathematically, $f(x, y)$ can also be derived from $F(u, v)$ (eqn [4]). After neglecting the normalization factor, it can be expressed as follows:

$$f(x, y) = \int\int_{-\infty}^{\infty} F(u, v)\exp\left[i\frac{2\pi}{\lambda f}(xu + yv) \right]\mathrm{d}u\,\mathrm{d}v \qquad [23]$$

The normalization factor $(1/\lambda f)^2$ is caused by the coordinate scale factor λf in the Fourier-transform plane (u, v).

Equations [22] and [23] are Fourier transform and inverse Fourier transform, respectively. The operation of Fourier transform is represented by a notation $F\{\cdot\}$, and inverse Fourier transform is $F^{-1}\{\cdot\}$.

Two Fourier transform lenses can be combined into a two-lens processor, which is better known as a 4-f optical processor, as shown in Figure 2. Figure 2 illustrates the principle of a 4-f optical processor, in which $f(x, y)$, $F(u, v)$, and $f(\xi, \eta)$ are amplitude functions in input, Fourier transform, and output planes, respectively. Since $F(u, v)$ is the function of spatial frequency, Fourier-transform plane is also called a frequency plane. Since a lens performs only Fourier transform, to perform inverse Fourier transform, the output plane must be rotated by 180°. Thus, orientations of ξ and η correspond to $-x$ and $-y$.

In practice, $f(x, y)$ is represented by an optical mask, which can be a film transparency or a spatial light modulator. The optical mask with $f(x, y)$ amplitude transmittance is illuminated with collimated coherent light to produce the amplitude function $f(x, y)$. The Fourier transform $F(u, v)$ is physically formed in the frequency plane. The frequency content of the input function $f(x, y)$ can be directly altered by placing an optical mask in the frequency plane. Finally, an image with amplitude function $f(\xi, \eta)$ is formed in the output plane. Notice that all detectors, including our eyes, detect intensity that is the square of amplitude.

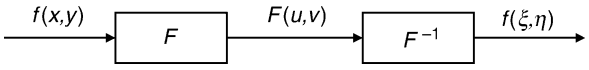


Figure 5 Block diagram of 4-f coherent optical processor.

According to eqn [22], $F(u, v)$ can be expressed in terms of $f(x, y)$ as

$$F(u, v) = F\{f(x, y)\} \quad [24]$$

Similarly, the amplitude function in the output plane must be

$$f(x', y') = F\{F(u, v)\} \quad [25]$$

where (x', y') is the coordinate of the output plane having the same orientation as (x, y) in the input plane. This function can be expressed in terms of inverse Fourier transform as follows:

$$f(-x', -y') = F^{-1}\{F(u, v)\} \quad [26]$$

A rotated coordinate (ξ, η), shown in Figure 2, substitutes $(-x', -y')$, such that

$$f(\xi, \eta) = F^{-1}\{F(u, v)\} \quad [27]$$

Thus, the successive operations can be simply illustrated by the diagram shown in Figure 5.

Spatial Filtering

The origin of optical information processing can be traced as far back as the work of Abbe of 125 years ago. His interest at that time concerned the theory of microscopy in order to establish a scientific approach to microscope design. He found that a larger aperture resulted in higher resolution. An object under the microscope, whose size was of the order of the wavelength of light, diffracted the illuminating light, i.e., the straight light was bent. If the aperture of the objective lens was not large enough to collect all of the diffracted light, the image did not correspond exactly to that object. Rather, it related to a fictitious object whose complete diffracted light pattern matched the one collected by the objective lens. These lost portions are associated with the higher spatial frequency. Their removal will result in a loss in image sharpness and resolution. The basic ideas of Abbe's theory were later clarified in a series of experiments by Porter in the early 1900s.

Since the late 1940s, the possibility of applying coherent optical techniques to image processing had been explored. In 1960, Cutrona *et al*. presented an optical system using lenses to perform Fourier transforms. Thus, every Fourier transformation in the processor can be realized by a lens. The power of this architecture is its capability of forming Fourier spectra and also being cascaded.

Consider an image formed by an optical system such as a camera; because of the error or limitation of the system, the perfect point in the object may smear. If we consider that the original object is the collection of a large number of very fine points, the formed image will be the collection of the same number of blurred spots, instead of fine points. Thus, the image formation can be expressed by convolution. The blurred spot is represented by the impulse response $h(x, y)$, which is the characteristic of the optical system.

When the input is a point (an impulse), its amplitude function is a delta function $\delta(x, y)$, and the output amplitude function is $h(x, y)$. The observed image is the intensity function $|h(x, y)|^2$. Since it indicates the spread of a point, $|h(x, y)|^2$ is called the point spread function.

The output image function $o(x, y)$ is the convolution of input function $f(x, y)$ and the impulse response $h(x, y)$ as follows:

$$o(x, y) = \int\int_{-\infty}^{\infty} f(p, q)h(x - p, y - q)\mathrm{d}p\ \mathrm{d}q \quad [28]$$

Notice that $h(x - p, y - q)$ is the blurred spot centered at (p, q). $H(u, v)$, $F(u, v)$, and $O(u, v)$ are Fourier transforms of $h(x, y)$, $f(x, y)$, and $o(x, y)$, respectively. Substitution of

$$h(x-p,y-q)=\int\int_{-\infty}^{\infty} H(u,v) \times\exp\left[i\frac{2\pi}{\lambda f}[(x-p)u+(y-q)v]\right]\mathrm{d}u\,\mathrm{d}v \quad [29]$$

then

$$F(u,v)=\int\int_{-\infty}^{\infty} f(p,q)\exp\left[-i\frac{2\pi}{\lambda f}(pu+qv)\right]\mathrm{d}p\,\mathrm{d}q \quad [30]$$

yields

$$o(x,y)=\int\int_{-\infty}^{\infty} F(u,v)H(u,v)\exp\left[i\frac{2\pi}{\lambda f}(xu+yv)\right]\mathrm{d}u\,\mathrm{d}v \quad [31]$$

Since

$$o(x,y)=\int\int_{-\infty}^{\infty} O(u,v)\exp\left[i\frac{2\pi}{\lambda f}(xu+yv)\right]\mathrm{d}u\,\mathrm{d}v \quad [32]$$

we obtain

$$O(u,v)=H(u,v)F(u,v) \quad [33]$$

The function $H(u,v)$, which is the Fourier transform of the impulse response $h(x,y)$, is called transfer function. Actually, $H(u,v)$ changes the spectrum of input function $F(u,v)$ to become $O(u,v)$ and the spectrum of the output function becomes $H(u,v)F(u,v)$.

The first application of spatial filtering may be the restoration of a photograph. The image formed by a camera is a collection of spots. If the image is in focus, the size of the spots is very fine, and thus the image looks sharp. If the image is out of focus, the spot size is fairly large and the image is blurred. Interestingly, an out-of-focus image can be restored using spatial filtering technique.

Consider that the original image is $f(x, y)$. Because it is out of focus, a point in the image becomes a spot expressed by $h(x, y)$. If the image is in focus, $h(x, y)$ is a delta function. If the in-focus image is $f(x, y)$, its frequency function is $F(u, v)$. The defocused image $o(x, y)$ is expressed by eqn [28], and its frequency function is $F(u,v)H(u,v)$. To produce the in-focus image, the frequency function must be $F(u, v)$. Thus $F(u,v)H(u,v)$ must be multiplied by the filter function of $1/H(u,v)$ to produce $F(u,v)$ only. The filter of $1/H(u,v)$ can be made from a photographic plate to provide its magnitude, and a glass plate having two values of thickness to provide $+1$ and -1 signs. The glass plate is essentially a phase filter. The photographic plate with varying density function is an amplitude filter. By passing frequency function $F(u,v)H(u,v)$ through the combination of amplitude and phase filters, it will result in $F(u,v)$ only. When $F(u,v)$ is again Fourier transformed by a lens, an in-focus image $f(x, y)$ will be produced.

Complex Matched Spatial Filtering

The restoration of a photograph mentioned above, utilizes the filter function $1/H(u,v)$, which has positive and negative real values. We will see an application in which the filter function is a complex function. A complex function filter can be produced by a single hologram, instead of using a combination of a phase filter and an amplitude filter. The complex matched spatial filter can perform a correlation, as first demonstrated by Vander Lugt. The complex filter is actually a Fourier transform hologram proposed by Leith and Upatnieks.

The synthesis of a complex filter is shown in Figure 6. First, a function $g(x, y)$ is displayed in the input plane by a transparency or spatial light modulator and illuminated with a collimated coherent light beam. Its Fourier-transform $G(u,v)$ is generated in the back focal plane of the lens. The hologram is made in the Fourier-transform plane by recording the interference pattern of $G(u,v)$ and the reference beam $R(u,v)$. $R(u,v)$ is an oblique collimated beam, having magnitude of unity, which is expressed mathematically as

$$R(u,v) = \exp(iau) \qquad [34]$$

and

$$a = \frac{2\pi \sin \alpha}{\lambda} \qquad [35]$$

where α is the angle of the reference beam and λ is the wavelength of light.

The transmittance of the recorded hologram will be

$$\begin{aligned} T(u,v) &= |G(u,v) + R(u,v)|^2 \\ &= |G(u,v)|^2 + 1 + G(u,v)\exp(-iau) \\ &\quad + G^*(u,v)\exp(iau) \end{aligned} \qquad [36]$$

where * denotes the complex conjugate. The hologram, which is called a complex matched spatial filter, is then placed back in the Fourier-transform plane, as shown in Figure 7. A new input function $f(x, y)$ is displayed in the input plane. The Fourier-transform of input function, $F(u,v)$ will be generated in the Fourier-transform plane. The amplitude of light,

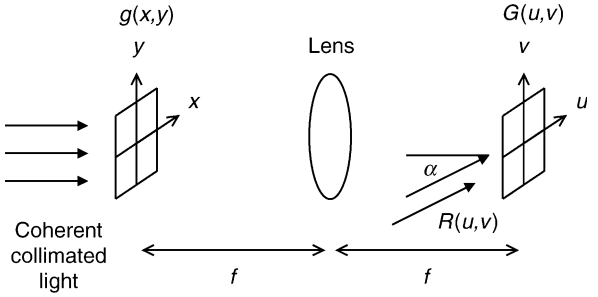


Figure 6 Synthesis of complex matched spatial filter is the recording of Fourier-transform hologram.

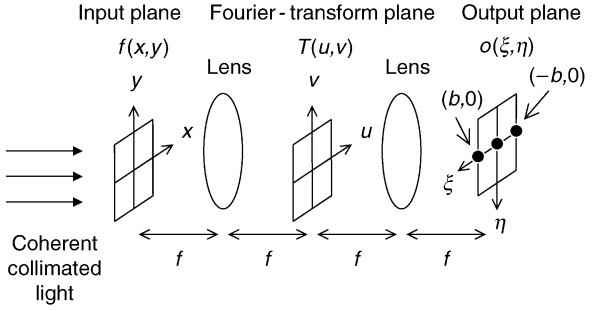


Figure 7 Optical processing using complex matched spatial filter in a 4-f processor.

immediately after the complex matched spatial filter, is

$$\begin{aligned} F(u,v)T(u,v) = {} & (|G(u,v)|^2 + 1)F(u,v) \\ & + F(u,v)G(u,v)\exp(-iau) \\ & + F(u,v)G^*(u,v)\exp(iau) \end{aligned} \quad [37]$$

The second lens then produces the inverse Fourier transform of $F(u,v)T(u,v)$ in the output plane:

$$\begin{aligned} F^{-1}\{F(u,v)T(u,v)\} = {} & F^{-1}\{(|G(u,v)|^2 + 1)F(u,v)\} \\ & + F^{-1}\{F(u,v)G(u,v)\exp(-iau)\} \\ & + F^{-1}\{F(u,v)G^*(u,v)\exp(iau)\} \end{aligned} \quad [38]$$

The first term will be at the center of the output plane. The second and the third terms will appear at $(b,0)$ and $(-b,0)$ in (ξ,η) plane, respectively, where

$$b = f\sin\alpha \quad [39]$$

and f is the focal length of lens.

We are now in a position to analyze the inverse Fourier transform of $F(u,v)G(u,v)$ and $F(u,v)G^*(u,v)$. Referring to eqns [28]–[33], the inverse Fourier transform of $F(u,v)G(u,v)$ is the convolution of $f(\xi,\eta)$ and $g(\xi,\eta)$. Thus, the complex matched spatial filtering can also be applied to the restoration of photography mentioned previously without using the combination of amplitude and phase filters.

The inverse Fourier transform of $F(u,v)G^*(u,v)$ is derived as follows:

$$\begin{aligned} o_3(\xi,\eta) = {} & \int\int_{-\infty}^{\infty} F(u,v)G^*(u,v) \\ & \times \exp\left[i\frac{2\pi}{\lambda f}(\xi u + \eta v) \right] du\, dv \end{aligned} \quad [40]$$

Substitute

$$\begin{aligned} F(u,v) = {} & \int\int_{-\infty}^{\infty} f(\alpha,\beta) \\ & \times \exp\left[-i\frac{2\pi}{\lambda f}(\alpha u + \beta v) \right] d\alpha\, d\beta \end{aligned} \quad [41]$$

and

$$\begin{aligned} G^*(u,v) = {} & \int\int_{-\infty}^{\infty} g^*(-p,-q) \\ & \times \exp\left[-i\frac{2\pi}{\lambda f}(pu + qv) \right] dp\, dq \end{aligned} \quad [42]$$

into eqn [40], one obtains

$$\begin{aligned} o_3(\xi,\eta) = {} & \int\int_{-\infty}^{\infty} f(\alpha,\beta) \int\int_{-\infty}^{\infty} g^*(-p,-q) \\ & \times \int\int_{-\infty}^{\infty} \exp\left[i\frac{2\pi}{\lambda f}(\xi - \alpha - p)u \right. \\ & \left. + (\eta - \beta - q)v \right] du\, dv\, dp\, dq\, d\alpha\, d\beta \end{aligned}$$

$$\begin{aligned} o_3(\xi,\eta) = {} & \int\int_{-\infty}^{\infty} f(\alpha,\beta) \int\int_{-\infty}^{\infty} g^*(-p,-q) \\ & \times \delta(\xi - \alpha - p, \eta - \beta - q) dp\, dq\, d\alpha\, d\beta \end{aligned}$$

$$o_3(\xi,\eta) = \int\int_{-\infty}^{\infty} f(\alpha,\beta) g^*(\alpha - \xi, \beta - \eta) d\alpha\, d\beta \quad [43]$$

Equation [43] is the correlation of $f(x,y)$ and $g(x,y)$. It is apparent that the correlation output $o_3(\xi,\eta)$ will achieve the maximum when $f(x,y)$ is identical to $g(x,y)$. If $f(x,y)$ and $g(x,y)$ are different, the correlation function will be a flat function with minor random fluctuation. Optical correlators using complex matched spatial filters have been applied to automatic pattern recognition.

Joint Transform Correlator

Alternatively, the joint transform correlator can perform a correlation without recording a complex matched spatial filter in advance. In addition to the correlation, the joint transform correlator can also be used as a general optical processor.

A transparency with amplitude transmittance $f(x,y)$ is centered at $(-a,0)$, and a second transparency with amplitude transmittance $g(x,y)$, is centered at $(a,0)$. The two transparencies are illuminated by a coherent collimated beam and jointly Fourier transformed by a single lens as shown in Figure 8. In the back focal plane of the lens, the amplitude function $S(u,v)$ is

$$\begin{aligned} S(u,v) = {} & \int\int_{-\infty}^{\infty} [f(x+a,y) + g(x-a,y)] \\ & \times \exp\left[-i\frac{2\pi}{\lambda f}(xu + yv) \right] dx\, dy \end{aligned} \quad [44]$$

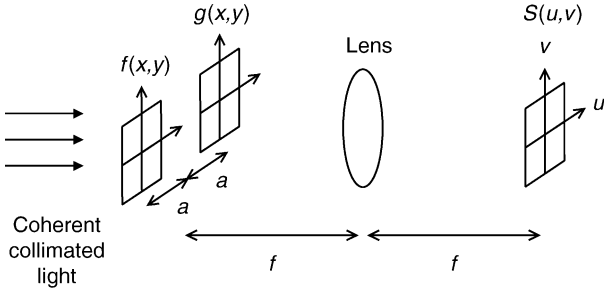


Figure 8 Joint Fourier transform of two input functions.

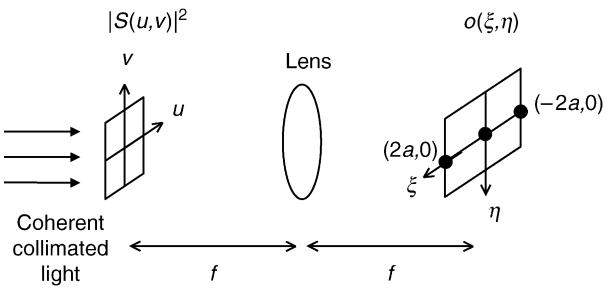


Figure 9 Fourier transform of the joint power spectrum.

$S(u,v)$ is the joint Fourier spectrum of $f(x,y)$ and $g(x,y)$. It can be written as

$$S(u,v) = F(u,v)\exp\left[i\frac{2\pi}{\lambda f}au\right] + G(u,v)\exp\left[-i\frac{2\pi}{\lambda f}au\right] \quad [45]$$

where $F(u,v)$ and $G(u,v)$ are Fourier transforms of $f(x,y)$ and $g(x,y)$, respectively.

If a square-law detector is placed in the back focal plane of the lens, the recorded pattern of joint power spectrum $|S(u,v)|^2$ is

$$|S(u,v)|^2 = |F(u,v)|^2 + F(u,v)G^*(u,v)\exp\left[i\frac{4\pi}{\lambda f}au\right] + F^*(u,v)G(u,v)\exp\left[-i\frac{4\pi}{\lambda f}au\right] + |G(u,v)|^2 \quad [46]$$

where * denotes the complex conjugate. A photographic plate is a square-law detector and can be used to record $|S(u,v)|^2$. When the developed plate is placed in the input plane, as shown in Figure 9, the input transmittance is $|S(u,v)|^2$ as given in eqn [46].

The inverse Fourier transform of the joint power spectrum $|S(u,v)|^2$, in the back focal plane of the lens, is

$$o(\xi,\eta)=\int\int_{-\infty}^{\infty}|S(u,v)|^2\exp\left[i\frac{2\pi}{\lambda f}(\xi u+\eta v)\right]\mathrm{d}u\,\mathrm{d}v \quad [47]$$

which consists of four terms, according to eqn [46]. We now analyze this equation term by term. The first term is

$$o_1(\xi,\eta)=\int\int_{-\infty}^{\infty}|F(u,v)|^2\exp\left[i\frac{2\pi}{\lambda f}(\xi u+\eta v)\right]\mathrm{d}u\,\mathrm{d}v \quad [48]$$

which is the auto-correlation of $f(x,y)$

$$o_1(\xi,\eta)=\int\int_{-\infty}^{\infty}f(\alpha,\beta)f^*(\alpha-\xi,\beta-\eta)\mathrm{d}\alpha\mathrm{d}\beta \quad [49]$$

The second term is

$$o_2(\xi,\eta)=\int\int_{-\infty}^{\infty}F(u,v)G^*(u,v)\exp\left[i\frac{4\pi}{\lambda f}au\right]\times\exp\left[i\frac{2\pi}{\lambda f}(\xi u+\eta v)\right]\mathrm{d}u\,\mathrm{d}v \quad [50]$$

which is

$$o_2(\xi,\eta)=\int\int_{-\infty}^{\infty}f(\alpha,\beta)g^*[\alpha-(\xi+2a),\beta-\eta]\mathrm{d}\alpha\mathrm{d}\beta \quad [51]$$

Equation [51] is the correlation between $f(x,y)$ and $g(x,y)$, shifted by $-2a$ in the ξ-axis direction.

Similarly, the third term will be the correlation between $g(x,y)$ and $f(x,y)$, shifted by $2a$ in the ξ-axis direction:

$$o_3(\xi,\eta)=\int\int_{-\infty}^{\infty}g(\alpha,\beta)f^*[\alpha-(\xi-2a),\beta-\eta]\mathrm{d}\alpha\mathrm{d}\beta \quad [52]$$

And the fourth term is

$$o_4(\xi,\eta)=\int\int_{-\infty}^{\infty}g(\alpha,\beta)g^*(\alpha-\xi,\beta-\eta)\mathrm{d}\alpha\mathrm{d}\beta \quad [53]$$

which is the auto-correlation of $g(x,y)$ and overlaps with the first term at the center of the output plane.

From the analysis given above, it is understood that the correlation of $f(x,y)$ and $g(x,y)$, and the correlation of $g(x,y)$ and $f(x,y)$, are shifted by $-2a$ and $2a$, respectively, along the ξ-axis. The correlation outputs are centered at $(-2a,0)$ and $(2a,0)$ in the back focal plane of the lens. If $f(x,y)$ and $g(x,y)$ are the same, the two correlation outputs are identical. The first and fourth terms are the autocorrelations of $f(x,y)$ and $g(x,y)$, centered at $(0,0)$. Thus the second and third terms, which are correlation signals, can be isolated. A versatile real-time joint transform

correlator, using inexpensive liquid crystal televisions, is widely used in laboratory demonstration.

See also

Information Processing: Coherent Analogue Optical Processors. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Abbe E (1873) Beitrage zur Theorie des Mikroskops und der mikroskopichen Wahrnehmung. *Archiv Mikroskopie Anatomie* 9: 413–468.

Bracewell RN (1965) *The Fourier Transform and Its Applications*. New York: McGraw-Hill.

Cutrona LJ, Leith EN, Palermo CJ and Porcello LJ (1960) Optical data processing and filtering systems. *IRE Transactions on Information Theory* IT-6: 386–400.

Goodman JW (1968) *Introduction to Fourier Optics*. New York: McGraw-Hill.

Jutamulia S (1992) Joint transform correlators and their applications. *Proceedings of the SPIE* 1812: 233–243.

Jutamulia S and Asakura T (2002) Optical Fourier-transform theory based on geometrical optics. *Optical Engineering* 41: 13–16.

Leith EN and Upatnieks J (1962) Reconstructed wavefronts and communication theory. *Journal of the Optical Society of America* 52: 1123–1130.

Porter AB (1906) On the diffraction theory of microscopic vision. *Philosophical Magazine* 11: 154–166.

Rau JE (1966) Detection of differences in real distributions. *Journal of the Optical Society of America* 56: 1490–1494.

Tsujiuchi J (1963) Correction of optical images by compensation of aberrations and by spatial frequency filtering. *Progress in Optics* 2: 133–180.

Vander Lugt A (1964) Signal detection by complex spatial filtering. *IEEE Transactions on Information Theory* IT-10: 139–145.

Weaver CS and Goodman JW (1966) A technique for optically convolving two functions. *Applied Optics* 5: 1248–1249.

Yu FTS (1983) *Optical Information Processing*. New York: Wiley.

Yu FTS and Ludman JE (1988) Joint Fourier transform processor. *Microwave Optical Technology Letters* 1: 374–377.

Yu FTS, Jutamulia S, Lin TW and Gregory DA (1987) Adaptive real-time pattern recognition using a liquid crystal TV based joint transform correlator. *Applied Optics* 26: 1370–1372.

G

GEOMETRICAL OPTICS

Contents

Lenses and Mirrors
Aberrations
Prisms

Lenses and Mirrors

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Lenses are carefully shaped pieces of transparent material used to form images. The term 'transparent' does not necessarily mean that the lens is transmitting visible light. The element germanium, well-known in the transistor industry, is a light-gray opaque crystal which forms images with infrared radiation, and provides the optics for night-vision instruments. The lenses used in cameras, telescopes, microscopes, and other well-known applications, are made from special glasses or – more recently – from plastics. Hundreds of different optical glass formulas are listed in the manufacturers' catalogues. Plastics have come into use in recent years because it is now possible to mold them with high precision. The way in which lenses form images has been studied for several hundred years, resulting in an enormous and very complicated literature. In this article, we show how this theory can be simplified by introducing the restriction that lenses are perfect. This limitation provides an easy introduction to the image forming process. A companion article **Geometrical Optics:** Aberrations, will deal with imperfect components, showing how their defects degrade images and how corrections are made.

The Cardinal Points and Planes of an Optical System

The location, size and orientation of an image are items of information which a designer needs to know. The nature of the image can be determined by a simple model of an optical system, based on two experimental observations. The first is well-known and is pictured in Figure 1. Parallel rays of light passing through a double convex lens (a simple magnifying glass, for example) should meet at the focal point or focus, designated as F'. For rays from the right side of the lens, another such point F exists at an equal distance from the lens. The other experiment uses the lens to magnify (Figure 2a). The amount of magnification decreases as the lens is brought closer to 'ISAAC NEWTON' (Figure 2b), and when the lens touches the paper, object and image are approximately the same size. The locations of object and image corresponding to equality in size are called the unit or principal points H and H', respectively, and the set of four points F, F', H, and H' are the cardinal points. Therefore, the planes through these points normal to the axis of the lens are the cardinal planes. Figure 3 shows an object whose image we can now determine. The coordinate orientation in this figure may appear strange at first. It is customary in optics to designate the lens axis as OZ, and the other two axes form a right-handed system. The x-axis then lies in the

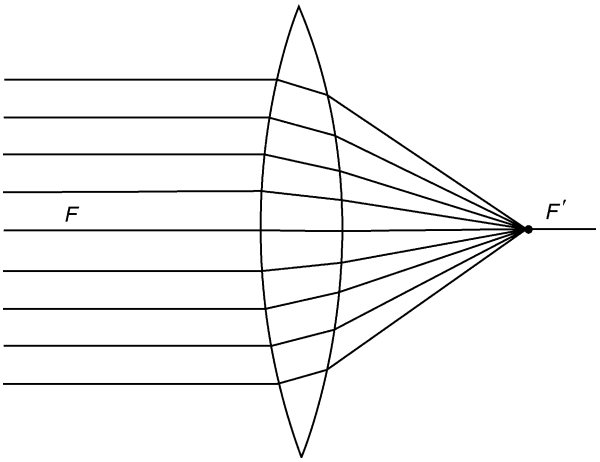


Figure 1 Focal points of a double convex lens.

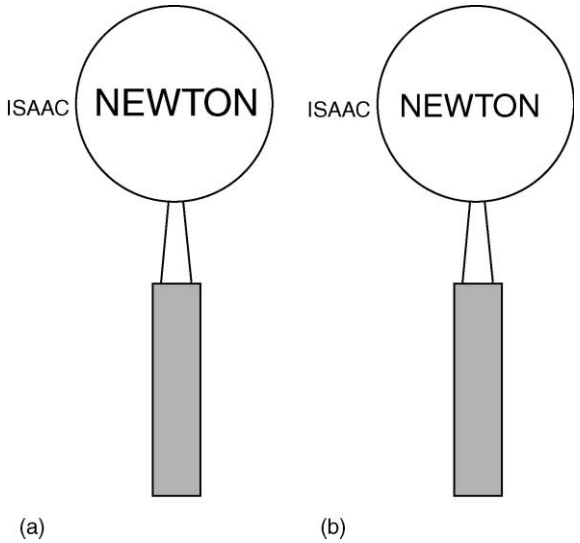


Figure 2 Magnification by a double convex lens. The magnification in (a) is reduced when the lens is brought closer to the image (b).

plane of the figure; the y-axis, coming out of the paper and perpendicular to it, is not shown. The object, of height x, is placed to the left of the focal point F. The cardinal points occur in the order F, H, H', F', since we expect – based on the experiment corresponding to Figure 2 – that the unit points are very close to the lens. Another fact to be demonstrated later is that H and H' are actually inside the lens. Two rays are shown leaving the point P at the top of the object. The ray which is parallel to the axis strikes the unit plane through H and continues to the right, completely missing the lens. The ray then meets the unit plane through H' at a point which must be a distance x from the axis. To understand why, we recognize that if the object were relocated so as to lie

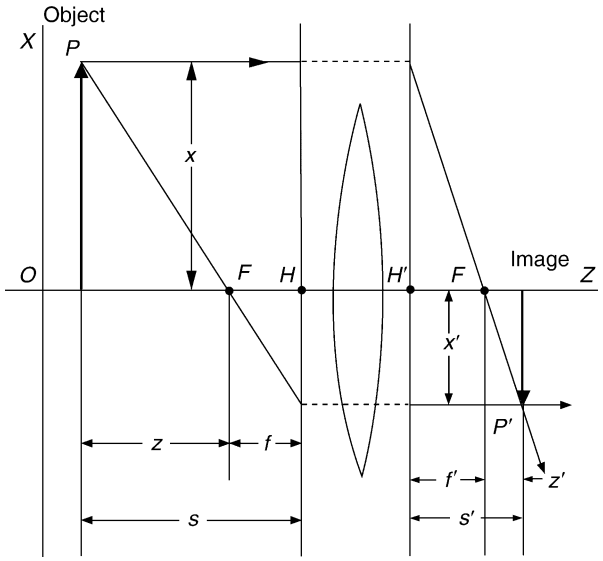


Figure 3 Ray tracing using cardinal points and planes.

on the object-space unit plane, the ray emerging at the other unit plane will be at a distance x from the axis; as required by the definition of these planes. We then assume that the ray is bent or refracted at the unit plane at H' and proceeds to, and beyond, the focal point F'. The second ray, from P through the focal point F, is easily traced since the lens is indifferent to the direction of the light. Assume temporarily that we know the location of the image point P'. Let a parallel ray leave this point and travel to the left. The procedure just given indicates that this ray will strike the unit plane at H', emerge at the other unit plane, be refracted so as to pass through F and reach the object point P. Then reversing the direction of this ray, it will start at P, emerge at the unit plane H' and travel parallel to the axis, intersecting the other ray at P'. The distance between F and H is called the object space focal length f, with f' being the corresponding image space quantity. Simple geometry, making use of the similar triangles and the distances indicated in this diagram, can be used to derive the Newton lens equation or the equivalent Gauss lens equation. Detailed derivations of these two equations will be found in either of the sources listed in the Further Reading at the end of this article. It is not usually made clear in physics texts that these equations apply only to the special case of a single, thin lens, and we shall not apply them here. The image in this figure is real, inverted, and reduced. That is, it is smaller than the object, oriented in the opposite direction, and can be projected onto a screen located at the image position. This result can be verified by using an ordinary magnifier to form the image of a distant light source on a sheet of white paper.

Paraxial Matrices

To consider how light rays behave in optical systems, we introduce the index of refraction n of a material medium, defined as the ratio of the velocity of light c in a vacuum to its velocity v in that medium, or:

$$n = \frac{c}{v} \qquad [1]$$

The velocity of light in air is approximately the same as it is in a vacuum. A light ray crossing the interface between air and a denser medium such as glass will be bent towards the normal to the surface. This is the phenomenon of refraction and the amount of bending is governed by Snell's law, which has the form:

$$n \sin \theta = n' \sin \theta' \qquad [2]$$

where n and n' are the indices of refraction of the two media and θ and θ' are the angles as measured from the normal. Figure 4 shows a ray of light leaving an object point P and striking the first surface of a lens at point P_1. It is refracted there, and proceeds to the second surface. All rays shown lie in the z, x-plane or meridional plane, which passes through the symmetry axis OZ. The amount of refraction is specified by Snell's law, eqn [2], which we shall now simplify. The sine of an angle can be approximated by the value of the angle itself, if this value is small when expressed in radians (less than 0.1 rad), and

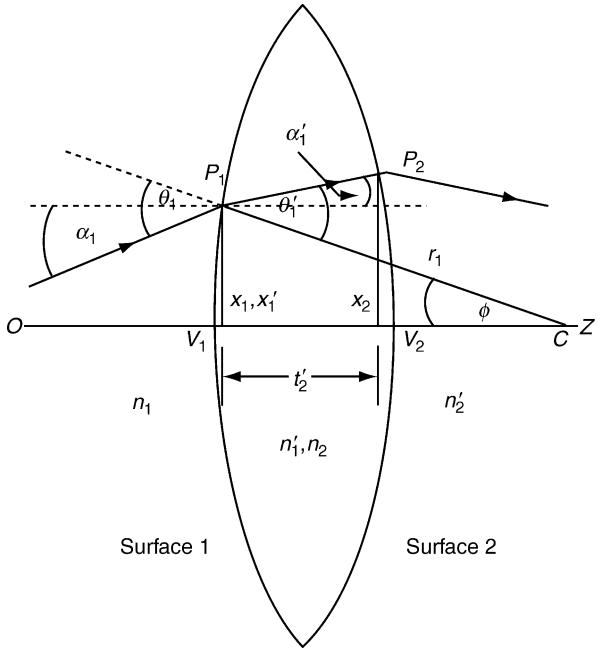


Figure 4 Ray passing through a double convex lens.

Snell's law simplifies to

$$n_1 \theta_1 = n_1' \theta_1' \qquad [3]$$

This is called the paraxial form of Snell's law; the word paraxial means 'close to the axis'. It turns out, however, that the Snell's law angles are not convenient to work with, and we eliminate them with the terms:

$$\theta_1 = \alpha_1 + \phi, \qquad \theta_1' = \alpha_1' + \phi \qquad [4]$$

where α_1 is the angle which the incident ray makes with the axis OZ, α_1' is the corresponding angle for the refracted ray, and ϕ is the angle which the radius r_1 (the line from C to P_1) of the lens surface makes at the center of curvature C. This particular angle can be specified as

$$\sin \phi = \frac{x_1}{r_1} \qquad [5]$$

but using the paraxial approximation simplifies this to

$$\phi = \frac{x_1}{r_1} \qquad [6]$$

Substituting for the angles gives

$$n_1' \alpha_1' = \frac{n_1 - n_1'}{r_1} + n_1 \alpha_1 \qquad [7]$$

Notice that the distance from the point P_1 to the axis is labeled as either x_1 or x_1'. This strange notation leads to the trivial relation

$$x_1 = x_1' \qquad [8]$$

and this can be combined with eqn [7] to obtain the matrix equation:

$$\begin{pmatrix} n_1' \alpha_1' \\ x_1' \end{pmatrix} = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} n_1 \alpha_1 \\ x_1 \end{pmatrix} \qquad [9]$$

where the constant k_1 is called the refracting power of surface 1 and is defined as

$$k_1 = \frac{n_1' - n_1}{r_1} \qquad [10]$$

To review the procedure for matrix multiplication, the first element $n_1' \alpha_1'$ in the one-column product matrix is calculated by multiplying the upper left-hand corner of the 2×2 matrix with the first element of the one-column matrix to its right, obtaining $1(n_1 \alpha_1)$, and to this is added the product $-k_1 x_1$ of the upper right-hand element in the 2×2 matrix and the second member of the one-column matrix. This square matrix is called

the refraction matrix R_1 for surface 1, defined as

$$R_1 = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} \qquad [11]$$

Next, we look at what happens to the ray as it travels from surface 1 to surface 2. As it goes from P_1 to P_2, its distance from the axis becomes

$$x_2 = x_1' + t_1' \tan \alpha_1' \qquad [12]$$

or using the paraxial approximation for small angles:

$$x_2 = x_1' + t_1' \alpha_1' \qquad [13]$$

Another approximation we shall use is to regard t_1' as being equal to the distance between the lens vertices V_1 and V_2. Using the identity:

$$\alpha_2 = \alpha_1' \qquad [14]$$

leads to a second matrix equation:

$$\begin{pmatrix} n_2\alpha_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t_1'/n_1' & 1 \end{pmatrix} \begin{pmatrix} n_1'\alpha_1' \\ x_1' \end{pmatrix} \qquad [15]$$

where the 2×2 translation matrix T_{21} is defined as

$$T_{21} = \begin{pmatrix} 1 & 0 \\ t_1'/n_1' & 1 \end{pmatrix} \qquad [16]$$

To get an equation which combines a refraction followed by a translation, we take the one-column matrix on the left side of eqn [9] and substitute it into eqn [15] to obtain

$$\begin{pmatrix} n_2\alpha_2 \\ x_2 \end{pmatrix} = T_{21} R_1 \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \qquad [17]$$

Note that the 2×2 matrices appear in right to left order and that the multiplication of two square matrices is an extension of the rule given above. This equation gives the location and slope of the ray which strikes surface 2 after a refraction and a translation. To continue the ray trace at surface 2, we introduce a second refraction matrix R_2 by expressing k_2 in terms of the two indices at this surface and the radius. Then eqn [17] is extended to give the relation:

$$\begin{pmatrix} n_2'\alpha_2' \\ x_2' \end{pmatrix} = R_2 T_{21} R_1 \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \qquad [18]$$

This process can obviously be applied to any number of components. The product of the three 2×2 matrices in the above equation is known as the system matrix S_{21} and it completely specifies the effect of the lens on the incident ray passing through it. It is also written as

$$S_{21} = R_2 T_{21} R_1 = \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \qquad [19]$$

where the four quantities a, b, c, and d are known as the Gaussian constants. They are used extensively in specifying the behavior of a lens or an optical system. We now state the sign and notation conventions which are necessary for the application of paraxial matrix methods:

1. Light normally travels from left to right.
2. The optical systems we deal with are symmetrical about the z-axis. The intersections of the refracting or reflecting surfaces with this axis are the vertices and are designated in the order encountered as V_1, V_2, etc.
3. Positive directions along the axes are measured from the origin in the usual Cartesian fashion, so that horizontal distances (that is, along the z-axis) are positive if measured from left to right. Angles are positive when measured up from the z-axis.
4. Quantities associated with the incident ray are unprimed; those for the refracted ray are primed.
5. A subscript denotes the associated surface.
6. If the center of curvature of a surface is to its right, the radius is positive, and vice versa.
7. There is a special rule for mirrors to be explained below.

Using the Gaussian Constants

Figure 5 shows the double convex lens of Figure 4 with the assumed locations of the cardinal points indicated. These positions, which we shall now determine accurately, are at distances designated as l_F, l_F', l_H, and l_H' and are measured from the associated vertex. The object position can be called t, a positive quantity which is measured to the right from object to the first vertex, or it can be called t_1 if measured in the opposite direction. For the first choice, the matrix specifying the translation from object to lens will have the quantity t/n_1 in its lower left-hand corner. However, it is both logical and convenient to use the first vertex as the reference point. Hence, we replace t/n_1 in the translation matrix with the quantity $-t_1/n_1$, and remember to specify t_1 as a negative number when calculating the image position. The equation connecting object and image is obtained by starting with this matrix, multiplying it by the system matrix of eqn [19], and finally by a translation matrix

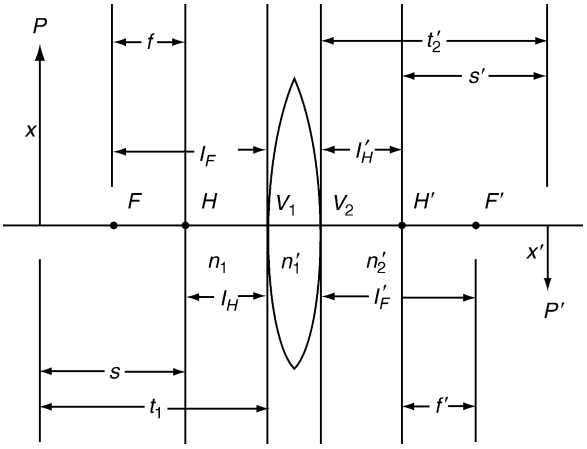


Figure 5 Definitions of cardinal plane locations.

corresponding to the translation t_2' from lens to image to obtain:

$$\begin{pmatrix} n_2'\alpha_2' \\ x' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t_2'/n_2' & 1 \end{pmatrix} \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \times \begin{pmatrix} 1 & 0 \\ -t_1/n_1 & 1 \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ x \end{pmatrix} \quad [20]$$

where α or α_1 is the angle that the ray from the top of the object makes with the z-axis. This equation takes the initial value of the inclination α and of the position x of the ray leaving the object, and determines the final values, α' and x' at the image. The product of the three 2×2 matrices on the right-hand side can be consolidated into a single matrix called the object-image matrix $S_{P'P}$. Then eqn [20] can be written as

$$\begin{pmatrix} n_2'\alpha_2' \\ x' \end{pmatrix} = S_{P'P} \begin{pmatrix} n_1\alpha_1 \\ x \end{pmatrix} \quad [21]$$

and this matrix has the complicated form:

$$S_{P'P} = \begin{pmatrix} b + \dfrac{at_1}{n_1} & -a \\ \dfrac{bt_2'}{n_2'} + \dfrac{at_1t_2'}{n_1n_2'} - d - \dfrac{ct_1}{n_1} & c - \dfrac{at_2'}{n_2'} \end{pmatrix} \quad [22]$$

If we put this matrix into the previous equation, we obtain an unsatisfactory result: the value of x' will depend on the angle α_1 made by the incident ray, as can be seen by multiplying the two matrices on the right side. The ratio of x' to x is called the magnification m; that is

$$m = \frac{x'}{x} \quad [23]$$

A perfect image can be formed only if the magnification is determined solely by the object distance and the constants of the lens. To eliminate this difficulty, the lower left-hand element in the matrix, eqn [22] is required to be equal to zero, and the magnification is then:

$$m = \frac{x'}{x} = c - \frac{at_2'}{n_2'} \quad [24]$$

The determinant of this matrix is unity, since it is the product of three matrices whose individual determinants are unity. It follows that

$$\frac{1}{m} = b + \frac{at_1}{n_1} \quad [25]$$

so that

$$\begin{pmatrix} n_2'\alpha_2' \\ x' \end{pmatrix} = \begin{pmatrix} 1/m & -a \\ 0 & m \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ x \end{pmatrix} \quad [26]$$

Using the fact that the lower left-hand element of the matrix eqn [22] is zero shows that

$$\frac{t_2'}{n_2'} = \frac{d + ct_1/n_1}{b + at_1/n_1}, \qquad \frac{t_1}{n_1} = \frac{d - bt_2'/n_2'}{-c + at_2'/n_2'} \quad [27]$$

This equation connects the object distance t_1 with the image distance t_2', both quantities being measured from the appropriate vertex. This relation is known as the generalized Gauss lens equation. It applies to all lens systems, no matter how complicated.

We can determine the location of the unit planes by letting $m = 1$. This t_2' in eqn [27] is the location l_H' of the image space unit plane, and it becomes

$$l_H' = \frac{n_2'(c-1)}{a} \quad [28]$$

This relation expresses the location of the unit plane on the image side as the distance from V_2 to H'. Similarly, the location of H with respect of V_1 is

$$l_H = \frac{n_1(1-b)}{a} \quad [29]$$

To locate the focal planes, consider a set of parallel rays coming from infinity and producing a point image at F'. The terms containing t_1 in eqn [27] are

much larger than d or b, and this relation becomes

$$l'_F = \frac{n'_2 c}{a} \qquad [30]$$

Letting t'_2 become infinite in the right-hand half of eqn [27], the location of F is given by the equation

$$l_F = \frac{-n_1 b}{a} \qquad [31]$$

The focal lengths f and f' shown in **Figure 5** were defined earlier as the distances between their respective focal and unit planes. Hence

$$f' = l'_F - l'_H = \frac{n'_2}{a} \qquad [32]$$

and

$$f = l_F - l_H = \frac{-n_1}{a} \qquad [33]$$

If we let $n_1 = n'_2 = 1$ for air, these become

$$-f = f' = \frac{1}{a} \qquad [34]$$

The Gaussian constant a is thus the reciprocal of the focal length in air. Even when the lens is not in air, it is still true that $f' = -f$ if the lens is in a single medium. It is also true regardless of whether or not the lens is symmetric.

It is customary in optics to work with a consistent set of units, such as centimeters or millimeters, and the sign conventions used for x and y are the standard Cartesian rules. As mentioned above, curved surfaces which open to the right have a positive radius and vice versa. As an example, let a double convex lens have radii of 2.0 and 1.0, respectively, an index of refraction of 1.5, and a thickness of 0.5. These quantities are then denoted as

$$r_1 = +2.0, \qquad r_2 = -1.0, \qquad t'_1 = t_2 = +0.5$$
$$n_1 = 1.0, \qquad n'_1 = 1.5 = n_2, \qquad n'_2 = 1.0 \qquad [35]$$

By eqn [10]:

$$k_1 = \frac{n'_1 - n_1}{r_1} = \frac{1.5 - 1.0}{2.0} = 0.25 \qquad [36]$$

and

$$k_2 = \frac{1.0 - 1.5}{-1.0} = 0.50 \qquad [37]$$

The system matrix S_{21} is

$$S_{21} = R_2 T_{21} R_1 = \begin{pmatrix} 1 & -0.50 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0.5/1.5 & 1 \end{pmatrix} \begin{pmatrix} 1 & -0.25 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.83 & -0.71 \\ 0.33 & 0.92 \end{pmatrix} \qquad [38]$$

and we note that the determinant is

$$(0.83)(0.92) - (0.71)(-0.33) = 1.00 \qquad [39]$$

This property of the system matrix provides a very useful check on the accuracy of matrix multiplications. The locations of the four cardinal points can be determined from the formulas given above and their locations are shown in the scale drawing of **Figure 6**. Since we know how to locate the cardinal points of a lens, we are in a position to ray trace in a precise manner, using the method of **Figure 3** and adding a third ray. After locating F, H, H', and F', we no longer need the lens; the cardinal planes with the proper separation are fully equivalent to the lens they replace. This is very much like what electrical engineers do when they replace a complicated circuit with a black box; this equivalent circuit behaves just like the original. Ray tracing becomes more complicated when we deal with diverging lenses (**Figure 7**), but the procedure gives above still applies. Parallel rays spread apart and do not come to a focus on the right-hand side of the lens. However, if the refracted

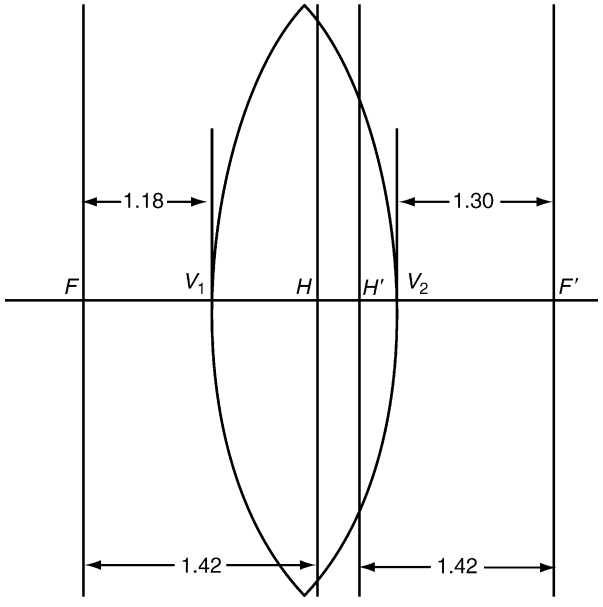


Figure 6 Positions of cardinal planes for an asymmetric lens.

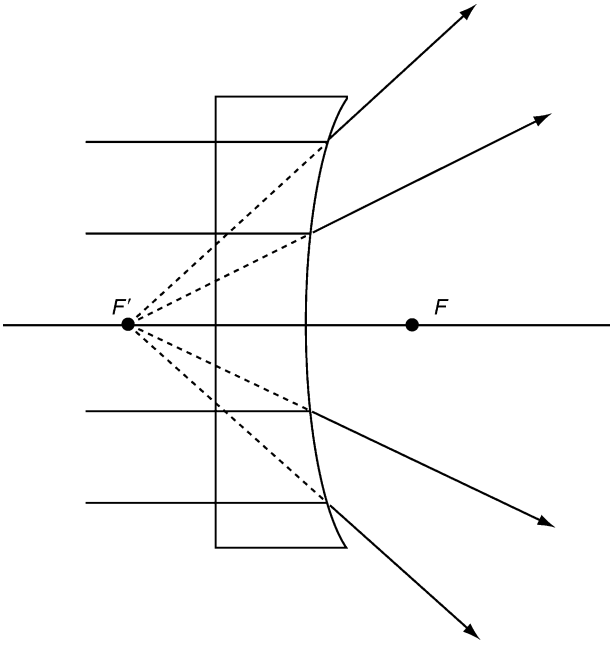


Figure 7 Ray tracing for a planar-concave lens.

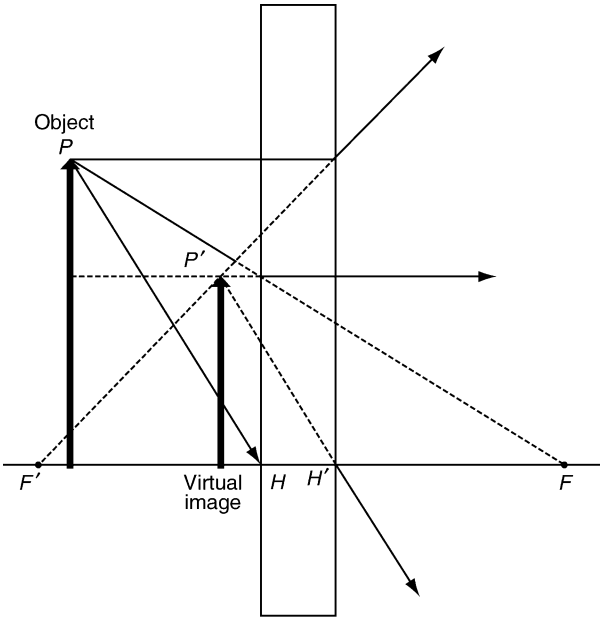


Figure 9 Formation of a virtual image by a concave lens.

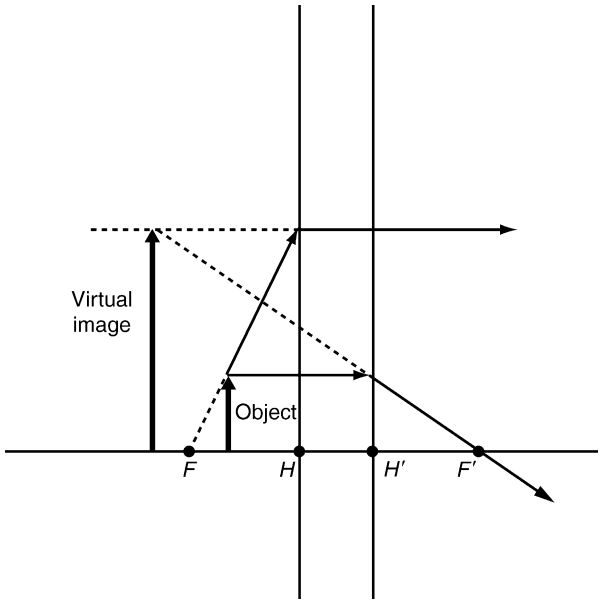


Figure 8 Formation of a virtual image by a convex lens.

rays are extended backwards, these extensions will meet as indicated. This behavior can be verified quantitatively by using the formulas developed above. It will be found that F and F' have exchanged places, but H and H' retain their original positions inside the lens. If an object is placed between the focal point F and the first vertex of a convex lens (Figure 8), then our usual procedure produces two rays which do not intersect in image space. The ray from the object which is parallel to the axis is refracted downward so that it does not intersect the other ray in image space. However, the extensions to the left of these two rays meet to form an image which is erect, magnified, and virtual. This is the normal action of a magnifying lens; the image can be seen but cannot be projected onto a screen, unlike the real, inverted image of Figure 3. This ray tracing diagram confirms what would be seen when a double convex lens is used to magnify an object lying close to the lens. For the double concave lens of Figure 9, a parallel ray leaves an object point P and is refracted upward at the unit plane H' in such a way that its extension, rather than the ray itself, passes through F'. The ray headed for F is refracted at H before it can reach the object-space focal plane and becomes parallel to the axis. The third ray, going from P to H, emerges at H' parallel to its original direction and its extension to the left passes through the image point P' already determined by the intersection of the other two rays. The resulting virtual image is upright and reduced. All three rays in this diagram behave exactly like the corresponding rays in Figure 3; the only change is the use of the extensions to locate the image. The eye receives the diverging rays from P and believes that they are coming from P'.

Nodal Points and Planes

Let a ray leave a point on the z-axis at an angle α_1, and have an angle α_2' when it reaches image space.

Since $x = 0$ at the starting point, eqn [26] shows that

$$n'_2\alpha'_2 = n_1\alpha_1/m \qquad [40]$$

The ratio of the final angle to the initial angle in this equation is the angular magnification μ, which is then

$$\mu = n_1/mn'_2 \qquad [41]$$

The locations of object and image for unit angular magnification are called the nodal points, labeled as N and N'. The above definition shows that the linear and angular magnification are reciprocals of one another when object and image are in air or the same medium. The procedure that located the unit points will show that the nodal points have positions given by

$$\frac{l_N}{n_1} = \frac{(n'_2/n_1) - b}{a} \qquad [42]$$

and

$$\frac{l'_N}{n'_2} = \frac{c - (n_1/n'_2)}{a} \qquad [43]$$

When the two indices are identical, these relations are identical to those for the points H and H'. That is why the rays from P to H and H' to P' in Figure 3 are parallel.

Compound Lenses

A great advantage of the paraxial matrix method is the ease of dealing with systems having a large number of lenses. To start simply, consider the cemented doublet of Figure 10: a pair of lenses for which surface 2 of the first element and surface 1 of the second element match perfectly. The parameters of this doublet are given in the manner shown below.

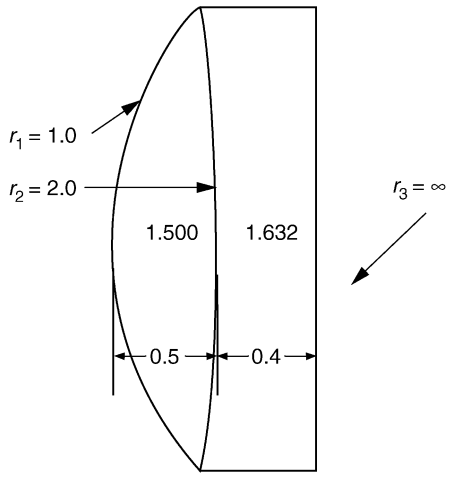


Figure 10 Specification for a cemented doublet.

It is understood that the first entry under r is r_1, the second entry is r_2, and so on. Note that this doublet has only three surfaces; if the two parts were separated by an air space, producing an air-spaced doublet, then there would be four surfaces to specify. The values of the indices n' and the spacings t' are placed between the values of r. The constants of the doublet are then

r	n'	t'
1.0		
	1.500	0.5
−2.0		
	1.632	0.4
∞		

[44]

Note that surface 3, which is flat, has a radius of infinity. Numbering the vertices in the usual manner, the system matrix is

$$S_{31} = R_3T_{32}R_2T_{21}R_1 \qquad [45]$$

where

$$k_2 = (n'_2 - n_2)/r_2 = (1.632 - 1.500)/-2.0 \qquad [46]$$

Approximations for Thin Lenses

An advantage of paraxial matrix optics is the ease of obtaining useful relations. For example, combining the several expressions for the Gaussian constant a, we obtain:

$$a = -\frac{n_1}{f} = \frac{n'_2}{f'} = k_1 + k_2 - \frac{k_1k_2t'_1}{n'_1} \qquad [47]$$

and using the definitions of k_1 and k_2 with the lens in air leads to

$$\frac{1}{f'} = (n'_1 - 1)\left[\frac{1}{r_1} - \frac{1}{r_2} + \frac{(n'_1 - 1)t'_1}{n'_1r_1r_2}\right] \qquad [48]$$

which is the lensmakers' equation. It tells how to find the focal length of a lens from a knowledge of its material and geometry. This derivation is much more direct than what you will usually find. This equation becomes simpler when we are dealing with thin lenses – those for which the third term in brackets can be neglected by assuming that the lens thickness

is approximately zero. Then

$$\frac{1}{f'} = (n'_1 - 1)\left[\frac{1}{r_1} - \frac{1}{r_2}\right] \qquad [49]$$

which is the form often seen in texts. We then realize that the original form is valid for lenses of any thickness, but the thin lens form can be used if the spacing is much less than the two radii. It is also seen that when the thickness is approximately zero, then

$$b = c = 1 \qquad [50]$$

and the unit planes have locations

$$l_H = l'_H = 0 \qquad [51]$$

Thus, they are now coincident at the center of the lens. A ray from any point on the object would pass undeviated through the geometrical center of the lens, as is often shown in the elementary books. These new values of the Gaussian constants give a system matrix of the form:

$$S_{21} = \begin{pmatrix} 1 & -1/f' \\ 0 & 1 \end{pmatrix} \qquad [52]$$

This matrix contains a single constant, the focal length of the lens, so that the preliminary design of an optical system is merely a matter of specifying the focal length of the lenses involved and using the resulting matrices to determine the object-image relationship. We also note that this matrix looks like a refraction matrix; the non-zero element is in the upper right-hand corner. This implies that the refraction–translation–refraction procedure that actually occurs can be replaced, for a thin lens, by a single refraction occurring at the coinciding unit planes. Ophthalmologists take advantage of the thin lens approximation by specifying focal lengths in units called diopters. The reciprocal of the focal length in meters determines its refraction power in diopters. For example, a lens with $f' = 10$ cm will be a 10 diopter lens. If two lenses are placed in contact, the combined power is the sum of the individual powers, for multiplying matrices of the above form gives the upper right-hand element as $(-1/f'_1 - 1/f'_2)$.

The Paraxial System for Design or Analysis

The material given in this article can serve as the basis for an organized way of taking the specifications of an optical system and using them to gain a full understanding of this system. We shall demonstrate it with a practical example which has some interesting features. Chemical engineers have long known that the index of refraction of a liquid can be determined by observing the empty bore of a thick glass tube with a telescope having a calibrated eyepiece and then measuring the magnification of the bore when the liquid flows through it. Consider a liquid with an index of 1.333 and a tube with bore radius of 3 cm, outer radius of 4 cm, and glass of index equal to 1.500. The first step is to specify the parameters of the optical system. These are the radii $r_1 = -3$, $r_2 = -4$ and the thickness $t'_1 = 1$, regarding the tube as a concentric lens with an object to its left; the indices $n_1 = 4/3$ (fractions are convenient), $n'_1 = 3/2$, $n'_2 = 1$. Calculating the elements of the two refraction matrices and the translation matrix, the system matrix is

$$S_{21} = \begin{pmatrix} 1 & -1/8 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 2/3 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1/18 \\ 0 & 1 \end{pmatrix} \qquad [53]$$

Multiplying, the Gaussian constants are

$$a = 2/27, \quad b = 11/12, \quad c = 28/27, \quad d = -2/3 \qquad [54]$$

An important check is to verify that $bc - ad = 1$, as is the case here. The expressions given previously for the positions of the cardinal points then lead to the values

$$\begin{aligned} &l_H = 3/2, \qquad l'_H = 1/2, \qquad l_F = -16\ 1/2, \\ &l'_F = 14, \qquad l_N = -3, \qquad l'_N = -4 \end{aligned} \qquad [55]$$

These points are shown in Figure 11. Although the same scale has been used for both horizontal and vertical distances, this is usually not necessary; the vertical scale, which merely serves to verify the magnification, can be arbitrary. The entire liquid column is the object in question, but it is simpler to use a vertical radius as the object. Then the image, generated as described in all the previous examples,

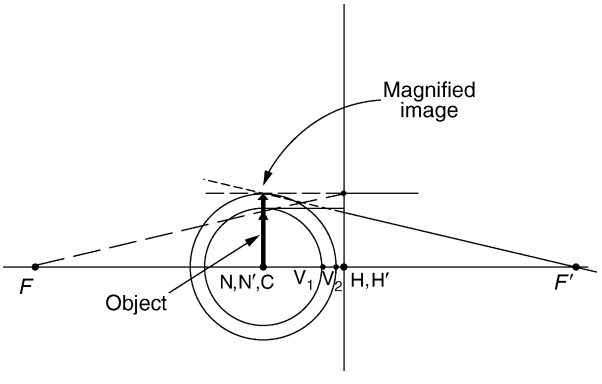


Figure 11 Ray tracing diagram for a thick-walled glass tube.

coincides with the object and it is enlarged, erect, and virtual. We confirm the image location shown by using the generalized Gaussian lens equation, giving

$$t' = [(-2/3) + (28/27)(-3)/(4/3)]/[(11/12) + (2/27)(-3)/(4/3)] = -4 \quad [56]$$

and this agrees with the sketch. The denominator of this expression was shown to be the reciprocal of the magnification $1/m$ with a value of 3/4, and m itself can be calculated from t' as $c - at' = 4/3$; this is another important check on the accuracy of the calculation; the sketch obeys this conclusion, and the purpose of this analysis is confirmed.

Reflectors

To show how reflecting surfaces are handled with matrices, consider a plane mirror located at the origin and normal to the z-axis (Figure 12). The ray which strikes it at an angle α_1 leaves at an equal angle, resulting in specular reflection. Since $k = 0$ for a flat surface, the refraction matrix reduces to the unit matrix and by eqn [9]:

$$n'_1\alpha'_1 = n_1\alpha_1 \quad [57]$$

This contradicts Figure 12, since α_1 and α'_1 are opposite in sign.

If, however, we specify that

$$n'_1 = -n_1 \quad [58]$$

then the difficulty is removed. Hence, reflections involve a reversal of the sign on the index of refraction, and two successive reflections restore the original sign. For a mirror, the refraction power is

$$k_1 = \frac{n'_1 - n_1}{r_1} = \frac{-1 - (1)}{r_1} = -\frac{2}{r_1} \quad [59]$$

The system matrix for a single refracting surface thus becomes

$$S_{11} = R_1 = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2/r_1 \\ 0 & 1 \end{pmatrix} \quad [60]$$

with

$$a = -2/r_1, \qquad b = c = 1, \qquad d = 0 \quad [61]$$

The connection between object and image can be expressed as

$$\begin{pmatrix} n'_1\alpha'_1 \\ x'_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t'/n'_1 & 1 \end{pmatrix} \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -t/n_1 & 1 \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \quad [62]$$

and using the same procedure that was applied to eqns [28]–[31], the locations of the six cardinal points are

$$l_F = -n_1 b/a, \quad l'_F = n'_1 c/a$$

$$l_H = n_1(1-b)/a, \quad l'_H = n'_1(c-1)/a \quad [63]$$

$$\frac{l_N}{n_1} = \frac{(n'_1/n_1) - b}{a}, \quad \frac{l'_N}{n'_1} = \frac{c - (n_1/n'_1)}{a}$$

so that for the spherical mirror:

$$l_F = \frac{(-1)(1)}{-2/r_1} = \frac{r_1}{2} = l'_F \quad [64]$$

$$l_H = l'_H = 0 \quad [65]$$

and

$$l_N = \frac{-1-1}{-2/r_1} = r_1 \quad [66]$$

but

$$\frac{l'_N}{-1} = \frac{1-(-1)}{-2/r_1}, \quad l'_N = r_1 \quad [67]$$

Since r_1 is negative, these equations show that the unit points lie on the vertex, the foci coincide halfway between the center of curvature and the

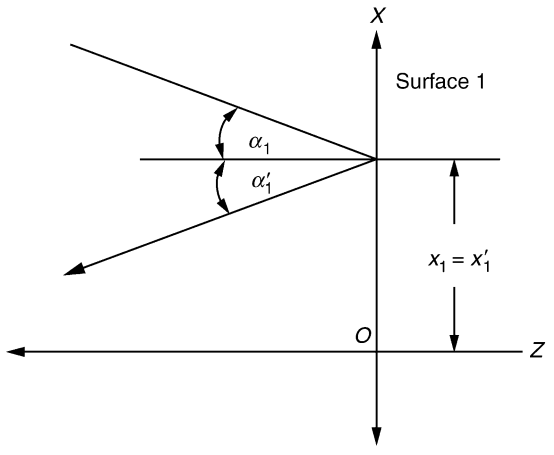


Figure 12 Specular reflection.

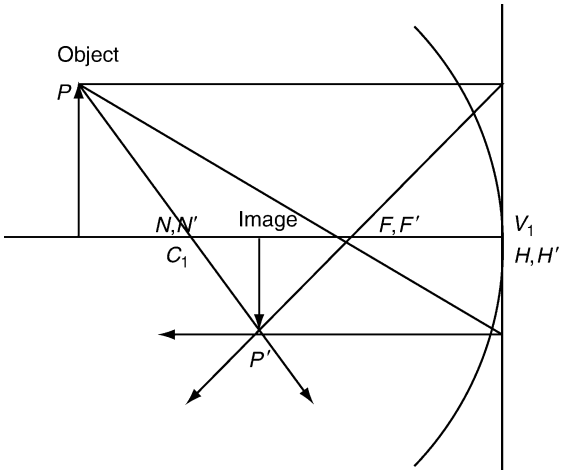


Figure 13 Ray tracing for a concave mirror.

vertex, and the nodal points are at this center. The perfect focusing is a consequence of the paraxial approximation. The true situation is presented in **Geometrical Optics:** Aberrations. Ray tracing for this mirror uses the procedure developed for lenses, which applies to optical systems of any complexity. Figure 13 shows an object to the left of the center of curvature. The ray from P parallel to the axis goes to H' and then through F', while the ray through F goes to H and then becomes parallel to the axis. Their intersection at P' produces a real, inverted image. The third set of rays represents something different: it requires a knowledge of the nodal point locations. The ray from P to N should be parallel to the ray from N' to P', by definition; in this example, they meet this requirement by being colinear.

As an example of the power of matrix optics to simplify the design or analysis of an optical system involving lenses and mirrors, consider an object that is 15 units to the left of a converging lens, with focal length $f' = 10$. A concave mirror, of radius $r_1 = 16$, is 20 units to the right of the lens. The refraction power of the mirror is $k_1 = (-1 - 1)/-16 = 1/8$ and using the thin lens form of the system matrix, the combined matrix is

$$\begin{pmatrix} 1 & -1/8 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 20 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1/10 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -3/2 & 1/40 \\ 20 & -1 \end{pmatrix} \quad [68]$$

This gives the constants a, b, c, and d, from which we find that the image is $-4\ 4/9$ units to the left of the mirror, its magnification is $-8/9$, and it is inverted. Compare the simplicity of this procedure with the usual way of doing this calculation, which involves first finding the image in the lens and then using that image as a virtual object.

Conclusion

It has been shown that the use of paraxial matrices provides a simple approach to an understanding of the kind of optical systems that provide perfect images. Further details and derivations of expressions given here will be found in the Further Reading below.

See also

Geometrical Optics: Aberrations.

Further Reading

Brouwer W (1964) *Matrix Methods in Optical Instrument Design*. New York: WA Benjamin.

Nussbaum A (1998) *Optical System Design*. Upper Saddle River, NJ: Prentice-Hall.

Aberrations

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The article **Geometrical Optics:** Lenses and Mirrors, describes image formation by components which are optically perfect. The images may be magnified or reduced and they may be erect or inverted, but they are otherwise faithful replicas of the object. This behavior is a consequence of ray tracing based on an approximate or paraxial form of Snell's law which we write as

$$n\theta = n'\theta' \quad [1]$$

where n and n' are the indices of refraction on either side of an interface and θ and θ' are the angles of incidence and refraction. If, however, the following exact form of Snell's law:

$$n \sin\theta = n' \sin\theta' \quad [2]$$

governs the ray behavior, as is the case when the rays have large inclinations with respect to the symmetry axis, then it is necessary to devise a more involved ray

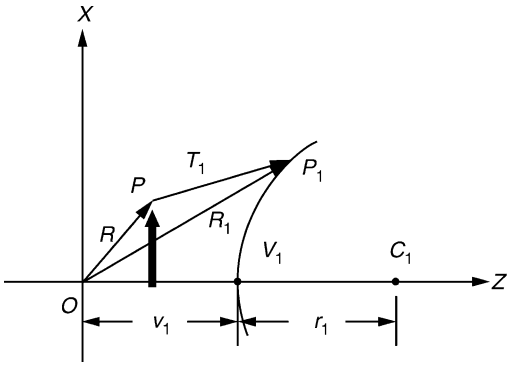


Figure 1 Ray tracing in the meridional plane.

tracing procedure. Figure 1, which shows the first surface of a lens, will be used to explain how this procedure works. A ray leaves the object point P and strikes the lens surface at the point P_1. Rays which lie completely in the plane ZOX are said to be meridional; this is the plane which passes through the symmetry axis, just as meridians on the Earth's surface are determined by planes through the geographic axis. By regarding the various distances in the figure as vectors, it is possible to start with the initial coordinates of the ray and with the direction-cosines which specify its slope and find the coordinates of the point P' where the ray strikes the lens surface. A derivation of the equation governing this process will be found in the text of A Nussbaum (see Further Reading). This derivation involves only elementary algebra and vector analysis, and will be found simpler than the usual textbooks treatment of ray tracing procedures. The equation is

$$T_1 = \frac{F}{-E + \sqrt{E^2 - c_1 F}} \qquad [3]$$

where

$$E = Bc_1/2 = c_1[(z - v_1)N + xL] - N \qquad [4]$$

and

$$F = Cc_1 = c_1\left[(z - v_1)^2 + x^2\right] - 2(z - v_1) \qquad [5]$$

Knowing the coordinates z and x of the starting point P of the ray, as well as its starting cosines N and L, plus the curvature c_1 or radius $r_1 = 1/c_1$ of the lens surface and the location v_1 of the surface's vertex, we can readily calculate the length T_1 of the ray. Then its components on the two axes give the coordinates of P'. Next, we need to find the slope of the ray after it is refracted at the lens surface, and this can be calculated from Snell's law combined with the behavior of the incident and refracted rays regarded as vectors. Again we invoke the reference cited above for a derivation of the equations giving the value of the direction-cosines after refraction; these expressions are

$$n'_1 N'_1 = K_1/c_1 + n_1 N_1 - K_1 z_1 \qquad [6]$$

and

$$n_1 L'_1 = n_1 L_1 - K_1 x_1 \qquad [7]$$

where K_1 is known as the refracting power and has the definition

$$K_1 = c_1(n'_1 \cos\theta' - n_1 \cos\theta) \qquad [8]$$

The ray can now be traced to the next surface knowing the lens thickness, and then to the image plane. These two sets of equations, when incorporated into a computer program of about two dozen lines, will trace meridional rays to any desired accuracy through an optical system containing an arbitrary number of lenses and spherical reflectors. As an example, consider a symmetric double convex lens with radii of 50 units at surface 1, −50 units at surface 2, a thickness of 15 units, and an index of 1.50. Rays parallel to the axis – one at 0.2 units above the axis, a second one at 2 units above the axis, and a third one at 20 units above the axis – are started well to the left and traced to the paraxial focal plane, which is at 47.368 units from the second surface. The ray that is 0.2 units above the axis will cross the focal plane at the axis; this is a paraxial ray, passing through the paraxial focal point F'. The ray which is 2 units above the axis crosses the focal plane slightly below the focal point, and is therefore not quite paraxial, while the ray which starts at 20 units above the axis falls well short of the focal point, intersecting the axis before it reaches the focal plane. This behavior indicates that the lens possesses a defect known as spherical aberration and these calculations imply that spherical aberration should be defined as the failure of nonparaxial rays to conform to the behavior of paraxial rays. This definition does not appear in optics texts, and the few definitions that are given appear to be incorrect. As clearly indicated in **Geometrical Optics:** Lenses and Mirrors, the focal point of a lens is strictly a paraxial concept and it is a constant, obtainable from the parameters of the lens or lens system. It is meaningless to speak of a focal point as defined by a pair of nonparaxial rays; this infinite set of ray-pairs determines a corresponding number of intersection points, all of which fall short of the true focal point F'. This situation is nicely illustrated by the next figure, which shows a large number of parallel rays, the central ones being paraxial and the others behaving differently. The rays which are very close to the axis will meet as

expected at the focal point F'. The rays a little farther out – the intermediate rays – will cross the axis just a little to the left of F', and those near the edge of the lens – the *marginal* rays – will fall very short, as indicated. Rotating this diagram about the axis, we realize that spherical aberration produces a very blurry circular image at the paraxial plane. The three-dimensional envelope of the group of rays produced by this rotation is known as the caustic surface and the narrowest cross-section of this surface, slightly to the left of the paraxial focal plane, is called the circle of least confusion. If this location is chosen as the image plane, then the amount of spherical aberration can be reduced somewhat. Other methods of improving the image will be considered below. Figure 2 was obtained by adding a graphics step to the computer program described above. Figure 3 shows how to define spherical aberration quantitatively. Two kinds of aberration are specified in this figure. The place where the ray crosses the axis, lying to the left of F', is at a distance, measured from the paraxial focus, called the longitudinal spherical aberration (LSA), while the distance below the axis at the focal plane is the transverse spherical aberration (TSA). To reduce this aberration, note in Figure 2 that the amount of refraction at the two surfaces of the lens for marginal rays is not the same. Equalizing the refraction at the two surfaces will improve the situation. Suppose we alter the lens while keeping the index, thickness, and focal length constant; only the radii will change. This can be done if the new radii satisfy the paraxial equation giving the focal length in terms of the lens constants. The process of varying the shape of lens, while holding the other parameters constant, is called *bending the lens*. This can be easily accomplished by a computer program, obtaining what are known as optimal shape lenses, which are commercially available. To study their effect on spherical aberration, define the shape factor σ of a lens as

$$\sigma = \frac{r_2 + r_1}{r_2 - r_1} \qquad [9]$$

from which $\sigma = 0$ for a symmetric lens ($r_1 = -r_2$). Figure 4a shows the monitor screen for a symmetric lens, and Figure 4b illustrates the result of lens bending. This lens has had the curvature of the first surface raised and the second surface has been flattened to keep the focal length constant. Let us now consider what lens designers can do about spherical aberration. Brouwer in his book (see Further Reading) gives the specifications for a cemented doublet consisting of a double convex lens followed by a lens with two concave surfaces. The longitudinal spherical aberration of the front lens alone varies as shown in Figure 5. The way that the designer arranged for the marginal rays to meet at the paraxial focus, rather than falling short, was to recognize that a diverging lens following the front element would refract the rays away from – rather than closer to – the axis. When this second component is added, we would expect the behavior shown in Figure 6. Note that the value of LSA for the doublet has been enormously reduced, as indicated by the horizontal scale change. Spherical aberration can be reduced in more elaborate systems as part of the design process. Telescopes, in particular, make use of corrector plates which are placed in front of the large mirrors.

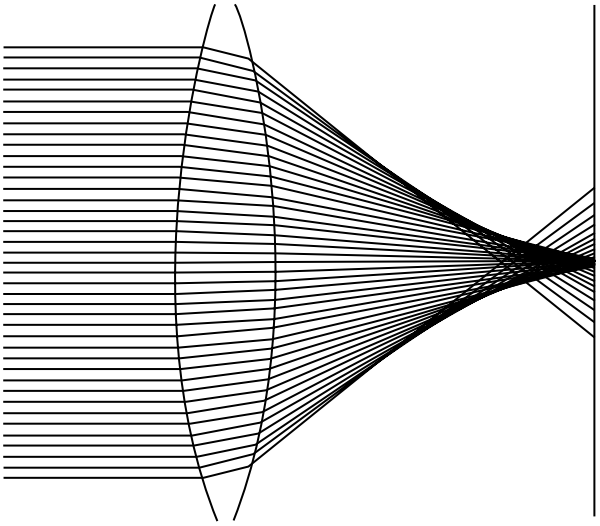


Figure 2 Spherical aberration for parallel meridional rays.

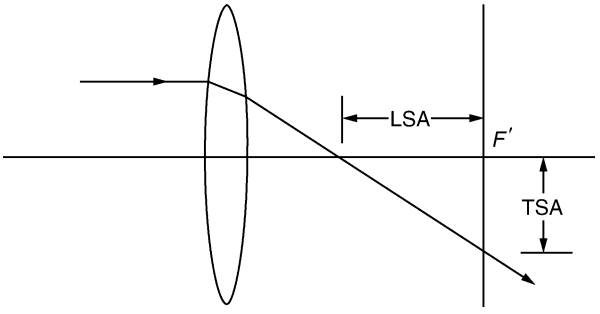


Figure 3 Definitions of transverse and longitudinal spherical aberration.

Coma

We have seen how to trace nonparaxial, meridional rays and found that spherical aberration appears because the intermediate and marginal rays do not behave like the central rays. We now wish to deal

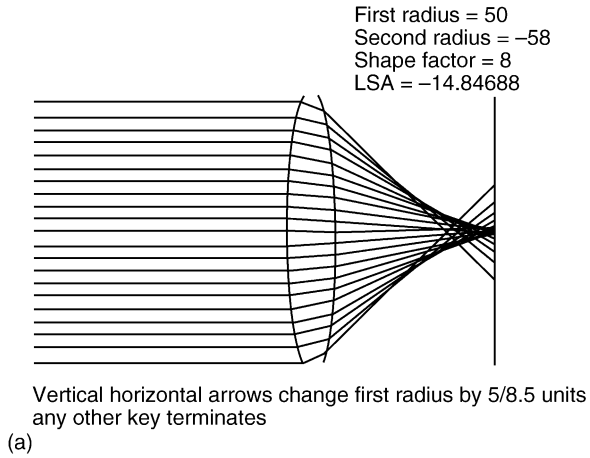


Figure 4 Lens bending as performed numerically.

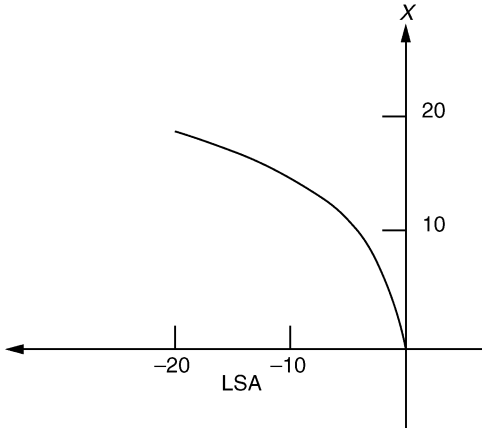


Figure 5 Spherical aberration for a single lens.

with rays that do not lie in a meridional plane; these are nonmeridional or skew rays. Calculating the behavior of such a ray would appear to be a complicated problem, but it turns out to be a simple extension of what has already been accomplished. Let a skew ray start at an arbitrary point in space with coordinates x, y, and z and let it have direction

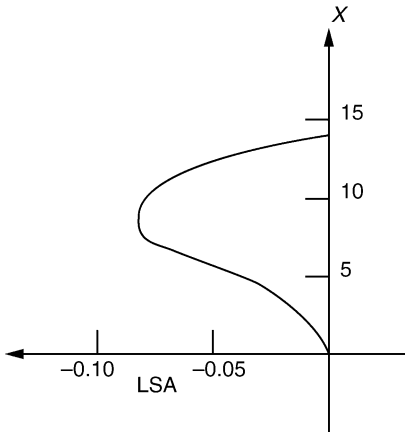


Figure 6 Reduction of spherical aberration by a doublet.

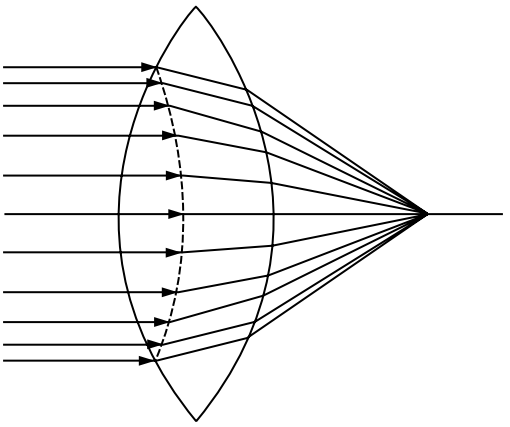


Figure 7 Configuration used to illustrate coma.

cosines L, M, and N with respect to OX, OY, and OZ. The points P and P_1 in Figure 1 are now out of the plane ZOX and when the vectors in this figure are expressed as the sum of three, rather than two, components, the equations given above for meridional rays acquire terms in y and M which have the same form as those for x and L. Now trace a set of skew rays chosen to lie on a cylinder (Figure 7) which is centered about the z-axis of a lens, so that all the rays are meridional. These rays are fairly far from the axis, so that they are also nonparaxial, but they all meet at a common image point, forming a cone whose apex is this image point. Then give this cylinder a downward displacement while holding fixed the intersection point of each ray with the dotted circle on the front of the lens. The tilting of the cylinder changes all its rays – except the top and the bottom rays – from meridional to skew. In addition, the cone on the image side will then tilt upwards; it should not change in any other way if the skew rays continue to meet at a well-defined apex. But this is not what happens in a simple lens, as will now be explained.

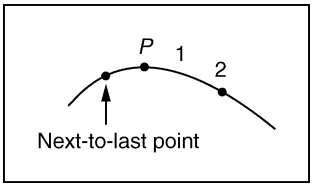


Figure 8 Starting a coma plot.

Let the top and bottom meridional rays meet at a point P in image space (Figure 8) and use this point to determine the location of an image plane. The skew ray just below the uppermost ray in Figure 7 will pass through this plane fairly close to the intersection point. Let us guess (and confirm later) that it will pass through a point slightly to the right of P and a little below it, as we observe this plane from the lens position. This is point 1 in Figure 8. The next skew ray on the front of the lens will strike the image plane still farther below and to the right, producing point 2 in the figure. As we continue to trace these rays down the front half of the circle, we realize that a ray almost at the bottom of Figure 7 will have to be very close to the lower meridional ray and that its image point is just below but to the left of P; this is the point labeled 'next-to-last' in Figure 8. All of these points join to form a closed curve. The back halves of the dotted circle and of the ray cylinder are not shown in the figure, but since we are again starting at P in Figure 8, the pattern will repeat itself. That is, we get a second closed curve which – by symmetry – should be identical to the first curve. This is a most unusual effect; double-valued phenomena are quite rare. To confirm this prediction, we use the skew ray equations, and trace the pattern formed by a set of concentric cylinders of different sizes. The appearance of the coma pattern generated by the computer is shown in Figure 9. Each half-circle on the front side of the lens produces a closed figure (an approximation to a circle) as an image. If coma were the only aberration (when it is called ideal coma), the plots would be the perfect, coinciding circles of Figure 10. The comet-like shape explains where the name comes from. We can picture the generation of these circles in the way shown in this figure. The ray through the geometrical center of the lens produces a point in the paraxial focal plane. The top and bottom rays (the meridional rays) from the circles of increasing size produce points which define image planes that are successively closer to the lens; this is a consequence of spherical aberration. Moving the top ray parallel to itself and along the half-circle generates the image plane pattern. The lines tangential to the coma circles make an angle of 60° with each other, as can be seen by measurement on the computer output. Most optics

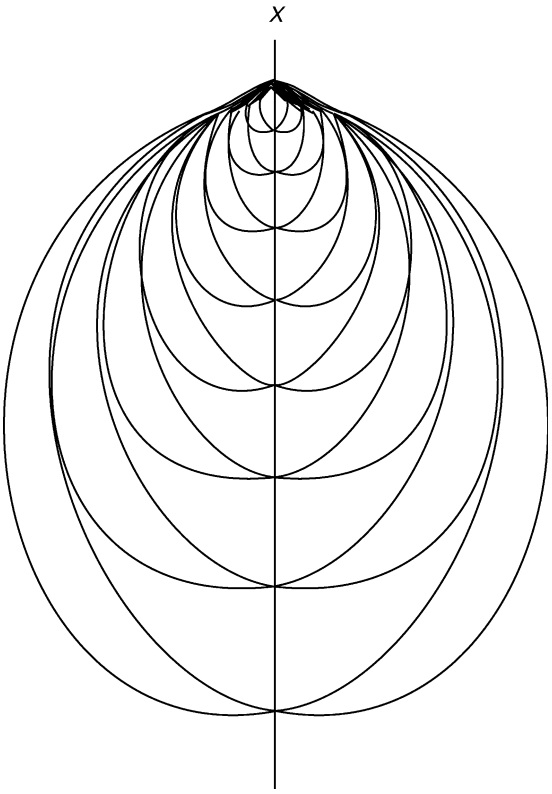


Figure 9 Coma as calculated numerically.

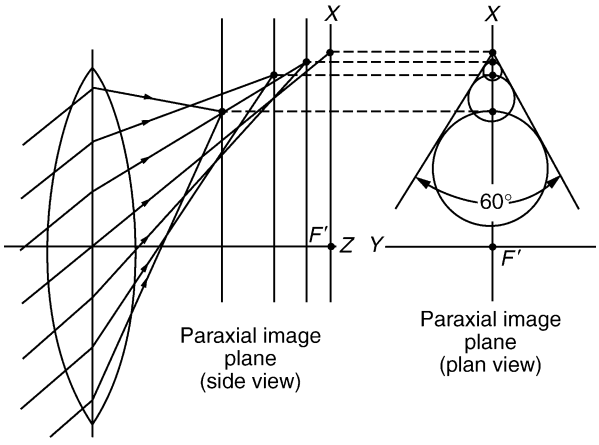


Figure 10 Customary way of illustrating coma.

books look at coma in a way different from that given here. Coma has been defined by WJ Smith (see Further Reading) as the variation of magnification with aperture, analogous to the definition of spherical aberration as the variation of focal length with aperture. And as was previously mentioned, focal length and magnification are purely paraxial concepts, not applicable to the description of aberrations. The definition that comes from the arguments

given above is that coma is the failure of skew rays to match the behavior of meridional rays, just as spherical aberration is the failure of meridional rays to match the behavior of paraxial rays. Most texts show coma in its ideal form.

Astigmatism

We have shown that spherical aberration is the failure of meridional rays to obey the paraxial approximation and coma is the failure of skew rays to match the behavior of meridional rays. To see the connection between these two aberrations, consider the object point P of Figure 11. This very complicated diagram can actually be very helpful in understanding the third of the five geometrical aberrations. Let P be the source of a meridional fan: this is the group of rays with the top ray labeled PA and the bottom ray is PB. If the lens is completely corrected for spherical aberration, this fan will have a sharp image point P_T' lying directly below the z-axis. Now let P also be the source of a fan bounded by the rays PC and PD. This fan is at right-angles to the other one; we can think of these rays as having the maximum skewness possible. If the lens has no coma, this fan also will produce a sharp image point P_s' which, for a converging lens, will be farther from the lens than the tangential focal point. In other words, even though the lens has been fully corrected for both aberrations, the two corrections will not necessarily produce a common image. In the figure, the meridional fan is called a tangential fan and the skew fan is called a sagittal fan; these are the terms commonly used by lens designers. The failure of the sagittal and tangential rays to produce a single image in a lens corrected for both spherical aberration and coma is known as astigmatism. Astigmatism, coma, and spherical aberration are the point aberrations that an optical system can have. We have already noted that spherical aberration is the only one that can be associated with a point on the z-axis; the others require an off-axis object. Astigmatism is very common in the human eye; to see how it shows up, look at the sagittal image point P_s' of Figure 11. This point is a distance z_s, from the x,y-plane. A plane through this point will have a line image on it, the sagittal line, due to the tangential fan, which has already come to a focus and is now diverging. The converse effect, producing a tangential line, occurs at the other image plane. If we locate an image plane halfway between the two, then both fans contribute to the image and in the ideal case it will be a circle. As the image plane is moved forwards or backwards, these images become ellipses and eventually reduce to a sagittal or a tangential line, as shown in Figure 12. Because there are two different

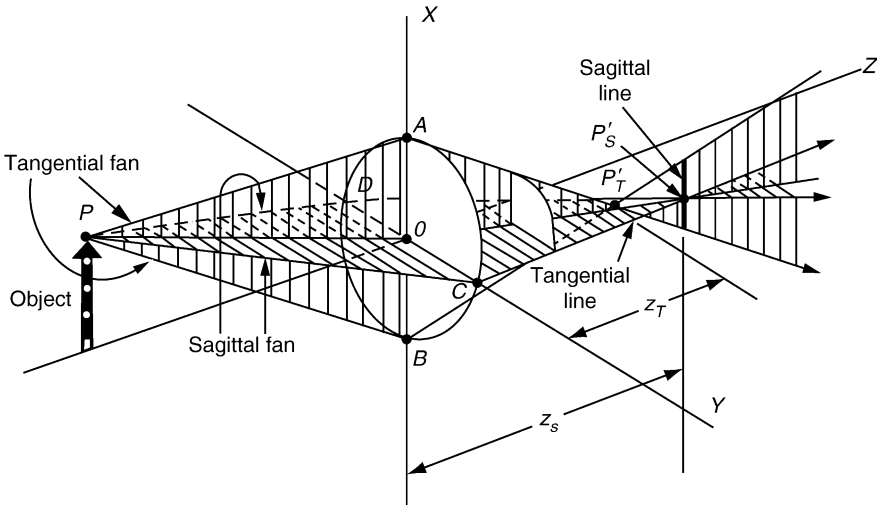


Figure 11 Tangential and sagittal fans displaying astigmatism.

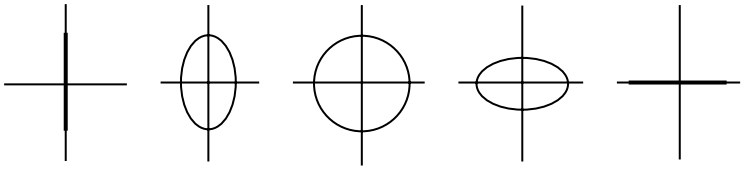


Figure 12 Astigmatic image patterns.

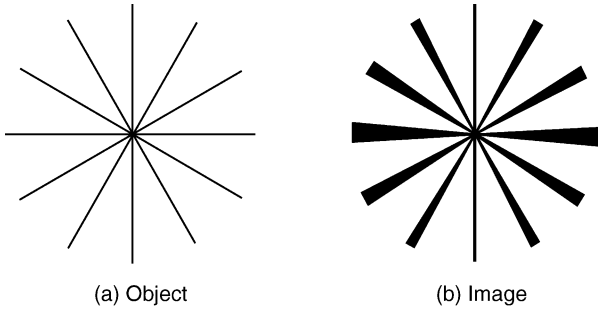


Figure 13 Image degraded by astigmatism.

image planes, an object with spokes (Figure 13) will have an image for which the vertical lines are sharp and the others get gradually poorer as they become more horizontal, or vice versa, depending on which image plane is chosen. Eye examinations detect astigmatism if the spokes appear to go from black to gray. This example of the effect of astigmatism was often based on a radial pattern of arrows, which explains the origin of the word sagittal, derived from the Latin 'sagitta' for arrow.

Curvature of Field and Distortion

Having covered the three point aberrations, there are two lens defects associated with extended objects. If we move the object point P in Figure 11 closer to or farther from the z-axis, we would expect the positions of the tangential and sagittal focal planes to shift, for it is only when the paraxial approximation holds that these image points are independent of x. Hence, we obtain the two curves of Figure 14, which shows what astigmatism does to the image of a two-dimensional object. If the astigmatism could be eliminated, the effect would be to make these curved image planes coincide, but we have no guarantee that the common image will be flat, or paraxial. The resulting defect is called Petzval curvature or curvature of field. For a single lens, the Petzval surface can be flattened by a stop in the proper place, and this is usually done in inexpensive cameras. Petzval curvature is associated with the z-axis. If we take the object in Figure 11 and move it along the y-axis, then all rays leaving it are skew and this introduces distortion, the aberration associated with the coordinates normal to the symmetry axis. Distortion is what causes vertical lines to bulge outward (barrel distortion) or inward (pincushion distortion) at the image plane, as shown in Figure 15. A pinhole camera will have no distortion, since there are no skew rays, and a single thin lens with a small aperture will have very little. Placing a stop near the lens to reduce astigmatism and curvature of field introduces distortion because, as

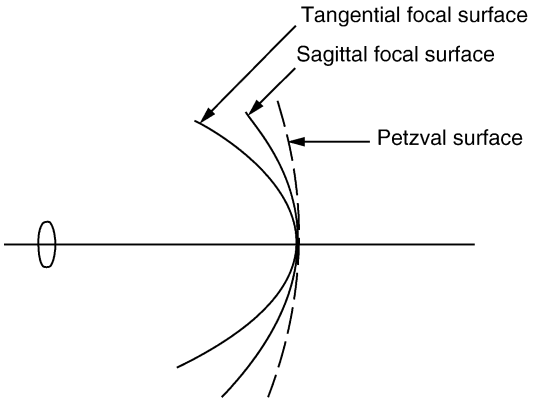


Figure 14 Petzval of field curvature.

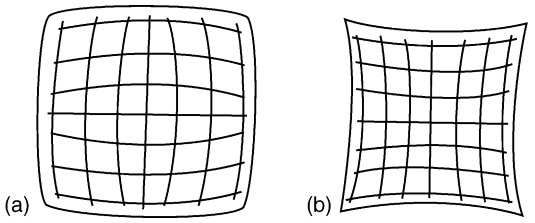


Figure 15 Barrel and pincushion distortion.

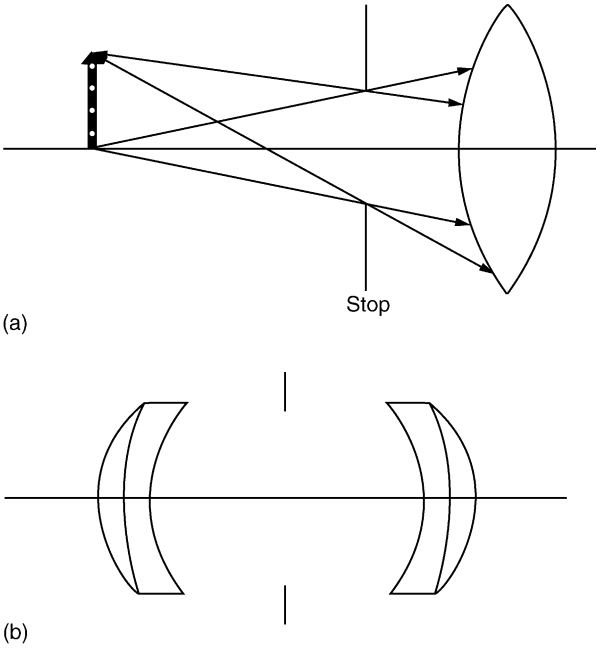


Figure 16 (a) Distortion-creating stop. (b) Symmetric-component lens that reduces distortion.

shown in Figure 16a, the rays for object points far from the axis are limited to off-center sections of the lens. The situation in this figure corresponds to barrel distortion; placing the stop on the other side of the lens produces pincushioning. Distortion can be

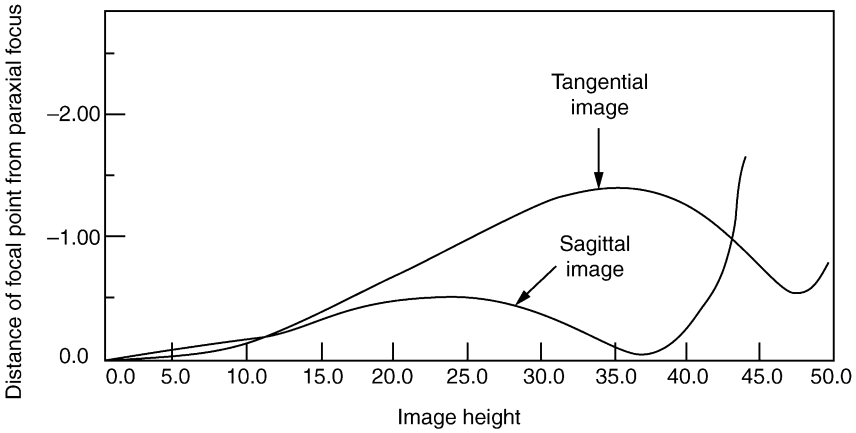


Figure 17 Reduction of astigmatism through design.

therefore reduced by a design which consists of two identical groupings (Figure 16b), with the iris diaphragm placed between them; the distortion of the front half cancels that of the back half. Figure 17 shows the output for a calculation using approximate formulas rather than exact ray tracing. Note that the designer has arranged for astigmatism to vanish at the lens margin. This example indicates the nature of astigmatism in a system for which the coma and the spherical aberration are low but not necessarily zero; it is the failure of the tangential and sagittal image planes to coincide.

Nonspherical Surfaces

So far, we have considered optical systems that use only spherical or plane reflecting and refracting surfaces. However, many astronomical telescopes use parabolic and hyperbolic surfaces. In recent years, nonspherical or aspheric glass lenses have become more common because they can be mass produced by computer-controlled machinery and plastic lenses can be molded with very strict tolerances. Ray tracing procedures given above can easily be extended to the simplest kind of aspheric surfaces: the conic sections. Iterative procedures may be applied to surfaces more complicated than the conic sections. A well-known example is the curved mirror; a spherical reflecting surface will have all the aberrations mentioned here, but a parabolic mirror will bring parallel rays to a perfect focus at the paraxial focal point, thus eliminating the spherical aberration. Headlamp reflectors in automobiles are a well-known example. A plano-convex lens with this defect corrected will be one whose first surface is flat and whose second surface is a hyperbola with its eccentricity equal to the index of refraction of the glass composing the lens. One form of astronomical telescope, which has a large primary concave mirror to collect the light and reflect it back to a smaller secondary convex mirror, is the Ritchey–Chrétien design of 1924. Both mirrors are hyperbolic in shape, completely eliminating spherical aberration. This is the design of the Hubble space telescope. The two-meter diameter primary mirror, formed by starting with a spherical mirror and converted to a hyperbola by computer-controlled polishing, was originally fabricated incorrectly and caused spherical aberration. From a knowledge of the computer calculations, it was possible to design, build, and place in orbit a large correcting optical element which removed all the aberrations from the emerging image.

Conclusion

The three-point aberrations and the two extended object aberrations have been defined and illustrated. Simple ways of reducing them have been mentioned, but the best way of dealing with aberrations is in the process of lens design. In addition, there is a completely separate defect known as chromatic aberration. All the above calculations are based on ray tracing for a single wavelength (or color) of visible light. However, the index of refraction of a transparent material varies with the wavelength of the light passing through it, and a design good for one value of the wavelength will give an out-of-focus image for any other wavelength. It has been known for many years that an optical system composed of two different kinds of glass will provide a complete correction for two colors – usually red and blue – and other colors will be partially corrected. Such lenses are said to be achromatic. For highly demanding applications, such as faithful color reproductions

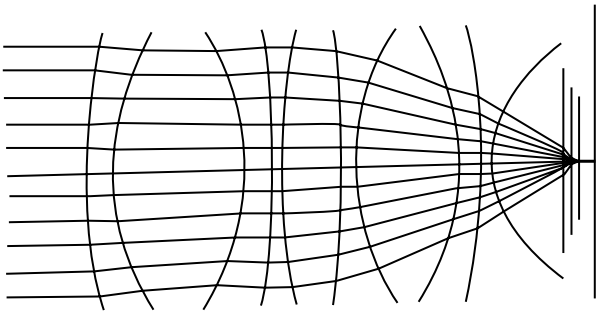


Figure 18 High-power high-resolution microscope objective.

of works of art, it is possible to design elaborate and expensive lenses which correct for yellow light as well. This kind of optical system is said to be apochromatic. With perfect imaging at the two ends and the middle of the visible spectrum, the color error elsewhere is virtually nonexistent. As an example of a design which is free of all aberrations – geometric and chromatic – Figure 18 shows the ray trace for a high-power aberration-free and apochromatic microscope objective. Note that the trace shows parallel rays brought to a focus inside the cover glass; in use, the light source would be a point on the slide and the light would converge at the eyepiece.

See also

Geometrical Optics: Lenses and Mirrors; Prisms.

Further Reading

Brouwer W (1964) *Matrix Methods in Optical Instrument Design*. New York: WA Benjamin.

Laikin M (1990) *Lens Design*. New York: Marcel Dekker.

Nussbaum A (1998) *Optical System Design*. Upper Saddle River, NJ: Prentice-Hall.

Nussbaum A and Phillips RA (1976) *Contemporary Optics for Scientists and Engineers*. Englewoods Cliffs, NJ: Prentice Hall.

Smith WJ (1990) *Modern Optical Engineering*, 2nd edn. New York: McGraw-Hill Book Co.

Prisms

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Prisms are solid structures made from a transparent material (usually glass). Figure 1 shows the simplest type of prism – one that has many uses. The vertical sides are two identical rectangles at right angles to each other and a larger third rectangle is at 45° to the other two. The top and bottom are 45°–45°–90° triangles. The most common applications of prisms are:

1. To bend light in a specified direction
2. To fold an optical system into a smaller space
3. To provide proper image orientation
4. To combine or split optical beams, using partially reflecting surfaces
5. To disperse light in optical instruments such as spectrographs.

Single Prisms as Reflectors

A prism like that of Figure 1 can act as a very effective reflector; it is more efficient than a mirror since there is no metal coating to absorb light. Figure 2 shows a light ray entering the prism at right angles to one of the smaller sides, reflected at the larger side, and emerging at right angles to the other small side. The path shown is determined by Snell's law. This law has the form:

$$n \sin \theta = n' \sin \theta' \qquad [1]$$

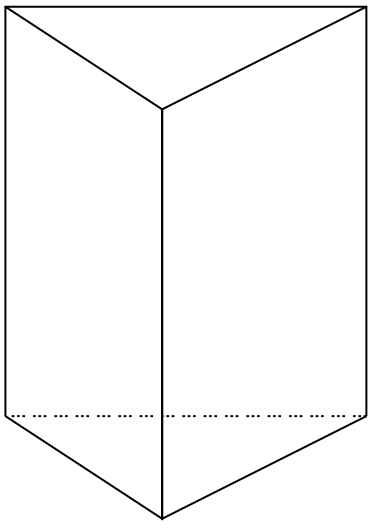


Figure 1 A 45–45–90 degree prism.

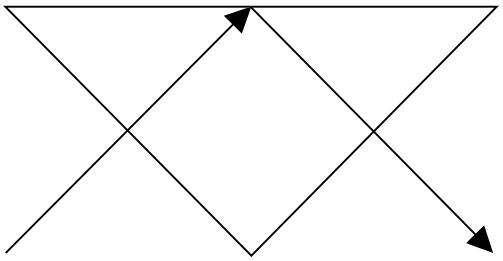


Figure 2 Total internal reflection by a prism.

This indicates that a ray of light making an angle θ to the normal at the interface between media of indices of refraction n and n', will be refracted and the angle to the normal becomes θ'. Notice that the light ray strikes the surface where the reflection takes place at an angle of 45° to the normal; this angle is too large to permit it to emerge from the prism after refraction. To understand why, we recognize that when the ray is able to leave the glass, Snell's law requires that the emerging angle be larger than the incident angle. The largest possible incident angle – known as the critical angle – will correspond to an emerging angle of 90°; that is, this ray will be parallel to the interface. For an index of 1.50, as an example, the critical angle θ_c satisfies Snell's law in the form:

$$1.5 \sin \theta_c = 1.0 \sin 90° \qquad [2]$$

from which $\theta_c = 41.8°$. The ray shown in the figure will therefore remain in the glass. Furthermore, since the indices appearing on both sides of Snell's law are now those of the glass, then the incident and refracted angles must be identical, and the prism reflects the ray in the same way that a mirror would. This phenomenon is known as total internal reflection.

This prism has an important application as an optical component of a fingerprint machine, whose action depends on modifying the total internal reflection which normally occurs. Figure 3 shows the prism and the reflecting behavior of Figure 2, with a person's finger pressed against the long side. When the ray strikes a place on the surface corresponding to a valley in the fingerprint pattern, the index at this region is that of air, and the ray is reflected in the usual manner. But a ray striking a portion of the prism which lies immediately below a raised section of the skin will make an angle with the normal which is now smaller than the critical angle, since the index is much greater than that of air, so that the ray will cross the interface and be absorbed by the finger. The reflected beam, with the fingerprint pattern sharply reproduced, is sent to a detector and the resulting signal goes to a computer, where it is processed and used to drive a printer. The fingerprint cards generated by this kind of equipment are sharper than those done by the traditional smeared-ink method. It is possible to observe the phenomenon just described by looking at your fingers while holding a glass of water – the fingerprints will show up in an enhanced manner. In addition to the better quality, the manufacturer of one version of this equipment has incorporated a feature in their computer program which can detect attempts by the person being fingerprinted to degrade the image.

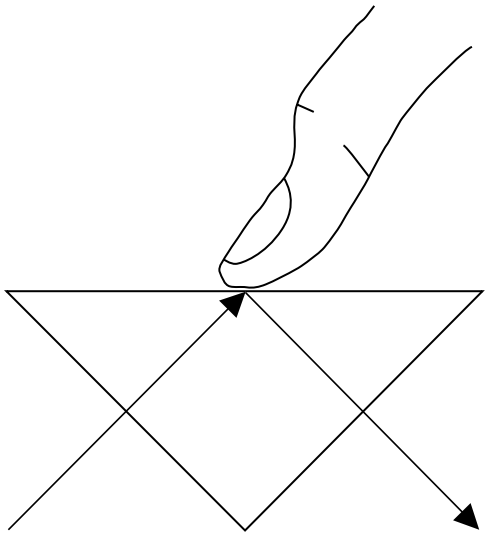


Figure 3 Use of a prism to produce a fingerprint.

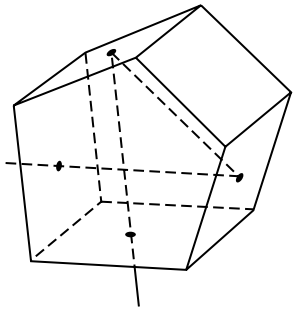


Figure 4 A pentaprism.

A more elaborate single prism – one with five sides – is the pentaprism of Figure 4, used extensively in single lens reflex cameras. This prism does not take advantage of the total internal reflection process, as described above. Below the bottom face is a mirror at a 45° angle. Light from the camera's lens is reflected upward and emerges from the left-hand face where the scene being photographed is observed by the eye.

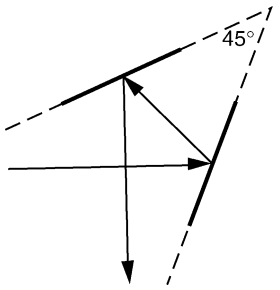


Figure 5 Action of coated faces of a pentaprism.

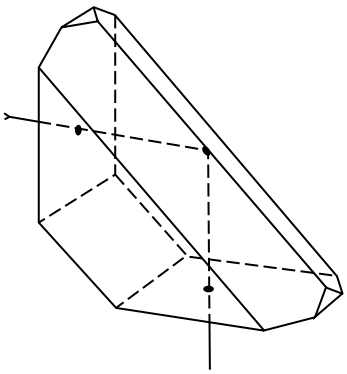


Figure 6 Amici prism.

Since there are two reflections, the reversal caused by a single mirror is eliminated. The camera's mirror is hinged and when the exposure button is pressed, it flies up and the incoming light can reach the film. The two reflecting faces, which lie at an angle of 45° to one another (Figure 5), are coated. To see why this is necessary, note that the two angles in the isosceles triangle formed by the light ray and the face extensions must equal (180° − 45°)/2 or 67.5°. This ray must then lie at an angle of 90° − 67.5° to the normal, or 22.5°, which is much less than the critical angle needed for reflection.

There are at least a dozen other forms of prism which have been designed and used. For example, the Amici prism (Figure 6) is a truncated right-angle prism with a roof section added to the hypotenuse face. It splits the image down the middle and thereby interchanges the right and left portions. These prisms must be very carefully made to avoid producing a double image. One use is in telescopes to correct the reversal of an image.

Double Prism Reflectors

Another extensive application of prisms is their use in pairs as components of telescopes and binoculars. The most widely used is a pair of Porro prisms: two right angle prisms arranged as shown in Figure 7. The double change in direction has two effects; it eliminates image reversal and it shortens the spacing between objective and eyepiece lenses, making the binoculars lighter and more compact. It is possible to calculate how light rays behave in Porro prisms with a ray tracing procedure. This method involves the use of 4×4 matrices and its development is rather complicated. A simple understanding of the way light passes through this optical component can be based on Figure 8, which shows the two prisms separated for clarity. An arbitrary ray (the heavy line) coming from the object will be reflected at a 45° angle at each of the four faces that it strikes. Using two line segments at right angles as an object, the ray from the top of the vertical arrow and the ray from the end of the horizontal arrow will behave as indicated. The image is then identical to the object but rotated by 180°. Since the optics of the binoculars – the combined effect of the objective and the eyepiece – result in an inverted image, the Porro prism corrects this problem.

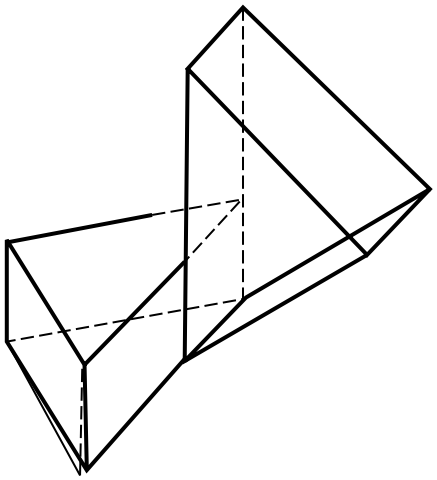


Figure 7 A pair of Porro prisms.

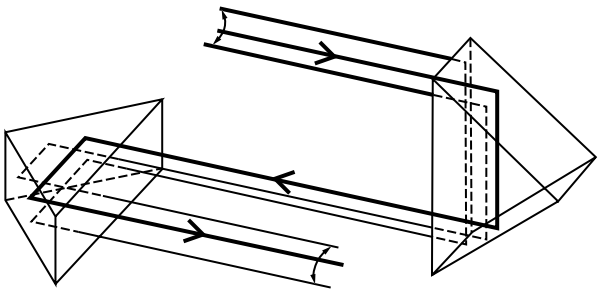


Figure 8 Action of Porro prisms.

Prisms as Instruments

Prisms are the crucial optical element in many laboratory instruments. To show how they can be used to measure the index of refraction of transparent materials – solids or liquids – consider the prism of Figure 9, with apex angle A. Light comes to the prism at an incident angle i, enters the prism at the angle of refraction r, crosses the prism and is refracted a second time when it emerges, the associated angles being designated as r' and i'. When the ray leaves the prism, it is traveling in a direction which represents an angular deviation δ from its original orientation. This deviation is the sum of the individual changes in direction at each face, or

$$\delta = (i - r) + (r' - i') \qquad [3]$$

For the triangle formed by the projected normals, which meet at the angle A:

$$r = r' + A \qquad [4]$$

Hence

$$\delta = i + i' - A \qquad [5]$$

The minimum value of the deviation is an important property of the prism; we find it by differentiation to obtain:

$$\mathrm{d}\delta = \mathrm{d}i + \mathrm{d}i' = 0 \qquad [6]$$

Differentiation of the form of Snell's law valid at each interface, namely:

$$\sin i = n \sin r, \quad \sin i' = n \sin r' \qquad [7]$$

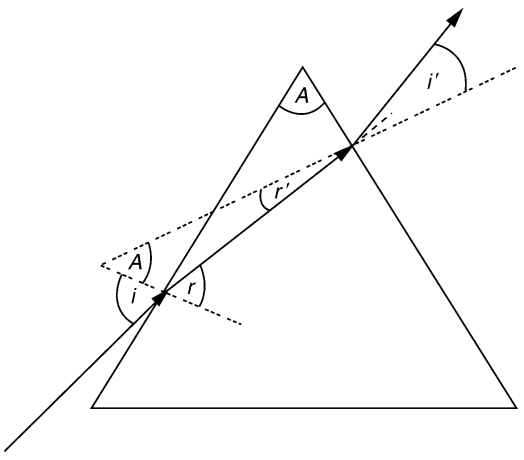


Figure 9 Calculation of prism deviation.

gives

$$\mathrm{d}i = n \cos r \, \mathrm{d}r / \cos i \qquad [8]$$

$$\mathrm{d}i' = n \cos r' \, \mathrm{d}r' / \cos i'$$

$$= n \cos r' \, \mathrm{d}r / \cos i' \qquad [9]$$

where $\mathrm{d}r'$ can be replaced by $\mathrm{d}r$, as obtained from the equation connecting r, r', and A. It now follows that:

$$\mathrm{d}\delta = n\Big[\{\cos r / \sqrt{(1 - n^2 \sin^2 r)}\} - \{\cos(A - r) / \sqrt{(1 - n^2 \sin^2(A - r)}\}\Big]\mathrm{d}r = 0 \qquad [10]$$

from which

$$\cos^2 r\{1 - n^2 \sin^2(A - r)\} = \cos^2(A - r)\{1 - n^2 \sin^2 r\} \qquad [11]$$

or:

$$\cos^2 r = \cos^2(A - r) \qquad [12]$$

and finally:

$$r = A/2, \quad r' = A/2, \quad i' = i \qquad [13]$$

The differentiation thus leads to the conclusion that the ray enters and leaves the prism in a symmetrical manner. The amount of deviation is now:

$$\delta = 2i - A \qquad [14]$$

or:

$$i = (\delta + A)/2 \qquad [15]$$

and:

$$n = \sin i / \sin r = \sin\{(\delta + A)/2\} / \sin(A/2) \qquad [16]$$

Using this expression to calculate n as a function of δ, it is found that δ is a minimum, so that a measurement of the minimum deviation produced by a prism gives the value of its index of refraction in a direct and accurate way. This method can also be used for liquids; a triangular glass cell with walls of reasonable thickness is made very accurately and used to measure δ. The refraction due to the walls does not affect the measurement, since the refraction occurring at the entrance wall is equal but opposite to that at the exit wall.

The description of prism behavior given so far assumes that we are dealing with monochromatic light; that is light of a single wavelength or single

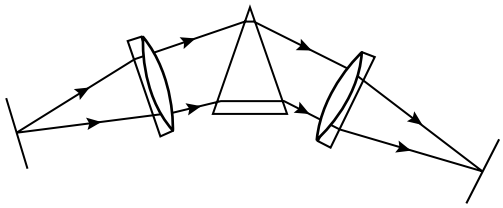


Figure 10 A prism spectrometer.

color in the visible spectrum. However, the index n of a prism varies with wavelength and white light passing through a prism will undergo dispersion – each component of a beam will emerge at a different angle and can be observed. The prisms that are part of expensive glass chandeliers display such an effect. This property of a prism is used in spectrometers – instruments which can measure the wavelength of light. Figure 10 shows a schematic diagram of this kind of instrument of high quality. Note that the light is brought into and taken out of the spectrometer with doublets. As shown in **Geometrical Optics:** Aberrations, doublets can be designed to be achromatic – that is, they focus light throughout the spectrum in approximately the same way. Various kinds of spectrometers have been put into use, with the ability to resolve closely spaced wavelengths being a crucial factor in the design.

See also

Geometrical Optics: Aberrations; Lenses and Mirrors.

Further Reading

Brouwer W (1964) *Matrix Methods in Optical Instrument Design*. New York: WA Benjamin.

Longhurst RS (1967) *Geometrical and Physical Optics*, 2nd edn. New York: John Wiley & Sons.

MIL-HDBK-141 (1962) *Military Standardization Handbook*. Washington, DC: US Government Printing Office.

H

HOLOGRAPHY, APPLICATIONS

Contents

Art Holography
High-Resolution Holographic Imaging and Subsea Holography
Holographic Recording Materials and Their Processing

Art Holography

A Pepper, Shearwater Foundation, Nottingham, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Holography captures the phase and amplitude of light reflected from, or transmitted through, an object. In visual terms it captures 'true' three-dimensional space and any objects which are placed within that space. This is very different from the various optical illusions, including perspective drawing and painting, which 'trick' the eye into thinking it is seeing a dimensional image on a flat surface. Not surprisingly, artists have been intrigued by display holography since reports about it became more public in the early 1960s.

Since much of the information about the holographic process was originally published in professional scientific and optics journals, which the artistic community rarely read, or had access to, the transfer of practical information from optical scientist to visual artist was slow. As less technical reports about holography began to appear in publications, such as *Scientific American*, and art journals, such as *Leonardo*, more artists became aware of the possibilities holography offered. It is now over 50 years since its invention by Dennis Gabor, and art holography can be found in public and private collections around the world, ranging from major museums, through academic art galleries to large individual collections.

For the purpose of this article, holography can be defined as a section of a broad optical field, encompassing display holograms produced by artists, creative individuals, and groups. These works are often produced through independent or self-driven research, but can also be the result of a commissioned project. Holographic art works tend not to be produced specifically for commercial gain (although some do sell for large amounts and can be editioned in the same way prints or photographs, to increase their salability or make them more accessible to a wider audience). It is incredibly difficult to assess the number of artists working in this field today, as no specific research has been carried out, but it can be estimated as between 50–100 individuals worldwide, which includes those who have shown work in curated group and solo exhibitions. This relatively small number has made an enormous impact and includes artists who have made holography their main medium of expression, as well as those, more established in other media, who have extended their creative vocabulary by embracing the holographic process. Salvador Dali for example, had holograms made of his sculptural compositions in the early 1970s, and Bruce Nauman used the process to capture a series of self-portraits in the 1960s. Apart from these well known and established contemporary artists, much of the pioneering work has been driven by individuals

who have concentrated on holography as their medium of choice.

Some artists build their own optical studio in which to experiment in a hands-on environment, some collaborate with scientific and research establishments, where the work, facilities and expertise is shared, and others design and supervise the production of their work which is produced for them by a commercial company. All of them produce a unique contribution to the visual arts.

Early Exploration

Holography offers a unique method of recording and reproducing three dimensions on a flat surface without the need for special glasses (for stereoscopic viewing), lenses (such as lenticular viewing) or linear perspective which, since the Renaissance, has been our main method of tricking the eye into believing that an image recedes into the distance and displays three dimensions.

Many of the very early works produced by artists concentrated on the obvious still life nature of the medium and explored the unusual phenomenon of objects 'frozen' in holographic space. Almost all of these early works were recorded using continuous wave helium neon lasers, either in scientific/research facilities, which artists had arranged to use, or in collaboration with established scientists and optical imaging researchers.

As early as 1969, British artist Margaret Benyon was experimenting with this new medium. Her laser transmission hologram 'Picasso', is a recording of images taken from Pablo Picasso's painting 'Demoiselles D'Avignon'. Here Margaret Benyon took a copy of the painting, cut out the characters, placed them in a three-dimensional spatial arrangement, with real objects, and then made a hologram of this 'scene'. Picasso attempted to show, through many of his paintings, multiple views of the same subject on a two-dimensional surface, one of the main thrusts of Cubism in the late twentieth century. Benyon achieved this in 'real space' by recording a hologram of the flat reproductions. By doing so, she demonstrated that holography automatically achieved what Picasso, and many other key contemporary visual artists, had been trying to do for years.

Holographic Black Holes

It has been known, from the very early days of holographic research that, when using a continuous wave laser to record a hologram, if any object in the scene moves more than half the wavelength of light, no interference pattern will be produced and therefore no recording of the object will take place. Unlike photography, where if an object moves during the exposure, it will become blurred, in holography it will simply not be recorded. Clearly, if no recording can be made, optical researchers working with holography, while noting the phenomenon, would not necessarily feel the need to make holograms which would fail. Several artists, on the other hand, not confined by research deadlines, have acknowledged this strange phenomenon and exploited this limitation of the holographic process. One of the very early examples is 'Hot Air' (Figure 1), a laser transmission hologram by Margaret Benyon and now in the Australian National Gallery Collection. Produced in 1970, it shows a three-dimensional composition of a jug, cup, and the artist's hand. The jug and cup were filled with hot water and during the recording of the holographic plate, with a continuous wave laser, the heat from the hot water disturbed the air above these objects. This tiny amount of movement resulted in the area above the jug and cup not being recorded. Similarly, the artist's hand moved slightly during the exposure and therefore did not record. Strangely, when viewing the resulting laser transmission hologram, the hand is visible. What is displayed is a holographic hand-shaped black hole, the three-dimensional recording of the space once occupied by the hand; the dimensional recording of the absence of an object.



Figure 1 'Hot Air', Margaret Benyon 1970. Laser transmission hologram. Copyright © 1970 M. Benyon. Reproduced with permission.

> Artists in display holography place continual demands on the technology. Their natural tendency is to break rules. I believe that most successful holographic artists belong to this 'suck it and see' category of experimentalists. On hearing about the 'rules' of holography, they break them, just to see what happens. (Margaret Benyon, PhD Thesis, 1994).

Benyon started by breaking the rules and has continued to develop from these early pioneering laser transmission works, which helped establish a visual 'vocabulary' including many important social and political issues.

Other artists have exploited the 'limitations' of the holographic process to produce visually stunning works which question how we look at the three-dimensional world and question what is 'real'. Rick Silberman (USA) developed a shadowgram technique where the hologram is made of a laser illuminated ground glass screen. Objects are placed in front of this luminous surface and their three-dimensional silhouette is recorded. The viewer of the hologram is looking at the space where the object used to be. A pivotal work in this field is 'The Meeting' 1979 (Figure 2), a white light reflection hologram which had been recorded in such a way as to cause the three-dimensional shadow of a wine glass to protrude out of the holographic plate and be viewable several centimeters in front of it, between the observer and the holographic emulsion. The actual wine glass used in the recording has been broken and what remains (the base, stem, and small part of the bowl of the glass) is placed on a platform attached to the holographic plate. When viewing the work head on, the missing parts of the broken glass are recreated, in three dimensions, by the holographic shadow. When viewed obliquely, from the side for example, only the broken (real), glass is visible. Not only does this piece have an element of performance in the recording, breaking, mounting, and reconstruction of the wine glass, it also explores questions of visual perception and concepts of reality.

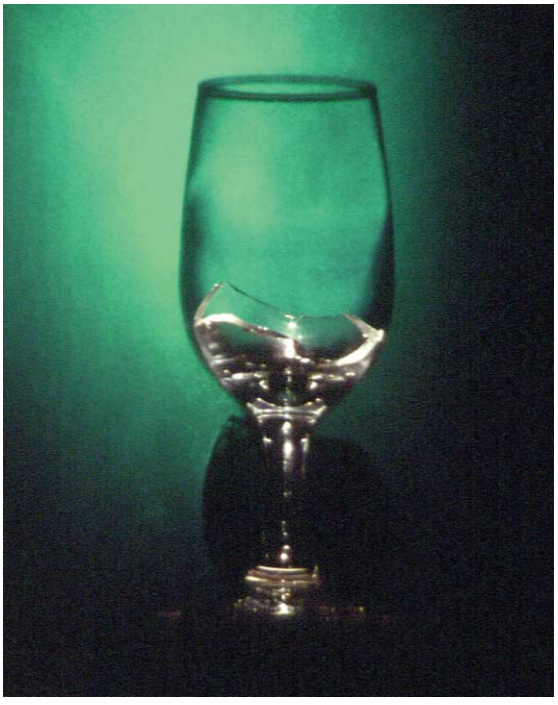


Figure 2 'The Meeting', Rick Silberman 1979. White light reflection hologram (shadowgram). Copyright © 1980 ATP. Reproduced with permission.

The recording of the shadow of an object, or objects, has now become commonplace in creative holography, with many artists employing techniques to mix different objects together, change their color and rearrange them in space. One advantage of this technique is that dependence on absolute stability of the objects used is reduced. The objects are 'encouraged' to move during the recording process and so create 'negative' images. Steve Weinstock (USA), Marie-Andrée Cossette (Canada), Michael Wenyon and Susan Gamble (USA/UK) have all contributed to this area in some of their early works.

Public Display

As the technology and optical processes of holography have developed, so greater flexibility, and public accessibility, have emerged. The single most important aspect has been the ability to display holograms in white light.

Emmett Leith and Uris Upatnieks invented laser transmission holography in 1964. They demonstrated the practicalities of Gabor's original invention, by utilizing laser light, to produce three-dimensional images. To display these recorded images, laser light needed to be shone through the holographic plate, at the original angle used by the recording reference beam (off-axis). The results were breathtaking, with full visual parallax in all directions. Although it was the laser light which made the recording and reconstruction of these unusual images possible, the laser was also a limiting factor. Each time a hologram was displayed, a laser was required. This relegated the technique to displays in scientific conferences, laboratories, or other locations with optical equipment. What was needed was a technique allowing holograms to be displayed without the need for this expensive and dangerous laser light. White light holography was the answer.

Although a laser is still required to record the original hologram, by using a copying technique, a

restricted aperture or one step process, the resulting holograms can be reconstructed using a small source of white light such as a spotlight or even direct sunlight. As expensive lasers were no longer needed to display holograms, large public exhibitions could be mounted. Gallery and museum spaces could include holography in their exhibitions and individuals could buy and display holography in domestic locations. This technique also benefited the commercial display market, where holograms could now be used for advertising and promotional events.

Stephen Benton (USA) (1942–2003), inventor of white light transmission (rainbow) holography in 1968, made a major contribution to the use of the medium by artists. His technique not only allowed holograms to be viewed in white light, but they were extremely bright and offered the possibility of creating multicolor images. His 'Crystal Beginnings' 1977, (Figure 3), shows just how much space can be created with very simple points of light. Benton even taught master classes for artists so that he could explain how his technique functioned and explain the mathematics involved. He also worked with several artists to help them realize their individual projects.

Harriet Casdin-Silver (USA) was one of the first pioneering artists to work with Benton and produced several key works now seen as integral to the art of holography. Their early collaboration 'Cobweb Space', 1972, sculpted pure light and was one of the very first art holograms to be displayed using Benton's rainbow hologram technique. A later work, 'Equivocal Forks', 1977, exists in different formats; as a laser transmission hologram showing the pseudoscopic display of domestic folks which protrude, inside out, from the holographic plate, and as a number of white light transmission holograms, of the same image, which were shown in an outdoor installation mounted by The Center for Advanced Visual Studies, part of the Massachusetts Institute of Technology. This demonstrated to a wide, nonscientific, audience, that holograms could be accessible, easily illuminated and weatherproof. Casdin-Silver has continued to produce art holography and has, most recently, worked with large-scale portraiture.

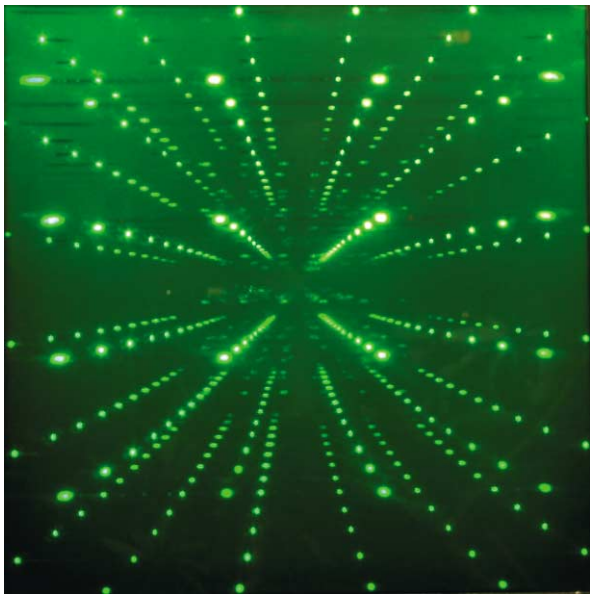


Figure 3 'Crystal Beginnings' Stephen Benton 1977, White light transmission (rainbow) hologram. Jonathan Ross Collection, London, UK. One of the best known works by the inventor of the rainbow hologram. Copyright © 2004 J. Ross. Reproduced with permission.

Although Benton's process increased the accessibility of holography, an earlier technique, invented by Yuri Denisyuk (Russia) in 1962, allowed holograms to be recorded in a single step and then displayed using white light. The political situation between the USSR and the rest of the world during the early 1960s meant that information and results from much of this early work was not openly accessible, and communication between Denisyuk and other key researchers in holography was limited. However, as soon as details of the Denisyuk technique became more widely available in the 1970s, artists began to embrace it as a practical alternative.

The work of Nicholas Phillips (UK), whose research improved chemical processing of display holograms during the late 1970s, helped artists and the rest of the holographic field to achieve much better results when making reflection holograms, a technique favored in Europe. He was also heavily involved with the rock band, 'The Who', and their Holoco company which mounted two very successful exhibitions of display holography at the Royal Academy of Art, London, at this time. Artists who might not have come across the holographic phenomenon became aware of its potential due to its presence in one of the UK's more traditional art venues. Other large exhibitions in the late 1970s and early 1980s in USA, Sweden, and Japan also highlighted the immense possibilities holography offered to a nonscientific community of gallery visitors.

As interest in all aspects of holography increased, dedicated spaces opened to sell, display, and archive the output from artists, scientists, and commercial producers. One of the most influential of these was the Museum of Holography, located in the SoHo art district of Lower Manhattan, New York (Figure 4). From 1976 through to 1992, it regularly mounted group and solo exhibitions by artists, giving many of



Figure 4 Dr Dennis Gabor, inventor of holography with, Posy Jackson, Director of the Museum of Holography, during his visit to the Museum in March, 1977. Copyright © 1977 Paul D. Barefoot. Reproduced with permission.

them their first major exposure to a museum-going public. There were displays relating to the scientific, commercial, and industrial usage of the medium, as well as exceptionally popular traveling exhibitions. 'Through the Looking Glass', which traveled worldwide, allowed early display holography to be experienced by thousands of people who would not normally have come into contact with the medium. In the Summer of 1981, over 300,000 visitors saw 'Through the Looking Glass' during the 11 weeks it was on view at the Israel Museum, Jerusalem.

After the Museum of Holography closed in 1992, its entire collection was purchased by the MIT Museum, Boston, USA, and now forms part of the largest museum collection of creative and scientific holography in the world. The New York museum was not an isolated venue. During the late 1970s and early 1980s, museums of holography opened in Chicago, Paris, and Pulheim (near Cologne). Galleries were also founded in many major cities to sell art and commercial holograms.

Stability Solutions

Although being able to display holograms using inexpensive sources of white light allowed the medium to become much more accepted as a display process, there has always been the cost and complexity of the optical laboratories needed to record the holograms in the first place. Few artists were able to equip their studios with 'off the shelf' optical components and precision-built isolation tables most often found in scientific and research labs. The solution was to replace the commercially built isolation table with much less expensive sand tables.

Sand box holography revolutionized the field for independent artists. By building a wood or brick box filled with sand, resting on car inner tubes, artists were able to hand-make their own isolation tables. Inexpensive optics, mounted onto plastic pipes were simply thrust into the sand, allowing complex optical recording setups to be constructed. Surprisingly, many key artworks have been produced using this technique, which opened the field to anyone with basic construction skills and a laser.

The laser has remained an expensive element in the holography studio but, most recently, the prospect of replacing this with laser diodes has offered a much more economic alternative. Not only can light from these diodes provide more compact optical setups to record holograms, they can also be used to display them. Significant progress has been made in both diode recording and display, which is opening up new opportunities for the visual artist. Almost all of the early practitioners in the field were attracted to holography because of the laser transmission holograms they saw. They have a depth and dimensional quality which is extremely different from the white light viewable holograms we tend to see today. Now, with the possibility of displaying holograms illuminated by inexpensive laser diodes, the laser transmission hologram is set, once again, to be a viable medium for artists.

Object Holography

Many artists were attracted to holography because of its unusual three-dimensional properties. However, they did not have the resources to make holograms of large objects or scenes, so scaled these down and constructed small sculptural settings for recording. One of the pioneering artists in this area was Dan Schweitzer (1946–2001). Originally an actor and familiar with theatrical staging, many of his holograms are populated by tiny figures placed in specially built 'theatrical' settings. Holography not only recorded these specially constructed miniature 'rooms', but the resulting images were realistic enough to suggest to the viewer that they were looking at a scaled-down life-size event. We do this regularly when looking at photographs and accept that they are a small-scale version of a larger scene. Traditional holography does not offer this 'scaling down' in the same way as photography. One celebrated example is Schweitzer's 'Thendara', 1978, where a sculpted, miniature figure sits and looks towards a window frame. He also produced a companion piece in which the physical model, used to make the earlier hologram, is on display as a piece of sculpture. Here the miniature window is the frame for a hologram displaying a previously recorded model of a landscape. As Schweitzer's technical skills

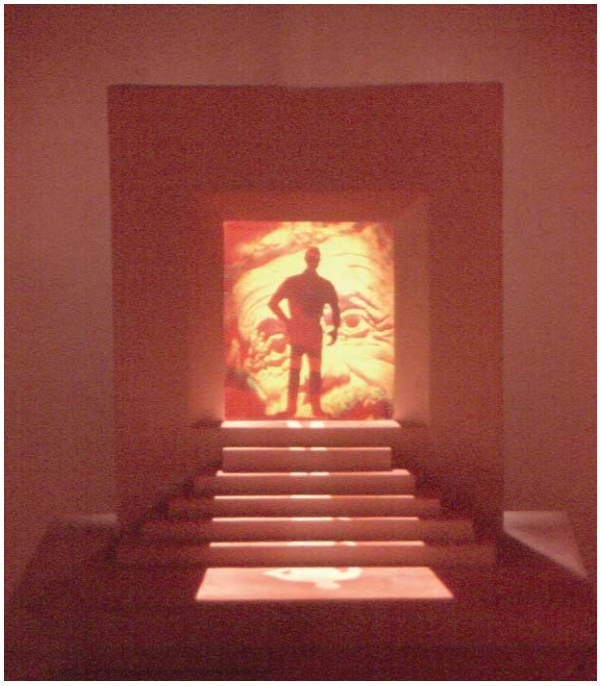


Figure 5 'Od Vent', Dan Schweitzer 1980, White light transmission hologram in sculptural mount. Copyright © 1980 ATP. Reproduced with permission.

developed, he began to produce complex multilayered works incorporating still-life models, hologram fragments, and optical animation to construct visual spaces which could not exist in the real world.

In his work 'Od Vent' (1980) (Figure 5), Schweitzer framed an exceptionally complex white light transmission hologram into a theatrical model of a doorway. The miniature stairs lead up to an opening, a portal into another world, in which the hologram is mounted. Illuminated from behind, the hologram displays multiple exposures, optical animation, and several superimpositions of objects as an observer moves from left to right in front of the piece. In this complex holographic world, the silhouette of the figure, standing in the door, casts a bright (negative) shadow down the miniature stairs. Not only are we able to look into an illusionary world through the holographic doorway, where at a certain point Einstein looks back at us, but the peculiarities of this holographic world extend into our own world as seen by the 'real' (negative) shadow cast down onto the stairs.

Another artist, who combined the use of multiple exposed objects displayed in a single space, is John Kaufman (USA). He established a reputation by recording objects (often rocks) collected from around his Californian holography studio. These found objects were captured in three dimensions and were then, through a multiple exposure and emulsion swelling technique, combined into a single still-life composition. He became expert at masking out areas of his objects so that they could appear in different, intense colors, and in many cases used the holographic process and multiple exposures to mix colors directly in the holographic space he was working with. In 'Stone Head' (1982) (Figure 6), for example, this relatively 'normal' stone is transformed by the use of intense color and the three-dimensional volume it occupies.

When considering how the hologram should be framed and displayed, several artists have utilized the mount in which the hologram sits as an integral part of a much more complex work. Frithioff Johansen (Denmark) recorded a hologram of smoke, freezing it in time and space using a pulsed laser (see below). This was then converted into a white light transmission hologram and mounted in a frame of granite. 'Hardware-Software 1' 1987–1988, allows the gallery visitor to not only view the massiveness of the granite structure, but also the delicate sculptural solidity of smoke 'frozen' in space.

Rebecca Deem (USA) recorded what appears to be an extremely simple hologram of a luminous tube, mounted onto a swivel at the end of an elegant metal support. In 'W/Hole Message' viewers can look into the small holographic plate and see the image of the tube extending away from them into the volume behind the plate. By flipping the hologram, on its swivel, to turn it over, the image in the hologram is also flipped and the tube appears in the viewer's space, between them and the plate. This three-dimensional pseudoscopic image, one of the attractive and unusual aspects of holographic display, is inside out, back to front, and upside down. The gallery visitor really does have a chance to view the whole message.

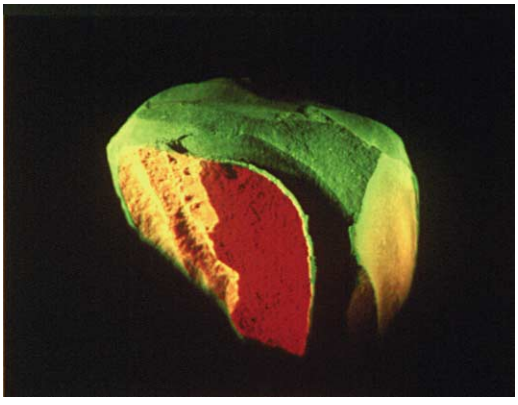


Figure 6 'Stone Head', John Kaufman 1982, White light reflection hologram. Copyright © 1982 John Kaufman. Reproduced with permission.



Figure 7 'Breakfast', Pearl John 2002. Laser etched toast displayed in a white light reflection hologram. Copyright © 2002 Pearl John. Reproduced with permission.

Text has become an important element in holography, as it is often incorporated in, onto, or around holographic displays. Eduardo Kac (Brazil) has produced several holograms, which display poems, words, and letters floating within the holographic space he has created.

Pearl John (UK) most recently produced a series of still-life holograms using a pulsed laser but with the added subtlety of writing onto the surface of her objects. In 'Breakfast' 2002 (Figure 7), for example, the slices of toast, making up the composition, have been laser etched with text which appears 'burnt' into the surface of the bread.

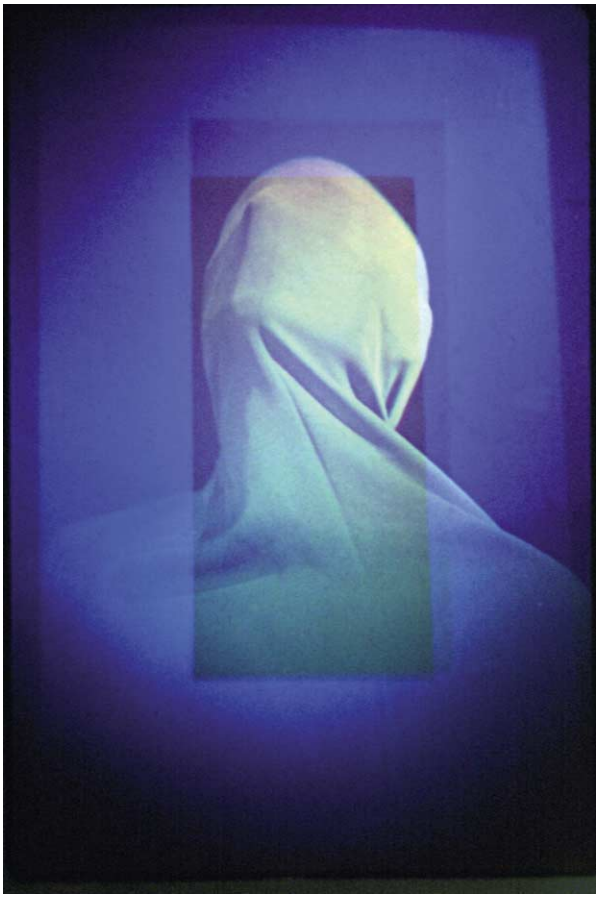


Figure 8 'Awakening', Ana Maria Nicholson, 1993. White light (pulsed) reflection hologram. Copyright © 1993 Ana Maria Nicholson. Reproduced with permission.

Pulsed Lasers

Continuous wave lasers placed limits on the holographic recording process and dictated that no movement in the recorded object could take place. We have seen that this 'problem' was exploited by artists at an early stage, but there was also a desire to capture living objects, in the form of the human figure. Pulsed laser holography was the solution. Because this type of laser generates a bright flash of coherent laser light, the intensity and duration of the pulse is enough to 'freeze' any physical movement during the recording process, allowing the recording of people, animals, or anything which moves and could be brought into an holography studio. Again, the cost and accessibility of this specialist equipment was prohibitive to artists, but they found ways of persuading scientists with pulsed lasers to help them. Early adopters of this technique were artists such as Ana Maria Nicholson (USA), Melissa Crenshaw and Sydney Dinsmore (Canada), Margaret Benyon (UK), Anait (USA), and Alexander (USA/UK). Recordings could now be made of live figurative compositions and individual portraits. Many of the artists, not content with producing the holographic version of a traditional photographic portrait, began to manipulate these recordings, distorting, multiple exposing and inverting these unusual three-dimensional images of people.

Holographic portraits do have an unusual quality, which painters or photographers have found difficult to deal with but art holographers have embraced. Some, like Ana Maria Nicholson (USA), have combined subtle processing techniques, to introduce color 'layers' into the portraits, combined with nonstandard views or compositions (see 'Awakening' 1993 (Figure 8)). Other artists, who emerged in the 1980s, produced complex, highly staged compositions, such as those from Patrick Boyd (UK), or they offered us a different view of the 'self portrait' like those by Martin Richardson (UK). Others, such as Shu-Min Lin (Taiwan), used the combination of multiple views of human figures recorded onto the same holographic plate. By walking past these displays, an animated effect could be experienced.

So much interest was generated in pulsed holography that several studios emerged for use by artists, including independent facilities and those within educational establishments. No longer was this the domain of the well funded scientific research facility. Several of these 'studios' still exist and are offering artists access to equipment through specialized artist-in-residence programs.

Abstraction

Although holography is an exceptional process for recording every detail of an object and the volume it occupies, many artists have wanted to work with the abstract and kinetic possibilities the process offers. Light can be recorded and displayed in three dimensions, it can be animated, inverted, compressed, sculpted, and projected. Douglas Tyler (USA), in his In the 'Dream Point Series' 1983 (Figure 9), recorded small laser transmission holograms but displayed them using normal white light, producing a smearing effect of the three-dimensional object. He then mounted these onto transparent surfaces (Plexiglas) and drew on the surface of the plastic to incorporate etched lines and shapes. Rudie Berkhout (Netherlands/USA) began his holographic exploration with stunning abstract landscapes made of luminous, multicolor, animated objects which took advantage of his mastery of the white light transmission technique. Most recently he has been recording reflection holograms which display undulating colored 'waves' of light as an observer moves past the hologram. These have been included in specially produced paintings for large-scale architectural commissions.

Sam Moree (USA) combines multiple images, exposures, diffraction gratings, and smaller holograms into a single work. The finished holograms display a complex series of image fragments and colors within the depth of the holographic space he has constructed. Ruben Nuñez (Venezuela), originally approached holography from a background in glass production and used the ability of the holographic process to 'capture' light reflecting and passing through his glass objects. The resulting white light transmission holograms display an abstract 'photonic' display of form and color. Fred Unterseher (USA), uses holography to record optical distortions and diffraction gratings which would display abstract patterns of light. Many of these early works were mandala images which displayed an intense kinetic effect as an observer walked past the surface of the holographic plate.



Figure 9 'Dream Passages': 32″ × 138″ (3 panels 32″ × 46″). Douglas E. Tyler 1983, Laser transmission holograms, tape and Plexiglas, illuminated with white light. Detail of center panel. Copyright © 1983 D.E. Tyler. Reproduced with permission.

Pseudo Holograms

In 1973, Lloyd Cross invented the multiplex (integral) hologram technique which used sequential two-dimensional images recorded on movie film combined with white light (rainbow) hologram techniques. The result was a holographic stereogram, which gave viewers the illusion of seeing animated images in three dimensions. The final holographic sheet was displayed either curved or as a complete cylinder, which was often motorized so that viewers could stand in front of the unit and watch the animation take place as the hologram rotated. Such a display does give the impression that an observer can walk all the way round the image, enhancing its perceived three-dimensional effect.

As the original images are recorded photographically, there are several advantages – anything which can be recorded on inexpensive black and white movie film can be converted into a holographic display. Short animated sequences can be presented while accepted film techniques such as zoom, pan, tilt, and transitions can be incorporated into the finished sequence of images. Although very popular for advertizing and promotional campaigns during the 1980s, particularly in the USA, several artists used this technique. They could make the movie film sequences themselves and then send this film to a commercial producer for a hologram to be made for them. This offered an interesting balance in the creative process. The artist could concentrate on the filmic image, using facilities and photographic techniques they were familiar with and not have to learn the holographic process or build a holographic laboratory themselves. Artists such as Anait (USA) produced several experiments in the area, one of which showed her inside the holographic display,



Figure 10 'Woman' Harriet Casdin-Silver 1979. Multiplex hologram (white light transmission). Copyright © 1980 ATP. Reproduced with permission.

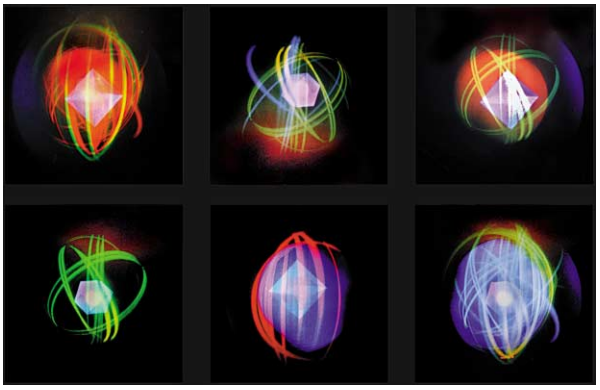


Figure 11 'Neuro Hologram', Ikuo Nakamura, 1993. White light transmission hologram with multiple, computer controlled, lighting. Copyright © 1993 Ikuo Nakamura. Reproduced with permission.

writing, with spray cream, seemingly onto the holographic surface. Dieter Jung (Germany), used this technique to display a specially written poem which appeared to spiral around inside of the display.

Harriet Casdin-Silver used the technique to produce an unusual self-portrait. 'Woman', 1979 (Figure 10), is a multiplex hologram of the artist's head. Black and white film images were recorded of Casdin-Silver's head, through 360 degrees, and then each frame of film was converted into a thin white light transmission hologram. The artist chose not to display the resulting image in a normal cylindrical display, but to 'flatten' it out so that the front, sides, and back of her head could be seen, distorted, in a single head-on view.

As video and digital imaging processes have developed over the years they too have been combined with this pseudo holographic technique, opening up a wide avenue of possibilities. Video footage, computer generated modeling, or any process which can produce sequential 2-D images, can now be converted into an animated 3-D hologram. As digital video cameras and computer 3-D modeling software drop in price, and increase in sophistication, a whole new area of display has become available to the digital and video artist.

Multi Media

As in other areas of the visual arts, more than one medium or process are often combined. Doris Vila (USA) has produced complex environments combining large-scale holographic images which, rather than being illuminated with a single source of light, are reconstructed with multiple spotlights or via video projectors. The lighting is activated by sensors in the exhibition space, which are computer controlled and made to trigger certain sequences dependent on the input they receive from the movement of visitors.

Traditional illumination of a hologram involves a single point source of light shining onto, or through, the holographic plate. This will reconstruct a single image. If multiple lights are shone onto the hologram, multiple reconstructions of the single image will result. By using many independently controlled lights on one hologram, Vila has been able to create animation and kinetic effects as the resulting multiple images overlap. Using light from a video projector also allows the artist to combine a reconstructed holographic image with two-dimensional animated footage from the video projector.

Other artists to combine holograms with structures, interactive lighting and environments, sound, and installational displays include: Georges Dyens (Canada), Marie-Christiane Mathieu (Canada), Matthew Schreiber (USA), Philippe Boissonnet (Canada) Ana MacArthur (USA) and Shunsuke Mitamura (Japan).

Ikuo Nakamura (Japan/USA) has used multiple lighting of a single computer generated hologram to produce animated effects controlled by the viewer. In his 'Neuro Hologram' 1993 (Figure 11), brain waves from the viewer, wearing a special set of sensors on their head, are measured and converted into control signals for a number of lights behind the hologram. By relaxing, or becoming agitated, the viewer can change their brain waves and actively control the appearance of the hologram, causing multiple views of the holographic image to fade in and out, giving an impression of abstract animation.

Architectural Scale

One of the great limitations of holography has been the size of the holographic plates on which the light

sensitive emulsion is coated. These tend to be in the region of 20×23 cm, 30×40 cm, or meter-square formats. Generally, the larger the finished hologram, the larger the laboratory needed to produce it. This impacts a great deal on artists who do not have the facilities, or finances, to produce 'large' holograms. A solution has been to combine smaller holographic plates together, as a mosaic, to give the size and scale needed. In recent years, several artists have completed public commissions for large-scale interior and exterior holographic installations.

Setsuko Ishii, (Japan) has produced several installations in public spaces where individual holographic plates are mounted together to form a larger composite structure. Many of these pieces use holograms recorded on dichromate gelatin plates which produce an exceptionally bright image, ideal for public spaces, where visibility, from multiple angles, is important.

Michael Bleyenberg (Germany) produced an entire exterior holographic wall (13×5 meters) during 2002 (Figure 12), for the extension of a research building in Bonn. The installed graphic wall, 'EyeFire' (AugenFeuer), is made up of 26 smaller panels, each constructed from tiny holographic 'pixels' (holographic optical elements, HOE) to give an intense color effect. Working in collaboration with the Institute for Light and Building Technology, Cologne, Bleyenberg was able to design his graphic panels on a computer screen and have this information translated into the thousands of tiny holographic pixels which, when seen together, make up the complex and high intensity colour design.

Dieter Jung (Germany) has produced commissioned art works for large interior architectural spaces. In 'Perpetuum Mobile' 2001, located at the European Patent Office, The Hague, holograms are placed on the floor, which can be walked over, and an extensive holographic mobile structure hangs overhead. The light, and color, from the abstract holograms in the mobile sections, can be reflected in the abstract holographic floor and combine to produce an interior lightscape impossible to produce with more traditional media.

Holograms can be mass produced using a mechanical embossing process which transfers the optical contents of the original master hologram onto plastic sheeting for duplication in large runs. One novel use of this material has been the interior of the Canadian pavilion at Expo '92 (Figure 13). Melissa Crenshaw (Canada/USA) used huge quantities of embossed holographic material to cover the 3000 square foot interior walls of the specially built pavilion. Lighting was installed to illuminate this optical diffracting wall and the entire structure was covered in glass, over which water flowed. The resulting interior appeared to be a wall of liquid light which changed color and hue depending on where it was viewed from.

Sally Weber (USA) used holographic embossed material for several outdoor commissions. Structures were clad in embossed holographic sheets and made water resistant. Sunlight illuminated the installations to produce surfaces of constantly changing color. In direct (bright) sunlight, intense rainbow colors would be diffracted by the holographic surfaces. On an overcast day, pastel, more muted,



Figure 12 'EyeFire/AugenFeuer', Michael Bleyenberg 2000. White light transmission holograms mounted onto building exterior. Copyright © 2000 M. Bleyenberg. Reproduced with permission.



Figure 13 Interior of the Canadian Pavilion Expo '92, Spain. Holographic wall water feature by Melissa Crenshaw. White light transmission (embossed) hologram foil mounted into pavilion walls. Copyright © 1992 M. Crenshaw. Reproduced with permission.

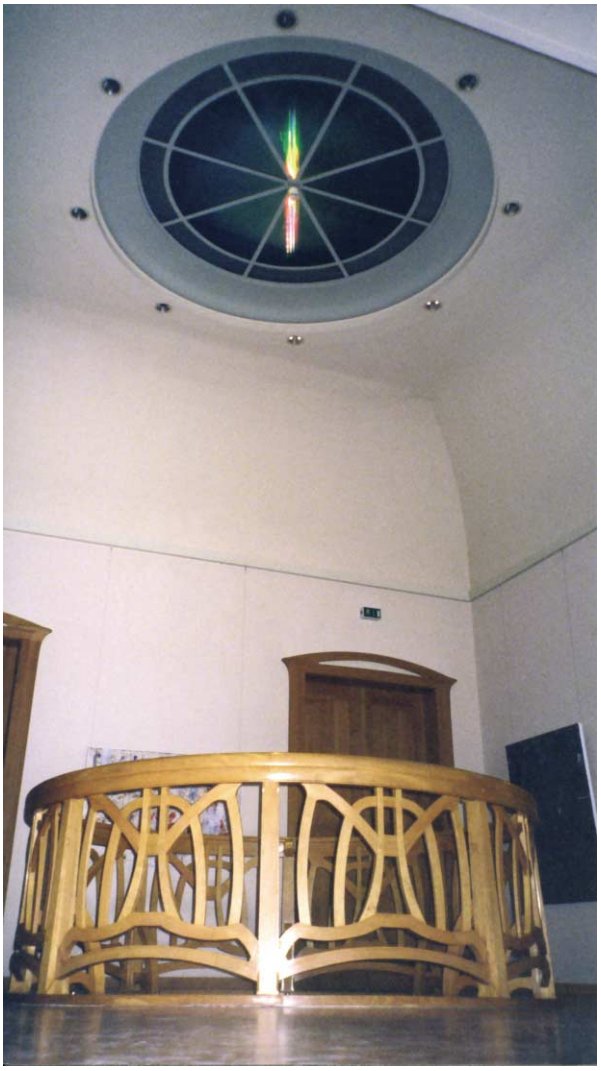


Figure 14 Signature of Source, Sally Weber 1997. Site specific window installation at the Karl Ernst Osthaus-Museum, Hagen, Germany. Copyright © 1997 Sally Weber. Reproduced with permission.

colors would appear. Although she is often known for these large-scale outdoor embossed works, one of her permanent commissions is housed in the Karl Ernst Osthaus-Museum, Hagen, Germany, where specially made holographic diffraction gratings are incorporated into an existing building. 'Signature of Source' 1997, (Figure 14), is a permanent, site-specific installation producing a line of light which enters the museum space from the skylight it is built into. This abstract line of pure color becomes a sculptural element within the space of the museum and is generated by the action of light from outside the building being manipulated and 'formed' as it passes through the holographic window.

One of the most celebrated and very early attempts to incorporate large-scale holograms into art works



Figure 15 'There's no Place Like Home', Paula Dawson 1979–1980. Laser transmission hologram of an entire room. Copyright © 1980 Paula Dawson. Reproduced with permission.

was the installation 'There's No Place Like Home' by Paula Dawson (Australia) in 1979–1980 (Figure 15). A laser transmission hologram of an entire room was recorded. The room was specially constructed, life size, and recorded onto a single holographic plate. The resulting hologram was then built into the window of a small 'house' constructed in the gallery. Visitors were able to approach the house, look into the window and 'see' the interior with its furniture, fixtures, and fittings. They would then enter the house and walk past the space where the room appeared to exist. When viewed from this vantage point there was nothing there except an empty space. Artists have often questioned and researched the concepts of 'inside and outside', voyeurism, reality, normality, and 'existence', but few have had the tenacity to produce an installation with such visual and conceptual impact.

Twenty-First Century Art

Art holography has developed and matured over the past 50 years to become a viable and significant element in the contemporary visual arts. Critics are still asking 'is it art?' in the same way they used to worry about whether photography was art. While they consider this problem, practitioners in the medium are continuing to find ways to push past its technical and visual boundaries (Figure 16). They continue to collaborate with the scientific and research world and use holography as a medium to comment on life and society in the early twenty-first century.

> …I hope, that one day I will be able to give you, artists, something really powerful; panoramic holograms,

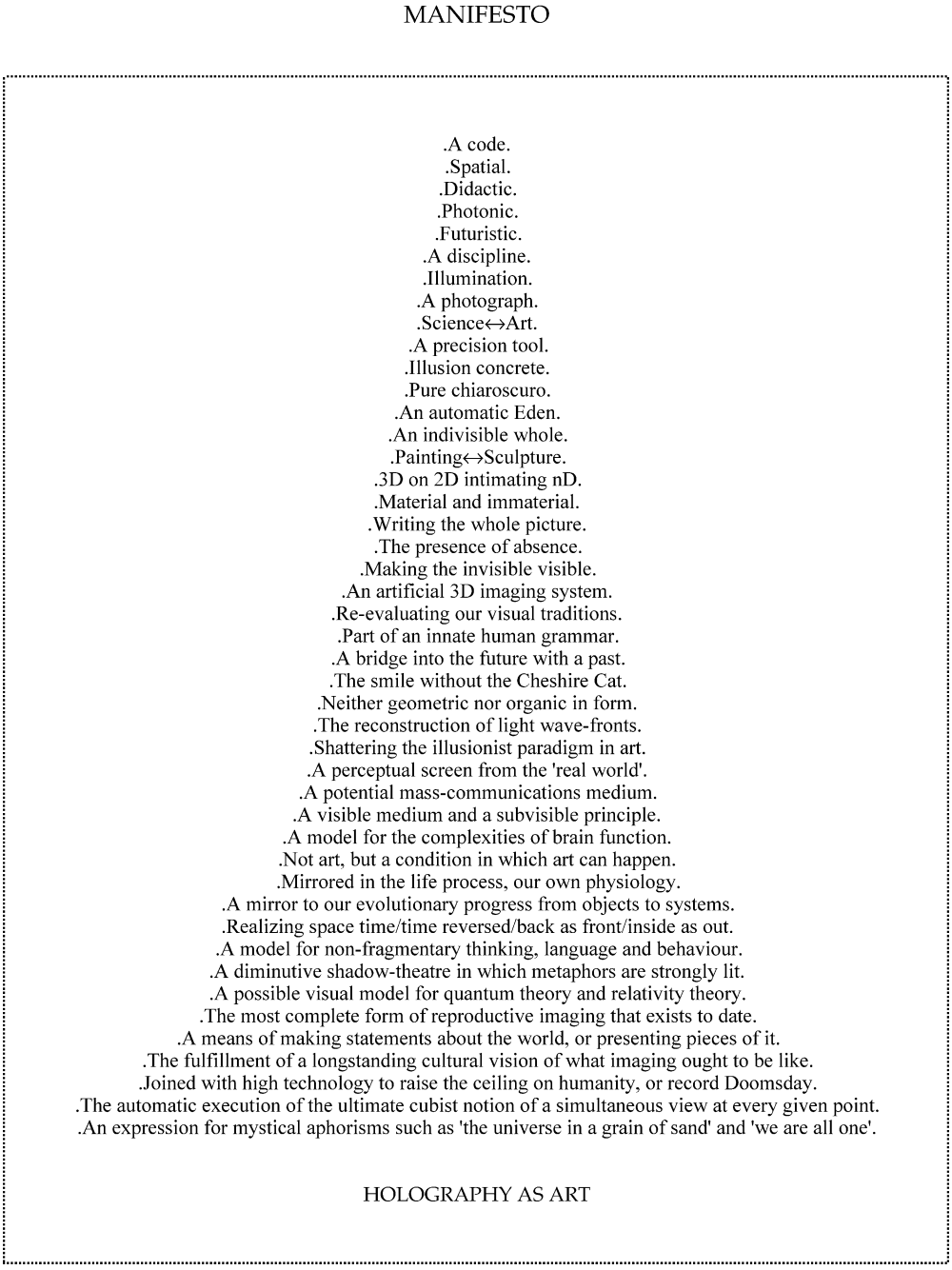


Figure 16 'Holography as Art', Margaret Benyon 1980. Manifesto. Copyright © 1980 M. Benyon. Reproduced with permission.

extending to infinity, in natural colors. Dennis Gabor, Inventor of holography in a letter to artist Margaret Benyon 28th February, 1973.

See also

Coherence: Coherence and Imaging. **Diffraction:** Diffraction Gratings. **Holography, Techniques:** Overview. **Imaging:** Multiplex Imaging. **Interferometry:** White Light Interferometry. **Introductory Article:** Early Development of the He–Ne Laser (Title TBC).

Further Reading

Benyon M (1994) *How is Holography Art?* PhD thesis, Royal College of Art, London.

Brill LM (ed.) (1989) Holography as an art medium, Leonardo. *Journal of the International Society of Arts, Sciences and Technology*, 22 (3 and 4).

Coyle R and Hayward P (1995) *Apparition, Holographic Art in Australia*. University of Sydney, NSW, Australia: Power Publications.

Casdin-Silver H (1998) *The Art of Holography Exhibition Catalogue*. Lincoln, MA, USA: DeCordova Museum and Sculpture Park.

Holographische Visionen, Bilder durch Licht zum Leben erweckt (1991) Exhibition Catalogue, Museum für Holographie und neue Visuelle Medien, Pulheim, Germany: Edition Braus.

Holosphere, Formerly published by the Museum of Holography. New York: Boston, USA: Archieved at the MIT Museum.

Jackson R and Burns J (1976) *Holografi: Det 3-Dimensionella Mediet, Exhibition Catalogue*. Museum of Holography, New York: Stockholm Cultural Center.

Jung D (ed.) (2003) *Holographic Network*. Germany: Rasch Verlag, Bramsach.

Orazem V and Schepers J (1992) *Olografia, Holografie, Holography, Exhibition Catalogue*. Rocca, Paolina, Italy: Centro Espositivo.

Lipp A and Zec P (1985) *Mehr Licht/More Light, Exhibition Catalogue*. Kabel-Verlag: Hamburger Kunsthalle.

Pepper AT (ed.) (1992) *The Creative Holography Index: The International Catalogue for Holography*. Germany: Monand Press.

Pepper AT (ed.) (1996) *Art in Holography 2*. International Symposium on Art in Holography. Nottingham, UK: University of Nottingham.

Margaret Benyon M (1989) *Phases: Exhibition Catalogue*. New York, USA: Museum of Holography.

Tyler DE (ed.) (1990) *Report, International Congress on Art in Holography*. Notre Dame, Indiana, USA: Saint Mary's College.

Tyler DE (ed.) (2003) *Leading Lights, Women in Holography, Exhibition Catalogue*. Notre Dame, Indiana, USA: Saint Mary's College.

High-Resolution Holographic Imaging and Subsea Holography

J Watson, University of Aberdeen, Aberdeen, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Holography is well-known to most of us: there can be few who are not familiar with the white-light display hologram seen in exhibitions and shops and on credit cards. The impact of holography, though, spreads far and wide into areas as diverse as vibration analysis, image processing, holographic-optical elements, and optical computers. One aspect often overlooked is its suitability for high-precision imaging and accurate dimensional measurement. It is the unique ability of holography to recreate, from a flat photographic film, a three-dimensional image that is optically indistinguishable from the original which sets it apart from other imaging techniques. In particular, holography comes into its own when recording objects in a hazardous environment or where access is difficult, such as underwater or in nuclear fuel element inspection.

We describe here the use of holography for imaging, identification, and measurement of aquatic organisms. A detailed knowledge of the distribution and dynamics of marine and freshwater organisms and particles, such as plankton, is crucial to our understanding of such systems and how they affect our environment. Traditional methods of gathering *in situ* data on aquatic particles, such as net-collection and electronic counting or photography, are not usually suited to observing precise spatial relationships and can destroy the more delicate organisms. The overriding benefit of holography is that it permits nonintrusive and nondestructive observation and analysis of organisms in their natural environment, while preserving their relative spatial distribution. Holography also offers a permanent archive. The use of a short-duration pulsed laser is essential, in order to reduce the effects of vibration or movement in any of the optical components or objects being recorded. Consequently, the object scene is effectively 'frozen' at the recording instant, thereby allowing fast moving particles to be imaged. Compared to a photograph, a hologram captures vastly more data. Sequential holograms can record changes within the volume over a defined period of time and the wide recording dynamic range afforded by a hologram allows images to be captured that would otherwise be lost in noise.

Methodology of Holographic Recording and Replay

Of the many possible recording methods of holography two in particular, 'in-line' and 'off-axis', find use for high-resolution *in situ* imaging underwater. The concepts and fundamental principles outlined here are generally applicable wherever high-precision measurement is needed from a hologram.

In an in-line reference beam hologram (ILH), a single laser beam (Figure 1) is directed through the

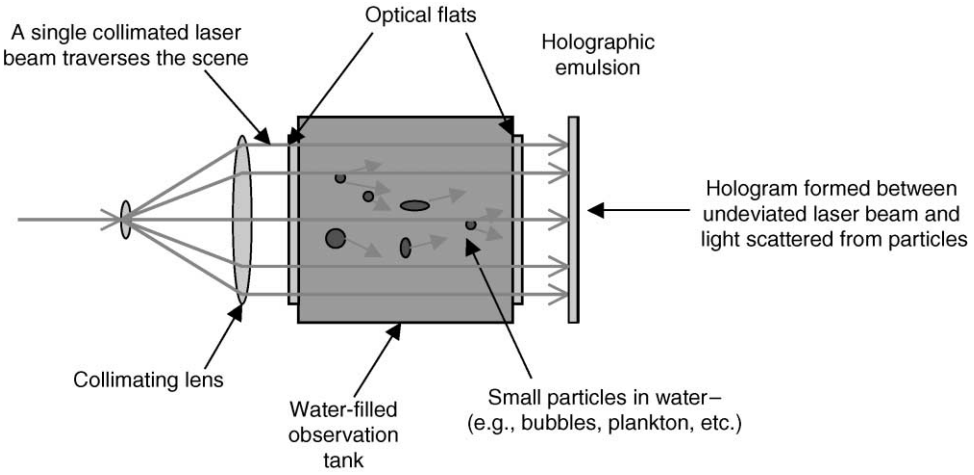


Figure 1 Recording an in-line hologram.

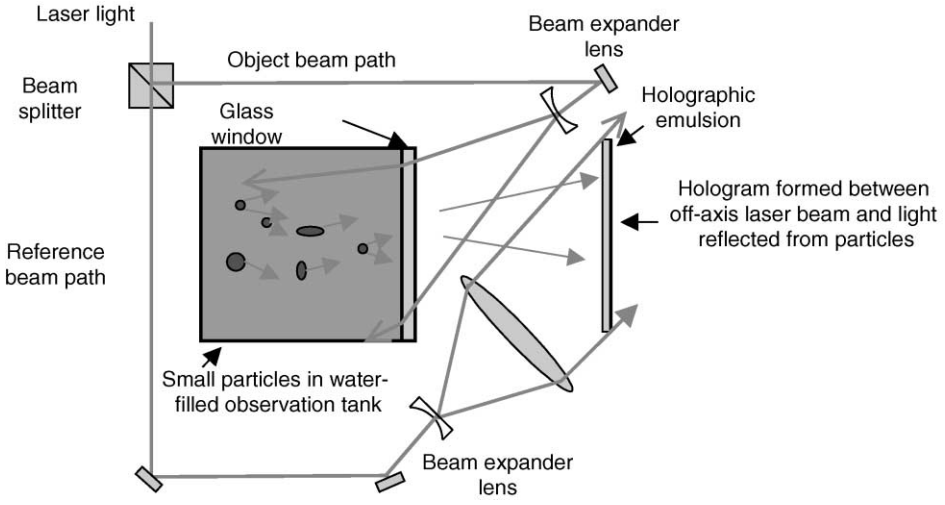


Figure 2 Recording an off-axis hologram.

sample volume towards the holographic plate and records the optical interference between light diffracted by the object and the undeviated portion of the illuminating beam. The subject volume is illuminated in transmission and the scene needs an overall transparency of about 80% so that speckle noise does not seriously degrade image quality. Particles down to around 5 μm dimension can be readily recorded and identified. The particular benefits of ILH are its geometric simplicity, minimization of laser coherence and energy requirements, and high resolution over a large depth of field.

For recording an off-axis reference beam hologram (OAH), a two beam geometry is utilized: one beam illuminates the scene and the other directly illuminates the holographic film at an oblique incidence angle (Figure 2). Interference occurs between diffuse light reflected from the scene and the angularly separated reference beam. On replay, the real and virtual images are similarly angularly separated which makes their interrogation easier. OAH is primarily applied to opaque subjects of large volume. Although there is no real upper limit to the size of particles that can be recorded (this is determined by available energy and coherence of the laser) the practical lower limit to the resolution that can be achieved is around 50 to 100 μm. The scene can be front, side or back illuminated (or all three combined) allowing high concentrations of large particles to be recorded, provided that they reflect sufficient light to the film, and allowing subject visibility to be optimized. OAH provides a more complete record than ILH, in the sense that it gives information on the surface

structure of the object, while ILH records only a silhouette.

Holograms are replayed or reconstructed in a variety of ways, depending on the recording parameters and subsequent processing of the recording medium. In scientific and engineering applications the requirements of high resolution and image fidelity demand that they are replayed with a laser, and viewed with some auxiliary optical system and (usually) a video camera. Depending on the orientation of the hologram, either the virtual image, which is apparently located behind the hologram and seen through it, or the real image, which is formed in space in front of the hologram, will be visible.

Considering first the off-axis hologram in the virtual image mode, the processed hologram is replaced in the position in which it was recorded and illuminated with an exact duplicate of the original reference wave, in terms of its wavelength, curvature, and beam angle. In this mode, which might be termed 'normal' viewing of an off-axis hologram, a virtual image (**Figure 3**) is observed as if viewing the scene through a window. Other than the colour, this image is almost optically indistinguishable from the original scene; it retains features such as parallax and three-dimensionality and does not suffer from perspective distortion or defocusing of the image.

For data extraction and measurement from a hologram, creation of the virtual image is not the best method of observation. Here the projected real image (**Figure 4**) is most useful. If we illuminate an off-axis hologram from behind, with a wave which is the exact conjugate of the original (i.e., one which possesses the same wavelength as the original but with the opposite direction and curvature), we generate an image that appears to float in the real space in front of the observer. Often a collimated reference beam is used in recording, since reconstruction is then a simple case of turning the hologram around. No lens is needed to form this image, and it is optically identical to the original wave that it is reversed left-to-right and back-to-front (pseudoscopic). Planar sections of the real image can be optically interrogated by directly projecting onto a video camera (often with the lens removed) which can be moved throughout the 3D volume. It is this 'optical sectioning' feature of a holographic recording that is so useful when nondestructive (noninvasive) evaluation must be made, at high resolution, of collections of objects which extend over a considerable depth.

A series of images taken from the real image of one hologram and captured on a video camera (**Figure 5**) illustrates how the viewing plane can be moved sequentially through the reconstruction to show firstly one marine organism and then another. The axial separation of the organisms is about 35 mm.

To replay an in-line hologram, it is also placed in a conjugate of the recording beam (**Figure 6**). The replayed hologram simultaneously forms two images, one virtual, the other real, which are located on the optic axis on opposite sides of the holographic plate. These images are co-axial, and located (for the typical case of collimated recording and replay beams) at equal distances in front of and behind the hologram plane. Although a true 3D real image is produced it has almost no parallax, and cannot easily be seen by the unaided eye. This image has inverted depth perspective and is seen in darkfield against the out-of-focus virtual image.

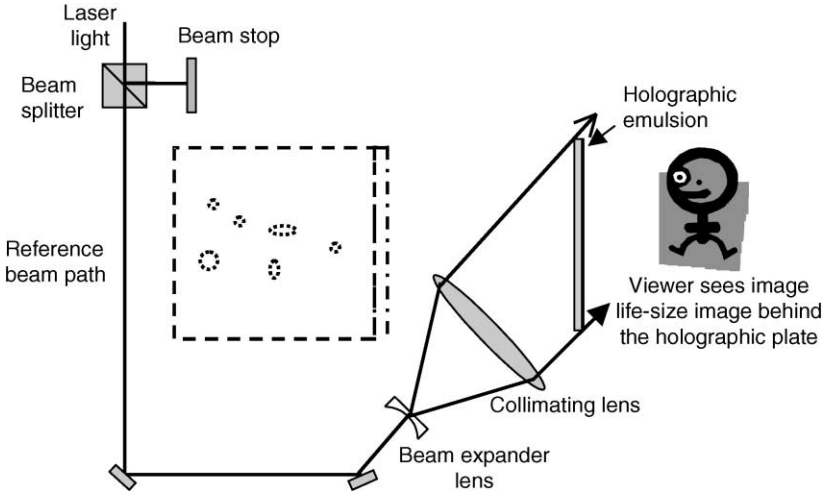


Figure 3 Replay of an off-axis hologram in virtual image mode.

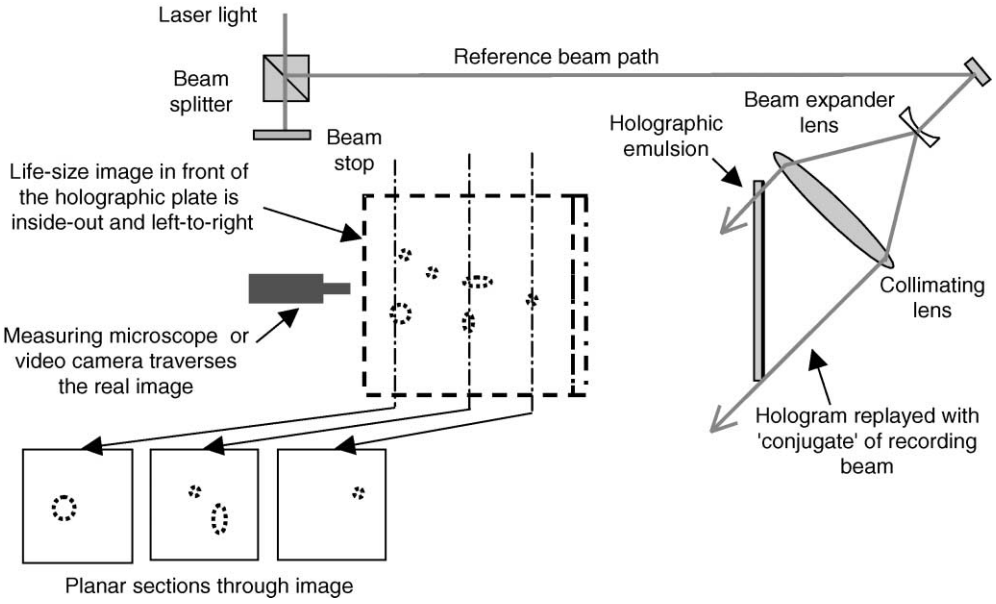


Figure 4 Replay of an off-axis hologram in real image mode.

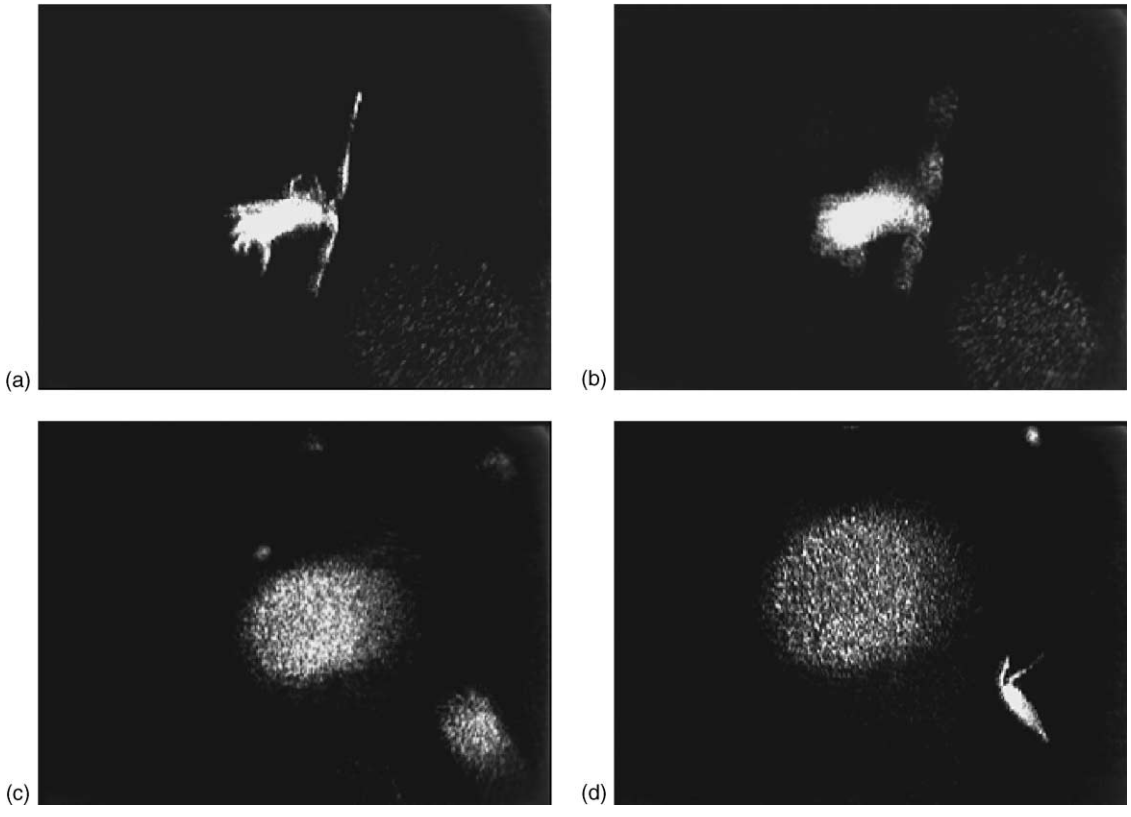


Figure 5 A series of images taken from a single hologram showing the concept of moving the viewing plane through the real image.

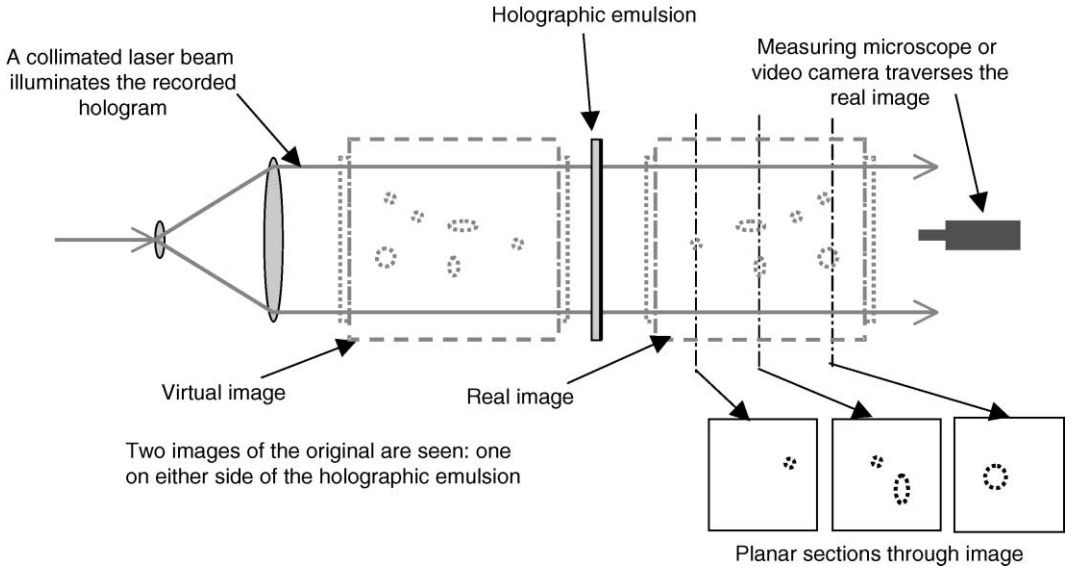


Figure 6 Replay of an in-line hologram.

Holographic Image Quality

In general, the usefulness of holography for accurate inspection and measurement is dependent on its ability to reproduce an image of the object which is low in optical aberrations and high in image resolving power. Regardless of whether or not the hologram is recorded underwater, high resolution, high contrast, and low noise are the dominant requirements, rather than image brightness. In practice, loss of resolution and image degradation can occur at any stage in the recording and replay of a hologram. All the primary monochromatic aberrations to be found in any optical system (spherical aberration, astigmatism, coma, distortion, and field curvature), may be present in the holographic image. The most significant factors affecting this are departure from conjugacy of the reconstruction beam; positional mismatch and misalignment between reference beam and film; degradation of the reconstruction beam due to poor optical quality of the film substrate or optical components; and, distortion of the fringe pattern recorded in the holographic film as a result of chemical processing. For precise and accurate reconstruction of the real image, the reconstruction reference beam should be the exact phase conjugate of that used in recording. Under these conditions, the lateral, longitudinal, and angular magnifications of the real image with all equal unity and aberrations will be reduced to a minimum.

Diffraction-limited resolution of a holographically produced real image, in the absence of all aberrations, is usually defined from the ability to distinguish two points in the image separated by a distance r. The resolution as defined in this way is dependent upon reconstruction wavelength, λ_c, the separation between hologram and reconstructed image, s, and the effective aperture, D, of the hologram:

$$r = 1.22s\lambda_c/D \qquad [1]$$

This relationship is well-known in imaging optics: it defines the ideal resolving power of a lens and is equally applicable to holography. In holography, though, the presence of speckle effects introduced by the coherence of the light and the finite aperture of the viewing system influence the resolving power. In practice, the minimum resolution is increased by a factor of two to three to take account of speckle. For off-axis holograms the choice of film size (effective aperture) has a considerable bearing on image resolution. Typically, for 100 mm diameter film the theoretical image resolution (neglecting speckle effects) is around 6 μm (150 lp/mm) with a 532 nm laser at a target distance of 1 m. For in-line holograms (with collimated illumination) the film size also defines the transverse dimensions of the volume recorded.

Wavelength matching between recording and reconstruction beams require that, if possible, the same type of laser be used in both replay and recording. Of course, this is not possible in field situations when the use of a pulsed laser is essential to ensure that the fringe pattern is not blurred by any movement in the system. In such applications, laboratory replay of the hologram may require the use of a tuneable laser to match wavelengths

precisely, although in practice an argon laser provides a useful selection of wavelengths. A mismatch between record and replay wavelengths will lead to distortion in the image, because longitudinal and lateral magnifications will no longer have the same value.

In contrast to off-axis holograms, the interference patterns on an in-line hologram are localized about the geometric shadow of the objects. The lack of a spatially separate reference beam leads to a requirement that there be a large fraction (>80%) of the beam unobscured. This limits the size of particles that can be recorded and sets an upper limit on the recordable concentration of particles (around 40 particles cm^{-3} for 20 μm diameter particles and a recorded depth of 1 m). There is also a need to balance the size of the particles with the object-to-hologram distance. The range of object size recordable on an in-line hologram has an upper limit set by the requirement to record in the far-field of the object, so that the conditions of Fraunhofer recording are satisfied:

$$z > d^2/\lambda \qquad [2]$$

where z is the object-to-film distance and d is the maximum dimension of the object to be recorded. In-line holograms are relatively easy to replay as the only requirement is to place the hologram normal to the replay beam. Following this criterion, the upper limit for good size measurement is about 2 to 3 mm particle size over a recording distance of a meter. Unlike off-axis holograms which are extremely sensitive to misalignment, the in-line geometry minimizes the effects of film rotation about its axes. A collimated recording beam is by far the easiest to duplicate upon replay. Small deviations from collimation can significantly affect conjugacy and hence resolution, introducing aberrations (mainly spherical) and a depth-dependent magnification (e.g., a cuboid replays as a trapezoidal prism).

In general, the interference field is captured on photographic emulsions; these may be laid on acetate or polyester film bases or on glass plates. If film is used, it is essential that it be held as flat as possible between thin flat glass plates or under vacuum, and should not be stretched or put under any strain during exposure. An in-line hologram does not require an extremely high-resolution recording medium. The fringe spacing is not as fine as the typical carrier fringe frequency used in off-axis holograms, and holograms can be made on conventional slow photographic emulsions. In practice, however, ultra-fine grained holographic film is almost always used since the intensity of the forward scattered light (a function of grain size) is then reduced: this is important in increasing the signal-to-noise ratio, as the replayed image is viewed against a background of the attenuated replay beam. Since the proportion of the illuminating beam that goes into the images is very low, this lack of angular separation requires that the average optical density of the hologram is high. The high optical density attenuates the undiffracted light, thus the image appears bright against a dark background, resulting in contrast reversal.

The exposed film is chemically processed to render the interference field permanent. This process is a crucial step in the holographic procedure, and especially critical for good off-axis holograms. Some of the factors, which have to be considered, include image brightness, image resolution, reconstruction wavelength, emulsion shrinkage, humidity, and noise level. It has been well established that for bright off-axis holograms on silver halide film, bleaching of the emulsion with its consequent conversion into a phase hologram is essential to maximize diffraction efficiency. Some forms of bleaching produce a nonuniform shrinkage of the emulsion which will give rise to astigmatism in the reconstructed image. Bleaching significantly increases the background noise levels of the hologram and introduces a high degree of variability into the result. In practice, the best repeatability will be obtained with amplitude processing if the low brightness can be tolerated.

Underwater Holography

When holography is extended to the inspection of underwater scenes, the hologram is recorded in water but the image is replayed, in the laboratory, in air. Because of the refractive index difference between recording and replay spaces, optical aberrations will be introduced into the reconstructed image, and will seriously impair the potential for precision measurement, unless some method of aberration reduction is incorporated. In ILH, only spherical aberration is significant since both reference and object beam angles are normal to the recording plane. However, in OAH the dominant aberrations are astigmatism and coma, which increase with the field angle of the subject in the reconstructed image. These limit resolution and introduce uncertainty in co-ordinate location, since the focal 'point' is now distributed over a finite region of image space. A key point here is an appreciation that the additional aberrations introduced are entirely due to the refractive index mismatch between object and image spaces and are unconnected with the holographic process itself: they are in essence the same distortions seen when looking into a fish-tank. Furthermore, the water

itself may be expected to influence the quality of the images produced. An increase in the overall turbidity of the water will adversely affect both in-line and off-axis techniques and would be expected to create a background noise that will reduce image fidelity.

Techniques, such as immersing the holographic plate in water during recording or direct replay back into water, have been used to minimize underwater aberrations, but such methods are generally impractical in the field. One practical solution, unique to holography, compensates the change of effective wavelength of light as it passes through water, by allowing a deliberate mismatch of recording and replay reference beams to balance the refractive index mismatch.

For holograms recorded and replayed in air, a usual prerequisite is that the reconstruction wavelength should remain unchanged from that used in recording. By noting the dependence of wavelength on the refractive index, we can apply the more general condition that it is the ratio λ/n that must remain constant. This relationship suggests that a hologram, immersed in water, and recorded at a wavelength λ_a will produce an aberration-free image in air when replayed at a reconstruction wavelength, λ_c, which is itself equivalent to the wavelength of light when passing through water, λ_w, i.e.:

$$\lambda_c = \lambda_w = \lambda_a n_a / n_w \qquad [3]$$

where n_a and n_w are the respective refractive indices of air and water. Thus, in principle, the aberrations may be completely eliminated by the introduction of an appropriate wavelength change between recording and replay. If a green (532 nm) laser is used in recording, the ideal replay wavelength is about 400 nm (i.e., 532 nm/1.33). However, no convenient laser lines exist at exactly this wavelength. Furthermore, complete correction assumes that the entire recording system be located in water. Since this is both impractical and undesirable, holograms are usually recorded with the holographic film in air behind a planar glass window. The additional glass and air paths affect the compensation of aberrations. However, third-order aberration theory shows that if the window-to-air path length ratio is appropriately chosen for a specific replay wavelength, then aberration balancing occurs and residual aberrations are reduced to a minimum over a wide range of field angles and object locations. For an air gap of 120 mm and a BK7 window of 30 mm thickness, a blue replay wavelength of 442 nm (HeCd laser) achieves a good performance over a full field angle of about 40°.

In most underwater applications the objects (or the camera) are in motion during the exposure. The effect of this motion is to blur out the finer fringes, and thus reduce resolution and contrast. In-plane motion is the most severe, and adopting the experimentally verified criterion that the maximum allowable motion is less than one-tenth of the minimum required fringe spacing of the smallest object, then for in-line holograms the maximum object motion must be less than one-tenth of the object's dimension. For particles of 10 μm dimension, and a typical Q-switched YAG laser pulse duration of 20 ns, a high transverse velocity of up to 50 m s^{-1} can be tolerated. Off-axis holograms are more demanding in their requirements and the maximum allowed velocity is reduced to about 3 m s^{-1}. This is, however, more than adequate for most field applications of the technique.

Underwater Holographic Cameras

The first-known use of holography for recording living marine plankton was by Knox in 1966, who recorded in-line holograms (ILH), using a short coherence length ruby laser, of a variety of living marine plankton species in a tank. However, to avoid problems with refractive index and wavelength changes, the holographic plate was immersed directly in water, which is completely impractical for field use. It was not until the late 1970s that quantitative data from replayed holograms of marine particles recorded *in situ* were obtained by Carder and his team. Small volumes (10 or 20 cm^3) were recorded using ILH and a low-power HeNe laser, which limited the technique to slowly moving objects. Later, Heflinger used the off-axis geometry with both pulsed argon-ion (514 nm) and pulsed xenon (535 nm) lasers to record plankton on 35 mm film. In order to compensate for the aberrations introduced by replaying the real image in air, the real image was projected back into the water tank using a conjugate beam from a continuous-wave argon laser. The sides of their tank were flexible so that the real image could be placed close to the exit window and viewed via an external microscope. Knox later developed a ship-borne in-line holographic camera which was deployed from the *RV Ellen B Scripps*. A pulsed xenon laser was used to record water volumes of up to 10^5 cm^3 to a water depth of 100 m.

It was not until the late 1990s that significant use of underwater holography for plankton recording came to the fore again. Joseph Katz at Johns Hopkins University, Baltimore, USA developed a camera based

around a ruby laser to record in-line holograms on rolls of film, in a 63 mm diameter sample volume that can be varied from 100 to 680 mm in length. The camera housing utilizes a two-tube design with one tube holding the laser and another containing film holder and power units. The recording volume is situated 890 mm above the camera to minimize disturbance to flow. The camera is designed as a 'Lagrangian drifter' to move with the water current and was successfully deployed in Chesapeake Bay, Baltimore, and Solomon's Island, Maryland. Holograms were recorded of various species of phytoplankton and zooplankton and represented a significant advance in the deployment of sub-sea holography.

From the early 1980s, a team led by Aberdeen University, Scotland, investigated the use of underwater holography for offshore applications. Primary interest was in inspection of subsea pipelines for corrosion and damage, and much of the work concentrated on an analysis of the optical aberrations introduced by recording in water and replaying in air and methods of compensating for such aberrations. However, this work was later adapted for holography of plankton and led to the consequent development of the HoloMar system. The HoloMar camera was the first to incorporate simultaneous ILH and OAH recording.

More recent developments have utilized digital or electronic recording of holograms in the in-line mode, and it is to be expected that much more use will be made of 'e-holography' in the future.

For field applications of holography, it is essential to record the holograms using a pulsed laser in order to create stable interference fringes. For subsea purposes, a laser with a wavelength in the blue-green region of the spectrum will match the oft-quoted peak transmission window of seawater. In practice, a frequency-doubled Nd-YAG laser with a wavelength of 532 nm is a good choice, although ruby lasers operating at 694 nm are also acceptable if transmission distance is not an issue. Usually OAH requires more energy and places stricter constraints on coherence properties than ILH. Beam quality is governed by quality of the optical components in the beam path and also by the output beam of the laser itself. Because off-axis holography uses diffuse front/side illumination, energies of up to several hundred millijoules are often required for large (up to a meter in diameter) poorly reflective objects, and coherence lengths of greater than a meter are needed to record large scene depths. To obtain a long coherence length, and thus enable large-volume objects to be recorded, it is necessary that the laser operate in a single longitudinal cavity mode so that the bandwidth of the laser line is kept small. To enable smooth even illumination of both subject and holographic emulsion, good spatial beam quality is needed. This usually dictates laser output beams with either a Gaussian (TEM_{00}) mode profile or a flat profile, and spatial filtering during beam expansion. Compared to off-axis holography, the demands placed upon lasers for ILH are minimal. The only essential requirement is for a good spatial mode structure. Energy requirements are moderate: as little as 10 mJ is enough since little light is lost during the passage of the beam through the recording volume. The laser coherence length can be quite short, and only depends on the

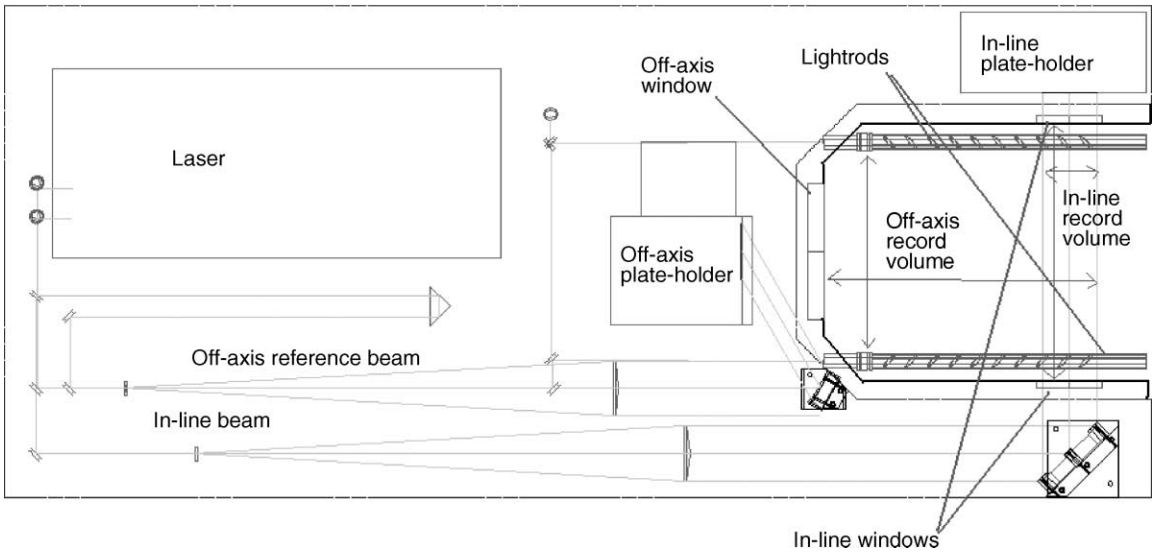


Figure 7 The HoloMar camera configuration showing the separate in-line and off-axis beam paths.

effective hologram area, thus lasers operating on more than one longitudinal mode are usually quite suitable.

The HoloMar System

The HoloMar system (shown diagrammatically in Figure 7) was developed to record simultaneous in-line and off-axis holograms of partially overlapping, but orthogonal, views of the water column. Although the recording views are orthogonal to each other, the arrangement provides an element of cross-correlation between holograms. It is designed around a Q-switched, frequency-doubled Nd-YAG laser with an output energy of 700 mJ in a single pulse of less than 10 ns duration and a coherence length in excess of 2 m. The laser output is split into two beams: one of 100 mJ and the other of 600 mJ energy. The 100 mJ beam is further split into two equal energy beams and expanded and collimated to a diameter of about 90 mm. One path forms the illuminating beam for the in-line mode and the other path forms the reference beam for the off-axis mode. The in-line path records a water column of 470 mm by 90 mm diameter (9,000 cm^3) at 400 mm from the front face of the housing, whereas the off-axis hologram records almost the entire volume (50,000 cm^3) delineated by the front window to the end of the arms that house the in-line system (see Figure 7). Three 'lightrods', encompassing hollow perspex tubes with 10 glass beam-splitter plates distributed equally along the length, positioned at 45° to the main axis, provide roughly even side illumination to the off-axis volume. The front face of the camera is shown in Figure 8; the lightrods are positioned on either side of the recording volume (two on one side, one on the other). The exit window for the in-line beam path and the off-axis window can also be seen.

Two motor-driven plate holders and transport mechanisms hold up to 20 plates for in-line and 24 for off-axis holograms. One pair of plates can be exposed every 10 seconds, or they can be exposed separately. The camera is remotely operated from the surface via a control console (PC). A network of micro-controllers serves as the backbone of the in-camera control system and communication between topside and camera modules is implemented using a Controller Area Network (CAN) bus via an umbilical cable. The entire camera is enclosed in a water-tight stainless steel housing of 1,000 mm diameter by 2,200 mm length. The prototype system is capable of either ship deployment or attachment to a fixed buoy and allows recording down to a water depth of 100 m.

The HoloMar camera is shown, in Figure 9, being lowered into Loch Etive, a sea loch on the west coast of Scotland, on its maiden dive. Figures 10 and 11 show holographic images of plankton recorded on this cruise. The first shows a view from a virtual

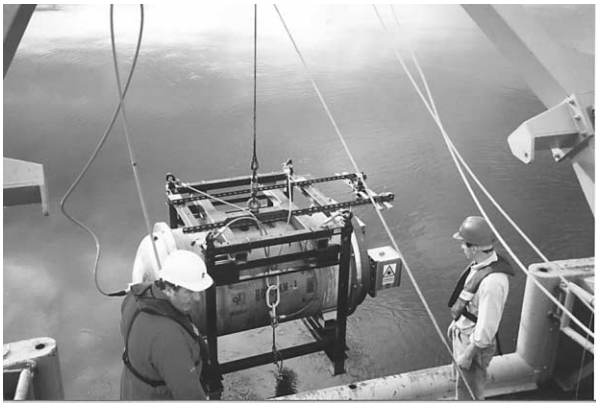


Figure 9 The HoloMar camera being lowered into Loch Etive, Scotland.

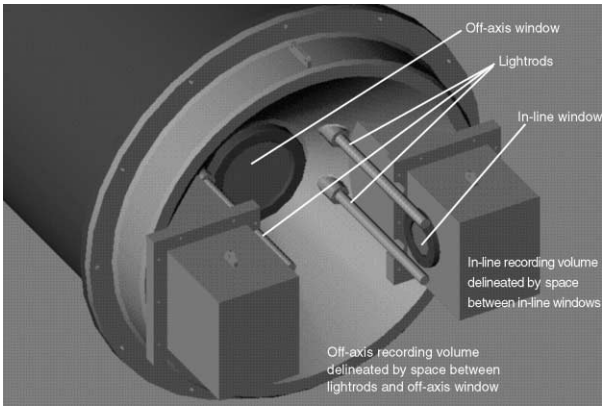


Figure 8 Front-end of the HoloMar camera showing the 'lightrods' and in-line and off-axis windows.



Figure 10 A photograph taken from a virtual image replay of an off-axis hologram recorded in Loch Etive.

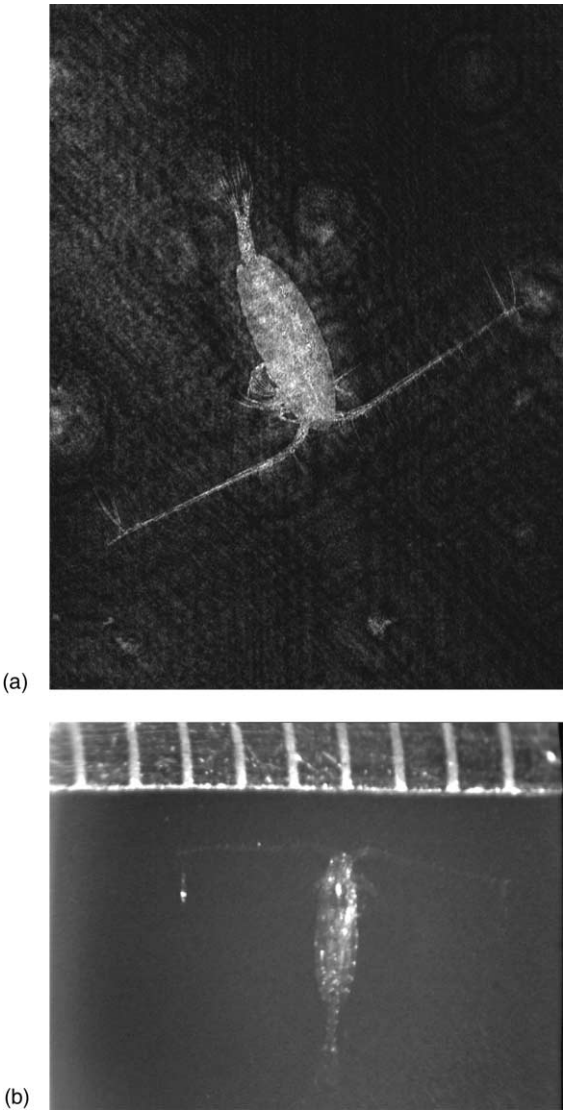


Figure 11 Images of calanoid copepods from (a) an in-line hologram recorded at 70 m depth and (b) an off-axis hologram recorded at 60 m depth, the millimeter scale is added at the replay stage. Each organism is approximately 4 mm long.

replay of an off-axis hologram at a wavelength of 514 nm: this is what a viewer would see when looking from the inside of the holocamera into the laser-illuminated water. Each bright dot signifies an object of interest; two fiducial wires can also be seen fixed diagonally across the arms at the end of the camera, some 300 mm from the off-axis window. In the lower left of the image a long 'string-like' object about 40 mm long is probably 'floc', the dead organic matter that forms the food-stuff of many zooplankton. Objects like this would be destroyed in net-collection. Figure 11 shows images taken from the real image projection of in-line and off-axis holograms. The holograms are replayed using a collimated beam at wavelengths of 514, 488, 457, or 442 nm, depending on location in the image volume. Some of the contrasting features between in-line and off-axis holograms are apparent: the in-line image shows clearly the outline of the organism with detail resolvable down to about 10 μm, whereas the off-axis hologram, although not achieving this resolution, reveals the body structure of the organism.

List of Units and Nomenclature

d	dimension of object to be recorded
D	effective aperture of hologram
n	refractive index
n_a	refractive index of air
n_w	refractive index of water
r	minimum resolvable separation between two points on the hologram
s	distance between hologram and image location
z	distance between hologram and object location
λ	wavelength
λ_a	in-air wavelength of recording beam
λ_c	reconstruction wavelength
λ_w	effective wavelength of recording beam in water

See also

Holography, Applications: Holographic Recording Materials and Their Processing. **Holography, Techniques:** Computer-Generated Holograms. **Phase Control:** Phase Conjugation and Image Correction.

Further Reading

Beers JR, Knox C and Strickland JDH (1970) A permanent record of plankton samples using holography. *Limnology and Oceanography* 15: 967–970.

Carder KL (1979) Holographic microvelocimeter for use in studying ocean particle dynamics. *Optical Engineering* 18: 524–525.

Hariharan P (1984) *Optical Holography*, 2nd edn. Cambridge, UK: Cambridge University Press.

Heflinger LO, Stewart GL and Booth CR (1978) Holographic motion pictures of microscopic plankton. *Applied Optics* 17: 951–954.

Hobson PR and Watson J (2002) The principles and practice of holographic recording of plankton. *Journal of Optics A: Pure and Applied Optics* 4: S34–S49.

Katz J, Donathay PL, Zhang J, King S and Russell K (1999) Submersible holocamera for detection of particle

characteristics and motions in the ocean. *Deep-Sea Research* 46: 1455–1481.
Kilpatrick JM and Watson J (1993) Underwater hologrammetry: reduction of aberrations by index compensation. *Journal of Physics D: Applied Physics* 26: 177–182.
Knox C (1966) Holographic microscopy as a technique for recording dynamic microscopic subjects. *Science* 153: 989–990.
Malkiel E, Alquaddoomi O and Katz J (1999) Measurements of plankton distribution in the ocean submersible holography. *Measurement Science and Technology* 10: 1141–1152.
Owen RB and Zozulya AA (2000) In-line digital holographic sensor for monitoring and characterizing marine particulates. *Optical Engineering* 39: 2187–2197.
Stewart GL, Beers JR and Knox C (1970) Application of holographic techniques to the study of marine plankton in the field and the laboratory. *Proceedings of SPIE* 41: 183–188.
Sun H, Dong H, Player MA, *et al.* (2002) In-line digital holography for the study of erosion processes in sediments. *Measurement Science and Technology* 13: L1–L6.
Watson J, Alexander S, Craig G, *et al.* (2001) Simultaneous in-line and off-axis subsea holographic recording of plankton and other marine particles. *Measurement Science and Technology* 12: L9–L15.

Holographic Recording Materials and Their Processing

H I Bjelkhagen, De Montfort University, Leicester, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The first off-axis transmission holograms in the USA were recorded on Kodak 649F spectroscopic silver halide material (now discontinued). For the first reflection holograms, which were recorded in Russia, a very fine-grain emulsion (Lippmann emulsion) had to be prepared to resolve the fine interference pattern associated with such holograms. Since then various types of recording materials for holography have been developed and manufactured, including improved silver halide emulsions.

The following provides a brief description of the holographic recording process. For recording holograms and holographic optical elements (HOEs), dichromated gelatin, photopolymer materials, and photoresist materials are employed, in addition to silver halide materials. Existing holographic recording materials are described, as well as suitable processing methods to obtain high-quality holograms and HOEs.

Holographic Recording Materials

To produce a hologram, a micropattern, caused by interference between the object beam and the reference beam, is recorded in a light-sensitive high-resolution material. The quality of a holographic image depends on a number of factors, such as: the geometry and stability of the recording setup; the coherence of the laser light; the reference and object beam ratio; the type of hologram produced; the size of the object and its distance from the recording material; the recording material and the emulsion substrate used; the processing technique applied as well as the reconstruction conditions. The material must comply with certain requirements to be suitable for this purpose. The most important of these concerns the resolving power of the material. The recording material must be able to resolve the highest spatial frequencies of the interference pattern created by the maximal angle θ between the reference and the object beams in the recording setup (Figure 1).

If λ is the wavelength of the laser light used for the recording of a hologram and n the refractive index of the emulsion, then the closest separation d_e between the fringes in the interference pattern created by the angle θ between the reference and the object beams in the recording setup is

$$d_e = \frac{\lambda}{2n \sin(\theta/2)} \qquad [1]$$

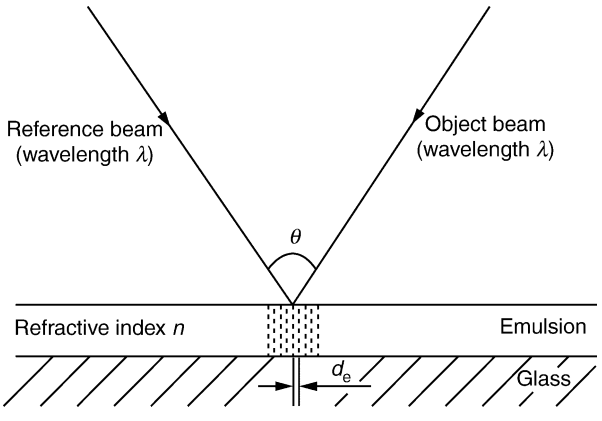


Figure 1 Demand on resolution for recording a hologram.

For high-quality holograms the resolution limit of the material must be higher than the minimum value obtained according to the above formula. For a reflection hologram recorded in blue light ($\lambda = 400\,\text{nm}$) with an angle of 180° between the beams, a minimum resolving power of 7600 lines $(\text{mm})^{-1}$ is required.

In holography, the resolution of the holographic image and the resolving power of the recording material are not directly related to the way in which they are in photography.

Equation [1] gives the minimum resolving power needed to record a hologram. The resolution of the image depends on the area (diameter D) of the recording material, the recording laser wavelength (λ), and the distance (L) between the recording material and the object (Figure 2).

Theoretically, the resolution of the holographic image is the diffraction-limited resolution that can be obtained when the information is collected over an aperture equal to the size of the recording holographic plate. In principle, the larger the holographic plate the better the resolution will be. To obtain high image resolution, it is important that the recorded interference fringes will not change or be distorted during processing. To prevent that from happening, a stable support for the emulsion (like a glass plate) is needed and the processing method applied must not affect the recorded fringe pattern in the emulsion. In coherent imaging systems laser speckles are present which may also have an effect upon image resolution.

There are two main types of holograms: amplitude and phase holograms. In a pure amplitude hologram only the absorption varies with the exposure (after processing), whereas in a pure phase hologram, either the refractive index or the emulsion thickness varies with the exposure. If the hologram is thin, $d \approx 0$, phase variations are then caused by surface relief variations only. If the hologram is thick and has a negligible surface relief ($\Delta d = 0$), phase variations are caused by index variations only. In many cases, the phase modulation in a hologram is a combination of the two different extreme types (a complex hologram).

For amplitude holograms, amplitude transmission T_a against exposure or log exposure is used (Figure 3). For a phase hologram, the corresponding curve is the phase shift against the log exposure relation (Figure 4).

One of the most important hologram characteristics is its diffraction efficiency η. It is defined as the diffracted intensity E_i of the wanted diffraction order of the hologram in relation to the incident

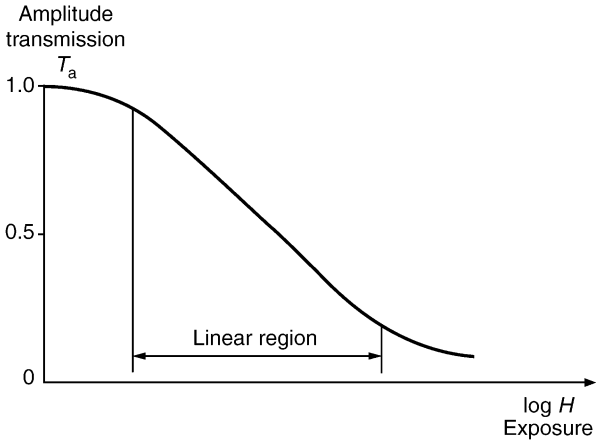


Figure 3 Amplitude transmission T_a versus the logarithm of exposure H used for the characterization of amplitude transmission holograms.

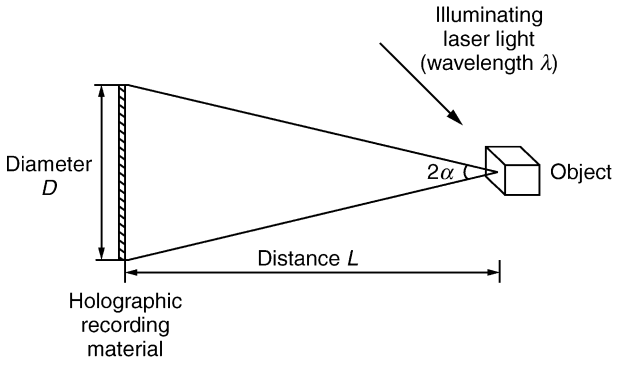


Figure 2 Resolution of a holographic image depends on the area (diameter D) of the recording material, the recording laser wavelength λ, and the distance L between the recording material and the object.

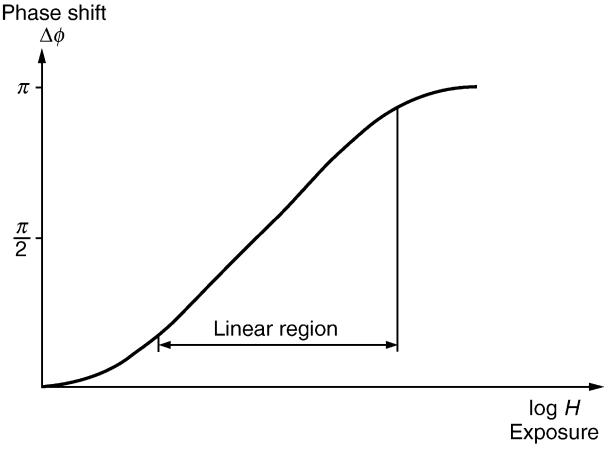


Figure 4 Phase shift $\Delta\phi$ versus the logarithm of exposure H used for the characterization of phase transmission holograms.

intensity of the reconstruction beam E_r : $\eta = E_i/E_r$. In the definition of η, the incidental light losses caused by surface reflection and base absorption should be subtracted, but are often ignored. Other important hologram characteristics are the signal-to-noise ratio (SNR) and the dynamic range of the recording material.

The resolution capability of an image reproduction process is normally described by the Modulation Transfer Function (MTF). Briefly, a test pattern containing a sinusoidal variation in illuminance combined with a continuous variation in spatial frequency along one direction is recorded. The modulation M of the pattern in the test target is

$$M = (H_{\mathrm{MAX}} - H_{\mathrm{MIN}})(\mathrm{H}_{MAX} + \mathrm{H}_{MIN})^{-1} \qquad [2]$$

where **H** is the exposure incident on the photographic material. When this pattern is recorded in the material, light scattering will take place in the emulsion, which will reduce the original contrast of the pattern. Therefore, the modulation of the pattern will be decreased, in particular at high spatial frequencies. The effective exposure modulation M' will then be

$$\mathrm{M}' = (H'_{\mathrm{MAX}} - H'_{\mathrm{MIN}})(H'_{\mathrm{MAX}} + H'_{\mathrm{MIN}})^{-1} \qquad [3]$$

where $\mathbf{H}'$ is exposure in the emulsion.

The original modulation M is constant and accurately known; it is also independent of the spatial frequency. After the tested emulsion has been processed, the corresponding 'exposed' modulation is obtained from the density variations. The ratio between modulation M' in the emulsion and modulation M of the incident exposure is called the modulation transfer factor, also called response

$$R = M(M')^{-1} \qquad [4]$$

If the response is plotted as the function of spatial frequency, this curve will then be the modulation transfer function of the material. The MTF for both a holographic emulsion and a conventional photographic emulsion are illustrated in Figure 5.

There are many ways of categorizing holograms but here we are mainly interested in definitions related to the recording material and its processing. One of the valid criteria is the thickness of the recording layer (as compared to the interference fringe spacing within the layer), i.e., the layer coated on the material substrate. Holograms can thus be classified into 'thin' or 'thick' (sometimes also called 'plane' or 'volume' holograms, respectively). To distinguish between the two types, the Q-parameter is normally used; it is defined in the following way:

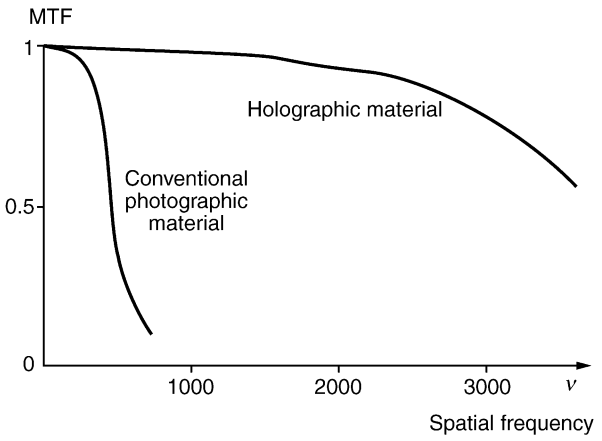


Figure 5 Graphical representation of the modulation transfer function (MTF) curves for both a conventional emulsion and a holographic emulsion are shown.

$$Q = 2\pi\lambda d(n\Lambda^2)^{-1} \qquad [5]$$

where λ is the wavelength of the illuminating light, d is the thickness of the layer, n is the refractive index of the emulsion, and Λ the spacing between the recorded fringes. A hologram is considered thick if $Q \geq 10$, and thin if $Q \leq 1$. Holograms with Q-values between 1 and 10 can be treated as either thin or thick.

Recording Materials

The most common holographic materials used for recording holograms are the following: silver halide emulsions; dichromated gelatin emulsions; photopolymer materials; photoresist materials; thermoplastic materials; bacteriorhodopsin; and photorefractive crystals.

The following materials can also be considered for holographic recordings, but not many practical applications have been reported so far: chalcogenide glasses; ferroelectric-photoconductors; liquid crystals; magneto-optic films; metal and organic dye ablative films; photochromic and photodichroic materials; photothermorefractive glass; and transparent electrophotographic films.

Sensitivity of Photographic and Holographic Materials

The time of exposure for a given material will depend on the sensitivity of the material used, as

well as on the intensity of the interference pattern. Some holographic materials must be processed after exposure in a specific way to obtain a hologram. The recorded intensity variations are converted during this processing step to local variations in optical density or refractive index and/or thickness of the recording layer.

Exposure H is defined as the incident intensity E times the time t of exposure of the recording material. If the intensity is constant during the whole exposure time, which is usually the case, then

$$H = Et \qquad [6]$$

Holographic materials are usually characterized using radiometric units. The radiometric equivalent of illuminance is irradiance. The unit of irradiance is $\mathrm{W(m)}^{-2}$ and the exposure will then be expressed in $\mathrm{J(m)}^{-2}$. The sensitivity of a holographic emulsion is most often expressed in $\mu\mathrm{J(cm)}^{-2}$ or $\mathrm{mJ(cm)}^{-2}$. Knowing the sensitivity of the material used and having measured the irradiance at the position of the holographic plate, the exposure time can be calculated using the above formula, i.e.,

$$\text{Exposure time} = \text{sensitivity/irradiance}$$

Holographic materials are sensitized in such a way that they are optimized for laser wavelengths commonly used in holography.

Silver Halide Materials

A silver halide recording photographic material is based on one type, or a combination of silver halide crystals embedded in a gelatin layer. The emulsion is coated on a flexible or stable substrate material. Silver-halide grain sizes vary from about 10 nanometers for the ultra-fine-grain holographic emulsions to a few micrometers for highly sensitive photographic emulsions (Table 1).

Silver salts are only sensitive to UV light, violet, and deep blue light. Therefore special sensitizers (dyes) must be added to the emulsion to make it sensitive to other parts of the spectrum. Orthochromatic emulsions are sensitive to green light. If the material has been sensitized to both green and red light, it is said to be panchromatic.

In the following, silver halide materials manufactured by photographic and holographic companies are described. The main difference between the new materials and the previous emulsions from Agfa, Ilford, etc., is that for the current silver halide emulsions, the grain sizes are smaller. The present holographic silver halide emulsion manufacturing process is more or less based on the Russian ultrafine-grain emulsion technology.

Table 1 Silver halide emulsion grain sizes

Type of emulsion	*Average grain diameter [nm]*
Ultra-fine-grain holographic emulsion	10–30
Fine-grain holographic emulsion	30–50
Fast holographic emulsion	50–100
Chlorobromide paper emulsion	200
Lithographic emulsion	200–350
Fine-grained photographic emulsion	350–700
Fast photographic emulsion	1000–2000
Fast medical X-ray emulsion	2500

Substrates for Holographic Emulsions

The material on which the emulsion is coated has a strong bearing on the final quality of the hologram. The best choice is often a glass plate as it is mechanically stable and optically inactive. High resolution imaging, hologram interferometry, holographic optical elements (HOEs) and spatial filters are a few examples where a very stable emulsion support is important. Mass produced holograms are mainly recorded on film substrates. However, the use of film has many advantages as compared to that of glass (breakage, weight, cost, size, etc.). Film substrates are used exclusively in the production of large-format holograms. Film substrates are of mainly two types: a polyester (polyethylene terephthalate); or a cellulose ester, commonly triacetate (cellulose triacetate) or acetate–butyrate. Polyester is mechanically more stable than triacetate film and it is also less sensitive to humidity. Because of the higher tensile strength, the polyester film can be made thinner than the triacetate film. On the other hand, polyester is birefringent, which can cause problems when recording holograms. In Table 2 most current commercial silver halide materials for holography are listed.

Processing of Silver Halide Emulsions

In silver halide emulsions a latent image is formed during exposure, which later is developed to a silver image. Processing (development, fixing, and bleaching) silver halide emulsions requires some expertise to obtain high-quality holograms. With the exception of DCG emulsions, other holographic materials are easier to process and more straightforward. Here it is not possible to describe all techniques employed and recipes used for silver halide processing. The reader is referred to the Further Reading section for more details.

Table 2 Commercial holographic recording materials

Material	*Thickness [μm]*	*Spectral sensitivity [nm]*	*Sensitivity [μJ (cm)$^{-2}$] at*				*Resolv. power [lp (mm)$^{-1}$]*	*Grain size [nm]*
			442	*514*	*663*	*694*		
SILVER HALIDE EMULSIONS:								
Slavich								
Red PFG-01	7	<700	–	–	80	–	>3000	35–40
Red PFG-03M	7	<700	–	–	1500	–	>5000	10–20
Green VRP-M	7	<550	–	80	–	–	>3000	35–40
Pan PFG-03C	9	400–700	1000	2000	1000	–	>5000	10–20
Colourholographic								
Red BB-700	7	<700	–	–	50	150	>2500	50–60
Red BB-640	7	<650	–	–	150	–	>4000	20–25
Green BB-520	7	<540	150	150	–	–	>4000	20–25
Blue BB-450	7	<470	150	–	–	–	>4000	20–25
Kodak								
Red 131PX	9	<650	2	–	0.5	–	>1250	70
Red 131CX	9	<650	2	–	0.5	–	>1250	70
Red 120PX	6	<750	60	–	40	40	>2500	58
Red 120CX	6	<750	60	–	40	40	>2500	50
FilmoTec-ORWO								
Red HF65	10	<650	–	–	1500	–	5000	30
Green HF53	10	<550	–	700	–	–	5000	30
Ultimate								
Ultimate 15	7	<700	–	150	150	150	>5000	15
Ultimate 08	7	<650	120	200	200	–	>7000	8
DICHROMATED GELATIN EMULSIONS								
Slavich								
Blue PFG-04	16	<515	1×10^5	2.5×10^5	–	–	10000	–
FilmoTec-ORWO								
Blue GF 40	6/20	*	*	–	–	–	–	–
* delivered: non-sensitized								
Holotec								
Custom orders only								
THERMOPLASTIC MATERIALS								
Tavex America								
Pan TCC-2	–	<800	1	1	1	1	1500	–
MD Diffusion								
Custom orders only								
PHOTORESIST MATERIALS								
Towne Technologies								
UV-Blue Shipley 1800	1.5–2.4	<450	1.5×10^5	–	–	–	1000	–
Hoya Corporation								
Custom orders only								
BACTERIORHODOPSIN MATERIALS								
MIB GmbH								
BR-WT B-type	30–100	<650	–	80×10^3	–	–	5000	–
BR-D96N M-type	30–100	<650	30×10^3	–	–	–	5000	–

The developing agent is the most important constituent in a holographic developer. In addition to the developing agent, a developer consists of the following components: a preservative (or antioxidant); a weak silver-solvent agent; an accelerator (or activator); a restrainer; and a solvent which is usually water. Employing a developer which contains a silver-solvent agent, solution-physical development

or semi-physical development can be performed. Such a developer is particularly suitable for processing the ultrafine-grain emulsions.

Holographic Developers

Silver halide materials for holography are all of the fine-grain type. Such materials require the application of special processing techniques to achieve the best possible results. For the processing of amplitude holograms a high-contrast developer is required, such as Kodak D-19. For holograms which are going to be bleached (phase holograms), developers based on ascorbic acid, hydroquinone, metol, pyrocatechol, or pyrogallol are frequently used. The following developers are recommended for the new silver halide materials: D-19; CW-C2; AAC; MAA-3; PAAP; and for pulsed laser recordings, the SM-6 developer. Colloidal development, employing a highly diluted developer which contains silver solvents, is recommended for the ultrafine-grain emulsions. Colloidal silver grains formed in the emulsion during such development are very small which means that very little absorption occurs. The method produces holograms with high diffraction efficiency combined with very low light scattering noise. The Russian GP-2 developer is recommended for colloidal development. Recipes for common holographic developers are found in Table 3.

Bleach Baths

Bleaching is used to convert a developed hologram (containing a silver image) into a phase hologram, to obtain a higher diffraction efficiency than is possible for amplitude holograms. During bleaching the silver image is converted into a water-soluble or an insoluble silver compound. Holographic bleaching techniques can be divided into three categories:

1. Conventional or direct (rehalogenating) bleaching;
2. Fixation-free rehalogenating bleaching;
3. Reversal (complementary) or solvent bleaching.

In conventional bleaching, the developed silver image is, after fixing, converted into a transparent silver halide compound, i.e., after the unexposed silver halide crystals have been removed. This method is not recommended for ultrafine-grain emulsions. Instead the fixation-free rehalogenating bleaching is preferred. The hologram is bleached directly after development, without fixing, leaving the unexposed silver halide crystals in the emulsion. A rehalogenating bleach bath consists of: an oxidizing agent; an alkali halide (often KBr); and a buffer. Among the rehalogenating bleaches, the following bleach baths work well: PBQ-bleach; Ferric EDTA; and the PBU-bleaches (PBU-amidol or the PBU-metol).

In reversal bleaching, the developed silver image is not fixed. Instead it is converted into a soluble silver complex which is removed from the emulsion during bleaching, leaving only the original unexposed silver halide grains in the emulsion. A reversal bleach bath consists of an oxidizing agent and a buffer. For ultrafine-grain silver halide emulsions, the reversal method is not recommended. The very small unexposed silver halide grains are also affected (partly dissolved) during reversal bleaching, resulting in low diffraction efficiency. For materials with larger grains, a more diluted version of the dichromate Kodak R-9 bleach can be

Table 3 Developers for silver halide emulsions

Constituents	*CWC2*	*GP-2*	*MAA-3*	*PAAP*	*AAC*	*SM6*
Methylphenidone		0.2 g				
Catechol	10 g					
Ascorbic acid	5 g		10 g	18 g	16 g	18 g
Hydroquinone		5 g				
Metol			2.5 g			
Phenidone				0.5 g		6 g
Sodium sulfite	5 g	100 g				
Sodium phosphate (diabasic)				28.4 g		28.4 g
Urea	50 g					
Potassium hydroxide		5 g				
Sodium hydroxide				12 g		12 g
Sodium carbonate	30 g		55.6 g		60 g	
Ammonium thiocyanate		12 g				
Distilled water	1 l	1 l	1 l	1 l	1 l	1 l
Dilution:	no	15 ml dev. + 400 ml water	no	no	no	no

Table 4 Bleach baths for silver halide emulsions

Constituents	*Ferric-EDTA (rehal.)*	*PBQ (rehal.)*	*PBU-(amidol/metol) (rehal.)*	*Copper (rehal.)*	*Chrome (reversal)*	*PSSB (reversal)*
Amidol *or* metol*			1 g			
Cupric bromide			1 g			
Potassium persulfate			10 g			20 g
Potassium dichromate					4 g	
p-Benzoquinone		2 g				
Ferric III sulfate	30 g					30 g
Copper sulfate				35 g		
Disodium EDTA	30 g					
Boric acid		1.5 g				
Citric acid			50 g			
Potassium bromide	30 g	30 g	20 g	30 g		
Sodium hydrogen sulfate	30 g			30 g		30 g
Sulfuric acid					4 ml	
Distilled water	1 l	1 l	1 l	1 l	1 l	1 l

*To be added after all other constituents have been dissolved.

used. Recipes for common holographic bleach baths are found in Table 4.

Commercial Silver Halide Emulsions

In Table 2 most of the current silver halide holographic emulsions are listed. The main manufacturer of holographic silver halide emulsions is the Micron branch of the SLAVICH Joint Stock Company photographic company located outside Moscow (2 pl. Mendeleeva, 152140 Pereslavl-Zalessky, Russia, www.slavich.com). Both film (triacetate) and glass plates are manufactured. Many different sizes are produced, including 115 cm by 10 m film rolls. Recommended holographic developers: CW-C2, SM-6. Many rehalogenating holographic bleach baths work well and, in particular, the PBU-type bleaches are recommended.

Colourholographic Ltd is a new company based in England (Braxted Park, Great Braxted, Witham CM8 3BX, England; www.colourholographic.com). The Colourholographic materials are based on the HRT emulsion, previously manufactured in Germany by Dr Birenheide. At the present time only glass plates are manufactured in England. Recommended processing solutions for the new HRT materials are a sodium-sulfite-free ascorbic acid–metol developer followed by a rehalogenating ferric EDTA or PBU bleaches. For reflection holograms, a pyrogallol developer and the ferric-EDTA bleach can be used.

Eastman Kodak's holographic materials are produced in America (343 State Street, Rochester, New York 14650, USA; www.kodak.com/go/PCB products). These plates are usually made to order only. Materials can be ordered with (marked -01) or without (marked -02) anti-halation layer. The 120 emulsion can be used for recording holograms using a pulsed ruby laser. Kodak recommends the developer Kodak D-19 for the processing.

FilmoTec is located in Germany (Chemie Park Bitterfeld-Wolfen, Röntgenstrasse, D-06766 Wolfen, Germany; www.filmotec.de). The company manufactures ORWO holographic emulsions which are coated only on triacetate film, 104 cm wide in 10 and 30 m lengths.

In France, M. Yves Gentet has started a small-scale emulsion manufacturing company: Atelier de Création d'Art en Holographie Yves Gentet (50, rue Dubourdieu, F-33800 Bordeaux, France; www.perso.wanadoo.fr/holographie) Monochromatic and panchromatic ultrafine-grain emulsions are produced called 'Ultimate'. Plates and film up to 60 cm by 80 cm are produced. An ascorbic acid–metol developer is recommended followed by a Ferric EDTA bleach. In addition, the Russian GP-2 developer can be used. Also a special developer can be ordered from the company.

Konica in Japan is manufacturing a green-sensitive emulsion for the use in the American holographic VOXEL 3D medical imaging system and, in addition, for the Japanese market only.

Dichromated Gelatin Materials

Dichromated gelatin (DCG) is an excellent recording material for volume phase holograms and HOEs. The grainless material has its highest sensitivity in the UV region, but extends into the blue and green parts of the spectrum. There are possibilities to sensitize the

emulsion in the red part of the spectrum as well. However, DCG is most often exposed with blue laser wavelengths. Depending on processing parameters, diffraction efficiency and bandwidth can be controlled. It is easy to obtain high diffraction efficiency combined with a large SNR.

During the exposure of a DCG emulsion to UV or blue light, the hexavalent chromium ion (Cr^{6+}) is photo-induced to trivalent chromium ion (Cr^{3+}) which causes cross-linking between neighboring gelatin molecules. The areas exposed to light are hardened and become less soluble than the unexposed areas. Developing consists of a water wash which removes the residual or unreacted chemical compounds. Dehydration of the swollen gelatin follows after the material has been immersed in isopropanol, which causes rapid shrinkage resulting in voids and cracks in the emulsion, thus creating a large refractive index modulation. The underlying mechanism is not completely clear because high modulation can also be caused by the binding of isopropanol molecules to chromium atoms at the cross-linked sites. The DCG material has a rather low sensitivity of about 100 mJ$(cm)^{-2}$. In many cases DCG plates are prepared by fixing an unexposed silver halide emulsion and then sensitizing it in a dichromate solution. It is also, for example, possible to make a 10 μm thick emulsion on a 8″ by 10″ glass plate using 1 g ammonium dichromate mixed with 3 g photographic grade gelatin in 25 ml deionized water. Then the emulsion is spin-coated on the glass substrate or the emulsion is applied by the doctor blade coating technique. Many HOE manufacturers prefer to produce their own emulsions but there are also commercial DCG emulsions available on the market.

Processing of DCG holograms are performed in graded alcohol (propanol) baths starting with water–propanol solutions of high water content and ending with pure propanol (cold or hot) baths. Depending on the wanted HOE characteristics, one has to carefully control the propanol/water mixtures as well as the temperature of the different baths. Cold baths produce better uniformity and lower noise. Warm baths can yield high index modulation but often with increased noise. The bandwidth can be controlled by the processing temperature and the ratio of propanol/water mixtures.

Commercial DCG Materials

Slavich in Russia is one manufacturer of presensitized dichromated plates for holography. The DCG emulsion is marked PFG-04. Plates up to a size of 30 by 40 cm can be ordered.

The FilmoTec DCG emulsions are only available on 190 μm triacetate film, 104 cm wide in 10 and 30 m lengths; emulsion thickness 6 or 20 μm. Note that the film needs to be sensitized in dichromate solution before recording. This means that the company only supplies large-format gelatin-coated film.

Holotec is a German-based company which offers presensitized DCG coated on both plates and film (Jülicher Strasse 191, D-52070 Aachen, Germany; www.holotec.de). The Holotec emulsion is the high-quality DCG emulsions developed by Professor Stojanoff in Aachen. The company can supply large-format DCG glass plates (m^2 size) or film (PET-Polyethylenterephtalat), pre-sensitized and ready to use.

Photopolymer Materials

Photopolymer materials have become popular for recording phase holograms and HOEs, in particular for mass production of holograms since some photopolymer materials only require dry processing techniques.

A photopolymer recording material, such as the DuPont material, consists of three parts: a photopolymerizable monomer; an initiator system (initiates polymerization upon exposure to light); and a polymer (the binder). First, an exposure is made to the information-carrying interference pattern. This exposure polymerizes a part of the monomer. Monomer concentration gradients, formed by variation in the amount of polymerization due to the variation in exposures, give rise to diffusion of monomer molecules from the regions of high concentration to the regions of lower concentration. The material is then exposed to regular light of uniform intensity until the remaining monomer is polymerized. A difference in the refractive index within the material is obtained. The DuPont material requires only a dry processing technique (exposure to UV light and a heat treatment) to obtain a hologram. The DuPont photopolymer materials have a coated film layer thickness of about 20 μm. The photopolymer film is generally coated in a 12.5″ width on a 14″ wide Mylar® polyester base which is .002″ thick. The film is protected with a .00092″ thick Mylar® polyester cover sheet.

The recording of a hologram on DuPont polymer is rather simple. The film has to be laminated to a piece of clean glass or attached to a glass plate using an index-matching liquid. Holograms can be recorded manually, but in order to produce large quantities of holograms, a special machine is required. For hologram replication a laser line scanning technique can provide the highest production rate.

The photopolymer material needs an exposure of about 10 mJ(cm)$^{-2}$.

After the exposure is finished, the film has to be exposed to strong white or UV light. DuPont recommends about 100 mJ(cm)$^{-2}$ exposure at 350–380 nm. After that, the hologram is put in an oven at a temperature of 120 °C for two hours in order to increase the brightness of the image. The process is simple and very suitable for machine processing using, for example, a baking scroll oven.

Commercial Pholymer Materials

Currently the only manufacturer of holographic photopolymer materials is E.I. du Pont de Nemours & Co. (DuPont Holographics, POB 80352, Wilmington, DE 19880, USA; www.dupont.com/holographics). In the past it was possible to obtain photopolymer materials from DuPont, but in 2002 the company changed its policy and today it is producing materials mainly for DuPont Holographics, but approved customers may still be able to buy material from DuPont. The reason for the restrictions is that many applications of mass-produced holograms are in the field of document security, where DuPont's photopolymers are used to produce optical variable devices (OVDs).

Photoresist Materials

Photoresist materials in holography are used for making masters for display and security holograms. The recorded relief image in the photoresist plate is then used to make the nickel shim needed as the tool for embossing holograms into plastic materials.

The exposure to actinic radiation produces changes in the photoresist layer that result in a solvency differentiation as a function of exposure. Processing is done with a suitable solvent dissolving either the unexposed or exposed regions, depending on whether the resist is of the negative or the positive type. For optical recording, positive photoresist (exposed resist removed during development) is preferred to the negative type because of the higher resolving power and low scatter. On the resulting surface, relief pattern particles of nickel are deposited by electrolysis to make a mold, which can then be used as an embossing tool. The photoresist process can be used for making transmission holograms only. If an embossed hologram is mirror-backed by using, for example, an aluminum coating process, it can be used in the reflection reconstruction mode as well. Photoresists are UV and deep-blue sensitive. The material is spincoated on glass substrates to obtain a thickness of between 0.5 and 2 μm, then it is baked at about 75 °C for 15 minutes. A typical photoresist for holography (e.g., Shipley Microposit 1350) has a sensitivity of about 10 mJ(cm)$^{-2}$. The most common laser wavelengths for recording photoresist holograms are the following wavelengths: 413 nm (Krypton-ion), 442 nm (Helium–Cadmium), and 458 nm (Argon-ion). At higher wavelengths, e.g., 488 nm, the sensitivity is rather low. Development of Shipley resist is performed in Shipley 303A developer, diluted 1 part developer + 5 parts de-ionized water. Normally only a 10 second development time is needed (under agitation). After that the plate has to be rinsed in de-ionized water for 2 minutes. The plate needs to be dried quickly with a high pressure steam of compressed nitrogen or dry, filtered compressed air.

Commercial Resist Materials

The main American manufacturer of photoresist is Shipley Co. (1457 Macarthur Rd., #101, Whitehall, PA 18052, USA). There are a few companies which make plates based on Shipley resist.

Towne Technologies, Inc. (6–10 Bell Ave., Somerville, NJ 08876, USA; www.townetech.com) produces spin-coated plates using Shipley resist (Shipley S-1800 series). Towne plates have a sublayer of iron oxide which enhances adhesion of the resist during electroplating operations. In addition, the sublayer eliminates unwanted backscatter from the glass substrate. The plates are developed in Shipley 303A developer.

Hoya Corporation (7-5 Naka-Ochiai 2-chome, Shinjuku-ku, Tokyo, Japan; www.hoya.co.jp) is a producer of mask blanks coated with 0.5 μm thick Shipley S-1800 resist (SLW plates). Different types of glass as well as quartz plates are produced. Hoya supply plates for holography, however, custom orders only.

Thermoplastic Materials

The holographic thermoplastic is a multilayer structure coated on glass or film. The substrate is first coated with a conducting layer, e.g., evaporated gold, then a photoconductor, e.g., poly-n-vinyl carbazole, PVK, that has been sensitized with, e.g., 2,4,7-trinitro-9-fluorenone (TNF), and on top of this layer a thermoplastic coating is deposited (usually a styrene-methacrylate material). The recording of a hologram starts with a uniform charging of the surface of the thermoplastic material using a corona charger. The charge is divided between the photoconductor and the thermoplastic layer. Exposure and the consequent photogeneration in the photoconductor cause charges of opposite signs to migrate to the interface with the thermoplastic layer and

the substrate. This will not change the charge, but only decrease the surface potential. Before the image can be developed, the material has to be recharged with the uniform corona charger. This step adds additional charges to the exposed areas, in proportion to the reduced potential which is also proportional to the exposure. The material is then heated to a softening temperature of the thermoplastic layer which will be deformed due to the electrostatic forces acting on it. The material is then cooled and the image is fixed as a relief surface pattern in the thermoplastic layer. The processing time after the exposure necessary to obtain the image is between ten seconds and half a minute. If the material is heated to a somewhat higher temperature, the pattern disappears and the image is erased. The thermoplastic material can now be used for the recording of another hologram, repeating the recording procedure. Such a cycling technique can be repeated hundreds of times, without any serious effect on the quality of the image. That, and the fast dry processing made such materials popular for the use in holographic cameras for nondestructive testing. The sensitivity is between 10 and 100 $\mu J(cm)^{-2}$ over the whole visible electromagnetic spectrum.

Commercial Thermoplastic Materials

In the past, several companies produced thermoplastic materials and recording equipment for holography. Since there is very little demand for such materials today, there is only one company which can deliver thermoplastic equipment for holography: Tavex America Inc. (14 Garden Road, Natick, MA 01760, USA; www.tavexamerica.com). Tavex America manufactures a thermoplastic camera, TCC-2 in which 40 mm by 40 mm plates are used. For obsolete systems, thermoplastic film can still be obtained from MD Diffusion in France (93 rue d'Adelshoffen, F67300 Schiltigheim, France).

Bacteriorhodopsin

Photochromic materials, a class to which bacteriorhodopsin (BR) belongs, have not been used much in holography. Photochromics are real-time recyclable materials which need no processing for development. They can easily be erased and reused over and over again. In holography the only successful material has been the BR film which was introduced only a few years ago. BR is a biomolecule with several applications in photonics, comprising a light-driven molecular proton pump with exceptional photoelectric and photochromic properties. BR is a photochromic protein, found in the photosynthetic system of a salt-marsh bacterium, termed purple membrane (PM), of the cell membrane of *Halobacterium salinarum*. BR thin films are made from polymer suspensions of the isolated purple membrane. Some of the desirable properties of BR films are real-time writing and erasing in microsecond time-scales, durability allowing millions of write/erase cycles without degradation, very high spatial resolution of 5000 $lp(mm)^{-1}$, reasonably good light sensitivity of approximately 10 $mJ(cm)^{-2}$, the encoding of both amplitude and phase patterns, and the use of visible laser light for writing and reading. The composition of the films may also be altered by chemical means to control the optical characteristics, and the BR molecules themselves may be mutagenically changed for performance optimization. Particular properties that can be enhanced in these ways are image lifetimes and the diffraction efficiency of holograms recorded in the films. One advantage of this material is that it is reversible. Therefore it has been considered to be used in holographic storage systems and for real-time holographic interferometry applications. BR film can be used for recording polarization holograms because of BR films inherent photo-inducable optical anisotropy. The material is also possible to expose with pulsed lasers, which makes it suitable for fast holographic recording.

The BR film works in the following way. When illuminated with green light, the BR molecule undergoes an absorption resulting in a maximum shift of about 160 nm towards the red and its color changes from purple to yellow. The yellow image can be erased by illuminating the BR material with blue light. The recording–erasing cycle can be repeated about a million times. Suitable BR-films for holography are obtained by embedding purple membranes into inert matrices like polyvinylalcohol or polyacrylamide. For hologram recording and erasure two photochemical conversions $B \rightarrow M$ (B-type holograms) and $M \rightarrow B$ (M-type holograms) can be employed. After excitation by light, BR cycles through a sequence of spectrally distinguishable intermediates and returns to the initial state. The time associated with this thermal relaxation process depends on several aspects, for example, type of BR (wildtype or variant), pH value of film-matrix, film-temperature, etc. For example, the M intermediate is populated from the B state (absorption spectrum peaked at 570 nm) by a green pumping beam and information is recorded with blue light, which initiates the photochemical $M \rightarrow B$ transition. The M state has its absorption peak at 410 nm. In BR variants in which the thermal relaxation of the M state has been prolonged, photocontrolled switching between these two states can be achieved. Only when

the pumping beam is incident upon the BR film can information be recorded with blue light. The pumping beam can be used as a readout beam for the recorded hologram.

Commercial BR Materials

A holographic camera based on a BR film is manufactured in Germany. The holographic system, FringeMaker-*plus*, developed at the University of Marburg is marketed by Munich Innovation Biomaterials (MIB) GmbH. (Hans-Meerwein-Str., D-35032 Marburg, Germany; www.mib-biotech.de). MIB is also a manufacturer of both wildtype and genetically modified BR film for holography and other scientific/technical applications. BR-films offered by MIB can be used in two different recording configurations: the transition from the initial B to M-state (induced with yellow light) serves as basis for optical data recording and processing (B-type recording); or the photochemically induced transition from M to B state (M-type recording). In the latter case, yellow light is used to control the population of the M-state and blue light is used to write the information. MIB offers BR-films with normal and slow thermal relaxation characteristics. Typical relaxation time ranges (RTR), after which the photochemically excited BR molecules have returned to the original state, are between 0.3 and 80 s. The BR film is sealed between two windows of high-quality optical glass. The BR film layer has a thickness of 30–100 μm, dependent on the optical density of the film. The diameter of the film device is 25 mm and the clear aperture is 19 mm.

Photorefractive Crystal Materials

Photorefractive crystals are electro-optic materials which are suitable for recording volume phase holograms in real-time. These crystals are important components in holographic storage systems and are seldom used in other holographic image applications. Holograms recorded in crystals, such as: lithium niobate ($LiNbO_3$); barium titanate ($BaTiO_3$); potassium tantalate niobate (KTN); barium sodium neobate (SBN); bismuth silicon oxide (BSO); and bismuth germanium oxide (BGO), consist of bulk space charge patterns. An interference pattern acting upon a crystal will generate a pattern of electronic charge carriers that are free to move. Carriers are moving to areas of low optical illumination where they get trapped. This effect will form patterns of net space charge, creating an electrical field pattern. As these crystals have strong electro-optic properties, the electrical field pattern will create a corresponding refractive index variation pattern, which means a phase hologram. These holograms can be immediately reconstructed with a laser beam differing from the beams creating the hologram. A hologram can also be erased and a new hologram can be recorded in the crystal. Holograms can be fixed by converting the charge patterns to the patterns of ions which are not sensitive to light. In some materials the hologram can be fixed by heating the crystal to above 100 °C based on thermally activated iconic conductivity. The resulting iconic is frozen upon cooling the crystal back to room temperature. More details about these materials and storage applications are found in the Further Reading.

See also

Holography, Techniques: Overview.

Further Reading

Bjelkhagen HI (1993) *Silver Halide Recording Materials for Holography and Their Processing*. Springer Series in Optical Sciences, vol. 66. Heidelberg: Springer-Verlag.

Bjelkhagen HI (ed.) (1996) *Selected Papers on Holographic Recording Materials*. SPIE Milestone Series, vol. MS130. Bellingham, WA, USA.

Lessard RA (ed.) (1995) *Selected Papers on Photopolymers, Physics, Chemistry, and Applications*. SPIE Milestone Series, vol. MS114. Bellingham, WA, USA.

Sincerbox GT (ed.) (1994) *Selected Papers on Holographic Storage*. SPIE Milestone Series, vol. MS95. Bellingham, WA, USA.

Smith HM (ed.) (1977) *Holographic Recording Materials*, Springer Topics in Applied Physics, vol. 20. Berlin: Springer-Verlag.

HOLOGRAPHY, TECHNIQUES

Contents

Overview
Color Holography
Computer-Generated Holograms
Digital Holography
Holographic Interferometry
Sandwich Holography and Light in Flight

Overview

C Shakher and A K Ghatak, Indian Institute of Technology, New Delhi, India

© 2005, Elsevier Ltd. All Rights Reserved.

Major Milestones

1948:	Essential concept for holographic recording by Dennis Gabor.
1960:	The first successful operation of a laser device by Theodore Maiman.
1962:	Off-axis technique of holography by Leith and Upatnieks.
1962:	Yu N Denisyuk suggested the idea of three-dimensional holograms based on thick photoemulsion layers. His holograms can be reconstructed in ordinary sunlight. These holograms are called Lippmann–Bragg holograms.
1964:	Leith and Upatnieks pointed out that a multicolor image can be produced by a hologram recorded with three suitably chosen wavelengths.
1969:	S A Benton invented 'Rainbow Holography' for display of holograms in white light. This was a vital step to make holography suitable for display applications.

Holography is the science of recording an entire optical wavefront, both amplitude and phase information, on appropriate recording material. The record is called a hologram. Unlike conventional photography, which records a three-dimensional scene in a two-dimensional format, holography records true three-dimensional information about the scene.

Holography was invented by Gabor in 1948 and his first paper introduced the essential concept for holographic recording – the reference beam. Holography is based on the interference between waves and, it provides us with a way of storing all the light information arriving at the film in such a way that it can be regenerated later.

The use of the reference beam is utilized because the physical detectors and recorders are sensitive only to light intensity. The phase is not recorded but is manifest only when two coherent waves of the same frequency are simultaneously present at the same location. In that case, the waves combine to form a single wave whose intensity depends not only on intensities of the two individual waves, but also on the phase difference between them. This is key to holography. The film record, or hologram, can be considered as a complicated diffraction grating. Holograms bear no resemblance to conventional photographs in that an image is not actually recorded. In fact, the interferometric fringes which are recorded on the recording material are not visible to an unaided eye because of extremely fine interfringe spacing (~ 0.5 micrometer). The fringes which are visible on the recording material are the result of dust particles in the optical system used to produce the hologram.

Gabor's original technique is now known as in-line holography. In this arrangement, the coherent light source as well as the object, which is a transparency containing small opaque details on a clear background, is located along the axis normal to the photographic plate. With such a system, an observer focusing on one image observes it superposed on the out-of-focus twin image as well as a strong coherent background. This constitutes the most serious problem of Gabor's original technique.

The first successful technique for separating the twin images was developed by Leith and Upatnieks in 1962. This is called the off-axis or side band technique of holography. We shall consider mainly the off-axis technique here.

Basic Holography Principle

Light from a laser or any light beam is characterized by spatial coherence (l_c) and temporal coherence (τ_c), which is discussed in all textbooks on optics. For typical laboratory hologram recording, the required degree of coherence of laser radiation is determined by the type of object and geometrical arrangement being used for the recording. The following condition must hold:

$$L \ll c\tau_c \qquad [1]$$

where τ_c is the coherence time, and L the maximum path length difference between the two waves chosen to record the hologram. To begin recording, two wavefronts are derived from the same laser source. One wavefront is used to illuminate the object and another is used as a reference wavefront. The wavefront derived by illuminating the object is superimposed with the reference wave and the interference pattern is recorded. These object and reference waves must:

- be coherent (derived from the same laser); and
- have a fixed relative phase at each point on the recording medium.

If the above conditions are not met, the fringes will move during the exposure and the holographic record will be smeared. Therefore, to record the hologram one should avoid air currents and vibrations. The recording medium must have sufficient resolution to record the fine details in the fringe pattern. This will be typically of the order of the wavelength. To create a three-dimensional image from the holographic process one then has to record and reconstruct the hologram. The object and reference waves are derived from the same laser.

Holograms record an interference pattern formed by interference of object and reference wavefronts, as explained above. We may mention here that for the two plane waves propagating at an angle θ between them, the spacing of the interference fringes is given by

$$\alpha = \frac{\lambda_0}{2\sin(\theta/2)} \qquad [2]$$

where λ_0 is the free space wavelength.

Processing of the hologram (developing, fixing, and washing of the recording material) yields a plate with alternating transparent and opaque parts, variation of refractive index, or variation of height corresponding to intensity variation in the fringe pattern. Such a plate can be regarded as a complicated diffraction grating. In this process, the hologram is illuminated with monochromatic, coherent light. The hologram diffracts this light into wavefronts which are essentially indistinguishable from the original waves which were diffracted from the object. These diffracted waves produce all the optical phenomena that can be produced by the original waves. They can be collected by a lens and brought into focus, thereby forming an image of the original object, even though the object has since been removed. If the reconstructed waves are intercepted by the eye of an observer, the effect is exactly as if the original waves were being observed; the observer sees what appears to be the original object in true three-dimensional form. As the observer changes his viewing position, the perspective of the image scene changes; parallactic effects are evident and the observer must refocus when the observation point is changed from a near to a distant object in the scene. Assuming that both the construction and reconstruction of the hologram are made with the same monochromatic light source, there is no visual test which can be made to distinguish between the real object and the reconstructed image of the object. It is as if the hologram were a window through which the apparent object is viewed.

Hologram of a Point Object

Consider a hologram recorded with a collimated reference wave normal to the recording plate and a point object inclined at a certain angle. If the hologram is illuminated once again with the same collimated reference wave, it reconstructs two images, one virtual true image and the other real image. However, the two images differ in one very important respect.

While the virtual image is located in the same position as the object and exhibits the same parallax properties, the real image is formed at the same distance from the hologram but in front of it. Corresponding points on the real and virtual images are located at equal distances from the plane of the hologram; the real image has the curious property that its depth is inverted. Such an image is not formed with a normal optical system; it is therefore called a pseudo image as opposed to a normal or orthoscopic image.

This depth inversion results in conflicting visual clues, which make viewing of the real image

psychologically unsatisfactory. Thus, if O_1 and O_2 are two elements in the object field, and if O_1 blocks the light scattered by O_2 at a certain angle, the hologram records information only on the element O_1 at this angle and records no information about this part of O_2. An observer viewing the real image from the corresponding direction then cannot see this part of O_2, which, contrary to normal experience, is obscured by O_1, even though O_2 is in front of O_1.

Production of an Orthoscopic Real Image

An orthoscopic real image of an object can be produced by recording two holograms in succession. In the first step, a hologram is recorded of the object with a collimated reference beam. When this hologram is illuminated once again with the collimated reference beam that was used to record it, it reconstructs two images of the object at unit magnification, one of them being an orthoscopic virtual image, while the other is a pseudoscopic real image. A second hologram is then recorded of this real image with a second collimated reference beam.

When the second hologram is illuminated with a collimated beam it reconstructs a pseudoscopic virtual image located in the same position as the real image formed by the second hologram is an orthoscopic image. Since a collimated reference beam is used throughout, the final real image is the same size as the original object and free from aberrations.

In addition to these characteristic linked intensities with the three-dimensional nature of the reconstruction, the holographic recording has several other properties. Each portion of the hologram can reproduce the entire image scene. If a smaller and smaller portion of the hologram is used for reconstruction, there is loss of image intensity and reconstruction. When a hologram is reversed, such as in contact printing processes, it will still reconstruct a positive image indistinguishable from the image produced by the original hologram.

Simple Mathematical Description of Holography

Let the object and reference waves be given by

$$U_o = O(x, y)e^{i\phi(x,y)} \qquad [3]$$

and

$$U_r = R\, e^{iky\sin\theta} \qquad [4]$$

U_r describes the reference (plane wave) propagating at angle θ to the z-axis.

The intensity at the hologram plane is

$$\begin{aligned} I &= |U_r + U_o|^2 \\ &= |U_r|^2 + |U_o|^2 + U_r^* U_o + U_r U_o^* \end{aligned} \qquad [5]$$

As can be seen from the above equation, the amplitude and phase of the wave are encoded as the amplitude and phase modulation of a set of interference fringes. The material used to record the patterns of interference fringes described above is assumed to provide linear mapping of the intensity incident during the reconstruction process into amplitude transmitted by or reflected from the recorded material. Usually both light detection and wavefront modulation are performed by photographic plate/film. We assume the amplitude transmission properties of the plate/film after processing to be described by

$$T = T_o - b(E - E_o) \qquad [6]$$

where the exposure E, at the film is $E = I\tau$; here τ is the exposure time. T_0 is the transmittance of the unexposed plate.

If the hologram is reilluminated by the reference wave, the transmitted wave amplitude will be

$$U_t = U_r(T_o - b\tau(|U_o|^2 + U_r^* U_o + U_r U_o^*)) \qquad [7]$$

The first term is reference wave times a constant. The second term is the reference wave modulated by $|U_o|^2 = O(x, y)^2$; it gives small-angle scattering about the reference wave direction. The third term is proportional to U_o. It is the same as the original object wave (note this is only so if the reconstruction wave is identical to the reference wave). The fourth term is

$$-b\tau U_r^2 U_0^* = -b\tau R^2\, e^{iky2\sin\theta} O(x, y)e^{-i\phi(x,y)} \qquad [8]$$

This is essentially a wave traveling in a direction $\sin^{-1}(2\sin\theta)$ to the z-axis, with the correct object amplitude modulation but its phase reversed in sign, producing a conjugate wave.

Types of Holograms

Primarily, the holograms are classified as thin and thick, based on the thickness of the recording medium. When the thickness of the recording medium is small compared with the average spacing of the interference fringes, the hologram can be treated as a thin hologram. Such holograms are characterized by

spatially varying complex amplitude transmittance:

$$\mathbf{t}(x, y) = [\mathbf{t}(x, y)]\exp[-i\phi(x, y)] \qquad [9]$$

Thin holograms can be further categorized as thin amplitude holograms or thin phase holograms. If amplitude transmittance of the hologram is such that $\phi(x, y)$ is constant while $\mathbf{t}(x, y)$ varies over the hologram, the hologram is termed an amplitude hologram. For a lossless phase hologram, $|\mathbf{t}(x, y)| = 1$, so that the complex amplitude transmittance is caused by variation in phase.

When the thickness of the hologram recording material is large compared to the fringe spacing of the interference fringes, the holograms may be considered as volume holograms. These may be treated as a three-dimensional system of layers corresponding to a periodic variation of absorption coefficient or refractive index, and the diffracted amplitude is at a maximum when Bragg's condition is satisfied. In general, the behavior of thin and thick holograms corresponds to Raman Nath and Bragg diffraction regimes. The distinctions between two regimes is made on the basis of a parameter Q, which is defined by the relation:

$$Q = \frac{2\pi\lambda_o d}{n_o \Lambda^2} \qquad [10]$$

where

Λ = spacing of fringe on hologram, measured normal to the surface;
d = thickness of recording medium;
n_o = mean refractive index of the medium; and
λ_o = wavelength of light.

Small values of Q ($Q < 1$) correspond to thin gratings, while large values of Q ($Q > 10$) correspond to volume gratings. For values of Q between 1 and 10, intermediate properties are exhibited. However, this criterion is not always adequate. The boundaries between the thin and thick holograms are given by the value of

$$P = \frac{\lambda_o^2}{\Lambda^2 n_o n_1} \qquad [11]$$

where P^{-2} is the relative power diffracted into higher orders, and $P < 1$ for thin holograms and $P > 10$ for thick holograms. For P having values between 1 and 10, the hologram may be thin or thick, depending on the other parameters involved.

Holograms can also be classified based on whether they are reconstructed in transmitted light or reflected light. For transmission holograms, during the recording stage, two interfering wavefronts make equal but opposite angles to the surfaces of recording medium and are incident on it from the same side. Reflection holograms reconstruct images in reflected light. They are recorded such that the interfering wavefronts are symmetrical with respect to the surface of the recording medium but are incident on it from opposite sides. When the angle between the interfering wavefronts is maximum (180°), the spacing between the fringe planes of the recording medium is minimum. Under such conditions, reflection holograms may have wavelength sensitivity high enough to be reconstructed, even with white light.

An important development in the field of holography was the invention of rainbow holograms by Benton in 1969. This provides a method for utilizing white light for illumination when viewing the holograms. The technique does so by minimizing the blur introduced by color dispersion in transmission holograms, at the price of giving up parallax information in one dimension.

Recording Materials

An ideal recording medium for holography should have a well matching spectral sensitivity corresponding to available laser wavelengths, linear transfer characteristics, high resolution, and low noise. It should also be either inexpensive or indefinitely recyclable. Toward achieving the above-mentioned properties, several materials have been studied, but none has been found so far that meets all the requirements. Materials investigated include: silver halide photographic emulsion; dichromated gelatin plates/films; silver halide sensitized gelatin plates/films; photo-resists; photo polymer systems; photochromics; photo thermoplastics; and ferro-electric crystals. Recently, the use of storage and processing capabilities of computers, together with CCD cameras, has been used for recording holograms.

Silver halide photographic emulsions are a commonly used recording material for holography, mainly because of relatively high sensitivity and easy availability. Manufacture, use, and processing of these emulsions have been well standardized. The need for wet processing and drying may constitute a major drawback. Dichromated gelatin can be considered an almost ideal recording material for volume phase holograms, as it has a large refractive index modulation capability, high resolution, low absorption and scattering.

Because of these features, dichromated gelatin has been extensively investigated. The most significant disadvantage is its comparatively low sensitivity. Pennington *et al.* developed an alternative technique, which combines the advantage of silver halide emulsions (high sensitivity) with that of dichromated

gelatin (low absorption scattering and high stability). It involves exposing silver halide photographic emulsion and then processing it, so as to obtain a volume phase hologram consisting solely of hardened gelatin.

Thin phase holograms can be recorded on photoresists, which are light-sensitive organic films yielding relief image after exposure and development. They offer advantages of easy replication using thermoplastics but are slow in response and undergo nonlinear effects at diffraction efficiency greater than ~0.05. Shipley AZ-1350 is a widely used photoresist, with maximum sensitivity in the ultraviolet, dropping rapidly for longer wavelengths towards the blue.

Photopolymers are being keenly investigated because they offer advantages such as ease of handling, low cost, and real-time recording for the application of holography and nonlinear optics. However, they have low sensitivity and short shelf life.

Thin phase surface relief holograms can be recorded in a thin layer of thermoplastics. These materials have a reasonably high sensitivity over the whole visible spectrum, fairly high diffraction efficiency, and do not require wet processing. Their application in holographic interferometry, optical information processing, and in making compact holographic devices has been widely reported.

Photorefractive crystals such as $Bi_{12}SiO_{20}$, $LiNbO_3$, $Bi_{12}GeO_{20}$, $BaTiO_3$, $LiTaO_3$, etc., as recording materials, offer high-angular sensitivity and provide capability to read and write volume holographic data in real time. Besides the materials discussed here, several other materials have been investigated for recording holograms; these include photocromics, elastomere devices, magneto-optic materials, etc.

Application of Holography

Holography can be constructed not only with the light waves of lasers, but also with sound waves, microwaves, and other waves in the electromagnetic spectrum of radiation. Holograms made with ultraviolet light or X-rays can record images of objects/particles smaller than the wavelength of visible light, e.g., atoms or molecules. Acoustical holography uses sound waves to see through solid objects. Holography has a vast scope of practical applications, which have been classified into two major categories:

1. applications requiring three-dimensional images for visual perception; and
2. applications in which holography is used as a measuring tool.

The unique ability of holography – to record and reconstruct both electromagnetic and sound waves – makes it a valuable tool for education, science, industry, and business. Below are some of the important applications:

1. Holographic interferometry (HI) is one of the most powerful measuring tools. In HI, two states of an object, i.e., initial and deformed state are recorded on the same photographic plate. After reconstruction of the light wave corresponding to two states of an object, interferences and deformations are displayed in terms of the interference pattern. The change of distance of one tenth of a micron, or lower, can be resolved. HI provides scientists/engineers with crucial data for design of critical machine parts of power-generating equipment, in the aircraft industry, automobile industry, and nuclear installations (say, for example, in the design of containers used to transport nuclear materials, improve the design of aircraft wings and turbine blades, etc.). Presently, HI is being widely used in mechanical engineering, acoustics, and aerodynamics, for nondestructive testing, to investigate oscillation in diaphragms and flow around various objects, respectively.
2. Microwave holography can detect objects deep within spaces, by recording the radio waves they emit.
3. Another important application of holography is the design of optical elements, which possess special properties. A holographic recording of a concave mirror behaves in much the same way as the mirror itself, i.e., it can focus the light. In some cases chromatism can be introduced in the design of elements so that a location of the point, where the beams are focused, depends on the wavelength. This can be achieved by accurately choosing the recording arrangement of the focusing elements, these elements are, in fact, diffraction gratings. These can have low noise levels, freedom from astigmatism, and have other useful properties. Holographic optical elements have found applications in supermarket scanners to read barcodes, headup displays in fighter aircraft to observe critical cockpit instruments, etc.
4. A telephone credit card used in Europe and other developed countries has embossed surface holograms which carry a monetary value. When the card is inserted into the telephone, a card reader discerns the amount due and deducts the appropriate amount to cover the cost of the call.

5. Holography is having applications in analog and digital computers, offering remarkable opportunities to realize various logical operators, devices for identifying images based on matched filtering, and in computer memory units. The basic advantage of holographic memory is that a relatively large volume of information can be stored and that there are a limited number of ways to change the record. The arrival of the first prototype of optical computers, which use holograms as storage material for data, could have a dramatic impact on the holographic market. The yet-to-be-unveiled optical computers will be able to deliver trillions of bits of information faster than the current generation of computers.
6. Optical tweezers are going to become an important tool for the study of micro-organisms/bacteria, etc.
7. Holograms can be used to locate and retrieve information without knowing its location in the storage medium, only needing to know some of its content.
8. The links between computer science and holography are now well-established and are developing, with at least two aspects making computer-generated holograms extremely interesting. First, such holograms enable us to obtain visual 3-dimensional reconstructions of imagined objects. For example, one can reconstruct three dimensions of a model of an object still in the design stage. Second, computer-generated holograms can be used to reconstruct lightwaves with specified wave fronts. This means specially computed and manufactured holograms may function as optical elements that transform the incident light wave into desired wavefronts.
9. Another important applications of holography is its utilization to compensate for the distortion that occurs when viewing objects through optically heterogeneous mediums. It can be achieved based on the principle of beam reversal by the hologram.
10. Finally, let us consider the application of holography to art. The development of holography gives very effective ways of creating qualitative three-dimensional images. Thus, a new independent area of holographic creative work representational/artistic holography has appeared. The art of holographic depiction has developed along two major routes. The first is creation of the view hologram, used as holograms of natural objects that are to be displayed in exhibitions and museums; these are also known as artistic holograms. Portrait holography is also classified under this category. The progress in portrait holography is hampered partly because of imperfection of pulsed lasers and partly because of the deterioration of photographic material when exposed to pulsed electromagnetic radiation.

The principle behind the creation of elusion by using composite holograms is also very convincing for the display of objects. To synthesize the composite holograms, the photographs of various aspects of a scene are printed onto the photographic plate. The synthesis techniques for preparing the composite hologram are very complicated, while the images created by holograms are still far from perfect. However, there is no doubt that composite holography opens holography up as an artistic technique. Recently rainbow holography has been very popular to display such objects.

See also

Holography, Applications: Art Holography. **Holography, Techniques:** Color Holography; Computer-Generated Holograms; Digital Holography; Holographic Interferometry. **Phase Control:** Phase Conjugation and Image Correction.

Further Reading

Benton SA (1969) On a method for reducing the information content of holograms. *Journal of the Optical Society of America* 59: 1545.

Denisyuk YuN (1962) Photographic reconstruction of the optical properties of an object in its own scattered radiation field. *Sov. Phys.-Dokl.* 7: 543.

Denisyuk YuN (1963) On the reproduction of the optical properties of an object by the wave field of its scattered radiation, Pt. I. *Optical Spectroscopy* 15: 279.

Denisyuk YuN (1965) On the reproduction of the optical properties of an object by the wave field of its scattered radiation, Pt II. *Optical Spectroscopy* 18: 152.

Gabor SD (1948) A new microscopic principle. *Nature* 161: 777.

Gabor D (1949) Microscopy by reconstructed wavefronts. *Proceedings of the Royal Society* A197: 454.

Gabor D (1951) Microscopy by reconstructed wavefronts: II. *Proceedings of the Physical Society* B64: 449.

Hariharan P (1983) *Optical Holography: Principle, Techniques and Applications*. Cambridge, UK: Cambridge University Press.

Leith EN and Upatnieks J (1962) Reconstructed wavefronts and communication theory. *Journal Optical of the Society of America* 52: 1123.

Leith EN and Upatnieks J (1963) Wavefront reconstruction with continuous-tone objects. *Journal of the Optical Society of America* 53: 1377.

Leith EN and Upatnieks J (1964) Wavefront reconstruction with diffused illumination and three-dimensional objects. *Journal of the Optical Society of America* 54: 1295.

Pennington KS and Lin LH (1965) Multicolor wavefront reconstruction. *Applied Physics Letters* 7: 56.

Stroke GW and Labeyrie AE (1966) White-light reconstruction of holographic images using the Lippmann-Bragg diffraction effect. *Physics Letters* 20: 368.

Color Holography

H I Bjelkhagen, De Montfort University, Leicester, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since the appearance of the first laser-recorded monochromatic holograms in the 1960s, the possibilities of recording full-color high-quality holograms have now become a reality. Over many years holographic techniques have been created to produce holograms with different colors, referred to as pseudo- or multicolor holograms. For example, in the field of display holography, it is common to make multiple-exposed reflection holograms using a single-wavelength laser and emulsion thickness manipulation to obtain an image with different colors. In this way many beautiful pseudocolor holograms have been made by artists. Applying Benton's rainbow hologram technique (for example, the one used for embossing holograms on credit cards), it has been possible to record transmission holograms as well as mass-produce embossed holograms with 'natural' colors. However, a rainbow hologram has a very limited viewing position from which a correct color image can be seen.

This article will acquaint the reader with the topic of color holography (full-color or true-color holography). This describes a holographic technique to obtain a color 3D image of an object where the color rendition is as close as possible to the color of the real object. A brief review of the development of color holography, the current status and the prospects of this new 3D imaging technique, are presented. Also included are recording techniques using three lasers providing red, green and blue (RGB) laser light as well as computer-generating techniques for producing full parallax 3D color images.

The Development of Color Holography

In theory, the first methods for recording color holograms were established in the early 1960s. Already Leith and Upatnieks (1964) proposed multicolor wavefront reconstruction in one of their first papers on holography. The early methods concerned mainly transmission holograms recorded with three different wavelengths from a laser or lasers, combined with different reference directions to avoid cross-talk. Such a color hologram was then reconstructed (displayed) by using the original laser wavelengths from the corresponding reference directions. However, the complicated and expensive reconstruction setup prevented this technique from becoming popular. More suitable for holographic color recording is reflection holography, which can be reconstructed and viewed in ordinary white light.

Most likely, the three-beam transmission technique might eventually become equally applicable, provided inexpensive multicolor semiconductor lasers become widely available. Such RGB lasers can be used to display a very realistic, deep, and sharp holographic image of, for example, a room.

Relatively few improvements in color holography were made during the 1960s and 1970s. Not until the end of the 1970s did a few examples of color holograms appear. During the 1980s, new and improved techniques were introduced. A review of various transmission and reflection techniques for color holography was published by Hariharan (1983). From 1990 until today, many high-quality color holograms have been recorded, mainly due to the introduction of new and improved panchromatic recording materials. On the market are ultra-fine-grain silver halide emulsions (manufactured by Slavich in Russia) as well as photopolymer materials (manufactured by E. I. du Pont de Nemours & Co. in the USA).

Color reflection holography presents no problems with regard to the geometry of the recording setup, but the final result is highly dependent on the recording material used and the processing techniques applied. There are some problems associated with recording color holograms in general and specifically in silver halide emulsions:

- Scattering occurring in the blue part of the spectrum typical of many holographic coarse-grain silver halide emulsions makes them unsuitable for the recording of color holograms.

- Multiple storage of interference patterns in a single emulsion reduces the diffraction efficiency of each individual recording. The diffraction efficiency of a three-color recording in a single-layer emulsion is lower than a single wavelength recording in the same emulsion.
- During wet processing, emulsion shrinkage frequently occurs, causing a wavelength shift. White-light-illuminated reflection holograms normally show an increased bandwidth upon reconstruction, thus affecting the color rendition.
- The fourth problem, related to some extent to the recording material itself, is the selection of appropriate laser wavelengths and their combination in order to obtain the best possible color rendition of the object.

In the beginning, when no suitable panchromatic emulsion existed, the sandwich technique was used to make color reflection holograms. Two plates were sandwiched together, in which, for example, two different types of recording materials were used. The most successful demonstration of the sandwich recording technique was made by Kubota (1986) in Japan. He used a dichromated gelatin (DCG) plate for the green (515 nm) and the blue (488 nm) components, and an Agfa 8E75 silver halide plate for the red (633 nm) component of the image. Kubota's sandwich color hologram of a Japanese doll recorded in 1986, demonstrated the potential of high-quality color holography for the first time. Color holograms have also been recorded in red-sensitized DCG materials.

Extensive work in the field of reflection color holography was performed by Hubel and Solymar (1991). Primarily they employed Ilford silver halide materials for the recording of color holograms applying the sandwich technique.

Very important is the research on recording materials by Usanov and Shevtsov (1990) in Russia. Their work is based on the formation of a microcavity structure in gelatin, by applying a special processing technique to silver halide emulsions. Such holograms recorded in a silver halide emulsion have a high diffraction efficiency and exhibit a typical DCG hologram quality.

Not until panchromatic ultra-fine-grain silver halide emulsions were introduced in Russia in the early 1990s, was it possible to record high-quality color holograms in a single emulsion layer as demonstrated by the present author. Although it is now possible to produce color holograms, they need to be displayed correctly. An important improvement would be to find an edge-illuminating technique to make color holograms easier to display. The display problem still remains the main obstacle preventing color holograms, as well as monochrome holograms, being more widely used.

Recording of Color Holograms

Currently, most transmission color holograms are of the rainbow type. Large-format holographic stereograms made from color-separated movie or video recordings have been produced. There are also some examples of embossed holograms in which a correct color image is reproduced. In order to generate a true color rainbow hologram, a special setup is required in which the direction of the reference beam can be changed in between the recordings of the color-separated RGB primary images. However, as already mentioned, holographic color images of the rainbow type can reconstruct a correct color image only along a horizontal line in front of the film or plate and, thus, of less interest for serious color holography applications. Reflection holography can offer full parallax, large field of view 3D color images where the colors do not change when observed from different directions.

Silver Halide Materials

To be able to record high-quality color reflection holograms it is necessary to use extremely low light-scattering recording materials. This means, for example, the use of ultra-fine-grain silver halide emulsions (grain size about 10 nm). Currently, the only producer of a commercial holographic panchromatic ultra-fine-grain silver halide material is the Micron branch of the Slavich photographic company located outside Moscow. The PFG-03c emulsion comes in standard sizes from 63 mm × 63 mm format up to 300 mm × 406 mm glass plates and is also available on film. Some characteristics of the Slavich material are presented in Table 1.

By using suitable processing chemistry for the PFG-03c emulsion it has been possible to obtain high-quality color holograms. These holograms are recorded in a single-layer emulsion, which greatly simplifies the recording process as compared with many earlier techniques. The aforesaid single-layer technique is described below.

Table 1 Characteristics of the Slavich panchromatic emulsion

Silver halide material	PFG-03C
Emulsion thickness	7 μm
Grain size	12–20 nm
Resolution	~10,000 lp(mm)$^{-1}$
Blue sensitivity	~1.0–1.5 × 10^{-3} J(cm)$^{-2}$
Green sensitivity	~1.2–1.6 × 10^{-3} J(cm)$^{-2}$
Red sensitivity	~0.8–1.2 × 10^{-3} J(cm)$^{-2}$
Color sensitivity peaked at	633 nm and 530 nm

Photopolymer Materials

The color holography photopolymer material from DuPont is another alternative recording material for color holograms. In particular, this type of material is suitable for mass production of color holograms. Although, being less sensitive than the ultra-fine-grain silver halide emulsion, it has special advantages of easy handling and dry processing (only UV-curing and baking). The DuPont color photopolymer material has a coated film layer thickness of about 20 μm. The photopolymer film is generally coated in a 12.5-inch width on a 14-inch wide Mylar® polyester base which is 0.002-inch thick. The film is protected with a 0.00092-inch thick Mylar® polyester cover sheet. However, the panchromatic polymer material is only supplied to specially selected and approved hologram producers. The reason for this is that some applications of color holograms are in the field of document security where optically variable devices (OVDs) are produced.

The recording of a color hologram on DuPont polymer is simple. The film has to be laminated to a piece of clean glass or attached to a glass plate using an index-matching liquid. To obtain the right color balance, the RGB sensitivity depends on the particular material, but typically red sensitivity is lower than green and blue sensitivities. It is difficult to obtain high red-sensitivity of photopolymer materials. Simultaneous exposure is the best recording technique for photopolymer materials. Holograms can be recorded manually, but in order to produce large quantities of holograms, a special machine is required. For hologram replication the scanning technique can provide the highest production rate. In this case, three scanning laser lines are needed, which can be adjusted in such a way that all three simultaneously can scan the film. The color photopolymer material needs an overall exposure of about 10 $mJ(cm)^{-2}$.

After the exposure is finished, the film has to be exposed to strong white or UV light. DuPont recommends about 100 $mJ(cm)^{-2}$ exposure at 350–380 nm. After that, the hologram is put in an oven at a temperature of 120 °C for two hours in order to increase the brightness of the image.

Laser Wavelengths for Color Holograms

The problem of choosing optimal primary laser wavelengths for color holography is illustrated in the 1976 CIE chromaticity diagram (Figure 1) which indicates suitable laser wavelengths for color holograms.

It may seem that the main aim of choosing the recording wavelengths for color holograms would be to cover as large an area of the chromaticity diagram as possible. However, there are many other

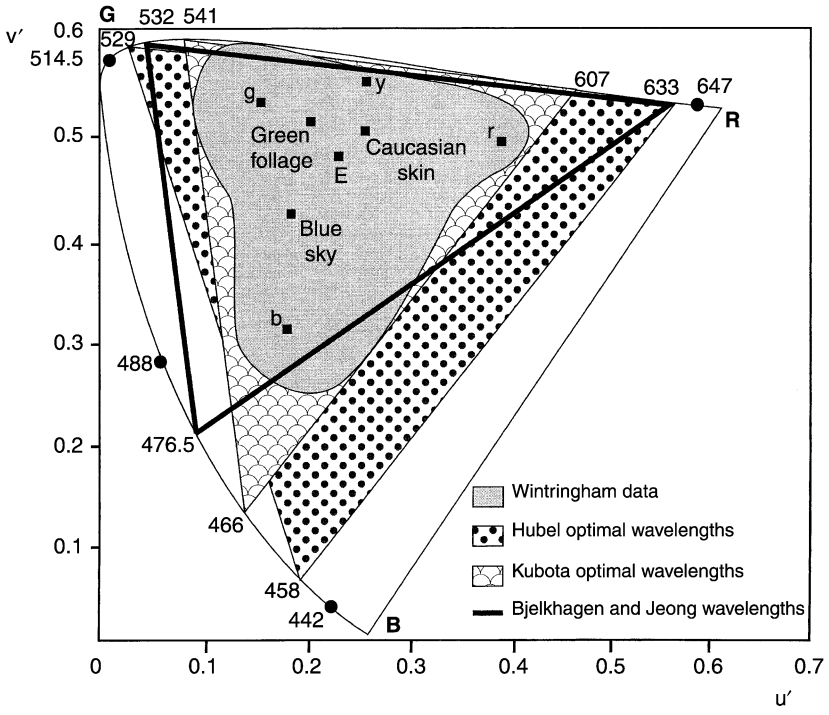


Figure 1 The 1976 CIE uniform scales chromaticity diagram shows the gamut of surface colors and positions of common laser wavelengths. Optimal color-recording laser wavelengths are also indicated.

considerations that must be taken into account when choosing the wavelengths for color holograms. One of these important considerations is the question of whether three wavelengths are really sufficient for color holography. The wavelength selection problem for color holography has been treated in several papers, for example by Peercy and Hesselink (1994).

Hubel and Solymar (1991) provided a definition of color recording in holography:

> A holographic technique is said to reproduce 'true' colors if the average vector length of a standard set of colored surfaces is less than 0.015 chromaticity coordinate units, and the gamut area obtained by these surfaces is within 40% of the reference gamut. Average vector length and gamut area should both be computed using a suitable white light standard reference illuminant.

An important factor to bear in mind when working with silver halide materials is that a slightly longer blue wavelength than the optimal one might give higher-quality holograms (better signal-to-noise ratio) because of reduced Rayleigh scattering during the recording. Another important factor to consider is the reflectivity of the object at primary spectral wavelengths. It has been shown that the reflectivity of an object at three wavelength bands, peaked at 450, 540, and 610 nm, has an important bearing on color reconstruction. These wavelengths can also be considered optimal for the recording of color holograms even though the laser lines are very narrow-band. A triangle larger than necessary can be considered in order to compensate for color desaturation (color shifting towards white) that takes place when reconstructing color reflection holograms in white light. According to Hubel's color rendering analysis, the optimal wavelengths are 464, 527, and 606 nm for the sandwich silver halide recording technique. If the calculations are performed to maximize the gamut area instead, the following set of wavelengths is obtained: 456, 532, and 624 nm. However, when approached from a different viewpoint, the optimal trio of wavelengths based on the reconstructing light source of 3400 K, a 6 μm thick emulsion with a refractive index of 1.63 and an angle of 30° between the object and the reference beam the following wavelengths were obtained: 466.0, 540.9, and 606.6 nm. Peercy and Hesselink (1994) discussed wavelength selection by investigating the sampling nature of the holographic process. During the recording of a color hologram the chosen wavelengths point-sample the surface-reflectance functions of the object. This sampling on color perception can be investigated by the tristimulus value of points in the reconstructed hologram which is mathematically equivalent to integral approximations for the tristimulus integrals. Peercy and Hesselink used both Gaussian quadrature and Riemann summation for the approximation of the tristimulus integrals. In the first case they found the wavelengths to be 437, 547, and 665 nm. In the second case the wavelengths were 475, 550, and 625 nm. According to Peercy and Hesselink, the sampling approach indicates that three monochromatic sources are almost always insufficient to preserve all of the object's spectral information accurately. They claim that four or five laser wavelengths are required.

Only further experiments will show how many wavelengths are necessary and which combination is the best for practical purposes. Another factor that may influence the choice of the recording wavelengths is the availability of cw lasers currently in use in holographic recording, for example, argon ion, krypton ion, diode-pumped solid state (DPSS) frequency-doubled Nd:YAG, helium neon, and helium cadmium lasers (Table 2).

The recent progress in DPSS laser technology has made available both red and blue DPSS lasers. These lasers are air-cooled, small, and each laser requires less than a hundred watts of electric power to operate. Usually, a set of three DPSS lasers will be the best choice of cw lasers for color holography in the future.

Setup for Recording Color Holograms

A typical reflection hologram recording setup is illustrated in Figure 2.

Table 2 Wavelengths from cw lasers

Wavelength [nm]	*Laser type*	*Single line power [mW]*
442	Helium cadmium	<100
457	DPSS blue	<500
458	Argon ion	<500
468	Krypton ion	<250
476	Krypton ion	<500
477	Argon ion	<500
488	Argon ion	<2000
497	Argon ion	<500
502	Argon ion	<400
514	Argon ion	<5000
521	Krypton ion	<100
529	Argon ion	<600
531	Krypton ion	<250
532	DPSS green	<3000
543	Green neon	<10
568	Krypton ion	<100
633	Helium neon	<80
647	Krypton ion	<2000
656	DPSS red	<1000

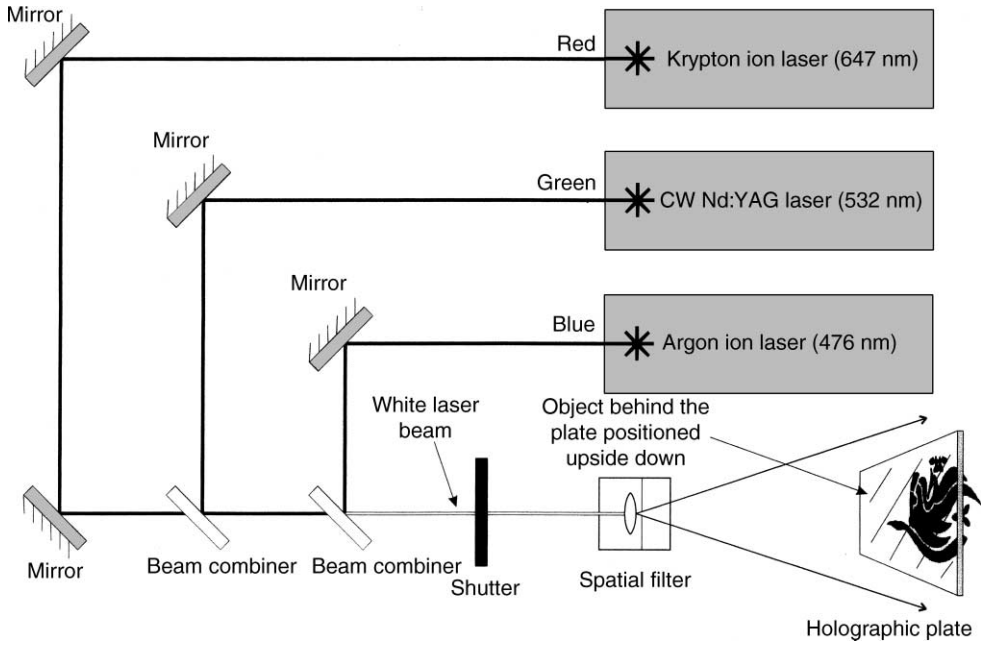


Figure 2 The setup for recording color reflection holograms.

The different laser beams necessary for the exposure of the object pass through the same beam expander and spatial filter. A single-beam Denisyuk arrangement is used, i.e., the object is illuminated through a recording holographic plate. The light reflected from the object constitutes the object beam of the hologram. The reference beam is formed by the three expanded laser beams. This 'white' laser beam illuminates both the holographic plate and the object itself through the plate. Each of the three primary laser wavelengths forms its individual interference pattern in the emulsion, all of which are recorded simultaneously during the exposure. In this way, three holographic images (a red, a green, and a blue image) are superimposed upon one another in the emulsion.

Three laser wavelengths are employed for the recording: 476 nm, provided by an argon ion laser, 532 nm, provided by a cw frequency-doubled Nd:YAG laser, and 647 nm, provided by a krypton laser. Two dichroic filters are used for the combining of the three laser beams. The 'white' laser beam goes through a spatial filter, illuminating the object through the holographic plate.

By using the dichroic filter beam combination technique, it is possible to perform simultaneous exposure recording, which makes it possible to control independently the RGB ratio and the overall exposure energy in the emulsion. The RGB ratio can be varied by individually changing the output power of the lasers, while the overall exposure energy is controlled solely by the exposure time. The overall energy density for exposure is about 3 $mJ(cm)^{-2}$.

In the initial experiments performed by the author in 1993, a specially designed test object consisting of the 1931 CIE chromaticity diagram, a rainbow ribbon cable, pure yellow dots, and a cloisonné elephant, was used for the color balance adjustments and exposure tests. Later, another test target was employed – the Macbeth ColorChecker® chart, which was used for color rendering tests.

Color reflection holograms can also be produced using copying techniques so that projected real images can be obtained, however, normally associated with a restricted field of view. For many display purposes, the very large field of view obtainable in a Denisyuk color hologram is often more attractive.

Processing of Color Holograms

The dry processing of color holograms recorded on photopolymer materials has already been described. The process is simple and very suitable for machine processing using, for example, a baking scroll oven. The processing of silver halide emulsions is more difficult and critical. The Slavich emulsion is rather soft, and it is important to harden the emulsion before the development and bleaching takes place. Emulsion shrinkage and other emulsion distortions caused by the active solutions used for the processing must be avoided. In particular, when recording master color holograms intended for photopolymer replication, shrinkage control is extremely important. The processing steps are summarized in Table 3.

Table 3 Color holography processing steps

1.	Tanning in a formaldehyde solution	6 min
2.	Short rinse	5 s
3.	Development in the CWC2 developer	3 min
4.	Wash	5 min
5.	Bleaching in the PBU-amidol bleach	~5 min
6.	Wash	10 min
7.	Soaking in acetic acid bath (printout prevention)	1 min
8.	Short rinse	1 min
9.	Washing in distilled water with wetting agent added	1 min
10.	Air drying	

It is very important to employ a suitable bleach bath to convert the developed silver hologram into a phase hologram. The bleach must create an almost stain-free clear emulsion so as not to affect the color image. In addition, no emulsion shrinkage can be permitted, as it would change the colors of the image. Washing and drying must also be done so that no shrinkage occurs. Finally, to prevent any potential emulsion thickness variation by humidity variations, the emulsion needs to be protected by a glass plate sealed onto the hologram plate.

Evaluation of Color Holograms

Recorded color holograms of the two test targets are presented here. The illuminating spotlight to reconstruct the recorded color holograms was a 12-Volt 50-Watt halogen lamp. This type of spotlight is suitable for displaying color holograms. The selection of a suitable lamp for the reconstruction of color holograms is much more important than the selection of lamps for monochrome hologram display. The color balance for the recording of a color hologram must be adjusted to the type of spotlight that is going to be used for the display of the finished hologram. Figure 3 shows a typical normalized spectrum obtained from a white area of the color test target hologram.

This means that the diffraction efficiency of each color component is obtained assuming a flat spectrum of the illuminating source. One should note the high diffraction efficiency in blue, needed to compensate for the low blue light emission of the halogen spotlight. The noise level, mainly in the blue part of the spectrum, is visible and low. The three peaks are exactly at the recording wavelengths; i.e., 647, 532, and 476 nm.

In Table 4 some results of the Macbeth ColorChecker® investigation are presented. The 1931 CIE x and y coordinates are measured at both the actual target and the holographic image of the target. The measured fields are indicated in the table by color and the corresponding field number.

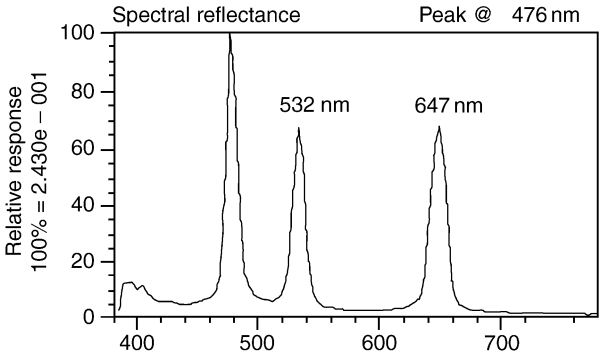


Figure 3 Normalized spectrum from a white area of a color test target hologram.

Color reproductions of two color holograms by the author are presented. A photograph of the 100 mm by 120 mm hologram of the CIE test object is shown in Figure 4. Featured in Figure 5 is a 200 mm by 250 mm hologram of a large Russian egg.

Computer-Generated Color Holograms

Today it is not possible to obtain a computer-generated hologram (CGH) with the same high information content as the laser-recorded ones of real objects. What may become possible in the future is a technique to compute and record the interference pattern that is stored in a Denisyuk color hologram which, upon illumination, can produce an image like the laser-recorded ones of real objects. Therefore, the best compromise is to use holographic stereograms which can provide high-quality computer-generated images. A holographic stereogram is created by using a series of 2D photographic images or 2D images displayed on a LCD screen from which the hologram is recorded. In order to obtain a high-quality holographic image with a large field of view, many 2D images are needed. As mentioned earlier, some color CGHs of the rainbow transmission type have been produced. However, image color changes as a function of viewer position in that type of hologram, making them less attractive than computer-generated reflection holograms. Over the last few years there has been rapid progress in color CGHs of the reflection type. Remarkable results have been achieved by Klug *et al.* (1997) at Zebra Imaging Inc. A new technique, to record very large full-parallax color reflection CGHs, has been developed. Color holograms can be produced, having both vertical and horizontal parallax, with a 100° field of view. The generation of a holographic hardcopy of either digitized images or computer graphics models is based on the following technique. The 'object'

Table 4 Chromaticity coordinates from color hologram recording tests using the Macbeth ColorChecker®

Object	*White #19*	*Blue #13*	*Green #14*	*Red #15*	*Yellow #16*	*Magenta #17*	*Cyan #18*
CIE *x y*	*x*/*y*	*x*/*y*	*x*/*y*	*x*/*y*	*x*/*y*	*x*/*y*	*x*/*y*
Target	0.435/0.405	0.295/0.260	0.389/0.514	0.615/0.335	0.517/0.450	0.524/0.322	0.285/0.380
Image	0.354/0.419	0.335/0.362	0.337/0.449	0.476/0.357	0.416/0.437	0.448/0.338	0.295/0.366

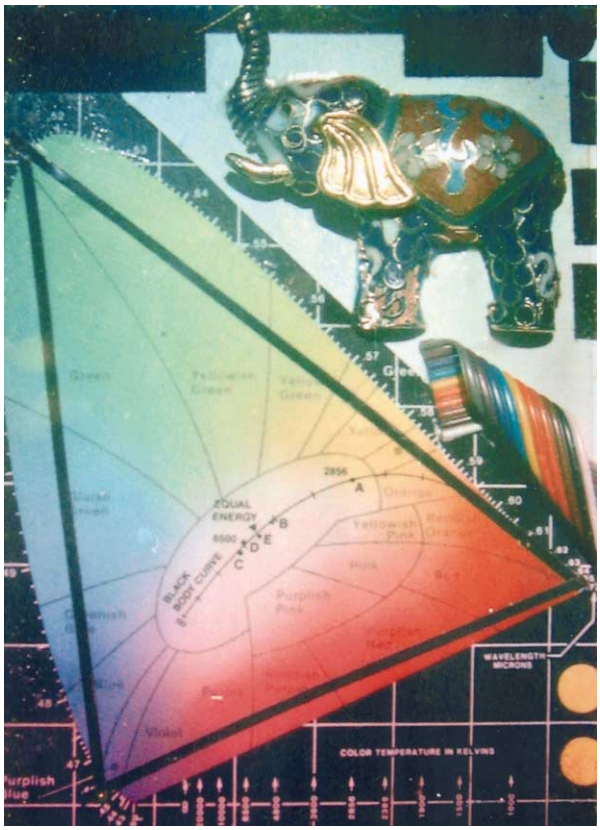


Figure 4 Hologram of the CIE test object recorded with 476, 532, and 647 nm laser wavelengths. Notice the rainbow ribbon cable, the pure yellow dots, the full range of colors, and a balanced white in the center of the CIE color test target.



Figure 5 Color hologram of a large Russian egg, size 100 mm by 120 mm.

sub-beam is directed through a sequence of digital images on a liquid-crystal screen. Each resulting exposure, about two millimeters square, is called a 'hogel'. The full-color hogels are the holographic building blocks of a finished CGH image. In an automated step-and-repeat process, 9×10^4 hogels are formed on a flat square tile of DuPont panchromatic photopolymer film by simultaneous RGB laser exposures. The 60 cm by 60 cm tile itself is the finished hologram, or to obtain larger holograms, the 3D image is made in 60 cm by 60 cm tiled increments. So far, the largest image created was of Ford's P2000 Prodigy concept car in which ten such holograms tiled together made up the large color reflection hologram (1.2 m by 3 m) which is reproduced in Figure 6.

This technique has opened the door to real 3D computer graphics. However, generating such a large hologram is a very time-consuming process. The Ford P2000 hologram took almost two weeks (300 hours) to produce, since each individual panel requires a 24-hour recording time.

The Future of Color Holography

The manufacturing of large-format color reflection holograms is now possible. Mass production of color holograms on photopolymer materials has started in Japan. Although good color rendition can be obtained, some problems remain to be solved, such

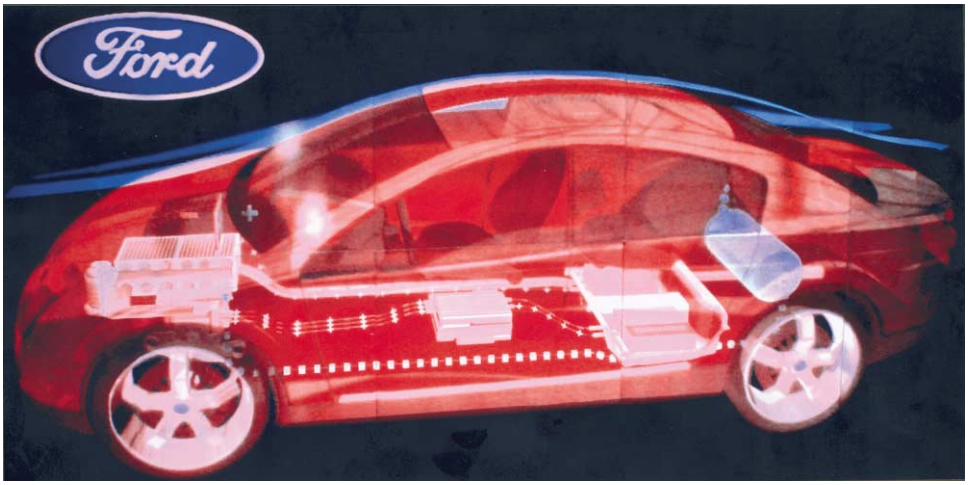


Figure 6 Computer-generated color hologram of Ford's P2000 Prodigy concept car. A ten tile full parallax Mosaic hologram produced by Zebra Imaging, Inc., size 1.2 m by 3 m.

as desaturation, image resolution, signal-to-noise ratio and dynamic range. Other limitations concerning holographic color recording include the fact that some colors we see are the result of fluorescence, which cannot be recorded in a hologram. There are some differences in the recorded colors in one of the test charts shown here. However, color rendition is a very subjective matter. Different ways of rendition may be preferable for different applications, and different people may have different color preferences.

At the present time, research on processing color reflection holograms recorded on ultra-fine-grain materials is still in progress. Work is carried out on microcavity structure holograms in order to produce low-noise reflection holograms with high diffraction efficiency. For producing large quantities of color holograms, the holographic photopolymer materials are most suitable and will be further improved in the future.

Another important field of research is the development of a three-wavelength pulsed laser. Employing such a laser, dynamic events, as well as portraits, can be recorded in a holographic form. Currently, research work is in progress at the French German Research Institute ISL, Saint Louis, France, as well as at the Geola company in Lithuania.

The virtual color image behind a color holographic plate represents the most realistic image of an object that can be obtained today. The extensive field of view adds to the illusion of beholding a real object rather than an image of it. The wavefront reconstruction process recreates accurately the three-wavelength light scattered off the object during the recording of the color hologram. This 3D imaging technique has many obvious applications, in particular, in displaying unique and expensive artifacts. There are also many potential commercial applications of this new feature of holographic imaging, provided that the display problem can be solved using some sort of integrated lighting.

Today, it is technologically possible to record and replay acoustical waves with very high fidelity. Hopefully, holographic techniques will be able to offer the same possibility in the field of optical waves, wavefront storage, and reconstruction. Further development of computer-generated holographic images will make it possible to display extremely realistic full-parallax 3D color images of nonexisting objects, which could be applied in various spheres, for example, in product prototyping, as well as in other applications in 3D visualization and three-dimensional art. Eventually it will become possible to generate true-holographic 3D color images in real-time, as computers become faster and possess greater storage capacity. However, the most important issue is the development of extremely high-resolution electronic display devices which are necessary for electronic holographic real-time imaging.

See also

Holography, Techniques: Computer-Generated Holograms.

Further Reading

Bjelkhagen HI (1993) *Silver Halide Recording Materials for Holography and Their Processing*. Springer Series in Optical Sciences, vol. 66. New York: Springer-Verlag.

Bjelkhagen HI, Jeong TH and Vukičević D (1996) Color reflection holograms recorded in a panchromatic ultrahigh-resolution single-layer silver halide emulsion. *Journal of Imaging Science & Technology* 40: 134–146.

Bjelkhagen HI, Huang Q and Jeong TH (1998) Progress in color reflection holography. In: Jeong TH (ed.) *Sixth International Symposium on Display Holography*, Proc. SPIE, vol. 3358, pp. 104–113. Bellingham, WA: The International Society for Optical Engineering.

Hariharan P (1983) Colour holography. In: Wolf E (ed.) *Progress in Optics*, vol. 20, pp. 263–324. Amsterdam: North-Holland.

Hubel PM and Solymar L (1991) Color reflection holography: theory and experiment. *Applied Optics* 30: 4190–4203.

Klug MA, Klein A, Plesniak W, Kropp A and Chen B (1997) Optics for full parallax holographic stereograms. In: Benton SA and Trout TJ (eds) *Practical Holography XI and Holographic Materials III*, Proc. SPIE, vol. 3011, pp. 78–88. Bellingham, WA: The International Society for Optical Engineering.

Kubota T (1986) Recording of high quality color holograms. *Applied Optics* 25: 4141–4145.

Leith EN and Upatnieks J (1964) Wavefront reconstruction with diffused illumination and three-dimensional objects. *Journal of the Optical Society of America* 54: 1295–1301.

Peercy MS and Hesselink L (1994) Wavelength selection for true-color holography. *Applied Optics* 33: 6811–6817.

Usanov YuE and Shevtsov MK (1990) Principles of fabricating micropore silver-halide-gelatin holograms. *Optics and Spectroscopy (USSR)* 69: 112–114.

Watanabe M, Matsuyama T, Kodama D and Hotta T (1999) Mass-produced color graphic arts holograms. In: Benton SA (ed.) *Practical Holography XIII*, Proc. SPIE, vol. 3637, pp. 204–212. Bellingham, WA: The International Society for Optical Engineering.

Computer-Generated Holograms

W J Dallas, University of Arizona, Tucson, AZ, USA

A W Lohmann, University of Erlangen-Nürnberg, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A hologram is a tool for shaping the amplitude and the phase of a light wave. Shaping the amplitude distribution is easy, one way is by a photographic plate, another way is by binary pixels as used in 'halftone printing'. The phase distribution can be shaped by RE-fraction on a thin film of variable thickness, or by DIF-fraction on a diffraction grating that has been deliberately distorted.

Both kinds of holograms can be manufactured based on a computer design, the refractive version being called a 'kinoform', and the diffractive version being known as a computer-generated hologram (CGH). We will explain the fundamentals of computer holography with emphasis on the 'Fourier–CGH', and will also highlight the large variety of applications.

From the Classical Hologram to the Computer-Generated Hologram: CGH

Holography is a method to form images, consisting of two steps: recording and reconstruction. In the recording step, some interference fringes are recorded on a photographic plate (Figure 1a). The two interfering waves originate from the object and from a reference light source. After being developed by the usual photo-chemical treatment the photographic plate is called a hologram. In the reconstruction step, the hologram acts as a diffracting object that is illuminated by a replica of the reference light. Due to diffraction, the light behind the hologram is split into three parts (Figure 1b). One part proceeds to the observer who perceives a virtual object where the genuine object had been originally during the recording step.

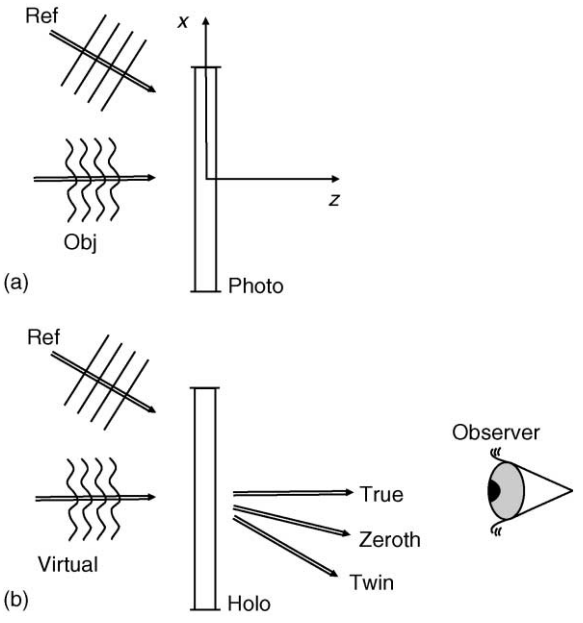


Figure 1 Holographic image formation in two steps: (a) recording the hologram, (b) reconstructing the object wave.

This was a brief description of 'classical-holography'. In 'computer holography' the first step, recording, is synthetic. In other words, the propagation of the light from the object to the photographic plate is simulated digitally. The simulation includes the addition of the object wave to the reference wave and the subsequent modulus square process, which describes the transition from the two complex amplitudes to the hologram irradiance:

$$|u_0(x) + u_R(x)|^2 = I_H(x) \qquad [1]$$

The amplitude transmittance of the hologram is

$$T_H(x) = c_0 + c_1 I_H(x) = c_1 u_0 u_R^* + \cdots \qquad [2]$$

The coefficients c_0 and c_1 represent the photochemical development. In the reconstruction process (Figure 1b) the transmittance T_H is illuminated by the reference beam. Hence, we get

$$u_R(x) T_H(x) = c_1 u_0(x) |u_R(x)|^2 + \cdots \qquad [3]$$

If the reference intensity $|u_R(x)|^2$ is constant, this term represents a reconstruction of the object wave $u_0(x)$. The omitted parts in eqns [2] and [3] describe those beams that are ignored by the observer (Figure 1b). The hologram transmittance $T_H(x)$ is an analog signal:

$$0 \le T_H \le 1 \qquad [4]$$

Before the 1960s, when computer holograms were invented, the available plotters could produce only binary transmittance patterns. Hence, it was necessary to replace the analog transmission $T_H(x)$ by a binary transmission $B_H(x)$, which is possible without loss of image quality. In addition, binary computer holograms have better light efficiency and a better signal-to-noise ratio. They are also more robust when being copied or printed.

Another advantage of a CGH is that the object does not have to exist in reality, which is important for some applications. The genuine object may be difficult to manufacture or, if it is three-dimensional, difficult to illuminate. The CGH may be used to implement an algorithm for optical image processing. In that case, the term 'object' becomes somewhat fictitious. We will discuss more about applications, towards the end of this article.

From a Diffraction Grating to a Fourier CGH

Now we will explain, in some detail, the 'Fourier type' CGH. Fourier holograms are more popular than two other types: 'image-plane' CGHs and 'Fresnel-type' CGHs. The Fresnel-type will be treated in the section on 'software' because of some interest in 3-D holograms. Certain hardware issues will also be described.

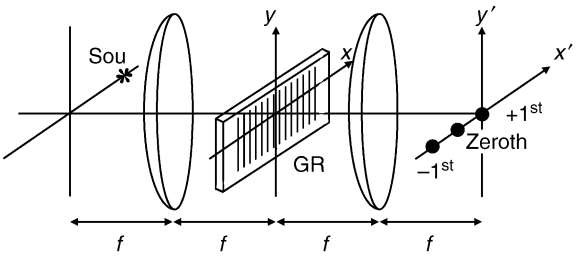


Figure 2 Grating diffraction, but with an off-axis source.

The explanation of Fourier CGHs begins with grating diffraction (Figure 2). The only uncommon feature in Figure 2 is the off-axis location of the source. It is arranged such that the plus-first diffraction order will hit the center of the output plane.

We will now convert a simple Ronchi grating (Figure 3a) into a Fourier CGH. The Ronchi grating transmittance can be expressed as the Fourier series:

$$G(x) = \sum_{(m)} C_m \exp\left(\frac{2\pi i m x}{D}\right); \quad C_0 = \frac{1}{2};$$

$$C_m = \frac{\sin(\pi m/2)}{\pi m} \qquad [5]$$

Now we shift the grating bars by an amount PD. Then we modify the width from $D/2$ to WD, as shown in Figure 3b. The Fourier coefficients are now

$$C_m = \frac{\exp(2\pi i m P)\sin(\pi m W)}{\pi m} \qquad [6]$$

The C_{+1} coefficient is responsible for the complex amplitude at the center of the readout plane:

$$C_{+1} = \frac{\exp(2\pi i P)\sin(\pi W)}{\pi} \qquad [7]$$

We are able now to control the beam magnitude, $A = (\frac{1}{\pi})\sin(\pi W)$, by the relative bar width W and the phase $\phi = 2\pi P$ by the relative bar shift P. This particular kind of phase shift is known as 'detour phase'.

So far, the light that ends up at the center of the exit plane, leaves the grating as a plane wave with a wave vector parallel to the optical axis. Now we distort the grating on purpose (Figure 3c):

$$W \to W(x, y); \qquad P \to P(x, y) \qquad [8]$$

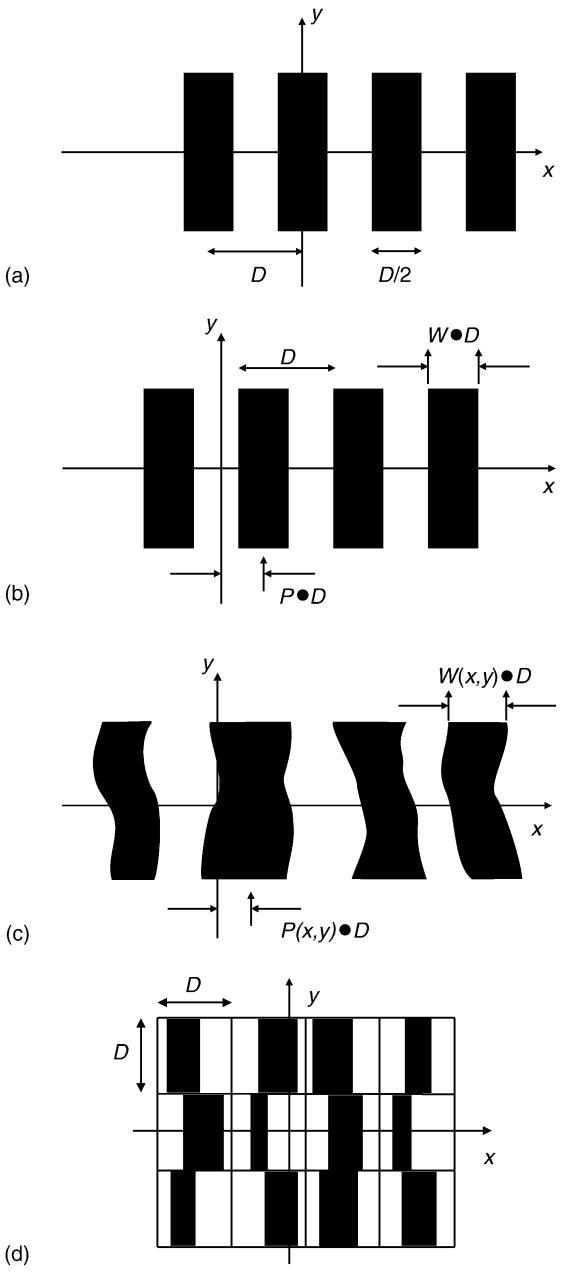


Figure 3 Modifying a grating into a hologram in 3 steps: (a) Ronchi grating, (b) Modified width and location of the grating bars, (c) Distorted grating bars, (d) as in **Figure 3c**, but now discrete distortions.

These distortions should be mild enough that the plus-first order light behind the distorted grating can be described by the complex amplitude:

$$\left(\frac{1}{\pi}\right)\sin[\pi W(x,y)]\,\exp[2\pi i P(x,y)] = u_{\mathrm{H}}(x,y) \qquad [9]$$

This is a generalization of eqn [7]. The phase $2\pi P(x,y)$ covers the range from $-\pi$ to $+\pi/2$ if the shift is bounded by $|P| \le \frac{1}{2}$. That is enough to cover the range of complex amplitudes within the circle of $|u_{\mathrm{H}}| \le \frac{1}{\pi}$. We will return shortly to the condition 'mild-enough distortions' in the context of **Figure 4**.

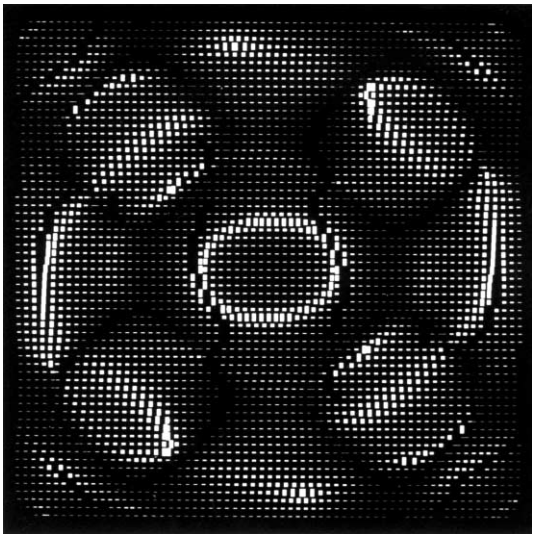


Figure 4 CGH with 64 × 64 cells.

Most plotters move their pens only in the x-direction and y-direction and printers place dots on a grid. Therefore it is convenient to replace the continuous (x, y) variations of W and P by piecewise constant variations, as in **Figure 3d**. That restriction turns out to be acceptable, as demonstrated by the first computer holograms. The theoretical proof involves the sampling theorem.

Recall that the light propagation through a 2-f system (**Figure 2**, between grating and output plane) can be described by a Fourier transformation:

$$\int\int u_{\mathrm{H}}(x,y)\,\exp\left[\frac{-2\pi i(xx' + yy')}{\lambda f}\right]\mathrm{d}x\,\mathrm{d}y = \tilde{u}_{\mathrm{H}}\left(\frac{x'}{\lambda f}, \frac{y'}{\lambda f}\right) \qquad [10]$$

Suppose now that we wish to see a particular image: $v(x', y')$. Hence, we request that

$$v(x', y') = \tilde{u}_{\mathrm{H}}\left(\frac{x'}{\lambda f}, \frac{y'}{\lambda f}\right)\mathrm{RECT}\left(\frac{x'}{\lambda f}\right) \qquad [11]$$

The RECT function indicates that only the PLUS first diffraction order is of interest. As a consequence we have to select the grating distortions $W(x, y)$ and $P(x, y)$ such that the hologram $u_{\mathrm{H}}(x, y)$ is the

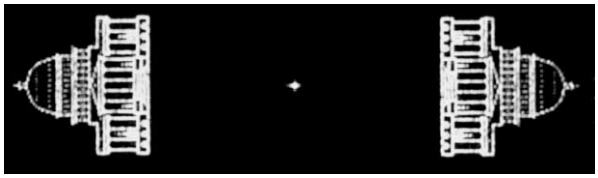


Figure 5 A reconstruction from a 1024 × 1024 CGH. The real image and the symmetrical twin image appear. The zeroth-order is blocked.

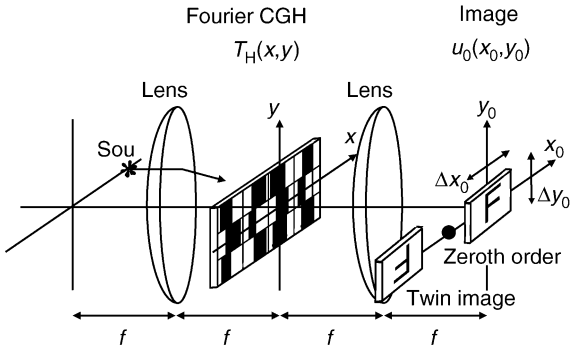


Figure 6 Reconstruction setup for a Fourier CGH.

Fourier transform of $v(x', y')$:

$$u_H(x,y) = \int\int v(x',y') \exp\left[\frac{2\pi i(xx'+yy')}{\lambda f}\right] dx'\,dy' \qquad [12]$$

The condition 'mild enough distortions' has been treated in the literature. There is not enough space here to present that part of the CGH theory. But we show an actual CGH with 64 × 64 cells (Figure 4). This CGH is no longer a grating, but, it shows enough resemblance to our heuristic derivation. Hence, the CGH theory appears to be trustworthy. A CGH looks less regular if the image $v(x', y')$ contains a random phase that simulates a diffuser. The diffuser spreads light across the CGH and so levels out unacceptably large fluctuations in the amplitude.

With 512 × 512 or 1024 × 1024 cells in a CGH one can obtain an image of the quality seen in Figure 5. The reconstruction setup is shown in Figure 6, here with the letter F as the image.

About Some CGH Algorithms

The computational effort for getting the desired hologram amplitude of a Fourier hologram u_H is reasonable due to the powerful FFT algorithm. However, to use it the data must be discrete. In other words, the CGH consists of cells, centered at $x_m = mD$, $y_n = nD$ as shown in Figure 3d. The questions now are: 'What is the proper cell size D?' and 'How many cells are needed?' The answers to those questions depend on the parameters of the image. If the zero order (see Figure 6) should be at the distance Δx_0, then the quasi-grating period D should obey:

$$\frac{\lambda f}{D} \geq \Delta x_0 \qquad [13]$$

And if the size of an image pixel should be as fine as δx_0, the size Δx_H of the hologram should obey:

$$\frac{\lambda f}{\Delta x_H} \leq \delta x_0 \qquad [14]$$

This condition is plausible because the finite size Δx_H of the hologram acts like a resolution-limiting aperture. The combination of eqns [13] and [14] yields

$$\frac{\Delta x_0}{\delta x_0} \leq \frac{\Delta x_H}{D} \qquad [15]$$

This means that the number of image pixels $\Delta x_0/\delta x_0$ is bounded by the number of CGH cells $\Delta x_H/D$. The generalization to two dimensions is straightforward. Again, Fourier CGHs benefit from the fact that the light propagation from the virtual object to the hologram plane can be described by a Fourier transformation. Hence, the FFT can be used.

And now we move to the synthesis of a Fresnel hologram. A Fresnel hologram is recorded at a finite distance z from the object. Hence, we have to simulate the wave propagation in free space from the object (at $z = 0$) to the hologram at a distance z. This free space propagation can be described (in the paraxial approximation) as a convolution of object and quasi-spherical wave:

$$u_0(x) \rightarrow \int_{-\infty}^{\infty} u_0(x') \exp\left[\frac{-i\pi(x-x')^2}{\lambda z}\right] dx' = u(x,z) \qquad [16]$$

In the Fourier domain the corresponding operation is:

$$\tilde{u}_0(\mu) \rightarrow \tilde{u}_0(\mu) \exp[-i\pi\lambda z\mu^2] = \tilde{u}(\mu,z) \qquad [17]$$

The indirect execution of eqn [16], based on eqn [17], is easy: a Fourier transformation of the object converts $u_0(x)$ into $\tilde{u}_0(\mu)$, which is then multiplied by the quadratic phase factor, yielding $\tilde{u}(\mu,z)$. An inverse Fourier transformation produces $u(x,z)$. The two Fourier transformations consist typically of $4N \log N$ multiply/add operations. N is the number of pixels: image size Δx_0, divided by the pixel size δx_0.

In the case of a rectangular object $u_0(x, y)$, the number of pixels is $\Delta x_0 \Delta y_0 / \delta x_0 \delta y_0$.

Now to describe the algorithmic effort for computing the convolution, the $u_0(x)$ has N pixels, and, the lateral size of the exponential is essentially $M = z/\delta z$, where δz represents 'depth of focus'. M describes the lateral size of the diffraction cone, measured in pixel size units δx_0, and the convolution involves MN multiply/adds. A comparison of these two algorithms is dictated by the ratio

$$\frac{M}{4 \log N} \qquad [18]$$

The direct convolution is preferable if the distance z is fairly small. The integration limits of eqn [16] depend on the oscillating phase $\pi x^2/\lambda z$. Effectively, nothing is contributed to this integral, when this phase varies by more than 2π over the length, δx_0, of an object pixel. Note that the pixel size of $u(x, z)$ does not change with the distance z, because the bandwidth $\Delta\mu(z) = \Delta\mu_0$ is z-independent. This is so because the power spectrum is z-invariant:

$$|\tilde{u}(\mu, z)|^2 = |\tilde{u}_0(\mu)|^2 \qquad [19]$$

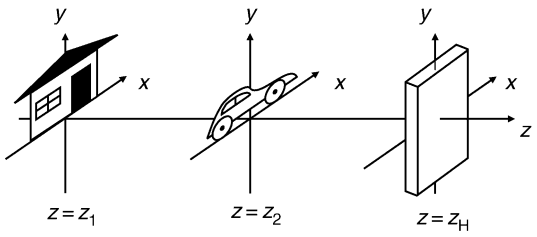


Figure 7 Synthesis of a hologram for a three-dimensional object.

It is easy to proceed from here to the CGH synthesis of a 3-D object. One starts with an object detail $u_1(x)$ at z_1, for example, the house as in Figure 7. Then, one propagates to the next object detail at z_2. This second detail may act as a multiplication (transmission) and as addition (source elements). This process is repeated as the wave moves towards the plane of the hologram. There, a reference wave has to be added, as in the Fourier case.

A very powerful algorithm for the CGH synthesis is 'IFTA' (iterative Fourier transform algorithm). Suppose, we want to design a Fourier CGH, which looks like a person and whose reconstructed image shows the signature of that person (Figure 8a,b). In other words, the two intensities

$$|u_H(x, y)|^2 \text{ and } |\tilde{u}_H(\mu, \nu)|^2 \qquad [20]$$

are determined. But, the phase distributions of u_H and $\tilde{u}_H$ are still at our disposal. The IFTA algorithm will yield those phases in many cases at the expense of ten, hundred or more Fourier transformations.

The IFTA algorithm is sometimes called 'Fourier Ping-Pong' because it bounces back and forth between the (x, y) and the (μ, ν) domain. The amplitudes $|u_H|$ and $|u_0|$ are enforced at every opportunity. But the associated phases are free to vary and to converge, eventually. If so, u_H and u_0 are completely known. The IFTA will not always converge, for example not if

$$|u_H(x, y)| = \delta(x, y) \text{ and } |\tilde{u}_H(\mu, \nu)| = \delta(\mu, \nu) \qquad [21]$$

But the chances are fairly good if both amplitudes are often close to one and seldom near zero.

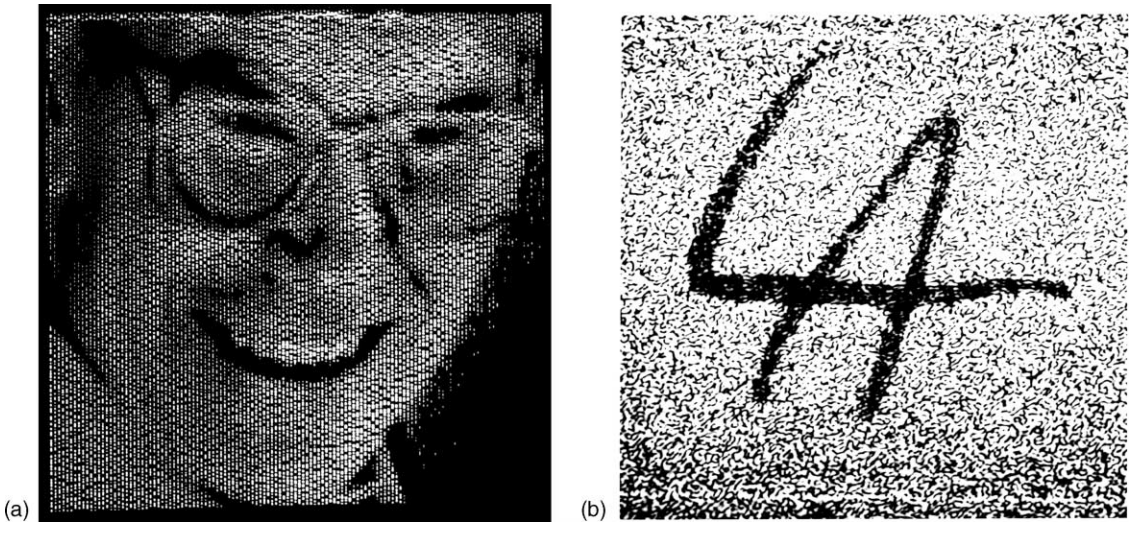


Figure 8 IFTA-CGH: (a) The Fourier CGH. (b) The optical reconstruction thereof. (Courtesy of H.O. Bartelt.)

Finally, an image plane CGH is trivial from an algorithmic point of view. The simulated propagation from object plane to image plane is usually an identity, apart from the finite bandwidth of the image-forming system. Nevertheless, image plane holograms can be quite useful, for example for interferometric testing of a nonspherical mirror of an astronomical telescope. The CGH is laid out to provide a wavefront that is suitable for an interferometric 'null test'. The CGH serves as a 'synthetic prototype'.

A fourth kind of CGH, to be named near-field CGH may emerge. It operates in the near-field and it employs evanescent waves as they occur, if object details are comparable in size with, or even smaller than, the wavelength. Conceivable topics are 'super resolution' and 'sub-lambda lithography'. Fundamentals and history of classical evanescent holography are reported in the Further Reading.

Some Hardware Aspects

We now describe some hardware aspects of the CGHs. Remembering the CGH cell structure in Figure 3d, we saw the amplitude encoded as relative width W (eqn [6]). Instead of the width, one may use the relative height $H \leq 1$ as the amplitude parameter. More efficient, but more difficult to manufacture are saw tooth-shaped hologram cells. The saw tooth prism may be approximated by a phase stair, which is possible with today's lithographic technology. The corresponding theory is called phase quantization.

Some CGH Applications

Some applications had been mentioned already before, e.g., 3-D display (see Figure 7). The data volume can be very high, hence, shortcuts such as eliminating the vertical perspective, are necessary. The largest CGHs for visual light are used for testing astronomical telescopes by means of interferometry. The IFTA algorithm has been used, among other things, for designing a CGH which acts as an optimal diffuser. Speckles and light efficiency are the critical issues in diffuser design. Another project, that also benefits from IFTA, is beam shaping. That is a broad topic with aims at, for example, homogenizing the output of laser, beam structuring for welding, cutting, and scanning.

A CGH can also be used as a spatial filter for image processing. For example, an input image may be Hilbert-transformed in the setup shown in Figure 9a. The filter (Figure 9b) is a grating, of which one sideband is shifted by one-half of the grating period. Such a geometric shift generates the π-phase shift that is needed for the Hilbert transformation. The same setup can also be used to implement Zernike's phase contrast. Now the grating is distorted such that the zero-frequency region at the optical axis is reduced in amplitude by (typically) a factor of 0.2. In addition, the phase is shifted by $\pi/2$ due to a fringe shift of a quarter period (Figure 10a). In the image plane one can see an ordinary image on-axis, and two phase-contrast images in the plus-minus first diffraction orders. One of them has positive phase contrast, the

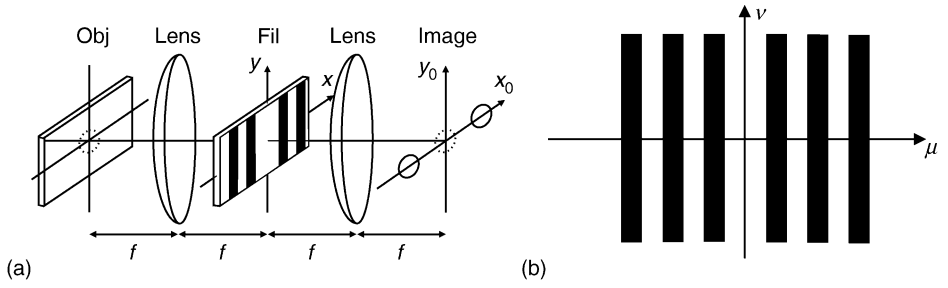


Figure 9 (a) Hilbert setup. (b) Hilbert filter.

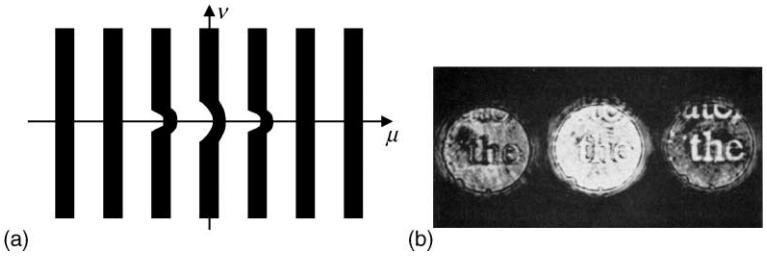


Figure 10 (a) Phase contrast filter. (b) Phase contrast output.

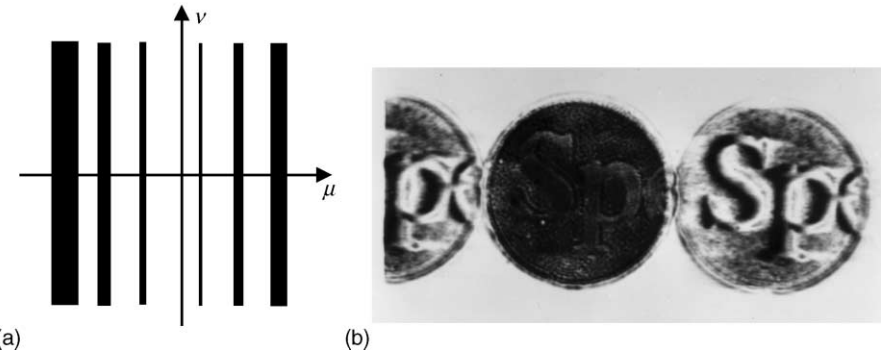


Figure 11 (a) Differentiation filter. (b) Differentiation output.

other one has negative phase contrast (Figure 10b). Another simple spatial filter (Figure 11a) produces a derivative of the input $u(x, y) : \partial u(x, y)/\partial x$ (Figure 11b). Equations [22]–[24] explain this differential filtering experiment.

$$u(x, y) \rightarrow \frac{\partial u(x, y)}{\partial x} = v(x, y) \qquad [22]$$

$$\tilde{u}(\mu, \nu) \rightarrow 2\pi i\mu \cdot \tilde{u}(\mu, \nu) = \tilde{v}(\mu, \nu) \qquad [23]$$

$$\tilde{p}(\mu, \nu) = 2\pi|\mu| \cdot \left\{ \begin{matrix} +i(\mu > 0) \\ -i(\mu > 0) \end{matrix} \right\} \qquad [24]$$

The phase portion of the filter (Figure 11a) is the same as for the Hilbert filter, but, the filter amplitude now enhances the higher frequencies. The object used for getting the results of Figure 10b and 11b was a phase-only object: a bleached photograph.

A very famous spatial filtering experiment is 'matched filtering', which can be used for pattern recognition. The original matched filters were classical Fourier holograms of the target pattern. Computer hologram's can do the same job, even with spatially incoherent objects, this work being initiated by Katyl.

Fourier holograms are also attractive for data storage, due to their robustness against local defects. However, a local error in the Fourier domain spreads out all over the image domain. The computed hologram has an advantage over classical holograms because of the freedom to incorporate any error detection and correction codes for reducing the bit error rate even further.

One of the most recent advances is in matched filtering with totally incoherent light. This step allows moving spatial filtering out of the well-guarded laboratory into hostile outdoor environments.

Finally, a word about terminology; several terms are used in place of 'CGH': synthetic hologram, digital hologram, holographic optical element (HOE), diffractive optical element (DOE) or digital optical element.

CGHs have also been made for other waves, such as acoustical waves, ocean waves, and microwaves. Digital antenna steering represents a case of computer holography.

See also

Diffraction: Fraunhofer Diffraction; Fresnel Diffraction. **Fourier Optics**. **Holography, Techniques:** Computer-Generated Holograms; Holographic Interferometry; Overview. **Imaging:** Information Theory in Imaging; Inverse Problems and Computational Imaging; Three-Dimensional Field Transformations. **Information Processing:** Coherent Analogue Optical Processors; Free-Space Optical Computing; Incoherent Analog Optical Processors. **Optical Communication Systems:** Free-Space Optical Communications. **Optical Processing Systems**. **Phase Control:** Wavefront Coding.

Further Reading

Andrés P (1999) *Opt. Lett.* 245: 1331.
Brown BR and Lohmann AW (1966) *Appl. Opt.* 5: 967.
Brown BR and Lohmann AW (1969) *IBM J. R&D* 13: 160.
Bryngdahl O (1973) *Progr. Opt.* 11: 168.
Bryngdahl O and Wyrowski F (1990) *Progr. Opt.* 28: 1–86.
Dallas WJ (1971) *Appl. Opt.* 10: 673.
Dallas WJ (1973) *Appl. Opt.* 12: 1179.
Dallas WJ (1980) *Topics in Appl. Phys.* 40: 291–366.
Dallas WJ (1999–2003) http://www.radiology.arizona.edu/dallas/CGH.htm. University of Arizona, Tucson, AZ.
Dresel T, Beyerlein M and Schwider J (1996) *Appl. Opt.* 35: 4615.
Fienup JR (1981) *Proc. SPIE* 373: 147.
Fienup JR (1982) *Appl. Opt.* 21: 2758.
Gerchberg RW and Saxton WO (1972) *Optik* 35: 237.
Goodman JW and Silvestri AM (1970) *IBM J. R&D* 14: 478.

Hansler RL (1968) *Appl. Opt.* 7: 1863.
Hauk D and Lohmann AW (1958) *Optik* 15: 275.
Ichioka Y and Lohmann AW (1972) *Appl. Opt.* 11: 2597.
Katyl RH (1972) *Appl. Opt.* 11: 1255.
Kozma A and Kelly DL (1965) *Appl. Opt.* 4: 387.
Lee WH (1978) *Progr. Opt.* 16: 219–232.
Lesem LB, Hirsch PM and Jordan JA (1969) *IBM J. R&D* 13: 150.
Lohmann AW and Paris DP (1967) *Appl. Opt.* 6: 1746.
Lohmann AW and Paris DP (1968) *Appl. Opt.* 7: 651.
Lohmann AW and Sinzinger S (1995) *Appl. Opt.* 34: 3172.
Lohmann AW and Werlich HW (1971) *Appl. Opt.* 10: 670.
McGovern AJ and Wyant JC (1971) *Appl. Opt.* 10: 619.
Pe'er A, Wang D, Lohmann AW and Friesem AA (1999) *Opt. Lett.* 24: 1469; (2000) 25: 776.
Sinzinger S and Jahns J (1999) *Micro Optics*. New York: Wiley.
Vander Lugt AB (1964) *IEEE, IT* 10: 139.
Wyrowski F and Bryngdahl O (1991) *Rep. Prog. Phys.* 1481–1571.
Yaroslavskii LP and Merzlyakov NS (1980) *Methods in Digital Holography*. New York: Plenum Press.

Digital Holography

W Osten, University of Stuttgart, Stuggart, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The sinusoidal nature of light, its ability to interfere, and the precise knowledge of its wavelength make it possible to measure several metrological quantities, such as distances and displacements, with high accuracy. However, the high frequency of light results in the difficulty that the primary quantity to be measured, i.e., the phase of the light wave, cannot be observed directly. All quadratic sensors (CCD, CMOS, photo plates) are only able to record intensities. Therefore, one makes use of a trick: the transformation of phase changes into recordable intensity changes as the basic principle of interferometry and holography. Holography provides the possibility to record and to reconstruct the complete information of optical wave fields. Numerous potential applications such as 3D-imaging of natural objects, fabrication of diffractive optical elements, and interferometric investigation of complex technical components become possible. Further progress has been made by recording holograms directly with electronic targets and not in photographic emulsions. In this way the independent numerical reconstruction of phase and intensity can be accomplished by the computer. One important consequence is that interferometric techniques, such as 3D-displacement analysis and 3D-contouring, can be implemented easily and with high flexibility. Furthermore, digitally recorded holograms can be transmitted via telecommunication networks and optically reconstructed at any desired place using an appropriate spatial light modulator. Consequently, the complete optical information of complex objects can be transported between different places with high velocity. This opens the possibility to compare nominal identical but physically different objects (master sample comparison) available at different locations with interferometric sensitivity. We call this remote metrology.

Direct Phase Reconstruction by Digital Holography

In digital speckle pattern interferometry (DSPI), the object is focused onto the target of an electronic sensor. Thus, an image plane hologram is formed as a result of the interference with an inline reference wave. In contrast with DSPI, a digital hologram is recorded without imaging. The target records the superposition of the reference and the object wave in the near-field region – a so-called Fresnel hologram. The basic optical setup in digital holography for recording holograms is the same as in conventional holography (**Figure 1a**). A laser beam is divided into two coherent beams, one illuminating the object and forming the object wave, the other entering the target directly and forming the reference wave. On this basis, very compact solutions are possible. **Figure 1b** shows a holographic camera where the interferometer, containing a beamsplitter and some wavefront shaping components, is mounted in front of the camera target.

For a description of the principle of digital Fresnel holography, we use a simplified version of the optical setup **(Figure 2a)**. The object is modeled by a plane rough surface that is located in the (x, y)-plane and illuminated by laser light. The scattered wavefield forms the object wave $u(x, y)$. The target of an electronic sensor (e.g., a CCD or a CMOS) used for recording the hologram is located in the (ξ, η)-plane at the distance d from the object. Following the basic principles of holography, the hologram $h(\xi, \eta)$ originates from the interference of the object wave $u(\xi, \eta)$ and the reference wave $r(\xi, \eta)$

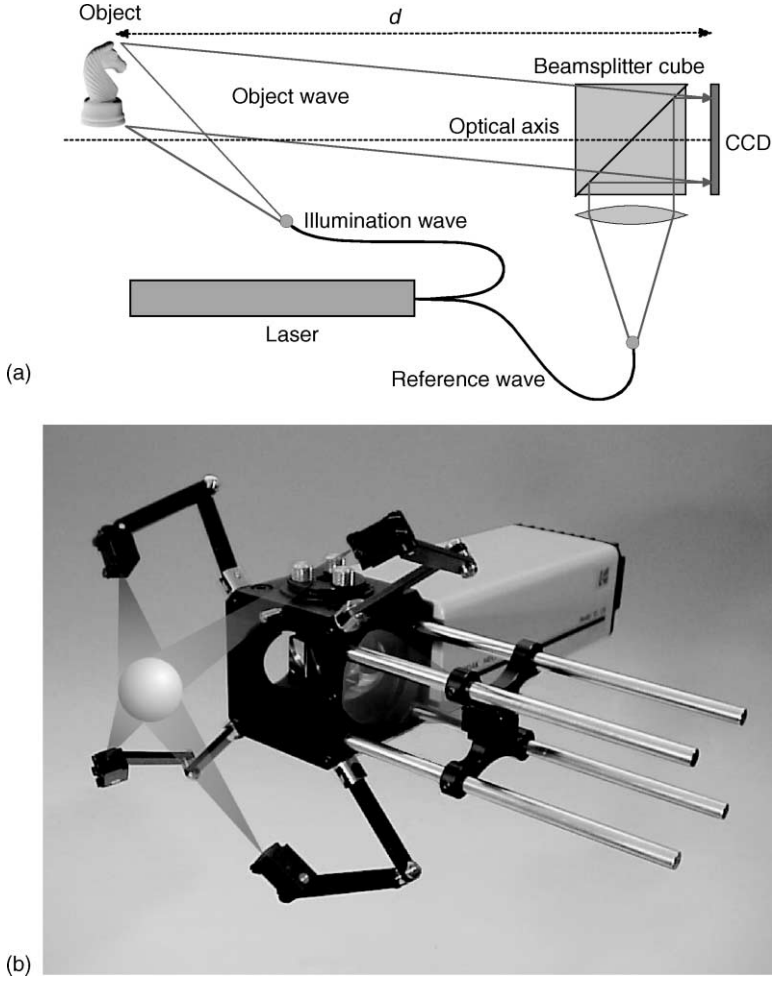


Figure 1 Setup for recording digital holograms. (a) Schematic setup for recording a digital hologram onto a CCD target. (b) Implementation of a holographic camera by mounting a compact holographic interferometer in front of a CCD-target (4 illumination directions are used). Reproduced from Seebacher S, Osten W, Baumbach Th and Jüptner W (2001) The determination of material parameters of microcomponents using digital holography. *Optics and Lasers in Engineering* 36(2): 103–126, copyright © 2001, with permission from Elsevier.

in the (ξ, η)-plane:

$$h(\xi,\eta) = |u(\xi,\eta) + r(\xi,\eta)|^2 = rr^* + ru^* + ur^* + uu^* \quad [1]$$

The transformation of the intensity distribution into a gray value distribution, that is stored in the image memory of the computer, is considered by the characteristic function t of the sensor. This function is generally only approximately linear:

$$T = t[h(\xi,\eta)] \quad [2]$$

Because the sensor has a limited spatial resolution, the spatial frequencies of the interference fringes in the hologram plane – the so-called micro-interferences – have to be considered. The fringe spacing g and the spatial frequency f_x, respectively, are determined by the angle β between the object and the reference wave (Figure 2b):

$$g = \frac{1}{f_x} = \frac{\lambda}{2\sin(\beta/2)} \quad [3]$$

with λ as the wavelength. If we assume that the discrete sensor has a pixel pitch (distance between adjacent pixels) $\Delta\xi$, the sampling theorem requires at least two pixels per fringe for a correct reconstruction of the periodic function:

$$2\Delta\xi < \frac{1}{f_x} \quad [4]$$

Consequently, we obtain for small angles β:

$$\beta < \frac{\lambda}{2\Delta\xi} \quad [5]$$

Modern high-resolution CCD or CMOS chips have a pitch, $\Delta\xi$, of about 4 μm. In this case, a maximum angle between the reference and the

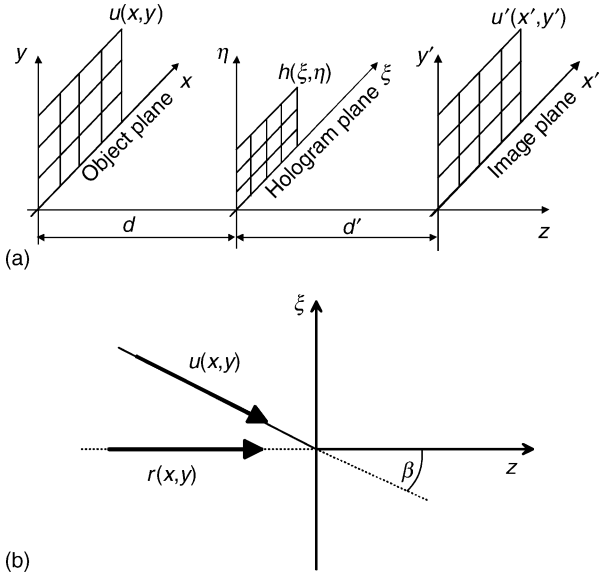


Figure 2 Geometry for recording and reconstruction of digital holograms. (a) Schematic setup for Fresnel holography. (b) Interference between the object and reference wave in the hologram plane.

object wave of only 4° is acceptable. The practical consequence of the restricted angle resolution in digital holography is a limitation of the effective object size that can be stored holographically by an electronic sensor. However, this is only a technical handicap. Larger objects can be placed at a sufficient distance from the hologram, or reduced optically by imaging with a negative lens.

The reconstruction is done by illuminating the hologram with a so-called reconstruction wave $c(\xi, \eta)$:

$$u'(x', y') = t[h(\xi, \eta)]c(\xi, \eta) \qquad [6]$$

The hologram, $t[h(\xi, \eta)]$, diffracts the wave, $c(\xi, \eta)$, in such a way that images of the object wave are reconstructed. Usually four terms are reconstructed if the wave, $u'(x', y')$, propagates in space. An assumed linear characteristic function, $t(h) = \alpha h + t_0$, delivers:

$$u' = Tc = t(h)c \qquad [7]$$

and

$$u' = \alpha\lfloor cu^2 + cr^2 + cur^* + cru^* \rfloor + ct_0 \qquad [8]$$

with two relevant image terms, $[cur^*]$ and $[cru^*]$. The appearance of the image terms depends on the concrete shape of the reconstruction wave. Usually the reference wave, $c = r$, or its conjugated version, $c = r^*$, is applied. In the case of the conjugated reference wave, a direct or real image will be reconstructed due to a converging image wave that can be imaged on a screen at the location of the original object.

However, in digital holography the reconstruction of the object wave in the image plane, $u'(x', y')$, is done by numerical simulation of the physical process, as shown in Figure 3a. The reconstruction wave with a well-known shape equal to the reference wave, $r(\xi, \eta)$, propagates through the hologram, $h(\xi, \eta)$. Following the Huygens principle, each point, $P(\xi, \eta)$, on the hologram, acts as the origin of a spherical elementary wave. The intensity of these elementary waves is modulated by the transparency, $h(\xi, \eta)$. In a given distance, $d' = d$, from the hologram, a sharp real image of the object can be reconstructed as the superposition of all elementary waves. For the reconstruction of a virtual image, $d' = -d$ is used.

Consequently, the calculation of the wavefield, $u'(x', y')$, in the image plane starts with the pointwise multiplication of the stored and transformed intensity values, $t[h(\xi, \eta)]$, with a numerical model of the reference wave, $r(\xi, \eta)$. If the hologram is assumed to be uniformly illuminated by a normally incident, monochromatic plane wave of amplitude 1 the reference wave can be modeled by $r(\xi, \eta) = 1$. After the multiplication the resulting field in the hologram plane is propagated in free space. In the distance, d', the diffracted field, $u'(x', y')$, can be found by solving the well-known Rayleigh–Sommerfeld diffraction formula, that is also known as the

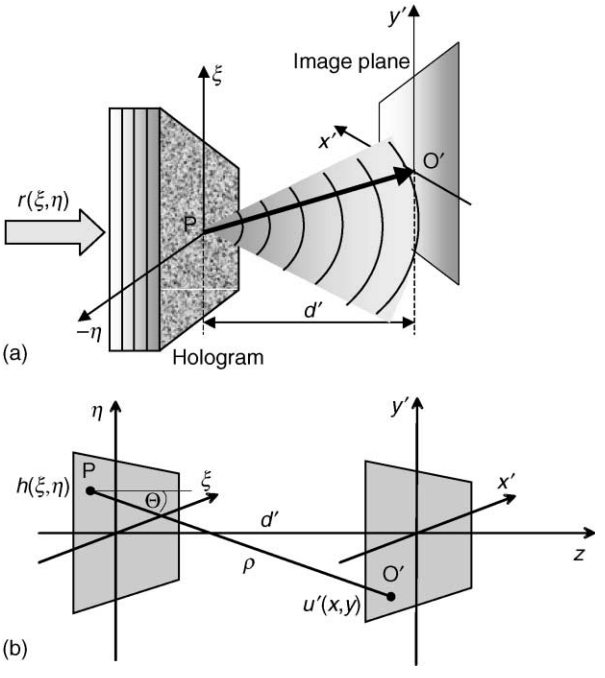


Figure 3 Reconstruction of digital holograms. (a) Principle of wave front reconstruction. (b) Light propagation by diffraction (Huygens–Fresnel principle).

Huygens–Fresnel principle:

$$u'(x',y') = \frac{1}{i\lambda}\int\int_{-\infty\cdots\infty} t[h(\xi,\eta)]r(\xi,\eta) \times \frac{\exp(ik\rho)}{\rho}\cos\theta\,d\xi\,d\eta \quad [9]$$

with

$$\rho(\xi-x',\eta-y') = \sqrt{d'^2+(\xi-x')^2+(\eta-y')^2} \quad [10]$$

as the distance between a point $O'(x',y',z=d')$ in the image plane and a point $P(\xi,\eta,z=0)$ in the hologram plane, and

$$k = \frac{2\pi}{\lambda} \quad [11]$$

as the wave number. The obliquity factor, $\cos\theta$, represents the cosine of the angle between the outward normal and the vector joining P to O′. This term is given as:

$$\cos\theta = \frac{d'}{\rho} \quad [12]$$

and therefore the Huygens–Fresnel principle can be rewritten:

$$u'(x',y') = \frac{d'}{i\lambda}\int\int_{-\infty\cdots\infty} t[h(\xi,\eta)]r(\xi,\eta)\frac{\exp(ik\rho)}{\rho^2}d\xi d\eta \quad [13]$$

This numerical reconstruction delivers the complex amplitude of the wavefront. Consequently, the phase distribution, $\phi(x',y')$, and the intensity, $I(x',y')$, can be calculated directly from the reconstructed complex function, $u'(x',y')$:

$$\phi(x',y') = \arctan\frac{\mathrm{Im}|u'(x',y')|}{\mathrm{Re}|u'(x',y')|} \qquad [-\pi,\pi] \quad [14]$$

$$I(x',y') = u'(x',y')u'^{*}(x',y') \quad [15]$$

The direct approach to the phase yields several advantages for imaging and metrology applications that are discussed later.

Reconstruction Principles in Digital Holography

Different numerical reconstruction principles have been investigated: the Fresnel approximation, the convolution approach, the lens-less Fourier approach, the phase shifting approach, and the phase-retrieval approach. In this section the main techniques are briefly described.

The Fresnel Approximation

If the distance, d, between the object and hologram plane, and equivalently, $d' = d$, between the hologram and image plane, is large compared to $(\xi - x')$ and $(\eta - y')$, then the denominator of eqn [13] can be replaced by d'^2 and the parameter ρ in the numerator can be approximated by a binomial expansion for the square root (10) where only the first two terms are considered:

$$\rho \approx d'\left[1 + \frac{(\xi-x')^2}{2d'^2} + \frac{(\eta-x')^2}{2d'^2}\right] \quad [16]$$

The resulting expression for the field at (x', y') becomes

$$u'(x',y') = \frac{\exp(ikd')}{i\lambda d'}\int\int_{-\infty\cdots\infty} t[h(\xi,\eta)]r(\xi,\eta) \times \exp\left[\frac{ik}{2d'}\{(\xi-x')^2+(\eta-y')^2\}d\xi\,d\eta\right] \quad [17]$$

Equation [17] is a convolution integral that can be expressed as

$$u'(x',y) = \int\int_{-\infty\cdots\infty} t[h(\xi,\eta)]r(\xi,\eta)H \times (\xi-x',\eta-y')d\xi\,d\eta \quad [18]$$

with the convolution kernel:

$$H(x',y') = \frac{\exp(ikd')}{i\lambda d'}\exp\left[\frac{ik}{2d'}(x'^2+y'^2)\right] \quad [19]$$

Another notation of eqn [17] is found if the term $\exp[(ik/2d')(x'^2+y'^2)]$ is taken out of the integral:

$$u'(x',y') = \frac{\exp(ikd')}{i\lambda d'}\exp\left[i\frac{k}{2d'}(x'^2+y'^2)\right]\int\int_{-\infty\cdots\infty} \times\left\{t[h(\xi,\eta)]r(\xi,\eta)\exp\left[i\frac{k}{2d'}(\xi^2+\eta^2)\right]\right\} \times\exp\left[-i\frac{2\pi}{\lambda d'}(\xi x'+\eta y')\right]d\xi\,d\eta \quad [20]$$

or

$$u'(x',y') = \frac{\exp(ikd')}{i\lambda d'}\exp\left[i\frac{k}{2d'}(x'^2+y'^2)\right] \times \mathrm{FT}_{\lambda d'}\left\{t[h(\xi,\eta)]r(\xi,\eta)\exp\left[i\frac{k}{2d'}(\xi^2+\eta^2)\right]\right\} \quad [21]$$

where $\mathrm{FT}_{\lambda d'}$ indicates the 2-dimensional Fourier transform that has been modified by a factor $1/(\lambda d')$. Equation [21] makes clear that the diffracted

wavefront consists of the Fourier transform (FT) of the digitally stored hologram, $t[h(\xi,\eta)]$, multiplied by the reference wave, $r(\xi,\eta)$, and the chirp function, $\exp\{(i\pi/\lambda d')(\xi^2+\eta^2)\}$. This Fourier transform is scaled by the constant factor, $1/(id'\lambda)$, and a phase factor that is independent of the processed hologram.

The discrete sampling in digital processing requires the transformation of the infinite continuous integral in eqn [20] into a finite discrete sum. This results in the finite Fresnel transform:

$$u'(m,n) = \sum_{j=0}^{M-1}\sum_{l=0}^{N-1} t[h(j\Delta\xi, l\Delta\eta)]r(j\Delta\xi, l\Delta\eta) \times \exp\left[\frac{i\pi}{d'\lambda}(j^2\Delta\xi^2 + l^2\Delta\eta^2)\right] \times \exp\left\{-i2\pi\left(\frac{jm}{M}+\frac{ln}{N}\right)\right\} \quad [22]$$

where constants and pure phase factors preceding the sums have been omitted. The main parameters are the pixel number $M \times N$ and the pixel pitches $\Delta\xi$ and $\Delta\eta$ in the two directions, which are defined by the used sensor chip.

The discrete phase distribution, $\phi(m,n)$, of the wavefront and the discrete intensity distribution, $I(m,n)$, on the rectangular grid of $M \times N$ sample points can be calculated from the reconstructed complex function, $u'(m,n)$, by using eqns [14] and [15], respectively. The wrapped phase distribution is computed directly without the need of additional temporal or spatial phase modulation such as phase shifting or spatial heterodyning.

For interferometric applications with respect to displacement and shape measurement the phase difference, $\delta(m,n)$, of two reconstructed wave fields, $u'_1(m,n)$ and $u'_2(m,n)$, needs to be calculated. The algorithm can be reduced to the following equation:

$$\delta(m,n) = \phi_1(m,n) - \phi_2(m,n) \quad [23]$$

Figure 4 shows the principle of displacement measurement by digital holography. Besides the investigation of surface displacements in the region of the wavelength, digital holography can be applied advantageously for the measurement of the shape of complex objects. In Figure 5, a turbine blade was investigated with the 2-wavelength contouring method. Both, the digitally reconstructed mod 2π-phase and its demodulated version after-phase unwrapping, are presented. Phase unwrapping is necessary in digital holography, since the fringe-counting problem still remains. However, there are some efficient approaches to overcome the difficulties of this process.

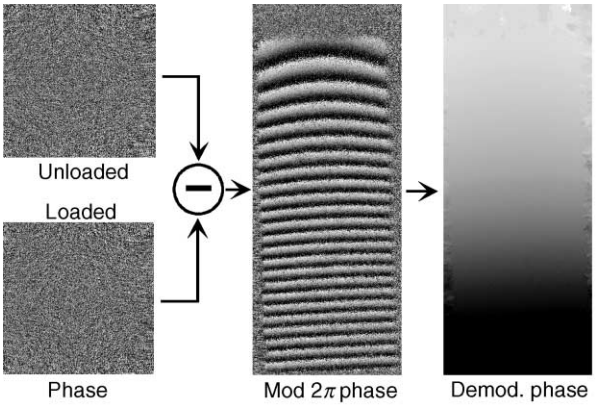


Figure 4 The principle of digital holographic interferometry demonstrated on an example of a loaded small beam. The interference phase is the result of the subtraction of the two phases which correspond to the digital holograms of both object states.

Following eqn [22] the pixel spacing in the reconstructed image is

$$\Delta x' = \frac{d'\lambda}{M\Delta\xi} \quad \text{and} \quad \Delta y' = \frac{d'\lambda}{N\Delta\eta} \quad [24]$$

Additional to the real image, a blurred virtual image and a bright dc-term, the zero-order diffracted field, is reconstructed. This term can effectively be eliminated by preprocessing the stored hologram or the different terms can be separated by using the off-axis instead of the in-line technique. However, the spatial separation between the object and the reference field requires a sufficient space-bandwidth product of the used CCD-chip, as discussed in the section on direct phase reconstruction above.

Numerical Reconstruction by the Convolution Approach

In the section above, we have already mentioned that the connection between the image term, $u'(x',y')$, and the product, $t[h(\xi,\eta)]r(\xi,\eta)$, can be described by a linear space-invariant system. The diffraction formula [17] is a superposition integral with the impulse response:

$$H(x'-\xi, y'-\eta) = \frac{\exp(ikd')}{i\lambda d'} \times \exp\left[\frac{ik}{2d'}\{(\xi-x')^2 + (\eta-y')^2\}\right] \quad [25]$$

A linear space-invariant system is characterized by a transfer function G, that can be calculated as the

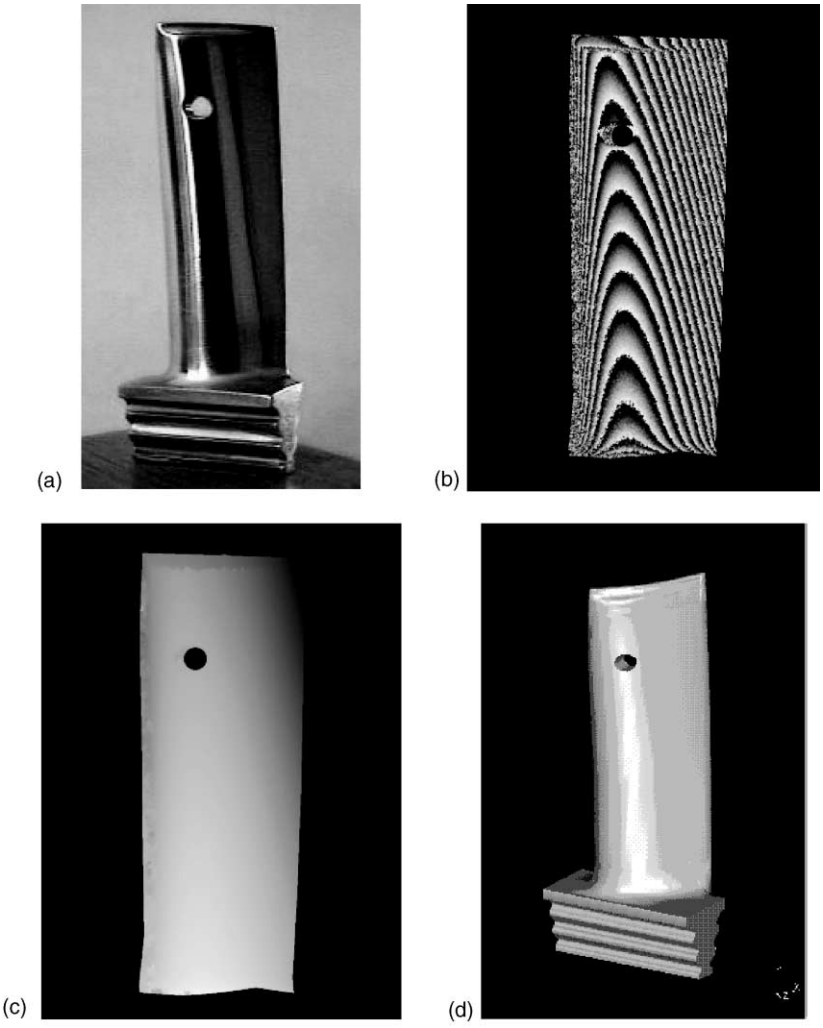


Figure 5 Phase reconstruction by digital holographic interferometry on an example of 2-wavelength contouring of a turbine blade. (a) Image of the turbine blade. (b) Mod-2π phase distribution. (c) Unwrapped phase distribution. (d) CAD reconstruction of the turbine blade.

Fourier-transform of the impulse response:

$$G(f_x, f_y) = \mathrm{FT}\{H(\xi, \eta, x', y')\} \qquad [26]$$

with the spatial frequencies f_x and f_y. Consequently, the convolution theorem can be applied, which states that the Fourier transform of the convolution, $t[h(\xi, \eta)]r(\xi, \eta)$, with H is the product of the individual Fourier transforms, $\mathrm{FT}\{t[h(\xi, \eta)]r(\xi, \eta)\}$ and $\mathrm{FT}\{H\}$. Thus $u'(x', y')$ can be calculated by the inverse Fourier transform of the product of the Fourier-transformed convolution partners:

$$u'(x', y') = \mathrm{FT}^{-1}\{\mathrm{FT}[t(h)r]\mathrm{FT}[H]\} \qquad [27]$$

$$u'(x', y') = [t(h)r] \otimes H \qquad [28]$$

with $\otimes$ as the convolution symbol. The computing effort is comparatively high: two complex multiplications and three Fourier transforms. The main difference to the Fresnel transform is the different pixel size in the reconstructed field:

$$\Delta x' = \Delta\xi \qquad \text{and} \qquad \Delta y' = \Delta\eta \qquad [29]$$

Numerical Reconstruction by the Lensless Fourier Approach

We have already mentioned that the limited spatial resolution of the sensor restricts the angle resolution of the digital hologram. The sampling theorem requires that the angle between the object beam and the reference beam at any point of the electronic sensor be limited in such a way that the

microinterference spacing is larger than double the pixel size. In general, the angle between the reference beam and the object beam varies over the sensor's surface, and so does the maximum spatial frequency. Thus, for most holographic setups, the full spatial bandwidth of the sensor cannot be used. However, it is important to use the entire spatial bandwidth of the sensor because the lateral resolution of the reconstructed image depends on a complete evaluation of the entire information obtained from the sensor.

Even if the speed of digital signal processing is increasing rapidly, algorithms should be as simple and as fast to compute as possible. For the Fresnel approach and the convolution approach, several fast Fourier transforms and complex multiplications are necessary. Therefore, a more effective approach, such as the subsequently described algorithm, seems promising.

The lensless Fourier approach is the fastest and most suitable algorithm for small objects. The corresponding setup is shown in Figure 6. It allows us to choose the lateral resolution in a range from a few microns to hundreds of microns without any additional optics. Each point, (ξ, η), on the hologram is again considered as a source point of a spherical elementary wavefront (Huygen's principle). The intensity of these elementary waves is modulated by $t[h(\xi, \eta)]$ – the amplitude transmission of the hologram. The reconstruction algorithm for lensless Fourier holography is based on the Fresnel reconstruction. Here again, $u(x, y)$ is the object wave in the object plane, $h(\xi, \eta)$ the hologram, $r(\xi, \eta)$ the reference wave in the hologram plane, and $u'(x', y')$ the reconstructed wavefield.

For the specific setup of lensless Fourier holography, a spherical reference wave is used with an origin at the same distance from the sensor as the object itself. In the case that $d' = -d$, $x = x'$, and $y = y'$, both virtual images are reconstructed and the reconstruction algorithm then reads:

$$u'(x', y') = \frac{\exp(\mathrm{i}kd)}{\mathrm{i}\lambda d} \mathrm{e}^{-\mathrm{i}\frac{k}{2d}(x^2+y^2)} \times \mathrm{FT}_{\lambda d'}\left\{t[h(\xi, \eta)] \cdot r(\xi, \eta) \cdot \mathrm{e}^{-\mathrm{i}\frac{k}{2d}(\xi^2+\eta^2)}\right\} \quad [30]$$

with $\mathrm{FT}_{\lambda d}$ as the 2-dimensional Fourier transformation which has been scaled by a factor $1/\lambda d$. In this recording configuration, the effect of the spherical phase factor associated with the Fresnel diffraction pattern of the object is eliminated by the use of a spherical reference wave, $r(\xi, \eta)$, with the same average curvature:

$$r(x, y) = \text{const.} \times \exp\left(\mathrm{i}\frac{\pi}{\lambda d}(\xi^2 + \eta^2)\right) \quad [31]$$

This results in a more simple reconstruction algorithm which can be described by

$$u'(x', y') = \text{const.} \times \exp\left(-\mathrm{i}\frac{\pi}{\lambda d}(x^2 + y^2)\right) \times \mathrm{FT}_{\lambda d}\{h(\xi, \eta)\} \quad [32]$$

Besides the simpler and faster reconstruction algorithm (only one Fourier transformation has to be computed), the Fourier algorithm uses the full space-bandwidth product (SBP) of the sensor chip because it adapts the curvature of the reference wave to the curvature of the object wave.

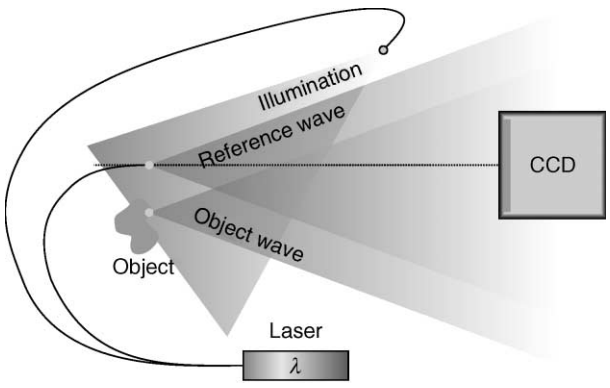


Figure 6 Scheme of a setup for digital lensless Fourier holography of diffusely reflecting objects.

Influences of Discretization

For the estimation of the lateral resolution in digital holography, three different effects related to discretization have to be considered: averaging, sampling, and the limited sensor size. We assume a quadratic sensor with $N \times M$ pixels of the size $\Delta\xi \times \Delta\xi$. Each pixel has a light-sensitive region with a side length $\gamma\Delta\xi = \Delta\xi_{\mathrm{eff}}$. γ^2 is the so-called fill factor $(0 \leq \gamma \leq 1)$, that indicates the active area of the pixel. The sensor averages the incident light which has to be considered because of possible intensity fluctuations over this area. The continuous expression for the intensity $I(\xi, \eta) = t[h(\xi, \eta)]$, registered by the sensor, has to be integrated over the light-sensitive area. This can be expressed mathematically by the convolution of the incident intensity $I(\xi, \eta)$, with a rectangle function:

$$I^1(\xi, \eta) \propto I \otimes \mathrm{rect}_{\Delta\xi_{\mathrm{eff}} \cdot \Delta\xi_{\mathrm{eff}}} \quad [33]$$

The discrete sampling of the light field is modeled by the multiplication of the continuous assumed hologram with the 2-dimensional comb-function:

$$I^2(\xi, \eta) \propto I^1(\xi, \eta) \times \mathrm{comb}_{\Delta\xi.\Delta\xi} \qquad [34]$$

Finally, the limited sensor size requires that the comb-function has to be truncated at the borders of the sensor. This is considered by the multiplication with a 2-dimensional rect-function of the size $N \times \Delta\xi$:

$$I^3(\xi, \eta) \propto I^2 \times \mathrm{rect}_{N\Delta\xi N.\Delta\xi} \qquad [35]$$

The consequences are amplitude distortions, aliasing and speckling that have to be considered in the reconstruction procedure.

Advantages of Digital Holography

Besides the electronic processing and the direct access to the phase, some further advantages recommend digital holography for several applications such as metrology and microscopy:

- The availability of the independently reconstructed phase distributions of each individual state of the object and interferometer, respectively, offers the possibility to record a series of digital holograms with increased load amplitude. In the evaluation process, the convenient states can be compared interferometrically. Furthermore, a series of digital holograms with increasing load can be applied to unwrap the mod 2π-phase temporally. In this method, the total object deformation is subdivided into many measurement steps, in which the phase differences are smaller than 2π. By adding up those intermediate results, the total phase change can be obtained without any further unwrapping. This is an important feature, since it is essential to have an unwrapped phase to be able to calculate the real deformation data from the phase map. Figure 7 shows an example of such a measurement. The left image shows the wrapped deformation phase for a clamped coin which was loaded with heat. The right image shows the temporal unwrapped phase which has been obtained by dividing the total deformation into 85 submeasurements. Thus, the displacement of the object can be observed almost in real time during the loading process.
- The independent recording and reconstruction of all states also gives a new degree of freedom for optical shape measurement. In the case of multiwavelength contouring, each hologram can be stored and reconstructed independently with its corresponding wavelength. This results in a drastic decrease of aberrations and makes it possible to use larger wavelength differences for the generation of shorter synthetic wavelengths.
- Because all states of the inspected object can be stored and evaluated independently, only seven digital holograms are necessary to measure the 3D-displacement field and shape of an object under test: one hologram for each illumination direction before and after the loading, respectively, and one hologram with a different wavelength (or a different source-point of illumination) which can interfere with one of the other holograms to perform two-wavelength-contouring (or two-source-point-contouring) for shape measurement. If four illumination directions are used, nine holograms are necessary.
- The direct access to all components of the wavefield makes it possible to correct wavefront aberrations effectively by computer.
- The digital hologram containing all information of the wavefront can be transmitted via the internet to any location and can be reconstructed digitally by computer or as an analogue version by a spatial light modulator. Consequently, remote access to wavefronts, that are generated at distant places, is possible.
- Finally, the possibility to miniaturize complex holographic sensor setups and to use this method for remote comparative interferometry, makes digital holography a versatile tool for the solution of numerous inspection and measurement problems.

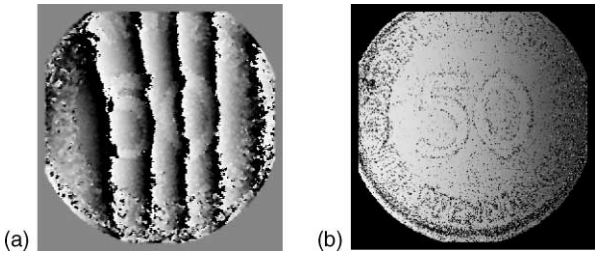


Figure 7 Temporal unwrapping by digital holography: (a) Wrapped phase of a thermal loaded coin. (b) Temporally unwrapped phase.

See also

Diffraction: Fresnel Diffraction. **Fourier Optics**. **Holography, Techniques:** Holographic Interferometry.

Further Reading

Butters JN and Leendertz JA (1971) Holographic and video techniques applied to engineering measurement. *Journal of Measurement Control* 4: 349–354.

Creath K (1993) Temporal phase measurement methods. In: Robinson DW and Reid GT (eds) *Interferogram Analysis*. Bristol and Philadelphia: IOP Publishing Ltd.

Demetrakopoulos T and Mittra T (1974) Digital and optical reconstruction of images from suboptical diffraction patterns. *Applied Optics* 13: 665–670.

Denisyuk YuN (1962) Photographic reconstruction of the optical properties of an object in its own scattered radiation field. *Doklady Akademii Nauk SSSR* 144: 1275–1279.

Gabor D (1949) Microscopy by reconstructed wavefronts. *Proceedings of the Royal Society, A* 197: 454–487.

Goodman JW (1996) *Introduction to Fourier Optics*. New York: McGraw Hill Comp. Inc.

Goodman JW and Lawrence RW (1967) Digital image formation from electronically detected holograms. *Applied Physics Letters* 11: 77–79.

Kolenovic E, Osten W, Klattenhoff R, Lai S, Kopylow C and Jüptner W (2003) Miniaturized digital holography sensor for distal three-dimensional endoscopy. *Applied Optics* 42(25): 5167–5172.

Kreis T (1996) *Holographic Interferometry*. Berlin: Akademie Verlag.

Kreis Th and Jüptner W (1997) Principles of digital holography. In: Jüptner W and Osten W (eds) *Proc. Fringe'97, 3rd International Workshop on Automatic Processing of Fringe Patterns*, pp. 353–363. Berlin: Akademie Verlag.

Kreis Th and Jüptner W (1997) The suppression of the dc-term in digital holography. *Optical Engineering* 36: 2357–2360.

Kujawinska M (1993) Spatial phase measurement methods. In: Robinson DW and Reid GT (eds) *Interferogram Analysis*. Bristol and Philadelphia: IOP Publishing Ltd.

Lee W-H (1978) Computer generated holograms: techniques and applications. In: Wolf E (ed.) *Progress in Optics*, vol. XVI, pp. 121–230. Amsterdam: North Holland.

Leith EN and Upatnieks J (1961) New techniques in wavefront reconstruction. *JOSA* 51: 1469–1473.

Osten W (2000) Holography and virtual 3D-testing. Proc. International Berlin Workshop HoloMet 2000, New Prospects of Holography and 3D-Metrology, *Strahltechnik* 14: 14–17.

Osten W (2003) Active metrology by digital holography. *Proceedings of SPIE* 4933: 96–110.

Osten W and Kujawinska M (2000) Active phase measuring metrology. In: Rastogi PK and Inaudi D (eds) *Trends in Optical Nondestructive Testing and Inspection*, pp. 45–69. Amsterdam: Elsevier Science B.V.

Osten W, Jüptner W and Seebacher S (1998) *The Qualification of Large Scale Approved Measurement Techniques for the Testing of Microcomponents*. Proc. 18th Symposium on Seebacher, Experimental Mechanics of Solids, Jachranka, pp. 43–55.

Osten W, Baumbach T, Seebacher S and Jüptner W (2001) Remote shape control by comparative digital holography. *Proc. Fringe 2001*, Bellingham: Elsevier Science, pp. 373–382.

Osten W, Seebacher S and Jüptner W (2001) The application of digital holography for the inspection of microcomponents. *Proceedings of SPIE* 4400: 1–15.

Osten W, Baumbach T and Jüptner W (2002) Comparative digital holography. *Optics Letters* 27(20): 1764–1766.

Pedrini G, Schedin S and Tiziani HJ (2001) Aberration compensation in digital holographic reconstruction of microscopic objects. *Journal of Modern Optics* 48: 1035–1041.

Powell RL and Stetson K (1965) Interferometric vibration analysis of three-dimensional objects by wavefront reconstruction. *JOSA* 55: 1593–1598.

Schnars U (1994) Direct phase determination in hologram interferometry with use of digitally recorded holograms. *Journal of the Optical Society of America A* 11: 2011–2015.

Schnars U and Jüptner W (1994) Direct recording of holograms by a CCD target and numerical reconstruction. *Applied Optics* 33(2): 179–181.

Seebacher S (2001) Application of digital holography for 3d-shape and deformation measurement of micro components. PhD thesis, University of Bremen.

Seebacher S, Osten W and Jüptner W (1997) 3-D deformation analysis of micro-components using digital holography. *Proceedings of SPIE* 3098: 382–391.

Seebacher S, Osten W, Baumbach Th and Jüptner W (2001) The determination of material parameters of microcomponents using digital holography. *Optics and Lasers in Engineering* 36(2): 103–126.

Takeda M, Taniguchi K, Hirayama T and Kogho H (1996) Single transform Fourier–Hartley fringe analysis for holographic interferometry. In: Füzessy Z, Jüptner W and Osten W (eds) *Simulation and Experiment in Laser Metrology*, pp. 67–73. Berlin: Akademie Verlag.

Wagner C, Seebacher S, Osten W and Jüptner W (1999) Digital recording and numerical reconstruction of lensless Fourier holograms in optical metrology. *Applied Optics* 38(22): 4812–4820.

Wagner C, Osten W and Seebacher S (2000) Direct shape measurement by digital wavefront reconstruction and multi-wavelength contouring. *Optical Engineering* 39(1): 79–85.

Yamaguchi I and Zhang T (1979) Phase-shifting digital holography. *Optical Letters* 22: 1268–1270.

Yaroslavsky LP and Merzlyakov NS (1980) *Methods of Digital Holography*. New York: Consultants Bureau.

Zhang Y, Pedrini G, Osten W and Tiziani H (2003) Image reconstruction for in-line holography with the Yang-Gu algorithm. *Applied Optics* 42(32): 6452–6457.

Holographic Interferometry

P Rastogi, Swiss Federal Institute of Technology, Lausanne, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The development of holographic interferometry has redefined our perception of optical interferometry. The unique property of holographic interferometry to bring a wavefront, generated at some earlier time, stored in a hologram and released at a later time, to interfere with a comparison wavefront, has made possible the interferometric comparison of a rough surface, which is subject to stress, with its normal state. Assuming the surface deformations to be very small, the interference of the two speckled wavefronts forms a set of interference fringes that are indicative of the amount of displacement and deformation undergone by the diffuse object.

Holographic interferometry is an important and exciting area of research and development. It has established itself as a highly promising noninvasive technique in the field of optical metrology. Techniques within its folds for the measurement of static and dynamic displacements, topographic contours, and flow fields have been developed and demonstrated with success in a wide range of problems. Significant and extensive contributions to the development of holographic interferometry have been realized both from the theoretical and experimental perspectives. A wide range of procedures have been developed which make it not only possible to measure surface displacements and deformations of engineering structures to an accuracy of a fraction of a micrometer but also to detect material flaws and inhomogeneities having escaped the manufacturing process. The methodology of quantitative analysis of holographic interferograms has undergone extensive development during the last decade. This article provides a brief review of the holographic techniques with emphasis placed on bringing out its relevance in experimental mechanics and nondestructive testing. The series of books given in the Further Reading section at the end of this article treat the field in detail.

Basic Methods in Holographic Interferometry

The ability of holography to store a wavefront and release it for reconstruction at a later time offers access to the possibility to compare wavefronts, which have albeit existed at different times. There are several schemes that have been established to obtain the interferometric comparison of the wavefronts.

Double-Exposure Holographic Interferometry

In frozen form of wavefront comparison, two holograms of the object are recorded on a photographic plate. The first hologram is made with the object in its initial, unstressed state and the second hologram is recorded with the object in its final, stressed state. Upon reconstruction, the hologram releases the two stored wavefronts, one corresponding to the object in its unstressed state and the other corresponding to the object in its stressed state (Figure 1). The virtual image is overlaid by a set of

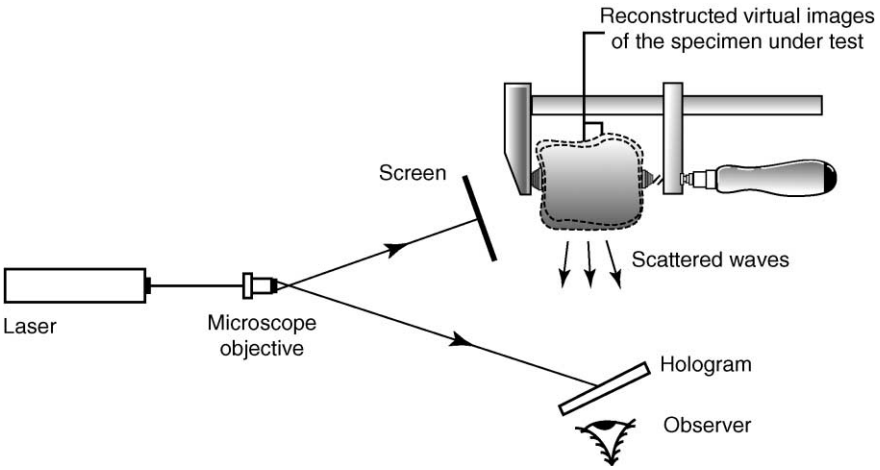


Figure 1 Layout for observing a double exposure holographic interferogram.

bright and dark interference fringes, which are due to the displacement of the object between the two exposures.

The fringe pattern is permanently stored on the holographic plate. The interference fringes denote the loci of points that have undergone an equal change of optical path between the light source and the observer. The image intensity of the holographic interferogram is

$$I(x, y) = I_0(x, y)\{1 + V(x, y)\cos(\varphi(x, y) - \varphi'(x, y))\} \quad [1]$$

where $I_0(x, y)$ is the background intensity, $V(x, y)$ is the fringe contrast, and φ and φ' are the phases of the waves from the object in its initial and deformed states, respectively.

A bright fringe is produced whenever

$$\varphi(x, y) - \varphi'(x, y) = 2n\pi \qquad n = 0, 1, 2, \ldots \quad [2]$$

where n is the fringe number. The wave scattered by a rough surface shows rapid variations of phase, which have little correlation across the wavefront. The phase difference varies randomly over the ray directions contained within the aperture of the viewing system, which would normally prevent the observation of fringes with large aperture optics near such points. However, a surface may exist at a distance from the object at which the value of phase difference is stationary over the cone of ray pairs defined by the viewing aperture. Interference fringes are localized in the region where the variation in phase difference is minimum over the range of viewing directions. This approach has been widely used to compute fringe localization for any object displacement and for any arbitrary illumination and observation geometry.

Real-Time Holographic Interferometry

A hologram is made of an arbitrarily shaped rough surface. After development, the hologram is placed back in exactly the same position. Upon reconstruction, the hologram produces the original wavefront. A person looking through the hologram sees a superposition of the original object and its reconstructed image (Figure 2). The object wave interferes with the reconstructed wave to produce a dark field due to destructive interference. If the object is now slightly deformed, interference fringes are produced which are related to the change in shape of the object. A dark fringe is produced, whenever:

$$\varphi(x, y) - \varphi'(x, y) = 2n\pi \qquad n = 0, 1, 2, \ldots \quad [3]$$

The method is very useful for determining the direction of the object displacement, and for compensating on the interferogram the influence of the rigid body motion of the object when subjected to a stress field.

The optical setup for real-time holography uses an immersion tank, which is mounted on a universal stage and contains the photographic plate. The holographic plate is followed by a vidicon camera, which visualizes the interferograms directly onto the TV monitor screen connected to a video-recording tape to memorize the information. The use of thermoplastic plates as recording material has gained in importance in recent years for implementing real-time and double exposure holographic interferometry setups, as it can be processed rapidly *in situ* using electrical and thermal processing.

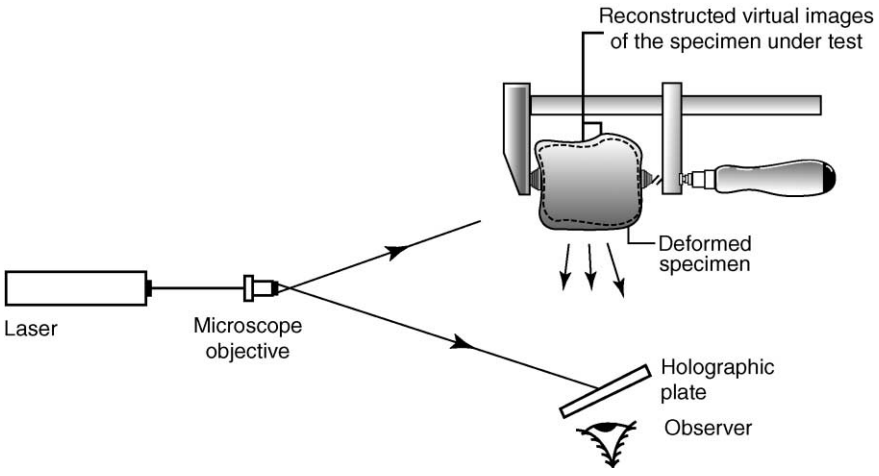


Figure 2 Layout for observing a holographic interferogram in real-time.

The plate is also cost-effective as it is erasable and reusable at least several hundred times. In thermoplastic recording the information is recorded as a thickness variation corresponding to a charge intensity pattern deposited on the thermoplastic layer. Self-developing or photorefractive crystals are other interesting alternatives to the use of film as the recording medium.

Digital Holographic Interferometry

The method consists of recording two holograms corresponding to undeformed and deformed states of an object using a CCD camera. These images are stored in a digital image processing system. The two holograms are then reconstructed separately by numerical methods from the digitaly stored holograms. The advantage of digital holography is that it allows for computing both the intensity and the phase of a holographically stored wavefront during the numerical reconstruction. The contours of constant optical phase change, due to deformation, are obtained by subtracting the reconstructed phases of the undeformed from the deformed object wavefield.

A hologram can be recorded on a CCD as long as the sampling theorem is fulfilled; that is to say, each period of the spatial variation of the hologram intensity is sampled by at least two pixels of the CCD array. The condition imposes a limit on the angle that the object beam can make with the reference beam, which must necessarily be small given that the resolution of CCDs is relatively low. The applications of the method are thus limited mainly by the pixel size in CCD array. The maximum spatial frequency $f_{\max}$ that must be resolved by the recording medium is determined by the maximum angle $\Delta\theta_{\max}$ between the reference and the object wave:

$$f_{\max} = \frac{2}{\lambda}\sin\frac{\Delta\theta_{\max}}{2} \qquad [4]$$

where λ is the wavelength of the laser source. Since the CCD cameras have resolutions of around 100 lines/mm, the maximum angle between the object and reference waves is limited to a few degrees.

Calculation of Phase Change on Object Loading

A diffusely reflecting object is deformed such that a point P on the surface moves to P′ as shown in Figure 3. In double-exposure or real-time holographic interferometry the two wavefronts related to the states of a surface before and after deformation are reconstructed simultaneously and compared. The change of optical path length due to displacement d is

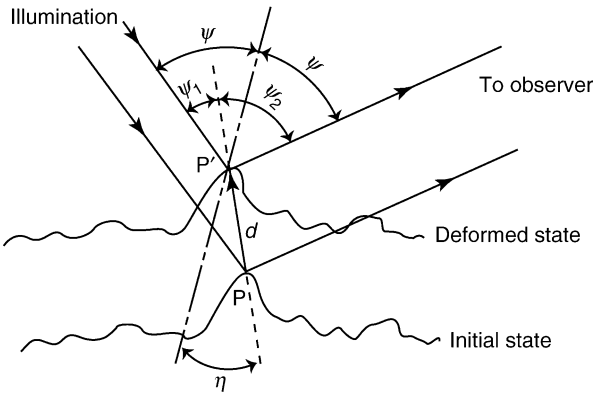


Figure 3 Calculation of the change in optical phase due to surface displacement.

$$\Delta\varphi = \frac{2\pi d}{\lambda}(\cos\psi_1 + \cos\psi_2) \qquad [5]$$

where ψ_1 and ψ_2 are the angles which the illumination and observation waves make with respect to the direction of displacement; eqn [5] can be written in the form:

$$\Delta\varphi = \frac{2\pi d}{\lambda}\cos\eta\cos\psi \qquad [6]$$

where ψ is the angle of bisection between the incident and viewing directions, and η is the angle which the bisector makes with the direction of displacement. The term $d\cos\eta$ being the resolved part of the displacement PP′ in the direction of the bisector, implies that the fringe pattern provides the measure of the displacement component along the bisector of the angle between the incident and viewing directions.

Digital Phase Measurement

The last decade has seen a rapid development of techniques for the automatic and precise reconstruction of phases from fringe patterns. These are based on the concepts of fringe tracking, Fourier transform, carrier frequency, and phase shift interferometry. Quantitative measurement of wavefront phase, for example, can be obtained in a few seconds using phase shift interferometry. With this technique one records a series of holograms by introducing artificially known steps of phase differences in the interference image in eqn [1]:

$$I_i(x,y) = I_0(x,y)\{1 + V(x,y)\cos(\Delta\varphi(x,y) + \Delta\psi_i)\};\quad i = 0,1,2,\ldots \qquad [7]$$

where $\Delta\psi_i$ is the phase shift which is produced, for example, by shifting the phase of the reference wave. A minimum of three intensity patterns is required to calculate the phase at each point on the object surface. Known phase shifts are commonly introduced using a piezoelectric transducer, which serves to shorten or elongate the reference beam path by a fraction of wavelength; the phase-shifter is placed in one arm of the interferometer. Numerous phase shifting techniques have been developed and which can be incorporated to holographic interferometry to generate a phase map, $\Delta\varphi(x, y)$, corresponding to the object displacement. The calculated phase $\Delta\varphi(x, y)$ is independent of the terms $I_0(x, y)$ and $V(x, y)$, and considerably reduces the dependence of the accuracy of measurements on the fringe quality.

The introduction of phase shift interferometry not only provides accurate phase measurements but also eliminates the problem of phase sign ambiguity of the interference fringes. Dynamic phase shifting techniques allow for producing time sequences of deformation maps for studying objects subjected to time-varying loads. The concept is based on considering the normal phase variations caused by a dynamic phenomenon as equivalent to the phase shifting that is introduced in a holographic interferometer in order to perform the phase evaluation task. This method extends the possibility of applying the method to continuous deformation measurements.

Basic Interferometers for Static Displacement Measurement

Equation [6] shows that observation along a single direction yields information only about the resolved part of the surface displacement in one particular direction. The approach to measuring three-dimensional vector displacement would be to record holograms from three different directions in order to obtain three independent resolved components of displacement. The computation of the resulting system of equations would then allow for determining the complete displacement vector.

Out-of-Plane Displacement Component Measurement

The out-of-plane component of displacement is usually measured by illuminating and viewing the object surface in a near normal direction. The fringe equation corresponding to the line of sight displacement component is given by

$$w = \frac{n\lambda}{2} \qquad [8]$$

where w is the out-of-plane component of displacement. An example of a fringe pattern depicting the out-of-plane displacements in a rectangular aluminum plate, clamped along the edges and drawn out at the center by means of a bolt, subjected to three-point loading, is shown in Figure 4.

In-Plane Displacement Component Measurement

The in-plane component of displacement can be measured by illuminating the object along two directions symmetric to the surface normal and observing along a common direction. This is schematized in Figure 5. The in-plane displacements, in a direction containing the two beams, are mapped as a moiré between the interferograms due to each illumination beam. The moiré appears as a family of fringes:

$$u = \frac{n\lambda}{2\sin\theta_e} \qquad [9]$$

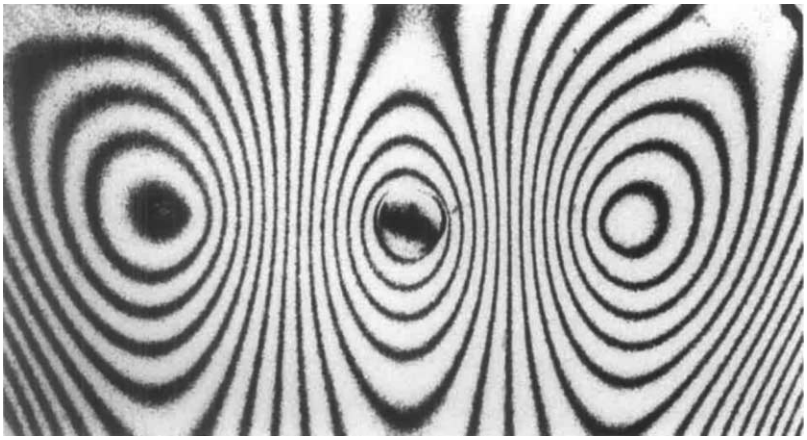


Figure 4 Example of interferogram corresponding to the whole field distribution of out-of-plane component of displacement, w, on an aluminum plate subjected to three-point bending.

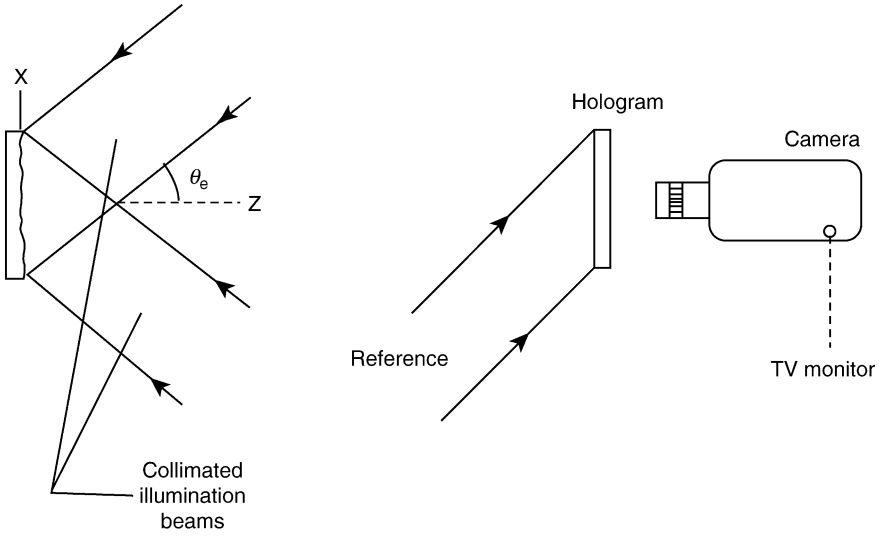


Figure 5 Schematic of the geometrical configuration for measuring the in-plane component of displacement, *u*.

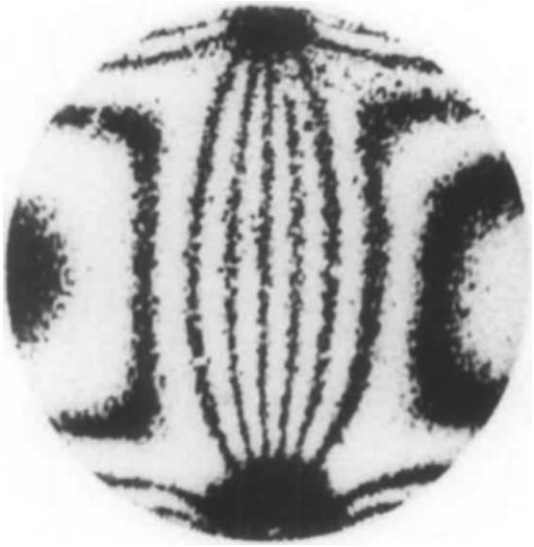


Figure 6 Fringe contours showing the in-plane displacement response of a disk under diametral compression.

where u is the displacement component in the x-direction and θ_e is the angle which the illumination beams make with the surface normal. Figure 6 shows the u pattern corresponding to a disk under diametral compression. The moiré pattern is obtained after optically filtering the interference image.

Comparative Holographic Interferometry

Another technique of significant interest in non-destructive testing is comparative holography that provides the contours of path variations related to the difference in displacements or shapes of two objects. The method provides information about the resolved part of the difference in displacement vector along the direction bisecting the illumination and observation rays. In the case of illumination and observation normal to the object surface, the fringe equation becomes

$$\Delta w = \frac{n\lambda}{2} \qquad [10]$$

where Δw is the difference in displacement component along z direction. The technique provides a tool to compare the mechanical responses of two nominally identical specimens subjected to the same loading and also the shapes of two nominally identical specimens. These features are useful in detecting anomalies in a test specimen with respect to the flaw-free master specimen. One such example is shown in Figure 7a, which displays the presence of a defect in a test specimen. In this case two plates, one flaw-free and the other with a flaw, are subjected to identical loadings. The fringe map in Figure 7b is related to the difference in displacements of test and master specimens subjected to unequal loadings. The presence of flaws is detected in both interferograms.

Basic Interferometers for Dynamic Displacement Measurement

Another important application of holographic interferometry is in the study of the resonant modes of vibration of an object.

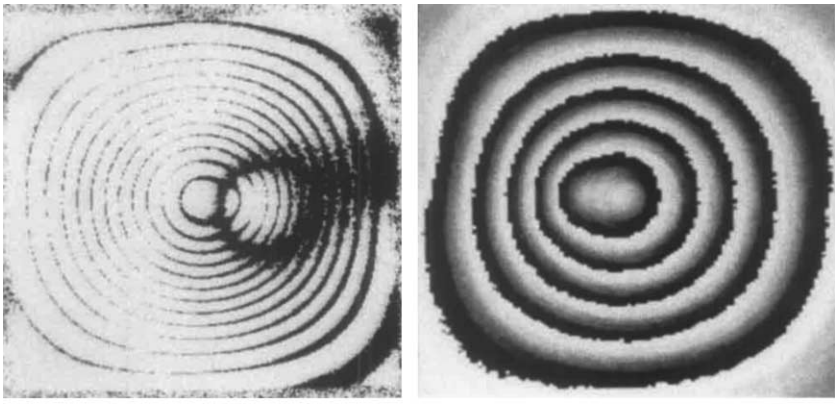


Figure 7 Comparative holography reveals the presence of defects in aluminum plates using (a) identical and (b) unequal loadings.

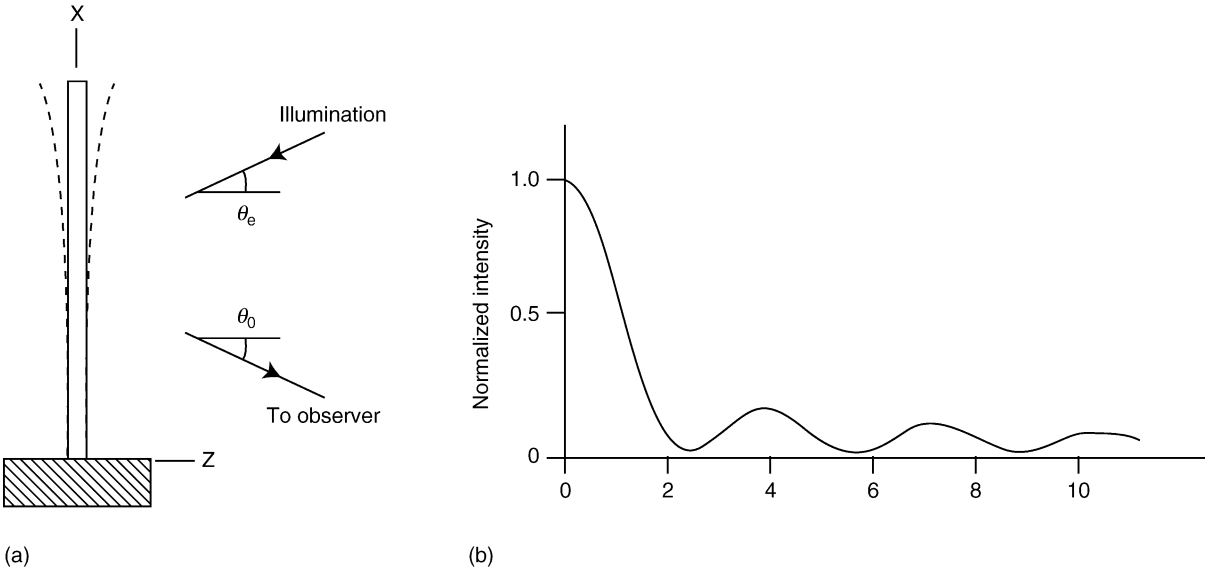


Figure 8 (a) Schematic of the geometrical configuration of a vibrating cantilever beam; (b) plot of fringe function for time-average holography.

Frozen Time-Average Holographic Interferometry

Consider that a hologram is recorded of an object, which is vibrating sinusoidally in a direction normal to its surface (Figure 8a). The exposure time is supposed to be much longer than the period of vibration. The intensity distribution of the image reconstructed by this hologram is

$$I(x, y) = I_o(x, y) J_o^2\left[\frac{2\pi}{\lambda} d(x, y)(\cos\theta_e + \cos\theta_o)\right] \quad [11]$$

where J_o is the zero-order Bessel function of the first kind. The virtual image is modulated by the $J_o^2(\xi)$ function. The dark fringes correspond to the zeros of the function $J_o^2(\xi)$. A plot of the function, shown in Figure 8b, is characterized by a comparatively brighter zero-order fringe, which corresponds to the nodes, a decreasing intensity, and an unequal spacing between the successive zeros. The interferogram in Figure 9 shows the vibrational modes of a helicopter component. The nodes that represent zero motion are clearly seen as the brightest areas in the time average reconstructed pattern.

Real-Time Time-Average Holographic Interferometry

The possibility of studying the response of vibrating objects in real-time extends the usefulness of the technique to identify the resonance states (resonant vibration mode shapes and resonant frequencies) of the object. A single exposure is made of the object in its state of rest. The plate is processed, returned to its original condition and reconstructed. The observer looking through the hologram at the sinusoidally

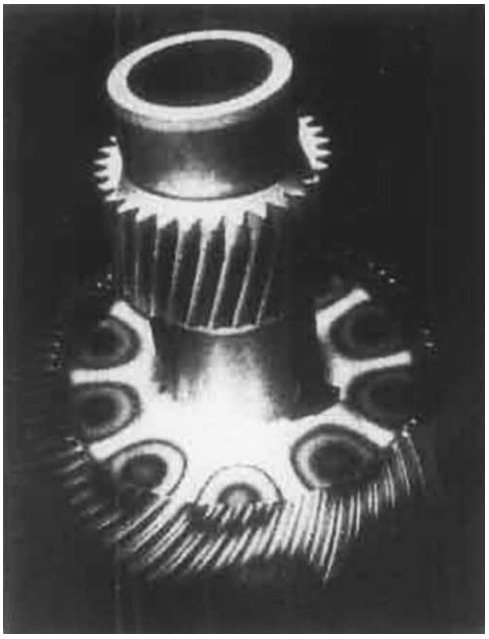


Figure 9 Fringe contours showing the mode shape of a vibrating helicopter component.

vibrating object sees the time-averaged intensity:

$$I(x,y) = I_o(x,y)\left[1 - J_o\left\{\frac{2\pi}{\lambda} d(x,y)(\cos\theta_e + \cos\theta_o)\right\}\right] \quad [12]$$

Stroboscopic Holographic Interferometry

Stroboscopic holography is another interesting variant in the study of vibrations. The hologram is recorded by exposing the photographic plate twice for short time intervals during a vibration cycle. The pulsed exposures are synchronized with the vibrating surface, which is equivalent to making the surface virtually stationary during the recording. Reconstruction of the hologram yields cosinusoidal fringes, which are characteristic of double exposure holography. In real-time stroboscopic holography, a hologram of a nonvibrating object is first recorded. If the vibrating object is illuminated stroboscopically and viewed through the hologram, the reconstructed image from the hologram interferes directly with the light scattered from the object to generate live fringes.

Flow Measurement

Application of holographic interferometry to flow visualization and the measurement of spatial refractive index, density or temperature distributions have led to an advancement of understanding in areas such as aerodynamics, plasma diagnostics, and heat transfer. In the holographic interferometer shown in Figure 10 two consecutive exposures are made, usually the first exposure without flow and the second in the presence of a flow field. Double pulsed holography is used if the flow field is changing rapidly. The optical phase change due to flow between the exposures is

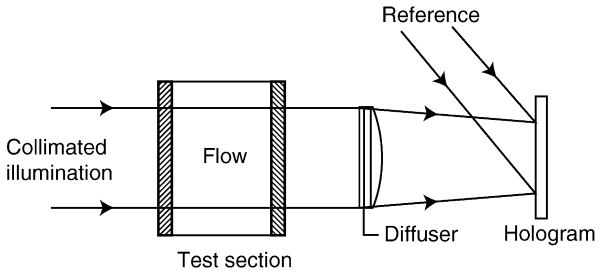


Figure 10 Schematic of a configuration for studying flow.

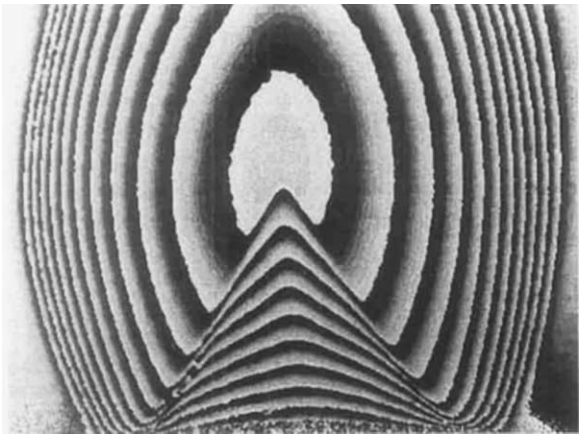


Figure 11 Fringe contours showing the phase distribution of a test flame.

$$\Delta\varphi(x,y) = \frac{2\pi}{\lambda}\int_0^t \{n'(x,y,z) - n'_o\}dz \quad [13]$$

where $n'(x,y,z)$ is the refractive index distribution during the second exposure, n'_o is the uniform refractive index during the first exposure, and t is the length of the test section. Assuming that the properties of the flow are constant in the z direction, the expression for phase change can be expressed as

$$\Delta\tau(x,y) = KL\{\rho(x,y) - \rho_o\} \quad [14]$$

where K is the Gladstone–Dale constant, ρ is the density of the gas, and ρ_o is the density in a no-flow situation. The interference pattern contours the change in the density field of the flow. The change in density per fringe is given by λ/Kt. An example of application related to temperature measurements in an axisymmetric premixed flame is illustrated in Figure 11.

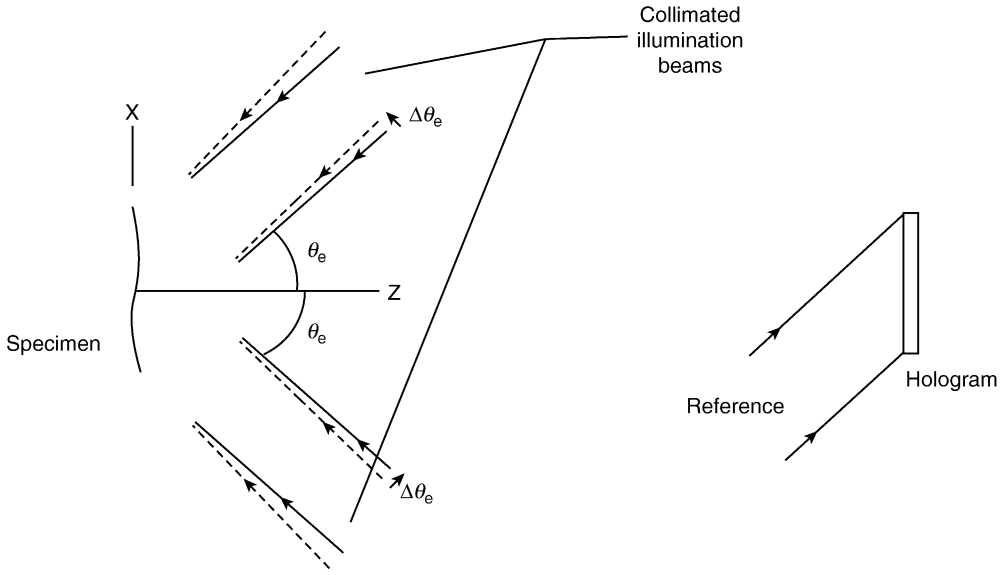


Figure 12 Schematic diagram of a two-beam method for shape measurement.

Surface Contouring

A practical way to display the shape of an object is to obtain contour maps showing the intersection of the object with a set of equidistant planes orthogonal to the line of sight.

The Dual Refractive Index Method

The method requires placing the model in a glass tank filled with a liquid (or gas) of refractive index n'_1. A hologram of the object is recorded, and the liquid contained in the tank is replaced by one of refractive index n'_2. An observer looking through the hologram sees the object surface modulated by a set of interference fringes arising from the change of optical phase in the light rays traversing the two liquids. The contour interval is given by

$$\Delta z = \frac{\lambda}{2(n'_1 - n'_2)} \qquad [15]$$

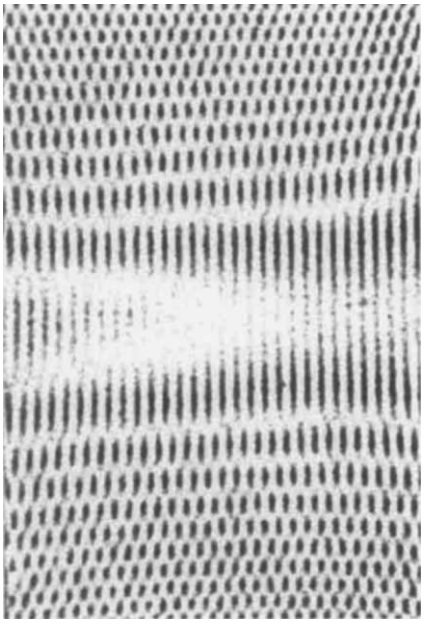


Figure 13 Example of a moiré topographic contour pattern obtained using the two-beam method.

The Dual Illumination Method

In this method, the object is illuminated obliquely by collimated beams along two directions symmetric to the surface normal. The scattered waves are recorded on a holographic plate and the object and its reconstructed image are viewed through the hologram. Two sets of parallel equidistant fringes are projected onto the object surface by introducing small tilts, $\Delta\theta_e$, to the illumination beams (Figure 12). The contour sensitivity per fringe is given by

$$\Delta z = \frac{\lambda}{2\sin\theta_e \sin\Delta\theta_e} \qquad [16]$$

The sensitivity of the method can be tuned in a wide range. An example of topographic contour pattern is shown in Figure 13. The distance between two adjacent moiré contour planes is 172 μm.

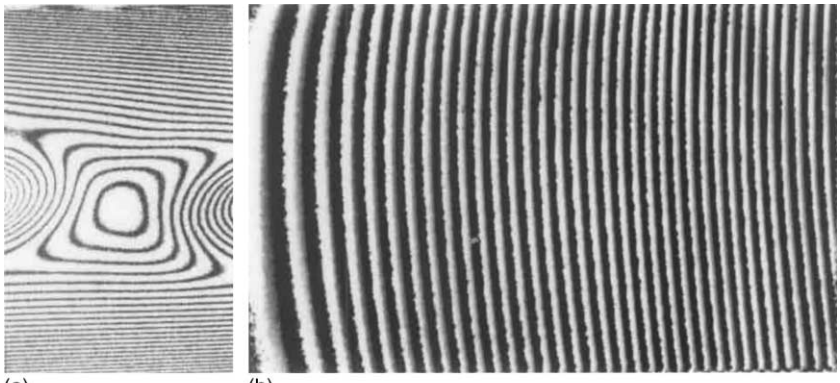


Figure 14 Examples of (a) shape measurement: the distance between two adjacent contour planes is 208 μm; and, (b) very large out-of-plane deformation measurement on a cantilever beam. The distance between two adjacent contour planes is 19 μm.

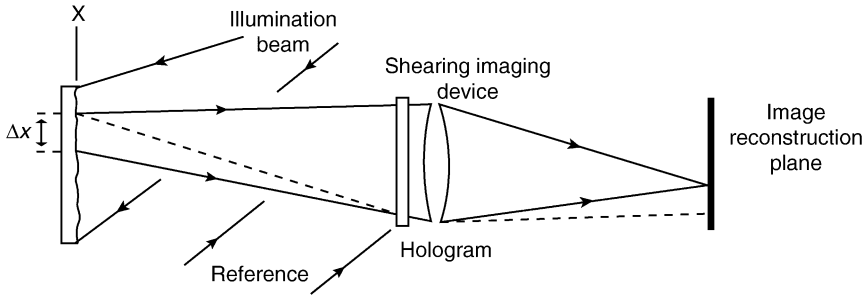


Figure 15 Schematic diagram of the principle used in holographic shearing interferometry.

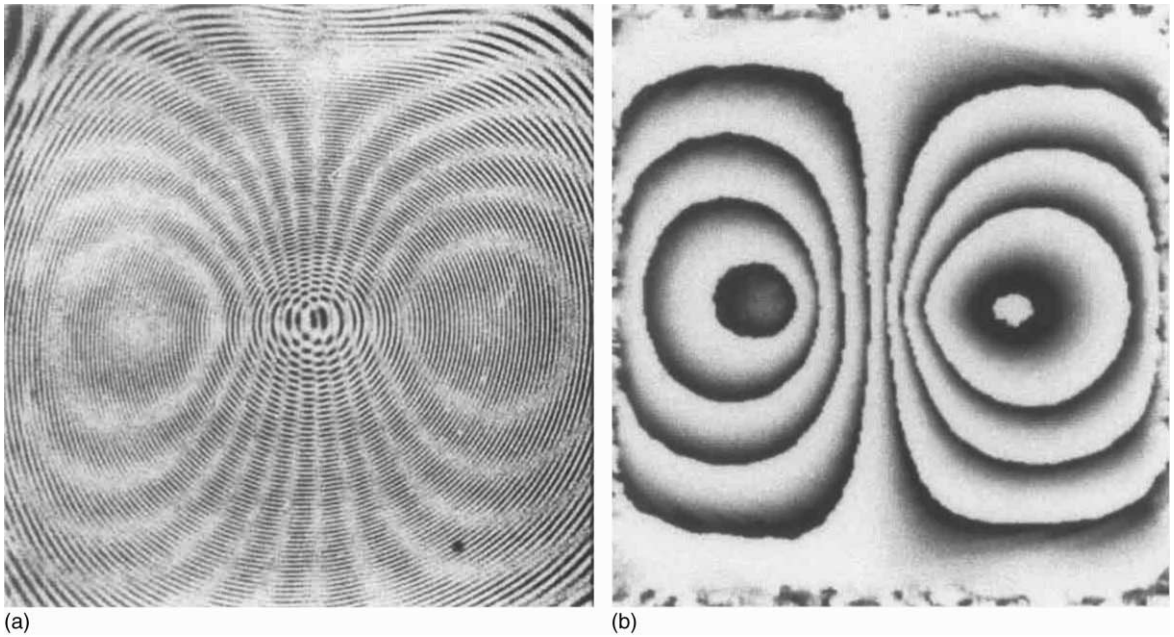


Figure 16 Examples of (a) moiré fringe contour, and (b) phase distribution corresponding to slope change produced on a centrally loaded aluminum plate clamped along its boundary.

In another method a collimated beam obliquely illuminates the object and a set of parallel equidistant fringes is projected onto the object surface by introducing a small tilt, $\Delta\theta_e$, to the illumination beam. Providing an appropriate reference beam rotation and a holographic plate translation generates contouring surfaces normal to the line of sight. A typical result using one beam illumination is shown in Figure 14a. The distance between two adjacent moiré contour planes is 208 μm.

This approach has been used to obtain measurement of large out-of-plane deformations undergone by a deformed object. The fringe pattern in Figure 14b corresponds to the out-of-plane displacement of a cantilever beam subjected to load at its free end. The distance between two adjacent contour planes is 19 μm.

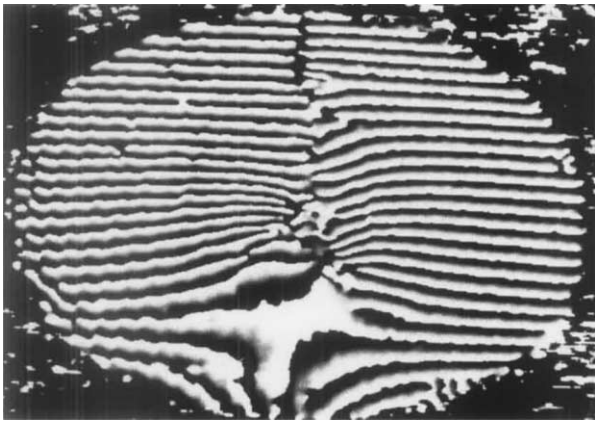


Figure 17 Example taken from the study of the fracture process zone in concrete: phase distribution corresponds to in-plane displacements on a notched concrete specimen subjected to a wedge splitting test.

Holographic Shearing Interferometry

Holographic shearing interferometry directly provides the patterns of slope change contours by laterally shearing wavefronts diffracted from the object surface. There are numerous ways to achieve shearing, such as, by introducing a Michelson type configuration, an inclined glass plate device, or a split-lens assembly in a holographic interferometer. The role of the shearing device is to enable the observation of a point on the object along two distinct neighboring directions. As a result of lateral wavefront shearing in the interferometer, a point in the image plane receives contributions from two different points on the object (Figure 15). Assuming that the illumination beam lies in the x–z plane and makes an angle θ_e with the z-axis, and the observation is carried along the direction normal to the object surface, the change of optical phase due to deformation is

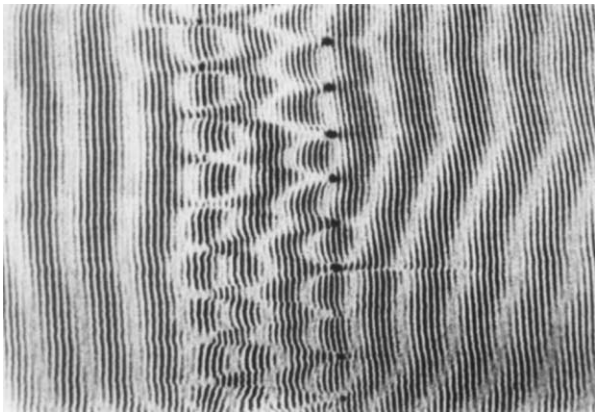


Figure 18 Example of application to nondestructive testing on a fingertip-joint wood specimen under tension using an in-plane sensitive holographic moiré setup.

$$\Delta\varphi = \frac{2\pi}{\lambda}\left[\frac{\partial u}{\partial x}\sin\theta_e + \frac{\partial w}{\partial x}(1+\cos\theta_e)\right]\Delta x \qquad [17]$$

where Δx is the object shear along the x-direction. For $\theta_e = 0$, eqn [11] reduces to

$$\frac{\partial w}{\partial x} = \frac{n\lambda}{2\Delta x} \qquad [18]$$

Equation [18] displays a family of holographic moiré fringes corresponding to the contours of constant slope change. The fringe pattern shown in Figure 16a illustrates a moiré corresponding to contours of constant slope change for a centrally loaded square aluminum plate clamped along its boundary. An example of phase distribution

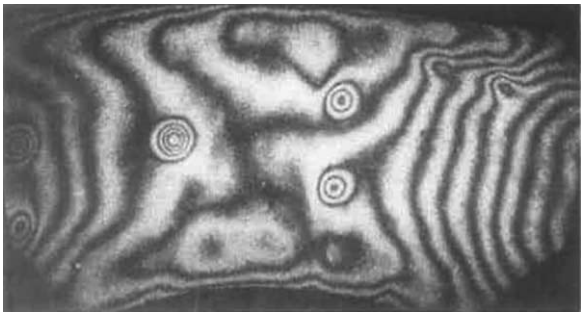


Figure 19 Application to the detection of defects in a tyre. Courtesy of H. Steinbichler, Steinbichler GmbH.

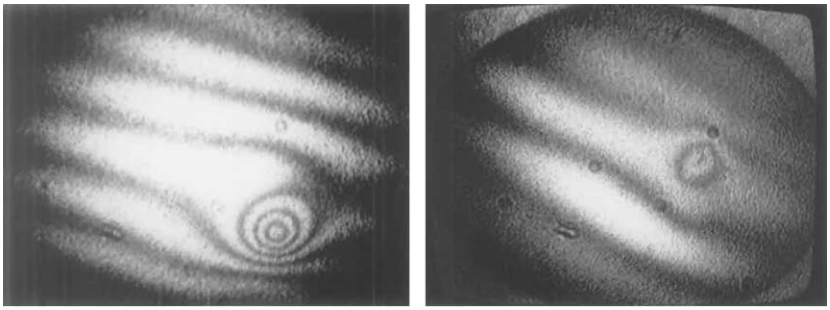


Figure 20 Fringe contours showing (a) a specific limb movement of a chick embryo during incubation, and (b) the hatching behavior. The point on the shell receiving blows is clearly visible.

corresponding to slope change along x-direction, is shown in Figure 16b.

Examples of Applications

Holographic interferometry has firmly established itself as an attractive and useful class of technique in nondestructive testing. Defects in a specimen are made apparent by the presence of anomalies in the interference fringe distribution of the deformed state of the specimen. The presence of flaws such as voids, disbonds, delaminations, and cracks weakens the strength of the specimen. The resulting strain concentration areas produced by the presence of flaws in the specimen are detected through their effect on the surface deformation map. The flaws in the specimen manifest themselves as distortion of the fringes, pockets of fringe concentrations, or localized discontinuities of the fringes in the full-field display of the deformation field of a test specimen in response to the applied load.

Figure 17 shows phase distribution corresponding to in-plane displacements of a notched fiber-reinforced concrete specimen subjected to a wedge-splitting test. The application is taken from the study of a fracture process zone in fiber-reinforced concrete. Another example of the application of holography in nondestructive testing is shown in Figure 18. The fringe pattern corresponds to in-plane displacements on a fingertip joint wood specimen under tension. Figure 19 shows an example of an application of holography in the tyre industry. The interferogram reveals, through the changes in the fringe pattern, defects in the tyre. The picture in Figure 20a is taken from the study of the embryonic behavior of embryos during incubation. The photograph shows a specific limb movement near the air space. The method also allows for a lively display of the hatching behavior of the chick. The point on the shell receiving blows is clearly visible in Figure 20b.

Conclusions

The aim of this article is to introduce readers to holographic techniques available for the noninvasive and whole-field investigation of the time-evolution and spatial distribution of the displacement and refractive index fields. Diverse techniques to obtain contourgrams of three-dimensional objects are also presented. The method offers strong potential in nondestructive testing applications, such as in the detection of flaws, in the conception of prototypes, in the evaluation of the performances of structural components, and in the verification of the reliability of finite-element computer codes.

See also

Holography, Techniques: Digital Holography; Overview. **Interferometry:** Overview; Phase Measurement Interferometry.

Further Reading

Erf RK (1974) *Holographic Non-Destructive Testing*. New York: Academic.

Ostrovsky YI, Butusov MM and Ostrovskaya GV (1980) *Interferometry by Holography*. Berlin: Springer-Verlag.

Ostrovsky YI, Schepinov VP and Yakovlev VV (1991) *Holographic Interferometry in Experimental Analysis*. Berlin: Springer-Verlag.

Rastogi PK (1994) *Holographic Interferometry – Principles and Methods*. Berlin: Springer-Verlag.

Schumann W and Dubas M (1979) *Holographic Interferometry*. Berlin: Springer-Verlag.

Schumann W, Zürcher JP and Cuche D (1985) *Holography and Deformation Analysis*. Berlin: Springer-Verlag.

Vest CM (1979) *Holographic Interferometry*. New York: Interscience.

Sandwich Holography and Light in Flight

N Abramson, Royal Institute of Technology, Stockholm, Sweden

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction to Holographic Interferometry

Holography is a photographic principle that produces 3D images that are so real that it is possible to take measurements inside this image with a precision of a fraction of a thousandth of a millimeter. Contours are formed within the holographic image that, like the relief lines of a map, indicate the displacements of every point on the object surface. The method is named holographic interferometry (*see* **Holography, Techniques:** Holographic Interferometry) and it can be used to measure deformation, vibration and also, with slightly lower resolution, dimension.

In real time holography the deformed object is directly compared to the holographic image of its undeformed state. In double exposure holography the comparison is made between two holographic recordings on the same plate, while in sandwich holography the two states are recorded on two different plates that are sandwiched together for comparison.

Sandwich holography is useful in industrial environments, because it can measure much larger deformations and also compensate for unwanted movement of the object or the setup between the two exposures. Further on in the process it will produce any sign of deformation and comparison can be made at any time without referring back to original setup. However, processing of the two plates is needed before the comparison can be made and the plates have to be positioned with high accuracy.

Interferometric Sensitivity

To calculate the interferometric sensitivity let us name (A) the point from which the laser light is emitted, (B) the point behind the hologram plate from which the observation is made, and (C) the displaced point on the object surface to be measured. Draw an ellipsoid through (C) with (A) and (B) as focal points. If (C) is displaced along the surface of the ellipsoid, the path length ACB will remain unchanged and no interference fringes are formed, while a displacement normal to this surface produces maximal interference. A diagram, the holodiagram, is based on these ellipsoids and used for the making and evaluation of holograms.

The bisector to ACB will be the normal to the ellipsoids. If the path length ACB is repeatedly increased by δL a diagram like Figure 1 is formed where the separation (s) of the ellipses, and thus the interferometric sensitivity, varies over the diagram. If δL is the wavelength of the laser light, one interference fringe is formed each time an object point has moved the distance (s) because of a deformation of the object between the two exposures. The displacement (d) can be calculated:

$$d = ns \qquad [1]$$

$$s = k0.5\delta L \qquad [2]$$

$$k = 1/\cos\alpha \qquad [3]$$

where n is number of fringes between fixed and studied point, k is constant along arcs of circles through A and B, and α is half the angle ACB.

When two slightly dissimilar patterns are placed one on top of the other, a moire pattern is formed that shows up the difference. The moire analogy to holographic interferometry is based on the fact that one fringe is formed for every ellipse that is crossed by an object point displaced between the two exposures.

If δL is the wavelength (λ) of the laser light the displacement (d) is

$$d = nk0.5\lambda \qquad [4]$$

If two-wavelength holography is used to measure the 3D shape of an undisturbed object, it will be seen intersected by ellipses where the separation (s) is

$$s = k0.5\lambda_1\lambda_2/(\lambda_1 - \lambda_2) \qquad [5]$$

If δL is the coherence length or the pulse length of the laser source, the object will be seen intersected by one single bright fringe, the thickness (s) of which is

$$s = k\delta L \qquad [6]$$

Everything outside that ellipse will be dark during reconstruction. This phenomenon can be used either to optimize the use of a limited coherence length or to explain the results of 'light-in-flight recordings', as described later.

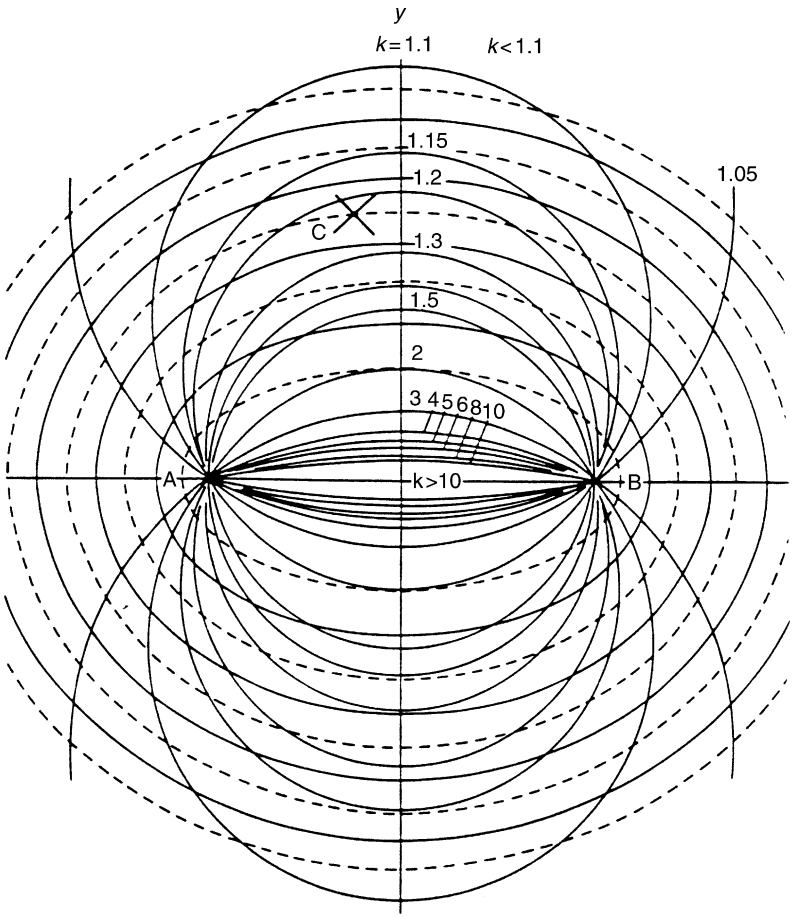


Figure 1 The holodiagram used for the evaluation of holograms.

The Principle of Sandwich Holography

In the following discussion on sandwich holography, we will only study the simplified example where the directions of illumination and observation are normal to the studied displacement and therefore α is equal to zero. The reflected light from an object point combined with the reference light produces a set of hyperbolas (Young's fringes) on the hologram plate.

When the object point is displaced between exposures on two different hologram plates, the hyperbolas will be positioned differently on the plates. There is a one-to-one relationship between the ellipsoids and the hyperbolas. Thus one point on the hologram plate will be crossed by one hyperbola if the corresponding object point is moved so that it crosses one of the ellipses of the holodiagram. Thus, the moire effect of the displaced hyperbolas at the hologram plate represents the moire effect of the ellipses in the object space. The interference pattern in the reconstructed object image can therefore be manipulated by moving one plate in relation to the other during reconstruction.

We have found that the simplest way to do this is to glue the two plates together separated by the glass thickness of one plate. To make the situation identical during exposure and reconstruction, the plates also have to be recorded in pairs. Thus, the first exposure can be made with one front plate and one back plate in contact in the hologram holder (Figure 2). The second exposure is made in exactly the same position with two new plates. After processing the front plate of the second exposure is glued in front of the back plate of the first exposure.

Then a reconstruction beam similar to the reference beam, is used to reconstruct the sandwich hologram. The interference fringes seen on the image of the object are manipulated by tilting the sandwich hologram, which results in an addition or subtraction of new fringes as in the following examples:

1. If the object is rotated by the small angle ϕ_1, the number of fringes can, during reconstruction, be

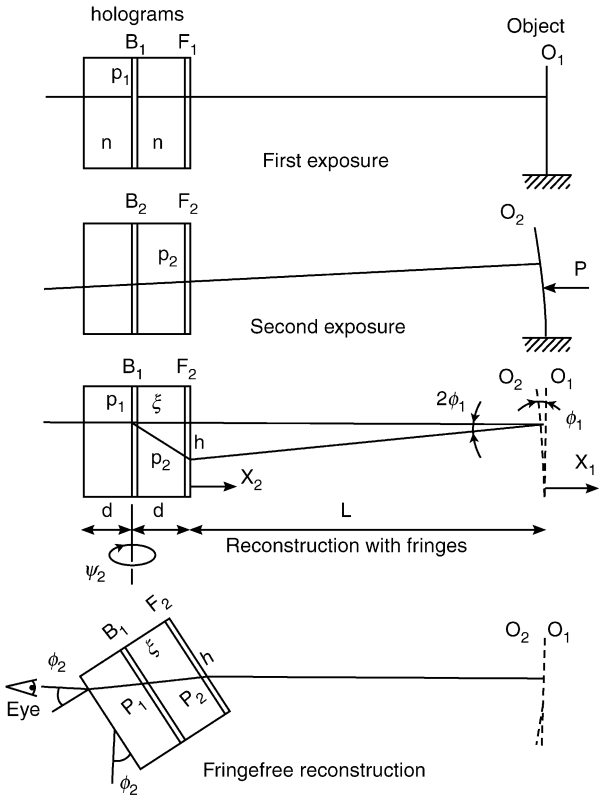


Figure 2 The exposures and reconstructions of a sandwich hologram.

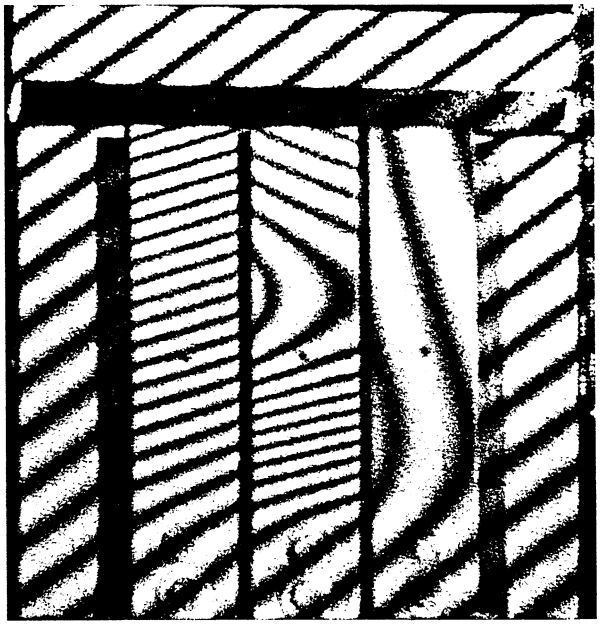


Figure 3 The movements of three steel bars in relation to a frame are to be measured by the fringes of sandwich holography. However, the unwanted motion of the total set-up makes the evaluation impossible.

eliminated by rotating the sandwich in the same direction by the much larger angle ϕ_2.

$$\phi_1 = \left[\frac{d}{2Ln}\right]\phi_2 \qquad [7]$$

2. If the sandwich during reconstruction is rotated around an axis perpendicular to that of the object rotation, new fringes are introduced at the angle γ, which is analogous to, but much larger than, the object rotation ϕ_1.

$$\phi_1 = \left[\frac{d}{2Ln}\right]\varphi_2 tg\gamma \qquad [8]$$

3. When the sandwich is rotated as described in (2) fringes are formed, the radius of curvature (r_2) of which are much larger than those of the bent object (r_1).

$$r_1 = \left[\frac{2Ln}{d}\right]\left[\frac{1}{\varphi_2}\right]r_2 \qquad [9]$$

Examples

The object consisted of three vertical steel bars (Figure 3) that were fixed by screws at their lower end to a rigid frame that surrounded all the bars. The middle bar was also supported at the upper end. Forces were applied at the middle of the bars. The right bar is deformed away from the observer while the other two are deformed towards him. After deformation of the three objects the second pair of plates were exposed. Then all four plates were processed.

The back plate of the first exposure (B_1 in Figure 2) was repositioned in the plate holder in front of the plate of the second exposure (F_2). The two plates were bonded together to form a sandwich hologram. O_1 and O_2 represent the two positions of one object, which between the two exposures was bent at angle ϕ_1 toward the hologram. (L) is the distance between the front hologram and the object (~1 m); (L) is also the distance between the point of illumination and the object. (D) is the distance between the point source of the reference beam and the hologram plate ($D = 2L$); (d) is the thickness of one hologram plate (~1.3 mm); (n) is the refractive index of the plates (~1.5 mm).

When the two plates were properly repositioned in the plate holder, fringes were seen everywhere, both on the steel bars and on the supposedly fixed frame which apparently had made an accidental movement (Figure 3). By tilting the plate holder in different directions, the fringes on the frame could be eliminated and the hologram could be evaluated as if the frame had remained fixed (Figure 4).

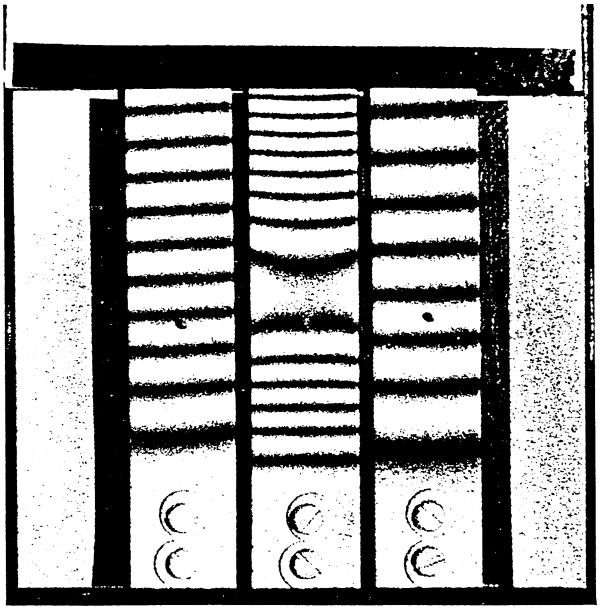


Figure 4 The unwanted fringes caused by the motion of the total set-up have been eliminated by a tilt of the sandwich hologram.

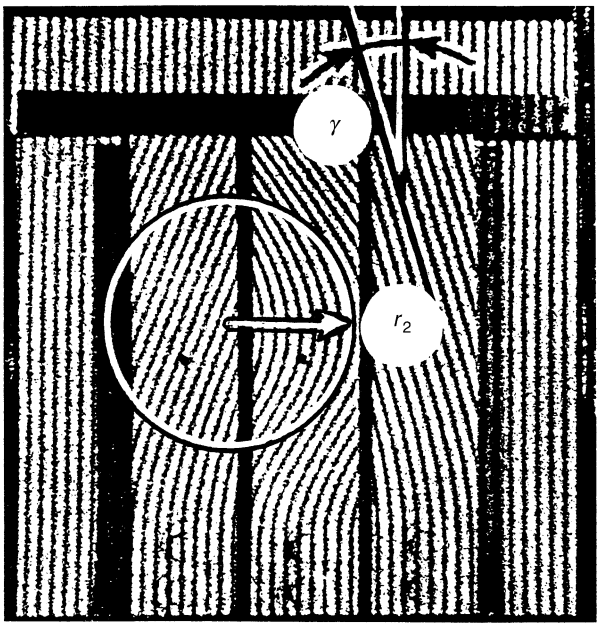


Figure 6 A rotation of the sandwich hologram around a vertical axis produces fringes that reveal direction of tilt and bending radius of all steel bars.

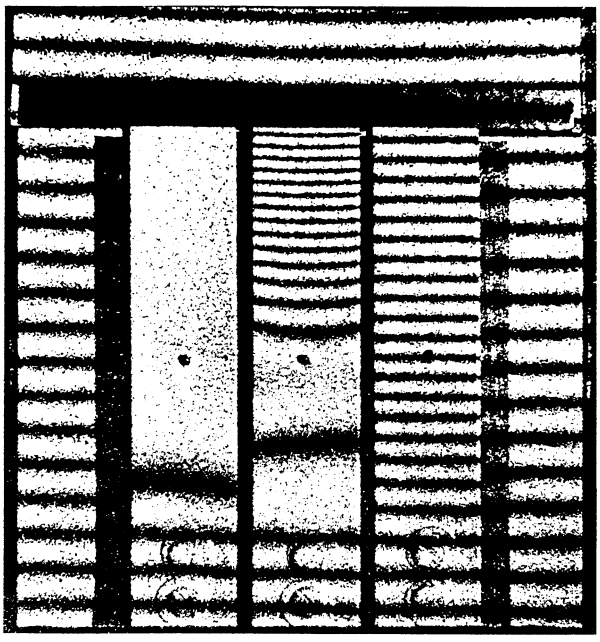


Figure 5 Another tilt of the sandwich hologram reveals that the left and the right bar have been deformed in opposite directions.

The plate holder was then tilted backwards at angle ϕ_2 and horizontal fringes were found on the frame while the top of the left bar became fringe-free (Figure 5). The number of fringes on the right bar increased, proving that it had been tilted in the opposite direction. If the sandwich were instead rotated around a vertical axis at angle φ_2, a set of vertical fringes were formed on the frame, while the bars became covered by fringes, the inclination and curvature of which represent a magnified view of the deformed bars as seen in profile (Figure 6).

Introduction to Light-in-flight Recording by Holography

Recording of light pulses in flight have been possible since the beginning of the 1990s, using two-photon fluorescence or ultrafast Kerr cells driven by laser pulses. Very fast optical phenomena, such as refractive-index changes in laser-produced plasmas, have been recorded by holography using short illumination pulses.

In 1968, Staselko *et al.* published a method of studying the time coherence pattern of wave trains from pulsed lasers by the use of autocorrelation in which the unknown pulse is compared to itself. The method we name 'light-in-flight recording by holography' (LIF) is based on studying both the temporal and the spatial shape of pulses by correlation, which means that we study the change in shape of a pulse before and after it has passed through the transformation we want to study. For this process, we want to start with as short a pulse (or short coherence length) as possible and then compare the transformed pulse to the original pulse.

In a hologram only those parts of an object will be recorded for which the path length from the laser to the holographic plate via the object does not differ from the path length of the reference beam by more than the pulse length (or coherence length) of the laser light used for the recording. If that length is short, during reconstruction a large object will be seen intersected by an imaginary surface in the form of an ellipsoid, one of its two foci being the point from which the spherical wavefronts of illumination are emitted (spatial filter (A) in Figure 1) and the other being the point of observation (B). A spherical wavefront appears distorted into an ellipsoid if it is divergent, and into a hyperboloid if it is convergent, while a flat wavefront appears distorted into a paraboloid. The general rule is that the apparent wavefront has the shape of a mirror that focuses light from (A) to (B).

By careful planning, it is possible to optimize the geometry of the holographic setup in such a way that the utilized portions of the ellipsoids and hyperboloids can be approximated into flat surfaces. Thus, the object is seen intersected by one flat imaginary interference surface, the depth of which can be altered during reconstruction by changing the point of observation at the hologram plate.

Light of short duration, or short coherence length, can, in this way, be used in a holographic method for finding depth contours. If, however, the 3D shape of the object is known a priori, e.g., if it is a flat surface, this method can just as well be used to study the 3D shape of the illuminating wavefront. As mentioned before, the intersections of the ellipsoids can be approximated by intersections of spherical wave fronts of illumination, if the distance between object and hologram plate is long and if the flat surface is perpendicular to the line of sight.

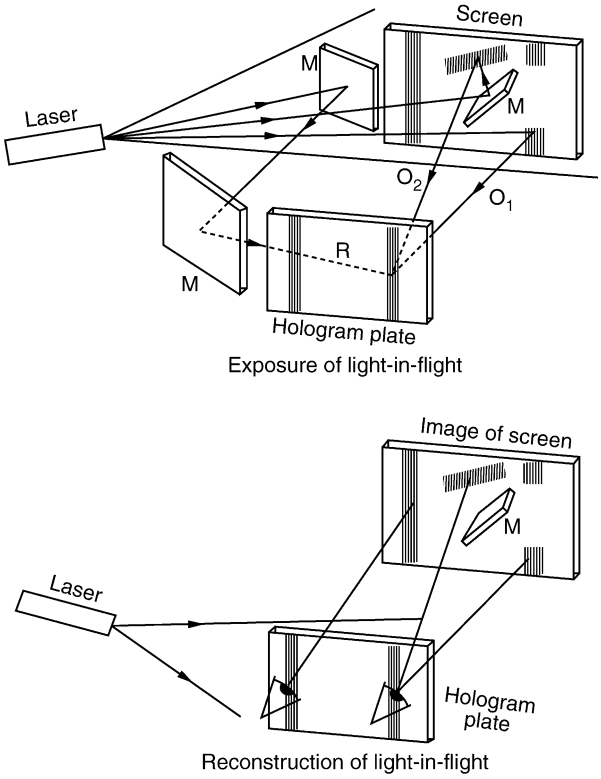


Figure 7 During exposure of the light-in-flight hologram, the short pulse from the laser is used to obliquely illuminate both the object and the hologram plate. Only those parts will be recorded where path length for object light and reference light are equal.

Examples of Wavefront Studies

As an example we will describe an experiment that consists of an opaque white diffusely reflecting flat screen, that is illuminated from the left at an oblique angle, as seen in Figure 7. A mirror is fixed to the screen so that some of the light is reflected upwards. During recording of the hologram, the reference pulse is reflected by the two mirrors (M) so that it arrives at the middle of the hologram plate at the same time as the object pulse, in the form of diffuse light, arrives from the middle of the screen. As the reference pulse passes over the hologram plate from left to right, it works like a curtain shutter (focal-plane shutter) recording early object light to the left and later light to the right. This 'light shutter' moves across the plate with a velocity faster than light which is possible as it is just an intersection between pulsefront and plate. The difference in time of the recordings made at the two ends of the hologram plate is the time it takes for the reference pulse to move across the plate, or about 800 picoseconds. During reconstruction with a continuous laser the early arriving object light is seen through the left part of the plate, while the later arriving light is seen through its right part. If the eye, or a TV camera scans from left to right, the pulse is seen moving continuously over the object screen, with some of it reflected upwards by the mirror (Figure 8a–d). The corresponding pulse length is 100 ps (~30 mm), and to minimize pulsefront distortions the screen is perpendicular to the line of sight.

The photo in Figure 9 shows the result of another experiment. A spherical wave expands from a point source at left and moves to the right. The spherical wavefront intersects a flat observation screen consisting of a white-coated aluminum plate of 20 × 30 cm. A cylindrical lens is fixed, with its axis normal to the screen, so that it focuses the arriving light almost parallel to the screen surface. Figure 8 is a composite of several photographs taken through different parts

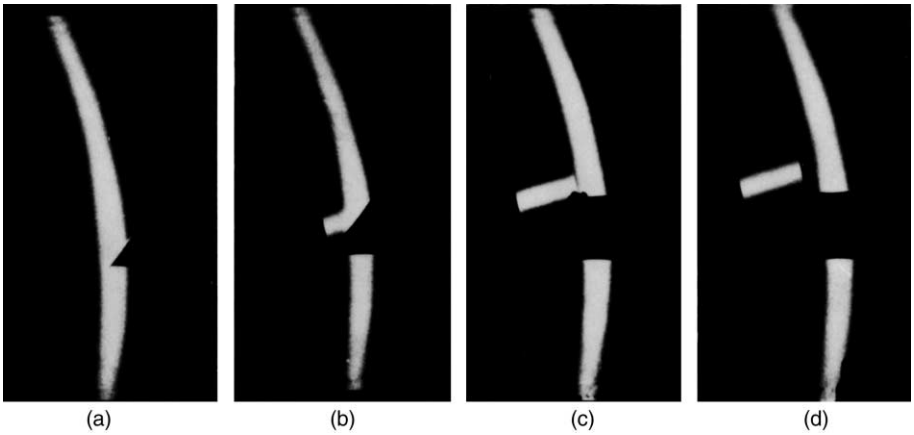


Figure 8 Resulting reconstructed images of the light reflected by a mirror at about 45°. These four photos were taken during reconstruction while the camera behind the hologram plate was moved from left to right in between each of the exposures.

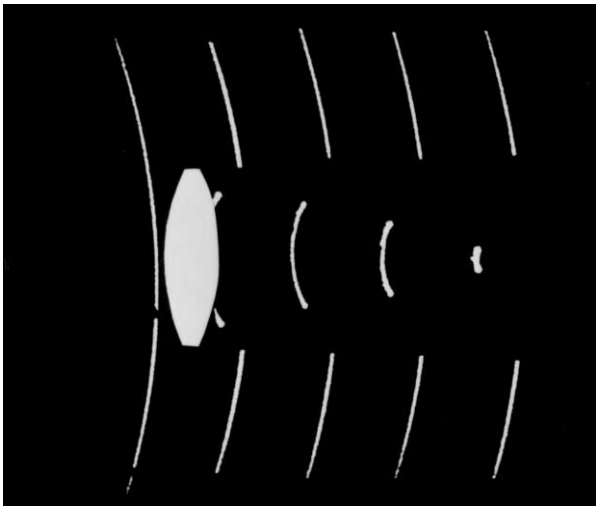


Figure 9 Light focused by a lens. A composite of several photos taken from different parts of the hologram plate. An image of the lens is included to make the picture clear.

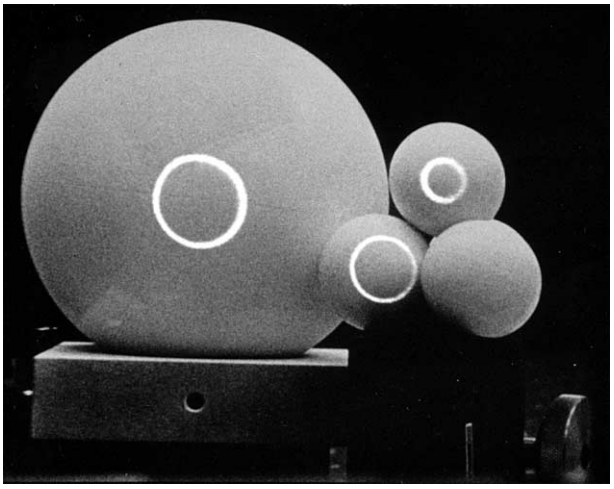


Figure 10 Spheres intersected by a thin sheet of light.

of the hologram plate. The pulse length was in this case 10 ps (~3 mm).

Measuring the Shape of 3D Objects

The wavefront (pulse front) of a short pulse forms a thin sheet of light. The intersection of this sheet by a 3D body produces on its surface a thin line representing a cross-section. It is advantageous if this cross-section represents a flat surface, which will be approximately the case if the illumination and observation beams are collimated or arrive from points at large distances.

By moving the point of observation along the hologram plate from left to right, the intersecting sheet of light can be moved in depth forward or backward. Determination of fringe order number (by finding the zero fringe) no longer represents any problem because only the zero order fringe exists. Therefore, this way of contouring has great advantages, especially when the observations and calculations are made by computer. Figure 10 shows a set of spheres intersected by a thin light sheet.

The intersecting light sheet can be manipulated in yet another way, i.e., it can be rotated. If the point of observation is moved away from the plate, the right part of an object will be seen through the right part of the plate were the recording made later than at the left part. During the time interval between left and right records, the intersecting light sheet has moved so that it intersects the object further away, resulting in a tilted cross-section of the object. This way it has been possible to rotate the cross-section up to 90 degrees, so that if, for example, a face is illuminated from the

front by a flat sheet of light, the reconstructed intersection will show the face in profile.

In a limited way it is also possible to produce results similar to those of LIF by using a CCD camera instead of the holographic plate. The great advantage is its compatibility with a PC, while the main disadvantage is a lower resolution.

Other Uses

Like all methods of 'gated viewing', LIF can be used to look through scattering media, e.g., to look through human tissue. By observing only the first light the influence of the later arriving scattered light is eliminated. Thus, LIF can be used as a diagnostic method based on the idea that the scattering, absorption, or refractive index of cancer is different from that of healthy tissue. Another use of LIF is to image just a thin section of a gas or liquid filled with particles. In this way it is possible to observe the distribution and velocity in a certain localization without influence of the scattering from particles in front of or behind the studied section.

Up to now we have discussed gated viewing by LIF in the laboratory using picosecond light pulses and distances of up to a few meters. Instead we could look at astronomical objects with light pulses of days (e.g., a supernova) and distances of many light years. Even in these cases we will see particles in space illuminated by ellipsoids with (A) being the supernova and (B) a telescope on Earth. While each ellipsoid in the laboratory may represent a pathlength of some nanoseconds, each ellipsoid in space might represent light years.

However, the methods we have shown here are not limited to the study of the apparent shape of spheres of light. They can also be used to study other objects moving at ultrahigh velocities. If an exploding supernova throws out particles in all directions at a rather slow velocity, it can be seen surrounded by an expanding sphere. If, instead, this sphere expanded with the speed of light (which is not possible) it would appear to us just like an ellipsoid of the holodiagram. If, however, it expanded at a very high velocity, but lower than that of light, the sphere would appear to us in the shape of an egg with its pointed end towards the observer. The velocity of expansion could, in this case, appear superluminal (higher than the speed of light). If the velocity is (v), its approaching velocity will to a first approximation appear to be the true velocity divided by ($1 - v/c$).

Finally, we will show a relationship between LIF and special relativity, by looking at a modification of the 'Minkowski light cones'. Figure 11 shows a diagram

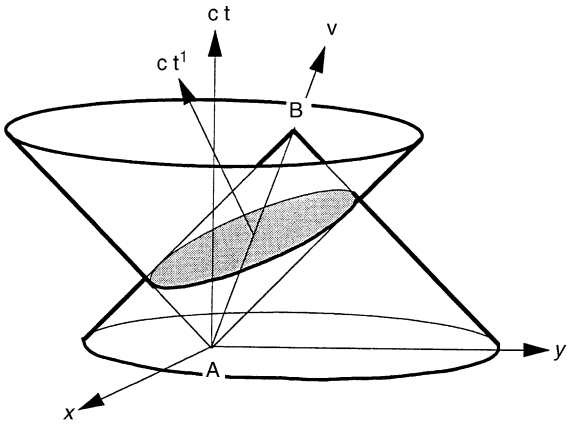


Figure 11 A modified Minkowski diagram consisting of one cone of illumination (A) and one cone of observation (B). The separation in time and space between (A) and (B) could either be static as in LIF or caused by a relativistic velocity of the observer from (A) to (B).

where (x) and (y) represent space coordinates, while (z) represents time multiplied by (c). A sphere from a short pulse of light that expands from the point A is represented by a cone with its apex at (A). A sphere from a short observation that converges to (B) is represented by a cone with its apex at (B). Only those object points, that are positioned at the elliptic intersection of the two cones, can be seen illuminated. This ellipse is identical to one of the ellipses of the holodiagram. The distance in time and space separating (A) and (B) could be either caused by the fixed distance between source of laser light and hologram plate in holographic LIF, or by a high velocity of the observer. Thus, the Lorentz contraction in special relativity could be explained as a measuring error because the high velocity of the observer traveling from (A) to (B) causes his sphere of observation to be transformed into an ellipsoid of observation. The tilt of the elliptic intersection indicates which observations that appear to be simultaneous depending on the separation of (A) and (B) along the y-axis, both in LIF and in relativity.

See also

Holography, Techniques: Holographic Interferometry.

Further Reading

Abramson N (1971) Moiré patterns and hologram interferometry. *Nature* 231: 65–67.

Abramson N (1981, 1986) *The Making and Evaluation of Holograms*. London: Academic Press.

Abramson N (1996) *Light in Flight, or, The Holodiagram – The Columbi Egg of Optics*. USA: SPIE Press, vol. PM 27.

An W and Carlsson T (1999) Digital measurement of three-dimensional shapes using light-in-flight speckle interferometry. *Optical Engineering* 38: 1366–1370.

Dugay MA and Mattick AT (1971) Ultrahigh speed photography of picosecond light pulses and echoes. *Applied Optics* 10: 2162–2170.

Garrington S (1992) Invisible jet sheds new light. *Nature* 355: 771–772.

Giordmaine J and Rentzepis P (1967) Two-photon excitation of fluorescence by picosecond light pulses. *Applied Physics Letters* 11: 216.

Hinrichs H and Hinsch KD (1997) Light-in-flight holography for flow visualization and velocimetry in three-dimensional flows. *Optics Letters* 22: 828–830.

Malin D and Allen D (1990) Echoes of the supernova. *Sky and Telescopes* 22–25, January.

Powell R and Stetson K (1965) Interferometric analysis by wavefront reconstruction. *Journal of the Optical Society of America* 55: 1593–1598.

Spears K and Serafin J (1989) Chrono-coherent imaging for medicine. *IEEE Transactions on Medical Engineering* 36: 1210–1220.

Staselko DI, Denisyuk YN and Smirnow AG (1969) Holographic recording of the time-coherence pattern of a wave train from a pulsed laser source. *Optics and Spectroscopy* 26: 413.

IMAGING

Contents

Information Theory in Imaging
Inverse Problems and Computational Imaging
Adaptive Optics
Hyperspectral Imaging
Imaging Through Scattering Media
Infrared Imaging
Interferometric Imaging
Lidar
Multiplex Imaging
Photon Density Wave Imaging
Three-Dimensional Field Transformations
Volume Holographic Imaging
Wavefront Sensors and Control (Imaging Through Turbulence)

Information Theory in Imaging

F O Huck and C L Fales, NASA Langley Research Center, Hampton, VA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Information theory is an intuitively attractive mathematical approach for combining observation with communication, or sensing with processing. So far, however, this approach has remained largely in the realm of signal processing and transmission, as first formalized by Shannon and Wiener over 50 years ago. Recent extensions of this approach to imaging have mostly addressed the interpretation of images, without explicitly accounting for the critical limiting factors that constrain image gathering and restoration. Yet, to be useful in the characterization of imaging, information theory must account for these constraints and lead to a close correlation between predicted and actual performance. The characterization that is outlined here traces the rate of information from the scene to the observer; and it pairs this rate, first, with the theoretical minimum data rate to assess the efficiency with which information can be transmitted, and, second, with the maximum-realizable fidelity to assess the quality with which images can be restored. Further details, especially for the mathematical development of figures of merit and for the design of imaging systems, can be found in the Further Reading.

Model of Imaging

Description

Modern imaging systems usually have the following three stages, as depicted in Figure 1:

1. *Image gathering*, to capture the radiance field that is either reflected or emitted by the scene and transform this field into a digital signal;

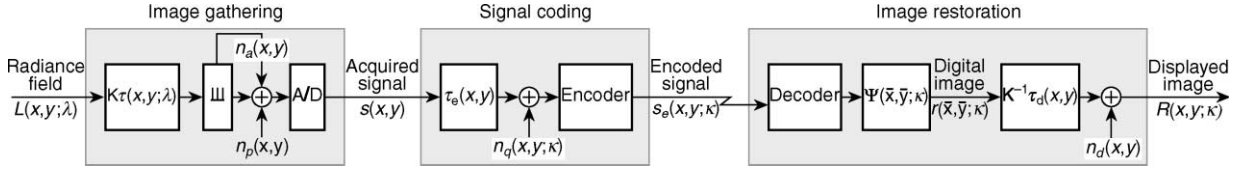


Figure 1 Model of digital imaging systems.

2. *Signal coding*, to encode the acquired signal for the efficient transmission and/or storage of data and decode the received signal for the restoration of images. The encoding may be either lossless (without loss of information) or lossy (with some loss of information), if a higher compression is required;
3. *Image restoration*, to produce a digital representation of the scene and transform this representation into a continuous image for the observer.

Human vision may be included in this model as a fourth stage, to characterize imaging systems in terms of the information rate that the eye of the observer conveys from the displayed image to the higher levels of the brain.

These stages are similar to each other in one respect: each contains one or more transfer functions, either continuous or discrete, followed by one or more sources of noise, also either continuous or discrete. The continuous transfer function of the image-gathering device (and of the human eye) is followed by sampling that transforms the captured radiance field into a discrete signal with analog magnitudes. In most devices the photodetection mechanism conveys this signal serially into an analog-to-digital (A/D) converter to produce a digital signal. However, analog signal processing in a parallel structure, akin to that in human vision, has many advantages that are increasingly emulated by neural networks.

This model of imaging differs from the classical model of communication that Shannon and Wiener addressed in two fundamental ways:

1. *The continuous-to-discrete transformation in image gathering*. Whereas communication is constrained critically only by bandwidth and noise, image gathering is constrained also by the compromise between blurring and aliasing, due to limitations in the response of optical apertures and in the sampling density of photodetection mechanisms. This additional constraint requires the rigorous treatment of insufficiently sampled signals throughout the imaging system.
2. *The sequence of image gathering followed by signal coding*. Whereas the characterization of communication addresses the perturbations that occur at the encoder and decoder or during transmission, the characterization of imaging systems must also address the perturbations in the image-gathering process prior to coding. These additional perturbations require a clear distinction between the information rate of the encoded signal and the associated theoretical minimum data rate or, more generally, between information and entropy.

The mathematical model of imaging that is outlined here accounts for the appropriate deterministic and stochastic (random) properties of natural scenes and for the critical limiting factors that constrain the performance of imaging systems. The corresponding formulations of the figures of merit are sufficiently general to account for a wide range of radiance field properties, photodetection mechanisms, and signal processing structures. However, the illustrations emphasize: (a) the salient properties of natural scenes that dominate the captured radiance field in the visible and near-infrared region under normal daylight conditions; and (b) the photodetector arrays that are used most frequently as the photodetection mechanism of image-gathering devices.

Radiance Field

The radiance field $L(x, y; \lambda)$, reflected by natural scenes, may be modeled approximately as

$$L(x, y; \lambda) = \frac{1}{\pi} I(\lambda)\rho(\lambda)\rho(x, y) \qquad [1]$$

where $I(\lambda)$ and $\rho(\lambda)$, respectively, represent the spectral irradiance and reflectance properties as a function of the wavelength λ, and $\rho(x, y)$ represents the random reflectance properties as a function of the spatial coordinates (x, y). Figure 2 illustrates a model of $\rho(x, y)$ for which the power spectral density (PSD) $\hat{\Phi}_\rho(\upsilon, \omega)$ corresponds closely to that typical of natural scenes, as given by

$$\hat{\Phi}_\rho(\upsilon, \omega) = \frac{2\pi\mu_\rho^2\sigma_\rho^2}{[1 + (2\pi\mu_\rho\zeta)^2]^{3/2}} \qquad [2]$$

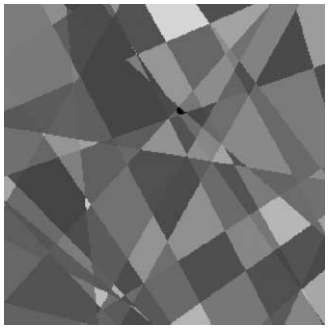


Figure 2 Model of reflectance $\rho(x, y)$.

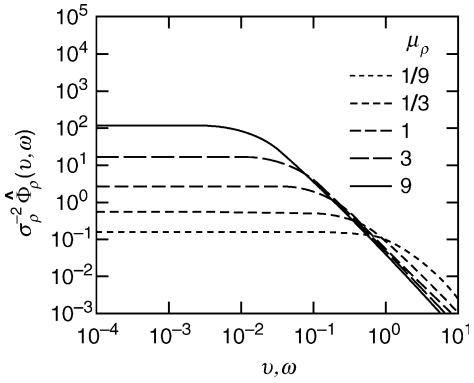


Figure 3 Normalized PSD of the reflectance $\rho(x, y)$.

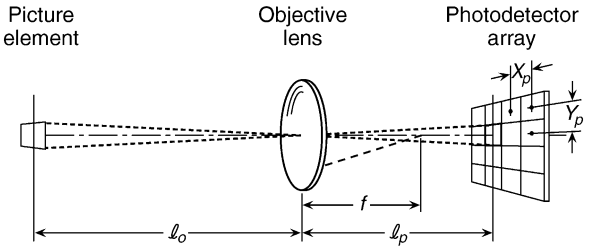


Figure 4 Optical geometry of image-gathering device.

where $\zeta^2 = \upsilon^2 + \omega^2$ and (υ, ω) are the spatial frequency coordinates of the Fourier-transform domain (see Figure 3). This model consists of random polygons, whose boundaries are distributed according to Poisson probability with a mean separation of μ_ρ, and whose magnitudes are distributed according to independent zero-mean Gaussian statistics of variance σ_ρ^2. The constraints associated with eqn [1] are satisfied by letting the mean reflectance $\bar{\rho} = 0.5$ and the standard deviation $\sigma_\rho \leq 0.2$.

Image Gathering

Figure 4 depicts the basic optical geometry of an image-gathering device, in which the objective lens forms an image of the captured radiance field across the photodetector array, and the array transforms this

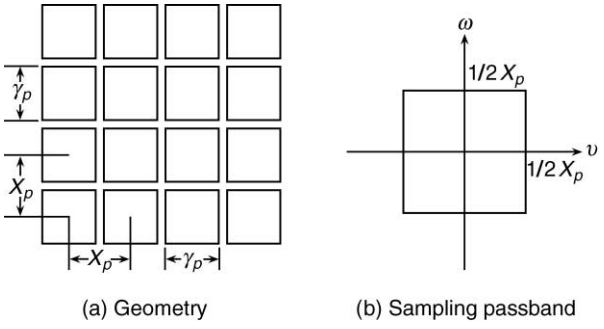


Figure 5 Photodetector-array geometry and sampling passband.

image into a discrete signal. The sampling lattice (X, Y), which determines the resolution of this device, is formed in angular units by the separation (X_p, Y_p) of the photodetector-array apertures and the distance ℓ_p as given for one of the dimensions by

$$X = 2\tan^{-1}\frac{X_p}{2\ell_p} \approx \frac{X_p}{\ell_p}\text{rad} = \frac{57.3X_p}{\ell_p}\text{deg} \qquad [3]$$

Normally, the sampling lattice is square, as shown in Figure 5, so that $X_p = Y_p$. Then the corresponding area of the photodetector apertures is $A_p = \gamma_p^2$, and the solid angle Ω of the instantaneous field of view (IFOV) is $\Omega \approx A_p/\ell_p^2$ ster.

The sampling lattice projected onto the scene at the distance ℓ_o has the dimensions $(X_o, Y_o) \approx (X\ell_o, Y\ell_o)$. Because these dimensions represent the finest spatial detail that normally can be restored, it is convenient to measure the mean spatial detail of the scene relative to the sampling lattice, as given by $\mu = \mu_\rho(X_o Y_o)^{-1/2}$, which reduces to $\mu = \mu_\rho X_o^{-1}$ for $X = Y$. Hence, $\mu = 1$ indicates that the sampling interval is equal to the mean spatial detail.

The image-gathering process that transforms the continuous radiance field $L(x, y; \lambda)$ into the discrete signal $s(\mathbf{x}, \mathbf{y})$ may be modeled by the expression

$$s(\mathbf{x}, \mathbf{y}) = K\rho(\mathbf{x}, \mathbf{y}) * \tau(\mathbf{x}, \mathbf{y}) + n_p(\mathbf{x}, \mathbf{y}) \qquad [4]$$

where the discrete coordinates $(\mathbf{x}, \mathbf{y}) = (mX, nY)$, and the symbol $*$ denotes continuous spatial convolution. This model accounts, in the simplest form, for the reflectance-to-signal conversion gain K, the reflectance $\rho(x, y)$ of the scene, the spatial response (SR) $\tau(x, y)$ of the image-gathering device, the continuous-to-discrete transformation by the photodetector-array lattice, and the discrete noise $n_p(\mathbf{x}, \mathbf{y})$ of the photodetectors. The discrete coordinates are

defined by the sampling function

$$\underline{\underline{\amalg}} \equiv \underline{\underline{\amalg}}(\mathbf{x}, \mathbf{y}) = XY \sum_{\mathbf{x},\mathbf{y}} \delta(x - \mathbf{x}, y - \mathbf{y}) \qquad [5]$$

where $\delta(x, y)$ is the Dirac delta function; and the gain is given by

$$K = \frac{1}{\pi} A_\ell \Omega \int I(\lambda)\rho(\lambda)r(\lambda)\mathrm{d}\lambda \qquad [6]$$

where $A_\ell = \pi D^2/4$ is the area of the objective lens with diameter D, and $r(\lambda)$ is the responsivity of the photodetectors. This gain relates the magnitude of the signal $s(\mathbf{x}, \mathbf{y})$ directly to that of the reflectance $\rho(x, y)$; and, together with the variances σ_ρ^2 and σ_p^2 of the reflectance and the photodetector noise, respectively, it also determines the rms-signal-to-rms-noise ratio (SNR) $K\sigma_\rho/\sigma_p$.

The DFT of $s(\mathbf{x}, \mathbf{y})$ is the periodic spectrum $\tilde{s}(v, \omega)$ given by

$$\tilde{s}(v, \omega) = K\hat{\rho}(v, \omega)\hat{\tau}(v, \omega) * \hat{\underline{\underline{\amalg}}} + \tilde{n}_p(v, \omega) \qquad [7]$$

where $\hat{\rho}(v, \omega)$ is the spectrum of the reflectance, $\hat{\tau}(v, \omega)$ is the spatial frequency response (SFR) of the image-gathering device, and $\tilde{n}_p(v, \omega)$ is the spectrum of the photodetector noise. The symbol $\hat{\underline{\underline{\amalg}}}$ is the Fourier transform of the sampling function $\underline{\underline{\amalg}}$. It accounts for the sidebands generated by the sampling process and is given by

$$\hat{\underline{\underline{\amalg}}} \equiv \hat{\underline{\underline{\amalg}}}(v, \omega) = \sum_{\mathbf{v},\mathbf{w}} \delta(v - \mathbf{v}, \omega - \mathbf{w}) \qquad [8]$$

where $(\mathbf{v}, \mathbf{w}) = (m/X, n/Y)$. The corresponding sampling passband is

$$\hat{B} = \left\{(v, \omega);\ |v| < \frac{1}{2X},\ |\omega| < \frac{1}{2Y}\right\} \qquad [9]$$

The above expression for $\tilde{s}(v, \omega)$ can be rewritten more conveniently as

$$\tilde{s}(v, \omega) = K\hat{\rho}(v, \omega)\hat{\tau}(v, \omega) + \hat{n}(v, \omega) \qquad [10]$$

where

$$\hat{n}(v, \omega) = \hat{n}_\mathrm{a}(v, \omega) + \tilde{n}_p(v, \omega) \qquad [11]$$

is the spectrum of the total noise in the acquired signal. The first term $\hat{n}_\mathrm{a}(v, \omega)$ represents the aliased signal components caused by insufficient sampling, as given by

$$\hat{n}_\mathrm{a}(v, \omega) = K\hat{\rho}(v, \omega)\hat{\tau}(v, \omega) * \hat{\underline{\underline{\amalg}}}_\mathrm{s} \qquad [12]$$

where

$$\hat{\underline{\underline{\amalg}}}_\mathrm{s} \equiv \hat{\underline{\underline{\amalg}}}_\mathrm{s}(v, \omega) = \sum_{\mathbf{v},\mathbf{w}\neq(0,0)} \delta(v - \mathbf{v}, \omega - \mathbf{w}) \qquad [13]$$

If, for spatial coordinates at either the scene or the photodetector array, the unit of (x, y) is meters, then the unit of (v, ω) is cycles m^{-1}; and if, for angular coordinates centered at the objective lens, the unit of (x, y) is radians, then the unit of (v, ω) is cycles rad^{-1}. Either way, it is convenient to normalize the sampling intervals to unity (i.e., $X = Y = 1$). Then the area of the sampling passband is $|\hat{B}| = 1$, the unit of (x, y) is samples, and the unit of (v, ω) is cycles sample^{-1}.

The SFR $\hat{\tau}(v, \omega)$ of the image-gathering device is

$$\hat{\tau}(v, \omega) = \hat{\tau}_\ell(v, \omega)\hat{\tau}_p(v, \omega) \qquad [14]$$

where $\hat{\tau}_\ell(v, \omega)$ and $\hat{\tau}_p(v, \omega)$ are the SFRs of the objective lens and photodetector apertures, respectively. This SFR may be modeled approximately by the circularly symmetric Gaussian shape

$$\hat{\tau}(v, \omega) = \exp[-(\zeta/\zeta_\mathrm{c})^2] \qquad [15]$$

where, as shown in Figure 6, the optical-response index ζ_c controls the trade-off between the blurring that is caused by the decrease of the SFR inside the sampling passband and the aliasing that is caused by the extension of the SFR beyond this passband.

Signal Coding

Signal coding may be divided into two stages, as depicted in Figure 1. The first stage, which is represented by the digital operator $\tau_\mathrm{e}(\mathbf{x}, \mathbf{y})$ and the quantization noise $n_q(\mathbf{x}, \mathbf{y}; \kappa)$, accounts for manipulations of the acquired signal that seek to extract those

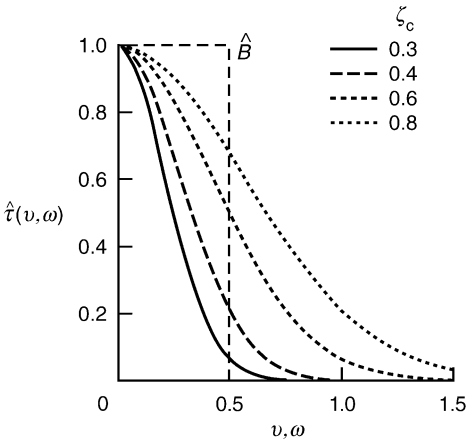


Figure 6 SFRs $\hat{\tau}(v, \omega)$ of the image-gathering device relative to the sampling passband $\hat{B}$ for unit sampling intervals (i.e., $X = Y = 1$).

features of the captured radiance field considered to be perceptually most significant, which usually entails some loss of information. The second stage, which is represented by the encoder, accounts for manipulations that seek to reduce redundancies in the acquired signal, without loss of information. The resultant encoded signal $s_e(\mathbf{x}, \mathbf{y}; \kappa)$ may be expressed in the same form as eqn [4] for the acquired signal $s(\mathbf{x}, \mathbf{y})$ as

$$s_e(\mathbf{x}, \mathbf{y}; \kappa) = s(\mathbf{x}, \mathbf{y}) \otimes \tau_e(\mathbf{x}, \mathbf{y}) + n_q(\mathbf{x}, \mathbf{y}; \kappa) \qquad [16]$$

where the symbol $\otimes$ denotes discrete spatial convolution, and $n_q(\mathbf{x}, \mathbf{y}; \kappa)$ is the quantization noise. If the data transmission and/or storage are error-free, as is assumed here, then the signal that leaves the decoder remains identical to the signal that enters the encoder.

The DFT of $s_e(\mathbf{x}, \mathbf{y}; \kappa)$ is the periodic spectrum $\tilde{s}_e(\upsilon, \omega; \kappa)$ given in the same form as eqn [10] by

$$\tilde{s}_e(\upsilon, \omega; \kappa) = K\hat{\rho}(\upsilon, \omega)\hat{\Gamma}_e(\upsilon, \omega) + \hat{n}_e(\upsilon, \omega; \kappa) \qquad [17]$$

where

$$\hat{\Gamma}_e(\upsilon, \omega) = \hat{\tau}(\upsilon, \omega)\tilde{\tau}_e(\upsilon, \omega) \qquad [18]$$

is the throughput SFR from image gathering to transmission, and

$$\hat{n}_e(\upsilon, \omega; \kappa) = \hat{n}(\upsilon, \omega)\tilde{\tau}_e(\upsilon, \omega) + \tilde{n}_q(\upsilon, \omega; \kappa) \qquad [19]$$

is the accumulated noise in the encoded signal. For η-bit quantization, where $\eta = \log \kappa$ and κ is the number of uniformly spaced quantization levels, $\tilde{n}_q(\upsilon, \omega; \kappa)$ may be treated (for $\eta \geq 4$) as independent Gaussian noise with the PSD

$$\tilde{\Phi}_q(\upsilon, \omega; \kappa) = \sigma_q^2 = (\sigma_{s_e}/\kappa)^2 \qquad [20]$$

where $\sigma_{s_e}^2$ is the variance given approximately by

$$\sigma_{s_e}^2 \approx K^2 \int_{\hat{B}} \int \hat{\Phi}_\rho(\upsilon, \omega)|\hat{\Gamma}_e(\upsilon, \omega)|^2 \mathrm{d}\upsilon \mathrm{d}\omega \qquad [21]$$

Image Restoration

Image restoration, as depicted in Figure 1, transforms the decoded signal $s_e(\mathbf{x}, \mathbf{y}; \kappa)$ into the digital image $r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa)$, as modeled by the expression

$$r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa) = \overline{\overline{\mathbb{III}}} s_e(\mathbf{x}, \mathbf{y}; \kappa) \otimes \Psi(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa) \qquad [22]$$

The operation indicated by $\overline{\overline{\mathbb{III}}} s_e(\mathbf{x}, \mathbf{y}; \kappa)$ (called zero padding) prepares the decoded signal for image restoration with interpolation, and the operator $\Psi(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa)$ produces the digital image $r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa)$ on the interpolation lattice $\overline{\overline{\mathbb{III}}}$ that is Z times denser than the sampling lattice $\underline{\mathbb{III}}$. If the interpolation in the image restoration is sufficiently dense (normally, for $Z \geq 4$), then the blurring and raster effects of the image-display process are suppressed entirely. This condition leads not only to the best realizable image quality but also to the simplest expression for the continuous displayed image, as given in the same form as eqns [10] and [17] by

$$\hat{R}(\upsilon, \omega; \kappa) = \hat{\rho}(\upsilon, \omega)\hat{\Gamma}_r(\upsilon, \omega; \kappa) + \hat{n}_r(\upsilon, \omega; \kappa) \qquad [23]$$

where

$$\hat{\Gamma}_r(\upsilon, \omega; \kappa) = \hat{\tau}(\upsilon, \omega)\tilde{\tau}_e(\upsilon, \omega)\hat{\Psi}(\upsilon, \omega; \kappa) \qquad [24]$$

is the throughput SFR from image gathering to display, and

$$\begin{aligned}\hat{n}_r(\upsilon, \omega; \kappa) &= K^{-1}\hat{n}_e(\upsilon, \omega; \kappa)\hat{\Psi}(\upsilon, \omega; \kappa) \\ &= \hat{n}_{ar}(\upsilon, \omega; \kappa) + \hat{n}_{pr}(\upsilon, \omega; \kappa) \\ &\quad + \hat{n}_{qr}(\upsilon, \omega; \kappa)\end{aligned} \qquad [25]$$

is the accumulated noise in the restored image. This noise may be expressed explicitly by substituting the noise components given above. The noise due to the granularity in image-display mediums is not accounted for here.

The Wiener filter, that minimizes the mean-square error due to the perturbations in image gathering and coding, is given by

$$\hat{\Psi}(\upsilon, \omega; \kappa) = \frac{\hat{\Phi}'_\rho(\upsilon, \omega)\hat{\Gamma}^*_e(\upsilon, \omega)}{\hat{\Phi}'_\rho(\upsilon, \omega)|\hat{\Gamma}_e(\upsilon, \omega)|^2 + \hat{\Phi}'_{n_e}(\upsilon, \omega; \kappa)} \qquad [26]$$

where $\hat{\Phi}'(\upsilon, \omega) = \sigma_\rho^{-2}\hat{\Phi}_\rho(\upsilon, \omega)$ is the normalized PSD of the radiance field and

$$\begin{aligned}\hat{\Phi}'_{n_e}(\upsilon, \omega; \kappa) = (K\sigma_\rho)^{-2}\Big[&\hat{\Phi}_n(\upsilon, \omega)|\tilde{\tau}_e(\upsilon, \omega)|^2 \\ &+ \tilde{\Phi}_q(\upsilon, \omega; \kappa)\Big]\end{aligned}$$

is the normalized PSD of the accumulated noise in the encoded signal. If the photodetector noise is white so that the PSD $\tilde{\Phi}_p(\upsilon, \omega)$ is equal to its variance σ_p^2 and the PSD $\tilde{\Phi}_q(\upsilon, \omega; \kappa)$ is equal to its variance σ_q^2, then the PSD of the noise simplifies to

$$\hat{\Phi}'_{n_e}(\upsilon, \omega; \kappa) = \hat{\Phi}'_{n_e}(\upsilon, \omega) + \tilde{\Phi}'_q(\upsilon, \omega; \kappa) \qquad [27]$$

where

$$\begin{aligned}\hat{\Phi}'_{n_e}(\upsilon, \omega) = \Big[&\hat{\Phi}'_\rho(\upsilon, \omega)|\hat{\tau}(\upsilon, \omega)|^2 * \underline{\hat{\mathbb{III}}}_s \\ &+ (K\sigma_\rho/\sigma_p)^{-2}\Big]|\tilde{\tau}_e(\upsilon, \omega)|^2\end{aligned} \qquad [28]$$

and

$$\hat{\Phi}'_q(\upsilon, \omega; \kappa) = (K\sigma_\rho/\sigma_{s_e})^{-2}\kappa^{-2} \qquad [29]$$

These simplifying conditions permit the Wiener filter given above and the figures of merit given below to be expressed as a function of the SNR $K\sigma_\rho/\sigma_p$, which can be specified without accounting explicitly for either the gain K or the variances σ_ρ^2 and σ_p^2.

Figures of Merit

Information Rate $\mathcal{H}$

The information rate $\mathcal{H}$ (in bits sample^{-1}) of the encoded signal $s_e(\mathbf{x}, \mathbf{y}; \kappa)$ is defined by

$$\mathcal{H} = \mathcal{E}[\tilde{s}_e(\upsilon, \omega; \kappa)] - \mathcal{E}[\tilde{s}_e(\upsilon, \omega; \kappa)|\hat{\rho}(\upsilon, \omega) : (\upsilon, \omega) \in \hat{B}] \quad [30]$$

where the first term, $\mathcal{E}[\cdot]$, represents the entropy of $\tilde{s}_e(\upsilon, \omega; \kappa)$ and the second term, $\mathcal{E}[\cdot|\cdot]$, represents the conditional entropy of this signal when the reflectance spectrum $\hat{\rho}(\upsilon, \omega)$ within the sampling passband $\hat{B}$ is known. For the assumptions that have been made about image gathering and coding, the conditional entropy becomes the entropy of the accumulated noise $\hat{n}_e(\upsilon, \omega; \kappa)$, so that $\mathcal{H}$ simplifies to

$$\mathcal{H} = \mathcal{E}[\tilde{s}_e(\upsilon, \omega; \kappa)] - \mathcal{E}[\hat{n}(\upsilon, \omega; \kappa) : (\upsilon, \omega) \in \hat{B}] = \frac{1}{2}\int_{\hat{B}}\int \log\left[1 + \frac{\hat{\Phi}'_\rho(\upsilon, \omega)|\hat{\Gamma}_e(\upsilon, \omega)|^2}{\hat{\Phi}'_{n_e}(\upsilon, \omega; \kappa)}\right] d\upsilon d\omega \quad [31]$$

The theoretical upper bound of $\mathcal{H}$ is Shannon's channel capacity $\mathcal{C}$ for a bandwidth-limited system, with white photodetector noise and an average power limitation (here the variance σ_ρ^2), as given by

$$\mathcal{C} = \frac{1}{2}\log\left[1 + (K\sigma_\rho/\sigma_p)^2\right] \quad [32]$$

$\mathcal{H}$ can reach $\mathcal{C}$ only when

$$\hat{\Phi}_\rho(\upsilon, \omega) = \begin{cases} \sigma_\rho^2, & (\upsilon, \omega) \in \hat{B} \\ 0, & \text{elsewhere} \end{cases}$$

$$\hat{\tau}(\upsilon, \omega) = \begin{cases} 1, & (\upsilon, \omega) \in \hat{B} \\ 0, & \text{elsewhere} \end{cases} \quad [33]$$

However, neither of these two conditions can occur in practice because both the PSD $\hat{\Phi}_\rho(\upsilon, \omega)$ of natural scenes and the SFR $\hat{\tau}(\upsilon, \omega)$ of image-gathering devices decrease gradually with increasing spatial frequency, as depicted in Figures 3 and 6, respectively. Hence, $\mathcal{H}$ can never reach $\mathcal{C}$.

Theoretical Minimum Data Rate $\mathcal{E}$

The theoretical minimum data rate $\mathcal{E}$ (also in bits sample^{-1}) that is required to convey the information rate $\mathcal{H}$ is defined by

$$\mathcal{E} = \mathcal{E}[\tilde{s}_e(\upsilon, \omega; \kappa)] - \mathcal{E}[\tilde{s}_e(\upsilon, \omega; \kappa)|\tilde{s}(\upsilon, \omega)] \quad [34]$$

where the second term, $\mathcal{E}[\cdot|\cdot]$ is the uncertainty of $\tilde{s}_e(\upsilon, \omega; \kappa)$, when $\tilde{s}(\upsilon, \omega)$ is known. For the assumptions that have been made about the coding process, the conditional entropy becomes the entropy $\mathcal{E}[\tilde{n}_q(\upsilon, \omega; \kappa)]$ of the quantization noise, so that $\mathcal{E}$ simplifies to

$$\mathcal{E} = \mathcal{E}[\tilde{s}_e(\upsilon, \omega; \kappa)] - \mathcal{E}[\tilde{n}_q(\upsilon, \omega; \kappa)] = \frac{1}{2}\int_{\hat{B}}\int \log\left[1 + \frac{\hat{\Phi}'_\rho(\upsilon, \omega)|\hat{\Gamma}_e(\upsilon, \omega)|^2 + \hat{\Phi}'_{n_e}(\upsilon, \omega)}{\tilde{\Phi}'_q(\upsilon, \omega; \kappa)}\right] d\upsilon d\omega \quad [35]$$

This expression for $\mathcal{E}$ represents the entropy of completely decorrelated data.

Information Efficiency $\mathcal{H}/\mathcal{E}$

The information efficiency of completely decorrelated data is defined by the ratio $\mathcal{H}/\mathcal{E}$. This ratio reaches its upper bound $\mathcal{H}/\mathcal{E} = 1$ when $\hat{\Phi}'_{n_e}(\upsilon, \omega) \ll \hat{\Phi}'_\rho(\upsilon, \omega)|\hat{\Gamma}_e(\upsilon, \omega)|^2$ and $\hat{\Phi}'_{n_e}(\upsilon, \omega) \ll \tilde{\Phi}'_q(\upsilon, \omega; \kappa)$, because then both eqn [31] for $\mathcal{H}$ and eqn [35] for $\mathcal{E}$ reduce to the same expression

$$\mathcal{H} = \mathcal{E} = \frac{1}{2}\int_{\hat{B}}\int \log\left[1 + \frac{\hat{\Phi}'_\rho(\upsilon, \omega)|\hat{\Gamma}_e(\upsilon, \omega)|^2}{\tilde{\Phi}'_q(\upsilon, \omega; \kappa)}\right] d\upsilon d\omega \quad [36]$$

This upper bound can be approached with a small loss in $\mathcal{H}$ only when the aliasing and photodetector noise are small. Hence, the electro-optical design that optimizes $\mathcal{H}$ also optimizes $\mathcal{H}/\mathcal{E}$. However, a compromise always remains between $\mathcal{H}$ and $\mathcal{H}/\mathcal{E}$ in the selection of the number of quantization levels: $\mathcal{H}$ favors fine quantization, whereas $\mathcal{H}/\mathcal{E}$ favors coarse quantization.

Maximum-Realizable Fidelity $\mathcal{F}$

The maximum-realizable fidelity $\mathcal{F}$ of the images restored with the Wiener filter $\hat{\Psi}(\upsilon, \omega; \kappa)$ can be expressed as a function of the throughput SFR $\hat{\Gamma}_r(\upsilon, \omega; \kappa)$ given by eqn [24] as

$$\mathcal{F} = \int\int \hat{\Phi}'_\rho(\upsilon, \omega)\hat{\Gamma}_r(\upsilon, \omega; \kappa)\, d\upsilon d\omega \quad [37]$$

or, equivalently, as a function of the integrand $\hat{\mathcal{H}}(\upsilon, \omega; \kappa)$ of the information rate $\mathcal{H}$ given by eqn [31] as

$$\mathcal{F} = \int\int \hat{\Phi}'_{\rho}(\upsilon, \omega)\left[1 - 2^{-\hat{\mathcal{H}}(\upsilon,\omega;\kappa)}\right] d\upsilon d\omega \qquad [38]$$

This relationship between fidelity and information rate can be extended to other goals. For example, the enhancement of spatial features (e.g., edges and boundaries), for robotic vision, can be accommodated by letting $\tilde{\tau}_e(\upsilon, \omega)$ be the desired feature-enhancement filter.

Performance and Design Trade-offs

Electro-optical Design

Figure 7 presents the information rate $\mathcal{H}$ (for lossless encoding) as a function of the optical-response index ζ_c and the mean spatial detail μ. The information rate $\mathcal{H}$ versus ζ_c is given for several SNRs $K\sigma_\rho/\sigma_p$ and $\mu = 1$, and $\mathcal{H}$ versus μ is given for the three informationally optimized designs specified in Table 1.

The curves show that $\mathcal{H}$ can reach its highest possible value only when both of the following conditions are met:

1. The relationship between the SFR $\hat{\tau}(\upsilon, \omega)$ and the sampling passband $\hat{B}$ is chosen to optimize $\mathcal{H}$ for the available SNR $K\sigma_\rho/\sigma_p$. This design is said to be informationally optimized.
2. The relationship between the sampling passband $\hat{B}$ and the PSD $\hat{\Phi}_\rho(\upsilon, \omega)$ is chosen to optimize $\mathcal{H}$. For the PSDs $\hat{\Phi}_\rho(\upsilon, \omega)$ typical of natural scenes, the best match occurs when the sampling interval is near the mean spatial detail of the scene (i.e., when $\mu \approx 1$).

These conditions are consistent with those for which the information rate $\mathcal{H}$ reaches Shannon's channel capacity $\mathcal{C}$. However, blurring and aliasing constrain $\mathcal{H}$ to values that are smaller than $\mathcal{C}$ by a factor of nearly two (i.e., $\mathcal{H} \lesssim \mathcal{C}/2$). Hence, these conditions replace the role of white and band-limited signals in classical communication theory, to establish an upper bound on the information rate in imaging systems.

The above two conditions appeal intuitively when the problem of image restoration is considered:

1. The result that the SFR $\hat{\tau}(\upsilon, \omega)$ that optimizes $\mathcal{H}$ depends on the available SNR is consistent with the observation that, in one extreme, when the SNR is low, substantial blurring should be avoided because the photodetector noise would constrain the enhancement of fine spatial detail even if aliasing were negligible. In the other extreme, when the SNR is high, substantial aliasing should be avoided so that this enhancement is relieved from any constraints except those that the sampling passband inevitably imposes.

Table 1 Informationally optimized designs

Design	$K\sigma_\rho/\sigma_p$	ζ_c	$\mathcal{H}^*$
1	256	0.3	4.4
2	64	0.4	3.3
3	16	0.5	2.2

*$\mu = 1$.

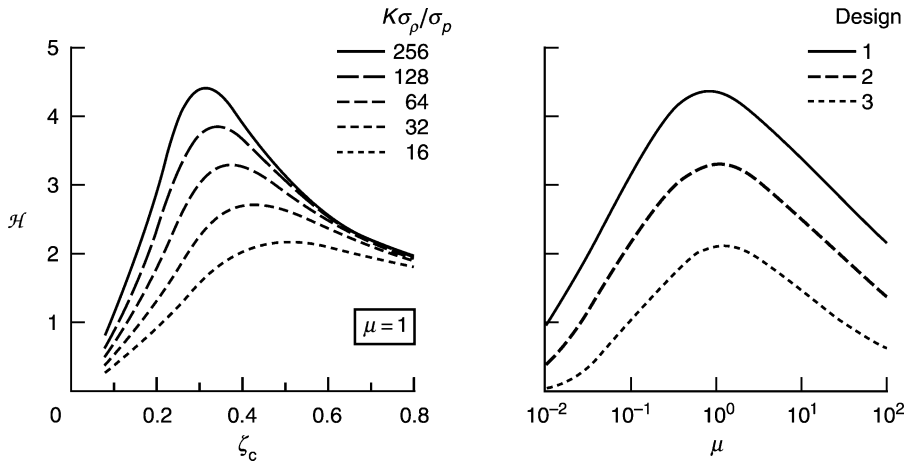


Figure 7 Information rate $\mathcal{H}$ versus the optical-response index ζ_c and the mean spatial detail μ (relative to the sampling interval). The designs are specified in **Table 1**, and the SFRs $\hat{\tau}(\upsilon, \omega)$ for ζ_c are shown in **Figure 6**.

2. The result that $\mathcal{H}$ reaches its highest value when the sampling interval is near the mean spatial detail of the scene is consistent with the observation that, ordinarily, it would not be possible to restore spatial detail that is finer than the sampling interval, whereas it would be possible to restore much coarser detail from fewer samples.

Information Rate and Data Rate

Figure 8 presents an information-entropy $\mathcal{H}(\mathcal{E})$ plot that characterizes the relationship between the information rate $\mathcal{H}$ of the encoded signal and the associated theoretical minimum data rate, or entropy, $\mathcal{E}$ for η-bit quantization. The encoder SFR $\tilde{\tau}_e(\upsilon, \omega)$ is assumed to be unity. The plot illustrates the trade-off between $\mathcal{H}$ and $\mathcal{E}$ in the selection of the number of quantization levels for encoding the acquired signal. It shows, in particular, that the design that realizes the highest information rate $\mathcal{H}$ also enables the encoder to realize the highest information efficiency $\mathcal{H}/\mathcal{E}$. This result appeals intuitively, because the perturbations due to aliasing and photodetector noise that interfere with the image restoration can also be expected to do so with the signal decorrelation.

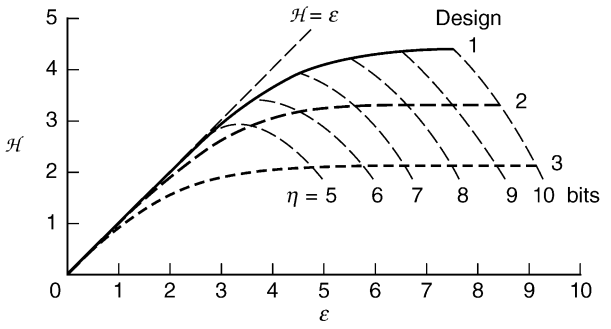


Figure 8 The information-entropy $\mathcal{H}(\mathcal{E})$ plot of the information rate $\mathcal{H}$ versus the associated theoretical minimum data rate $\mathcal{E}$ for η-bit quantization. The designs are specified in Table 1.

Throughput SFR

Figure 9 illustrates the SFRs $\hat{\tau}(\upsilon, \omega)$ and $\hat{\Psi}(\upsilon, \omega; \kappa)$ of the image-gathering device and Wiener filter, respectively, and of their product, the throughput SFR $\hat{\Gamma}_r(\upsilon, \omega; \kappa)$, for the three designs specified in Table 1. As the accumulated noise becomes negligible, $\hat{\Gamma}_r(\upsilon, \omega; \kappa)$ approaches the ideal SFR of the classical communication channel, which is unity inside the sampling passband and zero outside.

Information Rate, Fidelity and Robustness

The PSD $\hat{\Phi}_\rho(\upsilon, \omega)$ of a scene is seldom known when images are restored. Moreover, even if the PSD were known to be the one given by eqn [2], then the mean

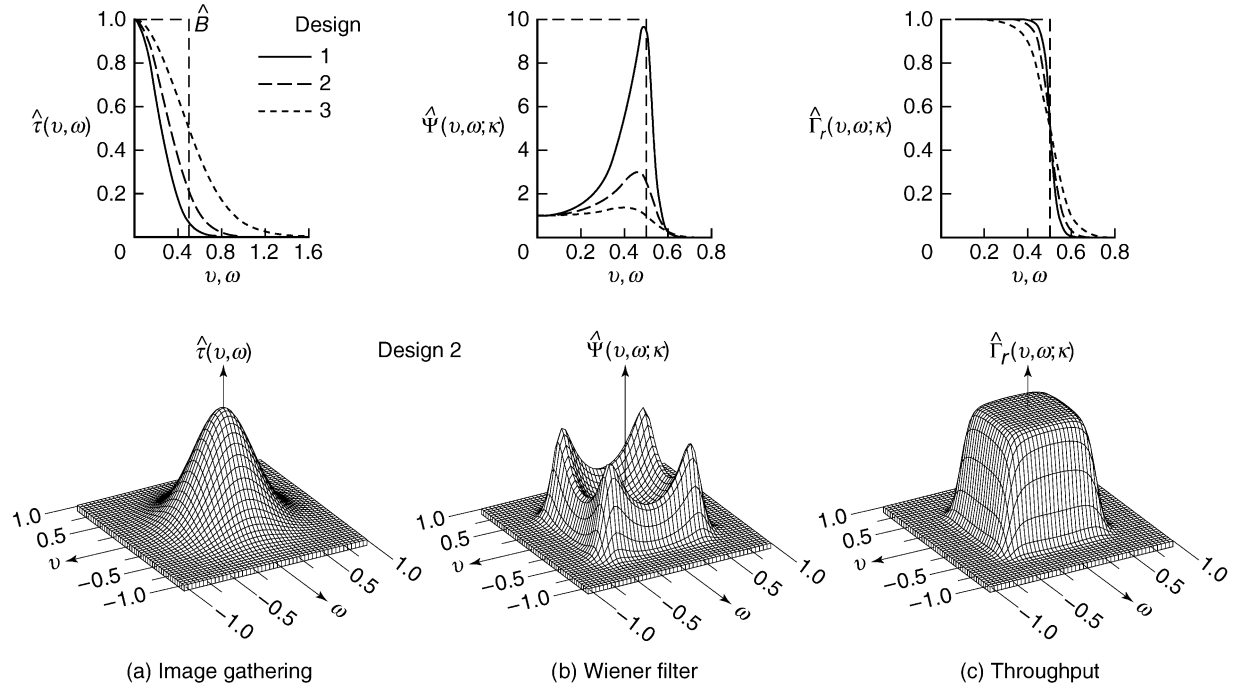


Figure 9 SFRs of image gathering, restoration, and throughput. The designs are specified in Table 1.

spatial detail μ would usually remain uncertain because it depends, not only on the properties of the scene, but also on the angular resolution and viewing distance of the image-gathering device.

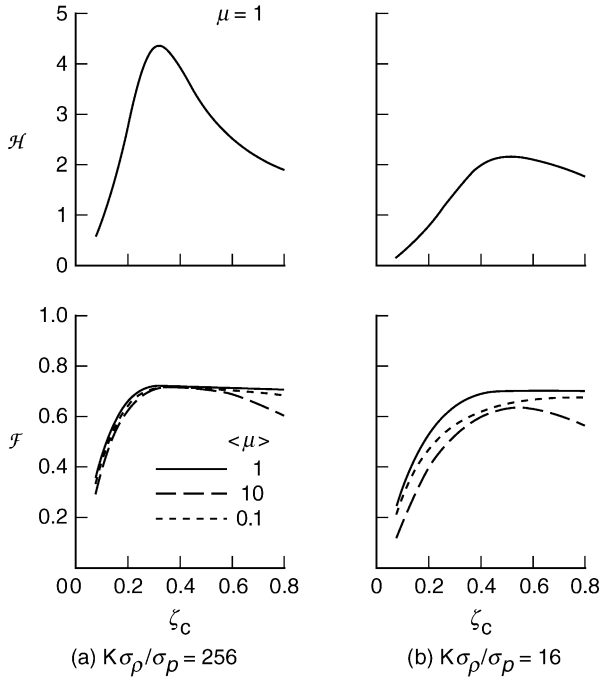


Figure 10 Information rate $\mathcal{H}$ and maximum-realizable fidelity $\mathcal{F}$ versus the optical-response index ζ_c for two SNRs $K\sigma_\rho/\sigma_p$. $\mathcal{F}$ is given for the matched ($\langle\mu\rangle = 1$) and two mismatched Wiener restorations.

In practice, therefore, image restorations have to rely on estimates of the statistical properties of scenes. The tolerance of the quality of the restored image to errors in these estimates is referred to as the robustness of the image restoration.

Figure 10 presents the information rate $\mathcal{H}$ and the corresponding maximum-realizable fidelity $\mathcal{F}$ for matched and mismatched Wiener restorations. These restorations represent the digital image $R(\bar{x}, \bar{y}; \kappa)$ or, equivalently, the continuous image $R(x, y; \kappa)$ displayed on a noise-free medium. The matched restorations use the correct value of μ, whereas the mismatched restorations use wrong estimates of μ. These curves reveal that the informationally optimized designs produce the highest fidelity and robustness, and that both improve with increasing $\mathcal{H}$. Moreover, these curves reveal that, ordinarily, one cannot go far wrong in Wiener restorations by assuming that the mean spatial detail is equal to the sampling interval (i.e., $\mu = 1$). This observation appeals intuitively because spatial detail that is much smaller cannot be resolved, and detail that is larger is not degraded substantially by blurring and aliasing.

Information Rate and Visual Quality

Figures 11 and 12 present Wiener restorations and their three components, as given by eqn [23], where $\rho(x, y) * \Gamma_r(x, y; \kappa)$ accounts for blurring, $n_{ar}(x, y; \kappa)$ accounts for aliasing, and $n_{pr}(x, y; \kappa)$ accounts for

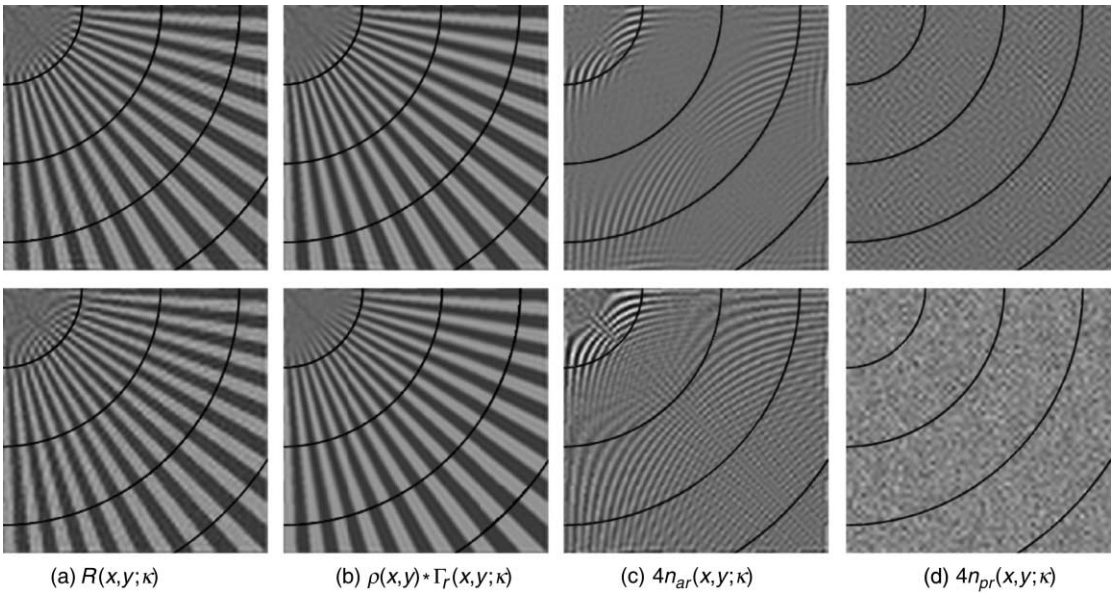


Figure 11 Images restored with the Wiener filter and their three components for designs 1 (top) and 3 (bottom).

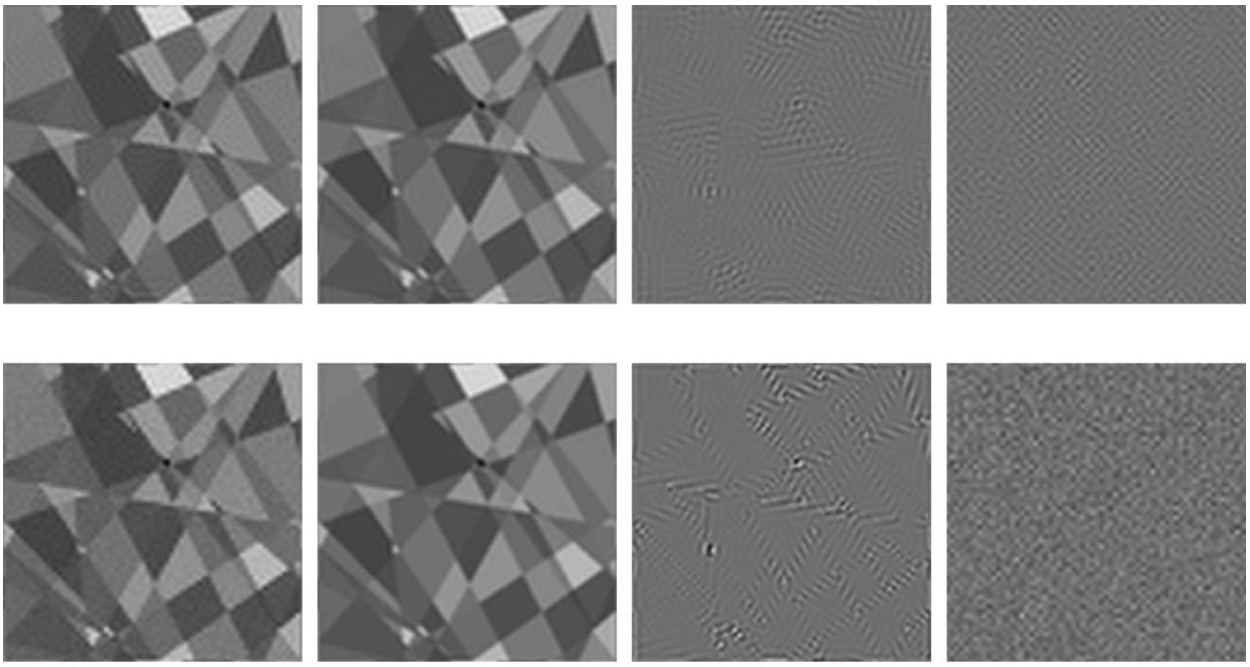


Figure 12 Images restored with the Wiener filter and their three components for designs 1 (top) and 3 (bottom).

photodetector noise. The contrast of the illustrations of the two noise components has been increased by a factor of four to improve their visibility. The extraneous structure that aliasing produces in the images of the resolution wedges is commonly referred to as Moiré pattern. The corresponding distortion in the images of the random polygons emerges more subtly as jagged (or staircase) edges.

The images are formed from a set of 64×64 samples with a $Z = 4$ interpolation and have a 4×4 cm format. If these images were displayed in a smaller format, then the jagged edges could not be resolved by the observer and would appear to be the result of blurring or photodetector noise instead of aliasing. Aliasing has, therefore, often been overlooked as a significant source of image degradation.

The Wiener restoration produces images in which fine spatial detail near the limit of resolution normally has a high contrast approaching that in the scene. However, these images also tend to contain visually annoying defects, due to the accumulated noise and the ringing near sharp edges (Gibbs phenomenon). Hence, it is often desirable to combine this restoration with an enhancement filter that gives the user some control over the trade-off among fidelity, sharpness and clarity.

Figure 13 presents images for which the Wiener restorations given in Figures 11 and 12 are enhanced for visual quality by suppressing some of the defects due to aliasing and ringing. This enhancement improves clarity at the cost of a small loss in sharpness. The images are produced again from 64×64 pixels, but they are displayed in three different formats. The smallest format, with a density of 64 pixels/cm, corresponds to a display of 256×256 pixels in a 4×4 cm area, which is typical of the images found in the prevalent digital image-processing literature. As can be observed, this small format leads to images that are indistinguishable from each other because they hide many of the distortions that larger formats clearly reveal.

In general, if the images are displayed in a format that is large enough to reveal the finest spatial detail near the limit of resolution of the imaging system, then distortions due to the perturbations in this system become also visible. When no effort is spared to reduce these distortions without a perceptual loss of resolution and sharpness, then it becomes strongly evident that the best visual quality with which images can be restored improves with increasing information rate, even after the fidelity has essentially reached its upper bound. This improvement continues until it is gradually ended by the unavoidable compromise among sharpness, aliasing, and ringing as well as by the granularity of the image display.

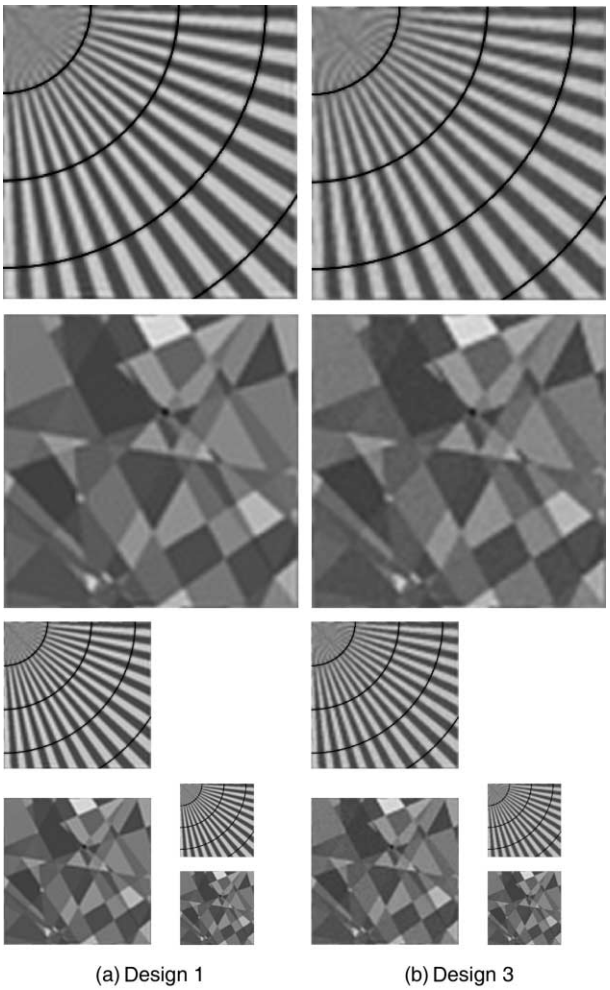


Figure 13 Images restored with the Wiener filter and enhanced for visual quality. The large, medium, and small formats contain 16, 32 and 64 pixels/cm, respectively.

Concluding Remarks

The performance of image-gathering devices and the human eye is constrained by the same critical limiting factors. When these factors are accounted for properly, then it emerges that the design of the image-gathering device that is optimized for the highest realizable information rate corresponds closely, under appropriate conditions, to the design of the human eye that evolution has optimized for viewing the real world. This convergence of information theory and evolution toward the same design clearly supports the extension of information theory to the characterization of imaging systems.

Numerous other information-theoretic characterizations of imaging are now emerging. However, none of these characterizations accounts for the critical limiting factors that constrain image gathering and restoration, as outlined here. Hence, it is not possible for them to address the efficiency and accuracy with which images can be conveyed. Moreover, none of these characterizations combines observation with communication, or sensing with processing, in a general formalization. Instead, each of them still deals with a particular measurement or imaging problem. Hence, the role of information theory in imaging is far from mature. Many challenges remain: to develop general formalizations, to account for physical phenomena, and to correlate predictions with measurements.

See also

Information Processing: Optical Digital Image Processing. **Modulators:** Electro-optics.

Further Reading

Fales CL, Huck FO, Alter-Gartenberg R and Rahman Z (1996) Image gathering and digital restoration. *Philosophical Transactions of the Royal Society of London* A354: 2249–2287.

Huck FO and Fales CL (2001) Characterization of image systems. *Encyclopedia of Imaging Science and Technology*. New York: John Wiley & Sons.

Huck FO, Fales CL and Rahman Z (1996) An information theory of visual communication. *Philosophical Transactions of the Royal Society of London* A354: 2193–2248.

Huck FO, Fales CL and Rahman Z (1997) *Visual Communication: An Information Theory Approach*. Boston, CA: Kluwer Academic Publishers.

Lesurf JCG (1995) *Information and Measurement*. Bristol, UK: Institute of Physics Pub.

O'Sullivan JA, Blahut RE and Snyder DL (1998) Information-theoretic image formation. *IEEE Transactions Information Theory* 44: 2094–2123.

Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423 and 28: 623–656.

Shannon CE and Weaver W (1964) *The Mathematical Theory of Communication*. Urbana, IL: U. Illinois Press.

Wiener N (1949) *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. New York: John Wiley & Sons.

Inverse Problems and Computational Imaging

M Bertero and P Boccacci, University of Genova, Genova, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A digital image is an array (2D image) or a cube (3D image) of quantized numbers, providing a representation of a physical object or a map of some distributed property of the object (density of matter, energy, etc.). It can be obtained by digitizing the original image (for instance, a photograph) or can be directly formed by the outputs of the imaging system (digital camera, microscope, telescope, etc.). When visualized on a suitable image display device, a digital image must provide a representation of the physical object which can be easily interpreted and analyzed. In many cases, however, this condition is not satisfied. We give two typical examples.

The first is that of an image degraded by blurring which can be introduced by defects of the optical systems (such as aberrations) or by image motion, image defocusing, etc. The representation of the object does not have the quality which should be desirable for its interpretation.

The second case is even more important and arises when the output of the imaging system is not directly related to the physical quantity to be imaged. For instance, in X-ray tomography, the output of a detector is the attenuation of an X-ray pencil crossing the body while the quantity to be imaged is the density of the body.

In the situations described above, the desired image can be obtained by solving a number of mathematical problems. If we denote f as the object, namely the target of the imaging system, and g as the image of f provided by the system, then the problems to be solved can be summarized as follows:

- develop a physical model of the system to obtain a mathematical relationship between the object f and the image g; the computation of g from a given f is usually called the solution of the direct problem;
- solve the problem of obtaining the physical quantity f from given outputs g of the system; this second step is usually called the solution of the inverse problem.

The digital image of f, obtained by solving an inverse problem, is a computed one, generated by the computer where the algorithm for the solution of the inverse problem has been implemented.

It may be convenient to formulate both the direct and the inverse problem in terms of functions rather than arrays, cubes, etc., because such a formulation allows the use of powerful tools of functional analysis for their investigation. Next, the results obtained in the continuous case can be used for understanding the features of the corresponding discrete problem. In most important applications, for instance, the relationship between f and g is linear so that, in the discrete version of the direct problem, one obtains g by applying a suitable matrix A to f while, in the discrete version of the inverse problem, one obtains f from g by solving a linear algebraic system. However, the solution of the inverse problem is not so simple and the difficulties can be understood by looking at the corresponding continuous problem.

The basic reason relies on the fact that inverse problems are ill-posed in the sense of Hadamard: the solution may not exist; even if it exists it may not be unique; and, even if it exists and is unique, it may not depend continuously on the data. The last statement means that a small perturbation of the data, such as that caused by noise or any kind of experimental error, can completely modify the solution.

Inverse problems are ill-posed because imaging systems do not transmit complete information about the physical object. Therefore, the problem is to extract the useful information contained in the data or, in mathematical terms, to look for approximate solutions which are stable against noise. This is the purpose of a mathematical theory introduced in its general form by the Russian mathematician, Tikhonov and known as regularization theory.

We illustrate the general framework outlined above by means of a specific example – image deconvolution, also known as image deblurring, image restoration, etc. The object f is the image which should be recorded in the absence of degradation, while g is the image corrupted by aberrations or other causes of blurring. Both f and g are functions of 2D (or 3D) space variables $\mathbf{x}$ which are the coordinates of a point in the image domain. If the imaging system is isoplanatic, then it is described by a space-invariant point spread function (PSF) $K(\mathbf{x})$, which is the image of a point source located in the center of the image domain. The PSF provides the response of the system to any point source wherever it is located. On the other hand, its Fourier transform, the transfer function $\hat{K}(\omega)$ (the hat denotes Fourier transform

and ω, the spatial frequencies associated to the space variables $\mathbf{x}$), tells us how a signal of a fixed frequency is propagated through the linear system, so that the blurring can also be viewed as a sort of frequency filtering.

Under these assumptions the image g is the convolution product of object f and PSF K:

$$g(\mathbf{x}) = \int K(\mathbf{x} - \mathbf{x}')f(\mathbf{x}')\mathrm{d}\mathbf{x}' \qquad [1]$$

and the solution of the direct problem is just the computation of a convolution product.

The image g, as given by eqn [1], is the so-called noise-free image. The actually recorded image g_r is affected by a noise term and, in general, it can be written in the following form:

$$g_r(\mathbf{x}) = g(\mathbf{x}) + n(\mathbf{x}) \qquad [2]$$

where $n(\mathbf{x})$ is a function describing noise contamination. It is a random function and, therefore, it is unknown even if one understands its statistical properties. We point out that eqn [2] does not imply that the noise is additive or signal independent. For instance, in microscopy and astronomy, the contamination of the image is due both to photon noise, which satisfies Poisson statistics, and read-out noise, which is basically Gaussian and white. Therefore the noise term in eqn [2] must only be intended as the difference between the noisy and the noise-free image.

In a first attempt at approaching the inverse problem, it seems quite natural to ignore the noise term $n(\mathbf{x})$ and to solve eqn [1] for $f(\mathbf{x})$ with $g(\mathbf{x})$ replaced by $g_r(\mathbf{x})$. Then, by taking the Fourier transform of both sides of this equation and using the well-known convolution theorem, one finds the following relationship:

$$\hat{g}_r(\omega) = \hat{K}(\omega)\hat{f}(\omega) \qquad [3]$$

which relates the Fourier transform of f and g to the transfer function of the imaging system.

The solution of this equation looks elementary. However, a first problem is due to the fact that an optical system is generally band-limited; its band Ω is the bounded domain of spatial frequencies where the transfer function is different from zero. Since the noise is not band-limited or, at least, has a band much broader than that of the optical system, outside Ω the right-hand side of eqn [3] is zero while the left-hand side is not; in other words, the solution of the deconvolution problem, as formulated above, does not exist because no object f satisfies eqn [3] everywhere. In this particular case, one can circumvent the problem by applying a suitable filter to $g_r(\mathbf{x})$, in order to suppress the out-of-band noise.

The second problem is the nonuniqueness of the solution of the problem with the filtered data. This is due to the so-called invisible objects, namely objects whose Fourier transform is zero on Ω, so that the corresponding images are exactly zero. If we find a solution of the deconvolution problem, by adding an arbitrary invisible object to this solution, we find another solution of the same problem. A well-defined solution can be obtained by requiring its Fourier transform to be zero outside Ω. In other words, we set to zero what is not transmitted by the imaging system.

The standard way for approaching the previous questions is to look for solutions in the least-squares sense, namely for objects which minimize the functional:

$$\varepsilon^2(f) = \|K * f - g_r\|^2 \qquad [4]$$

where $*$ denotes the convolution product, as defined in eqn [1], and the right-hand side is the square of the L^2-norm (which can also be interpreted as energy) of the discrepancy between the recorded image g_r and the computed image $K * f$. If one looks for a least-squares solution with minimal energy then one automatically re-obtains the solution discussed above, namely an object whose Fourier transform in Ω is given by

$$\hat{f}_{\mathrm{inv}}(\omega) = \frac{\hat{g}_r(\omega)}{\hat{K}(\omega)} \qquad [5]$$

and is zero outside Ω. Such a solution is that provided by the so-called inverse filter.

However, there is an additional difficulty in that the transfer function usually tends to zero at the boundary of the band while the Fourier transform of the noise, hence that of g_r, does not, or tends to zero in a different way. It follows that $\hat{f}_{\mathrm{inv}}(\omega)$ is divergent or very large at the boundary of the band. We conclude that its inverse Fourier transform does not exist or, if it exists, is affected by large artifacts due to noise propagation. The latter is the typical situation in the case of digital images. In **Figure 1** we give the result of a numerical simulation showing the typical effect produced by the inverse filter. The object is a beautiful picture of a galaxy recorded by the Hubble Space Telescope (HST), while the PSF is a computed one describing the blurring of HST before installation of the optics correction. The blurred image is obtained by convolving the object with the PSF and by adding white noise. The restored image produced by the inverse filter is completely corrupted by noise.

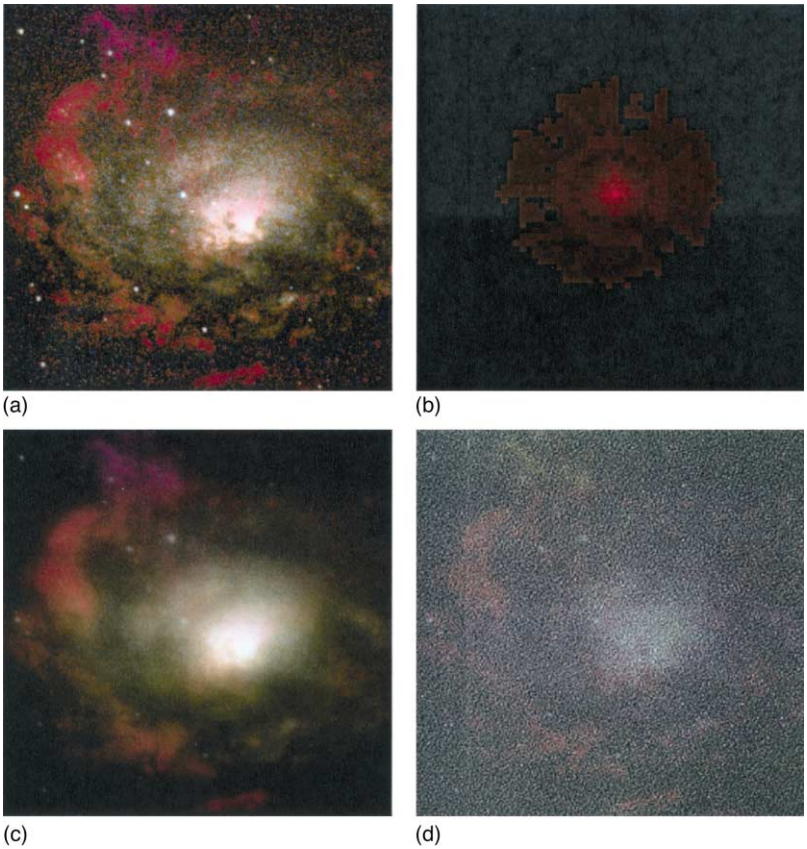


Figure 1 Illustrating the effect of the inverse filter: (a) Image of the Circinus galaxy taken on 10 April 1999, with the Wide Field Planetary Camera 2 of the Hubble Space Telescope; (b) the PSF used for blurring the RGB components of the image (here shown in red); (c) the blurred image obtained by convolving the components of the image in (a) with the corresponding PSFs and adding noise; (d) the restoration obtained by applying the inverse filter to the three components of (c). Part (a) courtesy of NASA/Space Telescope Science Institute.

The inverse filter provides a solution which fits the data in the best possible way and therefore also fits the noise in the frequency domains where the signal is weakly transmitted by the imaging system. This can be achieved at the cost of an energy (L^2-norm) of the solution which is too large. Therefore, such a conclusion suggests that one must search for a reasonable compromise between data fitting and energy of the solution. This is the basic idea of regularization theory, at least in its most simple form. The mathematical formulation is obtained by looking for objects which minimize the functional:

$$\Phi_\mu(f) = \|K * f - g_r\|^2 + \mu\|f\|^2 \qquad [6]$$

where μ, the so-called regularization parameter, is a parameter controlling the trade-off between data fitting and energy of the solution: when μ is small, the data fitting is good and the energy is large while, when μ is large, the data fitting is poor and the energy is small.

For a given μ, the function f_μ, which minimizes the functional of eqn [6] is called the regularized solution of the problem. It is easy to show that its Fourier transform is given by

$$\hat{f}_\mu(\omega) = \frac{\hat{K}^*(\omega)}{|\hat{K}(\omega)|^2 + \mu}\hat{g}_r(\omega) \qquad [7]$$

The regularized solutions form a one-parameter family of functions. When μ is small these functions are strongly contaminated by noise whose effect is gradually reduced when μ increases. For too large values of μ, one re-obtains blurred versions of the original object.

Such behavior is illustrated by the sequence of regularized solutions given in Figure 2 and corresponding to the example of Figure 1. The regularization parameter increases starting from left to right and from top to bottom. Its values are $\mu = 0$ (inverse filter), 0.001, 0.01, 0.05, 0.1, and 0.5 (the noise is of the order of a few percent and the PSF is normalized in such a way that the sum of all its value is 1). The sequence makes evident the well-established theoretical result of the existence of an optimal value of the regularization parameter. Such an optimal value can

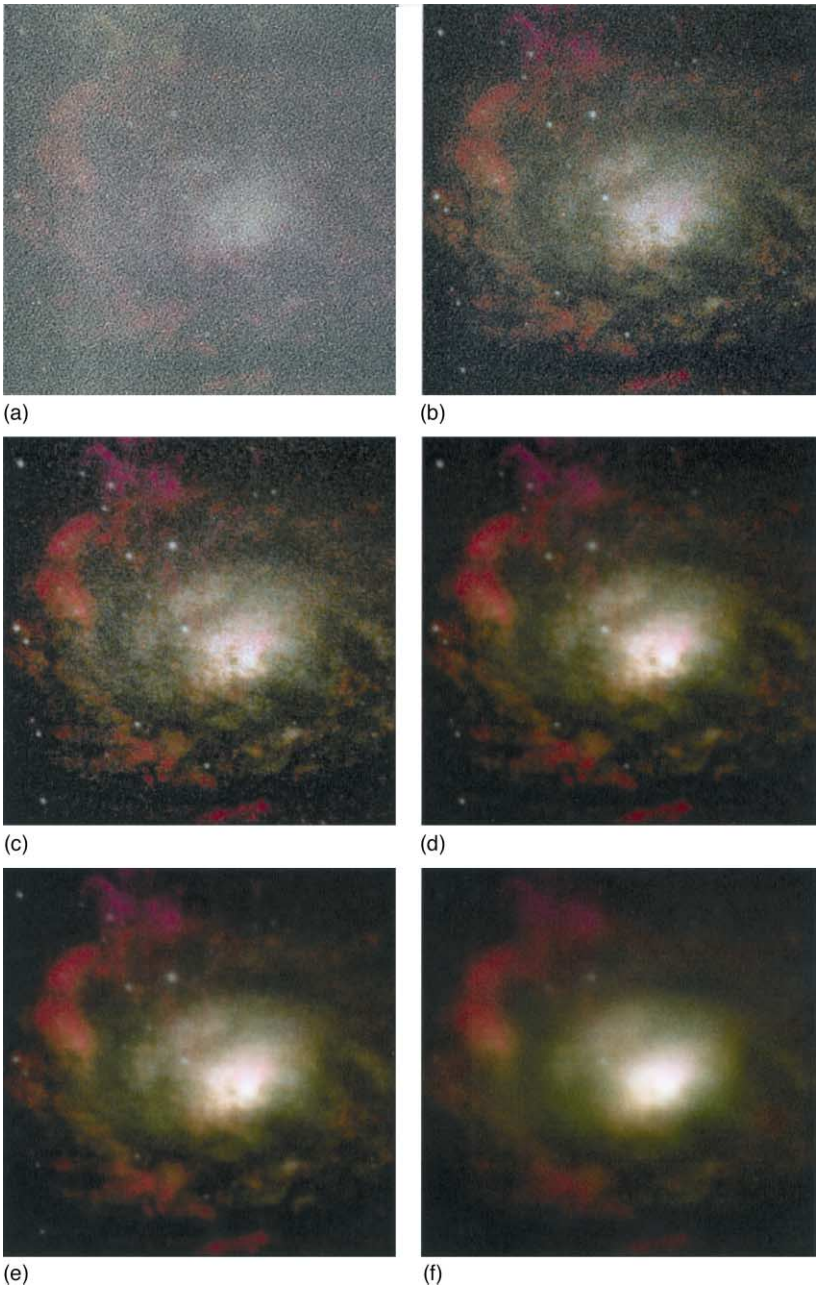


Figure 2 Illustrating the effect of the regularization parameter. The blurred image is that shown in **Figure 1c**. The sequence is obtained by increasing the value of the regularization parameter: (a) $\mu = 0$ (inverse filter); (b) $\mu = 0.001$; (c) $\mu = 0.01$; (d) $\mu = 0.05$; (e) $\mu = 0.1$; (f) $\mu = 0.5$. It is evident that the best restoration is that shown in (c).

be determined in the case of numerical simulations as that represented in **Figure 2**. However, its estimation in the case of real images is more difficult. Several methods for the selection of μ have been proposed. We only mention the so-called discrepancy principle. It consists of determining the value of μ, such that the value of the discrepancy functional at f_μ, $\varepsilon^2(f_\mu)$ coincides with an estimate of the energy of the noise. In other words, the data are fitted with an accuracy comparable with their uncertainty.

As shown by eqn [7], the regularized solution has the same band Ω of the imaging system, because $\hat{K}^*(\omega)$ is zero outside Ω. As is known, a band-limited function can be represented in terms of its samples taken at a rate which is roughly proportional to the size of the band. This is the content of the famous Shannon sampling theorem (more precisely its 2D extension). On the other hand, the sampling distance can be taken as a measure of the resolution provided by the restored image so that one concludes that the

restoration method discussed above does not produce an improvement of the resolution of the imaging system. It certainly produces an improvement of the quality of the image which makes possible a visual verification of that resolution.

In many circumstances, the term super-resolution is used for describing methods which allow an improvement of resolution beyond the limit provided by the sampling theorem. Therefore, a super-resolving method must produce a restored image with a band broader than that of the imaging system. Therefore, super-resolution is related to the problem of out-of-band extrapolation.

Such an approach was already proposed in the 1960s, when it was observed that the Fourier transform of an object with a finite spatial extent (all objects have this property) is analytic; then the theorem of unique analytic continuation implies that the Fourier transform of the object can be determined everywhere from its values on the band of the imaging system.

Unfortunately this problem is ill-posed, so that out-of-band extrapolation is not feasible in practice. Only recently it has been recognized that a considerable amount of super-resolution is possible when the spatial extent of the object is not much greater than the resolution distance of the instrument. For example, super-resolving methods could be used for detecting unresolved binary star in astronomical images.

The important point is that a super-resolving method must implement explicitly the finite extent of the object: the domain of the object must be estimated and the method must search for a solution which is zero outside this domain. This introduces an important question in object restoration and, more generally, in the theory of inverse problems, namely the design of regularized solutions satisfying additional conditions (constraints). This question is also important from the theoretical point of view: the problem is ill-posed because of insufficient information on the object as transmitted by the imaging system; the use of additional constraints reduces the class of the solutions which are compatible with the data and therefore can improve the restoration. This is the use of a priori information in the solution of inverse problems.

A constraint which has been widely investigated and used is the positivity of the solution, i.e., all the values of the restored image must be positive or zero. The physical meaning of the constraint is obvious. Its beneficial effect is to reduce ringing artifacts which affect the regularized solutions previously discussed, in cases where the object contains bright spots over a black background or sharp intensity variations from zero to some positive value. The ringing is essentially related to the well-known Gibbs effect in the truncation of the Fourier series.

The requirement of positivity is the requirement of a particular lower bound on the values of the solution. Therefore, one can consider more general constraints in requiring a given lower and/or upper bound on the values of the solution, eventually combined with a constraint on the domain to produce, for instance, a super-resolved positive solution. The most general form of these constraints requires that the solution belongs to a given closed and convex set $\mathcal{C}$ in some functional space. A general method producing regularized solutions in a convex set is an iterative method which is essentially a gradient method for the minimization of the least-squares functional, with a projection on the set $\mathcal{C}$ at each iteration. In the case of object deconvolution, the least-squares functional is that given in eqn [4] and the method has the following form: if f_k is the result of the k-th iteration, then f_{k+1} is given by

$$f_{k+1} = P_{\mathcal{C}}[f_k + \tau K^{\mathrm{T}} * (g_r - K * f_k)] \qquad [8]$$

where $P_{\mathcal{C}}$ is the convex projection onto the set $\mathcal{C}$, τ is a relaxation parameter and $K^{\mathrm{T}}(\mathbf{x}) = K(-x)$. In general, the algorithm, known as the projected Landweber method, is initialized with $f_0 = 0$. An important property is that it has a regularization effect, in the sense that iterations must not be pushed to convergence but stopped after a suitable number of iterations to avoid strong noise contamination. In Figure 3 we compare, in a specific example, the result obtained by means of the linear regularization method and that obtained by means of the projected iterative method with a lower and upper bound on the solution. The reduction of ringing is evident.

Among the methods producing positive solutions we must mention the most popular one in astronomy, the so-called Richardson–Lucy method, which is an iterative method for Maximum-Likelihood estimation. If the PSF has been normalized in such a way that the sum of all its values is one, it takes the following form:

$$f_{k+1} = f_k\left(K^{\mathrm{T}} * \frac{g_r}{K * f_k}\right) \qquad [9]$$

The method is normally initialized with a uniform image and it must not be pushed to convergence to avoid noise amplification. Another algorithm suggested by statistical arguments and producing positive solution is the so-called Maximum Entropy Method (MEM).

The image restoration methods described above have wide applications both in microscopy and

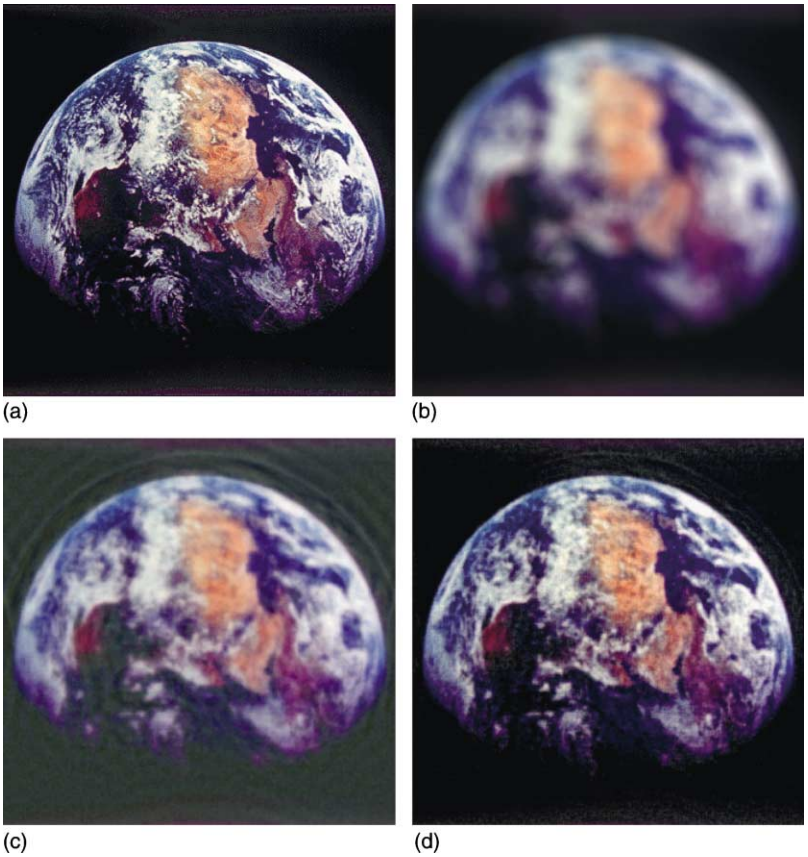


Figure 3 Illustrating the effect of constraints on image restoration: (a) Image of the Earth taken during the Apollo 11 mission; (b) blurred image of (a) obtained by assuming out-of-focus blur and adding noise; (c) optimal restoration provided by the regularization method; the ringing around the boundary of the Earth is evident; (d) optimal restoration obtained by means of the projected Landweber method with a lower and an upper bound on the solution at each iteration. Part (a) courtesy of NASA.

astronomy. For instance, the minimization of the functional eqn [6], with the additional constraint of positivity, has been used for deconvolving images in microscopy to reduce the effect of the missing cone. On the other hand, the Richardson–Lucy method was widely used for deconvolving the images of HST before installation of corrective optics.

Nowadays, image deconvolution is becoming more important for the ground-based telescopes equipped with adaptive optics. Indeed, even if adaptive optics is able to provide a considerable compensation of the atmospheric blur, a further improvement can be obtained by deconvolving the detected images. The PSF is provided by the image of a suitable guide star.

Finally, it is worth mentioning the Large Binocular Telescope, in construction on the Mount Graham in Arizona, because this instrument requires image restoration methods for a full exploitation of its imaging properties. The telescope (Figure 4) consists of two 8 m mirrors on a common mount: both mirrors are equipped with adaptive optics and are combined interferometrically to reach the resolution

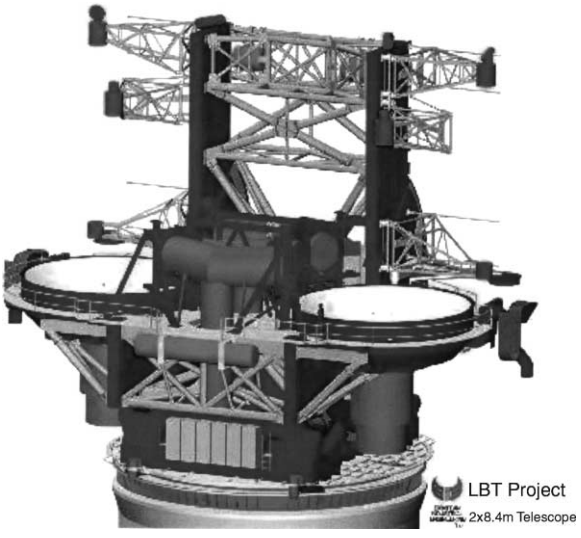


Figure 4 Picture of the Large Binocular Telescope (LBT) in construction on the Mount Graham in Arizona. It consists of two mirrors on the same mount. The images of the two mirrors are combined interferometrically to produce a high-resolution image in the direction of the line joining the centers of the mirrors (baseline).

of a 22.8 m mirror in the direction of the baseline. The telescope can be rotated to record images of the same astronomical target with different orientations of the baseline, i.e., the line joining the centers of the two mirrors. Finally the images must be processed by means of multiple images deconvolution methods to obtain a unique high-resolution image, equivalent to that produced by a 22.8 m mirror.

The applications of inverse problems with major social impact are in the area of medical imaging. The spectacular success of X-ray computerized tomography (CT), invented by GH Hounsfield at the beginning of the 1970s, has stimulated the development of other imaging modalities, based essentially on the same principle, such as positron emission tomography (PET) and single photon emission computerized tomography (SPECT). Computed images are also provided by magnetic resonance imaging (MRI). In such a case, however, the computational problem is rather simple since it consists essentially in Fourier transform inversion.

The basic principle of X-ray CT is described as a finely collimated source S (see Figure 5a) emitting a pencil of X-rays which propagates through the body along a straight line L up to a well collimated receiver R. If we know both the intensity I_0 of the source and the intensity I detected by the receiver, the logarithm of the ratio I_0/I is the integral of the linear attenuation function (roughly proportional to the density function of the body) along the line L. If the source and the receiver are moved along two parallel lines, having direction θ and defining the plane Π (in practice a planar slice of the body under consideration), one obtains the integrals of the linear attenuation function $f(\mathbf{x})$ along all parallel lines orthogonal to θ. The scanning variables in the plane Π are defined in Figure 5b: φ is the angle formed by the unit vector θ with the x-axis and s is the signed distance between the origin and the integration line L. The integrals along all parallel lines orthogonal to θ define a function of s which is called the projection of $f(\mathbf{x})$ in the direction θ:

$$P_\theta(s) = \int_{-\infty}^{+\infty} f(s\cos\varphi - u\sin\varphi, s\sin\varphi + u\cos\varphi)\mathrm{d}u \qquad [10]$$

The set of the projections of $f(\mathbf{x})$ for all possible θ is called the Radon transform of $f(\mathbf{x})$, in honor of the mathematician Johann Radon, who first investigated the problem of recovering a function of two variables from its line integrals. Today, such a problem is known as Radon transform inversion or object restoration from projections. Sampled values of the Radon transform, i.e., sampled values of the projections for a number of possible directions between 0 and π, are the outputs of the acquisition system of the medical equipment known as CT-scanners (Figure 6). Their data-processing system contains the implementation of an algorithm for Radon transform inversion and the final result is a set of computed images providing maps of slices of the human body.

The plot of the Radon transform of f in the (s, φ)-plane, obtained by representing its values as gray levels, is called the sinogram of f. In Figure 7, we give the image of the sinogram of a brain slice. The horizontal rows provide representations of the projections of the phantom. In the image, the angle φ defining the direction θ takes values between 0 and 2π

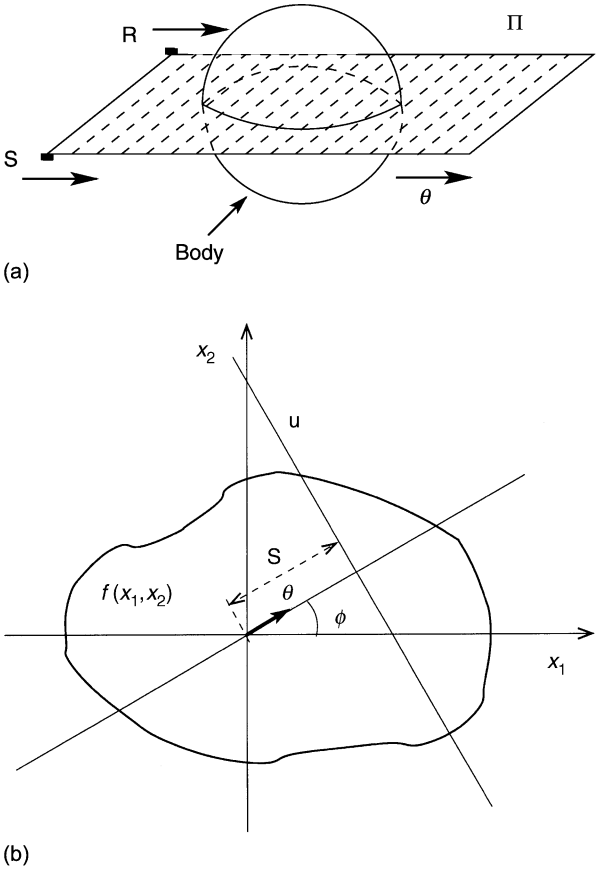


Figure 5 (a) Scheme of the scanning procedure in X-ray tomography: the source(S)–receiver(R) pair is moved along a direction θ orthogonal to the line S–R, defining the plane Π to be imaged; when the scanning has been completed the system is rotated by a certain angle, the scanning procedure is repeated, and so on. (Reproduced with permission from Bertero M and Boccacci P (1998) *Introduction to Inverse Problems in Imaging.* Bristol, UK: IOP Publishing.) (b) Definition of the variables used in the representation of the Radon transform; φ is the angle formed by the S–R line with the x_1-axis; s is the signed distance between the origin and the integration line (orthogonal to θ); and u is the coordinate on the integration line.

while the variable s takes values between $-a$ and a, if a is the radius of a disc containing the phantom. As is evident, the sinogram is symmetric with respect to the point $(0, \pi)$. Each point in the rectangle corresponds to a straight line crossing the phantom and the straight lines through a fixed point of the phantom describe a sinusoidal curve in the (s, φ)-plane. This property has given rise to the name sinogram.

The sinogram is, in a sense, the image of the body slice under investigation, as provided directly by a CT-scanner. However, an inspection of this image does not easily provide information about the body. Therefore, it must be processed to obtain a more significant image. Radon transform inversion is an example of inverse and ill-posed problem and its solution must be treated with extra care.

The crucial point in Radon transform inversion is the Fourier slice theorem, whose content is the following: the Fourier transform of the projection in the direction θ of the function f, is the Fourier transform of f on the straight line passing through the origin and having direction θ:

$$\hat{P}_\theta(\omega) = \hat{f}(\omega\theta) \qquad [11]$$

This theorem clarifies the information content of CT data: each projection provides the Fourier transform of the function f along a well-defined straight line in the plane of the spatial frequencies. It is also the starting point for deriving the basic algorithm of data inversion in tomography, the filtered back-projection (FBP). It is obtained from Fourier transform inversion in polar coordinates and is a two-step algorithm: the first step is a filtering of the projections by means of a ramp-filter; the second is the back-projection of the filtered sinogram. More precisely the two steps are as follows:

- *Filtering*: for each θ the projection $P_\theta(s)$ is replaced by a filtered projection given by

$$Q_\theta(s) = \int_{-\infty}^{+\infty} |\omega| \hat{P}_\theta(\omega) e^{i\omega s} \, \mathrm{d}\omega \qquad [12]$$

  where the multiplication by the ramp filter $|\omega|$ derives from the Jacobian of polar coordinates;
- *Back-projection*: this operation is the dual of the projection operation: indeed the projection assigns to a straight line L with coordinates (s, φ), hence to a point of the domain of definition of the Radon transform, the integral of f along L; on the other hand, the back-projection assigns the value of the Radon transform at (s, φ) to all points of the straight line L with the same coordinates (a pictorial representation of projection and back-projection is given in Figure 8); as a consequence the back-projection operator $R^\#$ assigns to each

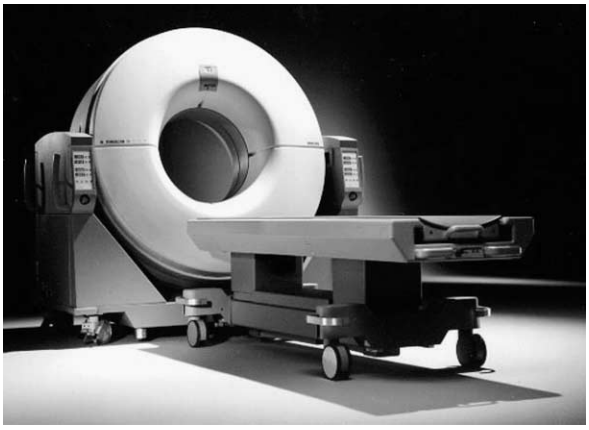
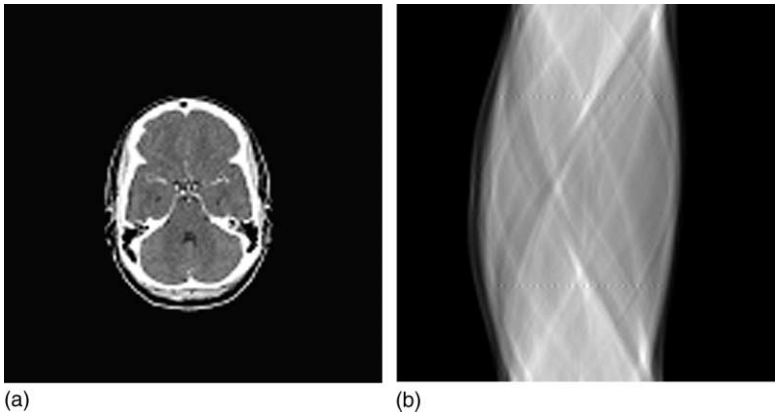


Figure 6 Picture of a scanner for X-ray tomography. The annular part contains the acquisition and scanning systems.



(a) (b)

Figure 7 (a) Picture of a brain slice; (b) the corresponding sinogram. As explained in the text, the sinogram is a gray-level representation of the Radon transform in the s, φ plane. Each horizontal line in (b) corresponds to a projection of (a).

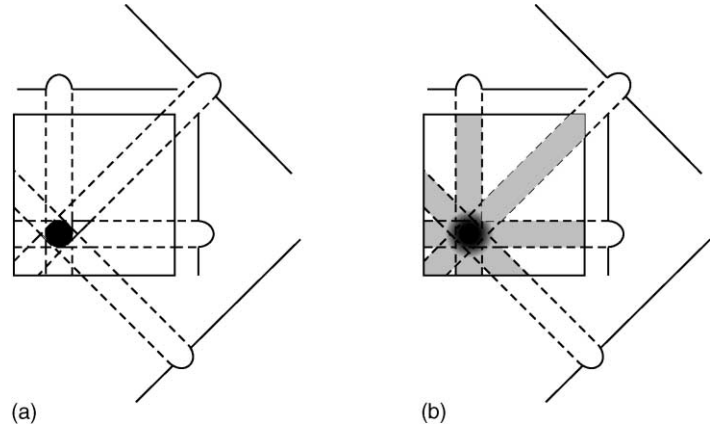


Figure 8 (a) The effect of the projection operation is shown by drawing the projections of a circular phantom corresponding to three directions. (b) shows the result obtained by applying the back-projection operation to the three projections given in (a). It is evident that the picture provided by this operation has a maximum on the domain of the circular phantom.

point $\mathbf{x}$ the sum (integral) of all values of the Radon transform corresponding to straight lines passing through $\mathbf{x}$:

$$R^{\#}P_{\theta}(\mathbf{x}) = \int_0^{2\pi} P_{\theta}(x\cos\varphi + y\sin\varphi)\mathrm{d}\varphi \qquad [13]$$

The back-projection operator, when applied directly to the projections as in eqn [13], provides a blurred image of f; in order to get a satisfactory restoration it must be applied to the filtered projections $Q_{\theta}(s)$. This point is illustrated in Figure 9.

We point out that the ramp filter introduced in eqn [12] amplifies the high frequency components of the projections hence the noise corrupting these components. This remark clarifies that Radon transform inversion is an ill-posed problem and requires some kind of regularization. This is obtained by introducing an additional low-pass filter which reduces the effect of the high frequencies. The most frequently used combination of ramp and low-pass filter is the so-called Shepp–Logan filter.

X-ray tomography is also called transmission computerized tomography (TCT) because the image is obtained by detecting the X-rays transmitted by the body. It provides information about anatomical details of human organs because the map of the linear attenuation function is essentially the map of the density of the tissues.

A different type of information is obtained by the so-called emission computerized tomography (ECT) which is based on the administration of radio-nuclide-labeled agents known as radio-pharmaceuticals. Their distribution in the body of the patient depends on factors such as blood flow, metabolic processes, etc. Then a map of this distribution is obtained by detecting the γ-rays produced by the decay of the radio-nuclides. Therefore, ECT yields functional information, in the sense that the images produced by ECT show the function of the tissues of the organs.

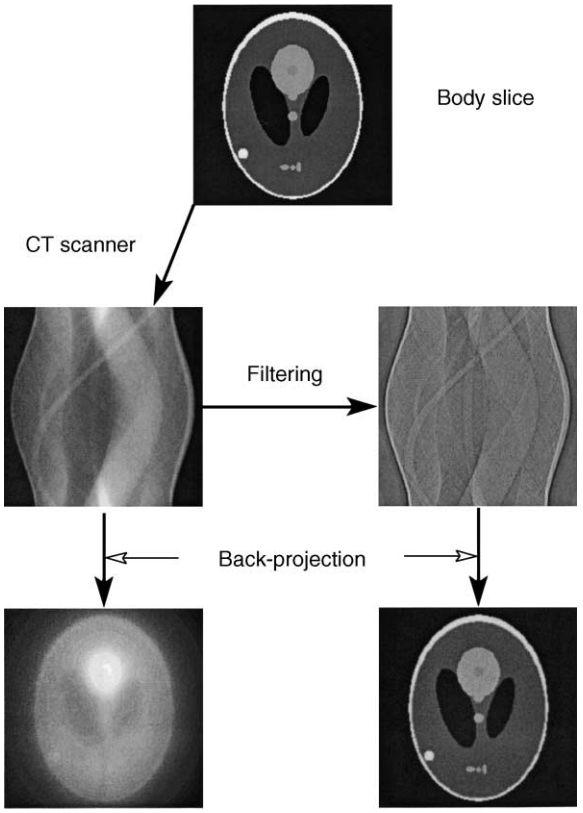


Figure 9 Illustrating the effect of the filtered back-projection. Shown here is: a slice of the so-called Shepp–Logan phantom; the corresponding sinogram (indicated by the arrow CT scanner); the filtered sinogram (indicated by the arrow filtering; the restorations obtained by applying the back-projection operation both to the original sinogram and to the filtered sinogram.

Two different modalities of ECT are usually considered:

- single photon emission computerized tomography (SPECT), which makes use of radio-isotopes where a single γ-ray is emitted per nuclear disintegration;
- positron emission tomography (PET), which makes use of β^+-emitters, where the final result of a nuclear disintegration is a pair of γ-rays, propagating in opposite directions, produced by the annihilation of the emitted positron in the tissues.

In both cases it is necessary to detect the γ-rays coming from well-defined regions of the body and, to this purpose, different methods are used in each case.

In SPECT, the discrimination of the γ-rays is obtained by means of a collimator, consisting of holes in a large lead slab covering the crystal detector face. In PET the collimation is obtained by means of pairs of detectors in coincidence. This technique, which is also called electronic collimation, is more accurate than the physical collimation used in SPECT and allows the design of systems with a great efficiency.

The basic reconstruction method in TCT, namely FBP, is often used in the reconstruction of SPECT and PET data. However, several corrections are necessary in practice. The principal ones are due to the collimator blur and to the scattering of photons in the body. If one develops a more accurate model of data acquisition by taking into account these effects, then the restoration problem implies the inversion of a very large matrix which is sparse and ill-conditioned. To this purpose iterative regularization methods are used such that the basic operation required at each step is matrix-vector multiplication. Other imaging modalities, which have been proposed for medical applications, are electrical impedance and microwave tomography. The underlying inverse problems are basically nonlinear so that the computational cost of the methods for their solution is, in general, too high for clinical applications. However the research in these areas is very active and, therefore, the situation could be improved in the near future.

See also

Tomography: Tomography and Optical Imaging.

Further Reading

Bertero M (1989) Linear inverse and ill-posed problems. In: Hawkes PW (ed.) *Advances in Electronics and Electron Physics*. vol. 75, pp. 1–120. New York: Academic Press.

Bertero M and Boccacci P (1998) *Introduction to Inverse Problems in Imaging*. Bristol, UK: IOP Publishing.

Bertero M and De Mol C (1996) Super-resolution by data inversion. In: Wolf E (ed.) *Progress in Optics*. vol. XXXVI, pp. 129–178. Amsterdam, Elsevier.

Engl HW, Hanke M and Neubauer A (1996) *Regularization of Inverse Problems*. Dordrecht, Germany: Kluwer.

Groetsch CW (1993) *Inverse Problems in Mathematical Sciences*. Braunsschweig, Germany: Vieweg.

Herman GT (ed.) (1979) *Image Reconstruction from Projections*. Berlin: Springer.

Herman GT (1980) *Image Reconstruction from Projections. The Fundamentals of Computerized Tomography*. New York: Academic Press.

Huang TS (ed.) (1975) *Picture Processing and Digital Filtering*. Berlin: Springer.

Kak AC and Slaney M (1988) *Principles of Computerized Tomographic Imaging*. New York: IEEE Press.

Krestel E (ed.) (1990) *Imaging Systems for Medical Diagnostics*. Berlin: Siemens Aktiengesellschaft.

Natterer F (1986) *The Mathematics of Computerized Tomography*. New York: Wiley.

Tikhonov AN and Arsenin VY (1977) *Solutions of Ill-Posed Problems*. Washington, DC: Winston/Wiley.

Adaptive Optics

C Pernechele, Astronomical Observatory of Padova, Padova, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The image quality of a perfect optical system is ideally limited by its diffraction figure. In the real world there are many effects which degrade that physical limit, resulting in poorer image quality at the focal plane. These effects may be static (e.g., time-independent, as, for example, error in the alignment of the optical train) or dynamic (e.g., variable with time, such as optical misalignment introduced by thermo-mechanical distortion). Active and adaptive optics, the topics of this article, are technologies able to correct for the optical aberrations induced by dynamical effects.

There are a variety of causes liable for dynamical image degradation, among these time-variable optical misalignments, pointing and tracking errors (for motorized instruments), and atmospheric perturbations. In practice, both atmospheric turbulence and misaligned instrument introduce random spatial and temporal perturbations on the optical beam, distorting the spherical (ideal) wavefront incoming on to the focal plane. This leads to loss of both the diffraction limited condition and image resolution. A typical situation for astronomical application is shown in Figure 1. A target (a star in this case) emits light whose wavefront, being far from the observing system, can be considered plane. When this wavefront passes through the atmosphere it is distorted with a temporal frequency of the order of hundreds of Hertz. This dynamically distorted wavefront fills the aperture of the optical system and will once more be distorted by the dynamic misalignment of the optical system itself. The final wavefront will be deformed with respect to the ideal (spherical) wavefront.

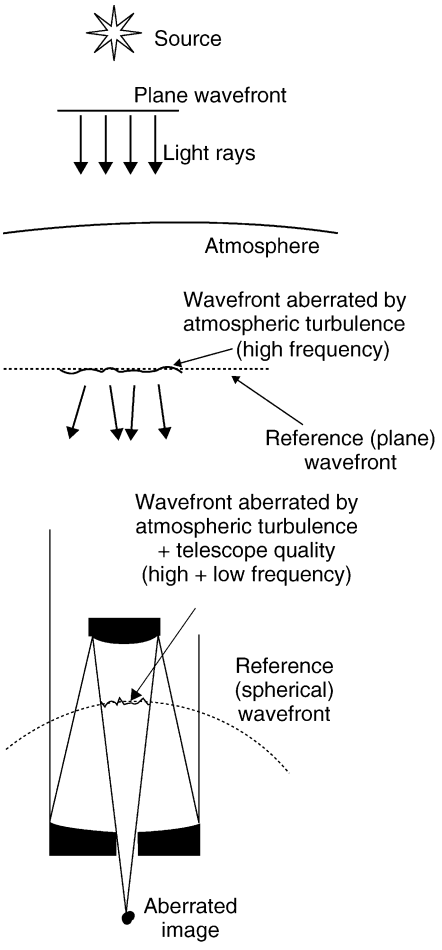


Figure 1 Dynamical wavefront aberration.

Dynamic optical misalignment may depend on gravitational and thermal deformation, while pointing error may appear as very slow (drift), slow (wander), or fast (jitter) tilt of the image on the focal plane. Atmospheric perturbations are produced by air turbulence and may be natural (as in the astronomical or satellite tracking cases) or induced by the experiment itself (thermal blooming in high power laser applications). All these effects appear on very different time-scales, as shown in Figure 2. As noticeable, atmospheric turbulence introduces aberrations of all orders, at a frequency above about 1 Hz. Pointing and tracking errors introduce an image tilt with different time-scale. Gravitational effects induce mechanical deformation of the optics, mainly in astronomical applications because of the massive components (the telescope primary mirror in particular). Thermal deformation produces optical misalignment among optical components mainly resulting in defocus. Figure 2 is purely illustrative, because some effects may produce wavefront aberrations at different frequencies than indicated there. For example, drift may be faster than indicated in the figure, and thermal blooming (a thermal effect) introduces aberrations at a higher frequency, and so on.

The correction of the wavefront errors is usually made by means of devices able to change the phase wave somewhere in the light path. Though the concept for correcting aberrated wavefront at slow or high frequency is the same, for example, the use of one or more deformable elements in the optical path, the technologies developed for each area are quite

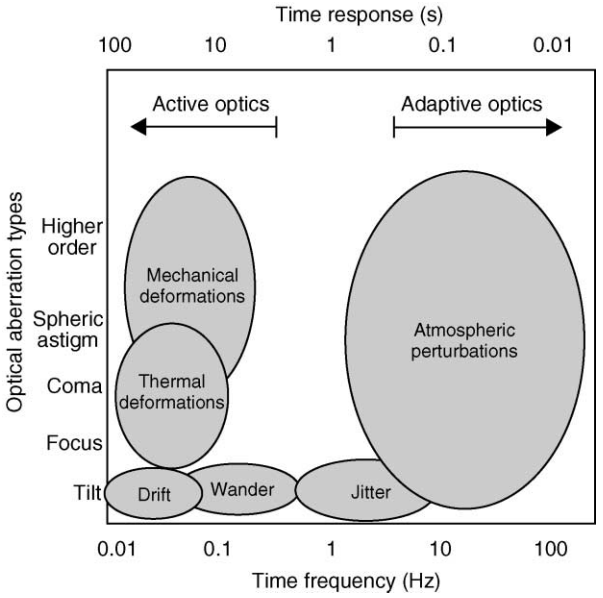


Figure 2 Wavefront aberration time-scale and associated aberrations.

different: variations at slow frequency (approximately below 0.1 Hz in Figure 2) are indicated as quasi-static errors and the techniques developed to overcome them are properly called active optics, while adaptive optics indicates the techniques to overcome the unwanted effects at high frequency (above about 1 Hz in frequency). The frequency value dividing the active from adaptive optics is not well defined and depends on the field of application and its scientific engineering communities. The important point is that a correct use of a combination of the two techniques may ultimately lead to bringing the imaging system close to its diffraction limit.

Slow Corrections: Active Optics

The analysis of the wavefront is required before its correction, and one or more deformable/movable mirrors and their control system, to operate the corrections. A typical system for low frequency correction is shown in Figure 3. A telescope contains three mirrors (M1, M2, and M3) and an instrument to detect its scientific target, usually placed on its optical axis. The dynamic deformations present in this kind of instrument are mainly third-order spherical (thermal deformation of M1), third-order astigmatism (a saddle-horse deformation of M1 due to gravity when the telescope is pointing away from the zenith), and third-order coma (due to gravitational decentering between M1 and M2). Since a telescope has a tracking system some pointing error may be present, causing motion of the image in the focal plane. The observation of an off-axis target (typically a star for astronomical applications) is made to measure the deformations of the incoming wavefront by means of a wavefront sensor (WS) system. This step is completed using some of the wavefront sensing technique described elsewhere in this encyclopedia. The WS must measure the information on the wavefront distortion with adequate spatial resolution to compensate for aberrations across the full aperture of the system. Once the wavefront distortion has been measured by the WS, the needed corrections are sent to the control system which apply the appropriate mirror deformation/movements. In the example of Figure 3 only the primary mirror (entrance pupil) is deformable and it is used for correcting spherical and astigmatism aberrations. M2 is used to compensate the decentering coma by means of lateral displacement. M3 may be tilted to correct for low-frequency pointing errors. In astronomical applications active optics is essential for telescopes with primary-mirror diameters larger than approximately four meters which cannot be manufactured rigidly at reasonable cost. Active optics requires a thin deformable or a segmented primary mirror maintained on the correct shape at every elevation. Using this method, low-frequency correction has been achieved to date in telescopes with primary mirrors of the order of ten meters.

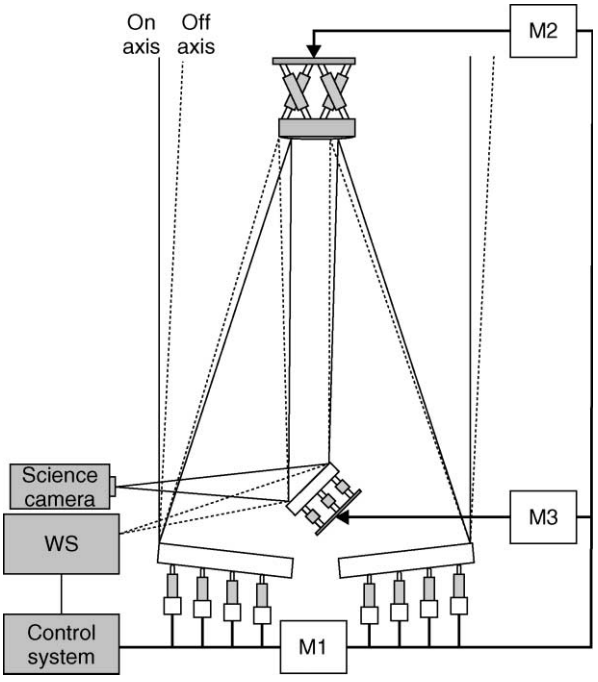


Figure 3 A typical system for correcting low-frequency dynamical aberrations (active optics).

Fast Corrections: Adaptive Optics

It is clear that corrections at high frequency are impossible using the approach discussed above, because of the large dimensions and weight of the mirror and the low response time required. For an adaptive correction it is necessary to image a smaller system pupil in the optical train and make the active correction of the wavefront. A typical configuration is shown in Figure 4. The light from the telescope focal plane (corrected for low frequency errors by means of active optics) is collimated and a pupil is located in this beam. Behind the active mirror a beamsplitter is used to send part of the light into the WS and measure wavefront distortions. The analysis results are sent to the control system which performs the active mirror deformation for the wavefront compensation. If the WS and the control system are sufficiently fast (typically faster than the time-variations of the wavefront distortion), this system is able to produce a plane wavefront which can be focused at the diffraction limit of the system. For this reason

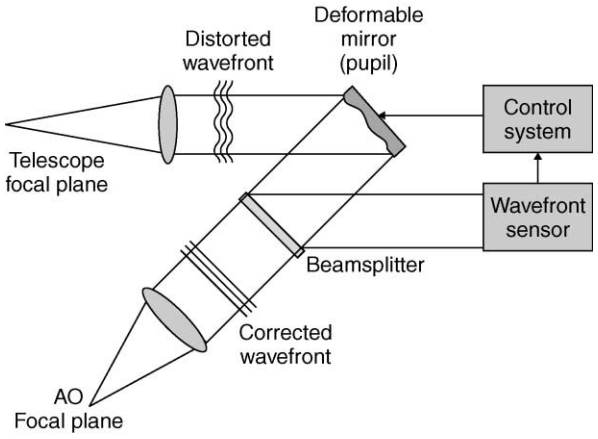


Figure 4 A typical setup for correcting high-frequency dynamical aberrations (adaptive optics).

adaptive optics are often referred to as real-time system.

The seminal paper on the adaptive optics concept was written by H.W. Babcock of the Mount Wilson and Palomar observatories in 1953. Since then a great improvement has been made, especially in military and astronomical fields. In recent years, however, many applications of adaptive optics have developed in fields different from traditional ones, such as precision manufacturing by means of high-power lasers, telecommunications, aero-optics, medical physics, and ophthalmology.

We describe adaptive optics with particular reference to astronomical applications. We provide a brief description of other applications at the end of this article.

Imaging Through the Atmosphere

The optical effects of the atmospheric turbulence are due to local temperature variations that induce fluctuations in the refractive index of air. The refractive index changes depend on the air density as well as the range of the temperature variations. Air density is highest at sea level and decreases exponentially with altitude. Optical effects of turbulence decrease with the altitude of the site, the reason for locating astronomical telescopes on mountains. The effects of turbulence on a point source image are depicted in Figure 5. Hereafter we consider an ideal telescope, for example, one without need for active correction at low frequencies. We consider the source 1 (continous line). The radiation emitted by the source is a spherical wavefront that can be treated as a plane, because of the very long distance to the Earth's atmosphere.

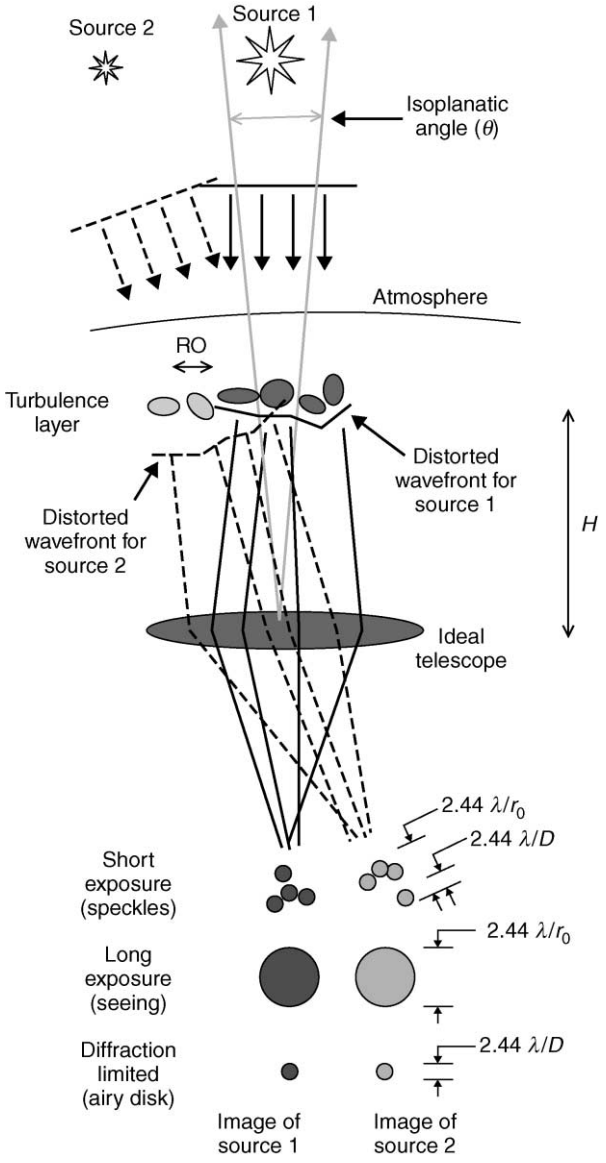


Figure 5 Imaging through the atmosphere.

In absence of wavefront distortion, the angular diameter of the image (Airy disk), is approximately 2.44 λ/D, with λ the wavelength of the observation and D the telescope aperture (the mirror diameter). In this case the image is said to be diffraction limited, and it is the best possible performance for the observation. This is the performance of a space-based ideal telescope.

In the presence of the atmosphere the incoming wavefront is distorted by the turbulence and when it reaches the telescope aperture it is no longer a plane (see in Figure 5). The so-called Fried's parameter r_0 (also called atmospheric coherence length) is the most common parameter used to quantify the distortion effects of atmospheric turbulence. It is defined as the

distance over which the wavefront phase is highly correlated (in this way it provides a measure of the turbulence scale size). For an undistorted plane wave, r_0 is infinity, while for atmospheric distortion it typically varies between 5 cm (bad sites) and 20 cm (excellent sites). We depict Fried's parameter schematically in Figure 5, as bubbles of dimension r_0 in the turbulent layer (referred to as Fried's cells). r_0 is a statistical parameter, which depends on time and wavelength ($r_0 = r_0(t, \lambda^{6/5})$). Its variability with time may be considerable, sometimes changing by a factor of 2 within a few seconds. A small telescope, with a diameter less than r_0, will always operate at this diffraction limit. When the r_0 is smaller than the telescope aperture, the angular size of the image is different if we consider long- or short-time exposures. If the exposure time is very short (less than hundredths of a second) it is possible to image a structure in the focal plane (see short exposure image in Figure 5). This structure, which varies with time, is due to high-order wavefront distortions that occur within the aperture. The short exposure image is broken into a large number of speckles of diameter $2.44\,\lambda/D$, which are randomly distributed over a circular region of approximately $2.44\,\lambda/r_0$ in angular diameter. A lower-order wavefront distortion is correlated with the size of the largest disturbances produced by atmospheric turbulence (the so-called outer scale) which varies considerably (from 1 m to over 1 km). Such disturbances produce a time-dependent tilt of the overall wavefront that causes fluctuations in the angle of arrival of the light from a star and randomly displace the image. For D of the order of the Fried parameter r_0, image motion dominates, and a simple tip-tilt mirror (not just deformable) provides useful correction. For D greater than r_0, speckles dominate and a deformable mirror is necessary to account for the wavefront distortion. For long exposures the dimension of the image is determined by the ratio λ/r_0 rather than λ/D. The angular resolution is diminished due to the atmospheric turbulence (see Figure 5). For a 4 m telescope, atmospheric distortion degrades the spatial resolution by more than one order of magnitude compared to the diffraction limit, and the intensity at the center of the star image is lowered by a factor of 100 or more. Even at the best astronomical sites, ground-based telescopes observing at visible wavelengths cannot, because of atmospheric turbulence alone, achieve an angular resolution better than telescopes of about 20 cm diameter.

Most of the distortions imposed by the atmosphere are in the wave-phase and an initially plane wavefront traveling 20 km through the turbulent atmosphere accumulates, across the diameter of a large telescope, phase errors of a few micrometers.

Turbulence is distributed along the propagation path through the atmosphere, and then the wavefront errors become uncorrelated as the field angle increases. This effect is known as angular isoplanatism and is one of the major limitations to the astronomical adaptive optics. The structure becomes completely uncorrelated out of the isoplanatic angle, approximately equal to r_0/H, where H is the altitude of the turbulence layer. This decorrelation is sketched in Figure 5, where the short exposure images for objects 1 and 2 have different speckles structure.

In Earth's atmosphere significant turbulence occurs at altitudes up to 20 km, often with large peaks in the vicinity of the tropopause, at around 10 km there exist, in practice, more than one parameter H in Figure 5, one for each turbulence layer. This three-dimensional distribution of the turbulence considerably restricts the angle over which adaptive compensation is effective and leads to the design of multiconjugate adaptive optics, as we next explain.

Adaptive System Design

As stated above, an active system is composed of at least one wavefront sensor and one deformable mirror. Typically, the wavefront is corrected by making independent tip, tilt, and piston movements of the deformable mirror surface, as shown in Figure 6. As a rule of thumb, a deformable mirror used for atmospheric correction should have one actuator for each Fried's cell, such that the total number of actuators should be equal to the number of Fried's cells filling the telescope aperture, i.e., D^2/r_0^2. For instance, a near-perfect correction for an observation in visible light (0.6 μm) with an 8 m telescope placed in a good astronomical site ($r_0 = 20$ cm), would require a deformable mirror with approximately 6400 actuators in its adaptive optics system. Recalling that r_0 is strongly dependent on the wavelength (i.e., proportional to $\lambda^{6/5}$), similar performance in the infrared at 2 μm requires only 250 actuators. This is one reason why adaptive correction is easier in the infrared. A large number

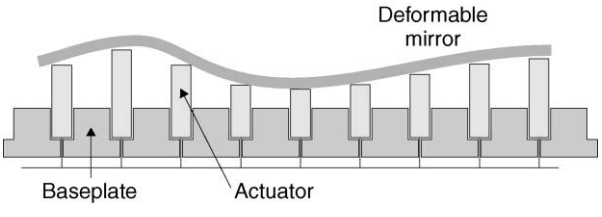


Figure 6 A sketch of a deformable mirror.

of Fried's cells requires a similarly large number of subapertures in the wavefront sensor. Commonly used WSs are based on modifications of the classic Hartmann screen test or of curvature type. In the former case the adaptive device is made with individual piezoelectric actuators (see Figure 6), while in the latter the correction is achieved with a bimorph adaptive mirror, made of two bonded piezoelectric plates. With both methods, wavefront sensing is done on a reference off-axis object (Figure 3), or even on the observed object itself (Figure 4) if it is bright enough. In the former case the angular distance of the reference object must fall inside the isoplanatic angle and must be sufficiently luminous. In typical astronomical sites the isoplanatic angle at a wavelength of 0.6 μm is ~5 arcsec, while it can be ~20 arcsec at 2 μm. This is an additional reason why the adaptive optics system is most efficient in the infrared range. Correction in the visible requires a reference star approximately 25 times brighter than in the infrared. Most current astronomical systems are designed to provide diffraction-limited images in the near-infrared (1 to 2 μm) with the capability for partial correction in the visible. However, some military systems for satellite observations do provide full correction in the visible on at least 1 m class telescopes.

The control system is generally a specialized computer that calculates, from the WS measurements, the commands to send the actuators of the deformable mirror. The calculation must be performed quickly (within 0.5 to 1 ms), otherwise the state of the atmosphere may have changed making the wavefront correction inaccurate. The required computing power can exceed several hundred million operations for each command set sent to a 250 actuator deformable mirror. As in active optics systems, zonal or modal control methods are used. In zonal control, each zone or segment of the mirror is controlled independently by wavefront signals that are measured for the subaperture of that zone. In modal control, the wavefront is expressed as the best-fit linear combination of polynomial modes of the wavefront distorted by the atmospheric perturbations.

Future Development

As explained above, a convenient method for measuring the wavefront distortion is to look for an off-axis reference star (also known as natural guide star, NGS) within the isoplanatic angle. However, it is generally impossible to find a bright reference star close to an arbitrary astronomical target. Conditions are far better in the infrared than in the visible because atmospheric turbulence (and in particular its high spatial frequencies) is, for a given AO correction, less pronounced at longer wavelengths. The spatial and temporal sampling of the disturbed wavefront can be reduced, which in turn can permit the use of fainter reference stars. Coupled with the larger isoplanatic angle in the infrared, this gives a much better outlook for AO correction than in the visible.

Nevertheless, even for observations at 2.2 μm, the sky coverage achievable by this technique (equal to the probability of finding a suitable reference star in the isoplanatic patch around the chosen target) is of the order of 0.5 to 1%. It is therefore unsurprising that most scientific applications of AO have to date been limited to the study of objects which provide their own reference object. The most promising way to overcome the isoplanatic angle limitation is by the use of artificial reference stars, also referred to as laser guide stars (LGS). These are patches of light created by the back-scattering of pulsed laser light of sodium atoms in the high mesosphere or by molecules and particles located in the low stratosphere. The laser beam is focused at an altitude of about 90 km in the first case (sodium resonance) and 10 to 20 km in the second case (Rayleigh diffusion). Such an artificial reference star can be created as close to the astronomical target as desired, and a wavefront sensor measuring the scattered laser light is used to correct the wavefront aberrations on the target object.

Military laboratories have reported the successful operation of adaptive optics devices at visible wavelengths using laser guide star on both a 60 cm telescope and a 1.5 meter telescope, achieving an image resolution of 0.15 arcsec. Nevertheless, there are still physical limitations of a LGS. The first problem, focus anisoplanatism, or cone-effect, was realized long ago. Since the artificial star is created at a relatively low altitude, back-scattered light collected by the telescope forms a conical beam, which does not cross the same turbulence-layer areas as light from the astronomical target. This leads to a phase estimation error, which in principle can be solved by the simultaneous use of several laser guide stars surrounding the targeted object. A more significant problem is image motion or tilt determination. Since the paths of the outgoing and returning light rays are the same, the centroid of the artificial light appears to be stationary in the sky, while the apparent position of an astronomical source suffers lateral motions (also known as tip/tilt). The simplest solution is to supplement the AO system and LGS with a tip/tilt corrector set on a (generally) faint close NGS. Performance is then limited by poor photon statistics in the tip/tilt correction.

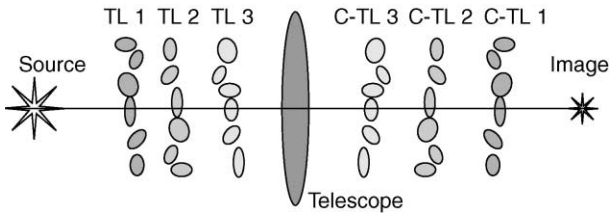


Figure 7 Conjugated planes of turbulent layers.

Adaptive optics with a multicolor laser probe is another concept that has been studied to solve the tilt determination problem of laser beacon based AO. Only applicable to sodium resonant scattering at 90 km, it excites different states of the sodium atoms and makes use of the slight variation in the refraction index of air with wavelength. Its main drawback is the limited returned flux, due to the saturation of mesospheric sodium layer.

The quality of a site for adaptive optics is determined not only by the median seeing, but also by the height and thickness of the turbulent layers (often only a few thin layers dominate). Some adaptive-optics systems can be adapted to correct for turbulence originating at a particular height by using several reference stars. This technique is known as multiconjugate adaptive optics (MCAO) and is still in his infancy. This concept deals with a 3-D view of the atmosphere, and for this reason it is also called turbulence tomography. Several wavefront sensors are needed for tomography. Turbulence tomography is also useful to correct for the cone effect when using LGS. When several deformable mirrors are placed on the appropriate conjugate planes, as shown in Figure 7, anisoplanatism reduces and the size of the corrected field of view increases. The situation is roughly depicted in Figure 7, where each turbulent layer (indicated as TL) has a proper conjugated plane (C-TL).

Another challenging optical solution is one of utilizing large secondary adaptive mirrors. This application seems especially interesting at thermal wavelengths, where any additional mirror raises the significant thermal background detected by instruments.

Other Adaptive Optics Applications

In recent years the adaptive optics technique has been developed in many applications beside astronomy and the military, for example, as in high-power laser welding or cutting, in ophthalmology and in aero-optics.

In high-power laser applications, thermal blooming takes place when laser light is absorbed by the gases and particles that make up the atmosphere. Since energy is conserved, the energy lost by the laser (its intensity diminishes) is gained by the atmosphere that heats up. This inhomogeneity at the boundary of the laser beam generates wavefront distortions by turbulence, as shown in Figure 8. This means that laser light that passes through the air at a later time interacts with a different atmosphere than the light that passed by earlier. The net effect is that, in the case of a Gaussian laser beam, light following a wavefront which passed at an earlier time, 'sees' a concave lens that deflects the successive rays out of the beam (away from the axis) causing the beam to appear to grow in radius or 'bloom'. Adaptive optics techniques for real-time compensation of blooming in laser beams were developed in the late 1960s. Given the high coherence and monochromaticity of a laser, the technique is more straightforward than for more common white-light applications of the astronomical applications.

Density fluctuations in a compressible shear layer are sufficiently high to cause a time varying index of refraction of the layer. This situation is common in a turbulent flowfield, and the study of the wavefront aberrations induced by this flowfield, when it is of relatively short propagation length in the near-field of the optical system, is referred to as aero-optics. Turbulent flowfield induces various types of degradations in the performances of an optical system. For example the jitter and defocusing of laser-based weapons cause a reduction in the amount of energy delivered to the target. Tracking systems experience bore-site errors and imaging systems experience blurring. Recent work suggests that adaptive optics compensation may successfully overcome these problems.

In ophthalmology the adaptive technique finds application in studying the retina of the eye. In this case the eye itself produces wavefront aberration, and the retina is seen as blurred. Techniques in ophthalmic laser surgery are being developed to use adaptive optics to correct for detected distortions.

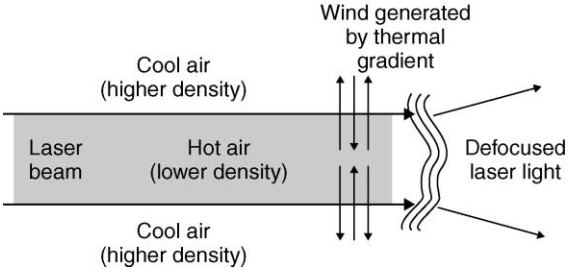


Figure 8 Thermal blooming effect in high-power laser applications.

In the field of telecommunications, studies are in progress for using adaptive techniques in data transmission between ground-based stations and satellites.

See also

Environmental Measurements: Doppler Lidar; Optical Transmission and Scatter of the Atmosphere. **Instrumentation:** Astronomical Instrumentation; Telescopes. **Phase Control:** Wavefront Coding; Wavefront Sensors and Control (Imaging Through Turbulence).

Further Reading

Beckers JM (1993) Adaptive optics for astronomy: principles, performance and application. *ARAA* 31: 13–62.

Fried DL (1965) Statistics of a geometric representation of wavefront distortion. *JOSA* 55: 1427–1435.

Hardy JW (1998) *Adaptive Optics for Astronomical Telescopes*. New York: Oxford University Press.

Jumper EJ and Fitzgerald EJ (2001) Recent advances in aero-optics. *Progress in Aerospace Sciences* 37: 299–339.

Liang J, Williams DR and Miller DT (1997) Supernormal vision and high-resolution retinal imaging through adaptive optics. *JOSA A* 14(11): 2884–2892.

Lukin VP (1995) *Atmospheric Adaptive Optics*. Bellingham: SPIE Optical Engineering Press.

Ragazzoni R, Marchetti E and Valente G (2000) Adaptive-optics corrections available for the whole sky. *Nature* 403(6765): 54–56.

Roddier F (1999) *Adaptive Optics in Astronomy*. Cambridge: Cambridge University Press.

Roggeman M and Welsh B (1996) *Imaging Through Turbulence*. Boca Raton, FL: CRC Press.

Tyson RK (2000) *Introduction to Adaptive Optics*. Bellingham: SPIE Optical Engineering Press.

Welford W (1974) *Aberrations of the Symmetrical Optical System*. New York: Academic Press.

Wilson RW and Jenkins CR (1996) Adaptive optics for astronomy: theoretical performance and limitations. *MNRAS* 268: 39.

Hyperspectral Imaging

M L Huebschman, R A Schultz and H R Garner, University of Texas, Dallas, TX, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Hyperspectral imaging techniques on a microscopic level have paralleled the hyperspectral remote astronomical sensing which has developed over the years. It is especially apparent in remote sensing of our own planet's land masses and atmosphere, such as in the LANDSAT satellite program. In biology, biochemistry, and medicine, hyperspectral imaging and multispectral imaging are often interchanged, although there is a distinct difference in the two techniques.

A multispectral imager produces images of each of a number of discrete but relatively wide spectral bands. Multispectral imagers employ several wide bandpass filters to collect the data for the images. A hyperspectral imager produces images over a contiguous range of spectral bands and, thus, offers the potential for spectroscopic analysis of data. Hyperspectral imagers may use gratings or interferometers to produce the contiguous range of spectral bands to record.

Fluorescent dyes (also referred to as fluorophores or fluorochromes) and recently quantum dots, as markers, are a very powerful tool for analyzing organic tissues for researchers in cytology (study of cell function), cytogenetics (study of the human chromosomes), histology (study of structure and function of tissues), microarray scanning (analysis of concentrations of multiple varied probes), and biological material science. Without the correlation of the spectral and spatial data in biological images, there would be little useful information for the researcher. Therefore, hyperspectral imaging in the biological sciences must combine imaging, spectroscopy, and fluorescent probe techniques. Such imaging requires high spatial resolution to present a true physical picture, and the spectroscopy must also have high spectral resolution in order to determine the quantities of biochemical probes present. The correlation must provide the color-position relation for identification of meaningful information.

This description will begin with an initial review of multispectral imaging to lead into a discussion of hyperspectral imaging. The discussion will examine aspects of hyperspectral imaging based on an interferometric method and on a grating dispersion method similar to the methods in multispectral imaging.

Image Parameters

The analysis of fluorophores involves the excitation of the dye with a shorter wavelength and thus higher energy than the emission spectrum. The use of the appropriate excitation wavelength will yield a maximum emission intensity at a characteristic wavelength. The quality of a fluorescent image generally combines contrast and brightness. Contrast depends on the ratio of emission intensity to that of background light. Consequently, the use of filters to block the excitation wavelength and prevent it from overwhelming the field is critical in providing high contrast for the detection of fluorophore emission. However, the use of narrow filters to restrict the background can severely limit the brightness of the fluorescent image. This is due to the barrier filters allowing only a portion of the actual emission spectrum to pass. Brightness is an intrinsic property of each individual fluorophore, but is also dependent on the ability to illuminate with the peak excitation wavelength for that fluorophore.

The balance between contrast and brightness may be further complicated by applications designed to facilitate the evaluation of multiple fluorophores. The selection of appropriate excitation and emission filters within a single set to allow for the simultaneous viewing of multiple fluorophores can be very inhibiting. Although quad-band beamsplitters are currently available, commercial advances to date have yielded only triple-pass combinations for simultaneous detection, such as one of which permits the simultaneous detection of the widely used fluorophore combination fluorescein-5-isothiocyanate (FITC), Texas Red and 4′,6-diamidino-2-phenylindole (DAPI) (e.g., Chroma Technology Corp.).

Although the capture of digital images has propelled this field, efforts to collect multicolored images as digital files also suffer from technical limitations. Color CCD cameras often utilize filters or liquid crystals to segregate the red, green, and blue wavelengths onto the imaging chip. The employment of these cameras is restricted to resolving colors as discriminated by the filters used. For scientific applications, then, the greatest sensitivity and selectivity has been achieved through black and white CCD cameras with optimized filters for a specific application. While this approach is widely used, it suffers from the inherent limits of the spectra of data that can be analyzed since it is defined by the characteristics of the optical filters employed. Thus, the system is restricted to the use of fluorophores that can be effectively resolved from each other by filters, and the filters selected limit the data collected to a portion of the total emission spectrum for any one fluorophore.

Multispectral Imaging

Multispectral imaging systems in biomedical research are composed of a microscope, filters and/or interferometer for spectral selection, a charged coupled device (CCD) camera to record the data, and computer control for acquisition, analysis, and display (Figure 1a). The introduction of CCD cameras has offered a means to document results and capture images as digital computer files. This has facilitated the analysis of results with more and more complex probe sets and has allowed the use of more sophisticated computer algorithms for data analysis. The other major component of the multispectral imaging system is a light source to excite the sample's fluorescent dyes whose emissions are selected by filters or an interferometer for recording on the CCD camera.

An example of the use of multispectral imaging is in the identification of the 24 human chromosomes. A scheme was presented for the analysis of human chromosomes by differentially labeling each chromosome with a unique combination of 5 fluorophore-labeled nucleotides (building blocks of DNA and RNA, consisting of a nitrogenous base, a five-carbon sugar, and a phosphate group). These complex probes are simultaneously hybridized to metaphase chromosomes and analyzed to determine which of the fluorophores are present. Combinations of one, two, or three of the five fluorophores are enough to distinguish a particular chromosome from the others.

An interferometric multispectral system used to perform this type of analysis is the Spectral Karyotyping (chromosome typing) or SKY System (Figure 1a). This technology does not represent hyperspectral imaging, but is presented for comparison. Illumination light passes through a triple band-pass dichroic filter which allows for the simultaneous excitation of all the fluorochromes in, for instance, Speicher's scheme. Modern multi-band pass filters add little to image distortion while effectively blocking unwanted autofluorescence or excitation illumination. The fluorescence emissions from the sample are split into two separate beams by the interferometer. The beams are then recombined with an optical path difference (OPD) between them such that the two beams are no longer in register and interfere with each other. By adjusting the OPD, a particular wavelength band can be selected for, which results in constructive interference and thus can be collected by the CCD camera pixels. After adjusting the OPD to record

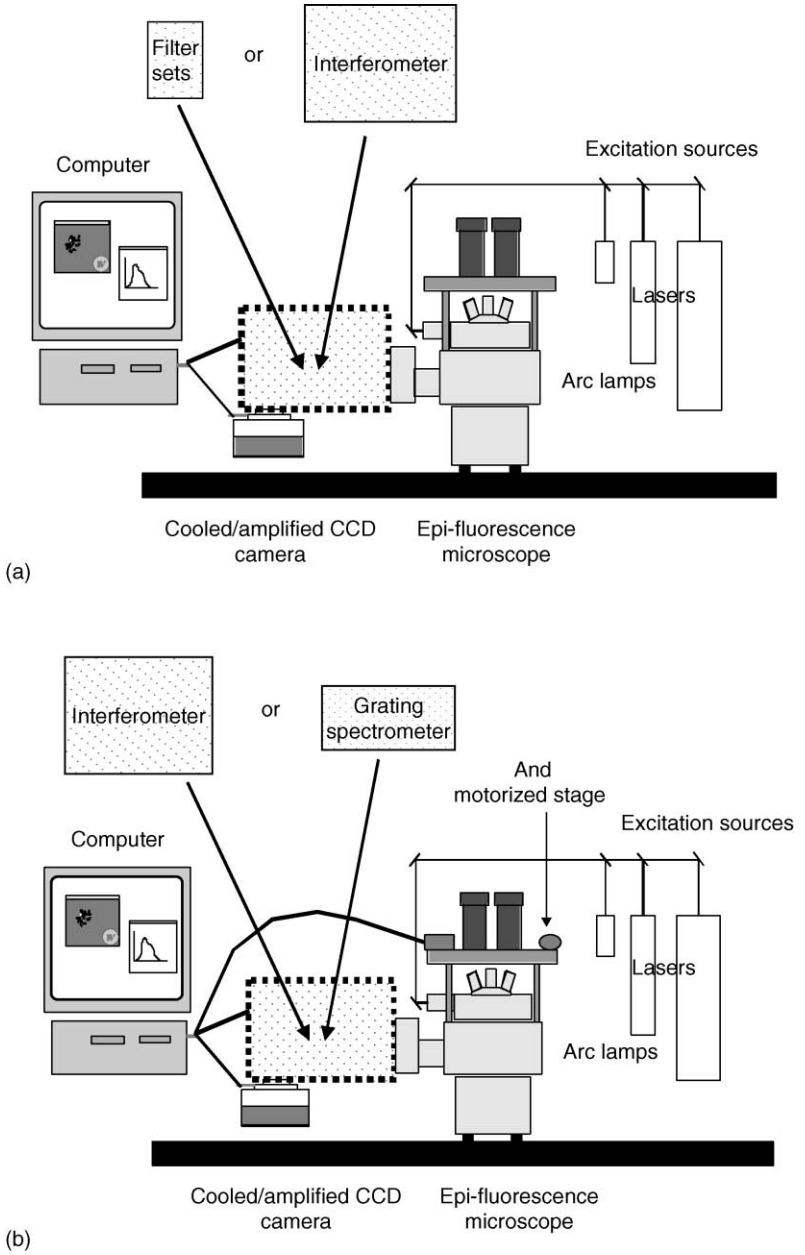


Figure 1 The instruments for multispectral and hyperspectral imaging are similar. (a) Multispectral imagers use interferometers or bandpass filters to collect noncontiguous emission wavelengths. (b) Hyperspectral imagers employ interferometers or gratings to collect a contiguous wavelength band. Computers, microscopes, CCD cameras, and excitation light sources are common to all the techniques. The systems which use filter sets or interferometers use motorized filter placement systems or automated interferometers. The gratings systems employ a motorized stage for the sample movement. Adapted with permission from Schultz RA, Nielsen T, Zavaleta JR, Ruch R, Wyatt R and Garner HR (2001) Hyperspectral imaging: a novel approach for microscopic analysis. *Cytometry* 43: 239–247.

the 5 dye wavelengths, each pixel's 5 records are then compared to the known probe-labeling scheme to identify the corresponding chromosome material. The image of the metaphase chromosome spread can be karyotyped using specific pseudo-colors for maximal visual discrimination of the chromosomes.

Another nonhyperspectral method is the Multi-Fluorescence *In Situ* Hybridization or M-FISH System. It uses five single-band-pass excitation/emission filter sets where each filter set is specific for one of the five fluorochromes used in the combinatorial labeling process (Figure 1a). Again, five separate exposures are taken (one exposure for each filter set) and combined to create a multicolor image. Most M-FISH systems now have motorized filter set selectors which minimize any image registration

shifts that may occur between exposures. For each pixel, the presence/absence of signal from each of the five fluorochromes is assessed and then matched to the probe-coding scheme to determine which chromosome material is present. As with the SKY System, an image of the metaphase spread is converted into a karyotype and specific pseudo-colors are used to maximize visual discrimination of the chromosomes.

Hyperspectral Imaging

As has been stated, the main difference between multi- and hyper-spectral imaging is the collection of a broad contiguous band of wavelengths. The apparatus and methods for hyperspectral imaging are similar to multispectral imaging (Figure 1). Instead of collecting sections of the emission spectrum using filters or interferometers, hyperspectral imaging collects a continuous emission spectrum by dispersing the wavelength of light using gratings and measuring the wavelength intensity, or by dispersing Fourier components of light using interferometers and measuring interference intensity (Figure 1b). Multispectral images are limited to measuring total intensity within the selected wavelength band. Determining the abundance of more than one dye per pixel is less accurate when only a small segment of each dyes' spectrum is passed by the emission filter. This has not been considered a limitation in the past because most research depended on just identifying the presence and not the abundance of several spectrally broad and overlapping dyes which also occupied relatively large spatial areas. With the advent of smaller and smaller featured microarrays, the desire to use more dyes in identification schemes, and the use of quantum-dot markers which can be used to indicate both genetic makeup and quantity, the limitations of multispectral imaging has led to a move to hyperspectral imaging.

In contrast to the fluorescence multispectral imaging approaches, a hyperspectral image is fully three-dimensional. Every XY spatial pixel of the image has an arbitrarily large number of wavelengths allied with it. An intensity cube of data is collected. Two of the axes of the cube are spatial and one axis is wavelength (see Figure 2). At each of the 3D points, there is an associated intensity of emitted energy. Therefore, it is a four-dimensional information space. The individual intensity signature of differing emitting bodies can then be separated to give a complete view. Spectral decomposition is a mathematical/algorithmical process by which individual emission spectra of the dyes can be uniquely separated from a composite (full spectral) image by knowing the signature of each dye emission spectra component measured separately. Spectral decomposition may be performed using standard algorithms with codes which are readily available.

Interferometric Hyperspectral Imaging

The system architecture of the interferometric hyperspectral imager resembles that of the SKY System (Figure 1). But, instead of just sequencing the interferometer to select a set of specific wavelengths to record as in the SKY System, the interferometer in the hyperspectral imager is sequenced through

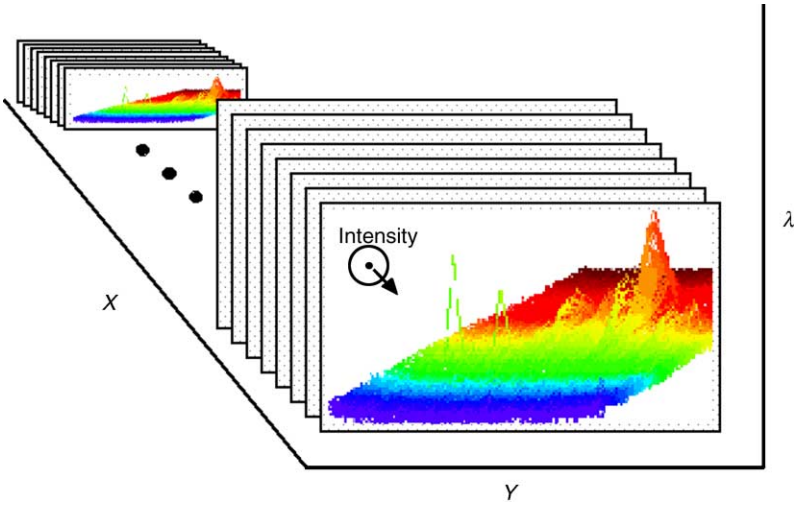


Figure 2 Intensity data cube. In this data cube, each CCD exposure captures a measurement in intensity (number of counts) by wavelength for a small spatial area in Y and ΔX. Many displacements in X with subsequent CCD acquisitions yield a 'cube' of data in X, Y, and λ. The intensity counts may be thought of as another dimension, $I(x,y,\lambda)$. Another form of the data cube is with each CCD acquisition comprising intensity counts for a spatial X and Y area for a small $\Delta\lambda$ and multiple displacements in λ for each acquisition.

discrete optical path differences to record the intensity at each pixel for each OPD. Fourier spectroscopy techniques are used to transform the OPD intensities at each pixel into the wavelength intensities at each pixel. Each pixel may be thought of as a Fourier spectrometer which is not dependent on any other pixel. Then, spectral decomposition algorithms can be run to determine which dyes are present (see the next section).

Like the multifilter methods, the interferometer approach provides a real two-dimensional image that can be observed, focused, and optimized prior to measurement. But like the filtering methods, the excitation source must be filtered for sufficient contrast to view the fluorescent emissions. Interferometers have high optical throughput, relative to other methods, and can have high and variable spectral resolution. The spectral range can cover from the ultraviolet to the infrared. The high and variable spectral resolution allows the user to trade off sensitivity, spectral resolution, and acquisition time in a way that is not possible with any other technique. And interferometers do not cause polarization of light as a filter may.

However, interferometric instruments are susceptible to image registration and lateral coherence difficulties. They are limited by the extreme precision required of the optical components which results in the continuous need for monitoring and adjustment, since aberrations are not tolerated in interferometric instruments. The throughput is reduced by loss due to the beamsplitter design and from the mirrors. Additionally, the Fourier transform instrument deconvolutes based on a two-step process; the correlation of the light passed by an interferometer followed by an inverse Fourier transform in software.

To mitigate the extreme precision and sensitivity of the most interferometric systems, Sagnac interferometers have been employed because of their common path intrinsic stability. The common path compensates any misalignment, vibration, or temperature affect. The OPD is provided by the Sagnac affect. When the Sagnac interferometer is rotated, the path traveled in one direction will be different from that of the light on the common path in the other direction. Although, this interferometer method is trading one difficulty for another, it is somewhat easier to manage the rotation of a physical system than to maintain the exacting precision of mirror alignments in a standard interferometer.

Spatial Resolution

The spatial resolution of a Sagnac Hyperspectral Imager is set by the size of the pixels of the CCD camera and the microscope system magnification, since the interferometric hyperspectral viewer sees the entire field of view at once. Let the x- and y- dimensions of the CCD pixel be δx and δy, respectively, and the magnification be M, then the spatial resolutions in the x and y directions are $2\delta x/M$ and $2\delta y/M$, respectively. For CCD pixel sizes of 10 microns square, binning together two pixels in the x- and two pixels in the y- direction, and an overall magnification of 60, the x and y resolutions are both 0.66 microns.

Field of View

For N pixel rows and C pixel columns, the field of view will be $N\delta y/M$ and $C\delta x/M$ for the length of the area along the y- and x- axis, respectively. For a CCD array of 1536 columns by 1024 rows, and the parameters above, the field of view is 171 microns by 256 microns, x and y, respectively.

The sensitivity, dynamic range, and spectral response are set by the CCD camera optical components.

Spectral Resolution

Spectral resolution will depend on the ability to vary the OPD and the Fourier transform. While the interferometer is used basically as a filter in the multispectral imager to select a wavelength, in the hyperspectral imager it is used to vary the OPD for all the wavelengths whose energy is integrated at the CCD camera pixels.

The electric field, E, for emissions at one wavelength, λ, or wave number, $k = 1/\lambda$, radial frequency, ω, and amplitude, A, can be written as

$$E(k) = A(k)\, \mathrm{e}^{-\mathrm{i}(2\pi kx - \omega t)} \qquad [1]$$

and the intensity which is measured is the magnitude squared of the electric field:

$$I = E^{*}E \qquad [2]$$

Emissions from fluorescent sources are usually noncoherent and composed of a range of wavelengths, even with only one dye present. The intensity measured at a pixel is the superposition of many wave numbers. After traveling a certain OPD, L, the intensity integrated over all wave numbers may be written as

$$I_{\mathrm{pxl}}(L) = \frac{1}{2}\left[\int I(k)\mathrm{d}k + \int I(k)\cos(2\pi kL)\mathrm{d}k\right] \qquad [3]$$

or

$$I_{\mathrm{pxl}}(L) = C_0 + C_1 \cdot \mathrm{Re}\{\mathrm{FT}[I(k)]\} \qquad [4]$$

where FT stands for Fourier transform.

The inverse transform is

$$\int I_{\text{pxl}}(L)\, e^{-i2\pi k' L}\, dL = C_0 \int e^{-i2\pi k' L}\, dL + C_1 \int\int I(k) e^{-i2\pi k L'}\, dk \cdot e^{i2\pi k' L}\, dL \quad [5]$$

$$\int I_{\text{pxl}}(L) e^{-i2\pi k' L}\, dL = C_3 + C_4 I(k)$$

The integration constant, C_3, can be ignored for large wave numbers, as in the case of the visible range, since C_3 will be proportional to the inverse of the wave number. The intensity as a function of wave number is then

$$I(k) = C_5 \int I_{\text{pxl}}(L) e^{-i2\pi k L}\, dL \quad [6]$$

For absolute quantitative measurements, the constant C_5 can be determined by calibration with a known emission. For discrete measurements made by changing the optical path length, then the working formula is

$$I_{\text{pxl}}(k) = \sum_{L_i} I_{\text{pxl}}(L_i) e^{-i2\pi k L_i}\, \Delta L \quad [7]$$

This is the discrete Fourier transform which must be preformed for each pixel to determine the intensity as a function of wavelength. Per the Nyquist sampling tenets:

$$\Delta L = \frac{1}{(N/2)\Delta k} \quad \text{or} \quad \Delta k = \frac{2}{N \Delta L} \quad [8]$$

where Δk is the discrete wave number interval and N the number of different OPDs. Sagnac interferometers can make rotations which give at least a change of $\Delta L = \lambda/4$, so substituting in eqn [8]:

$$\Delta k = k_1 - k_0 = \frac{1}{\lambda_1} - \frac{1}{\lambda_0} = \frac{\Delta\lambda}{\lambda^2 + \lambda\Delta\lambda} = \frac{8}{N\lambda} \quad [9]$$

$$\frac{\Delta\lambda}{\lambda^2} \approx \frac{8}{N\lambda} \quad \text{or} \quad \frac{\Delta\lambda}{\lambda} \approx \frac{8}{N}$$

A 5% resolution in wavelength (or 20 nm at 400 nm) requires $N = 160$ and, of course, a 2.5% resolution requires $N = 360$. Therefore, about 400 OPD changes and acquisitions of the intensity fringes will give calculated resolutions of less than 10 nm in wavelength.

Analysis of the spectrum based on this resolution can be conducted to determine which dyes are present. Software techniques for displaying data will be described in the next section.

Grating-Dispersion Hyperspectral Imaging

A hyperspectral imaging microscope collects the complete visible emission spectrum from a microscope slide by using a grating spectrometer to provide the wavelength dispersion. This microscope correlates spatial and spectral information with minimal use of optical filters. A hyperspectral microscope system is composed of a standard epi-fluorescence microscope (Olympus IX70) equipped with objectives lens (including 1.25 × and 10 × Plan Fluorite, 40 × Plan Apochromat Dry, and 60 ×, and 100 × Plan Apochromat Oil), as well as a set of standard filter cubes (Chroma Technology) to allow for traditional visualization of dyes through the eyepiece. (Prototype typical components are listed in parentheses.) Multiple excitation sources can be utilized including standard 100 W mercury and xenon arc lamps, tungsten brightfield illumination, a 532 nm solid state laser (Brimrose), a helium neon laser (Uniphase), an argon ion laser (Uniphase), and a pulsed-doubled nitrogen dye laser (Laser Science, Inc.) (Figure 1b).

One side port of microscope is optically coupled to a spectrograph (Acton Research SpectraPro® 556i). A narrow entrance slit into the spectrograph allows for only one line (ΔX) of the image to be captured at a time. A 50-micron slit permits 80% of maximum throughput for the spectrometer. The spectrograph separates the polychromatic light illuminated from this single line into its light components. Normal operation uses a grating with 50 grooves/mm and a 600 nm-blaze wavelength, which allows wavelengths of approximately 400–780 nm to be measured. Different spectrograph gratings can be used for special applications from near UV to near IR. The center wavelength and grating can be shifted under computer control. The output of the spectrograph is then captured with an air/liquid-cooled CCD camera (Photometrics-Quantix) and an optional image intensifier (Video Scope International VS4-1845). The camera has a resolution of 1536 × 1024 pixels with a depth of 12 bits per pixel. The 1536 pixels are aligned in the spatial direction to allow the greatest spatial resolution. During normal operation the wavelength resolution is 'binned' to 128 pixels, sufficient to resolve ~8 different spectra in a single excitation scan. Higher resolution can be employed for special applications.

A line of excitation light can be provided by a line generator when laser excitation is used or by a combination of slits and cylindrical lenses when using traditional lamps or burners (mercury or xenon). This line of light illuminates the sample in the same position that maps the resulting emission light

through the imaging spectrograph and onto the camera. Some advantages of illumination with a line of light include decreased sample bleaching, reduced background due to scattering, and efficient use of available excitation power. A traditional bandpass, a low-pass dichroic filter, or a linear variable filter (Raynard Corp.) may be employed to select or restrict the excitation wavelength during specific applications.

Linear Variable Filter (LVF)

The linear variable filter (LVF) (Raynard Corporation) can be used to replace the standard excitation filter cube to provide a capability for continuous excitation wavelength band selection. The LVF is $1'' \times 2.25''$ and is coated such that it continuously filters light from ~400 nm to ~800 nm along its length. Custom filters that span other wavelengths are available. The coating and therefore wavelength is uniform along a short distance, making it an ideal method for selecting an excitation wavelength band to paint across the slide, excite the slide's contents, and map the contents' emission into the slit of the imaging spectrograph. The bandwidth passed is about 21 nm wide. The LVF introduces ~50% attenuation of the intensity.

Spatial Resolution

The spatial resolution of a grating spectrograph Hyperspectral Imager is set by the size of the pixels of the CCD camera in the y- direction and the microscope system magnification. However, differing from the interferometric system, the spatial resolution in the x- direction depends on the spectrometer slit width and the microscope system magnification. As before, let the x- and y- dimensions of the CCD pixel be δx and δy, respectively, and the magnification be M. Let w_s be the slit width of the spectrometer, and note than the slit width is always going to be larger than a few δx. Then the spatial resolution in the y-direction is again $2\delta y/M$, but the spatial resolution in the x- direction is now $2w_s/M$. Again, for CCD pixel sizes of 10 microns square, binning together two pixels in the y-direction, and an overall magnification of 60, the y- resolution is 0.66 microns, as it was for the interferometer method. For a slit width of 50 microns, the x- resolution is 1.67 microns, independent of the number of pixels binned in the x-direction. The control software moves the stage the right number of steps to minimize the overlap or skipped areas. The slit width can be reduced, but at a loss of throughput or increased exposure time. The control software can compensate for different spatial resolution to display the image as 1:1.

Field of View

For N pixel rows (y- dimension) and C pixel columns (x- dimension), the field of view will be $N\delta y/M$ for the length of the area along the y-axis. If the spectrometer slit length is less than the height of the CCD, then the length is just h_s/M, where h_s is the slit length. We orient our CCD array so that the 1536-pixel dimension is along the y-axis. The y-view is thus 256 microns as in the interferometric example. To achieve the same field of view as the Sagnac method imager, we would need to take 103 stage movements and acquisitions. This highlights the difference in the two hyperspectral methods. The grating method collects the complete spectrum on each acquisition with no spectrometer adjustment but needs to step down the slide to produce the full spatial image. The interferometer method collects the complete spatial image on each acquisition with no position adjustment but needs to step the interferometer through a range of optical path differences to acquire multispectral content.

The sensitivity, dynamic range, and spectral response are set by the CCD camera optical components.

Spectral Resolution

Spectral resolution depends on spectrometer instrument function which includes the slit width, the dispersion of the grating, the CCD pixel size, and number of pixels binned together. The dispersion for normal operation was determined to be ~40 nm/mm or 0.4 nm/pixel in the 10 micron CCD pixel example. For example, binning of 8 pixels yields a spectral resolution of 3.2 nm at all wavelengths. To achieve this spectral resolution, the interferometer method would require 1000 OPD increments and acquisitions.

Analysis of the spectrum based on this resolution can be conducted to determine which dyes are present. Although the spatial resolutions differ considerably in the two methods, as long as the dye peaks are separated by the spectral resolution spacing, the spectral decomposition algorithms will find all dyes within the spatial resolution area. We have demonstrated that the analysis algorithms can resolve 5 dyes with peaks separated by more than 6 nm in various combinations and known quantity ratios against the dye standards. The program successfully identifies the dyes present with a 91% accuracy in the quantities.

Software

Software (*Hyper*Scope, a 'C' language program) controls all the hardware components for acquisition.

It includes features for analyzing and displaying the resultant $X-Y$ images and spectral information for the grating spectrometer Hyperspectral Imager. During a scan, individual pictures are taken in the $Y-\lambda$ plane, while the stage is incremented in the X-direction. A collection of such $Y-\lambda$ scan files, which can total from 50–1000, is merged to build an image cube (Figure 2). Another advantage of the single-line acquisition mode is that time-dependent features can be monitored continuously. The software was designed to permit repeated collection of a single line at variable time points. The utility of this feature has been demonstrated by capturing images of fluorescent objects moving through UV-transparent capillaries.

The software was designed to display the resulting $X-Y$ plane extracted in various ways from the image cube. A graph of the emission spectrum for any pixel within that image may be viewed by clicking on any pixel of interest in the $X-Y$ image (Figure 3c). A standard linear curve-fitting algorithm is used to determine the contribution of individual dyes to the measured emission spectrum. The software also features a windowing technique (integration of the emission spectra over a fixed window) which emulates standard filtering. An overlay feature within the software allows the curve fitting results to be displayed in false color and compared by viewing the composite for multiple fluorophores in a single image. Additionally, it permits the contribution of a fluorophore to be emphasized for visualization purposes by either displaying the contribution of a single fluorescent signature in the absence of the others that were present in the image or by scaling individual display intensity values up or down. Overlapping pseudo-colors assigned to represent each fluorophore (either by curve fitting or by windowing) may be displayed as added, averaged, maximum or minimum values. For each new fluorophore–excitation source combination, the characteristic emission spectra is acquired and added to a library of available spectra that can be used as components in the spectral decomposition. This permits the user to introduce new fluorophores at any time. Moreover, background spectra (defined as a region of interest by the user) can be acquired and used as one element of the decomposition. This background component can be turned off in the false color view (as any of the spectral components may be) resulting in enhanced noise reduction and visual contrast improvement.

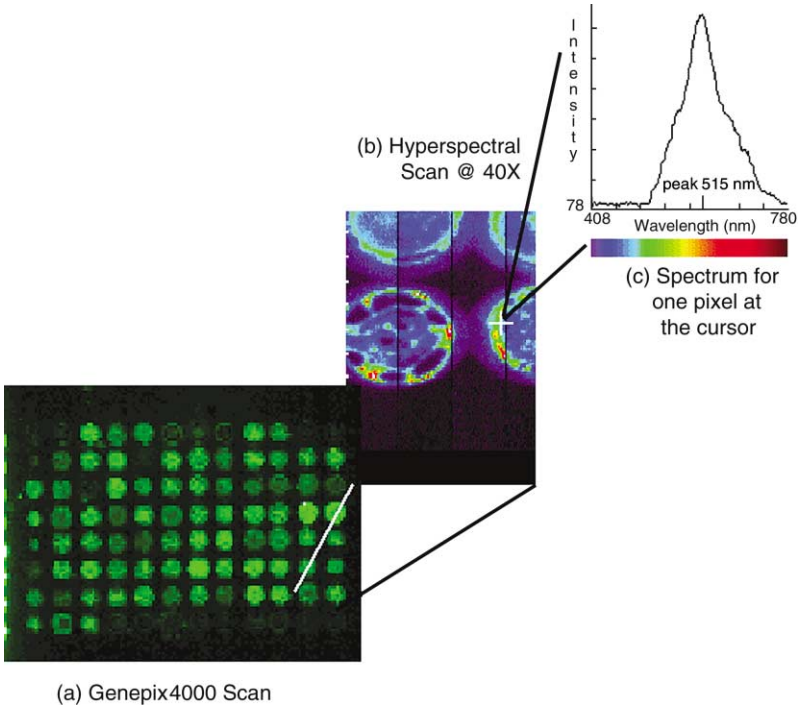


Figure 3 The benefit of spectral imaging is demonstrated in a spatial high-resolution scan of an expression microarray. (a) A portion of the scan of an expression microarray on a microscope slide using the Genepix4000 scanner. It records the intensity emitted by the Cy3 dye when excited with a 532 nm laser. (b) A portion of the scan of the same microarray with the Hyperspectral Imaging Microscope using a 40X objective and Hg excitation lamp. False color is displayed to indicate the intensity variations. (c) A *Hyper*Scope computer program display of the spectrum recorded when the cursor is placed over one pixel in a scan.

Description of Acquisition

In order to acquire a full image, a slide is placed on the microstage mover (Ludl Electronics) that allows for precise movement of the sample across the image acquisition plane. The user can position the stage via a joystick or via software to move to predetermined coordinates. The image is then scanned, with the number of scans being selected at the time of the image acquisition. Furthermore, the sample can be scanned at one magnification or resolution and then rescanned at another. Each $Y-\lambda$ image is saved as a single $Y-\lambda$ file for each increment along the sample. Thus, each image file corresponds to the excitation line of light and the emission collection region that maps through the imaging spectrograph to the CCD camera. Saving individual files permits unlimited $Y-\lambda$ image acquisition at the same Y coordinate, valuable for time-dependent studies. This feature also provides some robustness by enabling a partial $X-Y$ image to be displayed should acquisition of one or more Y scans fail for any reason. After acquiring each image, the stage is moved automatically, one or more steps in the X direction, to acquire emissions at the next line and generate the next $Y-\lambda$ file. Therefore, only the fluorescent probes in a line at a single X-location are examined at each acquisition time. Step sizes can be adjusted by the software to give: (i) the best resolution; (ii) the best 1:1 image; (iii) the minimum overlap; or (iv) the fast acquisition time.

After a variable number of rows of associated emission have been captured (typically 50–1000 scans), an image cube is loaded into memory comprising all the $Y-\lambda$ data files. A graph of the spectrum from any pixel (or bin) in the $X-Y$ plane may be displayed in a separate window by clicking on the cube (Figure 3c). A typical acquisition scan of 250 $Y-\lambda$ files, which are used to construct a $250 \times 768 \times 128$ ($\times 12$ bit) image cube, consists of approximately 100 Mb of information.

MicroArray Scanning and Customized Excitation

Another use of this technology is to analyze spotted microarrays. The Hyperspectral Imager has been compared to commercial two-color scanners with comparable sensitivity at approximately 0.5 fluors per square micron (Figure 3). Increased sensitivity (500–10 000 gain) can be provided by using a microchannel plate amplifier (Model VS4-1845, Videoscope International, Ltd.) installed between the CCD camera and the imaging spectrograph.

Since the hyperspectral imaging microscope collects images that reflect only the $Y-\lambda$ plane, it contains the potential to use a single line of light as a source for image illumination. Therefore, lights systems being developed, which are capable of generating a line of light composed of variable or complex wavelengths, could be employed to precisely control excitation. The incorporation of such an excitation source in hyperspectral imaging microscopy could be useful in many clinical and basic research applications.

See also

Environmental Measurements: Hyperspectral Remote Sensing of Land and the Atmosphere. **Imaging:** Interferometric Imaging.

Further Reading

Anguiano A (2000) Fluorescence in situ hybridization (FISH): Overview and medical applications. *Journal of Clinical Ligand Assay* 23: 33–42.

Berland K, Jacobson K and French T (1998) Electronic cameras for low-light microscopy. *Methods in Cell Biology* 56: 19–44.

Bernheim A, Vagner-Capodano AM and Couturier J (2000) From cytogenetics to cytogenomics of tumors. *S-Medical Science* 16: 528–539.

Bezrookove V, Hansson K, van der Burg M, *et al.* (2000) Individuals with abnormal phenotype and normal G-banding karyotype: improvement and limitations in the diagnosis by the use of 24-colour FISH. *Human Genetics* 106: 392–398.

Boland MV and Murphy RF (1999) Automated analysis of patterns in fluorescence-microscope images. *Trends in Cell Biology* 9: 201–202.

Cardullo RA and Alm EJ (1998) Introduction to image processing. *Methods in Cell Biology* 56: 91–115.

Garini Y, Katzir N, Cabib D, Buckwald RA, Soenksen D and Malik Z (1996) Spectral bio-imaging. In: Wang XF and Herman B (eds) *Fluorescence Imaging Spectroscopy and Microscopy*, pp. 87–122. New York: John Wiley and Sons.

Garini Y, Macville MV, du Manoir S, *et al.* (1996) Spectral Karyotyping. *Bioimaging* 4: 65–72.

Gothot A, Grosdent J and Palulus J (1996) A strategy for multiple immunophenotyping by image cytometry: Model studies using latex microbeads labeled with seven Streptaviden-bound fluorochromes. *Cytometry* 24: 214–225.

Gribble SM, Roberts I, Grace C, Andrews KM, Green AR and Nacheva EP (2000) Cytogenetics of the chronic myeloid leukemia-derived cell line K562: Karyotype clarification by multicolor fluorescence *in situ* hybridization, comparative genomic hybridization, and locus-specific fluorescence in situ hybridization. *Cancer Genet Cytogenetics* 118: 1–8.

Han H, Gao X, Su J and Nie S (2001) Quantum-dot-tagged microbeads for multiplexed optical coding of bimolecules. *Nat Biotech.* 19: 631–635.

Huebschman ML, Schultz RA and Garner HR (2002) Characteristics and capabilities of the hyperspectral imaging microscope. *Eng in Med and Bio*. 21: 104–117.

Inoue T and Gliksman N (1998) Techniques for optimizing microscopy and analysis through digital image processing. *Methods in Cell Biology* 56: 63–90.

Jain AK (1997) *Fundamental of Digital Image Processing*, pp. 80–128. New Jersey: Prentice-Hall.

Jalal SM and Law ME (1999) Utility of multicolor fluorescent in situ hybridization in clinical cytogenetics. *Genet Med*. 1: 181–186.

Johannes C, Chudoba I and Obe G (1999) Analysis of X-ray-induced aberrations in human chromosome 5 using high-resolution multicolour banding FISH (mBAND). *Chromosome Research* 7: 625–633.

Kruse F (1993) The Spectral Image Processing System (SIPS) – Interactive visualization and analysis of imaging spectrometer data. *Remote Sensing Environment* 44: 145–163.

Lu YJ, Morris JS, Edwards PAW and Shipley J (2000) Evaluation of 24-color multifluor-fluorescence *in-situ* hybridization (M-FISH) karyotyping by comparison with reverse chromosome painting of the human breast cancer cell line T-47. *Chromosome Research* 8: 127–132.

Malik Z, Cabib D, Buckwald RA, Talmi A, Garini Y and Lipson SG (1996) Fourier transform multipixel spectroscopy for quantitative cytology. *Journal of Microscopy* 182: 133–140.

Press WH, Flannery BP, Teukolsky SA and Vetterling WT (1986) *Numerical Recipes, the Art of Scientific Computing*, pp. 546. New York: Cambridge University Press.

Raap AK (1998) Advances in fluorescence in situ hybridization. *Mutat. Res. Fundam. Mol. Mech. Mutagen* 400: 287–298.

Sakamoto M, Pinkel D, Mascio L, *et al.* (1995) Semiautomated DNA probe mapping using digital imaging microscope: II. system performance. *Cytrometry* 19: 60–69.

Schrock E, duManoir S, Veldman T, *et al.* (1996) Multicolor spectral karyotyping of human chromosomes. *Science* 273: 494–497.

Schultz RA, Nielsen T, Zavaleta JR, Ruch R, Wyatt R and Garner HR (2001) Hyperspecteral imaging: a novel approach for microscopic analysis. *Cytometry* 43: 239–247.

Schwartz S (1999) Molecular cytogenetics: show me the colors. *Genet Med*. 1: 178–180.

Speicher MR, Ballard SG and Ward DC (1996) Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics* 12: 368–375.

Wang YL (1998) Digital deconvolution of fluorescence images for biologists. *Methods in Cell Biology* 56: 305–315.

Zhao L, Hayes K and Glassman A (2000) Enhanced detection of chromosomal abnormalities with the use of RxFISH multicolor banding technique. *Cancer Genet Cytogenetics* 118: 108–111.

Imaging Through Scattering Media

A C Boccara, Ecole Supérieure de Physique et Chimie Industrielles and Centre National de la Recherche Scientifique, Paris, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

We are often faced with the problem of seeing through scattering media such as smoke, fog, clothes, blood, tissues, ceramics, walls, etc. Tremendous effort has been made to introduce techniques which will allow us to overcome the intrinsic difficulty of selecting 'good' photons which are able to provide valuable information on structures which appear fuzzy or, for most of the time, completely hidden.

Although a number of concepts and techniques that we will discuss in this article could be applied to a large number of scattering samples, we will restrict ourselves to biological tissues.

Studying light propagation through biological tissues is becoming more and more popular among physicists, biologists, engineers, and doctors. Indeed this research domain appears to be growing rapidly as seen by the number of published articles, conferences, and international meetings. Optical techniques are attractive and are starting to compete with other well-established techniques (magnetic resonance imaging, X-ray imaging, ultrasonic imaging, positron emission tomography, etc.), because they are non-ionizing, nondestructive, they can reach high resolution and are usually much cheaper. Optical techniques reveal an optical contrast which is valuable in providing morphologic as well as functional information.

Most biological tissues are weakly absorbing in the deep red and near infrared region of the spectrum; the main difficulty in performing imaging arises from the high level of scattering. Figure 1 shows the picture of a small flash light bulb placed in the mouth: it appears fuzzy and red.

In Figure 2 we consider various photon trajectories between a short pulsed source S and a fast detector D.

Unscattered photons are called 'ballistic' photons, their number decreases exponentially, and they are completely damped after one or two millimeters.

Then there are scattered photons which propagate close to the ballistic trajectory reaching the detector; these are named 'snake-like' photons and can be used for imaging purposes.

Most of the energy which reaches D is related to the scattered photons which may exhibit very long trajectories (typically ~10 times the distance S–D).

The description of various experimental methods used to overcome the difficulty introduced by scattering will be associated with the various photon families that we have mentioned: ballistic, snake-like, and scattered photons. We will first give a few orders of magnitude of the parameters which will dictate the way light interacts with the tissues. Let us emphasize that these parameters also carry the information related to the contrast mechanism. The ultimate goal is to realize an 'optical biopsy', noninvasive but as close as possible to the regular biopsy followed by histopathology.



Figure 1 Biological tissues are highly scattering but the absorption is low in the red (as seen in this digital visible photography) and the near infrared.

Optical Parameters

Scattering Cross-Section

As shown in Figure 3, biological tissues are constituted of structures which exhibit various characteristic scales: from a few nanometers in membranes to about ten micrometers for an entire cell.

As we know from scattering theories these various scales correspond to various scattering properties. For small particles ('small' means much smaller than the optical wavelength λ) we are in the Rayleigh regime which is based on the evaluation of a time-varying dipole induced by a uniform time-varying (at the light frequency scale) electric field. This dipole radiates in the far field and its scattering cross-section, which is the ratio of the scattered power to the incoming irradiance for a sphere of radius r, is given by

$$\sigma = \frac{\langle P \rangle}{I} = \left(\frac{n^2 - 1}{n^2 + 2} \right)^2 \frac{8\pi}{3} r^6 k^4 \qquad [1]$$

n being the refractive index of the sphere divided by the index of the surrounding medium and $k = 2\pi/\lambda$. The scattering diagram for such particles is isotropic if we use unpolarized light. This cross-section is much smaller than the 'geometrical' section S of the particle; for instance, if $n = 1.5$, $a = 10$ nm, $\lambda = 0.6$ μm, $\sigma = 1510^{-8}$ μm^2, which is much smaller than the actual section (3×10^{-4} μm^2).

For larger particles the sphere (or the ellipse) interacting with a plane polarized wave is rather complicated because we are not restricted to the dipolar approximation (uniform field), but induced multipolar moments and their radiating fields have to be taken into account. This theory was proposed by Mie and can be found in optics textbooks. As an example, Figure 4 shows the scattering cross-section of a micron-size particle as a function of the

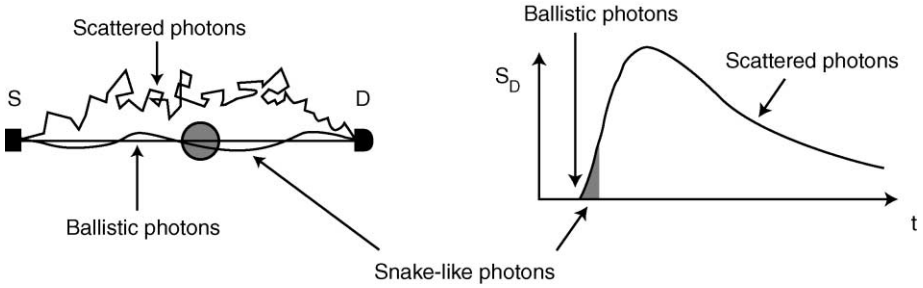


Figure 2 Between an ultrafast light source S and a detector D can be distinguished various classes of photons which can be used for imaging.

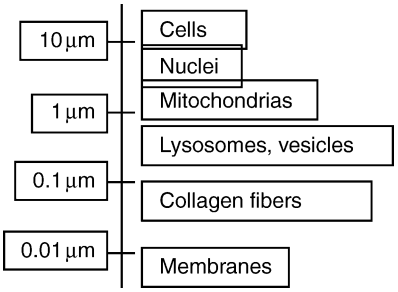


Figure 3 Various sizes of scatterers may be found in cell structures.

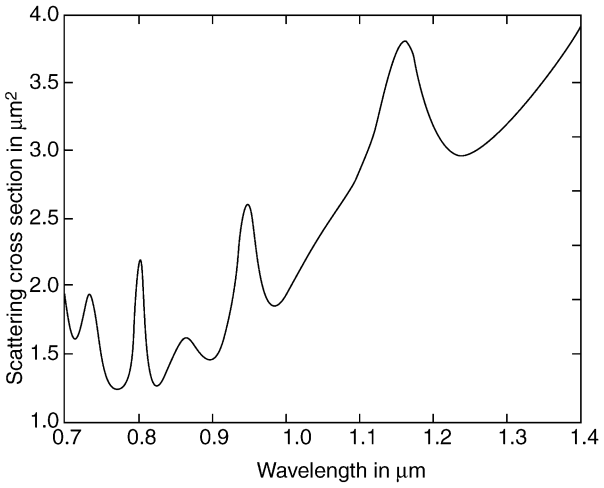


Figure 4 In the Mie scattering regime resonances are observed when varying the wavelength (here a 1 μm sphere, refractive index: 2).

wavelength: a number of 'resonances' can be observed.

The scattering diagram is no longer isotopic, it exhibits lobes, and forward scattering is much higher than back-scattering. Scattering is characterized by a scattering coefficient, μ_s, which is the inverse mean free path l_s of photons between scattering events:

$$l_s = 1/\mu_s = \sigma\rho \qquad [2]$$

where σ is the scattering cross-section and ρ the number of scatterers by unit volume.

The scattering coefficients μ_s of tissues are found within the range 100–1000 cm^{-1}. After a path length L, the number of ballistic photons will be damped by a factor $\exp(-L/l_s)$, so tissue imaging cannot rely only on ballistic photons as they survive only to a path length larger than about one mm. One needs to consider in more detail the scattering properties of the scatterers.

The dominant sizes of the scatterers in tissue (cells) demand consideration of the scattering in the Mie regime. Such a scattering event does not result in isotropic scattering angles. Instead, the scattering in tissue is biased in the forward direction.

Anisotropic scattering is quantified in a coefficient, g, which is defined as the mean cosine of the scattering angle θ, $p(\theta)$ being the probability of a particular scattering angle:

$$g \equiv \langle\cos(\theta)\rangle \equiv \frac{\int_0^\pi p(\theta)\cos(\theta)\sin(\theta)\mathrm{d}\theta}{\int_0^\pi p(\theta)\sin(\theta)\mathrm{d}\theta} \qquad [3]$$

For isotropic scattering, $g = 0$. For complete forward scattering, $g = 1$. In tissues, g varies from 0.7 to 0.95.

Photon migration, especially in the so-called diffusion regime, is often based on isotropic scattering such as for phonons (heat) or charge carriers scattering. For diffusion-like models, it has been shown that one may use an isotropic scattering model with a corrected scattering coefficient, μ_s', and obtain equivalent results using:

$$\mu_s' = \mu_s(1 - g) \qquad [4]$$

One can consider $l_s' = 1/\mu_s'$ as the length over which the incoming photon loses the memory of its initial direction, μ_s' is usually called the transport-corrected scattering coefficient.

Absorption Cross-Section

The absorption coefficient, μ_a (cm^{-1}) represents the inverse mean path length of photons before absorption. $1/\mu_a$ is also the distance in a medium at which intensity is attenuated by a factor of $1/e$ (Beer's Lambert law). Absorption in tissue is strongly wavelength dependent and is due to chromophores and water. Among the chromophores in tissue, the dominant component is the hemoglobin in blood. In the visible range, the blood absorption is relatively high, whereas it is much weaker in the near infrared. Water absorption is low in the visible and NIR region and increases above 1 μm, with tissues turning completely opaque above 2 μm. Thus, for greatest penetration of light in tissue, wavelengths in the 700–1300 nm spectrum are used. This region of the spectrum is called 'the therapeutic window'. Different spectra of chromophores allow one to separate the contribution of varying functional species in tissue such as quantification of oxy- and deoxy-hemoglobin to study tissue oxygenation.

Finally, when light is propagating in scattering media of thickness L, the damping of the incoming energy involves a combination of its absorption and

scattering properties, and is given by

$$\exp(-L/a) \qquad \text{with} \qquad a^2 = 1/3\mu_a(\mu_s' + \mu_a) \qquad [5]$$

This formula reflects the fact that the photon's pathlength is increased by scattering, leading to an increase of the damping, which is related to absorption.

These few examples provide a 'static' view of light behavior in scattering media. Studying in a more detailed way, the spatial and temporal distribution of light, is more difficult.

Photon Migration in Scattering Media

Photon migration models have been inspired by other fields, such as astrophysics, soft matter physics, and neutronics, where physical media are random in both space and time. To avoid difficulties linked to the microscopic complexity such investigations rely on a statistical basis.

The most widely used theory is the time-dependent diffusion approximation to the transport equation:

$$\vec{\nabla} \cdot (D\vec{\nabla}\Phi(\vec{r},t)) - \mu_a\Phi(\vec{r},t) = \frac{1}{c}\frac{\partial\Phi(\vec{r},t)}{\partial t} - S(\vec{r},t) \qquad [6]$$

where $\vec{r}$ and t are spatial and temporal coordinates, c is the speed of light in tissue, and D is the diffusion coefficient related to the absorption and scattering coefficients by the formula:

$$D = \frac{1}{3[\mu_a + \mu_s']} \qquad [7]$$

The quantity $\Phi(\vec{r},t)$ is called the fluence, defined as the power divided by the unit area. This equation does not reflect the scattering angular dependence; the use of μ_s' instead of μ_s ensures that isotropic scattering has been reached. Indeed, for the use of the diffusion theory which implies anisotropic scattering, the diffusion coefficient is expressed in terms of the transport-corrected scattering coefficient; $S(\vec{r},t)$ is the source term.

The diffusion equation has been solved analytically for different types of measurements such as reflection and transmission modes, assuming that the optical properties remain invariant throughout the tissue. To incorporate the finite boundaries, the method of images has been used.

The diffusion approximation equation in the frequency-domain is the Fourier transform of the time-domain equation. This leads to the equation:

$$\vec{\nabla} \cdot (D\vec{\nabla}\Phi(\vec{r},\omega)) - \left[\mu_a + \frac{i\omega}{c}\right]\Phi(\vec{r},\omega) + S(\vec{r},\omega) = 0 \qquad [8]$$

Here the time variable is replaced by the frequency ω which is the modulation angular frequency of the source. In this model, the fluence can be seen as a complex number describing the amplitude and phase of the so-called 'photon density wave' in addition to a DC component:

$$\begin{aligned}\Phi(\vec{r},\omega) &= \Phi_{AC}(\vec{r},\omega) + \Phi_{DC}(\vec{r},0) \\ &= I_{AC}\exp(i\theta) + \Phi_{DC}(\vec{r},0)\end{aligned} \qquad [9]$$

where θ is the phase shift of the diffusing wave whose wavelength is:

$$2\pi\sqrt{\frac{2c}{3\mu_s'\omega}} \qquad [10]$$

Although analytical techniques provide a rapid understanding of most phenomena involved in static or dynamic situations, real biological samples are much more complex than homogeneous isotropic media. In these cases, direct and inverse algorithms map the spatially varying optical properties.

Monte Carlo simulations are very popular nowadays; a number of useful programs are available on the websites of various groups. Results may include time or polarization dependence of the photons along their paths.

The other method used in tissue optics is the random walk theory (RWT) on a lattice. RWT models the diffusion-like motion of photons in turbid media in a probabilistic manner. Using RWT, an expression may be derived for the probability of a photon arriving at any point and time given a specific starting point and time.

Imaging in the Ballistic Regime: Taking Advantage of Various Coherence Properties

Shallow Tissue Imaging Through Selection of Ballistic Photons

When imaging thin (less than 2 mm) tissue samples, it is possible to use photons which have not been scattered, called the ballistic photons which can be used to form high-resolution images in the same manner as if no scattering had taken place.

Their number, however, decreases exponentially with propagation distance and this ballistic signal is usually hidden by multiply scattered photons that obscure the image. For relatively shallow tissue depths it is possible to use various spatial filtering techniques to block the multiply scattered photons and there is much current research that aims to

extend optical imaging and biopsy to depths of a few mm.

Confocal microscopy rejects much of the scattered light through spatial filtering and has been shown to form 3-D images at tissue depths of up to 200–300 μm, both *in vitro* and *in vivo*, which corresponds to about 50 dB of damping for the ballistic photons.

Two-photon fluorescence microscopy is particularly interesting in this regime since it ensures that all the detected fluorescence photons originate from the desired image plane. This technique also ensures a better penetration (using IR excitation) and a better resolution (nonlinear response) than single photon confocal microscopes.

Imaging to much greater tissue depths does not appear to be practical using optical microscopy and spatial filtering alone and more sophisticated techniques must also be employed to discriminate in favor of the ballistic photons. These techniques exploit either the fact that the ballistic signal photons retain spatial coherence of incident light or the fact that the ballistic photons keep the optical phase memory that multiply diffused photons have lost.

One of the most successful ballistic light-imaging techniques based on the latter filtering approach is optical coherence tomography (OCT) which uses coherence gating using low temporal coherence light and an interferometric detection. Coherent detection is used to discriminate against the multiply scattered photons that scatter back into the trajectories of ballistic photons. The ballistic light signal is measured by interference with a reference beam and the resulting pattern is detected with high sensitivity using homodyne or heterodyne detection. By using low coherence length sources and matching the time-of-flight of the reference beam signal to that of the desired ballistic light, OCT can acquire depth-resolved images in scattering media such as biological tissues. For powers in the mW range for tissue irradiation in the infrared, the unscattered ballistic (or more precisely single back scattered) component of the light will be limited by the shot noise detection limit after propagating through approximately 25 scattering mean free paths (MFP) of a scattering medium (or more than 100 dB). This corresponds to a 'typical' tissue thickness of about 1 mm tissue depth for the reflection geometry which is potentially useful for clinical applications such as screening for skin cancer, or for retinal examination.

OCT has proved clinically useful as a means of acquiring depth-resolved images in the eye and is being developed for application in strongly scattering biological tissue. Although the image acquisition requires pixel-by-pixel scanning, the use of new high-power superluminescent diodes in the 100 mW range or of ultrafast lasers to provide high average power, low coherence length radiation (sub-two cycle lasers) has resulted in an *in vivo* OCT system providing real-time imaging.

There are a number of other approaches to coherent imaging, including whole-field 2-D acquisition techniques which remove the need to scan transversely pixel by pixel.

Whole-field imaging using a temporally and spatially incoherent source and a well-balanced interferometric microscope intrinsically provides higher image acquisition rates and can take advantage of inexpensive spatially incoherent, broadband sources such as high-power LEDs, and white light thermal sources.

As an example, Figure 5 shows Linnik-type white light interferometer with optical path modulation and Figure 6 an image of an *ex-vivo* rat retina where one can recognize the main feature expected from histology.

Holography, which works in the Fourier space rather than in the object or image space, is also an interferometric technique able to discriminate between ballistic and scattered photons. Since the development of electronic holography, which takes advantage of available arrays of detectors with multimillion pixels, this approach is rapidly growing.

Real-time 3-D imaging systems based on photorefractive holography in multi-quantum-wells devices, which can potentially acquire depth-resolved images at thousands of frames per second, have been demonstrated.

These ballistic light techniques can extend the depth of tissue imaging to the mm range when

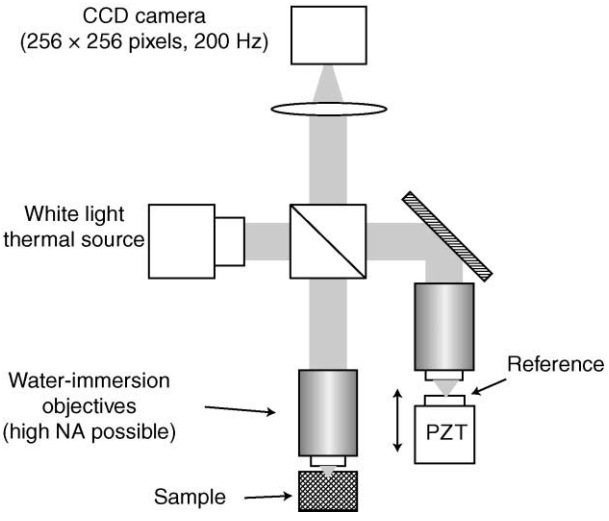


Figure 5 The full-field OCT setup based on a Linnik interferometer. The PZT actuator is used to modulate the path difference between the two arms.

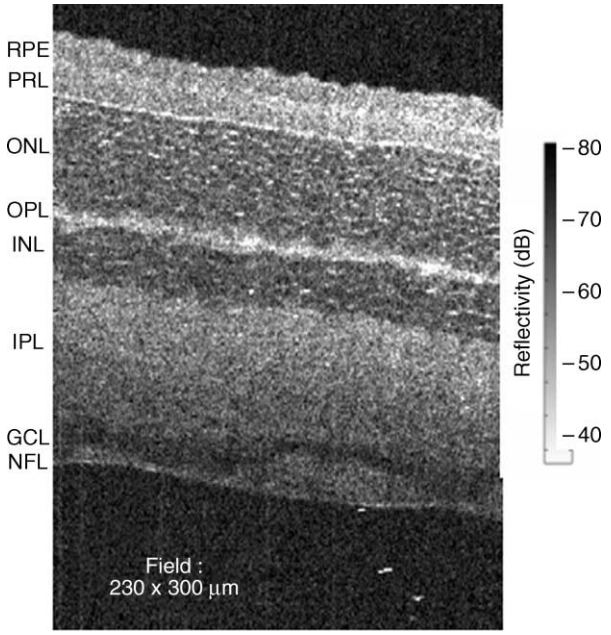


Figure 6 OCT rat retinal image: retinal pigmented epithelium. PRL: photoreceptor layer. ONL: outer nuclear layer. OPL: outer plexiform layer. INL: inner nuclear layer. IPL: inner plexiform layer. GCL: ganglion cell layer. NFL: nerve fiber layer.

imaging in reflection. When it is necessary to image through cms, rather than mms of tissue, however, there will be no detectable ballistic light signal and one must extract useful information from the scattered photons.

Imaging in the Snake-Like Regime: Taking Advantage of Time (or Frequency) Gating

The first approach is to use time gating to select the earliest arriving scattered light, which will provide the most useful image information.

For moderate tissue depths one can exploit the fact that biological tissue is highly forward scattering, with most photons being only slightly deviated from their original direction upon each scattering event. After a few cm of tissue, there can be a significant number of photons that have followed reasonably well-defined 'snake-like' paths about the original direction through the tissue, arriving at the detector after the ballistic light but before the fully diffuse photons whose trajectories are effectively randomized. This early arriving light can be selectively gated using ultrafast lasers and a variety of detectors such as streak cameras, time-gated optical image intensifiers, and time-gated fast photon counting systems, which provide picoseconds time gates.

An alternative, sometimes cheaper, approach to temporal gating is the use of (high) frequency signal discrimination (which is the Fourier transform of the pulse experiment). The periodic solution of the diffusion equation with a periodic source term gives rise to the so-called 'photon density waves'. Such diffusive waves, like thermal waves, are used in the near-field regime (their amplitudes are damped by $\exp(-2\pi)$ after propagating over a single wavelength path) and they can reveal subsurface structures through their refractive and diffusive behavior. Coherence gating and polarization gating can be used when the early arriving light still retains some degree of coherence of the incident light. Indeed polarization discrimination has been used to select those photons which undergo few scattering events and therefore preserve a fraction of their initial polarization state, from the photons which experience multiple scattering resulting in a complete loss of memory of their initial polarization state.

As an example, Figure 7 shows the image of sub-mm light sources (LEDs) taken through 1 cm of polystyrene solution ($\mu_s' \sim \mu_s \sim 10\ \mathrm{cm}^{-1}$) by selecting photons which kept the memory of their initial circular polarization (here a few less than 1%). One can see the sources quite clearly.

For many important biomedical applications, such as mammography or imaging brain function, it is necessary to penetrate through 5–10 cm of tissue, after which all the detected signal is diffuse. Detecting earlier arriving photons can provide more information but, if the time window is too narrow, the detected signal becomes too weak to use with acceptable data acquisition times.

This problem is, however, being addressed with some success using statistical models of photon transport, with different degrees of approximation ranging from full Monte Carlo simulation of photon propagation to the diffusion equation. The approach is to address the inverse problem, i.e. to measure the scattered light signal as comprehensively as is practical and to calculate what distribution of absorption and scattering properties would have produced the measured signal. This provides a means to quantify the optical properties of biological tissue, averaged over a volume corresponding to a particular distribution of photon paths, and to form relatively low-resolution tomographic images.

Multiplying the number of sources and detectors and correlating the various intensity distributions over the whole structure under examination gives, as expected, better precision in localizing the size and position of the local structure to be identified.

Despite the multiple input and multiple output approaches, real-life situations with the previously

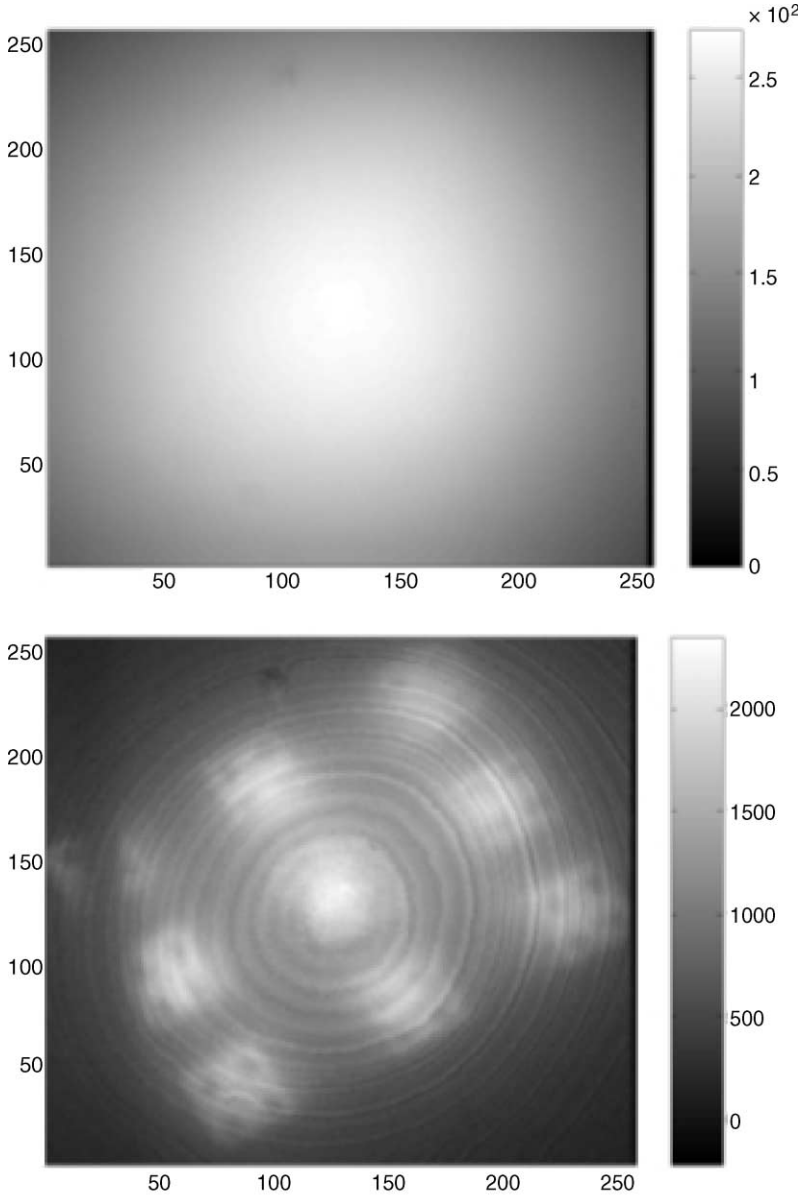


Figure 7 Polarization imaging: here the degree of circular polarization is measured in the upper image. The lower image shows the field without polarization filtering.

mentioned problem of the inhomogeneity of the optical properties of human tissues means that the results are still poor in term of resolution and sub-centimeter structures can hardly be observed in breast imaging, for instance.

One important exception to this limited success in finding precisely the position, the size, and the optical properties of hidden structures is the case of 'activation' studies. This corresponds to a 'differential' situation in which an unknown structure is locally changing with time. Here the goal is not to reach a full description of the unknown structure but only to characterize the local changes induced by the activation.

The main field of application of this approach is certainly in brain activation studies. A matrix of light emitters (usually at two wavelengths which differentiate oxy- and deoxy-hemoglobin, for instance 780 and 840 nm) is coupled to a matrix of light detectors. The signals are usually balanced and any local change which breaks the symmetry of the scattered photon paths induces a differential signal, easy to detect, because there is no background signal before activation takes place.

Moreover it has been demonstrated that for weak perturbations (weak variations of the optical parameters), the inverse problem can be rigorously solved. This geometrical selection provides a cheap and simple setup with rather spectacular results. Sometimes a modulation is added and the phase shift of the photon density wave is enhanced by the differential approach.

Imaging in the Multiple Scattering Regime: Taking Advantage of all the Photons

Light Only: Transillumination and Fluorescence Imaging

The optical systems used for CW measurements are rather simple and require only the alignment of a number of light sources and detectors. Compared to time-gated techniques due to multiple scattering a much stronger path length distribution is present in the signal and the result is a loss of localization and resolution. In the so-called transillumination geometry, collimated detection is used to isolate multiply scattered photons close to the normal emergence angle.

Direct imaging with the required resolution using CW techniques in very thick tissues, such as breast tissue, has not been established even with sophisticated multi-sources/multi-detectors combinations (typically a few tens of sources and detectors at three different wavelengths).

Despite rapid progress in the modeling of light behavior in scattering media it has not yet been demonstrated that CW imaging can provide images with suitable resolution in thick tissue and time-dependent measurement techniques discussed earlier are dominantly used.

For fluorescence imaging, one has to account for the strong attenuation of light as it passes through tissue. Fluorescent dyes (porphyrin, indo-cyanine, and more recently quantum dots) have been developed that are excited and re-emit in the 'biological window' at near-infrared (NIR) wavelengths, where scattering and absorption coefficients are lower, which is favorable for deep fluorescence imaging. Fluorescent intensity in deep tissues depends on the location, size, concentration, and intrinsic characteristics (e.g., lifetime, quantum efficiency) of the fluorophores, and on the scattering and absorption coefficients of the tissue at both the excitation and emission wavelengths. In order to extract intrinsic characteristics of fluorophores within tissue, it is necessary to describe the statistics of photon path lengths which depend on all these different parameters. The amount of light emitted at the site of the fluorophore is directly proportional to the total amount of light impinging on the fluorophore.

Using fluorescence contrast has the unique feature of carrying two spectroscopic contrasts related to both excitation and fluorescent photons.

The next two techniques take advantage of another way to handle the optical information by coupling optics and acoustics.

Coupling Light and Sound: Acousto-Optic and Photo-Acoustic

Although the two approaches that we will now describe are different in their basic principles as well as in the experimental setups, their names are similar – opto-acoustics (sometimes called photo-acoustics) and acousto-optics.

Opto- (or Photo-)Acoustics

In this technique a pulsed (or modulated) electromagnetic beam irradiates the structure under examination. Despite their tortuous paths (about 10 to 100 cm) photons propagate through the volume in a few nanoseconds. Local absorbing centers then absorb the electromagnetic energy, they experience a fast thermal expansion, and become sources of acoustic waves.

The optical problem of tomographic reconstruction is then transformed into a simpler problem of acoustic reconstruction of source distribution because the speed of the sound is much slower than the speed of the light. More precisely we can compute the signal generated by a short light pulse impinging on a scattering medium which contains an absorber (thickness e) as can be seen on Figure 8, when $e\mu_a \ll 1$.

The amount of power per unit surface which is deposited in the absorbing slice is $E\mu_a$ (E is the fluence in W/m^2). When the light pulse duration τ is short enough, despite the propagation in the scattering medium ($\tau \ll e/v$ where v is the speed of the sound),

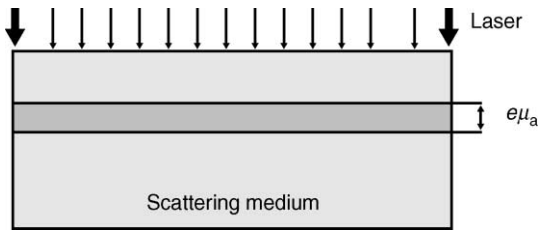


Figure 8 Photo acoustic signal generation: 1-D model.

we find a variation of the mass by volume unit:

$$\Delta\rho \sim \rho\beta\Delta T \sim \rho\frac{\beta E\mu_a e}{C_v e} \qquad [11]$$

The corresponding pressure is:

$$\Delta\rho = \frac{v^2}{\gamma}\Delta\rho \qquad \text{or} \qquad \Delta\rho \sim \rho\frac{v^2\beta E}{C_\rho}\mu_a \qquad [12]$$

Although this approach is new, compared to the purely optical transillumination approach, it has led to a large number of studies in which time gating brings new information to the accurate determination of the US emitter depth. It has been applied to breast tumor examination to complement standard imaging techniques such as echography and X-ray mammography. Recently this technique has demonstrated submillimeter resolution for *in vivo* blood distribution in rat brain imaging.

Acousto-Optic

Here a DC single frequency laser is used to illuminate the sample. For a typical biological sample a few cm thick the trajectories are distributed between the sample thickness value and more than ten times this value. Due to the good temporal coherence of the source all the scattered photons are likely to interfere when they emerge from the sample volume. These interferences are seen as a speckle field which can be observed at a suitable distance from the sample surface.

As seen in Figure 9, adding an ultrasonic field will mainly induce a small (smaller than the optical wavelength) displacement of the scatterers (it also induces local variation of the refractive index) which will cause a speckle modulation at the ultrasonic frequency.

If the zone irradiated by the ultrasonic field is optically absorbing, less photons will emerge from this zone than from a nonabsorbing one and thus less modulation will be seen when scanning across the sample volume. This modulation can be detected by selecting a single speckle grain and averaging over a number of uncorrelated speckle distributions to obtain a suitable average value (for instance with latex particles). This is not always possible, in particular because the overall geometry is rather stable in a semisolid tissue. Using a single detector which receives many speckle grains gives less contrast. A multiple detector and a parallel processing of many speckle grains improve greatly the speed and the signal-to-noise ratio: a typical image of absorbing spheres, embedded in a scattering phantom, is shown in Figure 10.

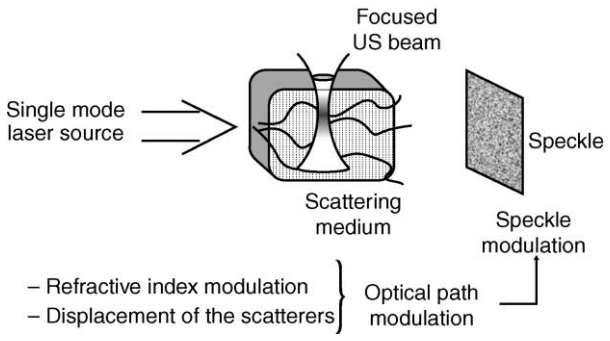


Figure 9 Acousto-optic imaging principle: photons are tagged by the acoustic field.

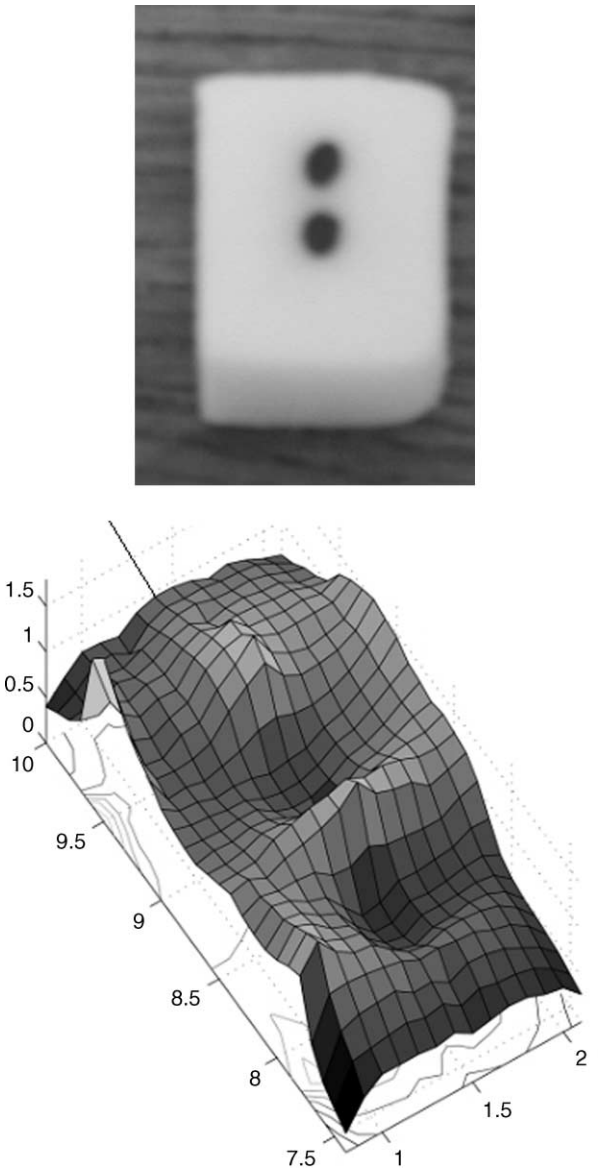


Figure 10 Acousto-optic imaging (lower image) of two absorbing spheres embedded in a scattering medium (upper image shows a section of the gel sample).

Conclusion

The complexity associated with multiple scattering in turbid media, such as tissues of the human body, prevents an easy use of light as a routine tool for in-depth examination, whereas superficial examinations have always been sources of diagnostics.

Optical techniques will obviously progress to better characterization of tissues' optical properties, improvement of laser properties and detection techniques, as well as new mathematical approaches of the inverse problem.

We believe that this field of research is still open to new ideas, and new experimental schemes leading to new breakthroughs in the delicate problem of using light in highly scattering media.

Acknowledgements

I would like to warmly thank all the research staff (Arnaud Dubois, Benoit Forget, François Ramaz, Vincent Loriette) and PhD students (Michael Atlan, Kate Grieve, Gael Moneron, Juliette Selb) at ESPCI whose results have been used to illustrate this article. I also acknowledge discussions with my colleagues Pr Paul French (Imperial College London) and Dr Amir Gandbjbakhche (NIH Bethesda).

See also

Coherence: Coherence and Imaging. **Holography, Techniques:** Digital Holography. **Imaging:** Interferometric Imaging. **Interferometry:** White Light Interferometry. **Microscopy:** Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy. **Tomography:** Optical Coherence Tomography; Tomography and Optical Imaging.

Further Reading

Boccara AC and Fink M (eds) (2001) *Optical and Acoustical Imaging of Biological Media*. Compte Rendu de l'Académie des Sciences, Physics and Astrophysics série IV tome 2, no. 8. Paris: Elsevier.

Born M and Wolf E (1997) *Principles of Optics*. New York: Oxford University Press.

Goodman JW (2000) *Statistical Optics*. New York: Wiley-Interscience.

Hecht E and Zajac A (1979) *Optics*. Reading, MA: Addison-Wesley.

Sebbah P (ed.) (2001) *Waves and Imaging through Complex Media*. Dordrecht: Kluwer Academic Publishers.

Van de Hulst HC (1981) *Light Scattering by Small Particles*. New York: Dover.

Van Tiggelen B and Skipetrov S (eds) (2003) *Wave Scattering in Complex Media: From Theory to Applications*. Dordrecht: Kluwer Academic Publishers.

Infrared Imaging

K Krapels and R G Driggers, US Army Night Vision and Electronic Sensors Directorate, Fort Belvoir, VA, USA

Published by Elsevier Ltd.

Introduction

All objects above 0 Kelvin emit electromagnetic radiation associated with the thermal activity on the surface of the object. For terrestrial temperatures (around 300 Kelvin), objects emit a good portion of the electromagnetic flux in the infrared part of the electromagnetic spectrum. The visible band spans wavelengths from 0.4 to 0.7 micrometers (infrared engineers typically use micrometer/μm for wavelength rather than the nanometer or wavenumber more commonly used in other fields). Infrared imaging devices convert energy in the infrared portion of the electromagnetic spectrum into images that can be created by the human eye. The unaided human eye cannot image infrared energy because the lens of the eye is opaque to infrared radiation.

The infrared spectrum begins at the red end of the visible spectrum where the eye can no longer sense energy. It spans from 0.7 μm to 100 μm. The infrared spectrum is, by common convention, broken into five different bands (this may vary according to the application). The bands are typically defined as: near infrared (NIR) from 0.7 to 1.0 μm; short wavelength infrared (SWIR) from 1.0 to 3.0 μm; mid-wavelength infrared (MWIR) from 3.0 to 5.0 μm; long wavelength infrared (LWIR) from 8.0 to 14.0 μm; and far infrared (FIR) from 14.0 to 100 μm. These bands are depicted graphically in Figure 1, which shows the atmospheric transmission for a 1-kilometer horizontal ground path for a 'standard' day in the United States. These types of transmission graphs can be tailored for any conditions, using sophisticated atmospheric models such as MODTRAN (from www.ontar.com). Note that

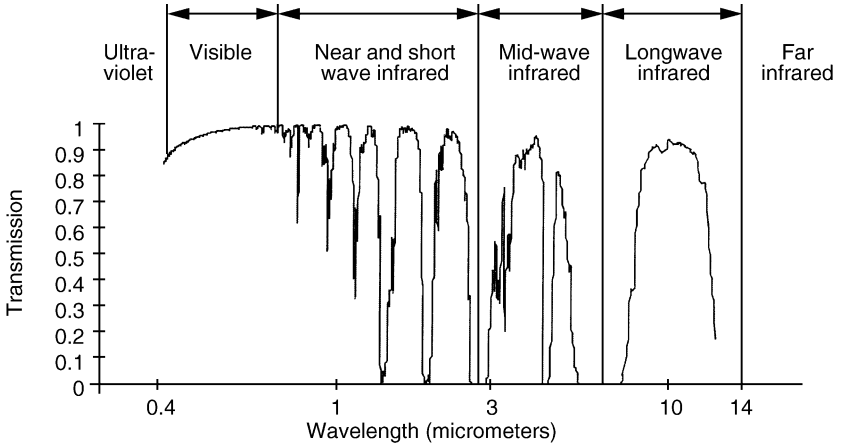


Figure 1 Visible and infrared spectra.



Figure 2 Visible image–reflected energy.

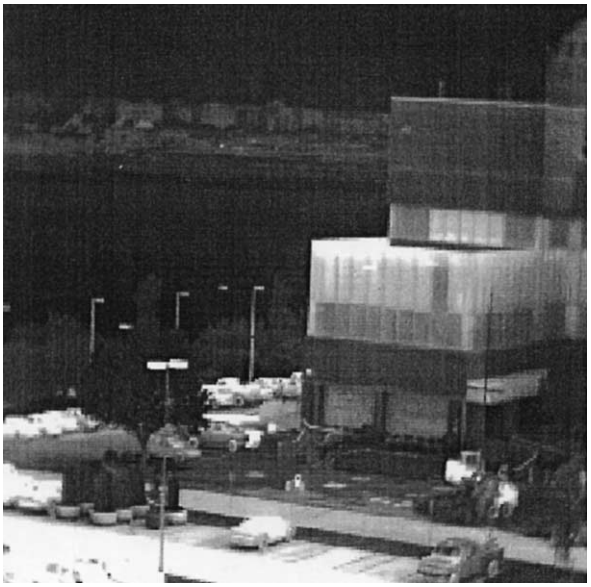


Figure 3 Infrared image–self-emitted energy.

there are many atmospheric 'windows', such that an imager designed with such a band selection can see through the atmosphere.

The primary difference between a visible spectrum camera and an infrared imager is the physical phenomonology of the radiation from the scene being imaged. The energy used by a visible camera is predominantly reflected solar, or some other illuminating energy, in the visible spectrum. The energy imaged by infrared imagers (Forward Looking Infrared, commonly known as FLIRs) in the MWIR and LWIR bands, is primarily self-emitted radiation. From Figure 1, the MWIR band has an atmospheric window in the 3 to 5 μm region and the LWIR band has an atmospheric window in the 8 to 12 μm region. The atmosphere is opaque in the 5 to 8 μm region, so it would be pointless to construct a camera that responds to this waveband.

Figures 2 and 3 are images that show the difference in the source of the radiation sensed by the two types of cameras. The visible image (Figure 2) is light that was provided by the Sun, propagated through Earth's atmosphere, reflected off the objects in the scene, traversed through a second atmospheric path to the sensor, and then imaged with a lens and a visible band detector. A key here is that the objects in the scene are represented by their reflectivity characteristics. The image characteristics can also change by any change in atmospheric path or source characteristic change. The atmospheric path characteristics from the Sun to the objects change frequently, because the Sun angle changes throughout the day and also weather and

cloud conditions change. The visible imager characterization model is therefore a multipath problem that is extremely difficult.

The LWIR image given in Figure 3 is obtained primarily by the emission of radiation by objects in the scene. The amount of electromagnetic flux depends on the temperature and emissivity of the objects. A higher temperature and a higher emissivity correspond to a higher flux. The image shown is 'white-hot' or the whiter a point in the image the higher the flux leaving the object. It is interesting to note that trees have a natural self-cooling process since a high temperature can damage foliage. Objects that have absorbed a large amount of solar energy are hot and are emitting large amounts of infrared radiation. This is sometimes called 'solar loading'.

The characteristics of the infrared radiation emitted by an object are described by Planck's black-body law in terms of spectral radiant emittance

$$M_\lambda = \varepsilon(\lambda)\frac{c_1}{\lambda^5(e^{c_2/\lambda T}-1)}\ \text{W/cm}^2\ \mu\text{m} \qquad [1]$$

where c_1 and c_2 are constants of 3.7418×10^4 W μm^4/cm^2 and 1.4388×10^4 μm K. The wavelength, λ, is provided in micrometers and $\varepsilon(\lambda)$ is the emissivity of the surface. A black-body source is defined as an object with an emissivity of 1.0, so it is a perfect emitter. Source emissions of black-bodies at typical terrestrial temperatures are shown in Figure 4. Often, in modeling, the terrestrial background temperature is assumed to be 300 K. The source emittance curves are shown for other temperatures for comparison purposes. One curve corresponds to an object colder than the background and two curves correspond to temperatures hotter than the background.

Planck's equation describes the spectral shape of the source as a function of wavelength. It is readily apparent that the peak shifts to the left (shorter wavelengths) as the body temperature increases. If the temperature of a black-body is increased to that of the Sun (5900 Kelvin), the peak of the spectral shape would decrease to 0.55 μm or green light. This peak wavelength is described by

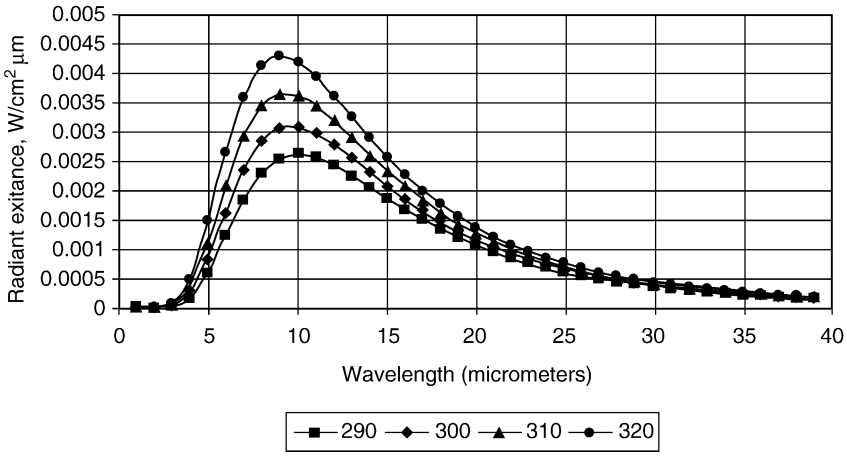


Figure 4 Planck's blackbody radiation curves for four temperatures from 290 K to 320 K.

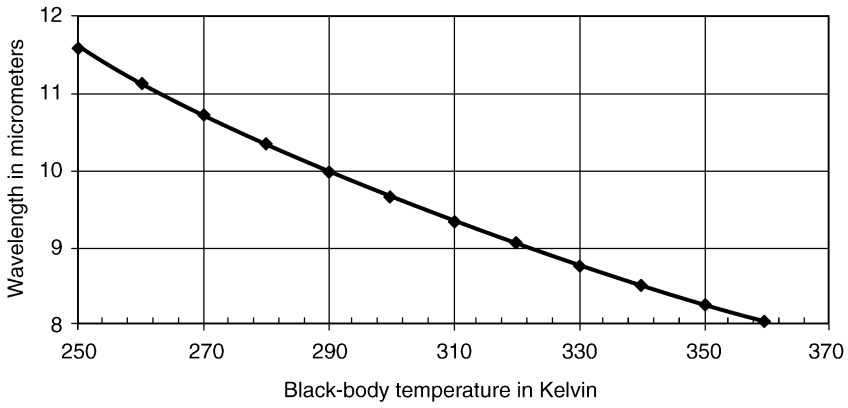


Figure 5 Location of peak of black-body radiation, Wien's Law.

Wien's displacement law

$$\lambda_{max} = 2898/T, \ \mu m \qquad [2]$$

Figure 5 shows the peak wavelength as a function of temperature through the longwave region. It is important to note that the difference between the black-body curves is the 'signal' in the infrared bands. For an infrared sensor, if the background is at 300 K and the target is at 302 K, the signal is the difference in flux between these curves. Signals in the infrared ride on very large amounts of background flux. However, in the visible band this is not the case. For example, consider the case of a white target on a black background. The black background is generating no signal, whereas the white target is generating a maximum signal (given that the sensor gain has been adjusted). Dynamic range may be fully utilized in a visible sensor. In the case of an IR sensor, a portion of the dynamic range is used by the large background flux radiated by everything in the scene. This flux never has a small value, hence sensitivity and dynamic range requirements are much more difficult to satisfy in IR sensors than in visible sensors. Table 1 summarizes this characteristic numerically for the 300 K background and targets shown in Figure 4.

A typical infrared imaging system scenario is depicted in Figure 6. The scene consists of two major components, the target and the background. In an IR scene, the majority of the energy is remitted from the constituents of the scene. This emitted energy is transmitted through the atmosphere to the sensor. As it propagates through the atmosphere it is degraded by absorption and scattering. Obscuration by intervening objects and additional energy emitted by the path also affect the target energy. All of these contributors, which are not the target of interest, essentially reduce one's ability to discriminate the target. So, at the entrance aperture for the sensor, the target signature is reduced from the values observed at the target. Then, the signal is degraded by the optics and scanner (as applicable) of the sensor. The energy is then sampled by the detector array and converted to electrical signals. Various electronics amplify and condition this signal before it is presented to either a display for human interpretation or an algorithm like an automatic target recognizer for machine interpretation. A linear systems approach to modeling allows the components' transfer functions to be treated separately as contributors to the overall system performance. This approach allows for straightforward modifications to a performance model for changes in the sensor or environment when performing trade-off analyses.

Table 1 Signal/dynamic range limitation in IR bands

Blackbody temperature	*Exitance in 8–12 μm band (w/cm^{-2})*	*Contrast*
290 ($-10\Delta T$ K target)	1.25E − 02	−8.48%
300 (background)	1.48E − 02	
310 ($+10\Delta T$ target)	1.74E − 02	7.97%
320 ($+20\Delta T$ target)	2.02E − 02	15.39%

History of Infrared Imaging

Night vision systems began to be developed extensively during World War II. Satisfying the requirement to image at night was approached with two different methodologies. The first method was image intensification (I^2). This method involves amplification of any small light that was available and displaying it directly to the eye. Typically I^2 devices are sensitive to visible light and a small portion of the short-wave IR (SWIR, 0.7–3.0 μm) band. They are often classified (along with visible band imagers) as electro-optical (EO) imagers. With an I^2 imager, there must be some source of illumination for them to function well (as little as starlight is sufficient

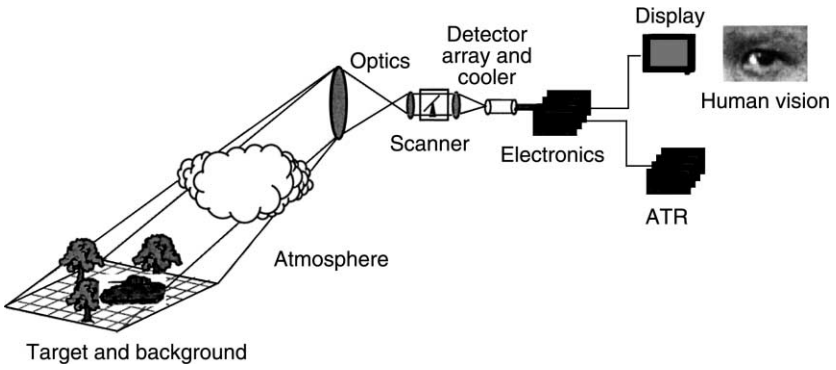


Figure 6 Typical infrared imaging scenario.

Table 2 Applications of infrared imaging

Category	*User*	*Application*
Military	Intelligence analyst	Intelligence surveillance and reconnaissance
	Pilot, driver	Navigation/pilotage
	Gunner, weapons officer	Targeting/fire control
	Air defense	Search and track
Commercial	*Civil government*	
	Police	Surveillance, fugitive tracking
	Fire	Rescue, hot-spot location
	Environmental	
	EPA	Emission tracking
	Interior department	Resource management
	NIMA\USGS	Mapping
	NOAA\MWS	Weather forecasting
	Industrial	
	Manufacturers	Machine vision
	Maintenance	Non-destructive testing
	Medical	
	Doctors	Diagnostic imaging

for approx 20/40 acuity with current systems). The second type of night vision device is the infrared or thermal imager (also known as FLIRs). FLIRs are sensitive to the radiation in either the mid-wave IR (MWIR, 3–5 μm) or long-wave IR (LWIR, 8–14 μm) bands.

The first infrared cameras used photographic film that was sensitive to infrared wavelengths. Intelligence and reconnaissance imaging from aircraft drove the infrared system development as a night augmentation to existing visible cameras. These imagers used either continuous moving film or stationary film, like an ordinary camera. A slit aperture was used in the continuous moving film cameras and aircraft motion provided the scan of the scene along the film to build a continuous image. The introduction of television, as electronic imaging in the visible spectrum, led to the development of electronic imagers in the infrared band. Electronic infrared imagers were not restricted to looking down and using aircraft motion for scan, hence these devices came to be known as FLIRs (Forward Looking InfraRed). Today, an electronic infrared imager may be called a FLIR or a thermal imager, with the trend being to use FLIR to describe military applications.

Early FLIRs were often accompanied by illumination devices. They were not very sensitive compared to modern FLIRs so active illumination was necessary. Infrared searchlights or illuminators were used to get higher signal-to-noise ratios. These types of FLlRs are called 'active' IR imagers, while FLIRs that do not use illuminators are 'passive' imagers. Generally, active imagers are used in civil applications and are declining in military use. Broadly, applications for thermal imagers fall into two categories; military and commercial. Table 2 lists some of the purposes and users for modern thermal imagers. The design and performance criteria vary widely for some of the applications.

Types of Infrared Imagers

Infrared imagers are classified by different characteristics: scan type, detector material/cooling requirements, and detector physics. The scan type refers to the method used to construct the electronic image. The camera may use a single detector which is raster scanned over the input scene to build an image. Alternatively, a parallel scan uses a linear array of detectors scanned across a scene to build an image. The latest advances in materials have led to staring arrays of detectors. In a staring system, a detector is present in two dimensions to represent each image pixel and no mechanical motion of the focal plane is necessary to construct the image. There are some hybrid FLIR types that combine the different imaging techniques. Usually this combination leads to an improvement in signal-to-noise ratio or to reduce undersampling.

The second classification, by detector material/cooling requirements, usually describes FLIRs built using differing materials. For example, a typical LWIR FLIR material is mercury cadmium telluride (HgCdTe or MCT). In order to achieve high sensitivity, these devices are cooled to decrease dark current. Usually the cooling is based on liquid nitrogen or a cryogenic cooler, and the detectors operate at 77 K. Another common detector material is indium antimonide (InSb), which is used for MWIR FLIRs and is also cooled. A newer class of infrared cameras is not cooled, being referred to as 'uncooled' FLIRs. These uncooled FLIRs are microbolometers (resistive elements) or pyrometers (capacitive elements) and have less sensitivity than cooled imagers. Typically, the cooled imagers are comprised of photon detectors while the uncooled imagers are based on thermal detectors, the uncooled FLIRs being used in lower-performance applications.

Infrared Imager Performance

There are three general categories of infrared sensor performance characterizations. The first is sensitivity and the second is resolution. When end-to-end, or human-in-the-loop (HITL), performance is required, the third type of performance characterization

describes the visual acuity of an observer through a sensor. The former two are both related to the hardware and software that comprises the system, while the latter includes both the sensor and the observer. Sensitivity is determined through radiometric analysis of the scene/environment and the quantum electronic properties of the detectors. Resolution is determined by analysis of the physical optical properties, the detector array geometry, and other degrading components of the system in much the same manner as complex electronic circuit/signals analysis.

Sensitivity describes how the sensor performs with respect to input signal level. It relates noise characteristics, responsivity of the detector, light gathering of the optics, and the dynamic range/quantization of the sensor. Radiometry describes how much light leaves the object and background and is collected by the detector. Optical design and detector characteristics are of considerable importance in sensor sensitivity analysis. In infrared systems, noise equivalent temperature difference (NETD) is often a first-order description of the system sensitivity. The 3D noise model describes more detailed representations of sensitivity parameters (see Further Reading).

The second type of measure is resolution. Resolution is the ability of the sensor to image small targets and to resolve fine detail in large targets. Modulation transfer function (MTF) is the most widely used resolution descriptor in infrared systems. Alternatively, it may be specified by a number of descriptive metrics such as the optical Rayleigh criterion or the instantaneous field-of-view of the detector. Where these metrics are component-level descriptions, the system MTF is an all-encompassing function that describes the system resolution.

Sensitivity and resolution can be competing system characteristics and they are the most important issues in initial studies for a design. For example, given a fixed sensor aperture diameter, an increase in focal length can provide an increase in resolution, but may decrease sensitivity. Typically, visible band systems have plenty of sensitivity and are resolution-limited, whereas infrared imagers have been more sensitivity-limited. With staring infrared sensors, the sensitivity has seen significant improvements.

Often metrics, such as NETD and MTF, are considered to be separable. However, in an actual sensor, sensitivity and resolution performance are not independent. As a result, minimum resolvable temperature difference (MRT or MRTD) has become the primary performance metric for infrared systems. MRT is a quantitative performance measure in terms of both sensitivity and resolution and a simple MRT curve is shown in Figure 7. The performance is bounded by the sensor's limits and the observer's limits. The temperature difference (or thermal contrast) required to image smaller details in a scene increases with detail size. The inclusion of observer performance yields a single sensor performance characterization, MRT. It describes sensitivity as a function of resolution, and includes the human visual system.

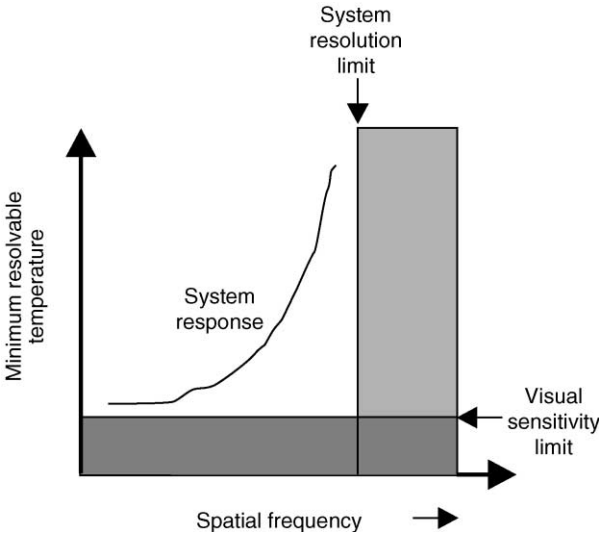


Figure 7 IR imager performance.

Many different sensor characteristics may be considered to increase the fidelity of a sensor model. A list of typical infrared imaging system parameters is given in Figure 8. When considering performance in the system sense, the groupings are often blocks in the system diagram, each with separate transfer functions. This approach works well unless the linear shift invariance assumption is not valid.

General Characteristics of Infrared Imagers

There are a number of imaging characteristics that make infrared systems a little different than conventional visible imaging systems. These include source flux levels, detector charge storage and sensitivity, detector size, diffraction blur, sampling, and uniformity characteristics.

First, we consider the source flux levels. The daytime and night-time flux levels (in photons per second per square centimeter) on Earth in the visible (0.3 to 0.7 μm) is 1.5×10^{17} and around 1×10^{10}, respectively. In the mid-wave (3 to 5 μm), the daytime and night-time flux levels are 4×10^{15} and 2×10^{15} μm where the flux is a combination of

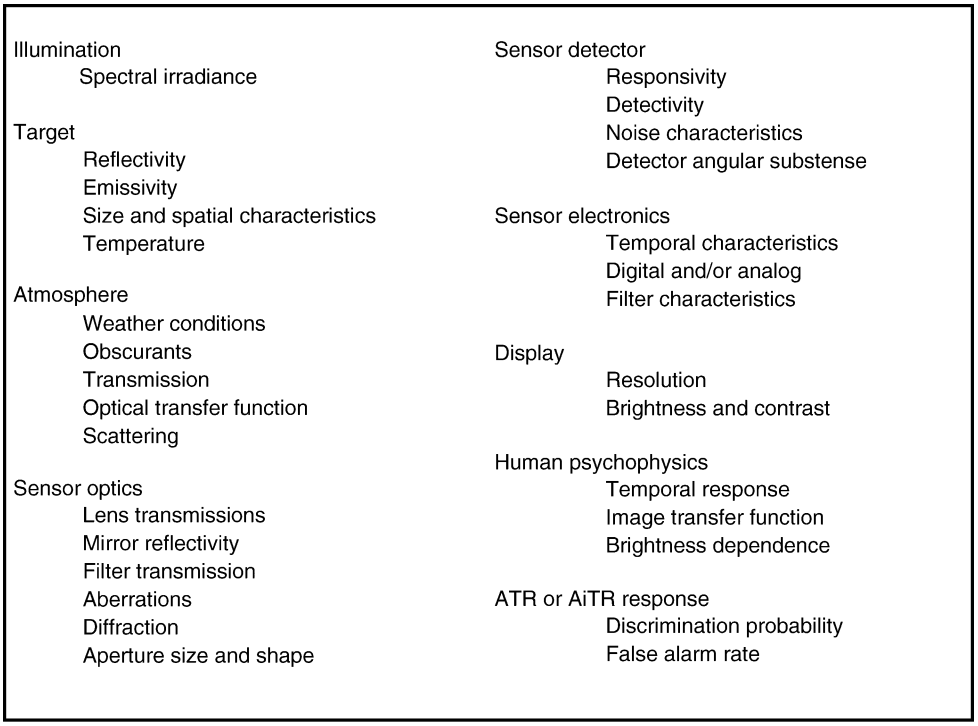


Figure 8 Factors in infrared sensor modeling.

emitted and solar reflected flux. In the longwave, the flux is primarily emitted where both day and night yield a 2×10^{17} μm frequency, the AMOP drops rapidly to zero. The effect is to extend the prediction of the MRT beyond the half sample rate and introduce a parameter called minimum temperature difference perceived (MTDP) to replace MRT in the range performance estimate and also in the lab to characterize an undersampled system.

The concept of MTDP follows from observation of standard 4-bar images as the pattern frequency passes the imager half sample rate. The image changes from four bars to three, two, and finally one (can be perceived). The phase of the target image on the detector array may need to be adjusted to observe this progression. The standard MRT measurement requires that all four bars be resolvable by the observer during the measurement process. For use in the lab, the MTDP is defined as the minimum temperature difference at which two, three, or four bars can be resolved in the image of the standard 4-bar test pattern by an observer, with the test pattern positioned at optimum phase. The optimum phase is the phase at which two, three, or four bars are 'best perceived'. TRM3 uses the standard definition of MRT as for adequately sampled imagers. The TRM3 Approach Model is implemented if the imager is considered undersampled as defined when the prefilter MTF at the half-sample frequency is larger than 10%. The MTDP equation is given by

$$\mathrm{MTDP}(f) = \frac{\frac{\pi}{2}\mathrm{SNR}_{\mathrm{th}}\Psi(f)}{\mathrm{AMOP}(f)} \qquad [3]$$

where $\mathrm{SNR}_{\mathrm{th}}$ is a threshold signal-to-noise ratio and Ψ is the total system filtered noise. MTDP is used in the same manner as MRTD in range calculations.

Another approach is the triangle orientation discrimination (TOD) threshold. In the TOD, the test pattern is a (nonperiodic) equilateral triangle in four possible orientations (apex up, down, left, or right), and the measurement procedure is a robust 4AFC (four alternative forced-choice) psychophysical procedure. In this procedure, the observer has to indicate which triangle orientation he sees, even if he is not sure. Variation of triangle contrast and/or size leads to a variation in the percentage correct between 25% (complete guess) and 100%, and by interpolation the exact 75% correct threshold can be obtained. A complete TOD curve (comparable to an MRTD curve) is obtained by plotting the contrast thresholds as a function of the reciprocal of the triangle angular size (Figure 9).

The TOD method has a large number of theoretical and practical advantages: it is suitable for undersampled and well-sampled electro-optical and optical

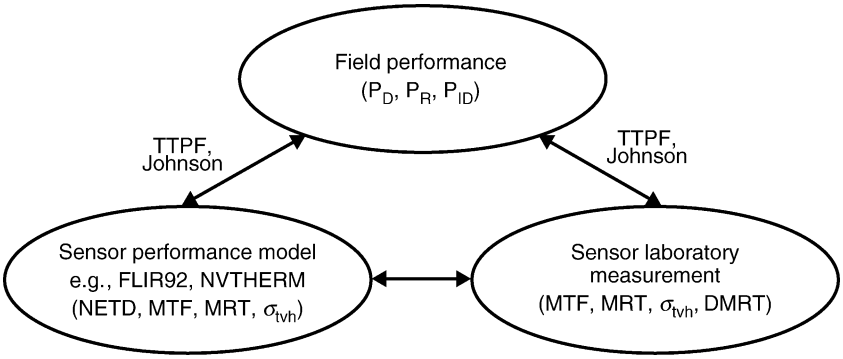


Figure 9 Modeling, testing and performance triangle.

imaging systems in both the thermal and visual domains, it has a close relationship to real target acquisition, and the observer task is easy. The results are free from observer bias and allow statistical significance tests. The method may be implemented in current MRTD test equipment with little effort, and the TOD curve can be used easily in a target acquisition (TA) model such as ACQUIRE. Two validation studies with real targets show that the TOD curve predicts TA performance for under-sampled and well-sampled imaging systems very well. Currently, a theoretical sensor model to predict the TOD (comparable to NVTherm or TRM3) is under development. The lab measurement and field performance appear sound, but the model is not yet available.

Testing Infrared Imagers: NETD, MTF, and MRTD (Sensitivity, Resolution and Acuity)

In general, imager sensitivity is a measure of the smallest signal that is detectable by a sensor. Sensitivity is determined using the principles of radiometry and the characteristics of the detector. For infrared imaging systems, noise equivalent temperature difference (NETD) is a measure of sensitivity. The system intensity transfer function (SITF) can be used to estimate the NETD, which is the system noise rms voltage over the noise differential output. NETD is the smallest measurable signal produced by a large target (extended source).

Equation [4] describes NETD as a function of noise voltage and the system intensity transfer function. The measured values are determined from a line of video stripped from the image of a test target, as depicted in Figure 10. A square test target is placed before a black-body source. The delta T is the difference between the black-body temperature and the mask. This target is then placed at the focal point of an off-axis parabolic mirror. The mirror serves the purpose of a long optical path length to the target, yet relieves the tester from concerns over atmospheric losses to the temperature difference. The image of the target is shown in Figure 10b. The SITF slope for the scan line in Figure 10c is the $\Delta S/\Delta T$ where ΔS is the signal measured for a given ΔT. The N_{rms} is the background signal on the same line.

$$\text{NETD} = \frac{N_{rms}[\text{volts}]}{\text{SITF\_Slope}[\text{volts/K}]} \qquad [4]$$

Resolution is a general term that describes the size of resolvable features in an imager's output. With infrared imaging systems, the resolution is described by the system modulation transfer function (MTF). Consider Figure 11 for the determination of MTF, from either point spread function or edge spread function (psf or esf). In Figure 11a, a test target is placed at the focal point of a collimator. The target may be a point, edge, or line as shown in Figure 11b. The IR system images this target. The output of a detector scan, or a line of detectors in the staring array case, is taken for analysis, in Figure 11c. For the edge spread function, the spatial derivative is taken to get the psf. The Fourier transform is then calculated to give the system MTF as shown in Figure 11d. The point spread function is really the impulse response of the system, so a smaller psf is desirable, where a wide MTF is desirable. Such a psf/MTF combination gives a higher resolution.

The MRTD of an infrared system is defined as the differential temperature of a four bar target that makes the target just resolvable by a particular observer. It is a measure of observer visual acuity when a typical observer is using the infrared imager. It results in a descriptor, which is a function not just of a single value. It is a plot of sensitivity as a function of spatial frequency.

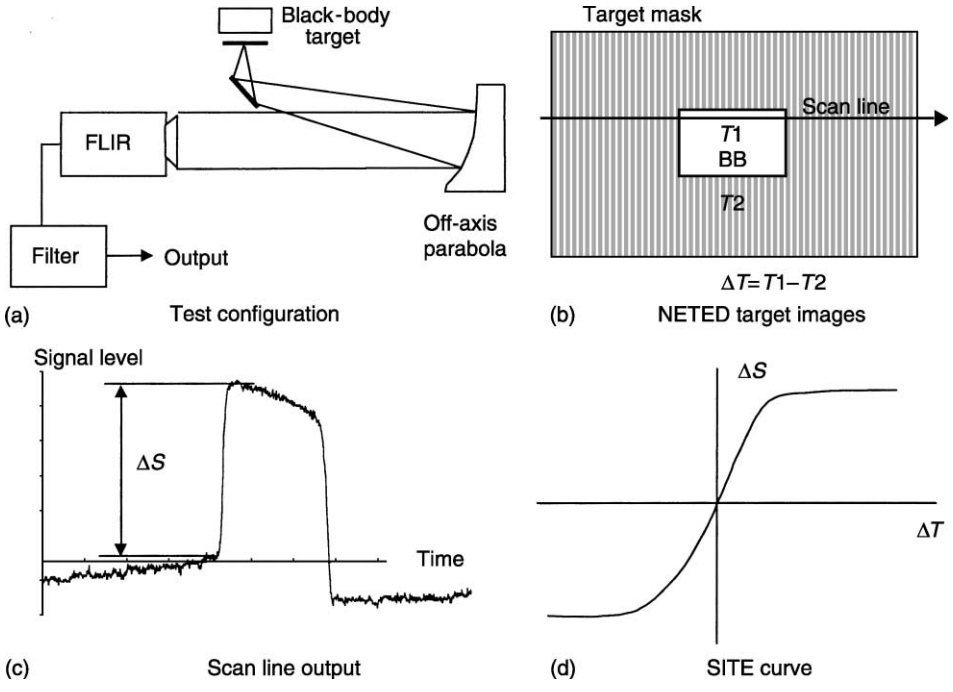


Figure 10 NETD measurement.

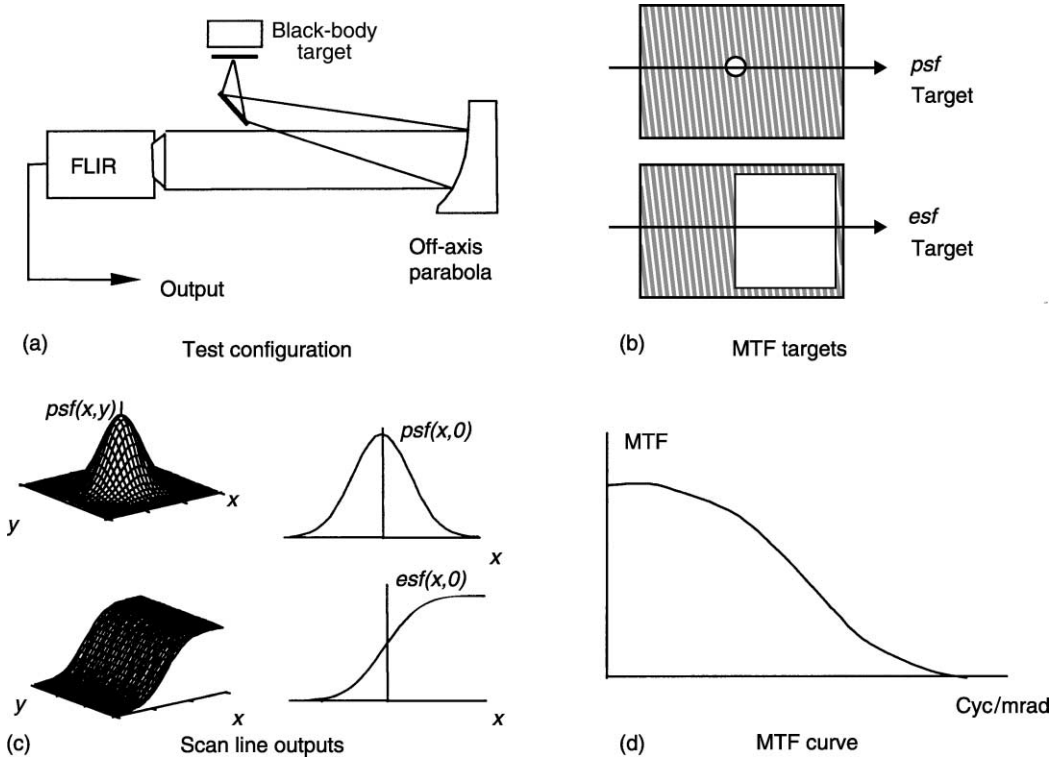


Figure 11 MTF measurement.

This parameter can be extended to field performance using Johnson's criteria.

As a laboratory measurement, it uses a 7:1 aspect ratio bar target. The procedure is shown in Figure 12. A bar target mask, with a black-body illuminator, is placed at the focal plane of a collimator (Figure 12a and 12b). The temperature is increased from a small value until the bars are just resolvable to a trained

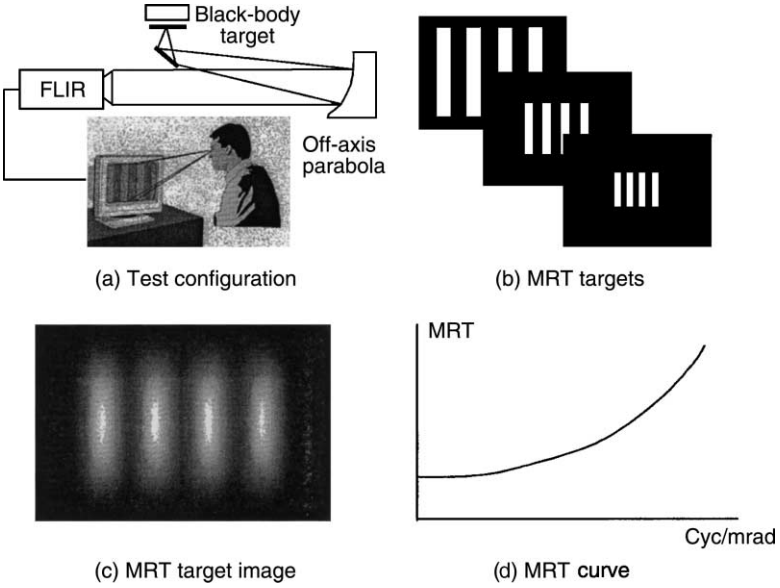


Figure 12 MRTD measurement.

observer (Figure 12c). Then the temperature difference is decreased until the bars are no longer visible. These are averaged for a particular target. The same procedure is performed with a negative contrast bar target. The average of the positive and negative values is the MRT for a particular spatial frequency. These data points are plotted and a curve may be fitted to interpolate/extrapolate performance at other than the discrete spatial frequencies of the targets, as in Figure 12d.

Summary

We have provided a general description of infrared imaging systems in terms of characteristics, modeling, field performance, and performance measurement. The characteristics of infrared imagers continue to change. Significant changes have occurred in the past five years, to include the development of higher performance uncooled imagers and ultra-narrow field of view photon systems. Large format detector arrays are commercially available in the mid-wave and are becoming more available in the longwave. Still in the research phase are dual-band focal plane level. At first sight, it appears that the longwave flux characteristics are as good as a daytime visible system; however, there are two other factors limiting performance. First, the energy bandgaps of infrared photons are much smaller than those photons in the visible, so the detectors suffer from higher dark current. The detectors are usually cooled to reduce this effect. Second, the reflected light in the visible is modulated with target and background reflectivities that typically range from 7 to 20%. In the infrared, all terrestrial objects are emitting with an emission temperature of around 300 Kelvin. Typically, a two-degree equivalent black-body difference in photon flux is considered a high contrast target to background flux. The flux difference between two black-bodies of 302 K compared to 300 K can be calculated in a manner similar to that shown in Figure 4. This is the difference in signal that provides an image, so note the difference in signal compared to the ambient background flux. In the longwave, the signal is three percent of the mean flux and in the mid-wave, it is six percent of the mean flux. This means that there is a large flux pedestal associated imaging in the infrared. Unfortunately, there are two problems accompanying this large pedestal. First, photon noise is the dominant noise and is determined by the mean of the pedestal and is compared to the small signal differences. Second, the charge well storage in infrared detectors is limited to around 10^7 charge carriers. A longwave system in a hot desert background would generate 10^{10} charge carriers in a 33 millisecond integration time. Smaller F-numbers, spectral bandwidths, and integration times are used so that the charge well does not saturate. This results in SNR of 10 to 30 times below the ideal. It has been suggested that some on-chip compression may be a solution to the well pedestal problem. In many scanning FLIR systems, the pedestal is eliminated by AC-coupling the detector signals. Infrared focal plane array (IRFPA) readout circuits have been previously designed and fabricated to perform charge skimming and charge partitioning.

Another major difference in infrared systems, from that of visible systems, is the size of the detector and diffraction blur. For infrared photon detectors, typical sizes range from 25 to 50 μm, where visible systems can be fabricated with 3 μm detectors. Also, the diffraction blur for the longwave is more than ten times larger than the visible blur and mid-wave blur is eight times larger than visible blur. Therefore, the image blur, due to diffraction and detector size, is much larger in an infrared system than a visible system. Also, it is common for infrared staring arrays to be sampling limited where the sample spacing is larger than the diffraction blur and the detector size. Dither and microscanning are frequently used to enhance performance.

Finally, both infrared photon and thermal detectors in staring arrays have responsivities that vary dramatically and it is current practice to correct for the resulting nonuniformity (nonuniformity correction or NUC). The nonuniformity can cause fixed pattern noise in the image that can limit the performance of the system even more than temporal noise (especially in static imaging or stabilized target acquisition imaging).

Modeling Infrared Imagers

Otto Schade Sr. developed the earliest imaging system models and performance measures. His work describing television in the 1950s and 1960s, pioneered the way for the characterizations of imaging sensors used by engineers today. His performance measure for still pictures, moving pictures and television systems was based on an observer resolving a three bar periodic target.

John Johnson developed the technique of relating the acquisition of military targets in the field to laboratory measurements of resolvable bar targets. He divided the discrimination tasks into four categories: detection, orientation, recognition, and identification. This technique was termed an 'equivalent bar pattern approach'. It related performance on a simple test target to performance with complex object targets. Johnson viewed scale models and bar targets in the laboratory against a bland background. The smallest discernible barchart target yielded the maximum resolvable bar pattern frequency. These results, tabulated in Table 3 as cycles across the minimum dimension, were the basis for the discrimination methodology in widespread use today. Note that detection took only one cycle, but as the tasks got more specific or difficult, the requirement went as high as 8 cycles across the target minimum dimension.

Recognizing that typical vehicle aspect ratios were somewhat limited, Johnson and Lawson conducted further experiments. Paul Moser conducted some of his own, reanalyzed Johnson and Lawson's, and developed the concept that this task was related to the average or critical dimension instead of the minimum dimension. This led to the conversion of line pairs to resolution elements, introducing the second dimension. Additionally, he performed a similar experiment on marine targets, which is summarized in Table 4. Lloyd and Sendall developed the MRT metric, which combines both resolution and sensitivity characteristics. This measurement of MRT is described in detail in a following article, but there is a theoretical model that is used to evaluate new sensor designs and concepts. It relates the minimum temperature difference between bar pairs at which they are just resolvable as a function of spatial frequency. MRT describes the infrared imager sensitivity as a function of resolution. More specifically, MRT is a measure of thermal contrast sensitivity as a function of spatial frequency.

Ratches *et al.* developed the NVL (US Army Night Vision and Electronic Sensors Directorate formerly known as the Night Vision Laboratory) Static Performance Model, which predicted the end-to-end performance. It started with the target signature and carried through to the observer. The MRT, a laboratory measurement or a modeled value, was related to the probabilities of discrimination through the target transfer probability function, cumulative percentages related to Johnson's criteria.

Table 3 Johnson's criteria

Target	*Resolution per minimum dimension*			
Broadside view	*Detection*	*Orientation*	*Recognition*	*Identification*
Truck	0.9	1.25	4.5	8
M-48 Tank	0.75	1.2	3.5	7
Stalin Tank	0.75	1.2	3.5	6
Centurion Tank	0.75	1.2	3.5	6
Half-track	1	1.5	4	5
Jeep	1.2	1.5	4.5	5.5
Command car	1.2	1.5	4.3	5.5
Soldier (standing)	1.5	4.8	3.8	8
105 howitzer	1	1.5	4.8	6
Average	1 ± 0.25	1.4 ± 0.35	4.0 ± 0.8	6.4 ± 1.5

Table 4 Moser's data

Discrimination task	Line pairs/critical dimension
Detect ship	Aperiodic treatment
Classify as combatant	4
Recognize type	10

D'Agostino and Webb created the 3D noise model, both spatial and temporal considerations, for analyzing system noise. This technique divided the noise components into manageable and understandable components. It also simplified the integration of these effects into models.

The introduction of focal plane array (FPA) imagers has created a requirement to improve the model performance to account for the difference between scanning arrays and staring arrays. This change has made sampling an important area for improvement. A new model called NVTHERM is a result of the NVESD model improvement effort, which addresses undersampled imaging system performance.

Sensor characterization is seen in three ways: theoretical models, field performance, and laboratory measurements (**Figure 9**). Theoretical models are developed that describe sensor sensitivity, resolution, and human performance for the purpose of evaluating new conceptual sensors. Acquisition models, and other models, are developed to relate the theoretical models to field performance. This link allows theoretical models to be converted to field performance quantities (e.g., probabilities of detection, recognition, and identification). Field performance is measured in the field so that the theoretical models can be refined and become more accurate with advanced sensor developments. Since field performance activities are so expensive, methods for the direct measurement of sensor performance are developed for the laboratory. Field performance testing of every infrared sensor built, including buy-off, acceptance, and life-cycle testing, is ridiculous and out of the question. Laboratory measurements of sensor performance are developed such that, given these measurements, the field performance of a system can be predicted. Sensor characterization programs require accurate sensor models, field performance estimates with acquisition models, and repeatable laboratory measurements.

There are a few alternatives to the US NVTherm model. One candidate approach to undersampled imager modeling is Germany's TRM3, or the MDTP approach. The impact of undersampling in the image of a 4-bar target can readily be seen by simply observing the change in the image as a function of spatial frequency and phase. The spatial frequency is defined as line pairs per angular extent of the target, and the phase specifies the relative location of the target image to the detector array raster. These effects are obviously not independent of each other, but for each target there can be found an optimal phase where the observer can see the maximum amplitude modulation in the image of the target. MRT measurements in the past have utilized this variation with phase by allowing the observer to optimize the displayed image through target or system line-of-sight changes during the measurement process. For undersampled imagers, TRM3 addresses the problem of the MRT calculation's inability to predict beyond the half sample rate of the sensor, by replacing the MTF in the denominator of the MRT equation with an appropriately scaled term called the average modulation at optimum phase (AMOP). AMOP is the average signal difference in the image of the 4-bar standard pattern, with the test pattern positioned at optimum phase. AMOP oscillates between the presample MTF and the bar modulation. Beyond a frequency of 0.8 times the sample rate or 1.6 times the half sample arrays and quantum well detector systems. These systems will find their place in applications to include both military and commercial sectors.

Further Reading

Bijl P and Valeton MJ (1998) Triangle orientation discrimination: the alternative to minimum resolvable temperature difference and minimum resolvable contrast. *Optical Engineering* 37(7): 1976–1983.

D'Agostino J and Webb C (1991) 3-D analysis framework and measurement methodology for imaging system noise. *Proceedings of SPIE* 1488: 110–121.

Driggers RG, Cox P and Edwards T (1998) *Introduction to Infrared and Electro-Optical Systems*. Boston, MA: Artech House, p. 8.

Holst GC (1995) *Electro-Optical Imaging System Performance*. Winter Park, FL: JCD Publishing, p. 347.

Holst GC (1995) *Electro-Optical Imaging System Performance*. Winter Park, FL: JCD Publishing, p. 432.

Johnson J (1958) Analysis of image forming systems. *Proceedings of IIS* 249–273.

Johnson J and Lawson W (1974) Performance modeling methods and problems. *Proceedings of IRIS*.

Ratches JA (1973) *NVL Static Performance Model for Thermal Viewing Systems*. USA Electronics Command Report ECOM 7043, AD-A011212.

Schade OH (1948) Electro-optical characteristics of television systems. *RCA Review* IX(1–4).

Sendall R and Lloyd JM (1970) Improved specifications for infrared imaging systems. *Proceedings of IRIS* 14(2): 109–129.

Vollmerhausen RH and Driggers RG (1999) NVTHERM: next generation night vision thermal model. *Proceedings of IRIS Passive Sensors* 1.

Wittenstein W (1998) Thermal range model TRM3. *Proceedings of SPIE* 3436.

Interferometric Imaging

D L Marks, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Interferometric imaging is image formation by measuring the interference between electromagnetic signals from incoherently emitting or illuminated sources. To learn about interferometric image formation using coherent light, *see* **Holography, Techniques:** Overview. To understand what optical coherence is, *see* **Coherence:** Overview. To learn what kinds of sources produce incoherent light, *see* **Incoherent Sources:** Lamps. To learn about interferometric instruments, *see* **Interferometry:** Overview. To learn about interferometric instruments utilizing incoherent light, *see* **Interferometry:** White Light Interferometry. To better understand the effect of the coherence of light sources on image formation, *see* **Coherence:** Coherence and Imaging.

Interferometric imaging attempts to improve the resolution of images by interferometrically combining the signals from several apertures. By doing so, a resolution can be achieved that is superior to that achievable through standard image formation by any of the individual apertures. This is especially useful in astronomy, where resolving more distant and smaller stellar objects is always desirable. It is also finding increasing use in microscopy. It has several advantages over conventional noninterferometric imaging methods. The method can combine the light gathered from several imaging instruments to form an image superior to that which can be formed by any one of the instruments. It can also obviate the need to produce a single very large aperture to achieve the equivalent resolution, which in many cases is of impractical size. In addition, because it can produce a phase-resolved measurement, the data can be processed more flexibly to form a computed image estimate. There are also considerable disadvantages to interferometric imaging compared to conventional imaging methods. The amount of signal gathered is far less than could be gathered with an equivalent single aperture. It is difficult to mechanically or feedback stabilize the time delay between two widely separated apertures to obtain an accurate phase measurement. Often only a small bandwidth of the source light can be utilized, further reducing the available signal. The interference component of the combined signal can be very small. Typically an interferometric image must be computed from the data, rather than being directly measured on photographic film or an electronic focal plane array. Even with these disadvantages, interferometric imaging is frequently more feasible than building a single large aperture to achieve the highest available resolution images.

In contrast to interferometric imaging, a standard image forming instrument such as a telescope has a resolution limited by the telescope mirror aperture in the absence of aberrations and atmospheric turbulence. As telescope mirrors become very large, they become bulky, extremely heavy, and difficult to manipulate, as well as deform under their own weight and temperature gradients. However, adaptive optics is being more frequently utilized to dynamically correct for instrument aberrations as well as atmospheric turbulence. The images captured by telescopes are typically intensity-only images, which are amenable to image processing and deconvolution, but nevertheless lack phase information. Interferometric imaging is an alternative, where the resolution of a large aperture can be attained by measuring the interference between individual points of the electromagnetic field within an area equivalent to the large aperture extent. The interferometric combining of light from sub-apertures to achieve the resolution of a larger aperture is called aperture synthesis.

Unlike image formation by a lens, interferometric images are never physically formed. Conventional images are formed when diverging spherical waves emanating from sources are focused by a lens to converging spherical waves on a photographic film or an electronic sensor. In this case, the image information is directly contained in the exposure at each position on the sensor. Interferometric measurements contain the image information as statistical correlations between the electromagnetic fields at pairs of spatial points. The statistical properties of electromagnetic waves are modeled by optical coherence theory. Interferometric imaging works by measuring the statistical correlations between various pairs of points in the electromagnetic field, and then uses optical coherence theory to infer what the image of the source of the radiation is.

The light that emanates from most natural and many artificial sources is incoherent. Incoherent radiation is optical random noise. It is analogous to the sound of static heard on a radio receiving no signal. Because it is random, the fluctuations of the electromagnetic field produced by an incoherent source are completely unrelated to the fluctuations of any other source. The origin of the randomness is

usually from microscopic processes such as the thermal motions of electrons or the spontaneous emission of an excited atom. Because these microscopic processes tend to act independently of one another, the light produced by different sources tends to be mutually incoherent. Because incoherent sources do not radiate a deterministic field, the average amount of power they radiate is usually specified rather than the field itself, which is random.

However, as the fields propagate away from their sources, the fields can become partially coherent. To see this, consider the light waves propagating away from a single point-like incoherent source. The field produced at the source point consists of random fluctuations in time, but these fluctuations travel outward in a spherical wave at the speed of light. The fluctuations of the field of any two points an equal distance from the source will be the same, or perfectly correlated, because they arrive from the source at the same time. This is illustrated in Figure 1a, where a random wave is emanated from a point incoherent source. Because of the complete correlation, we can say that these two points are fully spatially coherent.

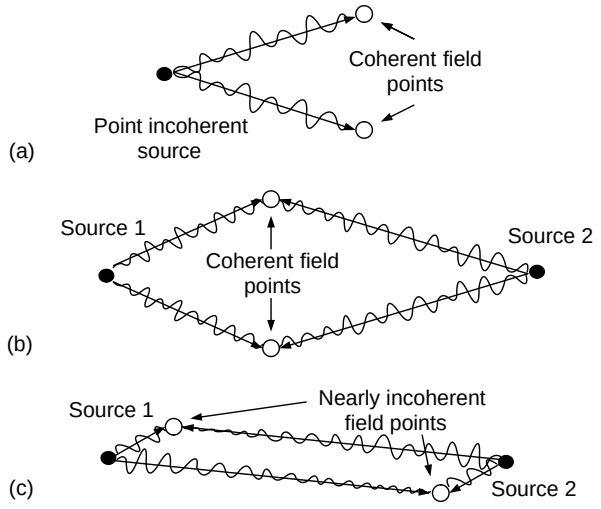


Figure 1 The distance between incoherent sources and two field points determines the coherence between the field points. (a) Because there is only one source, the field everywhere is determined by the radiation of that source. Therefore, the same fluctuations of the wave arrive at two points an equal distance from the source at the same time, and produce identical fields at both points, and so are completely coherent. (b) Two field points are now the same distance from two incoherent sources. Because both sources contribute identical fields to both points (because the fluctuations of the wave from both sources arrive at both points at the same time), the superposition of the fields from both sources at both field points are identical, and so the fields are coherent. (c) Two field points are now near two sources. Because each field point receives radiation mostly from its nearby source and very little from the other source, the fields at each field point are almost completely mutually incoherent.

So while the source itself produces random noise, the fluctuations of the field at points distant from the source become correlated because they originate from the same source.

Partially coherent waves occur when the measured field contains a superposition of the fields from more than one incoherent source. If there are now two point sources instead of one, each field point in space will receive radiation from both sources. The amount of the contribution from each source to each point in space depends on the distance between the field point and the source. In general, because both sources can be different distances from the two field points, the fluctuations at the field points will not be the same. For example, two field points that are the same distances from both sources will be perfectly correlated, or coherent, because they receive the same combination of fields from both sources. This situation is illustrated in Figure 1b. However, if one considers two points that are near both respective sources, the points will usually receive light from the nearby source. Because the sources are incoherent, and each point is mostly receiving radiation from its closer source, these points will be nearly uncorrelated, as shown in Figure 1c.

A spatial distribution of incoherent sources radiating various amounts of power will create a pattern of correlations in the field remote from the source. By using interferometry, the correlation between pairs of field points remote from the sources can be measured. The interference between two points in the electromagnetic field is achieved by relaying the two fields (e.g., with mirrors) to a common location that has an equal travel time for the fields from both points, where the fields are superimposed. The power of the superimposed beams is then detected, by a photocathode, for example. If the two signals being combined are incoherent, then no interference will take place, and the total power measured on the photodetector is simply the sum of the power of the two signals. However, if the two signals are partially coherent, there will be a deviation in the measured power from the sum of the power of the two signals. The magnitude of this deviation relative to the power of the constituent signals indicates the degree of partial coherence of the signal.

An interferometer, such as depicted in Figure 2, gathers the light from two areas of the electromagnetic field remote from an object and combines them together at the same point to form an interferogram. In this example, two telescopes of a known separation gather light from a star. The primary mirror of each telescope is of insufficient size to resolve the star as anything but a point. However, a telescope is used in practice to collect enough light to produce a

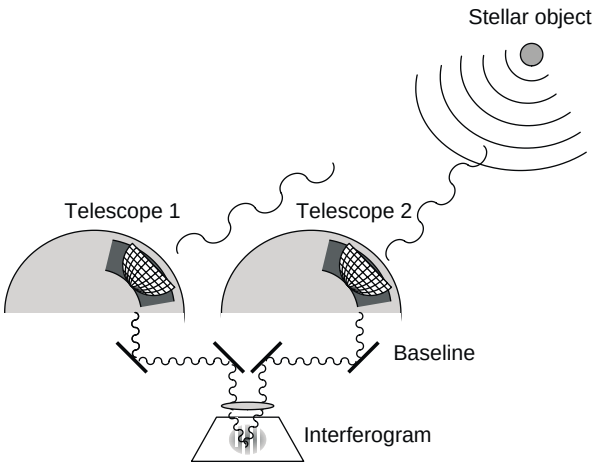


Figure 2 Two telescopes receive radiation from a stellar object. The light is collimated and directed down the baseline to a point halfway between the telescopes, where the two images are interfered together onto a focal plane.

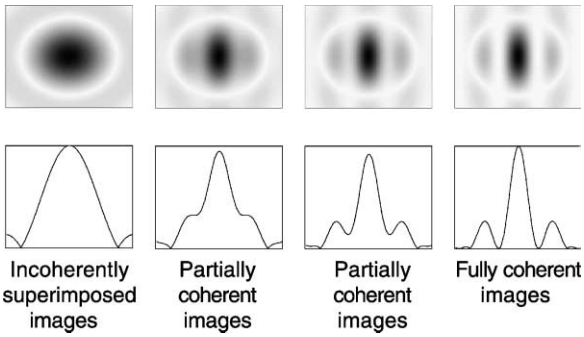


Figure 3 The image of two interfering wavefronts from a point object with various states of coherence.

measurable interference signal. Each telescope primary mirror forms a point image of the star. At the image plane of each telescope a mirror directs the point image to a measurement plane equidistant between the two telescopes. This line that joins the two telescopes is called the baseline, and it is the baseline length rather than the individual telescope mirror size that determines the minimally resolvable size. The image from each telescope is superimposed by mirrors (or perhaps beamsplitters) and imaged onto a focal plane, to get an image similar to that of Figure 3. If the light arriving from one telescope is blocked, then the image formed looks like an Airy ring pattern formed by an image, a point-like object, which is the incoherent image of Figure 3. However, when both signals are allowed to interfere, the image will have vertical stripes superimposed on it. If these stripes are very prominent, so that the image is fully darkened inside the stripes, then an image similar to the rightmost image of Figure 3 is obtained, complete intereference is occurring and the two field points at which the telescopes are situated are coherent or completely correlated. If the stripes disappear, then no interference is occurring and the two field points at which the telescopes reside are incoherent. These examples are shown in Figure 3, with images of two superimposed points with varying degrees of coherence on the top row, and a line plot of the measured field through the center of the images in the bottom row. The degree to which the contrast of the stripes is enhanced by coherence is referred to as the 'modulation' and indicates the degree or magnitude of partial coherence.

In addition to having a degree of coherence, the correlation between two electromagnetic field points also has a relative phase shift. This phase shift is given by the relative position of the interference maxima, as shown in Figure 4. As the relative phase between the

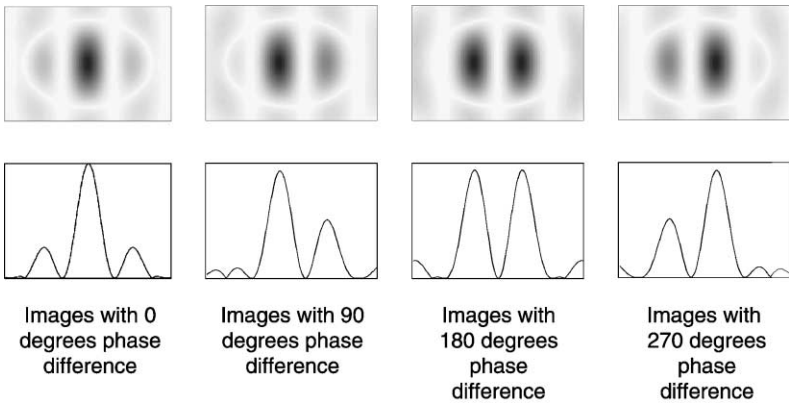


Figure 4 The image of two interfering wavefronts from a point object with various phase differences between the two wavefronts.

fields collected by the telescopes changes (which is a function of the path length difference between the light arriving at the two telescopes), the peaks of the interferogram between the two telescope images shift position. By observing the depth of the formed interference fringes and the position of the fringes, both the amplitude and the phase of the correlation between the two field points can be simultaneously measured. The amplitude of the modulation and the phase of the interference are used to compute the complex degree of coherence that is used in the computation of interferometric images.

In practice, the phase of the complex degree of coherence is much more difficult to measure than its magnitude. This is because while the amplitude tends to change very slowly as the positions of the two correlated points change, the phase is sensitive to changes in relative path length on the order of an optical wavelength. Even though a long baseline produces better resolution, the baseline length must be known to a small fraction of a wavelength to obtain a meaningful phase measurement. Minute vibrations in the positions of the mirrors on the baseline can easily cause fluctuations in the path length greater than a wavelength. This is observed as wild oscillations in the position of the fringes on the image plane. In addition, atmospheric turbulence can also produce large random phase shifts. Achieving a stable phase measurement is one of the most challenging engineering aspects of interferometric imaging.

To model realistic sources, we must determine the coherence produced by sources of a finite size. Finite size incoherent radiators can be regarded as an arrangement of an infinite number of point-like incoherent sources radiating various amounts of power. When an interferogram is made of a field which has two superimposed incoherent sources, the power of the interferograms that each would make alone is added together, rather than the fields of the interferograms. This is because when the fields of the two sources interfere, the phases of both sources are varying randomly over time. The two fields will randomly vary between constructively and destructively interfering with each other. Over a long time, the constructive and destructive interference averages out to no interference. We can then calculate what the interferogram of many incoherent sources combined is by computing the interferogram of a point source and adding the power of many such interferograms together.

To figure out what the power of an interferogram of an arbitrary object is, first consider what the interferogram of a single point source is. Consider a situation depicted in Figure 5, but with a single point

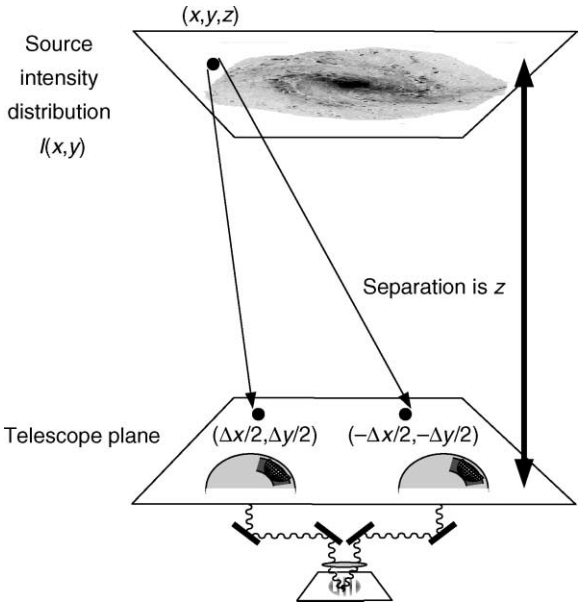


Figure 5 Diagram of source and telescope plane used for derivation of Van Cittert–Zernike theorem. A source at a distance z from the telescope plane has a power density $I(x,y)$. The telescope plane has two movable apertures symmetric about the origin of the telescope plane separated by $(\Delta x,\Delta y)$. The light from both apertures is directed to the center of the baseline at the origin of the telescope plane, where it is interfered on the sensor.

illuminator at the source plane positioned at (x,y,z) with radiant power I. We will have a receiving plane at a distance z from the source plane with two telescopes located symmetrically about the origin at $(\Delta x/2,\Delta y/2,0)$ and $(-\Delta x/2,-\Delta y/2,0)$. The fields received at the telescopes from the point will be interfered together at the center of the baseline, which is at the origin. We will assume that the point source produces a narrow band of radiation centered at wavelength λ. The point source produces a spherical wave that is received at both telescopes. The interference power is proportional to the two intensities measured at each telescope with the relative phase at the two telescopes causing constructive or destructive interference:

$$P \propto I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\phi \qquad [1]$$

where I_1 and I_2 are the intensities of the field at each telescope, and ϕ is the phase difference between the two fields at each telescope. We will assume that the point (x,y,z) is very far away from the telescopes, and so the intensities at the two telescopes are equal and proportional to the point intensity I, so that $P \propto I(1+\cos\phi)$. The phase ϕ is given by the difference in distance from the the source point at (x,y,z) to the receivers at $(\Delta x/2,\Delta y/2,0)$ and $(-\Delta x/2,-\Delta y/2,0)$. The phase is given by the difference in the Euclidean

distances:

$$\phi = \frac{2\pi}{\lambda}\left(\sqrt{z^2 + (x + \Delta x/2)^2 + (y + \Delta y/2)^2} - \sqrt{z^2 + (x - \Delta x/2)^2 + (y - \Delta y/2)^2}\right) \approx \frac{2\pi(x\Delta x + y\Delta y)}{z\lambda} \quad [2]$$

where the second term gives the difference in the distance to a first-order approximation in x and y. The interference intensity relation is then $P \propto I(1 + \cos(2\pi(x\Delta x + y\Delta y)/z\lambda))$.

If the source is now of a finite site, we can regard it as radiating an intensity (power per unit area) $I(x, y)$. To compute the total intensity of the interferogram, we integrate over all the sources in the source plane:

$$P(\Delta x,\Delta y) \propto \int I(x,y)(1+\cos(2\pi(x\Delta x+y\Delta y)/z\lambda))\mathrm{d}x\mathrm{d}y = \int I(x,y)\mathrm{d}x\mathrm{d}y + \int I(x,y)\cos(2\pi(x\Delta x+y\Delta y)/z\lambda)\mathrm{d}x\mathrm{d}y \quad [3]$$

This result can be interpreted as follows. The intensity of the interference when the two telescopes are separated by a distance $(\Delta x,\Delta y)$ is proportional to the real part of the two-dimensional Fourier transform of the intensity distribution $I(x,y)$ evaluated at spatial frequencies $(\Delta x/z\lambda,\Delta y/z\lambda)$, plus a constant intensity. This result, related to the Van Cittert–Zernike theorem, is a commonly used result in interferometric imaging. It is covered in more detail in **Coherence:** Overview.

By moving the telescopes to various horizontal and vertical separations and measuring the interference intensity, the real part of the Fourier transform of the intensity is inferred. In practice the imaginary part can also be found by varying the baseline optical path length difference by a quarter wavelength. With these samples of the Fourier transform, the inverse Fourier transform computes the image intensity distribution $I(x,y)$. This is how simple computational image formation occurs with interferometric data.

One application of the Van Cittert–Zernike theorem and interferometric imaging is the measurement of the diameter of stars. When observing coherence, the correlations between field points tend to decrease as the separation between the points grows. By observing this decrease of the coherence, the angular size of a star can be measured. This method was pioneered by Albert Michelson in 1890 with the invention of the Michelson Stellar Interferometer. It is needed because the aperture size required to measure the angular diameter of most nearby stars in the Milky Way is larger than the size of most telescope apertures. Stars usually have a circular profile that produces a circularly symmetric pattern of correlations. The coherence between field points as the distance between them is increased does not decrease monotonically, but disappears at certain separations, and has many revivals. The first separation distance at which interference disappears (and the degree of coherence is zero) for a circular profile object is $d_0 = 0.61\lambda/\alpha$, where λ is the wavelength of the light measured from the star, α is the angular diameter (in radians), and d_0 is the first separation from zero at which the interference disappears. In practice, each wavelength produces a null in interference at a different separation, so all but a small bandwidth of the star's light is filtered out before measurement to achieve a suitably strong interference null. For example, for the star Sirius, which has an angular diameter of 0.0068 arc seconds viewed from the Earth, measuring at a wavelength of $\lambda = 500$ nm (blue-green in the visible spectrum), the first separation distance at which no interference will be observed is at 18.5 m. This is larger than the largest available telescope aperture, and therefore an image cannot be directly formed that can resolve the size of Sirius.

Modern examples of stellar interferometry include the European Southern Observatory in Paranal, Chile, the CHARA Array at Mt Wilson, California, USA, and the NASA Jet Propulsion Laboratory Keck Interferometer in Hawaii, USA. The ESO main telescope system is the Very Large Telescope Interferometer (VLTI) consisting of four 8.2 m diameter telescopes as well as several smaller ones that can be moved independently, and will achieve milliarcsecond resolution through the interferometric combination of the telescopes. These telescopes are connected by underground tunnels through which the collected light from each telescope is combined. To measure the phase of the interference, extreme positioning precision is required in the relative delay between the light collected by the telescopes along the baseline, of 50 nm over 120 m, or 1 part in 2400 million, equivalent to 1.6 cm over the circumference of the Earth. Achieving this requires combinations of electronic and mechanical feedback systems with laser stabilized interferometric position measurement. The CHARA array has a very large 330-meter baseline, with 1-meter telescope apertures. The Keck Interferometer consists of two 10 m diameter telescopes with a baseline separation of 85 m. One of the main goals of achieving such high

resolution is to be able to observe planets and planet formation around distant stars, to identify possible signs of life in other solar systems.

A similar principle is used for interferometric image formation in radio astronomy. The chief differences are the methods used to collect, detect, and correlate the electromagnetic radiation. Radio astronomical observatories, such as the Very Large Array in New Mexico, USA, and MERLIN at the Jodrell Bank Observatory in Cheshire, UK, use collection dishes and radio antennas to directly measure the electric field, which is not feasible at optical frequencies. These fields are converted to electrical voltages, which can then be directly correlated using electronic mixers, rather than detecting the correlation indirectly through interference. Because the wavelength is much larger, the stability requirements are greatly relaxed. Radio and optical frequency interferometric images are formed in essentially the same way, using the Van Cittert–Zernike theorem.

Interferometric imaging methods have also become of interest for microscopy applications. Aperture synthesis is not needed in microscopy, because the objects of interest are very small and therefore single apertures can be used to achieve diffraction-limited resolution. However, interferometric detection of partially coherent light affords greater flexibility because an image may be computed from the data rather than being directly detected. Typically this is achieved by creating an incoherent hologram of the object, typically through shearing interferometry. For example, by using a Rotational Shearing Interferometer, one can measure an incoherent hologram that can be used to produce a representation of an object with a greatly increased depth-of-field. In addition, direct detection of the partial coherence of sources can allow the methods of computed tomography to be used to produce three-dimensional images of volumes of incoherently emitting objects. Because of the versatility that interferometric detection can provide, interferometric imaging will surely be increasingly used outside of astronomy.

See also

Coherence: Coherence and Imaging; Overview. **Holography, Techniques:** Overview. **Incoherent Sources:** Lamps. **Interferometry:** Overview; White Light Interferometry.

Further Reading

Baldwin JE and Haniff CA (2002) The application of interferometry to optical astronomical imaging. *Philosophical Transactions of the Royal Society of London A* 360: 969–986.

Beckers JM (1991) Interferometric imaging with the very large telescope. *Journal of Optics* 22: 73–83.

Born M and Wolf E (1997) *Principles of Optics*. New York: Cambridge University Press.

Goodman JW (1996) *Introduction to Fourier Optics*. New York: McGraw-Hill Inc.

Goodman JW (2000) *Statistical Optics*. New York: Wiley-Interscience.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. New York: Cambridge University Press.

Marks DL, Stack RA, Brady DJ, Munson DC Jr and Brady RB (1999) Visible cone-beam tomography with a lensless interferometric camera. *Science* 284: 2164–2166.

Monnier JD (2003) Optical interferometry in astronomy. *Reports on Progress in Physics* 66: 789–857.

Potuluri P, Fetterman MR and Brady DJ (2001) High depth-of-field microscopic imaging using an interferometric camera. *Optics Express* 8: 624–630.

Thompson AR, Moran JM and Swenson GW (2001) *Interferometry and Synthesis in Radio Astronomy*. New York: Wiley-Interscience.

Lidar

M L Simpson and D P Hutchinson, Oak Ridge National Laboratory, Oak Ridge, TN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Light detection and ranging (LIDAR) refers to an optical system that measures any or all of a variety of target parameters including range, velocity, and chemical constituents. Usually a LIDAR system consists of a transmitting laser source, transmitter optics, receiver optics, and a detector. Several configurations of LIDAR systems are possible, based on the type of laser source and detection scheme selected. Basically, sources can be continuous wave (CW), pulsed, or chirped. Sources that are CW are most often used to measure velocity or chemical constituents. Pulsed and chirped sources

provide the additional capability of measuring range. LIDAR receivers are either incoherent (direct detection – amplitude only) or coherent (heterodyne detection – amplitude and frequency). Generally, for velocity detection, a coherent receiver with a local oscillator (LO) laser is required. A coherent receiver is also needed if a chirped laser source is used, for instance, in frequency modulation continuous wave (FM-CW) LIDAR.

For the special case of imaging LIDAR, the detector has multiple elements. In an imaging LIDAR system, the detector, or focal plane array (FPA), and the receiver optics are designed to image independent volumes in object space onto individual elements in the FPA. Each detector element, therefore, contains information for a specific volume in space and can be thought of as an independent LIDAR system. The advantage of an imaging LIDAR system is that multiple measurements can be made on extended target(s) at the same time. Thus, characterization of rapidly changing events (like wake vortices), speckle reduction by spatial averaging, synthetic aperture radar, multiple target tracking, and target identification are all possible without the sophisticated algorithms and hardware platforms required in single channel scanning systems. In addition because the multiple pixels allow parallel image capture, a true imaging LIDAR can acquire data much more rapidly than a scanning system.

In this chapter, for clarity, the emphasis will be on imaging LIDAR systems applied to hard body targets for the measurement of range and velocity. The basic configurations presented here, though, are extensible to distributed targets and measurements of wind profiles and atmospheric constituents. Also, since general textbooks are available that cover LIDAR systems and incoherent imaging systems, the concentration will be on architectures, optics, detectors, and design considerations that are specifically applicable to coherent imaging LIDAR.

Frequency Modulated Continuous Wave (FM-CW) LIDAR

One LIDAR technique, which is particularly suited for imaging due to the simplicity of the processing, is FM-CW LIDAR. FM-CW LIDAR simultaneously measures range and velocity (Doppler shift) by using a frequency chirped waveform. Referring to Figure 1, assume that a continuous wave (CW) laser frequency is linearly modulated (chirped) by a triangular wave about an offset frequency, such that the modulated signal is given as:

$$f(t) = f_o + \alpha\tau \qquad [1]$$

where f_o is the subcarrier offset frequency and α is the linear modulation rate in frequency/second, and τ is a time variable. The transmitted beam is focused by the optical system onto the target surface, reflected (or absorbed or scattered) and arrives at the receiver after τ seconds, where

$$\tau = 2\frac{R}{c} \qquad [2]$$

and R is the distance to the target surface, and c is the speed of light. The returning signal is optically mixed with the LO on the detector and produces a sum/difference beat frequency (Δf). Figure 1 shows the incident wave with its linear or chirped frequency characteristics and the delayed version returned from the target. Only the difference frequency can be optically detected since only it falls within the detector's bandwidth. Assuming the target is not moving, the result is

$$\Delta f = f(t+\tau) - f(t) = f_o + \alpha(t+\tau) - (f_o + \alpha t) = \alpha\tau \qquad [3]$$

where $\alpha = B/T$, the chirp rate of the modulator in units of Hz/s, B is the modulation bandwidth, and T is the period of the chirp. Substituting into eqn [3] for

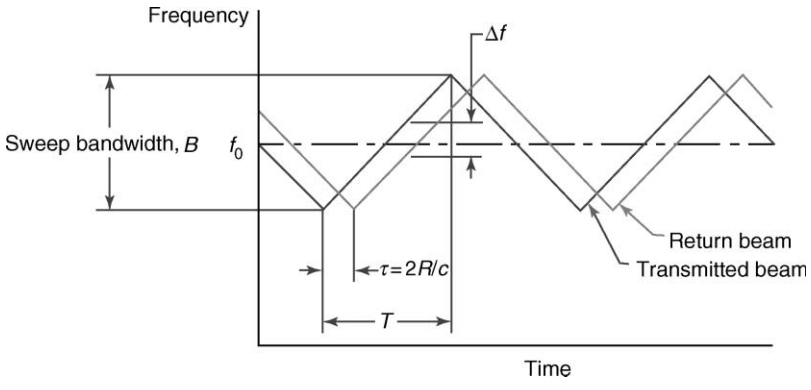


Figure 1 Triangular-wave modulation Envelope on the transmitted beam (black) and return beam (gray).

τ and α we have

$$\Delta f = \frac{B}{T}\left(2\frac{R}{c}\right) \quad [4]$$

Finally, one can solve for the distance to the target, R, as a function of the beat frequency and chirp rate:

$$R = \frac{cT}{2B}\Delta f \quad [5]$$

The difference frequency, Δf, is the same for the up and down ramp of the modulation signal. Figure 2 demonstrates the return signal for the case of a target moving toward the LIDAR system. The difference frequency now contains a component proportional to the target induced Doppler shift given by

$$f_d = 2\frac{v}{\lambda} \quad [6]$$

where v is the line-of-sight target velocity and λ is the laser wavelength. From Figure 2, the Doppler shift causes the difference frequency on the up ramp to be less than the range induced frequency shift given by eqn [5] and the net difference frequency is greater on the down slope. In this case, the difference frequency is given by

$$\Delta f = \frac{B}{T}\left(2\frac{R}{c}\right) + f_d \quad [7]$$

The range and Doppler contributions to the difference frequency may be separated by measuring the frequency difference over both the up and down chirp of the waveform. The Doppler frequency $f_d = \Delta f(\text{up}) - \Delta f(\text{down})$ and the range is given by the average of $\Delta f(\text{up})$ and $\Delta f(\text{down})$ over a complete cycle of the modulating triangular wave. Therefore, substituting in eqns [5] and [6], the range may be determined from the equation

$$R = \frac{cT}{2B}\left\{\Delta f(\text{up}) - \Delta f(\text{down})\right\} \quad [8]$$

and the line-of-sight (LOS) velocity is

$$v = \frac{\lambda}{2}\left\{\frac{\Delta f(\text{up}) + \Delta f(\text{down})}{2}\right\} \quad [9]$$

The range resolution of a FM-CW LIDAR system is the range for which the difference frequency equals 1 cycle in the sweep time interval of T or $\Delta f = 1/T$. Then, from eqn [5], the range resolution, ΔR, is given by

$$\Delta R = \frac{cT}{2B}\frac{1}{T} = \frac{c}{2B} \quad [10]$$

Therefore, the range resolution is only a function of the sweep bandwidth and not the laser wavelength. However, the Doppler shift is dependent on the laser frequency. The accuracy with which the difference frequency may be determined is a function of the number of cycles over which the measurement is made. The number of cycles is maximized for the case of $\tau = T/2$ or T should be set for twice the time delay for the maximum range of interest.

Numerical Example

A range resolution of 30 cm requires a chirp bandwidth, B, of 500 MHz. For a total range of interest of 10 km, the chirp time, T, should be 133.33 μs or the chirp rate should be 3750 Hz (1 up and 1 down ramp per cycle). The difference frequency for a 10 km range target will be 250 MHz, which demonstrates the need for wideband detectors to capture high-resolution range images (Table 1).

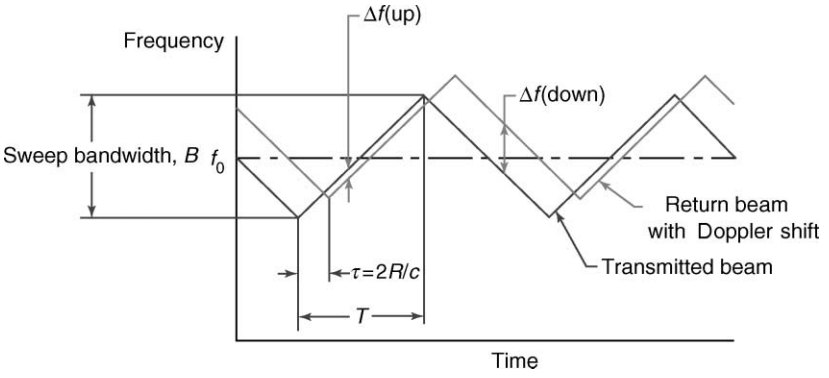


Figure 2 Lidar return with Doppler shift proportional to target velocity.

Table 1 Summary of FM-CW LIDAR characteristics

Parameter	*Value*
Range resolution (m)	0.3
Chirp rate (Hz)	3750
Chirp bandwidth (Hz)	500×10^6
Maximum range (km)	10

Optical Configuration for Imaging LIDAR

Imaging LIDAR systems can be constructed for virtually any wavelength range with appropriate sources, optical components, and detectors. The design of an optical receiver system capable of recording coherent images requires additional considerations over those necessary for a video camera or single-channel receiver. First, the collection optics must have resolution capabilities consistent with both the detector grid dimensions and the coherence properties of the source. Next, the local oscillator (LO) power must be applied to each array element. For a typical infrared (HgCdTe) detector, one to ten milliwatts of LO power might be required to reach the shot noise limit. For large arrays, this may result in unacceptable heat loads, even for momentary application. Thus, detector arrays with substantial spacing between detector elements benefit from the application of individual LO 'beamlets' to shield the insensitive portions of the array substrate. These beamlets can be generated by illuminating a mask containing an array of holes with the LO Gaussian laser beam. This scheme is illustrated in Figure 3.

In the configuration shown in the figure, the signal beam (from the target) and the LO beamlets are combined at the beamsplitter prior to being focused onto the detector plane by the objective lens. The detector plane coincides with the far-field image plane of the objective lens, and any additional optical elements utilized to enhance the camera's resolution are not shown. An expanded LO beam illuminates an aperture array located at the far-field object plane of the ancillary lens. For instance, if the objective lens and the ancillary lens are identical, the LO mask is imaged onto the detector plane with unity magnification. The detector surface is assumed to be flat with uniform video and heterodyne quantum efficiency.

Several aspects of this design are discussed below. We first discuss how the optical parameters may be adjusted to give nearly-Gaussian LO profiles at each detector position. The receiver antenna array is then shown to consist of nonoverlapping sub-beams spaced with the detector array geometry. Signal considerations for both incoherent and coherent sources are then discussed, and it is found that optimal detector size and array-spacing is dependent on the coherence properties of the intended object field.

Aperture Array

As an alternative to holographic LO beam array generation, the LO mask design presented here consists of an array of apertures with size and spacing equal to the detector dimensions. The LO is expanded to coherently illuminate the entire mask with approximately uniform intensity. The mask is located at the front focal plane of the ancillary lens and its image is located at the back focal plane (detector plane) of the objective lens. The coherent image of the mask formed at the detector plane by the ancillary-objective lens combination can be described by a two-stage Fraunhofer diffraction process.

Finite dimensions of the ancillary pupil results in spatial filtering which can affect the image profile of the LO mask. If desired, the ancillary pupil may be chosen to provide nearly-Gaussian LO beamlet profiles with beam waists at each detector location.

It is useful to first consider the coherent image of a single aperture uniformly illuminated with plane LO wavefronts. In one dimension, this wave train diffracts from the aperture to form a far-field diffraction pattern at the Fraunhofer plane of the ancillary lens given by

$$E_{\mathrm{F}}(\beta) = E_{\mathrm{F}} \frac{\sin(\beta)}{\beta} \qquad [11]$$

where $\beta = \frac{1}{2} kb \sin(\theta)$, $k = 2\pi/\lambda$, λ is the wavelength, b is the aperture width, and θ is measured from the optical axis. For small θ, $\beta = \pi b \nu/(\lambda f)$ where ν measures a point on the Fraunhofer plane and f is the focal length of the ancillary lens. The finite diameter D (determined by an aperture stop in the Fraunhofer plane) of the ancillary lens spatially filters the diffraction pattern and this limits the range of β values in eqn [11]. We define an aperture-limited pupil as one which limits β to the range $-\pi < \beta < \pi$; this will be the case when $f/D = b/2\lambda$.

Consider now a linear aperture array consisting of N individual apertures of width b and separation $a > b$. Assume that each aperture is illuminated by plane, uniform, mutually coherent LO wavefronts of equal intensity. In this case, the LO profile at the Fraunhofer plane is characterized by the diffraction pattern for a coherently illuminated linear diffraction grating:

$$E_{\mathrm{F}}(\theta) = E_0 \mathrm{sinc}\,(\beta) \frac{\sin(N\alpha)}{N \sin(\alpha)} \qquad [12]$$

where $\alpha = \frac{1}{2} ka \sin\theta$ and where β, θ, and k are defined above.

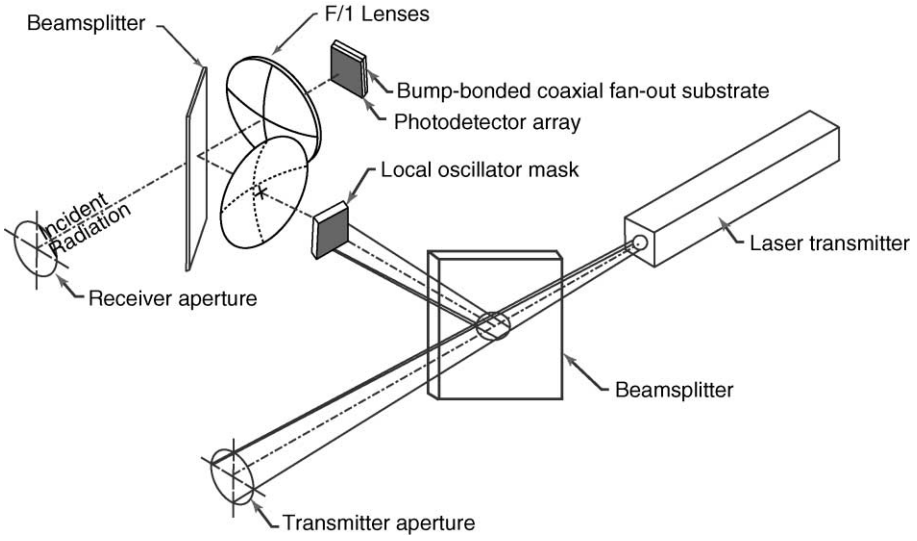


Figure 3 The individual beamlets are generated by the local oscillator or aperture mask. Reproduced with permission from Simpson ML, Bennett CA, Emery MS, *et al.* (1997) Coherent imaging with two-dimensional focal-plane arrays: design and applications. *Applied Optics* 36: 6913–6920.

Propagation of the LO from the ancillary lens Fraunhofer plane to the detector plane can be described by applying the Fraunhofer diffraction integral to eqn [12] with integration limits appropriate to the ancillary pupil dimensions. The larger pupil gives increased illumination at the detector edges while the aperture-limited pupil gives a nearly Gaussian LO profile with $1/e$ points close to the detector edges.

Receiver Antenna Array

The requirement of wavefront matching between the LO field and the signal field leads to the notion of an 'antenna beam' or equivalently, 'backward propagating LO' (BPLO) to define the field of view of an individual heterodyne receiver element. This beam is formed by allowing a reverse projection of the local oscillator field to diffract from the detector surface, through all collection optics in the signal beam path to a point on the object field of the imaging system. A receiver array will produce an array of such BPLO beams which must maintain the detector array geometry in the object field if the LO and signal beams are to mix efficiently.

As above, spatial filtering by the receiver pupil (assumed to be located at the Fraunhofer plane of the objective lens) will affect the BPLO profile. In the Fraunhofer plane of the objective lens, the uniformly illuminated single-detector BPLO profile is given by eqn [11], and we define an aperture-limited detector as one which limits the range of β to $\pm\pi$. Similarly, eqn [12] characterizes the BPLO profile resulting from a detector array uniformly illuminated with an extended uniform LO wavefront. The uniform illumination BPLO profiles in the object field differ from the detector plane profiles only in horizontal scale which now is consistent with the magnified detector dimensions.

For our purposes, we may define the far field of the objective lens according to the condition for maximum Gaussian beam collimation:

$$\omega_d \omega_0 = \frac{f\lambda}{\pi} \qquad [13]$$

where ω_0 is an effective BPLO Gaussian beam waist at the Fraunhofer plane and ω_d is an effective BPLO beam waist at the detector. Under this circumstance, the BPLO Rayleigh range is $Z_D = \pi\omega_0^2/\lambda = f(\omega_0/\omega_d)$. Identifying ω_0/ω_d as the magnification (notice that for the BPLO, the magnification is greater than unity) gives $d_o = (d_i/f)Z_D \sim Z_D$ where d_o is the object distance and d_i is the image distance. Thus the far-field of the objective lens is at least one BPLO Rayleigh range away.

The object field BPLO profiles for an aperture-limited detector array consists of nearly Gaussian sub-beams with $1/e$ points located approximately at the magnified detector edge locations. Since the far-field object distance exceeds Z_D, the divergence of each BPLO sub-beam is constant and remains distinct and nonoverlapping for all $d_o > Z_D$.

Imaging Applications Involving Incoherent Sources

For an object plane in the far field of aperture-limited collection optics, the nearly Gaussian-profile BPLO propagates from the receiver aperture according to

the condition for maximum Gaussian beam collimation. The waist is at the Fraunhofer plane where the BPLO wavefront is nearly plane, while at the object field the BPLO wavefronts are spherical with radius equal to

$$R(d_o) = d_o\left[1 + \left(\frac{Z_D}{d_o}\right)^2\right] \qquad [14]$$

where we have set $d_o - f \approx d_o$.

It will generally be the case that an imaging system will be used to analyze extended objects whose size exceeds the resolution limit. Following convention, we define a spatially coherent source as one which is illuminated with spatially coherent light with appreciable structure and roughness of depth within the coherence length of the illuminating light, and where the target motion does not appreciably perturb the speckle field during the measurement process. The speckle effects, which so seriously degrade the information content in images formed of spatially coherent objects, are not present in images formed of incoherent objects such as thermal emitters. Speckle may be reduced by time-averaging speckle fluctuations resulting from target motion, atmospheric effects, and independent transmitter pulses, and in the limit such objects become incoherent pseudothermal sources. Similar remarks apply to spatial averages over detector subarrays.

Independent radiators on the object field produce disturbances which can diffract to illuminate the entire receiver aperture as well as a finite coherence area in the image field A_{ci} given by

$$A_{ci} \gg \frac{\lambda^2}{\Omega_{ci}} \qquad [15]$$

where Ω_{ci} is the solid angle subtended by the receiver aperture at the image field. Alternatively, we may interpret A_{ci} as the area of an aperture in the object plane which, if illuminated with wavefronts characterized by eqn [14] (reciprocal illumination), would just fill the receiver aperture with a central diffraction order. Thus the combination of independent disturbances within the detector's field of view on the object plane, combine to form a spatially coherent disturbance over the receiver aperture and within A_{ci}.

According to the Van Cittert–Zernike theorem, the complex degree of spatial coherence μ_s is a unity amplitude wavetrain that propagates exactly as a signal beam resulting from the reciprocal coherent illumination of an object-field aperture of area A_{ci}. Thus, μ_s and the BPLO are Helmholtz-reciprocal pairs whose overlap determines the mixing efficiency with which heterodyne signal is generated from radiation collected from an extended incoherent source. Following convention, we refer to the overlap of μ_s and the BPLO as the effective receiver aperture.

Detectors larger than the aperture-limit, whose area exceeds A_{ci}, exhibit reduced receiver apertures and a corresponding decrease in heterodyne mixing efficiency. For example, varying the receiver pupil, while maintaining constant LO illumination, results in a heterodyne signal proportional to the received power for detector diameters much less than the aperture-limit, while detector diameters much greater than the aperture-limit produce heterodyne signal only in proportion to the square-root of the received power. Thus, for constant LO illumination and incoherent sources, increasing the entendue past the aperture limit will always increase the heterodyne signal, but less than for direct detection.

Optimal Detector Size

As noted above, an aperture-limited system is one where detectors of diameter d are matched with an objective lens with $f/D = d/2\lambda$. Commercially available detectors for use at 10 μm are commonly 100 μm wide suggesting the use of a $f/5$ lens, however aspheric lenses as fast as $f/1$ are also commercially available. Apart from aberrations, it would be advantageous to utilize a 5×5 subarray of 20 μm detectors with an $f/1$ lens in place of a single 100 μm detector for applications involving coherent sources. Applications involving incoherent thermal sources would generally favor a wide detector/fast lens combination since thermal sources provide weak heterodyne signals.

Noise Analysis and Performance

Radar Range Equation

The laser power requirements for the LIDAR system depend on the signal-to-noise ratio of the FPA, the i.f. amplifier performance, atmospheric transmission, and the optics system design. The following calculations are used to determine the performance of the LIDAR system for a single pixel of the FPA. The radar range equation is

$$P_r = P_t\left(\frac{G_t}{4\pi R^2}\right)\left(\frac{\sigma_{pt}}{4\pi R^2}\right)\left(\frac{\pi D^2}{4}\right)\eta_{atm}\eta_{sys} \qquad [16]$$

where P_r = received power; P_t = transmitted power; G_t = antenna gain; σ_{pt} = radar cross-section; R = range; D = transmitting and receiving aperture diameter; η_{atm} = atmospheric transmission; and η_{sys} = optical system efficiency.

The beam divergence of the transceiver is

$$\theta_t = \left[\left(\frac{D}{R}\right)^2 + \left(\frac{\lambda}{D}\right)^2\right]^{0.5} \quad [17]$$

where λ = wavelength and the antenna gain is

$$G_t = \frac{4\pi}{\theta_t^2} \quad [18]$$

The radar cross-section for a lambertian diffuse point target is $\sigma_{pt} = 4\rho_t dA$. The target reflectivity is ρ_t and dA is the target area, which is effectively the size of a pixel of the FPA projected onto the target. For purposes of the following calculations, a surface reflectivity of 0.1 is assumed. The pixel size at the target (assuming Gaussian beam optics) is given by

$$dA = \pi\left(\frac{D}{3}\right)^2\left[1 + \left[\frac{\lambda R}{\pi}\left(\frac{3}{D}\right)^2\right]^2\right] \quad [19]$$

where D = transmitter and receiver mirror diameter and it is assumed that the target is larger than dA.

Finally, the signal-to-noise (SNR) ratio for a coherent receiver is given by

$$SNR = \frac{\eta P_r}{h\nu B_{if}} \quad [20]$$

where η = detector quantum efficiency; P_r = received power from eqn [16]; h = Planck's constant; ν = frequency of laser; and B_{if} = post-detection bandwidth of signal processor.

Numerical example

Assuming an atmospheric transmission efficiency, η_{atm}, of 98%, an optical system efficiency, η_{sys}, of 30%, and a detector efficiency, η, of 10%, the SNR was calculated as a function of the transmitting/receiving telescope (assumed to be a mirror) diameter, D, with target range as a parameter and plotted in Figure 4. This calculation assumed

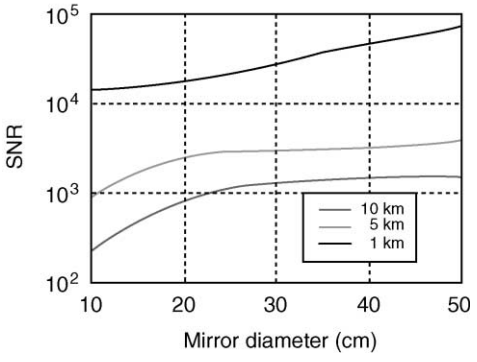


Figure 4 The SNR saturates for a mirror diameter greater than 30 cm at long range.

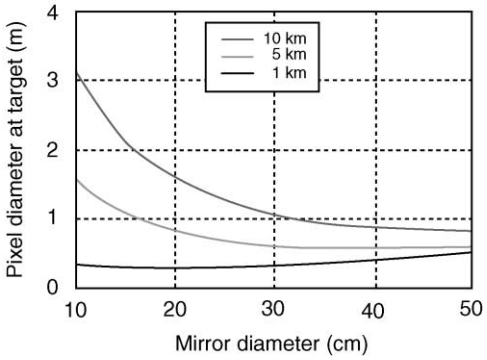


Figure 5 The pixel size determines the target cross-range image resolution.

that the pixel field-of-view (FOV) at the target was illuminated with 1 watt of power. Even at a range of 10 km, the 32 × 32 element imaging LIDAR receiver with a SNR of 100 for each pixel will require only 75 watts (approximately 74-milliwatts per pixel) of total transmitted power for a mirror diameter of 30 cm. From Figure 4, it appears that the telescope mirror size could be reduced to a diameter of 20 cm and still maintain an excellent SNR. However, the SNR is not the only consideration on the transmitting telescope diameter. The cross-range target resolution is determined by the projected pixel size at the target. This target pixel size is plotted versus the telescope mirror diameter in Figure 5. Diffraction from the transmitting and receiving optic sets the limiting resolution that can be achieved with the LIDAR system and is only a function of the diameter and wavelength. A transmit/receive telescope aperture of 30 cm seems to represent a good trade-off between cross-range resolution and SNR.

Applications and Future Directions

The remote sensing community has recognized the advantages of coherent receivers over conventional direct IR detection with respect to increased measurement sensitivity and multidimensional information capability. These advantages, however, come at a price of added system complexity, power, size, weight, and cost. In addition, choices of components, particularly in the long wavelength infrared (LWIR) region of the spectrum, have been limited to liquid nitrogen-cooled nonimaging detectors and table-top-sized gas lasers. Thus, implementations of imaging LIDAR systems have been primarily single detector, single source scanning systems which require high pulse repetition frequency lasers, complex tracking systems, and platforms, and augmentation of automated target recognition input to compensate for multiple, simultaneous measurements on the

same target. New technology developments in quantum well infrared photodetector (QWIP) detectors and arrays, and in HgCdZnTe technology are now paving the way toward thermo-electrically cooled LWIR imaging array detectors. Also waveguide, folded cavity, laser resonator designs are providing compact, lightweight packages for CO_2 lasers. Coherent receivers, that use the new focal plane arrays and compact sources, provide the capability of making parallel LIDAR measurements without the added hardware and software overhead of scanning systems. In the following section, several examples are provided that demonstrate the versatility of imaging LIDAR systems that use detector arrays for measurement applications such as chemical plumes, velocity gradients, and speckle reduction.

Wideband heterodyne receivers are useful in measuring chemical species whose absorption peaks are not well aligned with the laser source line. If the species absorption peak corresponds directly with the laser line, an absorption measurement can be made using direct detection (amplitude only). For off-line (off-resonance) measurements, heterodyning is needed, where the bandwidth of the heterodyne receiver dictates how far away from the laser line an absorption peak can be measured. For example, the CO_2 laser line spacing in the 9 to 12 micron region of the spectrum averages around 50 GHz. With the new QWIP detectors theoretically having heterodyne bandwidths of 30 GHz or more, full high-resolution spectral coverage for the long-wave IR is within grasp. For chemical detection with heterodyne receivers, there are two modes of operation. The first is termed passive heterodyne radiometry where the source is a black-body radiator and the absorption is measured for chemical species located between the blackbody and the receiver. The second mode of operation is thermoluminescence where a laser transmitter is used to excite a chemical species in the atmosphere and the resulting emission (luminescence) is measured with the heterodyne receiver. The advantage of imaging in both cases is the ability to capture chemical plumes in the presence of wind and other anomalies.

To illustrate the ability of heterodyne imaging receivers to spectroscopically resolve chemical plumes, **Figure 6** shows a series of 10×10 pixel images of a small bottle of concentrated ammonium hydroxide against a 212 °C blackbody recorded in both passive heterodyne radiometry and direct detection modes. **Figure 6a** shows a heterodyne image of the bottle with the top in place where, in **Figure 6b** the top is removed and the absorbing plume is clearly imaged. The heterodyne images in **Figures 6a** and **6b** were recorded with the LO adjusted to emit the 9R(30) line which is known to

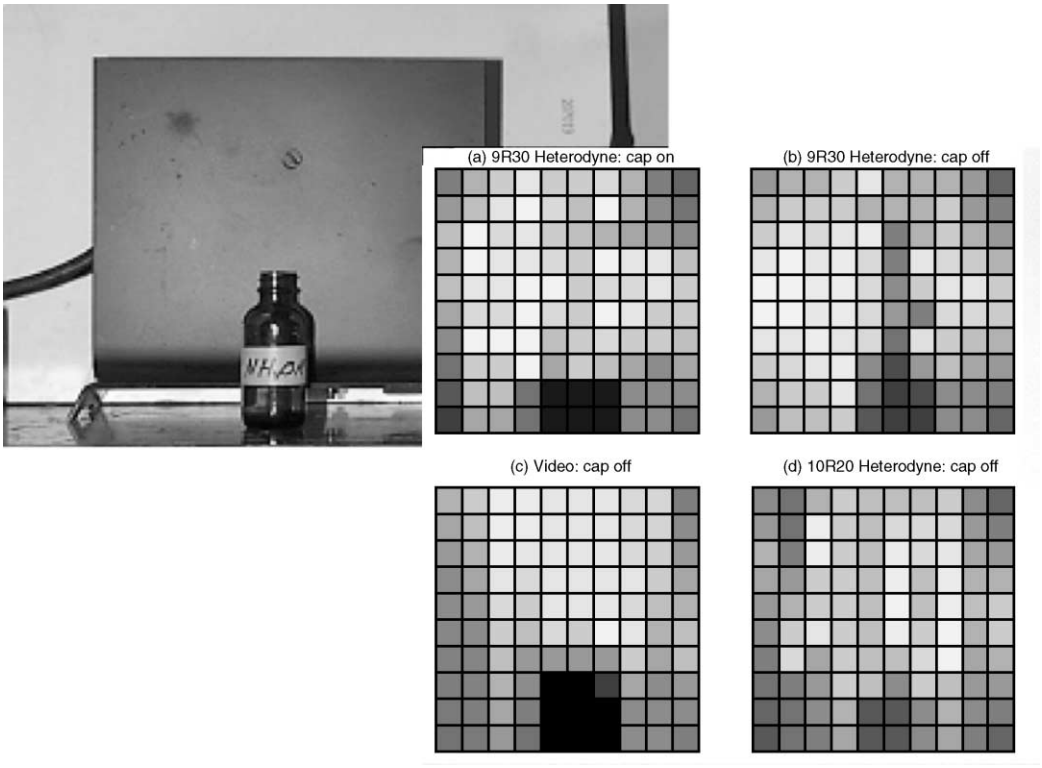


Figure 6 Ammonia plume characterization.

be absorbed strongly by NH_3. In Figure 6c, the same image was recorded by direct detection and fails to detect the ammonia plume. Similarly, in Figure 6d, the LO was adjusted to emit a nearby line known not to be absorbed by NH_3 and the plume is again not detected. Both the objective and ancillary lens were $f/1$ 38 mm, and the object field was about three meters away. The top of the bottle was approximately 2 cm in diameter and is located at the bottom of the field of view. Measurements utilizing hot ammonium hydroxide imaged against a cold background resulted in similar images (albeit at substantially reduced signal to noise ratios) where the plume was bright and the background was dark.

Another use of imaging LIDAR is the measurement of velocity profiles. Applications include wake vortex measurements for aircraft, wind shear detection, and vibration signature measurements for identification. In another experiment, active mode images were recorded by illuminating the target with a portion of the LO laser beam which had been shifted in frequency by 40 MHz using an acousto-optic modulator. The resulting Doppler image is recorded using an 8×8 HgCdTe array with discrete electronics for each pixel element in the array. Figure 7 shows the active image of a vertical squirrel cage fan where each pixel is rendered to represent the peak Doppler shift measured with the detector at the corresponding location in the image plane. The moving target scatters incoherently and thus speckle effects do not degrade the image. The measured Doppler shifts are consistent with the known rotational velocity of the target.

Speckle effects associated with coherent mixing of laser light reflected from an extended target (constructive and destructive interference) are a major source of noise in LIDAR measurements. These speckle effects can be reduced by signal averaging. For single detector LIDAR, averaging over many transmitter pulses can reduce speckle but this technique also reduces temporal resolution. The alternative considered here is to sacrifice spatial resolution by averaging over an array of detectors, all of which record the signal returned from a single transmitter pulse. Detectors separated by a coherence area diameter will record signals that approach statistical independence, and in this case speckle effects can be reduced by an amount approaching the square-root of the number of detectors in the subarray.

The following laboratory experiment demonstrates speckle reduction by spatial averaging with a fixed focal plane array. The receiver used in this experiment consists of a 3×3 array of HgCdTe detectors. The 50-micron diameter detectors are arranged in a square pattern, with 100 micron center-to-center spacing. Custom electronics multiplex the output of each detector for subsequent processing. As with the velocity profile experiment, LIDAR illumination is produced by frequency shifting a portion of the LO beam by 40 MHz with an acousto-optic modulator prior to illuminating the object field. Polarization of the illumination beam in the experiment is parallel to the LO prior to scattering from the target. The measured receiver bandwidth, limited mostly by the dewar design, is about 100 MHz, which permits the acquisition of the 40 MHz heterodyne signal. The collection optic consists of a 38 mm $f/2$ asphere which produces a 2 mm diameter pixel image at a 3 m object distance.

The target for this experiment is a 30 cm diameter disk coated with an aluminum powder. The particle size of this powder ranges from 1 to 100 microns as determined by electron microscopy. The disk is attached to a stepper motor that is incremented between frames to provide statistically independent speckle fields. Figure 8 shows the distribution of

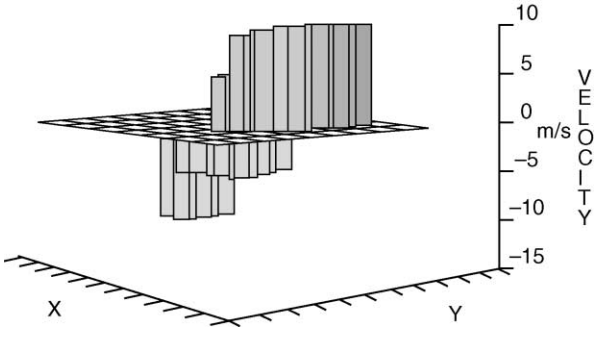


Figure 7 Doppler image of a rotating squirrel cage fan.

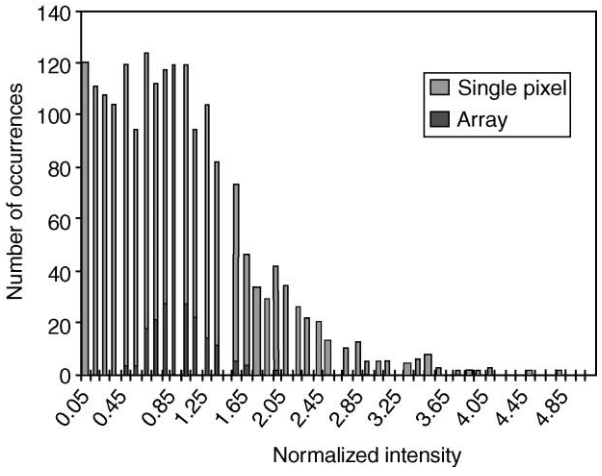


Figure 8 Speckle reduction using spatial averaging over a coherent array.

measurements taken with a single pixel superimposed on the distribution taken by averaging over the focal plane array. A reduction in speckle noise by a factor of the square root of the number of averaged pixels is obtained as predicted.

In this chapter, an overview of imaging LIDAR systems has been presented. Imaging LIDAR system design, performance, and application were discussed. With new technology developments in wide-bandwidth QWIP detector arrays, thermo-electrically cooled detectors, and compact sources, substantial potential exists for imaging LIDAR systems in the next generation in remote sensing. Due to the parallel, multifunction measurement capability of these new receivers a single sensor system can provide three-dimensional information, including range, shape, velocity, and chemical composition. In addition to the sensor fusion advantages, snapshot imaging allows the capture of high-speed, rapidly evolving events that elude present scanning systems.

See also

Environmental Measurements: Doppler lidar. **Imaging:** Infrared Imaging.

Further Reading

Born M and Wolf E (1964) *Principles of Optics*. London: Pergamon.

Forrestor AT (1956) On coherence properties of light waves. *American Journal of Physics* 24: 192–196.

Goodman JW (1975) Statistical properties of laser speckle patterns. In: Dainty JC (ed.) *Laser Speckle and Related Phenomena*, Chap. 2, pp. 9–75. New York: Springer-Verlag.

Haner AB and Isenor NR (1970) Intensity correlations from pseudothermal light sources. *American Journal of Physics* 38: 748–750.

Holzhauser E and Massig JH (1978) An analysis of optical mixing in plasma scattering experiments. *Plasma Physics* 20: 867–877.

Jelalian AV (1992) *Laser Radar Systems*, pp. 3–5. Norwood, WA: Artech House.

Kogelnik H and Li T (1966) Laser beams and resonators. *Applied Optics* 5: 1550–1566.

Kusters JF, Rye BJ and Walker AC (1989) Spatial weighting in laboratory light scattering experiments. *Applied Optics* 28: 657–664.

Martienssen W and Spiller E (1964) Coherence and fluctuations in light beams. *American Journal of Physics* 32: 919–926.

Measures RM (1984) *Laser Remote Sensing Fundamentals and Applications*. New York: John Wiley.

Rye BJ (1979) Antenna parameters for incoherent backscatter heterodyne lidar. *Applied Optics* 18: 1390–1398.

Shannon RR (1997) *The Art and Science of Optical Design*. New York: Cambridge University Press.

Siegman AE (1966) The antenna properties of optical heterodyne receivers. *Applied Optics* 5: 1588–1594.

Simpson ML, Bennett CA, Emery MS, *et al.* (1997) Coherent imaging with two-dimensional focal-plane arrays: design and applications. *Applied Optics* 36: 6913–6920.

Veldkamp WB (1983) Holographic local oscillator beam mutiplexing for array heterodyne detection. *Applied Optics* 22: 891–900.

Veldkamp WB and Van Allen EJ (1983) Binary holographic LO beam multiplexer for IR imaging detector arrays. *Applied Optics* 22: 1497–1506.

Zernike F (1938) The concept of degree of coherence and its application to optical problems. *Physica* V8: 785–795.

Multiplex Imaging

A Lacourt, Université de Franche-Comté, Besancon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The ability to project an image onto a screen in a darkened room has been known since antiquity and Leonardo da Vinci (1515) precisely described principles of the Camera oscura. When one exposes a paper sheet directly to sunlight, it is uniformly lighted. However, if one pierces a little hole through the paper sheet then an image of the sun can be displayed on a screen beyond the hole (Figure 1). Similarly, the shape of the light spot we can see in a dark room, from the rays passing through a small aperture of a closed shutter, is not that of the aperture. Generally, the spot has a circular shape, as an image of the sun. Actually, it is neither a shadow of the hole nor an image of the sun but a combination of both. Light incident on an aperture contains the entire information about the light distribution of the object, but this information has to be 'deblurred', or decoded. The aperture acts as an elementary spatial 2D-filter that selects and preserves the direction and intensity of light rays coming from each point of the scene. The projection on the screen of each pencil of light passing through the hole is a small light spot, the geometry of which is the same as the aperture, and

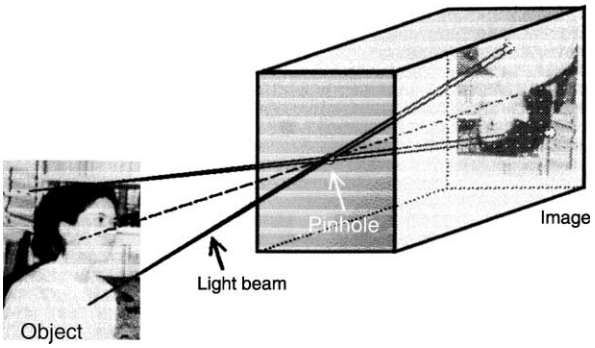


Figure 1 Principle of darkroom.

the intensity is proportional to that of the emitting point. The smaller the hole, the more detail the image exhibits, but with less luminosity. Thus, we must adopt a compromise between the resolution and the illumination of the image.

Now, if the mask has several holes, the amount of light reaching the screen, being proportional to the whole aperture area, is extensively increased. On the other hand, as each hole displays an image on the screen, the overlapping of these multiple images gives a new but blurred image. However, the information exists and we should be able to retrieve it if we know the characteristics of the complex aperture. This is the basic concept of multiplex imaging. The crucial question is: what is a good aperture design and a good way to reconstruct images having high resolution and with high illumination levels? The suitable masks with this aim are named synthetic apertures.

Porta (1589) and Kepler (1611) found the perfect solution, by setting up a lens in front of a large hole. Indeed, one can regard a lens as a wide aperture made of continuously juxtaposed microprisms (**Figure 2**). The lens obtains such refractive properties that the image displayed by each of these microprisms is moved toward a center. Moreover, as the angle of the microprisms varies continuously, it focuses on a single point, all the rays coming from a given object point. Thus, the blurring is suppressed and the illumination is increased. The way was opened toward the invention of optical instruments and photography and the 'pinhole' has fallen into oblivion for a long time.

Two new problems awake interest in synthetic apertures, around 1960. First, the domain of imaging has extended out of visible light: infrared and ultraviolet light, γ-rays, X-rays, ultrasound waves, and others. Materials suitable for making lenses or mirrors, operating with such waves, are expensive or often do not exist. For example, most of the transparent materials have a refractive index close to 1 for γ-rays and X-rays. However, some are opaque enough to enable the use of masks.

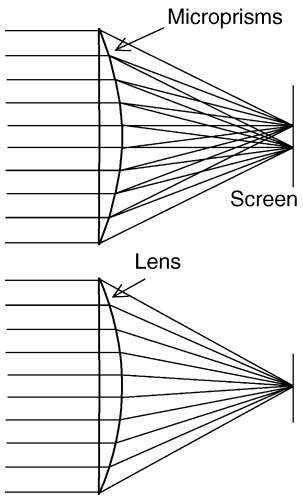


Figure 2 A lens can be modeled as a juxtaposition of microprisms with continously variable angles.

The second challenge appeared with progress in astronomy. In order to explore the universe more, further from Earth and closes to the Big Bang, astrophysicists needed telescopes providing very high luminosity and high resolution, since distant stars, nebulae, and other galaxies are extremely weak and small luminous objects. This exploration can be achieved by increasing the diameter of the primary mirror of telescopes. Indeed, due to diffraction of light by the pupil of the primary mirror, the theoretical limit of resolution of a telescope (i.e., the smaller distance between two points that one can discriminate), is inversely proportional to the diameter of the mirror. In addition, the illumination is proportional to its area. However, the working of large mirrors is a technical prowess that is very expensive and has high risks. Defect of a fraction of micrometer when polishing the surface can have dramatic consequences, e.g., Hubble. Also, such mirrors, being heavy, can bend out of shape under their own weight. Nevertheless, outstandingly large telescopes have been implemented, based on Adaptive optics: the deformations of the mirror are measured in real time by an interference method and compensated by activating a mosaic of jacks (New Technology Telescope – NTT – European program, 1989). Another way consists in replacing the solid-state mirror by a mosaic of 36 mirrors accurately directed (Keck telescope, USA). However, in either case, the mirror diameter does not exceed 8 to 10 meters. Synthetic apertures could be a solution to that difficulty, as earlier demonstrated by Fizeau (1868) and Michelson (1890).

In the first part of this article, any fundamentals are briefly recalled about properties of light and imaging.

The second part describes recent experimental methods and results obtained in synthetic aperture imaging. The last part reports the important programs in progress today relating to stellar interferometry.

Background

Electromagnetic Waves Propagation

Light power, which is involved in image formation, is carried by the electric field of an electromagnetic wave (Figure 3). The electric field $\vec{E}$ has properties of a vector orthogonal to the direction of propagation $\vec{N}$ of the light ray. It is characterized by its amplitude $\vec{E}_0$, the angular time frequency of vibration ω, and the wavevector $\vec{k}$, which specifies the propagating direction of the ray. Let $P(\vec{d})$ be a current point on a plane wavefront, defined by the position vector $\vec{d}$ from the origin of the reference frame. The electric field propagation of a monochromatic wave is expressed as:

$$\vec{E}(\vec{d},t) = \vec{E}_0 \exp[\mathrm{i}(\vec{k}\cdot\vec{d} - \omega t)] \qquad [1]$$

The light power carried by the electromagnetic field is proportional to $I = \|\vec{E}_0\|^2 = \vec{E}\cdot\vec{E}^*$ where $\vec{E}^*$ is the complex conjugate of $\vec{E}$. The time frequency $\nu = \omega/2\pi$, or the wavelength $\lambda = c/\nu$, where c is the electromagnetic wave celerity, characterize the hue of the light. The wavevector $\vec{k} = \vec{N}\omega/c$, where $\|\vec{N}\| = 1$ is an angular space frequency vector, as $\vec{n} = \vec{N}/\lambda = \vec{N}\nu/c$ is a space frequency vector. We have $\vec{k} = 2\pi\vec{n}$. As λ is a constant for a given monochromatic light, $\vec{n}$ defines the direction of propagation of a ray.

Spatially Coherent Light

Consider a very small source stated as the point source S_1, emitting uniform lightwaves in any direction (Figure 4). The wave reaching a farfield point, P, from S_1, propagates along a direction defined by the wavevector $\vec{k}_1$. The vectors $\vec{k}_1$ and $\vec{d}_1 = \overrightarrow{S_1P}$ are collinear. Another point source, S_2 is characterized from P by the wavevector $\vec{k}_2$, such that $\vec{k}_2$ and $\vec{d}_2 = \overrightarrow{S_2P}$ are collinear. When the distances d_1 and d_2 are close to each other, the amplitudes $\vec{E}_{01}$ and $\vec{E}_{02}$ of the waves at P are proportional to those of the corresponding emitting sources, S_1 and S_2. According to the principle of superposition, the resulting field at the point P, $\vec{E} = \vec{E}_1 + \vec{E}_2$, and the instant wavevector $\vec{k}$, are unique vectors.

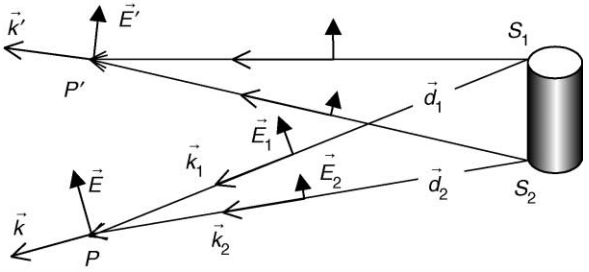


Figure 4 Spatially coherent light.

If S_1 and S_2 are two synchronous sources, the amplitude $\vec{E}_0$ of $\vec{E}$ and the wavevector $\vec{k}$ are constant versus time at any point P. This is still the case when an extended object is made of an arrangement of synchronous points. That situation defines a spatially coherent object. Thus, one cannot recover multiple data about the object's distribution, from the field framework at a single point P, since the relationship between the two is multi-unequivocal. Nevertheless, due to diffraction, one can retrieve them by knowing the complex amplitude distribution of light on a set of wavevectors available over an extensive area of a screen. Let $\vec{E}_0(\vec{r})$ be the amplitude of the electric field at the object point $M(\vec{r})$ and let $\vec{E}_0(\vec{n})$ be the one diffracted by the object to farfield toward the direction corresponding to the space frequency $\vec{n}$. Fourier transforms give the connection between $\vec{E}_0(\vec{r})$, and $\vec{E}_0(\vec{n})$:

$$\left\{\begin{array}{l} \vec{E}_0(\vec{n}) = \int \vec{E}_0(\vec{r}\,)\exp[\mathrm{i}2\pi\vec{n}\cdot\vec{r}]\mathrm{d}\vec{r}, \\ \vec{E}_0(\vec{r}) = \int \vec{E}_0(\vec{n}\,)\exp[-\mathrm{i}2\pi\vec{n}\cdot\vec{r}\,]\mathrm{d}\vec{n}. \end{array}\right\} \qquad [2]$$

This is the situation encountered when an object is illuminated by a laser or by a point source such as a single star.

Spatially Incoherent Light

When the point sources S_1 and S_2 are independent – spatially incoherent – both wavevectors $\vec{k}_1$ and $\vec{k}_2$

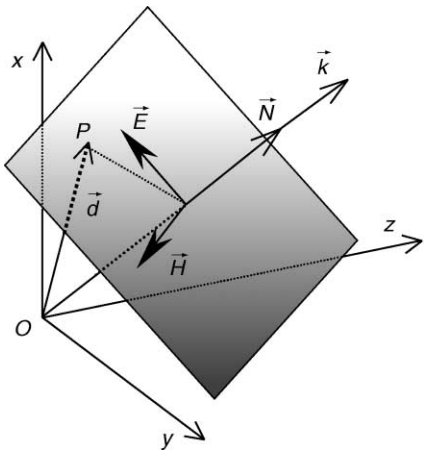


Figure 3 Electromagnetic wave propagation.

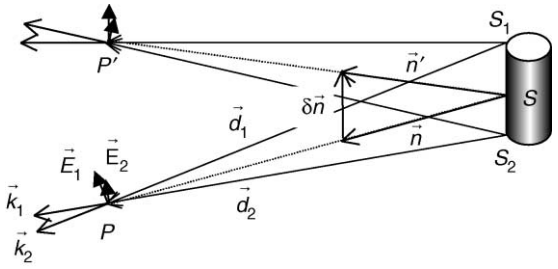


Figure 5 Spatially incoherent light.

coexist at the point P, each characterizing the angular position of S_1 and S_2 (Figure 5). If S_1 and S_2 are part of an extended intensity distribution $I_0(\vec{r})$, fans of wavevectors and fields are mixed at P, each coming from an independent point source S. Together, they carry the whole data on the object distribution. Those data are partially uncorrelated at any two points P and P' on a screen, related to the different optical paths from any point S to P and P' and to the finite time of coherence of light. The correlation between data available to farfield in two directions $\vec{n}$ and $\vec{n}'$ from the object, is given by the mutual intensity function, $\Gamma(\delta\vec{n}) = \int \vec{E}(\vec{n})\cdot\vec{E}^*(\vec{n}+\delta\vec{n})\mathrm{d}\vec{n}$, where $\delta\vec{n} = \vec{n}' - \vec{n}$. Fourier transforms give the relation between $I_0(\vec{r})$ and $\Gamma(\delta\vec{n})$:

$$\left\{\begin{array}{l} \Gamma(\delta\vec{n}) = \int I(\vec{r})\exp[\mathrm{i}2\pi\delta\vec{n}\cdot\vec{r}]\mathrm{d}\vec{r}, \\ I(\vec{r}) = \int \Gamma(\delta\vec{n})\exp[-\mathrm{i}2\pi\delta\vec{n}\cdot\vec{r}\,]\mathrm{d}\delta\vec{n} \end{array}\right\} \quad [3]$$

Therefore, by drawing a map of $\Gamma(\delta\vec{n})$, one can recover the object distribution $I_0(\vec{r})$ by a Fourier transformation. Most of time, one uses the spatial degree of coherence function, $\gamma(\delta\vec{n}) = \Gamma(\delta\vec{n})/\Gamma(\vec{0})$. When the object illuminates a screen, at a large distance d, the degree of coherence of light between two points P and P' on the screen is $\gamma(\delta\vec{\rho}/\lambda d)$, with $\delta\vec{\rho} = \overrightarrow{PP'}$. The 2D extension of $\gamma(\delta\vec{\rho}/\lambda d)$ defines the area of coherence of light on the screen. The width of the area of coherence, $\Delta\rho$, and that of the object distribution, Δr, are inversely proportional:

$$\Delta\rho\cdot\Delta r \sim \lambda d \quad [4]$$

If a mask perforated by two pinholes, separated by a distance $\delta\vec{\rho}$, replaces the screen, one can observe Young fringes of interference beyond the mask, the visibility of them being precisely $\gamma(\delta\vec{\rho}/\lambda d)$. Therefore, one can built a map of $\gamma(\delta\vec{\rho}/\lambda d)$ versus $\delta\vec{\rho}$ from a set of 2D measurements of visibilities of fringes, then reconstruct the object distribution by a Fourier transformation.

Notice that the entire data of the object distribution is available from a single coherence area. That is, a pinhole covering exactly one area of coherence would select the entire available data from the object and would display to farfield the most accurate image possible. Generally, such a pinhole is so small that practically no light passes through it. However, we can regard each area of coherence as an independent channel of information. Then an imaging device based on a juxtaposition of such reduntant channels can be considered as a multiplex imaging device.

Self-luminous objects such as extended or double stars, and artificial thermic sources are spatially incoherent sources.

Multiplex Imaging

Methods of multiplex – or lensless – imaging are performed either with coherent of incoherent light. The very different properties of the two types of light leads to completely different methods, even if all are related to holography. Holography with coherent light is an interference method for local recording of both the direction and modulus of the wavevectors and the amplitude of the field contained in many juxtaposed grains of speckle, scattered by the object. Each grain is an independent channel of imaging.

Synthetic Apertures with Spatially Incoherent Objects

As interferences are impossible by this method with incoherent light, we must improve other methods of imaging. Synthetic aperture is a concept consisting of using a redundant multichannel system to increase the luminosity of images without loss of resolution. The basic principle is that of a multiple holes darkroom. The processing is in two steps: recording and reading the image. The recording step consists in illuminating a photosensitive receptor with the object through a mask (Figure 6). Generally, the receptor is a

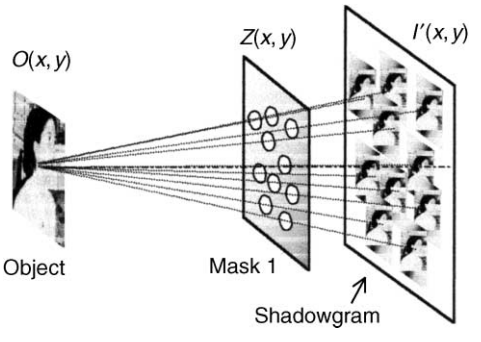


Figure 6 Principle of synthetic aperture: recording.

photographic plate. Obviously, the light distribution on the plate is blurred. Indeed, any elementary transparencies on the mask project on the receptor inverted images of the object located the shadow of the mask. Such a distribution is expressed by a convolution. Let $O(x, y)$ be the conical projection of the object on the receptor and $Z(x, y)$ that of the mask. The shadows distribution on the receptor is (the variable being normalized):

$$I'(x', y') = Z(x, y) * O(x, y) = \int Z(x, y)O(x' - x, y' - y)\mathrm{d}x\ \mathrm{d}y \quad [5]$$

The reading step is a decorrelation processing from the recorded intensity (Figure 7). It consists in achieving a new correlation of I' with an arbitrary function T:

$$I''(x, y) = T(x, y) * [Z(x, y) * O(x, y)] = [T(x, y) * Z(x, y)] * O(x, y) \quad [6]$$

$R[x, y] = [T(x, y) * Z(x, y)]$ is the system point spread function (SPSF) of the processing. It expresses the image distribution of a single-point. The perfect image is obtained when $R(x, y)$ is a 2D *Dirac* – or *delta*– distribution, $\delta(x, y)$:

$$R(x, y) = \delta(x, y), \quad \text{as} \quad \int \delta(x, y)O(x' - x, y' - y) \times \mathrm{d}x\ \mathrm{d}y = O(x', y') \quad [7]$$

Fresnel zone plates

First Mertz and Young used Fresnel zone plate (FZP) as a synthetic aperture for X-ray imaging purposes. FZPs are made of alternate transparent and opaque concentric crowns of equal area. The radius of the nth circle is $r_1\sqrt{n}$, r_1 being the smallest. When illuminated by a cylindrical light beam, the FZP is a multiple foci diffractive lens (Figure 8).

The reading step is either analogical or numerical. Through the analogical method, we consider the recorded plate as a pseudohologram (Figure 9). When the 'shadowgram' is lighted with a parallel beam of laser light, the FZPs recorded around each point concentrates the light on its foci. Then an image of the object is reconstructed on the focal planes; a method that is simple and elegant. Nevertheless, the processing is not linear since we record an intensity distribution at the recording step, while the reading step by diffraction involves amplitude. Moreover, the S/N ratio decreases as the amount of data increases, recorded on the photographic plate because of its finite dynamic.

Mertz and Young had the idea of using a second FZP to perform the deconvolution processing by optical methods. In this case, the SPSF, $R(x, y)$, is the autocorrelation function of the FZP transparency (Figure 10). It presents a narrow central peak, emerging from a pyramidal ground (Figure 11). The optical setup is the same as the recording step.

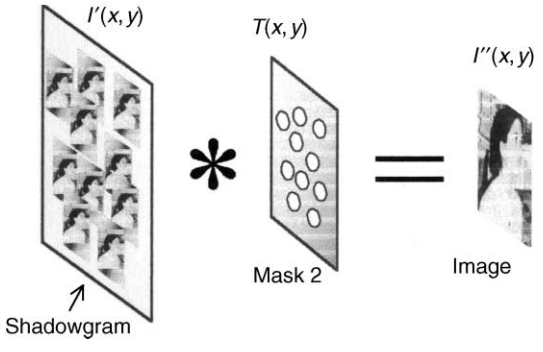


Figure 7 Synthetic aperture: reading.

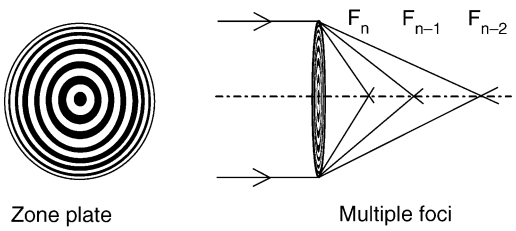


Figure 8 Properties of FZP.

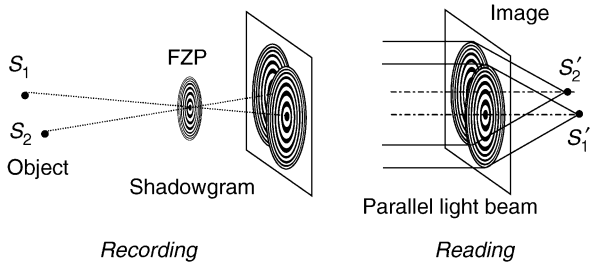


Figure 9 Recording and analogical reading with a FZP.

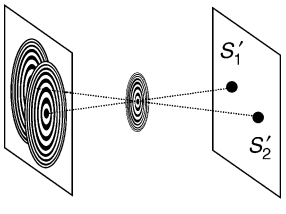


Figure 10 Optical decorrelation.

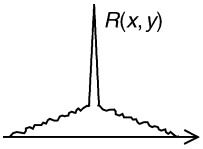


Figure 11 System point spread function with a Fresnel zone plate.

The FZP scale factor only must be adjusted to compensate for the magnification ratio generated by the conical projection at the recording step.

In either digital or analogical deconvolution, the image reconstructed is accurate, with good sharpness, if the number of zones is sufficient. Ceglio *et al.* obtained relevant images of 8 μm targets with FZP having 100 zones. Nevertheless, the multiple foci of FZP induces decreasing of the S/N ratio which is an important limitation. Besides, we pointed out above that the larger the object, the more the decrease in the S/N ratio.

Random and uniformly redundant arrays

Dicke and Ables substituted a random array to the FZP. The mask is made of randomly distributed holes and the expected advantage is doubled. First, one can reduce the transparent area of the mask by less than 50% in order to increase the S/N ratio when the object is wide. Second, the autocorrelation function of such an array is theoretically a Dirac distribution when its width is infinite. Thus, one can hope to narrow the system point response. However, when the width is finite, the pyramidal ground of the point response mentioned above is present and the advantage is not clearly demonstrated. In order to correct this effect, Chris Brown used a 'mismatched' function, for the postprocessing deconvolution. This function is calculated, first by achieving the complement of the recording mask function, then subtracting from it. Thus, the 1 transparencies result in +1 while the 0 transparencies are changed by −1. Brown shows images digitally reconstructed of X-ray sources, 30 sec. of arc wide. Fenimore and Cannon conceived a uniformly redundant array. The center random motif of the mask is half replicated on each side with only the center area being used for the recording step. The deconvolution is calculated by taking into account the whole area in a mismatched way. The result is an exact delta function. The authors show images of 200 μm-wide microspheres with a similar resolution but a signal 7100 times stronger than with a comparable pinhole camera.

Theta Rotating Interferometer

The theta shearing interferometer enables us to record complete data from the Fourier transform of a spatially incoherent object (Figure 12). The source can be any natural one, but monochromatic. This is easily practicable using a good filter or monochromator.

The device is based on a Michelson interferometer, the plane mirrors being replaced by two roof prisms. One of the roof prisms is fixed so the other can rotate around the axis of the interferometer. As a result, we have two images of the object. Let θ be the angle between the two edges of the prisms. The angle between the images is 2θ. Since the original object is noncoherent, and two homologous points being mutually coherent, the light they emit interferes, resulting in fringes.

The intensity, space frequency, and direction of the fringes characterize the intensity and the position of the point source. One shows that the farfield visibility distribution of fringes $V(\delta\vec{n})$ and the object distribution $O(\vec{r})$ are related by Fourier transforms:

$$\left\{\begin{aligned} V(\delta\vec{n}) &= \int O(\vec{r}\,)\exp\left[\mathrm{i}2\pi\frac{\delta\vec{n}\cdot\vec{r}}{2\sin\theta}\right]\frac{\mathrm{d}\vec{r}}{2\sin\theta}, \\ O(\vec{r}\,) &= \int V(\delta\vec{n})\exp\left[-\mathrm{i}2\pi\frac{\delta\vec{n}\cdot\vec{r}}{2\sin\theta}\right]\mathrm{d}\delta\vec{n} \end{aligned}\right\} \quad [8]$$

Then, one can see that the visibility of the interference pattern is an image of the degree of coherence $\gamma(\delta\vec{n})$, with a scale factor 1/2 sin θ.

Such a pattern recorded on a photographic plate is a Fourier hologram of the noncoherent object. An image can be reconstructed by diffraction at infinity or near a focus (see Figure 13, examples performed in 1972 by the author). The technique was successfully applied to the determination of the modulation transfer function of optical systems. C Roddier and F Roddier applied the method to astronomy.

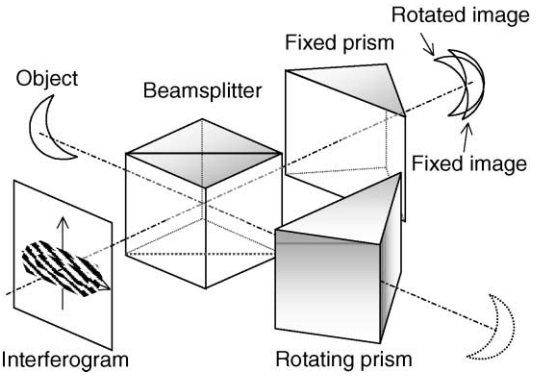


Figure 12 Theta shearing interferometer.

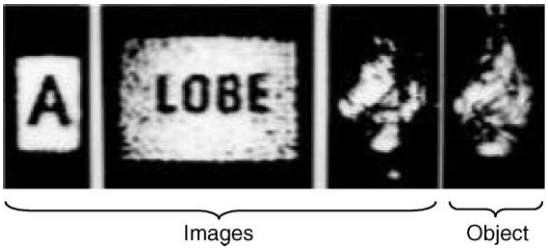


Figure 13 Incoherent light holograms.

For instance, they examined Alpha Orionis and discovered the evolution of its dust envelope and the possible existence of a companion.

High Resolution Imaging in Astronomy

In astronomy, the limit of resolution depends chiefly on two factors. The first is a technical one. Adaptive optics and very large telescopes (VLT) brought spectacular progress to clarity and resolution of telescopes. The 8.2 m VLT of Kueyen set up in 1989, shows images easily resolved of HIC 69495 Centauri, separated by 0.12 sec. of arc. Nevertheless, working of large mirrors for telescopes reaches its limits. Therefore, their clarity remains too weak. On the other hand, atmospheric turbulences make the refractive index of air unstable and the position of the images in the focal plane of telescopes unsettled. Thus, the point response is a granular spot named 'speckle'. Short exposure times relieve that inconvenience, but they prohibit the observation of weak stars. In 1970, Labeyrie proposed a method of speckle interferometry consisting of a statistical analysis of a set of speckles. The result is a display of a diffraction-limited image. He resolved many tens of single and double stars. Only Hubble, above the atmosphere, is able to do it by direct imaging. However, the primary mirror of Hubble being only 2.40 m wide, has clarity much lower. Stellar interferometry is a very promising alternative solution.

Principles of Stellar Interferometry

Stellar interferometry is based on eqn [3]. The light emitted by double or wide stars is noncoherent, so that one can access an image of them trough 2D measurements of the degree of coherence of light. We showed above that one could achieve this by using interference devices such as Young holes. Eqn [3] shows that if we measure the visibility of Young fringes at increasing distances $\delta\rho$ between the pinholes, the distance $\Delta\rho$ from which the visibility is zero gives the angular width of the star, $\Delta\alpha$, under the formula:

$$\Delta\rho \cdot \Delta\alpha \sim \lambda \qquad [9]$$

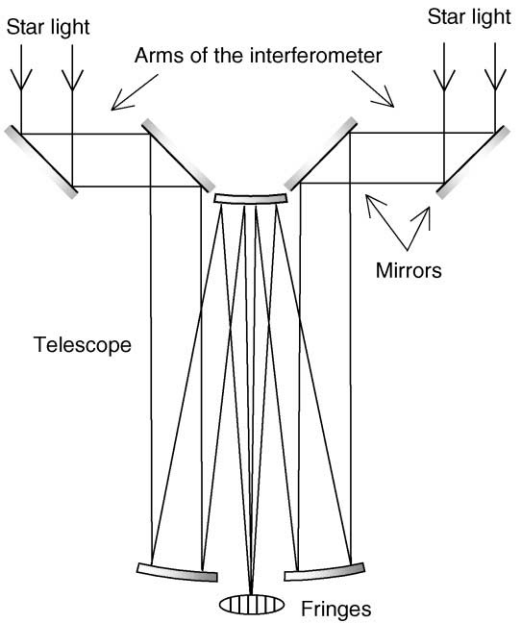


Figure 14 Stellar interferometer of Michelson.

where $\Delta\alpha = \Delta r/d$. In the case of a double star, the visibility varies periodically versus $\delta\rho$ when the pinholes are aligned along the axis of two stars. Let $\Delta\alpha$ be the angular distance between the stars: the smallest distance $\Delta\rho$ for which the visibility is zero is given by:

$$\Delta\rho \cdot \Delta\alpha = \lambda/2 \qquad [10]$$

The larger the uttermost distance between the holes, the more the limit of resolution of the processing is potentially sharp.

In order to exploit this idea, Stephan placed and rotated a mask with two holes in front of the objective of a refractive telescope. He established that the diameter of the stars he observed was much less than 0.16 sec. of arc. Then, Michelson built up, on top of a 30 cm refractive telescope, a stellar interferometer consisting of two arms catching the light at 7 m of distance from each other (Figure 14). He measured correctly the diameter of the satellites of Jupiter. Hale carried the distance of the base up to 15 m in 1920.

The Two Telescopes Interferometer of Labeyrie

In 1975, Labeyrie improved the idea by setting up two separate 25 cm diameter Cassegrain-coudé telescopes on a 12 m baseline, oriented north–south (Figure 15). The setting and mechanical stability of the device are extremely sensitive. The tolerances do not exceed any μm. A mobile table near the common focal plane of the two telescopes compensates for the optical path differences due to stars tracking. In spite

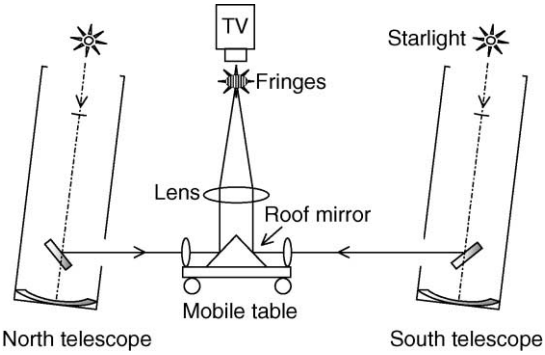


Figure 15 The two telescopes interferometer of Labeyrie.

of the acute difficulties, he succeeded in observing fringes on Vega. He obtained a resolution of 0.01 sec. of arc versus 0.5 sec. with a single 25 cm telescope. Later, one telescope was rendered mobile bringing the base width variable between 4 and 67 m. The resolution reaches up to 0.00015 sec. of arc. Note that the rotation of Earth also enables east–west analysis with resolution of 0.005 sec., and today, the 25 cm telescopes are replaced by 1.5 m ones.

The Very Large Telescope Interferometer (VLTI)

The two telescopes interferometer of Labeyrie is a first step toward a very ambitious project: the telescope at synthetic aperture and very large base. Proofs of reliability of such a principle was already given by the very large array (VLA) experimented in radio astronomy. Labeyrie with Lena, and the team of the ESO (Ecole Supérieure d'Optique, Paris), imagined a very large telescope interferometer (VLTI) made of 6 telescopes of 2 m, in a 2D structure on a base of 100 m. The European VLTI was built at Cerro Paranal. It is made of four 8.2 m telescopes corrected in real time by adaptive optics (NTT technology) and three 1.8 m telescopes. The fourth 8.2 m telescope was put in service in 2001. Some days, each of the 8.2 m telescopes used separately gives a resolution close to that of Hubble. However, the clarity is much better because of the larger primary mirror diameter (8.2 m against 2.4 m). In May 2003, the team of J-G Cuby observed a galaxy the farthest away ever seen, only 900 millions years old since the Big Bang. Obviously, a drastic gain of resolution (by a factor 10) is expected from the interference configuration.

See also

Coherent Lightwave Systems. **Diffraction:** Fresnel Diffraction. **Imaging:** Adaptive Optics; Interferometric imaging. **Instrumentation:** Telescopes. **Interferometry:** Overview.

Further Reading

Labeyrie A (1976) High resolution techniques in optical astronomy. In: Wolf E (ed.) Progress in Optics, vol. XIV, pp. 49–85. New York: North-Holland.

Mertz L (1965) *Transformations in Optics*. Amsterdam, Oxford: John Wiley & Sons, Inc.

Photon Density Wave Imaging

V Toronov, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Photon density wave imaging (PDWI) employs diffused light, usually in the near-infrared band, for imaging inhomogeneous structures in a highly scattering medium. The other names for PDWI are diffuse optical tomography and photon migration imaging. The term photon density wave (PDW) refers to the wave of light density spreading in the highly scattering medium from the source whose power changes in time. Therefore, rigorously speaking, photon density wave imaging uses intensity-modulated light sources, such as pulsed lasers, or sources whose power is harmonically modulated. However, many applications employ steady-state light sources. Most applications of PDWI are in the field of imaging of biological tissues, including medicine and cognitive neuroscience.

As in the other imaging methods, such as x-ray, computed tomography (CT) or ultrasound imaging, the problem of PDWI consists in the reconstruction of the internal structure of the medium from the measurements made outside the medium (on its surface). In the case of PDWI the internal structure of the medium is represented by the spatial distribution of its optical properties. The reconstruction procedure is essentially an iterative mathematical algorithm, which includes the modeling of the light propagation inside the medium assuming a certain distribution of the optical properties, and the

correction of this distribution to match the model prediction with the actual measurement. The modeling part of the imaging problem is usually referred to as the forward problem, and the correction procedure is known as an inverse problem.

Forward Problem: Photon Density Waves

Unlike optically transparent media, in a highly scattering medium the coherence of light from the coherent source very rapidly decays with distance. This happens due to the interference of many randomly and multiply scattered light waves at every point of the medium that is distanced from the source much farther than the photon mean free path l, which is the mean distance at which a light wave undergoes a single scattering event. For biological tissues, where light scattering occurs on cell membranes and other heterogeneities of the index of refraction, l is typically about 1 mm. Although high scattering eliminates the coherence of the electromagnetic waves at the short time intervals corresponding to oscillation at optical frequencies ($\sim 10^{-15}$ s), in the case of an intensity-modulated light source, the macroscopic wave of the density of light energy, called 'photon density wave', forms in the highly scattering medium. The properties of PDW depend on the type of the source modulation and on the local optical properties of the medium, which are the scattering coefficient μ_s [cm^{-1}] and the absorption coefficient μ_a [cm^{-1}].

Because of the complexity and incoherence of the macroscopic electromagnetic field in highly scattering media, i.e. media with $\mu_a \ll \mu_s$, the fundamental Maxwell equations of electrodynamics do not provide a convenient approach to the solution of the problem of light propagation in such media at distances much larger than l. Therefore, statistical approaches have been developed, such as the photon transport model and its diffusion approximation. In the photon transport model the light is characterized by the radiance $L(\mathbf{r}, \hat{\Omega}, t)$ [W cm^{-2} steradian^{-1}] (where the steradian is the unit solid angle), which is the power of radiation near the point $\mathbf{r}$ propagating within a small solid angle around the direction $\hat{\Omega}$ at time t. The photon transport model accounts for the possible nonisotropy of the radiation flux. For the given spatio-temporal distribution of the optical properties, μ_a and μ_s, and for the given spatial and angular distribution of light source power $S(\mathbf{r}, \hat{\Omega}, t)$ [W cm^{-3} steradian^{-1}] the radiance $L(\mathbf{r}, \hat{\Omega}, t)$ can be found from the linear radiation transfer equation, also called the photon transport equation, or simply the transport equation (TE):

$$\frac{1}{v}\frac{\partial L(\mathbf{r}, \hat{\Omega}, t)}{\partial t} + \nabla \cdot L(\mathbf{r}, \hat{\Omega}, t)\hat{\Omega} + (\mu_s + \mu_a)L(\mathbf{r}, \hat{\Omega}, t) = \mu_s \int L(\mathbf{r}, \hat{\Omega}', t) f(\hat{\Omega}, \hat{\Omega}') d\hat{\Omega}' + S(\mathbf{r}, \hat{\Omega}, t) \quad [1]$$

where $f(\hat{\Omega}, \hat{\Omega}')$ is the phase function, which represents the portion of light energy scattered from the direction $\hat{\Omega}$ to the direction $\hat{\Omega}'$, and v is the speed of light in the medium. TE is equivalent to the Boltzmann transport equation describing the transport of particles undergoing scattering.

Equation [1] is difficult to solve even numerically. Therefore, a hierarchy of approximations to the PTM was developed. This hierarchy, known as the P_N approximation, is based on the expansion of the radiance in a series of spherical harmonics $Y_{nm}(\hat{\Omega})$ truncated at $n = N$. In the P_1 approximation the coefficients of the series are proportional to the radiation fluence

$$\Phi(\mathbf{r}, t) = \int L(\mathbf{r}, \hat{\Omega}, t) d\hat{\Omega}, \quad [2]$$

and the flux

$$J(\mathbf{r}, t) = \int L(\mathbf{r}, \hat{\Omega}, t)\hat{\Omega} d\hat{\Omega}, \quad [3]$$

both having units of [W cm^{-2}]. By substituting the P_1 series for $L(\mathbf{r}, \hat{\Omega}, t)$ and $S(\mathbf{r}, \hat{\Omega}, t)$ into eqn [1] and assuming isotropic light sources and sufficiently slow change of source intensity, one can obtain the equation that includes only $\Phi(\mathbf{r}, t)$ as the unknown. Usually, this equation

$$v\mu_a U(\mathbf{r}, t) + \frac{\partial U(\mathbf{r}, t)}{\partial t} - D\Delta U(\mathbf{r}, t) = Q(\mathbf{r}, t) \quad [4]$$

is written in terms of the value $U(\mathbf{r}, t) = \frac{1}{v}\Phi(\mathbf{r}, t)$ [J cm^{-3}], which has units of energy density, and is called the photon density. In eqn [4] $D = v/(3\mu_a + 3\mu_s')$ [cm^2/s] is the light diffusion coefficient, and $\mu_s' = \mu_s(1 - g_1)$ [cm^{-1}] is the reduced scattering coefficient; g_1 is the average cosine of the scattering angle. The source term $Q(\mathbf{r}, t)$ [J cm^{-3} s^{-1}] in the right-hand side of the equation [4] describes the spatial distribution of light sources (in the isotropic approximation, i.e., neglecting the dipole and higher momenta of $S(\mathbf{r}, \hat{\Omega}, t)$) and the temporal modulation of source intensity.

Equation [4] is mathematically equivalent to the equation describing the density of the medium consisting of the particles undergoing linear diffusion. Therefore, it is called the diffusion equation (DE), and the P_1 approximation is also known as the diffusion

approximation. The diffusion approximation is valid under the following three conditions: (a) the light sources are isotropic or the distance between the source and the detector is much larger than l; (b) characteristic source modulation frequencies ν are low enough that $\nu\mu_s'/(2\pi\nu) \gg 1$; and (c) $\mu_a \ll \mu_s'$. In the biomedical applications of PDWI the condition (a) is fulfilled when the source–detector distance is larger than 1 cm. The condition (b) is valid when the characteristic modulation frequencies of the light source are less than 1 GHz. The condition (c) is fulfilled in most biological tissues, for which the scattering is highly isotropic, i.e. $g_1 \ll 1$. Thus, the forward problem of PDWI, i.e., the modeling of light propagation, can be solved using the diffusion approximation in a large variety of situations.

PDWs described by equation [4] are the scalar damped traveling waves, characterized by the coherent changes of light energy (photon density) in different locations on time intervals from microseconds to nanoseconds. Usually pulsed or harmonically modulated light sources are used to generate PDWs. The properties of the PDWs from a harmonic source are fundamental, because the PDW from the source of any type of modulation can be considered as a superposition of harmonic waves from sources having different frequencies using the Fourier transform.

In order to get an idea of the basic properties of PDW, one can consider the analytical solution of [4] for the case of a harmonically modulated point-like source placed in an infinite homogeneous scattering medium. This solution is

$$U(r,t) = U_{dc}(r) + U_{ac}(r)\exp[-i\omega t + i\phi(r)] \qquad [5]$$

where U_{ac}, U_{dc}, and ϕ are the ac, dc and phase, respectively, of the wave measured at the distance r from the source, and $\omega = 2\pi\nu$. U_{ac}, U_{dc}, and ϕ depend on the medium optical properties, source intensity Q_0, modulation amplitude A_0, and initial phase ψ_0 as

$$U_{ac}(r) = \frac{Q_0 A_0}{4\pi D} \times \frac{\exp\left[-r\left(\frac{\nu\mu_a}{2D}\right)^{1/2}((1+x^2)^{1/2}+1)^{1/2}\right]}{r} \qquad [6a]$$

$$U_{dc}(r) = \frac{Q_0}{4\pi D}\,\frac{\exp\left[-r\left(\frac{\nu\mu_a}{D}\right)^{1/2}\right]}{r} \qquad [6b]$$

$$\phi(r) = r\left(\frac{\nu\mu_a}{2D}\right)^{1/2}\sqrt{(1+x^2)^{1/2}-1} \qquad [6c]$$

where $x = \omega/(\nu\mu_a)$. Equations [5]–[6] describe a spherical wave whose wavelength at $\mu_a = 0.1\ \mathrm{cm}^{-1}$ and $\mu_s' = 8\ \mathrm{cm}^{-1}$ (typical for biological tissues) and at modulation frequency $\nu = 100$ MHz, is close to 35 cm, so that the phase of the wave changes approximately by 10° per cm. In the same time, the wave decays exponentially at a distance of about 3 cm, so that PDWs are strongly damped. A similar analytical solution in the form of a spherical wave can be obtained for the case of a semi-infinite homogeneous medium and point source placed on the surface of the medium (see Figure 1). These analytical solutions are employed to measure μ_a and μ_s' of the homogeneous medium. The method is based on fitting measured ac, dc, and phase of the PDW as functions of source–detector distance by the analytical solution.

Inhomogeneities in the optical properties of a turbid medium cause distortion of the PDW from sphericity (see Figure 1). PDWs in homogeneous media of limited size or with strong surface curvature also usually have complex structure. For a limited number of configurations one can obtain solutions to the DE using analytical methods of mathematical physics, such as the method of Green's functions. However, in practice the complexity of the inhomogeneities and surfaces usually requires numerical solution of the DE. This can be achieved using numerical methods for the solution of partial differential equations, such as the finite element method and the finite difference method. The finite difference method can also be used to solve the TE. This may be necessary in situations when the DE is not valid, for example to accurately account for the influence of the cerebrospinal fluid (which has relatively low scattering) on light transport in the human skull. However, the solution of the TE requires formidable computational resources, which increase proportionally to the volume of the medium. Therefore, hybrid DE-TE methods were developed, in

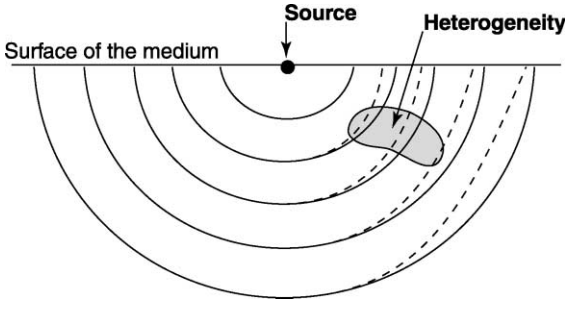


Figure 1 Equiphase surfaces of the PDW induced by a point source in a semi-infinite homogeneous medium (solid lines) and in a medium including heterogeneity (dashed lines).

which DE is used to describe light propagation in most of the medium except small low-scattering areas described by the TE.

One should note that although the term 'photon' suggests quantum properties of light, no consistent application of quantum theory to problems related to PDWI has been developed so far. Also, it should be noted that the notion of a photon as a particle of light having certain energy, momentum, and localization at the same time is not completely consistent with quantum electrodynamics. However, since the theoretical models of light transport in a scattering medium (such as TE and DE) are mathematically equivalent to the models describing systems of randomly moving particles, the notion of a photon as a 'particle of light' may be accepted as adequate and is widely used in the limits of PDWI.

Inverse Problem and Imaging Algorithm

Mathematically the forward problem can be expressed in the form of a general nonlinear operator F:

$$G = F[\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}] \qquad [7]$$

where G is the set of measured values (ac, dc, and phase), $\mathbf{r}$ is the coordinate vector of a medium element (voxel), and $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}$ denotes the set of tissue optical properties corresponding to whole collection of voxels. The problem of finding the spatial distribution of medium optical properties can be formalized as

$$\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\} = F^{-1}[G] \qquad [8]$$

Neither the transport model nor the diffusion model allow inversion in a closed analytical form for a general boundary configuration and spatial distribution of optical properties. The basic reason for the complexity of the inverse problem is in the complexity of the light paths in the highly scattering medium. This situation is opposite to that in CT, where the x-ray radiation propagates through the medium along straight lines almost without scattering, which significantly simplifies the forward problem and allows its analytical inversion. From the mathematical point of view, the complexity of the inverse problem of PDWI is due to its nonlinearity and ill-posedness. Only approximate methods of inversion have been developed so far.

The simplest method is called backprojection by analogy with the method of the same name used in CT. The method attempts to reconstruct the actual spatial distribution of the optical properties μ_a and/or μ_s' from the 'measured' bulk values μ_a^m and $\mu_s'^m$ obtained at a variety of locations assuming homogeneity of the medium. The measured values are considered as a result of the volumetric 'projection'

$$\{\mu_a^m(\mathbf{r}), \mu_s'^m(\mathbf{r})\} = \int \{\mu_a(\mathbf{r}'), \mu_s'(\mathbf{r}')\} h(\mathbf{r} - \mathbf{r}') d^3\mathbf{r}' \qquad [9]$$

of the actual distribution, where $\mathbf{r}$ and $\mathbf{r}'$ are the locations of the measurement and reconstruction, respectively. A weight function $h(\mathbf{r})$ accounts for the contribution of the particular region to a measured value and should be derived from the forward model. Different versions of the backprojection method use different weight functions $h(\mathbf{r})$. In the backprojection method the values of μ_a and/or μ_s' are obtained by inverting the convolution [9]. In practice the convolution integral is usually estimated by a sum over a limited number of voxels for a limited number of measurements. Therefore, the inversion of the convolution reduces to the solution of the set of algebraic equations corresponding to [9]. The advantage of the method is high speed, which allows real-time imaging of the dynamic processes, for example, functional hemodynamic changes in the brain. However, the backprojection method provides a phenomenological rather than a self-consistent inversion of the forward problem, and its quantitative accuracy is low.

A more consistent method is based on the perturbation approach to the solution of the inverse problem [8]. If we have an estimate $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$ that is close to the ideal solution, then $G_e = F[\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e]$ (see eqn [7]) is close to S. Then the difference $G - G_e$ can be written as a Taylor series:

$$G - G_e = J[\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e]\{\Delta\mu_a(\mathbf{r}), \Delta\mu_s'(\mathbf{r})\} + \ldots \qquad [10]$$

where $J[\ldots]$ is the Jacobian, and $\{\Delta\mu_a(\mathbf{r}), \Delta\mu_s'(\mathbf{r})\} = \{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\} - \{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$. Neglecting terms after the first one in [10], the problem of finding the correction $\{\Delta\mu_a(\mathbf{r}), \Delta\mu_s'(\mathbf{r})\}$ to the estimate $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$ reduces to the calculation of the Jacobian and to the solution of a linear algebraic equation.

This approach provides a basis for the PDWI algorithm, which is the following. First one makes an initial guess at the optical properties of the medium. Then one calculates the ac, dc, and phase of the PDW using a model of light propagation through the tissue and the geometry of the source–detector configuration. The calculated values are compared with the measured ones. If the error is above a set tolerance level, one calculates corrections to the current values of μ_a and μ_s' by solving eqn [10]. The algorithm is completed when the required error level is achieved. Thus, the PDWI algorithm iteratively applies both

the forward and inverse problems. Since at each of many iterations one should solve the forward problem, which in the simplest case is a partial differential equation, the algorithm typically requires considerable computational resources.

One should note that the success of the method is largely dependent on the closeness of the initial guess to the actual values of the optical properties. In many studies the 'first guess' of the tissue optical properties was obtained by measuring the bulk absorption and reduced scattering coefficients. However, because of the ill-posedness of the inverse problem, homogeneous initial conditions did not always lead to the correct reconstruction. A more productive approach to the problem of the initial conditions is to assign approximate initial values of the optical properties known for example from *in vitro* measurements to different structures revealed by nonoptical imaging techniques, such as magnetic resonance imaging (MRI) or CT.

Applications

PDWI was proposed and developed as a noninvasive biomedical imaging technique. Although the spatial resolution of PDWI cannot compete with that of CT or MRI, PDWI has the advantages of low cost, high temporal resolution, and high biochemical specificity. The latter is based on the optical spectroscopic approach. Biological tissues are relatively transparent to light in the near-infrared band between 700 and 1000 nm (a near-infrared window). The major biological chromophores in this window are oxy- and deoxyhemoglobin, and water. One can separate contributions of these chromophores into light absorption by employing light sources at different wavelengths. Multi-wavelength PDWI can also be used to measure the redox state of the cytochrome oxidase, which is the terminal electron acceptor of the mitochondrial electron transport chain and responsible for 90% of cellular oxygen consumption.

One of the hot topics in cognitive neuroscience is the hemodynamic mechanism of the oxygen supply to neurons and removal of products of cerebral metabolism. High temporal resolution and biochemical specificity make PDWI a unique noninvasive tool for studies of functional cerebral hemodynamics, although light can penetrate no deeper than the very surface of the adult brain. Figure 2 displays a typical arrangement of light sources and detectors used in PDWI of functional cerebral hemodynamics, and the images of the activated area in the brain.

Because of its low cost, noninvasiveness, and compact size of the optical sensors, PDWI is also considered to be a promising tool for mammography and structural imaging of anomalies in the brain. PDWI is an especially promising tool for neonatology, because the infant tissues have low absorption.

PDWI instruments employ lasers, light-emitting diodes and gas-discharge tubes as sources of light. Among the detectors there are photomultiplier

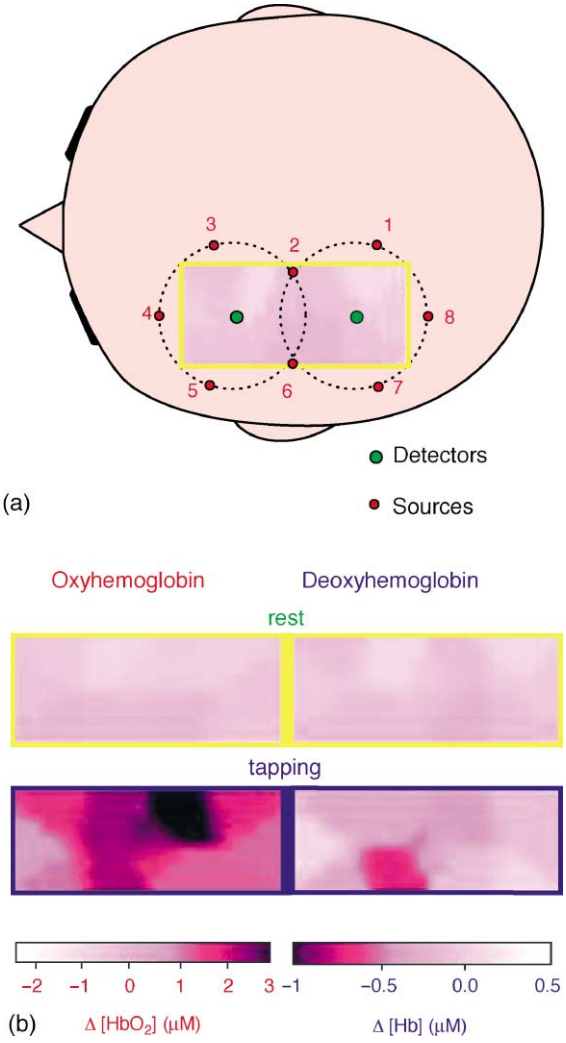


Figure 2 (a) Geometrical arrangement of the 16 sources and two detectors on the head of a human subject. Each one of the eight small red circles (numbered 1–8) represents one pair of source fibers at 758 and 830 nm. Two larger green circles represent detector fiber bundles. The rectangle inside the head outline is the 4×9 cm^2 imaged area. (b) Maps of oxyhemoglobin ([HbO_2], left panels) and deoxyhemoglobin ([Hb] right panels) concentration changes at rest (top) and at the time of maximum response during the third tapping period (bottom). The measurement protocol involved hand-tapping with the right hand (contralateral to the imaged brain area). Adapted with permission from Franceschini MA, Toronov V, Filiaci M, Gratton E, Fantini S (2000) On-line optical imaging of the human brain with 160-ms temporal resolution. *Optics Express* **6**(3), 49–57 (Figures 1 and 4).

tubes, avalanche photodiodes, and CCD cameras. Both sources and detectors can be coupled with the tissue by means of optical fibers. A number of commercially produced near-infrared instruments for PDWI are available on the market. According to the type of source modulation, there are three groups of instruments: frequency domain (harmonically modulated sources), time domain (pulsed sources), and steady state. Steady-state instruments are simpler and cheaper than frequency-domain and time-domain ones, but their capabilities are much lower. Time-domain instruments can theoretically provide a maximum of information about the tissue structure, but technically it is difficult to build a time-domain instrument with high temporal resolution and high signal-to-noise ratio. Frequency-domain instruments provide temporal resolution up to several milliseconds and have high quantitative accuracy in the determination of optical properties.

Nomenclature

Absorption coefficient, μ_a [cm^{-1}]
Alternating part of PDW, alternating current (ac)
Average cosine of the scattering angle, g_1
Circular modulation frequency, ω [s^{-1}]
Computed tomography (CT)
Difference between the estimate and actual optical properties, $\{\Delta\mu_a(\mathbf{r}), \Delta\mu_s'(\mathbf{r})\} = \{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\} - \{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$ [cm^{-1}]
Diffusion coefficient, $D = v/(3\mu_a + 3\mu_s')$ [cm^2/s]
Diffusion equation (DE)
Distribution of light source power $S(\mathbf{r}, \hat{\Omega}, t)$ [W cm^{-3} $steradian^{-1}$]
Divergence operator, ∇ [cm^{-1}]
Estimate of optical properties, $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$ [cm^{-1}]
General nonlinear operator, F
Jacobian, J
Laplace operator, Δ [cm^{-2}]
Magnetic resonance imaging (MRI)
Measured bulk optical properties (absorption and reduced scattering coefficients) $\{\mu_a^m \mu_s'^m\}$ [cm^{-1}]
Modulation frequency, ν [s^{-1}]
Non-alternating part of PDW, direct current (dc)
Optical properties of the medium (absorption and reduced scattering coefficients) as functions of spatial coordinate, $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}$ [cm^{-1}]
Phase as a function of source–detector distance $\phi(r)$
Phase function, $f(\hat{\Omega}, \hat{\Omega}')$
Photon density, $U(\mathbf{r}, t)$ [J cm^{-3}]
Photon density wave imaging (PDWI)
Photon density wave (PDW)
Photon mean free path, l [cm]
Radiance, $L(\mathbf{r}, \hat{\Omega}, t)$ [W cm^{-2} $steradian^{-1}$]
Radiation fluence, $\Phi(\mathbf{r}, t)$ [W cm^{-2}]
Radiation flux, $J(\mathbf{r}, t)$ [W cm^{-2}]
Reduced scattering coefficient, μ_s' [cm^{-1}]
Scattering coefficient, μ_s [cm^{-1}]
Set of measured values (ac, dc, and phase), G
Set of predicted values (ac, dc, and phase), G_e
Source term $Q(r, t)$ [J cm^{-3} s^{-1}]
Source–detector distance, r [cm]
Spatial coordinate, $\mathbf{r}$ [cm]
Speed of light in medium, v [cm/s]
Spherical harmonics $Y_{nm}(\hat{\Omega})$
Transport equation (TE)
Weight function, $h(\mathbf{r})$
$x = \omega/(v\mu_a)$

See also

Coherence: Coherence and Imaging. **Fiber and Guided Wave Optics:** Light Propagation. **Imaging:** Imaging Through Scattering Media. **Optical Materials:** Heterogeneous Materials. **Scattering:** Scattering Theory.

Further Reading

Arridge SR (1999) Optical tomography in medical imaging. *Inverse Problems* 15: R41–R93.

Boas DA (1996) *Diffuse Photon Probes of Structural and Dynamical Properties of Turbid Media: Theory and Biomedical Applications*. PhD Thesis, University of Pennsylvania.

National Research Council (US), Institute of Medicine (US) (1996) *Mathematics and Physics of Emerging Biomedical Imaging*. Washington, DC: National Academy Press.

O'Leary MA (1996) *Imaging with Diffuse Photon Density Waves*. PhD Thesis, University of Pennsylvania.

Special Issue (1997) Near-infrared spectroscopy and imaging of living systems. *Philosophical Transactions of the Royal Society B* 352(1354).

Tuchin V (2002) *Handbook of Optical Biomedical Diagnostics*. Bellingham, WA: SPIE Press.

Tuchin VV, Luo Q and Ulyanov SS (2001) *Imaging of Tissue Structure and Function*. Proceedings of SPIE 4427. Bellingham, WA: SPIE Press.

Villringer A and Chance B (1997) Non-invasive optical spectroscopy and imaging of human brain function. *Trends in Neuroscience* 20: 435–442.

Villringer A and Dirnagl U (1993) *Optical Imaging of Brain Function and Metabolism*. New York: Plenum Press.

Villringer A and Dirnagl U (1997) *Optical Imaging of Brain Function and Metabolism 2: Physiological Basis and Comparison to Other Functional Neuroimaging Methods*. New York: Plenum Press.

Three-Dimensional Field Transformations

R Piestun, University of Colorado, Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Conventional optical imaging is understood as an isomorphic mapping from an object space onto an image space. The object space is composed of scattering or emitting objects. The scattering objects can be opaque, transparent, or partially absorbing, while emitting objects can be as diverse as stars in astronomical imaging or fluorescent molecules in microscopic imaging. The image space is an optical field representation of the object.

In most classical applications of optical imaging the goal is to optimize the mapping between one plane in the object space and one plane in the image space. Modern optical systems, though, are aimed at acquiring three-dimensional information regarding the object. The image is thus a three-dimensional field representation of the object.

While many imaging systems are aimed at obtaining visually appealing images, many modern imaging systems register the field information in some type of sensitive device such as film or CCD cameras for later processing. This processing is typically performed electronically by a computer and the final image is ultimately a mathematical digital representation of the object. This type of imaging is called computational imaging. An example of computational imaging is optical tomography, which involves the registration of a set of images that do not necessarily resemble the object, and off-line digital processing to obtain a three-dimensional representation of the object.

In this article we will focus on imaging in the conventional sense but will also focus on the three-dimensional aspects of the image.

Geometrical Optics Transformations

In many optical problems it is appropriate to consider the wavelength of the electromagnetic wave as infinitely small. This approximation is the origin of the field of geometrical optics in which the propagation of light is formulated in terms of geometrical laws. Accordingly, the energy is transported along light rays and their trajectories are determined by certain differential or integral equations that can be directly derived from Maxwell's equations. One of these formulations uses Fermat's principle that states the principle of the shortest optical path. The optical path between two points P_1 and P_2 is defined as:

$$OP = \int_{P_1}^{P_2} n(P)\mathrm{d}s \qquad [1]$$

where n is the refractive index of the medium. Fermat's principle states that the optical path of a real ray is always a local minimum.

In ideal optical imaging systems, every point P_o of the so-called object space will generate a sharp image point P_i, that is a point where infinite rays emerging from P_o pass through P_i. Perfect imaging occurs when every curve defined by object points is geometrically similar to its image. Perfect imaging can occur between three-dimensional spaces or between surfaces.

An interesting result of geometrical optics is Maxwell's theorem. This theorem states that if an optical instrument produces a sharp image of a three-dimensional domain, the optical length of any curve in the object is equal to the optical length of its image. More explicitly, if the object space has index n_o and the image space index n_i, the theorem states that:

$$\int_{C_o} n_o \, \mathrm{d}s_o = \int_{C_i} n_i \, \mathrm{d}s_i \qquad [2]$$

where C_i is the image of the object curve C_o.

An important consequence of Maxwell's theorem is a theorem by Carathéodory which states that any imaging transformation has to be a projective transformation, an inversion, or a combination of them. Another consequence of this theorem is that if the two media are homogeneous, perfect imaging can only occur if the magnification is equal to the ratio of the refractive indices in the object and image space. If the media have the same index, the imaging is trivial with unit magnification. The simplest example of a perfect imaging instrument is a planar mirror. Therefore, if we want to achieve nontrivial imaging between homogeneous spaces with equal index, we must give up either the perfect imaging (similarity) or the sharpness of the image.

The geometrical optics approximation is very useful in the design of optical imaging systems. However, it fails to provide a complete description

of wave optical effects, notably diffraction. Therefore, the following sections provide a description of field transformations based on wave optics.

Systems Approach to 3D Imaging

In this article we deal with linear imaging systems. Hence the relation between three-dimensional (3D) input (object) signals and 3D output (image) signals can be completely specified by the 3D impulse response or point spread function (PSF) of the system $h(x, y, z; x'y', z')$. This function specifies the response of the system at image space coordinates (x, y, z) to a δ-function input (point source) at object coordinates (x', y', z'). The optical system object and image signals are related by the superposition integral:

$$\mathrm{i}(x, y, z) = \int\int\int_{\Omega} \mathrm{o}(x', y', z') h(x, y, z; x'y', z') \mathrm{d}x'\, \mathrm{d}y'\, \mathrm{d}z' \qquad [3]$$

where Ω is the object domain. It is clear that for a generic linear imaging system the image signal is completely specified by the response to input impulses located at all possible points of the image domain.

If the system is shift invariant, the PSF due to a point source located anywhere in the object domain will be a linearly shifted version of the response due to a reference point source. This approximation is in general not valid and the system's PSF is a function of the position of the point source in the object domain. Shift invariance is traditionally divided into radial and axial shift invariance. Radial shift variance is the change in the image of a point source as a function of its radial distance from the optical axis. Radial shift variance is caused by lens aberrations. Axial shift variance is the change in the image of a point source as a function of its axial position, z', in the object space. Axial shift variance is inherent to any rotationally symmetric lens system because they can only have one pair of conjugate surfaces, i.e., object–image corresponding surfaces.

In most practical cases the image is detected at a single plane, $z = z_1$ transverse to the optical axis Oz. In those situations, we need to specify the impulse response on this plane:

$$h(x, y, z_1; x'y', z') = h_{z'}(x, y; x'y') \qquad [4]$$

The condition for lateral shift invariance is then expressed as $h_{z'}(x, y; x'y') = \tilde{h}_{z'}(x - x', y - y')$. In this case, the superposition integral is a two-dimensional (2D) convolution and it is possible to define the transfer function of the system that is the 2D Fourier transform of the impulse response $\tilde{h}_{z'}(x - x', y - y')$.

Point Spread Function for Coherent, Incoherent, and Partially Coherent Illumination

The input–output relations of an optical imaging system relate different magnitudes according to the type of illumination. If the illumination is coherent, the system is linear in the complex amplitude, while if the illumination is fully incoherent the system is linear in intensity. The 3D impulse response of the incoherent (h_{Inc}) and coherent (h_{Coh}) systems are related as follows:

$$h_{\mathrm{Inc}}(x, y, z; x'y', z') = \left|h_{\mathrm{Coh}}(x, y, z; x'y', z')\right|^2 \qquad [5]$$

Notwithstanding, many optical systems cannot be considered fully coherent or fully incoherent. In these situations the system is said partially coherent. If the optical path differences within the optical system are much shorter than the coherence length of the illumination, the light effectively has complete temporal coherence and the magnitude of interest is the mutual intensity. The mutual intensity is defined as the statistical correlation of the complex wavefunction $U(\mathbf{r}, t)$ between different points in space:

$$J(\mathbf{r}_1, \mathbf{r}_2) = \langle U^*(\mathbf{r}_1, t) U(\mathbf{r}_2, t) \rangle \qquad [6]$$

where $\mathbf{r}_i = (x_i, y_i, z_i), i = 1, 2$. The illumination is then called quasi-monochromatic. For the case of partially coherent quasi-monochromatic illumination, the system is linear in the mutual intensity. The partially coherent impulse response (h_{PC}) is also related to the coherent impulse response as follows:

$$h_{\mathrm{PC}}(\mathbf{r}_1, \mathbf{r}_2; \mathbf{r}'_1, \mathbf{r}'_2) = h^*_{\mathrm{Coh}}(\mathbf{r}_1; \mathbf{r}'_1) h_{\mathrm{Coh}}(\mathbf{r}_2; \mathbf{r}'_2) \qquad [7]$$

The 3D superposition integral that relate object and image mutual intensities is now:

$$J_{\mathrm{image}}(\mathbf{r}_1, \mathbf{r}_2) = \int\int\int_{\Omega} h_{\mathrm{PC}}(\mathbf{r}_1, \mathbf{r}_2; \mathbf{r}'_1, \mathbf{r}'_2) \times J_{\mathrm{object}}(\mathbf{r}'_1, \mathbf{r}'_2) \mathrm{d}\mathbf{r}'_1\, \mathrm{d}\mathbf{r}'_2 \qquad [8]$$

Therefore, we conclude that knowledge of the coherent PSF is enough to characterize the optical system for illumination of various degrees of coherence.

Spatial Frequency Content of the 3D Point Spread Function

As previously stated, the coherent impulse response of the optical system completely defines its imaging characteristics. Therefore, it is justified to wonder

what the limitations are in the specification of this response. In this section we analyze the frequency domain of coherent and incoherent impulse responses.

The impulse response of the optical system will satisfy the corresponding wave equations regardless of the particulars of the optical system. In particular, assuming a homogeneous isotropic medium in the imaging domain, the coherent response $h(\mathbf{r})$ to an impulse at $\mathbf{r}'$ will satisfy the homogeneous Helmholtz equation in the image domain:

$$\nabla^2 h(\mathbf{r}) + k^2 h(\mathbf{r}) = 0 \qquad [9]$$

where $h(\mathbf{r})$ is the complex amplitude and $k = 2\pi/\lambda$, with λ the wavelength in the medium. If the field at a given origin of coordinates $z = 0$, $f(x, y)$, is known, $h(\mathbf{r})$ can be calculated using the angular spectrum representation as follows:

$$h(x, y, z) = \int\int_{-\infty}^{\infty} F(f_x, f_y) \times \exp[i2\pi(f_x x + f_y y + f_z z)]\mathrm{d}f_x\, \mathrm{d}f_y \qquad [10]$$

where $f_z = \sqrt{\lambda^{-2} - (f_x^2 + f_y^2)}$ and $F(f_x, f_y) = \mathrm{FT}\{f(x, y)\}$; FT represents a 2D Fourier transformation and f_x, f_y are the spatial frequencies along the x and y axes respectively.

Equation [10] can be written as a 3D integral as follows:

$$h(x, y, z) = \int\int_{-\infty}^{\infty} F(f_x, f_y)\cdot\delta\left[f_z - \sqrt{\lambda^{-2} - f_x^2 - f_y^2}\right] \times \exp[i2\pi(f_x x + f_y y + f_z z)]\mathrm{d}f_x\, \mathrm{d}f_y\, \mathrm{d}f_z \qquad [11]$$

where δ represents the dirac impulse. Next we represent the 3D impulse response $h(x, y, z)$ as a 3D Fourier decomposition:

$$h(x, y, z) = \int\int_{-\infty}^{\infty} H(f_x, f_y, f_z) \times \exp[i2\pi(f_x x + f_y y + f_z z)]\mathrm{d}f_x\, \mathrm{d}f_y\, \mathrm{d}f_z \qquad [12]$$

where $H(f_x, f_y, f_z) = \mathrm{FT}^{3\mathrm{D}}\{h(x, y, z)\}$. Comparing eqns [11] and [12] we can infer that:

$$H(f_x, f_y, f_z) = F(f_x, f_y)\cdot\delta\left[f_z - \sqrt{\lambda^{-2} - f_x^2 - f_y^2}\right] \qquad [13]$$

This relation implies that the domain of the Fourier transform of the impulse response is confined to a spherical shell of radius $1/\lambda$. Alternatively, $H(f_x, f_y, f_z)$ is zero everywhere, except from the surface of the sphere.

The previous derivation did not consider any limitation in the range of spatial frequencies allowed through the system. We now assume a finite aperture that will introduce an effective radial cutoff frequency $f_c < 1/\lambda$ and will further limit the spatial spectral domain of $h(x, y, z)$. This is depicted in Figure 1. Note that the transverse and longitudinal spatial frequencies are related, hence the longitudinal spatial frequencies are located within a bandpass of width $\Delta f_z = 1 - \sqrt{\lambda^{-2} - f_c^2}$.

Let us now consider the frequency content of possible incoherent PSFs. From eqn [5] we derive the 3D Fourier transform of the incoherent impulse response:

$$\begin{aligned} H_{\mathrm{Inc}}(f_x, f_y, f_z) &= \mathrm{FT}\left\{h_{\mathrm{Inc}}(x, y, z; x'y', z')\right\} \\ &= \mathrm{FT}\left\{|h_{\mathrm{Coh}}(x, y, z; x'y', z')|^2\right\} \\ &= H \otimes H(f_x, f_y, f_z) \end{aligned} \qquad [14]$$

where $\otimes$ represents a 3D correlation. The domain of H_{Inc} now becomes a volume of revolution as shown in Figure 2. Moreover, because h_{Inc} is real, H_{Inc} is conjugate symmetric, i.e., $H_{\mathrm{Inc}} = H_{\mathrm{Inc}}^*$. Now, the transverse cutoff frequency is $f_{\mathrm{c}}^{\mathrm{Inc}} = 2f_{\mathrm{c}}$ while the longitudinal cutoff frequency is $f_{\mathrm{cl}}^{\mathrm{Inc}} = 2\Delta f_z$.

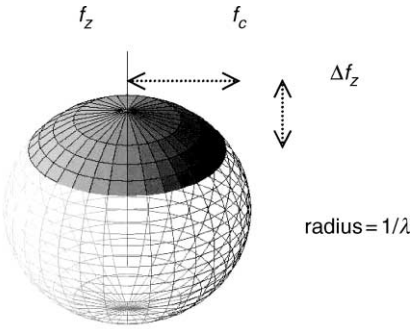


Figure 1 Three-dimensional spatial-frequency domain of the coherent point spread function.

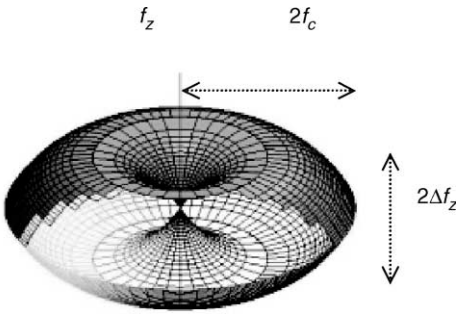


Figure 2 Three-dimensional spatial-frequency domain of the incoherent point spread function.

Diffraction Calculation of the 3D Point Spread Function

The 3D-image field created by a point source in the object space can be predicted by proper modeling of the optical system. In the absence of aberrations, the distortion of the PSF is originated in the diffraction created by the limiting aperture of the optical system. Diffraction is thus the ultimate limitation to the resolution of the optical system. Under the design conditions of the system, i.e., assuming all aberrations are corrected, the image of a point source is the diffraction pattern produced when a spherical wave converging to the (geometrical) image point is diffracted by the limiting aperture of the system. This is called a diffraction limited image.

There are many formulations to describe the 3D diffraction pattern obtained with a converging wave through an aperture. Here we review the Kirchhoff formulation. Let us consider the aperture **A** and the field at a point P_d originated from a converging spherical wave focused at point P_F (see Figure 3). We define the vectors **r** joining a generic point on the aperture P_0 and P_F, and **s** joining P_0 and P_d. The Kirchhoff formulation states that the complex amplitude $U(P_d/P_F)$ is the integral over the aperture of the complex amplitude at each point propagated to P_d by a spherical wave $\exp(iks)/s$ modified by the directional factor $\cos(\hat{\mathbf{n}}, \mathbf{r}) - \cos(\hat{\mathbf{n}}, \mathbf{s})$. $\hat{\mathbf{n}}$ is the unit vector normal to the aperture:

$$U(P_d/P_F) = \frac{-iC}{2\lambda} \int\int_A \frac{\exp[-ik(r-s)]}{rs} \times [\cos(\hat{\mathbf{n}}, \mathbf{r}) - \cos(\hat{\mathbf{n}}, \mathbf{s})] dS \qquad [15]$$

where $r = \|\mathbf{r}\|, s = \|\mathbf{s}\|$, and C is the amplitude of the spherical wave.

The directional factor varies slowly and in many situations can be approximated by a constant equal to 2 for small image-field angles. The amplitude over the aperture is also slowly varying and can be approximated by a constant, i.e., $C/r \approx C_0$. Different formulations of the Kirchhoff integral differ in the estimates of the factor $1/s$ and the phase $k(r-s)$. For example, one such formulation leads to the following expression for the 3D diffraction pattern for a focal point on the axis of an optical system with circular aperture:

$$U(P_d/P_F) = \frac{-ika^2 C_0}{z_d} \exp\left[ik\left(z_d - z_F + \frac{r_d^2}{2z_d}\right)\right] \times \int_0^1 J_0\left(\frac{ka\rho r_d}{z_d}\right) \exp\left[\frac{-ika^2\rho^2(z_d - z_F)}{2z_d z_F}\right] \rho d\rho \qquad [16]$$

where a is the radius of the aperture, r_d is the radial coordinate of P_d, and ρ is the radial coordinate of P_0.

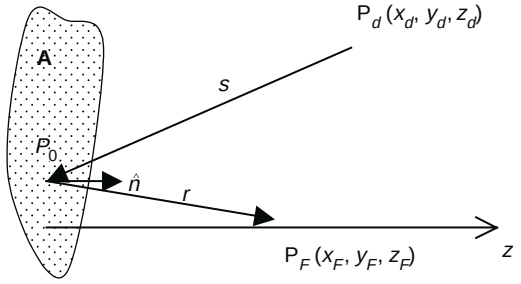


Figure 3 Diffraction of a spherical wave by a planar aperture.

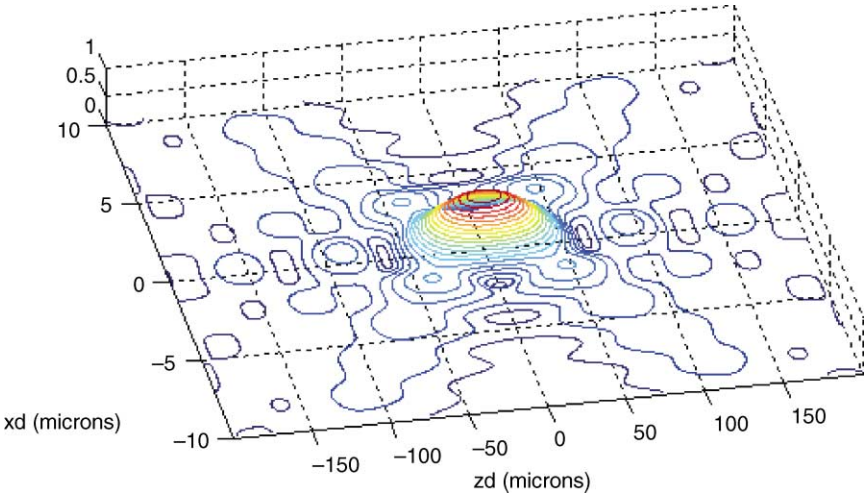


Figure 4 Lines of equal intensity in a meridional plane near the focus of a converging spherical wave diffracted at a circular aperture ($\lambda = 0.5$ μ, $a = 0.5$ cm, $z_F = 4$ cm).

To illustrate the use of eqn [16], we calculated the intensity around the focus of a spherical wave diffracted through a circular aperture of radius $a = 0.5$ cm, focusing at distance $z_F = 4$ cm from the screen. We assumed the wavelength to be $\lambda = 0.5$ μ. Figure 4 depicts the magnitude of the complex amplitude in a meridional plane near the focus of the converging spherical wave.

Models that take into account off-axis focal points have also been developed. For the most accurate calculations, it is necessary to consider the full electromagnetic nature of the optical field.

Conclusions

Classical imaging theory can be applied to the 3D imaging transformation from object to image. The systems approach to the imaging problem implies that knowledge of the 3D coherent PSF suffices to predict the 3D response of the system under illumination of varying degrees of coherence. The PSF response can be predicted using well-known formulations of diffraction.

See also

Coherent Lightwave Systems. **Fourier Optics**.

Further Reading

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. New York: Cambridge University Press.

Gibson SF and Lanni F (1989) Diffraction by a circular aperture as a model for three-dimensional optical microscopy. *Journal of the Optical Society of America A* 6: 1357–1367.

Goodman JW (1996) *Introduction to Fourier Optics*. New York: McGraw-Hill.

Li Y and Wolf E (1984) Three-dimensional intensity distributions near the focus in systems of different Fresnel numbers. *Journal of the Optical Society of America A* 1: 801–808.

Piestun R and Shamir J (2002) Synthesis of 3D light fields and applications. *Proceedings of the IEEE* 90: 222–244.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: Wiley.

Stamnes JJ (1986) *Waves in Focal Regions*. Bristol: Hilger.

Volume Holographic Imaging

G Barbastathis, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005 Published by Elsevier Ltd.

Volume holographic imaging (VHI) refers to a class of imaging techniques which use a volume hologram in at least one location within the optical path. The volume hologram acts as a 'smart lens,' which processes the optical field to extract spatial information in three dimensions (lateral as well as longitudinal) and spectral information.

Figure 1 is a generic VHI system. The object is either illuminated by a light source (e.g., sunlight or a pump laser) as shown in the figure, or it may be self-luminous. Light scattered or emitted by the object is first transformed by an objective lens and then illuminates the volume hologram. The role of the objective is to form an intermediate image which serves as input to the volume hologram. The volume hologram itself is modeled as a three-dimensional (3D) modulation $\Delta\varepsilon(\mathbf{r})$ of the dielectric index within a finite region of space. The light entering the hologram is diffracted by $\Delta\varepsilon(\mathbf{r})$ with efficiency η, defined as

$$\eta = \frac{\text{Power diffracted by the volume hologram}}{\text{Power incident to the volume hologram}} \quad [1]$$

We assume that diffraction occurs in the Bragg regime. The diffracted field is Fourier-transformed by the collector lens, and the result is sampled and measured by an intensity detector array (such as a CCD or CMOS camera).

Intuitively, we expect that a fraction of the illumination incident upon the volume hologram is Bragg-matched and is diffracted towards the Fourier-transforming lens. The remainder of the incident illumination is Bragg-mismatched, and as result is transmitted through the volume hologram un-diffracted. Therefore, the volume hologram acts as a filter which admits the Bragg-matched portion of the object and rejects the rest. When appropriately designed, this 'Bragg imaging filter' can exhibit very rich behavior, spanning the three spatial dimensions and the spectral dimension of the object.

To keep the discussion simple, we consider the specific case of a transmission geometry volume

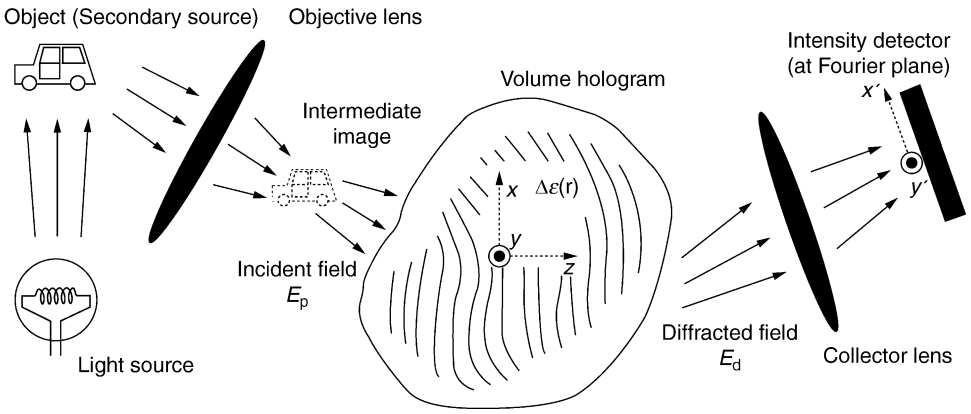


Figure 1 Volume holographic imaging (VHI) system.

hologram, described in Figure 2. The volume hologram is created by interfering a spherical wave and a planewave, as shown in Figure 2a. The spherical wave originates at the coordinate origin. The planewave is off-axis and its wavevector lies on the xz plane. As in most common holographic systems, the two beams are assumed to be at the same wavelength λ, and mutually coherent. The volume hologram results from exposure of a photosensitive material to the interference of these two beams.

First, assume that the object is a simple point source. The intermediate image is also approximately a point source, that we refer to as 'probe,' located somewhere in the vicinity of the reference point source. Assuming the wavelength of the probe source is the same as that of the reference and signal beams, volume diffraction theory shows that:

(i) If the probe point source is displaced in the y direction relative to the reference point source, the image formed by the volume hologram is also displaced by a proportional amount. Most common imaging systems would be expected to operate this way.

(ii) If the probe point source is displaced in the x direction relative to the reference point source, the image disappears (i.e., the detector plane remains dark).

(iii) If the probe point source is displaced in the z direction relative to the reference point source, a defocused and faint image is formed on the detector plane. 'Faint' here means that the fraction of energy of the defocused probe transmitted to the detector plane is much smaller than the fraction that would have been transmitted if the probe had been at the origin.

Now consider an extended, monochromatic, spatially incoherent object and intermediate image, as in Figure 2f. According to the above description, the volume hologram acts as a 'Bragg slit' in this case: because of Bragg selectivity, the volume hologram transmits light originating from the vicinity of the y axis, and rejects light originating anywhere else. For the same reason, the volume hologram affords depth selectivity (like the pinhole of a confocal microscope). The width of the slit is determined by the recording geometry, and the thickness of the volume hologram. For example, in the transmission recording geometry of Figure 2a, where the reference beam originates a distance z_0 away from the hologram, the planewave propagates at angle θ with respect to the optical axis z (assuming $\theta \ll 1$ radian), and the hologram thickness is L (assuming $L \ll z_0$), the width of the slit is found to be

$$\Delta x \approx \frac{\lambda z_0}{L\theta} \qquad [2]$$

The imaging function becomes richer if the object is polychromatic. In addition to its Bragg slit function, the volume hologram exhibits then dispersive behavior, like all diffractive elements. In this particular case, dispersion causes the hologram to image simultaneously multiple Bragg slits, each at a different color and parallel to the original slit at wavelength λ, but displaced along the z axis. Light from all these slits finds itself in focus at the detector plane, thus forming a 'rainbow image' of an entire slice through the object, as shown in Figure 2g.

To further exploit the capabilities of volume holograms, we recall that in general it is possible to 'multiplex' (superimpose) several volume gratings

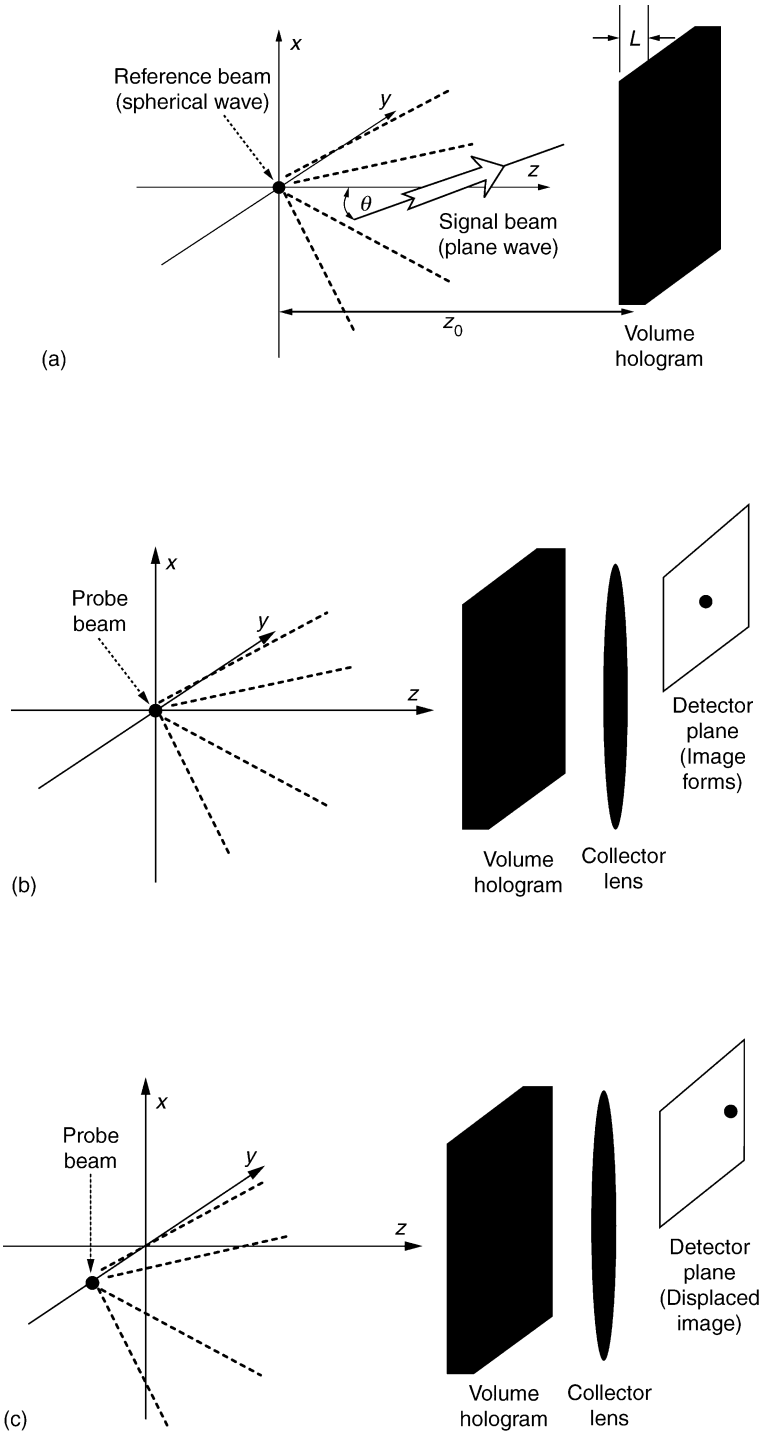


Figure 2 (a) Recoding of a transmission-geometry volume hologram with a spherical wave and an off-axis plane wave with its wave-vector on the *xz* plane. (b) Imaging of a probe point source that replicates the location and wavelength of the reference point source using the volume hologram recorded in part (a). (c) Imaging of a probe point source at the same wavelength but displaced in the *y* direction relative to the reference point source. (d) Imaging of a probe point source at the same wavelength but displaced in the *x* direction relative to the reference point source. (e) Imaging of a probe point source at the same wavelength but displaced in the *z* direction relative to the reference point source. (f) 'Bragg slitting:' Imaging of an extended monochromatic, spatially incoherent object using a volume hologram recorded as in (a). (g) Joining Bragg slits from different colors to form rainbow slices: Imaging of an extended polychromatic object using a volume hologram recorded as in (a). (h) Multiplex imaging of several slices using a volume hologram formed by multiple exposures.

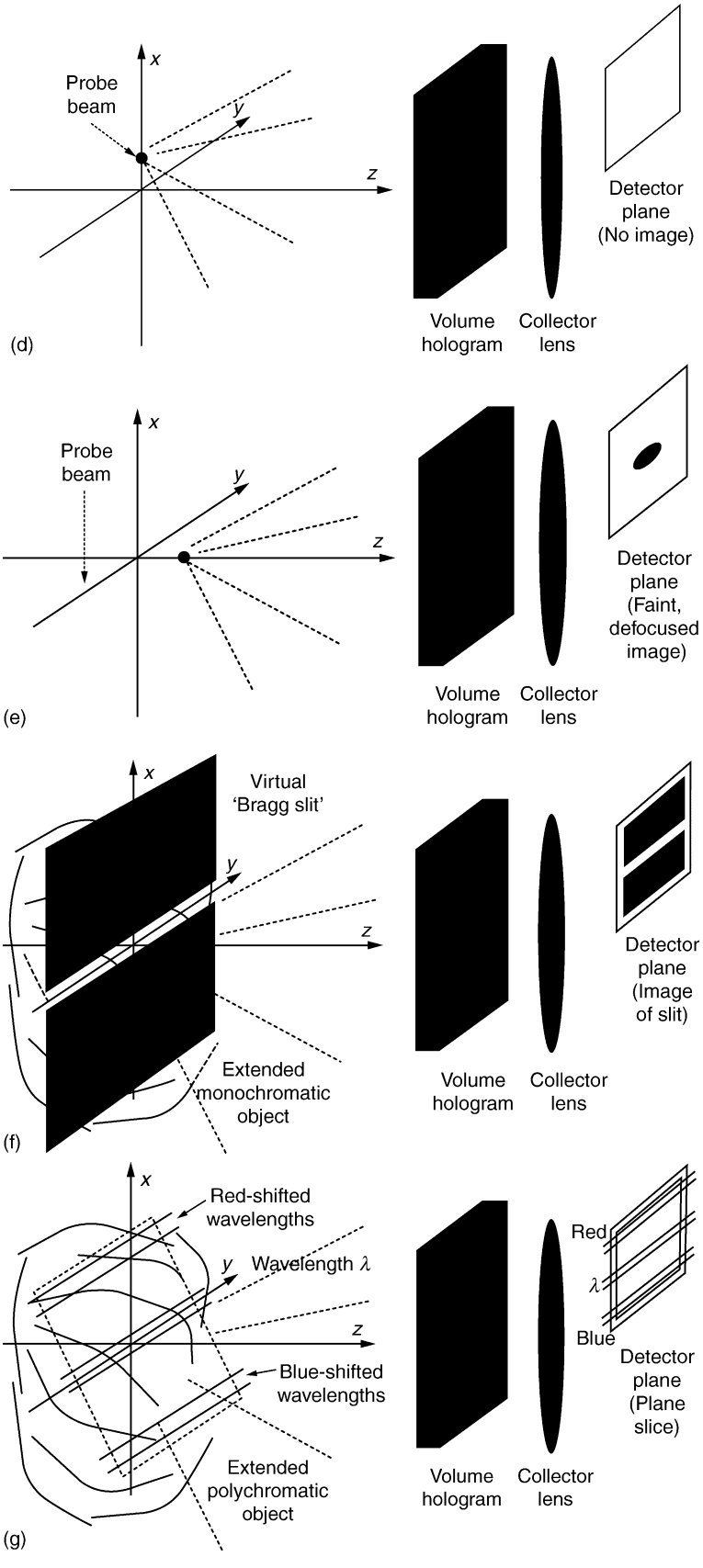


Figure 2 *Continued.*

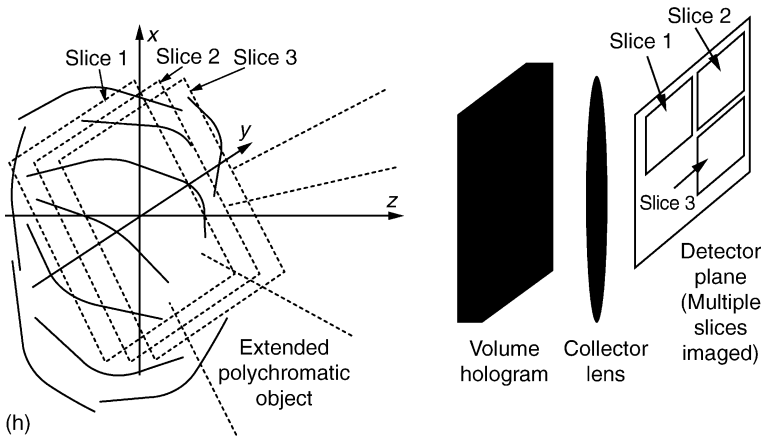


Figure 2 *Continued.*

within the same volume by successive exposures. In the imaging context, suppose that we multiplex several gratings similar to the grating described in Figure 2a but with spherical reference waves originating at different locations and plane signal waves at different orientations. When the multiplexed volume hologram is illuminated by an extended polychromatic source, each grating forms a separate image of a rainbow slice, as described earlier. By spacing appropriately the angles of propagation of the plane signal waves, we can ensure that the rainbow images are formed on nonoverlapping areas on the detector plane, as shown in Figure 2h. This device is now performing true four-dimensional (4D) imaging: it is separating the spatial and spectral components of the object illumination so that they can be measured independently by the detector array. Assuming the photon count is sufficiently high and the number of detector pixels is sufficient, this 'spatio-spectral slicing' operation can be performed in real time, without need for mechanical scanning.

The complete theory of VHI in the spectral and spatial domains is given in the Further Reading section below. Experimental demonstrations of VHI have been performed in the context of a confocal microscope where the volume hologram performs the function of a 'Bragg pinhole' as well as a real-time 4D imager. The limited diffraction efficiency η of volume holograms poses a major concern for VHI systems. Holograms, however, are known to act as matched filters. It has been shown that the matched filtering nature of volume holograms as imaging elements is superior to other filtering elements (e.g., pinholes or slits) in an information-theoretic sense. Efforts are currently underway to strengthen this property by design of volume holographic imaging elements which are more elaborate than described here.

See also

Holography, Techniques: Overview. **Imaging:** Three-Dimensional Field Transformations.

Further Reading

Barbastathis G and Brady DJ (1999) Multidimensional tomographic imaging using volume holography. *Proceedings of the IEEE* 87(12): 2098–2120.

Barbastathis G and Sinha A (2001) Information content of volume holographic images. *Trends in Biotechnology* 19(10): 383–392.

Barbastathis G, Balberg M and Brady DJ (1999) Confocal microscopy with a volume holographic filter. *Optics Letters* 24(12): 811–813.

Barbastathis G, Sinha A and Neifeld MA (2001) Information-theoretic treatment of axial imaging. In: *OSA Topical Meeting on Integrated Computational Imaging Systems*, pp. 18–20. Albuquerque, NM.

Liu W, Psaltis D and Barbastathis G (2004) Real time spectral imaging in three spatial dimensions. *Optics Letters* in press.

van-der Lugt A (1964) Signal detection by complex spatial filtering. *IEEE Transactions on Information Theory* IT-10(2): 139–145.

Wavefront Sensors and Control (Imaging Through Turbulence)

C L Matson, Air Force Research Laboratory, Kirtland, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In many situations, distant objects are imaged using optical or near-infrared imaging systems. Examples include terrestrial surveillance from space, tactical surveillance from airborne imaging systems, and ground-based astronomical imaging. Ideally, the resolution in these images is limited only by the wavelength of light λ used to create the images divided by the imaging system's effective diameter D. The ratio λ/D is called the diffraction limit of the imaging system. Unfortunately, when the imaging system is located in the Earth's atmosphere, it is often atmospheric turbulence fluctuations that limit the level of detail in the long-exposure images, not the diffraction limit of the telescope. For example, the Keck telescopes, located on top of Mauna Kea in Hawaii, are 10 m diameter telescopes that have a diffraction-limited angular resolution of 0.2 μrad for an imaging wavelength of 2 μm. However, atmospheric turbulence limits the achievable resolution in long-exposure images to the order of 1–5 μrad at this wavelength.

The important discovery by Anton Labeyrie in 1970 that short-exposure images of objects retain meaningful diffraction-limited information about the object, where 'short' means that the exposure time is less than the atmospheric turbulence coherence time, has revolutionized imaging through turbulence. Atmospheric coherence times depend upon a number of factors, including wind speeds, wind altitudes, and observation wavelength. At astronomical observatories for visible observation wavelengths, a rule of thumb is that atmospheric turbulence stays essentially constant for exposure times of one to ten milliseconds under average clear-sky weather conditions. A variety of techniques that exploit this discovery have been developed. These techniques can be grouped into three different categories: techniques that remove blurring from the detected intensity images (post-detection compensation), techniques that modify the optical field prior to detecting the light (pre-detection compensation), and hybrid techniques that use both post-detection and pre-detection compensation. In this article, a variety of these techniques are presented that are representative of the myriad of methods that have been developed during the past decades. To simplify the discussion, it is assumed that the blurring due to atmospheric turbulence is the same at every point in the image (isoplanaticity), that the images are formed using incoherent light as emitted by the sun and stars, and that a single filled aperture is used to create the images. Interferometric techniques that use multiple apertures to estimate the mutual coherence function of the optical field and calculate an image from it, are discussed in a separate article (*see* **Imaging:** Interferometric Imaging).

The emphasis in this article is on identifying and removing atmospheric blurring in images. As part of this process, regularization techniques must be included in the image reconstruction algorithms in order to obtain the best possible resolution for an acceptable level of noise. Although regularization techniques are not discussed in this article, information can be found in the Further Reading section at the end of this article (*see* **Imaging:** Inverse Problems and Computational Imaging).

Imaging and Atmospheric Turbulence

In Figure 1, a conceptual diagram illustrates the impact that atmospheric turbulence has upon resolution in classical imaging systems. A distant star is being imaged by a telescope/detector system. In the absence of atmospheric turbulence, the field emitted by the star has a planar wavefront over the extent of the lens since the star is a long distance from the imaging system and thus unresolved by the telescope. When imaged by the telescope, the diffraction pattern of the telescope is obtained, resulting in a diffraction-limited image (Figure 2). However, in the presence of atmospheric turbulence near the telescope, the planar wavefront is randomly distorted by the turbulence, resulting in a corrugated wavefront with a spatial correlation scale given by the Fried parameter r_o. For visible wavelengths, at good astronomical observing sites, r_o ranges from 5 cm to 20 cm. When this corrugated wavefront is converted to an image by the telescope, a distorted image is obtained (Figure 3). Because atmospheric turbulence is time-varying, the distortions in the image also change over time. If the imaging shutter on the telescope remains open for many occurrences of atmospheric turbulence, the resulting image will be an average of all the individually distorted images, resulting in what is known as the seeing disk (Figure 4). It is the seeing disk that typically limits spatial resolution in long-exposure images, not the diffraction pattern of the

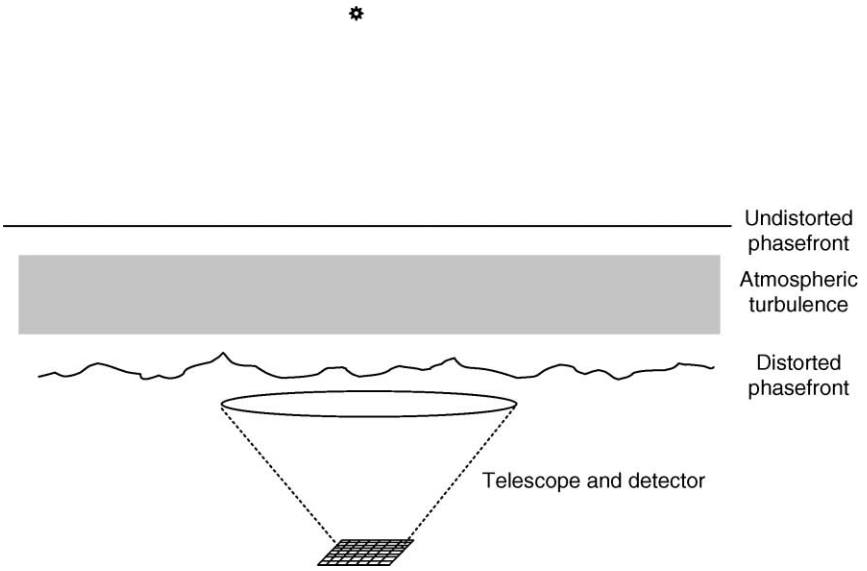


Figure 1 Conceptual diagram of imaging through atmospheric turbulence.

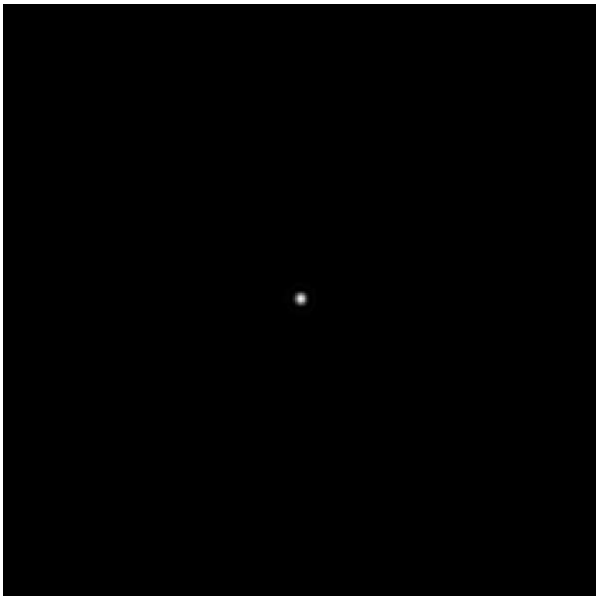


Figure 2 Computer-simulated diffraction-limited image of a point source.

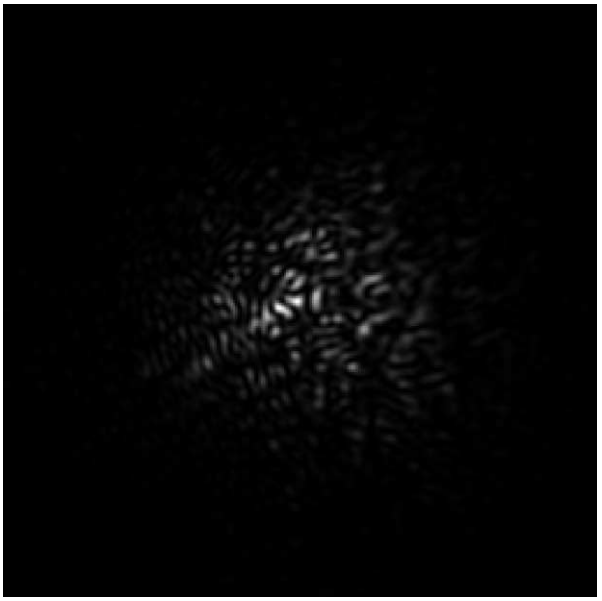


Figure 3 Computer-simulated short-exposure image of a point source through atmospheric turbulence.

telescope. As a result, the effective spatial resolution in long-exposure images in the presence of atmospheric turbulence is given by λ/r_o, not λ/D.

Atmospheric turbulence distorts a field's wavefront by delaying it in a way that is spatially and temporally random. The origin of these delays is the random variation in the index of refraction of the atmosphere as a result of turbulent air motion brought about by temperature variations. A Fourier optics imaging model is useful in characterizing the effects of atmospheric turbulence on spatial resolution in images. Let $i(\vec{x})$ be an image of an object $o(\vec{x})$, where $\vec{x}$ is a two-dimensional spatial vector, and let $I(\vec{f})$ and $O(\vec{f})$ be their spatial Fourier transforms, where $\vec{f}$ is a two-dimensional spatial frequency vector. In the absence of atmospheric turbulence, the diffraction-limited Fourier transform $I(\vec{f})$ is given by

$$I_d(\vec{f}) = O(\vec{f})H_d(\vec{f}) \qquad [1]$$

where the subscript 'd' denotes the diffraction-limited case and where $H_d(\vec{f})$ is the diffraction-limited

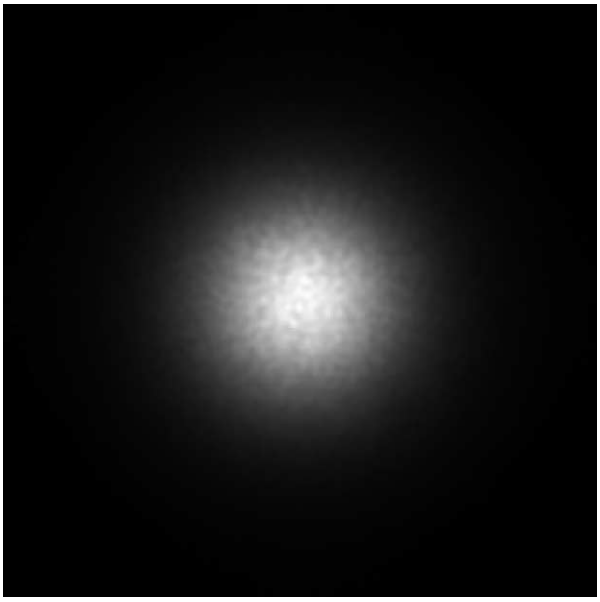


Figure 4 Computer-simulated long-exposure image of a point source through atmospheric turbulence.

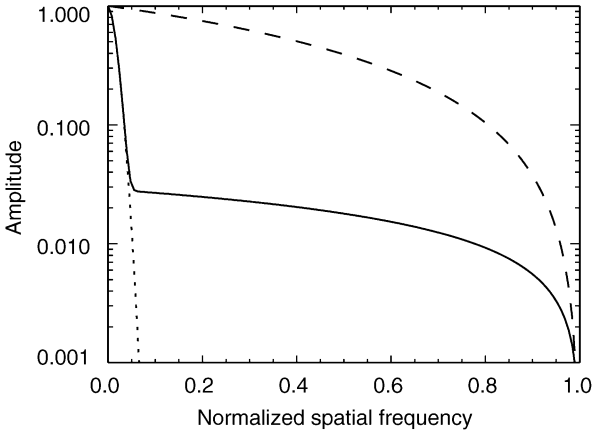


Figure 5 Plots of the Fourier amplitudes of the diffraction-limited transfer function (dashed line), Fourier amplitudes of the long-exposure transfer function (dotted line), and the square root of the average short exposure energy spectrum. The spatial frequency axis is normalized to one at the diffraction limit of the telescope.

telescope transfer function. In the presence of atmospheric turbulence, the Fourier transform of a long exposure image $i_{\mathrm{le}}(\vec{x})$ is given by

$$I_{\mathrm{le}}(\vec{f}) = O(\vec{f})H_{\mathrm{d}}(\vec{f})H_{\mathrm{le}}(\vec{f}) = I_{\mathrm{d}}(\vec{f})H_{\mathrm{le}}(\vec{f}) \qquad [2]$$

where the subscript 'le' denotes long exposure. If the turbulence satisfies Kolmogorov statistics, as is typically assumed, then:

$$H_{\mathrm{le}}(\vec{f}) = \exp\left[-3.44\left(\frac{\lambda d|\vec{f}|}{r_{\mathrm{o}}}\right)^{5/3}\right] \qquad [3]$$

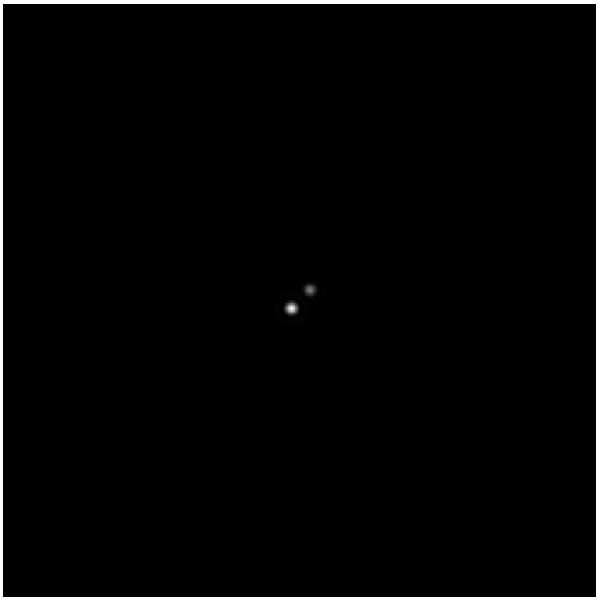


Figure 6 Computer-simulated diffraction-limited image of a binary star.

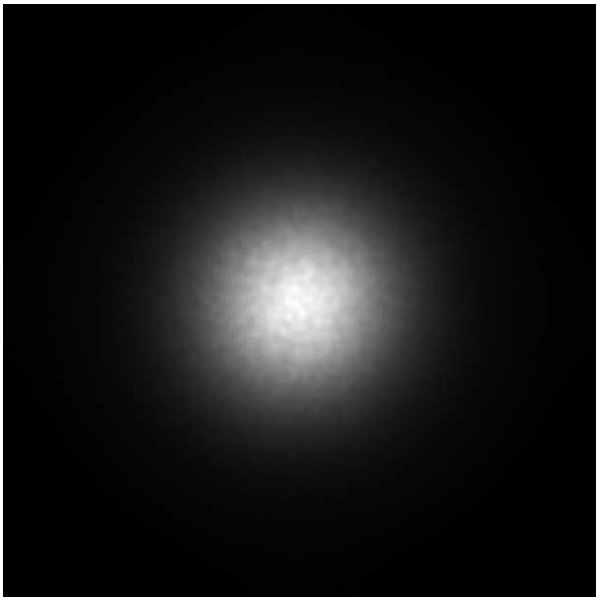


Figure 7 Computer-simulated long-exposure image of a binary star.

where d is the distance between the imaging system's exit pupil and detector plane. As can be seen from eqn [2], the effect of atmospheric turbulence on a long-exposure image is a multiplication of the diffraction-limited image's Fourier transform by a transfer function due to atmospheric turbulence. In Figure 5, an amplitude plot of a diffraction-limited transfer function is displayed along with an amplitude plot of a long-exposure transfer function for a 2.3 m

diameter telescope, an imaging wavelength of 0.5 μm, and an r_o value of 10 cm. Notice that the long exposure transfer function significantly reduces the amount of high spatial frequency information in an image. Diffraction-limited and long-exposure turbulence-limited images of a binary star are shown in Figures 6 and 7, illustrating the significant loss of spatial resolution brought on by atmospheric turbulence.

Post-Detection Compensation

Speckle Imaging

Labeyrie's 1970 discovery was motivated by his analysis of short-exposure star images. In these images (see Figure 3), a number of speckles are present, where the individual speckle sizes are roughly the size of the diffraction-limited spot of the telescope. From this, he realized that high spatial frequency information is retained in short-exposure images, albeit distorted. He then showed that the energy spectra $E_I(\vec{f})$ of the individual short-exposure images can be averaged to obtain an estimate of the average energy spectrum that retains high signal-to-noise ratio (SNR) information at high spatial frequencies. The square root of the average energy spectrum (the amplitude spectrum) of a star, calculated using Labeyrie's method (also known as speckle interferometry), is shown in Figure 5, along with the diffraction-limited and long-exposure amplitude spectra. Notice that meaningful high spatial frequency information is retained in the average amplitude spectrum out to the diffraction limit of the telescope. Because the atmospherically corrupted amplitude spectrum estimate is significantly lower than the diffraction-limited amplitude, a number of short exposure images must be used to build up the SNRs at the higher spatial frequencies.

To reconstruct an image, both the Fourier amplitudes and Fourier phases of the image must be known. In fact, it is well known that the Fourier phase of an image carries most of the information in the image. Because only the Fourier amplitudes are available using Labeyrie's method, other techniques were subsequently developed to calculate the Fourier phase spectrum from a sequence of short-exposure images. Techniques that combine amplitude spectrum estimation using Labeyrie's method along with phase spectrum estimation are called speckle imaging techniques. The two most widely used phase spectrum estimation methods are the Knox–Thompson (or cross spectrum) and bispectrum algorithms. Both of these algorithms calculate Fourier-domain quantities that retain high spatial frequency information about the Fourier phase spectrum when averaged over multiple short-exposure images. The cross spectrum and bispectrum techniques calculate the average cross spectrum $C(\vec{f}, \Delta\vec{f})$ and bispectrum $B(\vec{f}_1, \vec{f}_2)$, respectively, that are given by

$$C(\vec{f}, \Delta\vec{f}) = I(\vec{f})I^*(\vec{f} + \Delta\vec{f}) \quad [4]$$

$$B(\vec{f}_1, \vec{f}_2) = I(\vec{f}_1)I(\vec{f}_2)I^*(\vec{f}_1 + \vec{f}_2) \quad [5]$$

where the superscript $*$ denotes complex conjugate. Although the cross spectrum is defined for all values of $\vec{f}$ and $\Delta\vec{f}$, only for values of $\Delta\vec{f}$ or $\vec{f}$ less than r_o/λ are the cross spectrum elements obtained with sufficiently high SNRs to be useful in the reconstruction process. The equivalent constraint on the bispectrum elements is that either $\vec{f}_1, \vec{f}_2$, or $\vec{f}_1 + \vec{f}_2$ must be less than r_o/λ. Once the average cross spectrum or bispectrum is calculated from a series of short-exposure images, the phase spectrum $\phi_I(\vec{f})$ of the underlying image can be calculated from their phases using one of a number of algorithms based upon the following equations:

$$\phi_I(\vec{f} + \Delta\vec{f}) = \phi_I(\vec{f}) - \phi_C(\vec{f} + \Delta\vec{f}) \quad [6]$$

for the cross spectrum and:

$$\phi_I(\vec{f}_1 + \vec{f}_2) = \phi_I(\vec{f}_1) + \phi_I(\vec{f}_2) - \phi_B(\vec{f}_1 + \vec{f}_2) \quad [7]$$

for the bispectrum. Because the phases are only known for units of 2π, phasor versions of eqns [6] and [7] must be used. For the cross spectrum technique, each short-exposure image must be centroided prior to calculating its cross spectrum and adding it to the total cross spectrum. The bispectrum technique is insensitive to tilt, so no centroiding is necessary; however, the bispectrum technique requires setting two image phase spectrum values to arbitrary values to initialize the phase reconstruction process. Typically, the image phase spectrum values at $\vec{f} = (0, 1)$ and $\vec{f} = (1, 0)$ are set equal to zero, which centroids the reconstructed image but does not affect its morphology.

To obtain an estimated Fourier transform $I_e(\vec{f})$ of the image from the energy and phase spectrum estimates, the square root of the average energy spectrum is combined with the average phase spectrum calculated from either the bispectrum or cross spectrum. The result is given by

$$I_e(\vec{f}) = [E_I(\vec{f})]^{1/2} \exp[j\phi_I(\vec{f})] \quad [8]$$

where $j = \sqrt{-1}$. The result in eqn [8] includes atmospheric attenuation of the Fourier amplitudes, so it must be divided by the amplitude spectrum of the atmosphere to obtain the desired object spectrum estimate. The atmospheric amplitude spectrum estimate is obtained by using Labeyrie's technique on a sequence of short-exposure images of a nearby unresolved star. No phase spectrum correction is necessary because it has been shown theoretically and verified experimentally that atmospheric turbulence has a negligible effect upon $\phi_I(\vec{f})$.

Blind Deconvolution

Speckle imaging techniques are effective as well as easy and fast to implement. However, it is necessary to obtain a separate estimate of the atmospheric energy spectrum using an unresolved star in order to reconstruct a high-resolution image. Usually, the star is not co-located with the object and the star images are not collected simultaneously with the images of the object. As a result, the atmospheric energy spectrum estimate may be inaccurate. At the very least, collecting the additional star data expends valuable observation time. In addition, there is often meaningful prior knowledge about the object and the data, such as the fact that measured intensities are positive (positivity), that many astronomical objects are imaged against a black background (support), and that the measured data are band-limited, among others. Speckle imaging algorithms do not exploit this additional information. As a result, a number of techniques have been developed to jointly estimate the atmospheric blurring effects as well as an estimate of the object being imaged using only the short-exposure images of the object and available prior knowledge. These techniques are called blind deconvolution techniques. Two of the most common techniques are known as multi-frame blind deconvolution (MFBD) and projection-based blind deconvolution (PBBD).

The MFBD algorithm is based upon maximizing a figure of merit to determine the best object and atmospheric short-exposure point spread function (PSF) estimates given the data and available prior knowledge. This figure of merit is created using a maximum likelihood problem formulation. Conceptually, this technique produces estimates of the object and the atmospheric turbulence short-exposure PSFs that most likely generated the measured short-exposure images. Mathematically, the MFBD algorithm obtains these estimates by maximizing the conditional probability density function (pdf) $\mathrm{p}[i_m(\vec{x})|o(\vec{x}); h_m(\vec{x}),\ m = 1, \ldots, M]$, where $i_m(\vec{x})$ is the mth of M short exposure images and is given by

$$i_m(\vec{x}) = h_m(\vec{x}) * o(\vec{x}) + n_m(\vec{x}) \qquad [9]$$

where $h_m(\vec{x})$ and $n_m(\vec{x})$ are the atmospheric/system blur and detector noise functions for the mth image, respectively, and the asterisk denotes convolution. The noise is assumed to be zero mean with variance $\sigma^2(\vec{x})$. The conditional pdf $\mathrm{p}[i_m(\vec{x})|o(\vec{x}); h_m(\vec{x}),\ m = 1, \ldots, M]$ is the pdf of the noise with a mean equal to $h_m(\vec{x}) * o(\vec{x})$ and, as a function of $o(\vec{x})$ and $h_m(\vec{x})$, it is known as a likelihood function. For spatially independent Gaussian noise, as is the case for camera read noise, the likelihood function is given by

$$\mathrm{L_G}[o(\vec{x}); h_m(\vec{x}),\ m = 1, \ldots, M] = \prod_m \prod_{\vec{x}} \frac{1}{\sqrt{2\pi\sigma^2(\vec{x})}} \times \exp\left\{ -\left[i(\vec{x}) - (o * h_m)(\vec{x})\right]^2 / 2\sigma^2(\vec{x}) \right\} \qquad [10]$$

For Poisson noise, the likelihood function is given by

$$\mathrm{L_P}[o(\vec{x}); h_m(\vec{x}), m = 1, \ldots, M] = \prod_m \prod_{\vec{x}} \frac{\left[(o * h_m)(\vec{x})\right]^{i(\vec{x})} \exp\left\{ -\left[(o * h_m)(\vec{x})\right] \right\}}{i(\vec{x})!} \qquad [11]$$

In practice, logarithms of eqns [10] and [11] are taken prior to searching for the estimates of $o(\vec{x})$ and $h_m(\vec{x})$ that maximize the likelihood function.

In general, there are an infinite number of estimates of $o(\vec{x})$ and $h_m(\vec{x})$ that reproduce the original dataset when convolved together. However, when additional prior knowledge is included to constrain the solution space, such as positivity, support, and the knowledge that $h_m(\vec{x})$ is band-limited, along with regularization, the MFBD algorithm almost always converges to the correct estimates of $o(\vec{x})$ and $h_m(\vec{x})$.

The PBBD technique also uses prior knowledge constraints to jointly estimate $o(\vec{x})$ and $h_m(\vec{x})$. However, instead of creating and maximizing a likelihood function, the PBBD technique uses the method of convex projections to find estimates of $o(\vec{x})$ and $h_m(\vec{x})$ that are consistent with the measured short-exposure images and all the prior knowledge constraints. The data consistency constraint and the prior knowledge constraints are mathematically formulated as convex sets and the PBBD technique repeatedly takes the current estimates of $o(\vec{x})$ and $h_m(\vec{x})$ and projects them onto the convex constraint sets. As long as all the sets are truly convex, the

algorithm is guaranteed to converge to a solution that is in the intersection of all the sets as long as the algorithm doesn't stagnate at erroneous solutions called traps. These traps are the projection equivalent of local minima in cost function minimization algorithms.

A key component of this algorithm is the ability to formulate the measured short-exposure images and prior knowledge constraints into convex sets. As an example, consider the convex set $C_{\text{pos}}[u(\vec{x})]$ that corresponds to the positivity constraint. This set is given by

$$C_{\text{pos}}[u(\vec{x})] \triangleq \{u(\vec{x}) \text{ such that } u(\vec{x}) \geq 0\} \qquad [12]$$

It is straightforward to show that $C_{\text{pos}}[u(\vec{x})]$ is a convex set. Other constraints can be converted into convex sets in a similar manner. Although all constraints can be formulated into sets, most but not all of these sets are convex.

The projection process is straightforward to implement. For example, the projections of the current estimates of $o(\vec{x})$ and $h_{\text{m}}(\vec{x})$ onto $C_{\text{pos}}[u(\vec{x})]$ are carried out by zeroing out the negative values in these estimates. Similarly, the projections of $o(\vec{x})$ and $h_{\text{m}}(\vec{x})$ onto the convex sets that correspond to their support constraints are carried out by zeroing out all their values outside of their supports. In practice, some versions of PBBD algorithms have been formulated to use minimization routines as part of the projection process in order to minimize the possibility of stagnation.

Deconvolution from Wavefront Sensing

A key distinguishing feature of post-detection compensation algorithms is the type of additional data required to estimate and remove atmospheric blurring in the measured short-exposure images. Blind deconvolution techniques need no additional data. Speckle imaging techniques estimate average atmospheric blurring quantities and need an estimate of the average atmospheric energy spectrum. Because average quantities are calculated when using speckle imaging techniques, only an estimate of the average atmospheric energy spectrum is needed, not the specific energy spectra that corrupted the measured short-exposure images. As a result, the additional data need not be collected simultaneously with the short-exposure images. The technique described in this section, deconvolution from wavefront sensing (DWFS), seeks to deconvolve the atmospheric blurring in each short-exposure image by estimating the specific atmospheric field phase perturbations that blur each short-exposure image. Thus, the additional data must be collected simultaneously with the short-exposure images. The atmospheric field phase perturbations are calculated from data obtained with a wavefront sensor using a different region of the optical spectrum than used for imaging so that no light is taken from the imaging sensor. Most commonly, the wavefront sensor data consists of the first derivatives of the phasefront (phase differences) as obtained with Shack–Hartmann or shearing interferometer wavefront sensors, but curvature sensing wavefront sensors are also used that produce data that are proportional to the second derivatives of the phasefront.

The original algorithm to implement the DWFS technique produced an estimate $\hat{O}(\vec{f})$ of the Fourier transform $O(\vec{f})$ of the object being imaged using the following estimator:

$$\hat{O}(\vec{f}) = \frac{\sum_m I_{\text{m}}(\vec{f})\tilde{H}^*_{\text{m}}(\vec{f})}{\sum_m |\tilde{H}_{\text{m}}(\vec{f})|^2} \qquad [13]$$

where $\tilde{H}_{\text{m}}(\vec{f})$ is the overall system transfer function (including the telescope and atmospheric effects) for the mth short-exposure image estimated from the wavefront sensor data. To create $\tilde{H}_{\text{m}}(\vec{f})$, the atmospheric field phase perturbations corrupting the mth image must be calculated from the wavefront sensor measurements. This topic is discussed in more detail in **Imaging**: Adaptive Optics. Then $\tilde{H}_{\text{m}}(\vec{f})$ is obtained using standard Fourier optics techniques. Although eqn [13] is easily implemented given an existing system that includes a wavefront sensor, it suffers from the fact that it is a biased estimator. An unbiased version of eqn [13] is given by

$$\hat{O}(\vec{f}) = \frac{\sum_m I_{\text{m}}(\vec{f})\tilde{H}^*_{\text{m}}(\vec{f})}{\sum_n H_{\text{n,ref}}(\vec{f})\tilde{H}^*_{\text{n,ref}}(\vec{f})} \qquad [14]$$

where $\tilde{H}_{\text{n,ref}}(\vec{f})$ and $H_{\text{n,ref}}(\vec{f})$ are estimates of the overall system transfer function obtained from data collected on a reference star using wavefront sensor data and image-plane data, respectively. Thus, to obtain an unbiased estimator, two more datasets must be collected. The number of frames in the reference star dataset need not be the same as for the imaging dataset. As for speckle imaging, the atmospheric statistics must be the same for the reference star and imaging datasets.

Another technique that uses at least one additional data channel along with the measured image is called

phase diversity. In the additional data channels, images are also recorded, but with known field phase aberrations intentionally introduced to add known diversity to the unknown field phase corrupting the measurements. The most commonly introduced field phases aberration in an additional channel is defocus due to its ease of implementation. In the case of defocus, the additional data channel along with the imaging channel can be viewed as samples of the focal volume of the imaging system. Both the noise-free object and the unknown field phase aberrations can be recovered from the measured data using optimization techniques to enforce data consistency along with other prior knowledge constraints. These types of systems can use higher-order aberrations as well. The phase diversity concept has been generalized to an approach known as computational imaging, where an imaging system is designed by taking into account its intended use, the expected image degradation mechanisms, and the fact that the image will be the product of both an optical system and a postprocessing algorithm.

Pre-Detection Compensation

The primary pre-detection-only compensation technique in use today employs dynamical electro-optical systems to sense and remove the phases in the impinging wavefront due to atmospheric turbulence. This technique, known as adaptive optics, provides much higher-quality data than post-detection compensation techniques, when there is enough light from the object being imaged or a guidestar (artificial or natural) to drive the adaptive optics system at a bandwidth high enough to keep up with the temporal changes in atmospheric turbulence. It consists of a wavefront sensor that measures the portion of the phasefront in the pupil of the telescope that is due to atmospheric turbulence, a wavefront reconstructor that generates the deformable mirror drive signals from the wavefront sensor data, and a deformable mirror that applies the conjugate of the atmospheric phase distortions to a reimaged pupil. Some of the light reflected off the deformable mirror is then sent to the imaging sensor. Adaptive optics systems are discussed in much greater detail in **Imaging**: Adaptive Optics. Here, it suffices to say that no adaptive optics system produces imagery that is diffraction limited in quality, although, under the right conditions, the images can be near diffraction limited. However, atmospheric seeing conditions fluctuate significantly over a night as well as seasonally. Therefore, virtually all adaptive optics images benefit from applying post-compensation techniques, as is discussed next.

Hybrid Techniques

Adaptive Optics and Post-Processing

Often, an adaptive optics system does not fully correct for atmospheric turbulence. This can occur when atmospheric conditions become worse than the system is designed to handle. Also, an adaptive optics system may be intentionally designed to only partially correct for atmospheric turbulence in order to lower the cost of the system. As a result, there still may be significant atmospheric blurring in systems that use adaptive optics. For this reason, additional post-processing of adaptive optics imagery can result in significant improvement in image quality. All of the techniques discussed in the section on post-detection compensation can be applied to adaptive optics imagery. Applying these techniques should never degrade image quality, and will improve the quality if the adaptive optics system doesn't fully correct for atmospheric turbulence. Generally speaking, the amount of improvement decreases as the performance of the adaptive optics system increases. Because the adaptive optics system's performance parameters can vary as atmospheric conditions vary, the blind deconvolution algorithms are especially applicable to processing adaptive optics imagery since they estimate the overall system transfer function from the imaging data. Because speckle imaging techniques require a separate estimate of the overall system transfer function, it can be difficult and time-consuming to obtain all of the additional data necessary to apply these techniques to adaptive optics data. Since these post-detection compensation techniques require short-exposure images, the adaptive optics system needs to have a camera capable of collecting short-exposure images. Finally, even if an adaptive optics system fully corrects for the atmosphere, deconvolution techniques that remove the diffraction-limited transfer function amplitude attenuation can improve image quality.

Partially Redundant Pupil Masking

Although adaptive optics systems can function very well, currently they are expensive and require significant maintenance. The primary benefit of adaptive optics systems is that they remove atmospheric turbulence phase distortions in the pupil of the telescope, thus permitting light from all areas of the pupil to interfere in phase when creating an image. An interferometric view of imaging reveals that multiple regions of the light in the pupil contribute to a single spatial frequency in the image. When the phasefront in the pupil is distorted

by the atmosphere, these multiple regions tend to not interfere in phase, reducing the amplitude of the image's Fourier component at that spatial frequency. This amplitude reduction effect occurs for all spatial frequencies greater than the limit imposed by the atmosphere, r_o/λ. Therefore, techniques have been developed to modify the wavefront by masking part of the pupil prior to measuring the field with an imaging sensor. This class of techniques has various names including optical aperture synthesis, pupil masking, and aperture masking. Here this class of techniques will be referred to as partially redundant pupil masking (PRPM).

PRPM techniques function by masking the pupil in such a way as to minimize the amount of out-of-phase interference of light on the imaging detector. Early efforts used pupil masks that were opaque plates with r_o-sized holes in them. Because the atmospheric coherence length is r_o, all the light from an r_o-sized hole is essentially in phase. By choosing the holes in the mask so that only two holes contribute to a single spatial frequency in the image, out-of-phase interference problems are removed. This type of mask is called a nonredundant pupil mask. Of course, a significant amount of light is lost due to this masking, so only bright objects benefit from this particular implementation. Often, the pupil mask needs to be rotated in order to obtain enough spatial frequencies in the image to be able to reconstruct a high-quality estimate of the object. To increase the light levels in the image, masks have been made to allow some out-of-phase interference for a given spatial frequency, thus trading off the amplitude of the spatial frequencies allowed in the image with the amount of light in the image. Masks with this property are called partially redundant pupil masks.

One version of PRPM is called the variable geometry pupil (VGP) technique. It is envisioned to work in conjunction with an adaptive optics system that does not fully correct the wavefront for atmospheric phase distortions. It employs a spatial light modulator or a micro-electromechanical mirror between the deformable mirror and the imaging sensor. Either of these two devices is driven by signals from the wavefront sensor in the adaptive optics system in such a way that the portions of the pupil where the phasefront has a sufficiently large second derivative is blocked, while the remainder of the phasefront is transmitted.

All PRPM techniques benefit from post-compensation processing. For the case of nonredundant pupil masks, post-processing techniques are required to convert the discrete-spatial-frequency information into an image with reasonable quality. These techniques are essentially the same techniques as used in interferometric imaging, not those described in this article. When partially redundant pupil masks are used (including VGP masks), the image may include enough spatial frequency content that the methods described in the post-detection compensation section can be applied.

See also

Imaging: Adaptive Optics; Interferometric Imaging; Inverse Problems and Computational Imaging.

Further Reading

Ayers GR, Northcott MJ and Dainty JC (1988) Knox–Thompson and triple-correlation imaging through atmospheric turbulence. *Journal of the Optical Society of America A* 5: 963–985.

Bertero M and Boccacci P (1998) *Introduction to Inverse Problems in Imaging*. Bristol, UK: Institute of Physics Publishing.

Craig IJD and Brown JC (1986) *Inverse Problems in Astronomy*. Bristol, UK: Adam Hilger Ltd.

Dainty JC (1984) Stellar speckle interferometry. In: Dainty JC (ed.) *Topics in Applied Physics, vol. 9: Laser Speckle and Related Phenomena*, 2nd edn, pp. 255–320. Berlin: Springer-Verlag.

Goodman JW (1985) *Statistical Optics*. New York: John Wiley and Sons.

Haniff CA and Buscher D (1992) Diffraction-limited imaging with partially redundant masks I: Infrared imaging of bright objects. *Journal of the Optical Society of America A* 9: 203–218.

Katsaggelos AK and Lay KT (1991) Maximum likelihood identification and restoration of images using the expectation-maximization algorithm. In: Katsaggelos AK (ed.) *Digital Image Restoration*, pp. 143–176. Berlin: Springer-Verlag.

Nakajima T and Haniff CA (1993) Partial adaptive compensation and passive interferometry with large ground-based telescopes. *Publications of the Astronomical Society of the Pacific* 105: 509–520.

Roggemann MC and Welsh B (1996) *Imaging Through Turbulence*. Boca Raton, FL: CRC Press.

Schulz TJ (1993) Multiframe blind deconvolution of astronomical images. *Journal of the Optical Society of America A* 10: 1064–1073.

Tyson RK (1991) *Principles of Adaptive Optics*. San Diego, CA: Academic Press.

Youla DC and Webb H (1982) Image restoration by the method of convex projections: Part 1 – Theory. *IEEE Transactions on Medical Imaging* M1-1: 81–94.

INCOHERENT SOURCES

Contents

Lamps
Synchrotrons

Lamps

V Roberts, Roberts Research and Consulting, Inc., Burnt Hills, NY, USA

© 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

Introduction

Electric lamps are by far the most common source of light used on Earth, other than our sun. These sources are composed of a very large number of individual radiating elements, all acting independently. This independence causes the light generated by electric lamps to have a very low degree of spatial and temporal coherence. When compared to lasers, lamps are usually considered to be incoherent sources.

Electric lamps are widely used for general lighting, such as residential, commercial and industrial space, and task lighting. A much less frequent use of incoherent sources, but perhaps of more interest to the reader, is as wavelength, intensity, and total light output standard used to calibrate various coherent light sources. We will discuss the general characteristics of incandescent, gas discharge, and solid state lamps, with sources in the Further Reading section at the end of this article, which are used as calibration standards for coherent sources.

Definitions

Before discussing individual lamp technologies, we need to define some of the key terms that are used for describing the performance of electric lamps. Most of these definitions are based on or abstracted from the IESNA Lighting Handbook:

- *Efficacy*: The total light output of a lamp, measured in lumens, divided by the power input in watts. Efficacy has the units of lumens per watt, or $\mathrm{l\,m\,W^{-1}}$, and can be greater than 100. The efficacy of common light sources varies from $10\,\mathrm{l\,m\,W^{-1}}$ to over $200\,\mathrm{l\,m\,W^{-1}}$.
- *Correlated Color Temperature* (*of a light source*) (*CCT*): the absolute temperature of a blackbody whose chromaticity most nearly resembles that of the light source. CCT is measured in Kelvin (K). The CCT of typical light sources varies from 2000 K to 6000 K, with the most popular sources being in the range from 2700 K to perhaps 4500 K.
- *Color Rendition Index* (*of a light source*) (*CRI*): a measure of the degree of color shift objects undergo when illuminated by the light source as compared with those same objects when illuminated by a reference source of comparable color temperature. The highest possible CRI is 100. The lowest theoretical CRI is less than zero, but CRI less than 50 is generally considered to have no value in determining the color quality of a light source.
- *Photopic Luminous Efficiency Function* (V_λ): a measure of the sensitivity of the human eye in lumens per watt of radiation received, as a function of wavelength, for luminance levels higher than $\sim 3\,\mathrm{cd\,m^{-2}}$. V_λ is based on relatively few observations, and individuals may differ significantly from the standard V_λ function. V_λ provides the basis for measuring the efficacy of all light sources.
- *Visible Light*: generally accepted to mean radiation between 380 nm and 780 nm.

Incandescent Lamps

Incandescent lamps are quasi-blackbody thermal radiators. A true blackbody radiator would generate light according to Planck's Law, as shown in eqn [1] and Figure 1, for various temperatures. Most incandescent lamps use tungsten filaments, and the emissivity of tungsten, the ratio of its radiative output to that of a blackbody at the same temperature, is about 0.44 in the visible region of the spectrum and about 0.33 in the infrared. The radiative output of a tungsten incandescent lamp is, therefore, less than half that of a blackbody of the same temperature and the intensity versus wavelength curve does not exactly follow Plank's law, being slightly favorable to visible

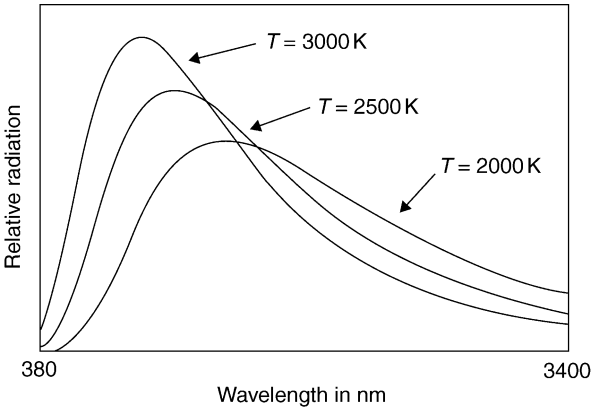


Figure 1 Radiation from blackbody source at various temperatures. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

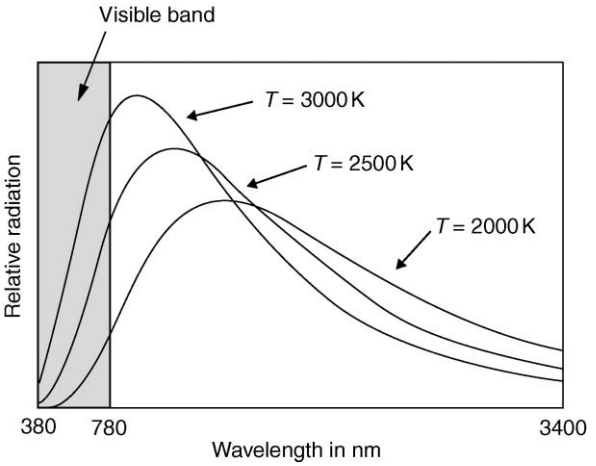


Figure 2 Radiation from blackbody source at various temperatures with visible band. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

radiation:

$$E(\lambda) = \frac{2\pi hc^2 10^{-9}}{\lambda^5(\exp^{hc/k\lambda T} - 1)} \qquad [1]$$

where:

E = power per square meter of surface area per nm of wavelength
λ = wavelength in meters
h = Plank's constant, 6.63×10^{-34} Joule seconds
c = speed of light, 3×10^{8} meters sec^{-1}
κ = Boltzman's constant, 1.38×10^{-23} Joule K^{-1}
T = temperature of the source in K

Tungsten lamps have excellent color quality, reflected in a perfect CRI of 100, are inexpensive, available in a great variety of sizes, shapes, and operating voltages, have a small source size that permits excellent light control with relatively small reflectors and lenses, run equally well on AC or DC power, can be easily dimmed, can have very long life, are insensitive to operating position, are relatively insensitive to ambient temperature, contain no hazardous materials, and light virtually instantly, though not as fast as LEDs. However, incandescent lamps have one major disadvantage; they are very inefficient. Most of the energy consumed by incandescent lamps is radiated in the infrared, while only 5% to 10% is radiated in the visible portion of the spectrum. The reason for this is shown in Figure 2, which shows the visible region of the spectrum superimposed on the blackbody radiation curves.

There are a number of techniques that have been used or are being investigated to increase the efficacy of incandescent lamps. First, as can be seen in Figure 2, increasing the filament temperature shifts the wavelength for peak emission toward lower wavelengths, and will result in a larger fraction of the radiated energy being in the visible region. Unfortunately, increasing filament temperature also increases the rate of tungsten evaporation and therefore shortens the life of tungsten lamps, since evaporation of tungsten is the dominant incandescent lamp failure mode. In a similar manner, the life of incandescent lamps can be increased to arbitrarily long times by decreasing filament temperature. This technique, which is the means used in far too many 'long life' incandescent lamps, unfortunately leads to a decrease in the already low efficacy of incandescent lamps. The approximate dependences of life, efficacy, and color temperature on operating voltage, for common incandescent lamps operating between 90% to 110% of rated voltage, are shown in eqns [2], [3], and [4]:

$$\frac{\text{LIFE}_1}{\text{LIFE}_2} = \left(\frac{\text{VOLTS}_2}{\text{VOLTS}_1}\right)^{13} \qquad [2]$$

$$\frac{\text{EFFICACY}_1}{\text{EFFICACY}_2} = \left(\frac{\text{VOLTS}_1}{\text{VOLTS}_2}\right)^{1.9} \qquad [3]$$

$$\frac{\text{COLOR\_TEMPERATURE}_1}{\text{COLOR\_TEMPERATURE}_2} = \left(\frac{\text{VOLTS}_1}{\text{VOLTS}_2}\right)^{0.42} \qquad [4]$$

Combining eqns [2] and [3] yields eqn [5]:

$$\frac{\text{LIFE}_1}{\text{LIFE}_2} = \left(\frac{\text{EFFICACY}_2}{\text{EFFACACY}_1}\right)^{6.84} \qquad [5]$$

Under the assumption that the color temperature of an incandescent lamp is approximately the same as the operating temperature of the filament, we can combine eqns [2] and [4] to obtain eqn [6], and combine eqns [3] and [4] to obtain eqn [7]:

$$\frac{\text{LIFE}_1}{\text{LIFE}_2} = \left(\frac{\text{TEMPERATURE}_2}{\text{TEMPERATURE}_1}\right)^{30.95} \quad [6]$$

$$\frac{\text{EFFACACY}_1}{\text{EFFICACY}_2} = \left(\frac{\text{TEMPERATURE}_1}{\text{TEMPERATURE}_2}\right)^{4.52} \quad [7]$$

One way to allow operation with higher filament temperatures, while not decreasing filament life, is to fill the lamp with a high molecular weight and/or high pressure gas, as either will retard tungsten evaporation. Lamps using krypton or xenon in place of the normal argon fill allow for a moderate increase in filament temperature. Better yet is a fill of halogen gas, such as HBr, CH_3Br, or CH_2Br^2 plus Kr or Xe at total pressures that range from 2 to 15 atmospheres. The high total pressure inhibits tungsten evaporation from the filament, and the halogen fill gas reacts with tungsten that deposits on the wall of the lamp to remove this tungsten and maintain light output over the life of the lamp. The tungsten removed from the wall of the lamp is redeposited on the tungsten filament, increasing lamp life under some circumstances. Due to the high gas pressures involved, tungsten halogen incandescent lamps are constructed in small 'filament tubes' to minimize both the amount of gas needed and the total explosive energy. The filament tubes are manufactured from quartz or special high temperature glass. For general purpose applications, the quartz or high temperature glass filament tube is enclosed in a larger glass envelope to protect the user should the filament tube explode and to protect the filament tube from environmental contaminants, such as salts from human fingers.

Since 90% to 95% of the energy consumed by an incandescent lamp is radiated as infrared energy, it is easy to see that the potential exists to dramatically increase the efficacy of incandescent lamps by reflecting some of this infrared energy back onto the filament, thus allowing the filament to be maintained at its normal operating temperature with far less electrical power. The reflector must have very low absorption in the visible region of the spectrum, high reflectivity in the IR and a reasonably sharp transition between the two, because the greatest IR emission is just below the visible band. The reflector and filament must also be precisely positioned so that the reflected IR radiation hits the filament, and the reflector must also be low enough in cost to be commercially viable. Successful IR reflectors have been made from multilayer dielectric films. These are similar to the coatings used to make high reflectivity mirrors for He–Ne lasers, but they have much broader bandwidth and lower reflectivity. The reflector-filament alignment problem is most easily solved using the geometry of a tungsten halogen filament tube as the carrier for the IR reflector. The typical efficacy gain is about 50%: a 60-watt IR-halogen lamp produces about as much light as a 90-watt halogen lamp.

Other methods being investigated to increase the efficacy of incandescent lamps involve modification of the filament to decrease its emissivity in the IR and/or increase its emissivity in the visible, and hence turn it into more of a 'selective radiator' than natural tungsten. Three general methods are under investigation: one involves creating a dense pattern of submicron size cavities on the surface of the tungsten to inhibit IR radiation; a second involves coating the tungsten with a material that is itself a selective radiator that favors the visible portion of the spectrum, while a third involves replacing the tungsten filament with a material that is a natural selective radiator. No products exist that employ these techniques, but active R&D projects are underway that may yield a more efficient incandescent lamp in the future.

Incandescent lamps can be used as standards for total flux and color temperature if carefully constructed and operated. The National Institutes for Standards and Technology (NIST) in the US sells 1000-watt tungsten halogen lamps calibrated to an accuracy of 0.6% in luminous intensity and 8 K in color temperature at a color temperature of 2856 K. Similar service are provided by standards organizations in other countries.

Discharge Lamps

The most efficient lamps produced today are based on electric discharges in gasses. These include 'low pressure' lamps such as fluorescent lamps and low pressure sodium lamps; and 'high pressure' or high intensity discharge (HID) lamps such as metal halide, high pressure sodium, high pressure mercury, and sulfur lamps. These lamps are all more efficient than incandescent lamps because they are 'selective radiators' instead of blackbody sources and have been designed to generate as much radiant energy as possible in the visible band while minimizing energy produced outside the visible band. The best white light discharge lamps have conversion efficiencies from electric power to visible light of 25% to 30%, making them almost five times as efficient as the typical incandescent lamp.

Fluorescent Lamps

Fluorescent lamps, as shown in Figure 3, are generally constructed in a cylindrical soda lime glass envelope with a length to diameter ratio significant greater than one. Most contain thermionic electrodes at each end to generate an electric field that establishes and sustains the discharge. The lamps are filled with about 2 Torr of Argon and/or another rare gas, and a few mg of liquid mercury, while the inside surface of the tube is coated with a phosphor that converts UV radiation generated by excited mercury atoms into visible light.

As with any other low pressure gas discharge, various discharge regions are aligned along the axis of the tube:

- Aston Dark Space
- Cathode Glow
- Cathode Dark Space
- Negative Glow
- Faraday Dark Space
- Positive Column
- Anode Glow
- Anode Dark Space

The axial lengths of all these regions, except for the positive column, are fixed by the properties of the cathode and anode, the fill gas type and pressure, and the discharge current. The length of the positive column is variable and fills the space between the Faraday Dark Space and the Anode Glow. The cathode and anode regions of the discharge have a total axial length of about twice the diameter of the discharge tube. Therefore, if the length-to-diameter ratio of the discharge tube is substantially greater than one, the majority of the tube is filled by the Positive Column, as shown in Figure 3. The positive column is the part of the discharge that creates the UV photons that excite the phosphor and cause it to generate visible light. The remainder of this discussion will therefore focus on the positive column.

The basic operation of a fluorescent lamp positive is shown in Figure 4. Free electrons are accelerated by the electric field established between the two electrodes. These electrons collide with both the mercury and rare gas atoms, but selectively excite and ionize mercury atoms, since the excitation and ionization energies of mercury are significantly lower than that

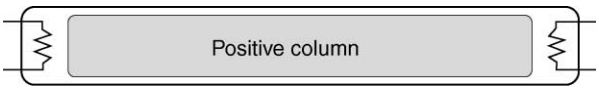


Figure 3 Linear fluorescent lamp. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

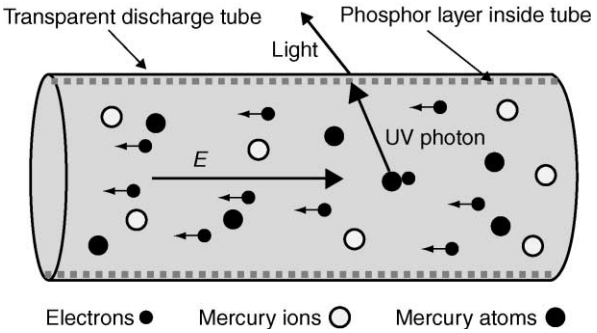


Figure 4 Operation of fluorescent lamp positive column. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

of any rare gas used in the lamps. Ionization of mercury creates more free electrons to sustain the discharge. The gas density in the lamp is low enough to require that most of the recombination of mercury ions and free electrons takes place on the wall of the tube, since this recombination requires a three-body collision. The excited mercury atoms generate both UV and visible photons, and the UV photons generate visible light once they reach the phosphor coated on the inner surface of the lamp.

The high efficiency of fluorescent lamps is directly related to the fact that 60% to 65% of the electric power consumed by the positive column of the fluorescent lamp discharge is converted into 245 nm UV photons as a result of the resonance transition from the 6^3P_1 state of mercury at 4.86 eV to the 6^1S_0 ground state. The resonance transition from the 6^1P_1 state of mercury generates 185 nm UV photons that carry about 10% of the energy. About 3% of the positive column energy ends up in mercury visible lines at 405 nm, 436 nm, 546 nm, 577 nm, and 579 nm.

The 245 nm and 185 nm photons generated by the excited mercury atoms travel only a fraction of a mm before they are absorbed by a ground-state mercury atom. However, due to the relative absence of other de-excitation mechanisms, a large fraction of these absorbed photons are re-emitted by the absorbing atoms, and most of the UV photons generated in the core of the discharge eventually reach the phosphor coated interior wall of the discharge tube where they are converted into visible light (and waste heat).

The phosphors used in modern fluorescent lamps are a blend of three rare-earth phosphors:

- Blue: $BaMg_2Al_{16}O_{27}:Eu^{2+}$
- Green: $CeMgAl_{11}O_{19}(Ce^{3+}):Tb^{3+}$
- Red: $Y_2O_3:Eu^{3+}$

Rare earth phosphors provide a combination of high efficiency, high color rendition index, and high resistance to degradation that far exceeds the performance of the halophosphate phosphors used in older generations of fluorescent lamps.

However, due to the Stokes Shift, there is a substantial energy loss converting the UV photons produced by excited mercury atoms into visible light, even using a 'perfect' phosphor that has quantum efficiency of 1.0. The average energy of a visible photon is 2.43 eV, using the 380 nm to 780 nm limits for visible radiation discussed at the start of this article. When one 245 nm photon that has 4.885 eV of energy is converted to one visible photon with an average energy of 2.43 eV, a bit over 50% of the photon's energy is lost. The situation is worse for the 185 nm photons, which each have 6.707 eV of energy. They lose 64% of their energy when they are converted to visible photons, so it is good that the fluorescent lamp discharge generates six to seven 245 nm photons for each 185 nm photon. The Stokes shift is responsible for the largest single energy loss mechanism in fluorescent lamps. Work is underway to develop phosphors with a quantum efficiency greater than 1.0, so less energy will be lost in the UV-to-visible light conversion process, but no practical phosphors have been developed to date.

The mercury pressure for optimum UV production is about 6×10^{-3} Torr, which is the vapor pressure of mercury over a liquid pool at 40 °C. At higher mercury pressure there is too much absorption by ground state mercury atoms, and at lower pressures there is an insufficient number of mercury atoms for excitation by the electrons. The diameter of most linear lamps is chosen so that the coolest spot on the interior of the lamp wall will be at 40 °C when the lamp is operated at its design power in an ambient temperature of 25 °C. Usually this 'cold spot' will be near the middle of the lamp, but in some higher power linear lamps the wall temperature is well over 40 °C. In these lamps a cold spot is often created at the end of the lamp by use of a longer electrode mount and a heat shield mounted between the electrode and the end of the lamp. Obviously, the performance of fluorescent lamps designed in this manner is highly sensitive to ambient temperature. In some smaller, high-power density lamps, most commonly compact fluorescent lamps (CFLs), it is not possible to create a region with a temperature as low as 40 °C. In these lamps the mercury pressure is controlled by amalgamating the mercury with Pb, Bi, or Sn, creating a compound that both reduces the mercury vapor pressure at any given temperature and also reduces the variability of mercury vapor pressure with temperature. These mercury amalgams have the secondary advantage of making the fluorescent lamp less sensitive to ambient temperature.

Mercury, which is the key ingredient in fluorescent lamps, is considered to be a hazardous material by regulatory agencies in many countries. At least as early as the mid-1980s researchers began to study ways to reduce or eliminate mercury from fluorescent lamps.

Development of a mercury-free fluorescent lamp is very difficult and the only mercury-free fluorescent lamp that exists, as of mid-2003 has an efficacy of about $10\,\mathrm{l\,mW^{-1}}$, one-tenth the efficacy of the best mercury-based fluorescent lamp (Osram Sylvania LINEX® Linear Excimer Lamp System, Model LINEX A3-10W40). There are three key problems that have to be solved to develop an efficient mercury-free fluorescent lamp system:

(i) Identify a nontoxic material that generates UV photons and has its optimum gas density at a convenient and safe temperature. The most promising candidate is xenon, used as either an atomic radiator or an excimer.

(ii) Develop a phosphor that produces high-quality white light with high efficiency when excited by UV photons from the chosen mercury-replacement material. This has proven a major challenge. The resonance line of atomic Xe has a wavelength of 147 nm, while the output from a xenon excimer discharge is at 172 nm. The Stokes shift loss converting each of these to visible light is 71% for atomic xenon and 66% for xenon excimers. To reduce the Stokes shift loss, a quantum-splitting phosphor, one that generates more than one visible photon for each UV photon, is needed. Most of the effort to develop mercury-free fluorescent lamps has been concentrated in this area. No viable phosphor has been developed, but work continues.

(iii) Develop a 'reservoir' for xenon or the other working gas. The mass of 'active' gas needed for a fluorescent lamp is small; a 4-foot, 1.5-inch diameter lamp needs 0.1 mg of mercury in the gas phase to operate properly. However, during life, some of this gas is consumed through physical attachment or chemical reaction and must be replaced. The drop of liquid mercury or small mass of mercury amalgam used in conventional fluorescent lamps provides a reservoir that maintains the mercury gas pressure in the presence of these consumption mechanisms.

In contrast to the failure to develop a mercury-free fluorescent lamp, there has been great success in

reducing the amount of mercury used in each lamp. A typical 4-foot, 1.5-inch diameter fluorescent lamp produced in the 1970s would have 50 mg to 100 mg of mercury. This high amount of mercury was mostly the result of the type of manufacturing equipment used, but mercury consumption required that each lamp had 15 mg to 20 mg of mercury in order to survive for its rated 20 000-hour life. The development of coatings and other materials that reduce mercury consumption, when combined with new manufacturing technology, has allowed the amount of mercury used in modern 4-foot fluorescent lamps to be reduced to 3 mg to 5 mg, and further reductions are possible with improved technology.

As stated above, the length-to-diameter ratio of most fluorescent lamps is significantly greater than one. The reason for this is that there is a 10 to 15 volt drop at the electrodes, and the power lost in these regions does not produce a significant amount of light. A large length-to-diameter ratio increases the voltage and power of the positive column, where the bulk of the UV photons are generated, relative to the power lost in the electrode regions. When the energy crunch hit in the mid-1970s, there was strong interest in developing a fluorescent lamp that could replace the inefficient incandescent lamp. However, the need for high length-to-diameter ratios meant that it would not be practical to make a fluorescent lamp shaped like an incandescent lamp. This need created the technology to manufacture CFLs which are constructed of long tubes bent or coiled into shapes that allow them to fit into spaces occupied by incandescent lamps.

Fluorescent lamps and other discharge lamps have a negative incremental voltage vs. current (V/I) characteristic, which means that their operating voltage decreases as the operating current increases. Due to this, such discharges need to be operated from a current limited power source, commonly known as a 'ballast' in the lighting industry. Ballasts were originally simple series inductors and lamps operated at the power line or mains frequency, though many line frequency ballasts are now considerable more complex than an inductor. In the 1940s, it was discovered that the efficacy of fluorescent lamps could be increased by operation at frequencies of a few kHz. It was not until the 1970s, however, until the cost of power electronics came down to the point where it was possible to make affordable high frequency electronic ballasts. Electronic ballasts operating in the 20 kHz to 100 kHz range are now used with almost all CFLs and in most commercial applications using linear fluorescent lamps.

The desire for fluorescent lamps shaped like incandescent lamps also sparked the development of electrodeless fluorescent lamps. Since the need for high length-to-diameter ratios is driven by the losses at the lamp electrodes, researchers realized they could make fluorescent lamps in short, wide shapes if they eliminated the electrodes. As shown in Figure 5, an electrodeless lamp is just like a transformer, except it has a single-turn discharge secondary instead of a multiturn copper or aluminum secondary. An open core version of an electrodeless fluorescent lamp with integral ballast is shown in Figure 6. The electric field is created by a time-varying magnetic field. Because of the relative high voltage-per-turn imposed by the discharge secondary, these lamps operate only at high frequency. There are three electrodeless fluorescent lamps on the market. The closed core design

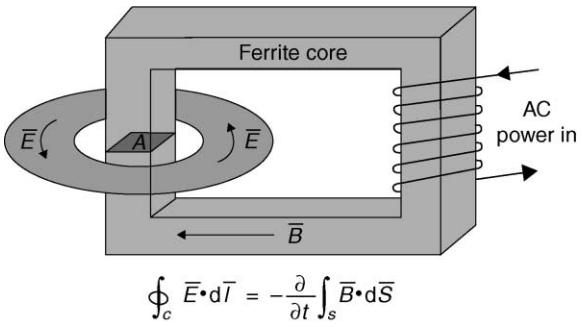


Figure 5 Basic electrodeless fluorescent lamp with closed ferrite core. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

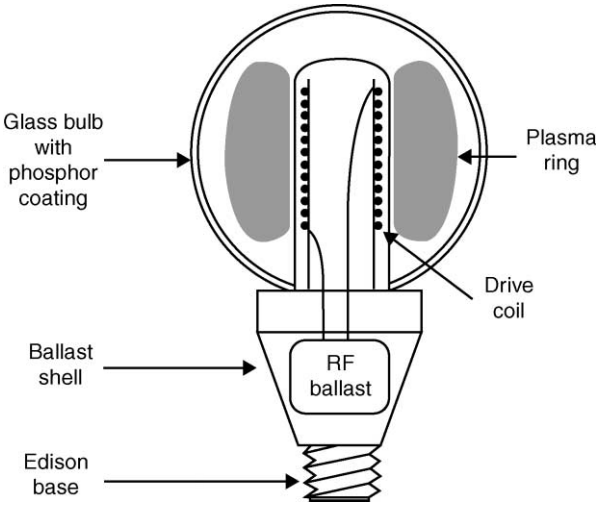


Figure 6 Open core electrodeless lamp with integral ballast. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

shown in Figure 7 operates near 250 kHz with an external ballast, while the open core design shown in Figure 8 operates near 2.5 MHz. Figure 9 shows an open core design with an integral ballast that also operates near 2.5 MHz. Some people have suggested that electrodeless fluorescent lamps are fundamentally different from normal fluorescent lamps and the high frequency field drives the phosphor to emit light. This is incorrect. Other than the manner in which the electric field is created, and the new shapes this allows, electrodeless fluorescent lamps operate in the same manner as normal fluorescent lamps.

Due to the strong dependence of fluorescent performance on mercury temperature, these lamps are not useful as flux standards. However, the five mercury visible lines listed above, plus the 365 nm line of mercury, make the ordinary fluorescent lamp an ideal source for wavelength calibration of spectrometers. Germicidal lamps, which do not use phosphor but are constructed in quartz envelopes and use the same type of mercury-rare gas discharge as fluorescent lamps, are a convenient source of 254 nm radiation that can be used to calibrate UV spectrometers. Care should be taken when using these lamps, as the 254 nm radiation can damage human eyes and skin.

High Intensity Discharge Lamps

If the mercury pressure in a discharge lamp is raised from 6×10^{-3} Torr to an atmosphere or

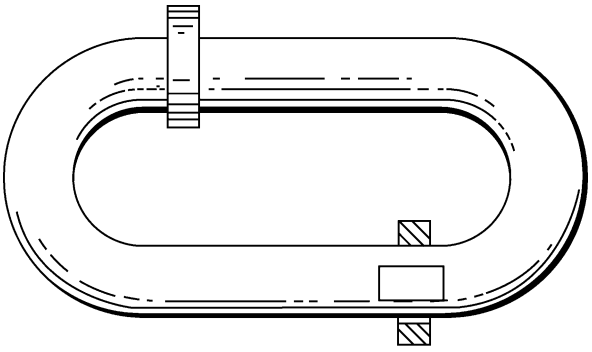


Figure 7 Closed core electrodeless fluorescent lamp, Osram Sylvania Icetron/Endura Lamp, from US Patent 5,834,905.

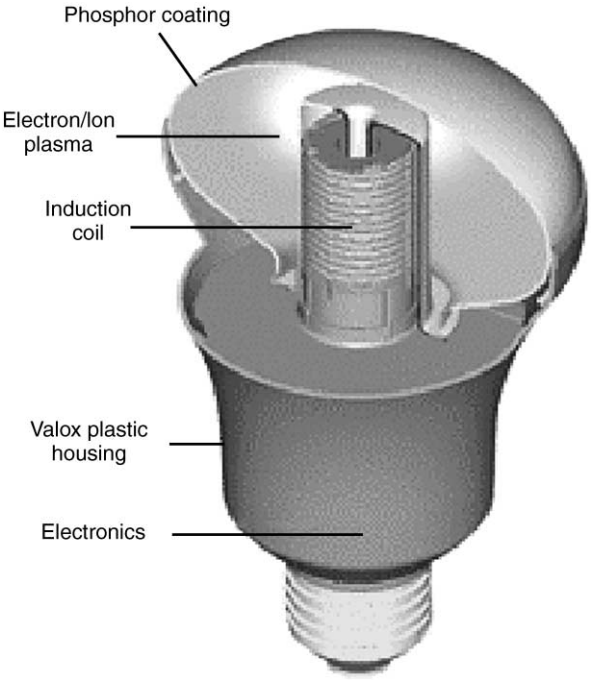


Figure 9 GE Genura® open core electrodeless fluorescent lamp with integral ballast. Courtesy of General Electric Company.

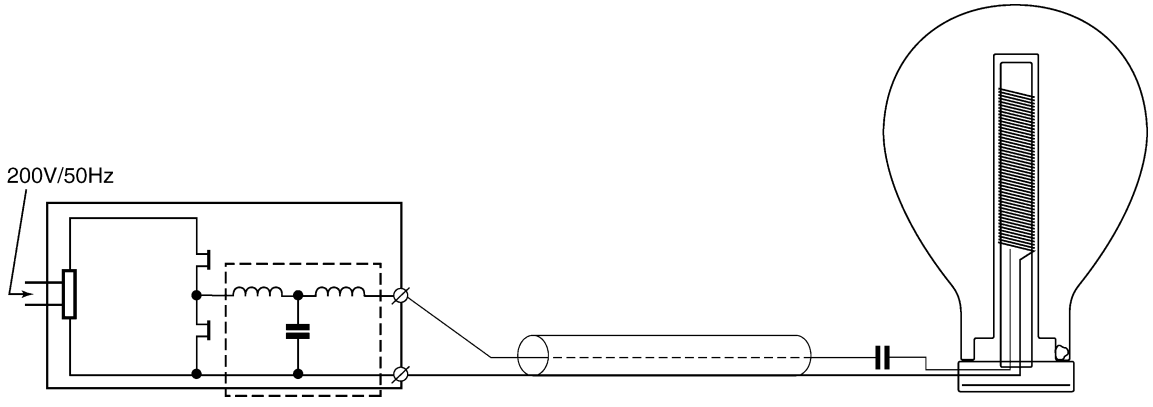


Figure 8 Open core electrodeless fluorescent lamp with separate ballast, Philips QL lamp, from US Patent 6,373,198 B1.

greater, the 245 nm and 185 nm resonance lines are trapped via absorption by ground-state mercury atoms and the radiation output from the discharge shifts to the five optically thin visible lines listed in the discussion of fluorescent lamps, plus the 365 nm near-UV line and a number of lines near 300 nm. The result is a high pressure mercury lamp that operates with an efficacy of about $55\,\mathrm{l\,m\,W^{-1}}$ and has a CRI of about 50. The high-mercury vapor pressure requires a wall temperature of at least 360 °C, which in turn requires that the arc tube be constructed from quartz and operated at high power density. A 400-watt high-pressure mercury arc tube has a diameter of 22 mm and length of about 100 mm. In contrast, a 32-watt fluorescent lamp has a diameter of 25.4 mm and a length of about 1200 mm. This accounts for the classification of these high pressure lamps as high intensity discharge (HID) lamps. Due to the high temperature of the arc tube, most HID lamps are constructed with the arc tube mounted inside a hard glass outer jacket that protects both the user and the arc tube. The high-pressure mercury lamp is a direct visible radiator, but the arc tube does generate a small amount of radiation in the 254 nm to 365 nm region, that radiation is trapped by the outer glass jacket. Some high-pressure mercury lamps have a phosphor coating on the inside of the outer jacket to convert this UV radiation into visible radiation, mostly in regions of the spectrum where the high-pressure mercury arc is deficient in radiation, thus improving the color of the lamp.

Most high-pressure mercury lamps operate at a mercury pressure of about four atmospheres. Because their spectrum is dominated by mercury line emission, the output of these lamps is deficient in the red portion of the spectrum and the lamps have poor color rendition. In 1992, Philips received a US patent for a new type of ultra-high pressure mercury lamp that they call UHP for 'Ultra High Performance'. This lamp operates at a mercury pressure of 200 atmospheres or more. The output spectrum is characterized by a significant amount of continuum radiation from quasi-molecular states of mercury and this continuum includes energy in the red portion of the spectrum that is missing from conventional high pressure mercury lamps. UHP lamps also include a small amount of halogen, such as bromine, which creates a tungsten-halogen cycle similar to that discussed earlier in relation to tungsten-halogen incandescent lamps. During lamp operation, tungsten is sputtered and evaporated from the electrodes. In the absence of the halogen fill, this tungsten rapidly deposits on the walls, leading to very short lamp life. The halogen combines with most of the tungsten before it has a chance to reach the wall of the arc tube. The halogen will also combine with any tungsten that has deposited on the wall of the arc tube. The tungsten-halogen compound formed is gaseous at the operating temperature of the lamp and will diffuse back toward the core of the discharge where the high temperature of the arc dissociates the tungsten-halogen compound. The tungsten is deposited back on the electrodes, thereby greatly extending the life of the lamp. Due to their very high pressures, UHP lamps are constructed in small discharge tubes and have a very short discharge length, typically 1 to 2 mm, making these lamps a virtual point source. Due to their excellent spectral characteristics and small source size, UHP lamps have found application in various projection applications, including LCD-based electronic projectors.

The performance of high-pressure mercury lamps can be substantially improved by adding metals such as sodium and scandium to the high-pressure mercury arc. The sodium emits near the peak of the eye sensitivity curve, providing high efficacy, while the scandium has many transitions in the visible, which fill out the spectrum and provide high CRI. If metallic elements are added directly to the high-pressure mercury arc they would deposit on or react with the wall of the arc tube. To prevent this they are added as iodides, such as NaI or ScI_3. This creates a regenerative cycle in which the NaI and ScI_3 are heated and dissociated in the 5000 K core of the arc, allowing the free metals to radiate. As the metals diffuse toward the cooler sections of the arc near the walls, they recombine with the iodine to reform the iodides. These metal halide lamps have produce light with an efficacy of about $90\,\mathrm{l\,m\,W^{-1}}$ to $100\,\mathrm{l\,m\,W^{-1}}$ and a CRI of about 65 to 75.

The high-pressure sodium (HPS) lamp is an HID lamp that uses hot sodium vapor to generate a golden white light. The lamp uses a mixture of sodium, mercury, and rare gas such as xenon. The core temperature of the sodium-mercury-xenon discharge is typically about 4000 K, while the temperature at the edge of the discharge and the walls of the ceramic arc tube is typically about 1500 K. At low pressure, emission from sodium is dominated by the resonance lines at 589.0 and 589.6 nm, so the light is virtually monochromatic. However, in the HPS lamp, the sodium pressure is typically 75 Torr, which is high enough to almost completely trap emission from the resonance lines, forcing radiant energy to escape in the wings of

these lines, thereby increasing the color quality of the light. Since sodium will attack quartz, this mixture is contained in a translucent alumina (aluminum oxide) ceramic arc tube. As with other HID lamps, the alumina arc tube is contained in a hard glass outer bulb. HPS lamps are the most efficient generators of 'white light' available. A typical 400 watt HPS lamp produces light with an efficacy of $125\,\mathrm{l\,m\,W^{-1}}$, this high efficacy being due to the fact that the emission matches the peak of the eye sensitivity curve. The light has a CRI of only 22; however, the HPS lamp is widely used for roadway and other outdoor lighting applications.

Solid State Sources

The newest incoherent sources to be used for general purpose lighting are solid state devices: light emitting diodes (LEDs), and organic light emitting diodes (OLEDs). LEDs are very similar to laser diodes but without the two mirrors used to form the laser cavity in laser diods. Photons are generated as a result of electron-hole recombination as electrons move across the energy gap of a biased p-n junction. By appropriate selection of materials, photons can be created at visible wavelengths. Low brightness infrared and red LEDs have been available for quite some time. It was not until the mid-1980s that high brightness red LEDs were developed. These were followed by increasingly more challenging yellow, green, and then blue high brightness LEDs, developed in 1994 by Shuji Nakamura, of Nichia Chemical Industries. Nakamura added phosphor to his high brightness blue LEDs in 1996 and developed the first 'white' LED, thus triggering the current level of excitement concerning the use of LEDs to replace incandescent and discharge lamps. Monochromatic high brightness LEDs have already replaced color-filtered incandescent lamps in many applications, such as traffic lights, vehicle signal lights, and signage. In these applications, LEDs are the clear winner in life and efficacy. However, when compared to unfiltered sources, the best white LEDs are about as efficient as tungsten halogen incandescent lamps and far less efficient than fluorescent and metal halide lamps. They also cost far more per lumen than even the most expensive discharge lamp. LEDs have high internal quantum efficiency, so the main technical challenges are to fabricate devices to minimize internal light trapping in order to improve device efficiency and reduce the cost per lumen produced.

Organic light emitting diodes (OLEDs) generate light by the passage of electric current through special organic molecules, using a process known as electrophosphorescence. The electrophosphorescent material is usually coated on a transparent and perhaps flexible substrate, and overcoated with a transparent conductor. OLEDs currently have lower efficacy and shorter life than LEDs, but they also have lower manufacturing costs. If LEDs are the 'new' point sources that replace incandescent lamps and perhaps CFLs, then OLEDs are the 'new' diffuse sources that some say will replace liner fluorescent lamps.

See also

Light Emitting Diodes.

Further Reading

Cobine JD (1958) *Gaseous Conductors: Theory and Engineering Applications*. New York: Dover Publications.

de Groot JJ and van Vliet JAJM (1986) *The High Pressure Sodium Lamp*. Philips Technical Library. Deventer–Antwerprn: MacMillian Education.

Elenbass W (1951) *The High Pressure Mercury Vapor Discharge*. Amsterdam: North-Holland Publishing Company.

Fisher HE (1996) *High Pressure Mercury Discharge Lamp*, U.S. Patent, 5,497,049.

Fisher HE and Hörster H (1998) *High-Pressure Mercury-Vapor Discharge Lamp*, U.S. Patent 5,109,181.

Godyak VA (2002) Bright Idea: Radio-frequency light sources. *IEEE Industry Applications Magazine* 8: 42–49.

Hirsh MN and Oskam HJ (1978) *Gaseous Electronics, Volume 1, Electrical Discharges*. New York: Academic Press.

Mitchel AGC and Zemansky MW (1971) *Resonance Radiation and Excited Atoms*, 2nd edn. London: Cambridge University Press.

Nichia's Shuji Nakamura: Dream of the Blue Laser Diode, ScienceWatch, January/February 2000. Available at: http://www.sciencewatch.com/jan-feb2000/sw_jan-feb2000_page3.htm

Pekarski P, Hechtfischer U, Heusler G, Koerber A, Moonch H, Retsch JP, Weichmann U and Kitz A (2003) *UHP Lamps for Projection Systems*, XXVI International Conference on Phenomena in Ionized Gasses, Greifswald, Germany. Available at: http://www.icpig.uni-greifswald.de/proceedings/data/Pekarski_1

Raizer YP (1991) *Gas Discharge Physics*. Berlin: Springer-Verlag.

Rea MS (2000) *IESNA Lighting Handbook*, 9th edn. New York: Illuminating Engineering Society of North America.

Roberts VD (2003) *Radio-Frequency Lighting*. McGraw-Hill Yearbook of Science & Technology, pp. 349–352. New York: The McGraw-Hill Companies, Inc.

Schewe PF and Stein B (1999) *Physics News Update*, Number 437 (Story #2), The American Institute of Physics Bulletin of Physics News. Available at: http://www.aip.org/enews/physnews/1999/split/pnu437-2.htm

Waymouth JF (1971) *Electric Discharge Lamps*. Cambridge, Massachusetts: MIT Press.

Wharmby DO (1993) Electrodeless Lamps for Lighting: A Review. *IEE Proceedings-A* 140(6): 465–473.

Synchrotrons

R Clarke, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Since their discovery by Röntgen in 1895, X-rays have been used in a wide variety of applications ranging from medical and industrial radiography to crystallographic studies of the structure of biomolecules. The short wavelength ($\lambda \sim 0.1$ nm) and high penetrating power of hard X-rays makes them an ideal probe of the internal structure and composition of condensed matter. Indeed, Watson and Crick's elucidation of the double helix structure of DNA in the early 1950s, based on the X-ray diffraction patterns of Rosalind Franklin, has revolutionized biology and provided an atomistic view of the stuff of life.

The quest for ever higher spatial resolution, and the desire to obtain high-quality X-ray data on shorter time-scales, have both been prime motivations for the development of high intensity hard X-ray sources since the mid-1970s. This evolutionary trend has led to the design and construction of large synchrotron radiation facilities, serving many thousands of users and enabling forefront experimental studies across a broad range of disciplines. One such example of a national facility, dedicated to high-brightness X-ray synchrotron radiation, is the Advanced Photon Source (APS), located at Argonne National Laboratory in the United States (**Figure 1**). Facilities of a similar scope and purpose have also been constructed in Europe and Japan, and a number of new high brightness sources are currently under construction in several countries around the world (see **Table 1**).

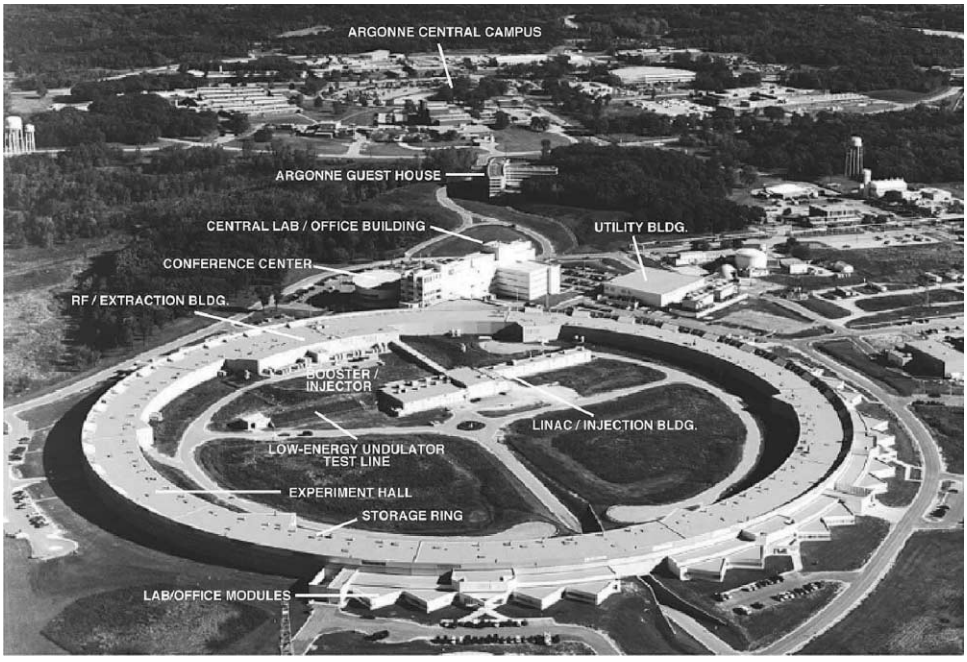


Figure 1 Aerial view of the Advanced Photon Source, Argonne National Laboratory. Courtesy of Advanced Photon Source.

Table 1 Third-generation synchrotron radiation sources

Location	*Facility*	*Electron energy (GeV)*
Grenoble, France	European Synchrotron Radiation Facility (ESRF)	6
Argonne, Illinois, USA	Advanced Photon Source (APS)	7
Nishi Harima, Japan	8-GeV Super Photon Ring (SPring-8)	8
Berkeley, CA, USA	Advanced Light Source (ALS)	1.5–1.9
Canberra, Australia	Boomerang[a]	3
Didcot, UK	DIAMOND[a]	3
Orsay, France	SOLEIL (LURE)	2.75
Karlsruhe, Germany	Angströmquelle-Karlsruhe (ANKA)	2.5
Hamburg, Germany	PETRA III[b]	6
Indore, India	INDUS II[a]	2.5
Villigen, Switzerland	Swiss Light Source (SLS)	2.4
Saskatoon, Canada	Canadian Light Source[a]	2.9
Stanford, CA, USA	SPEARS	3
Upton, New York, USA	NSLS II[b]	2.5–2.8
Trieste, Italy	Elettra	2.0–2.4

[a]Under construction as of March 2004.
[b]Planning stage.

In this article we will draw on examples from APS to illustrate the characteristics and performance of the current generation of synchrotron sources in the hard X-ray region.

From the earliest days of particle physics studies in the 1930s and 40s, it was realized that copious amounts of X-rays could be produced by accelerating charged particles in a curved trajectory, as in a cyclotron or a synchrotron. For particle physics such radiation is highly detrimental because it limits the ultimate energy of particle accelerators and forces their circumference to become inordinately large in order to reduce radiation losses.

Originally considered to be a major source of concern in particle accelerator design, synchrotron radiation soon began to emerge as a boon to X-ray science. This came about in the early 1970s after the capability was developed to produce and store significant circulating currents of high-energy charged particles (electrons or positrons) for colliding beam experiments. At that time X-ray scientists were tolerated as 'parasitic' experimenters at such storage rings, taking X-ray radiation from the bending magnets which maintained the stored current in an overall circular trajectory. The PETRA (Positron–Electron Tandem Ring Accelerator) facility at HASYLAB/DESY in Hamburg, Germany, and the Cornell Electron Storage Ring (CESR)/Cornell High-Energy Synchrotron Source (CHESS) at Cornell University, Ithaca, New York, are examples of accelerator facilities that were primarily intended for particle physics experiments and then became partly dedicated to X-ray production. Such shared synchrotron facilities are referred to as 'first-generation sources'.

The first generation sources demonstrated several attractive features of synchrotron radiation, including a high degree of collimation in the vertical direction with a divergence half-angle given approximately by the relativistic factor $\gamma^{-1} = m_0c^2/E$, where m_0c^2 is the rest-mass energy (0.51 MeV) of the electron (or positron) and E is the energy of the stored particle beam. For a bending magnet source with $E = 2-3$ GeV the beam would have a vertical collimation of ~ 0.3 mrad (17 mdeg). This would produce an X-ray beam with a vertical extent of ~ 3 mm at a distance of 10 m from the bending magnet source. The horizontal fan of bend magnet radiation is broad and is usually limited by a mask or a slit to a few mrad. The most useful characteristic of bend magnet radiation is its large bandwidth whereby the X-ray spectrum is continuous and increasing in intensity (see **Figure 2**) until some critical energy above which the intensity begins to fall off. Bending magnet radiation is therefore useful for X-ray measurements requiring a 'white' beam, for example, Laue X-ray diffraction, energy dispersive diffraction, and Extended X-ray Absorption Fine Structure (EXAFS) analysis.

Beginning in the 1980s, several second-generation storage rings were constructed in various locations around the world. The design of these facilities for the first time was optimized for the production of bright beams of X-ray and XUV (soft X-ray to ultraviolet) and their operation was dedicated exclusively to synchrotron radiation science. The increase in brightness provided by these new storage ring sources, compared to standard laboratory X-ray sources (e.g., rotating anode generators), was dramatic (5 or 6 orders of magnitude) and this opened up many new experiments that were not previously feasible. However, it was soon realized that the capabilities of such sources could be extended to

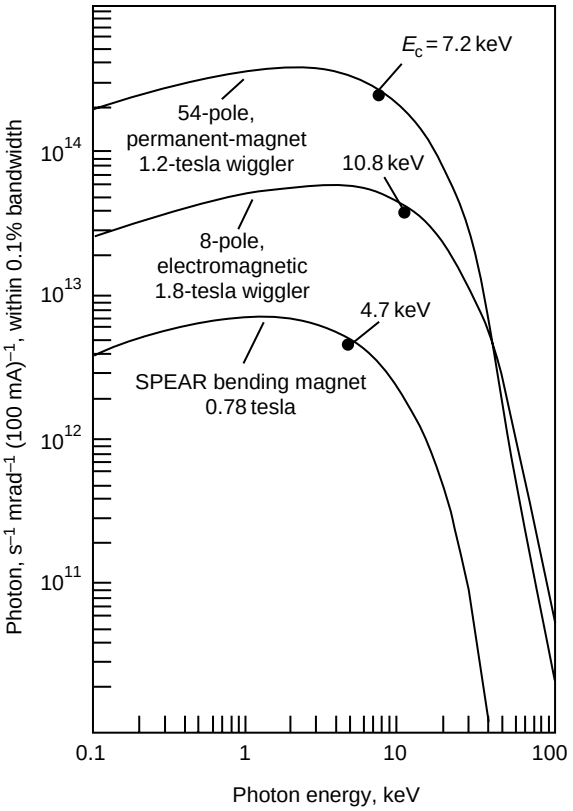


Figure 2 Comparison of bending magnet and wiggler sources. With permission from Stanford Synchrotron Radiation Laboratory, Stanford Linear Accelerator Center, Menlo Park, CA.

even higher levels of intensity and to significantly higher X-ray energies.

One of the limitations of bending magnet sources relates to the X-ray critical energy alluded to above:

$$E_c = 3hc\gamma^3/4\pi\rho \qquad [1]$$

where ρ is the radius of curvature imposed on the particle beam trajectory by the bending magnet. For a given particle acceleration energy, E, eqn [1] leads to a critical X-ray energy given by:

$$E_c\,(\text{keV}) = 2.2\,E^3\,(\text{GeV})/\rho\,(\text{m}) \qquad [2]$$

For a typical second-generation machine with $E = 2.5$ GeV and $\rho = 7$ m (limited by the magnetic field generated by the dipole electromagnet), the critical energy is only about 5 keV. There is useable flux well above this energy, but X-ray photons become scarce in the energy range above 10 keV for a bending magnet source with these characteristics. Since E is effectively fixed by the design parameters of the storage ring, the only possibility to achieve higher photon energies is to install additional magnetic devices which will introduce perturbations of the trajectory with effectively smaller radius of curvature in the denominator of eqn [2]. Such a device, known as a 'wiggler', shifts the effective critical X-ray energy to higher values in proportion to the strength of the magnetic field B (see Figure 2). Using wigglers based on superconducting magnets ($B \sim 5$ Tesla), E_c can be raised to >20 keV, a value which would require accelerator energies of more than 6 GeV using conventional bending magnets alone. In addition to shifting the output spectrum to higher energies, it is possible to enhance the intensity at a given energy by producing multiple wiggles with no net displacement of the particle trajectory. Multipole wigglers are able to produce significant increases in flux in the hard X-ray range. The 27-pole hybrid wiggler installed on beamline X25 at the National Synchrotron Light Source, a second generation facility at Brookhaven National Laboratory, Upton, New York, is one of the most successful designs of its kind, providing $\sim 6 \times 10^{14}$ photons/sec per 1% bandwidth at 8 keV. When focused with a toroidal X-ray mirror, this beam is about an order of magnitude brighter than the standard bending magnet sources at NSLS. For reasons that will become clear later in this article, the radiation from multiple wiggles in such a device adds incoherently.

Insertion Devices

The idea of using wigglers to enhance the characteristics of synchrotron radiation sources has led to the design and construction of storage rings with many straight sections to accommodate wigglers and other types of 'insertion devices'. Such facilities are referred to as third-generation sources. Coming on-line in the mid-1990s, several such sources are now in mature operation and are supporting a large international community of X-ray scientists.

One of the most important types of insertion device to emerge from the development of third-generation facilities is the undulator. Conceptually, the idea behind an undulator is similar to that of the wiggler, i.e., it provides additional tightly curved transverse displacements in the particle beam. There are two important differences: in an undulator the displacements are smaller and they are more numerous, being produced by a linear periodic array of many short alternating dipoles (Figure 3). Under these conditions the synchrotron radiation emitted from each 'undulation' of the particle beam trajectory adds coherently and therefore the peak intensity from such a device can be enhanced by a factor of N^2 where N is the number of dipole magnet pairs, or periods.

A simple theoretical description of the basic operation of an undulator can be discussed in terms of radiation from a charged particle moving at

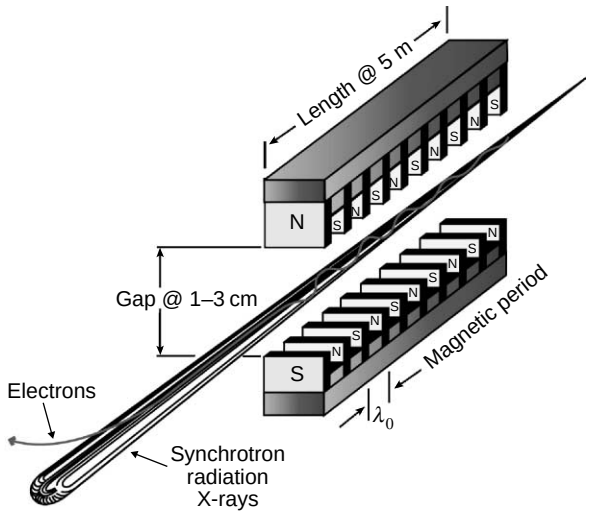


Figure 3 Schematic of undulator. Courtesy of Advanced Photon Source.

relativistic speeds through an alternating magnetic field. The magnetic field in the vertical direction (y) is taken to be periodic:

$$B_y = B_0 \sin[(2\pi/\lambda_u)z] \qquad [3]$$

where z is in the direction of average particle motion.

The transverse particle trajectory is described by:

$$x = (\lambda_u \phi_{max}/2\pi) \sin[(2\pi/\lambda_u)z] \qquad [4]$$

where λ_u is the magnet period, and ϕ_{max} is the maximum particle deflection angle:

$$\phi_{max} = (eB_0\lambda_u c)/2\pi m_0 c^2 \gamma = K/\gamma \qquad [5]$$

where

$$K = 93.4B_0 \text{ (Tesla)}\lambda_u \text{ (meters)} \qquad [6]$$

The field strength parameter K serves as a useful quantity for discussing the characteristics of various insertion device designs. For $K \leq 1$, which defines the undulator regime, the transverse angular deflections are less than or on the order of the angular spread of the emitted radiation and in this case the radiation from each of the deflections adds coherently. This is distinguished from the case of a multipole wiggler where $K \gg 1$ leading to incoherent superposition.

Most of the interesting features of undulator radiation are revealed by considering the trajectory of the particle in a Lorentz frame moving at speed v^* with respect to the laboratory frame. In this frame the particle will execute oscillatory motion with the period of oscillations T' given by the Lorentz transformation:

$$T' = \gamma^*(T - \beta^*\lambda_u/c) = \lambda_u/(v^*\gamma^*) \qquad [7]$$

with a corresponding oscillation frequency $\omega' = 2\pi/T' = (2\pi c/\lambda_u)\gamma^*$.

For $K \ll 1$, the motion in the co-moving frame is purely transverse and nonrelativistic. However, for $K \sim 1$, the oscillatory motion in this frame can have a longitudinal component and in this case the transverse motion becomes nonharmonic with frequency components at odd multiples of the fundamental frequency ω'. This nonharmonic motion is a consequence of the fact that in the laboratory frame the particle is moving with constant speed $\approx c$ but its path length is slightly longer than λ_u, owing to the transverse motion. This constrains the motion in the co-moving frame to be nonharmonic.

Undulator Spectrum

It now remains to transform ω' into the laboratory frame so that we can obtain the spectral distribution of the output of the undulator. Invoking the Doppler transformation:

$$\omega = \gamma^*\omega'(1 + \beta^* \cos\theta') \qquad [8a]$$

$$\tan\theta = (\sin\theta')/[\gamma^*(\cos\theta' + \beta^*)] \qquad [8b]$$

where θ is the angle at which the radiation is observed:

$$\omega_n(\theta) = (4\pi cn\gamma^2)/(\lambda_u[1 + \tfrac{1}{2}K^2 + (\gamma\theta)^2]) \qquad [9a]$$

or in terms of wavelength:

$$\lambda_n(\theta) = (\lambda_u/2n\gamma^2)[1 + \tfrac{1}{2}K^2 + (\gamma\theta)^2] \qquad [9b]$$

Equation [9] is the central expression describing the spectral output of an undulator. Notice how the macroscopic period of the undulator λ_u is transformed into a microscopic wavelength $\lambda_n(\theta)$ by the relativistic factor γ^2. For third-generation hard X-ray sources with particle energies in the 6–8 GeV range, $\gamma \sim 14\,000$, taking cm scales down to the angstrom level.

Equation [9] describes a spectrum of monochromatic peaks, for on-axis radiation ($\theta = 0$), with wavelengths that increase with increasing K. Since $K \sim B_0\lambda_u$ (eqn [6]) any particular harmonic can be tuned either by changing B_0 or λ_u. In most undulator designs, λ_u is fixed by the construction of the permanent magnet array, so the tuning is conveniently accomplished by varying B_0. In practice this is done by changing the vertical distance between the poles

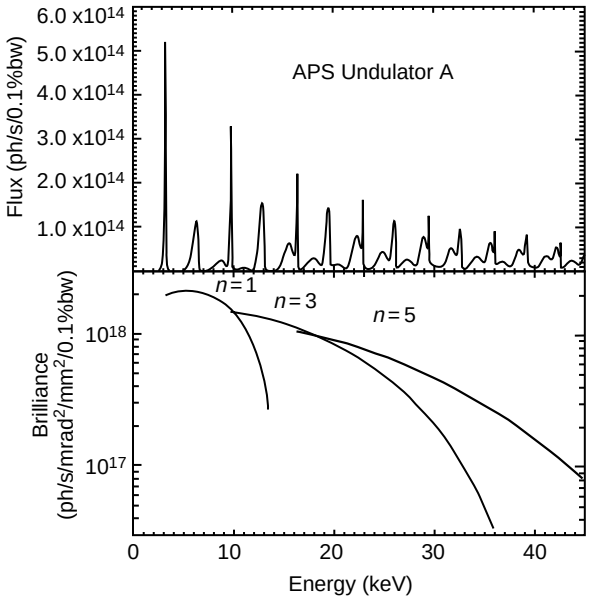


Figure 4 Undulator spectrum (upper panel) and tuning range for first, third and fifth harmonics (lower panel) of Advanced Photon Source Undulator A.

of the magnets (the 'undulator gap') similar to opening the jaws of a vise. Typical tuning ranges for the $n = 1$, 3, and 5 harmonics are illustrated in Figure 4.

Since the number of undulator oscillations of the particle beam is finite, the nth harmonic is not strictly monochromatic and has a spectral broadening of order $1/nN$, where N is the number of oscillator periods. In practice with N typically ~ 100, a bandwidth of $\sim 1\%$ is expected for the first harmonic. Additional broadening of a similar magnitude might also be expected from the θ-dependence in eqn [9], depending on the acceptance angle of the optics that is used in a particular experiment. This quasi-monochromatic radiation is sometimes referred to as 'pink beam'.

The storage ring energy of the APS was carefully chosen so that the tuning ranges of the different harmonics would overlap in energy, thus providing a continuous tuning of the X-ray energy over a wide range (Figure 4). Note also that the discrete spectral characteristics of the undulator can be deliberately smeared out to mimic the broadband (white beam) characteristics of a bend magnet (with consequent reduction in brightness). This is done simply by tilting the jaws of the undulator with respect to each other so that B_0 is made to vary along the length of the undulator. The ability to taper the undulator in this way is a standard feature of most designs.

The radiation from an undulator is sharply peaked in the forward direction in both the vertical and horizontal directions. The half-angle of the emission cone is given by:

$$\theta_{1/2} = (1/\gamma)[(1 + \tfrac{1}{2}K^2)/2Nn]^{1/2} = [\lambda_n/L]^{1/2} \qquad [10]$$

where L is length of the undulator. For a typical third-generation source undulator, such as the Type A undulator at the Advanced Photon Source, $L = 2.4$ m, giving $\theta_{1/2} \approx 6$ μrad for the first harmonic tuned to 12 keV. This corresponds to an X-ray beam size of ~ 180 μm at a distance 30 m from the source.

The very tight collimation of undulator radiation (eqn [10]) highlights the importance of distinguishing between the brightness and the 'brilliance' of the beam. The latter term takes into account the natural collimation of the beam, as reflected in its units of photons/sec/mm^2/mrad^2/0.1% bandwidth. Undulator radiation is well matched to the acceptance angle of ideal crystal monochromators and is $\sim 10\,000$ times more brilliant than the radiation from a bend magnet source. The high brilliance of undulator X-ray beams is extremely advantageous in X-ray imaging applications, X-ray microprobe measurements, and in experiments where the sample size is very small (e.g., high-pressure diamond anvil cell measurements).

In practice the emission cone of the source might be broadened a little when the undulator characteristics are convolved with those of the storage ring. The largest effect is in the horizontal direction where the photon beam divergence is dominated by the electron beam divergence. This is a consequence of the particular conditions under which the storage ring is operated in order to stabilize the particle beam orbit.

Equation [10] demonstrates that preserving the exquisite performance capabilities of an undulator insertion device places severe demands on the stability and precision of the storage ring orbit. An important measure of this precision is characterized by the 'emittance' of the particle beam defined in terms of the area of phase space (position and angular divergence) occupied by the stored beam. In general, storage rings are designed to achieve very low emittances and in current third-generation sources they are close to being diffraction limited. Table 2 summarizes the characteristics of the APS storage ring operated in 'high-emittance' and 'low-emittance' modes.

The large power densities resulting from the extreme brightness of undulator sources (up to 160 kW/mrad^2 for APS Undulator A) necessitates careful design of the X-ray optical components that are used to condition the beam. For example, if a double-bounce Silicon crystal pair is used as the first optical component to monochromatize the undulator

Table 2 Advanced photon source storage ring parameters

Parameter	*Low emittance*	*High emittance*
Storage ring energy, E	7.0 GeV	7.0 GeV
Storage ring current, I	100 mA	100 mA
Beam energy spread, $\delta E/E$	0.096%	0.096%
Horizontal emittance, ε_x	3.5×10^{-9} m-rad	7.7×10^{-9} m-rad
Vertical emittance, ε_y	3.5×10^{-11} m-rad	7.7×10^{-11} m-rad
Coupling constant	1%	1%
Horizontal beta function, β_x	14.4 m	16.1 m
Vertical beta function, β_y	4.0 m	4.3 m
Dispersion function, η_x	0.124 m	0.0 m
Horizontal beam size, σ_x	254 μm	351 μm
Vertical beam size, σ_y	12 μm	18 μm
Horizontal beam divergence, $\sigma_{x'}$	15.6 μrad	21.8 μrad
Vertical beam divergence, $\sigma_{y'}$	3.0 μrad	4.2 μrad

beam, the first crystal is normally cooled with liquid nitrogen to dissipate the heat. Alternately, high thermal conductivity diamond crystals are sometimes used as monochromators.

Helical Undulators

The use of undulators to produce extremely bright, tunable X-ray beams has enabled many new applications of X-ray science, as well as offering significant improvements to existing techniques in terms of quality of data and the time required to accumulate a high degree of statistical accuracy. One of the areas that has benefited significantly from novel undulator design is research on magnetic materials and other spin-dependent phenomena. It is often advantageous to probe such systems with circularly polarized X-rays, for example to study magnetic X-ray circular dichroism (MXCD). Early MXCD measurements utilized the left and right elliptical polarization of bending magnet radiation above and below the plane of the ring. Such measurements are difficult because the intensity falls off dramatically out of the plane. The problem to produce high-brightness beams of circularly polarized X-rays is effectively solved by designing an undulator which produces a transverse periodic helical magnetic field of constant magnitude. Several helical undulators are now in operation at third-generation sources. The soft X-ray region (500–1500 eV) is particularly interesting because it covers the L-edges of the 3d transition elements.

Coherence

While this article describes what would be nominally considered as 'incoherent' sources in the parlance of conventional optics, it should be pointed out that undulators are in fact excellent sources of coherent X-rays. This is by virtue of the high degree of collimation inherent in the emission of radiation from the undulations of a highly relativistic charged particle (eqn [10]). The spatial (transverse) coherence angle is given by:

$$\lambda_n/2d \qquad [11]$$

where d is the source size. This is the angle within which the beam is spatially coherent.

For APS undulator A, the source size is ~15 μm (vertical) × 240 μm (horizontal), giving a coherence angle in the vertical direction of 6.5 μrad and 0.4 μrad in the horizontal direction. The X-ray beam divergence is approximately 4 μrad (vertical) by 15 μrad (horizontal). This means that the beam has complete spatial coherence in the vertical direction and is coherent over about 3% of the horizontal fan. At a distance of 30 m from the source, an on-axis pinhole of ~12 micron diameter will be completely illuminated by coherent X-rays. Transversely coherent X-ray beams are finding novel applications in X-ray photon correlation spectroscopy, coherent X-ray scattering, and phase reconstruction (lensless) imaging.

Timing Structure

X-ray storage rings produce pulsed X-rays. In the Advanced Photon Source, one or more long trains of electron bunches (15 ns long) are accelerated by the linac to 325 MeV. The electron bunches are injected into a booster that ramps the single bunch from 325 MeV to 7 GeV in about 225 ms. At the end of the booster cycle, the bunch is extracted and injected into the storage ring. The bunch can be injected into any of 1296 'buckets' that are spaced by 2.842 ns, forming any arbitrary pattern after several injection cycles.

The APS storage ring rf frequency of 351.927 MHz determines the bucket spacing of 2.842 ns. The circumference of 1104 m gives a revolution frequency of 271.5 kHz which allows for 1296 buckets in 3683 ns. The storage ring is filled with a selected bunch pattern depending on the requirements for a particular operations cycle. The standard 'singlets' bunch pattern at APS is 102 mA of average current in a train of 23 bunches (4.43 mA per bunch) each spaced by 153 ns (54 buckets). There is a gap of 306 ns (108 buckets) in this pattern, useful as

a timing guide. The singlets pattern has a lower lifetime than patterns with less charge per bunch, due to natural internal scattering processes. A 'top-up' mode has been devised to overcome the problem of overall intensity fall-off due to the exponential decay of the stored current. In top-up mode, new bunches are injected into the storage ring every few minutes to maintain the stored current at 102 mA.

A second fill pattern, known as the 'hybrid' pattern, is again 102 mA but distributed in one bunch of 5 mA, plus 8 groups of 7 consecutive bunches (1.73 mA/bunch), spaced by 68 ns (24 buckets). The 8 groups are diametrically opposite the 5 mA bunch, allowing for a 1.5 μs gap on both sides of the 5 mA bunch. This hybrid mode permits a single bunch to be isolated using a mechanical chopper or a fast framing camera. Since the RMS width of the single bunch is ~36 ps, relatively fast time-resolved experiments can be performed stroboscopically using this pulsed X-ray beam. In another implementation of time-resolved experiments, the singlet pattern is used, isolating each bunch with a fast avalanche photodiode detector, or an X-ray streak camera. With the streak camera it is possible to achieve time resolutions of a few picoseconds. This approach has been used to study impulsive phonon excitations and ultrafast strain wave propagation in solids.

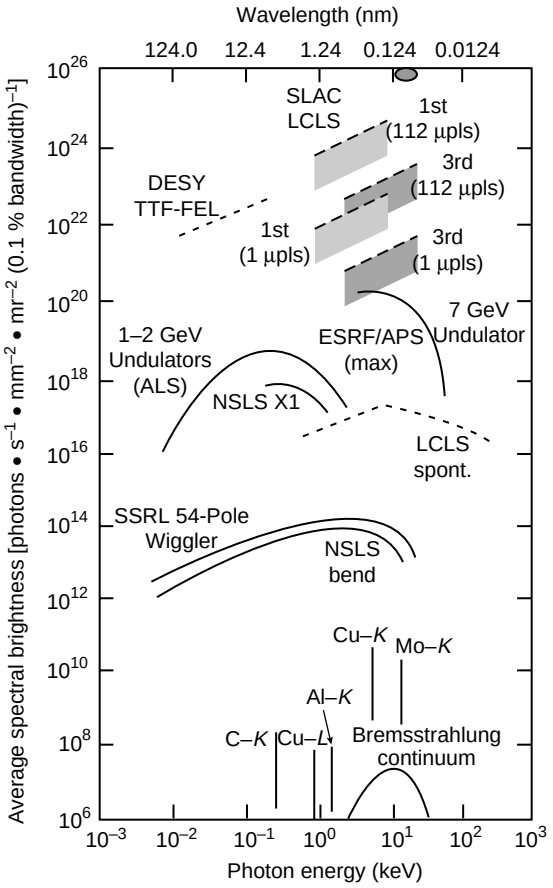


Figure 5 Comparison of spectral brilliance of X-ray sources. Courtesy of Linear Coherent Light Source, Stanford Linear Acceleration Center.

The Future

This article highlights current (third-generation) storage ring sources based on undulator-type insertion devices. These sources will provide high-brightness tunable X-rays in support of a wide range of experiments for many years to come. The brightness of these sources will increase steadily into the future as continual improvements are made, such as further reducing the emittance of the particle beam, improved undulator design, increasing the average storage ring current, and making changes to the storage ring lattice. These incremental improvements will be important to maintain the vitality and usefulness of national user facilities based on third-generation sources. New storage ring sources with designs and specifications optimized for low emittance and new types of undulators (e.g., in-vacuum undulators) will come online and will undoubtedly lead to new technical advances, supporting new and exciting science. It seems likely that linear accelerator (linac) based undulator sources will provide a promising route towards higher brilliance in the future. The superconducting Energy Recovery Linac (ERL) proposed by CHESS is an interesting concept which uses regeneration of the electron beam to overcome the high energy cost of single pass particle acceleration.

A new generation of X-ray sources (the 'fourth generation') is envisioned to take on the task of providing vastly brighter beams (>10 orders more peak brilliance than at present, see Figure 5) with much shorter pulses (~100 fs). These facilities will also be based on linear accelerators. Two such *Free Electron Laser* sources are currently being designed and should be operational towards 2010: the Linac Coherent Light Source (LCLS), to be built at the Stanford Linear Accelerator Center (SLAC), and the X-ray Free Electron Laser (XFEL) Source, to be constructed by the TESLA collaboration at HASYLAB, Hamburg.

See also

Coherence: Overview. **Physical Applications of Lasers:** Free Electron Lasers in Physics.

Further Reading

Brown G and Lavender W (1991) Synchrotron radiation spectra. In: Brown GS and Moncton DE (eds) *Handbook on Synchrotron Radiation*, vol. 3. Amsterdam: North Holland.

Duke PJ (1993) Synchrotron radiation sources. In: Michette AG and Buckley CJ (eds) *X-Ray Science and Technology*. Bristol: Institute of Physics Publishing Ltd.

Duke PJ (2000) *Synchrotron Radiation: Production and Properties*. Oxford, UK: OUP.

Mills DM (ed.) (2002) *Third Generation Hard X-ray Synchrotron Radiation Sources*. Berlin: Wiley, VCH.

INFORMATION PROCESSING

Contents

All-Optical Multiplexing/Demultiplexing
Coherent Analog Optical Processors
Free-Space Optical Computing
Incoherent Analog Optical Processors
Optical Bit-Serial Computing
Optical Digital Image Processing
Optical Neural Networks

All-Optical Multiplexing/Demultiplexing

Z Ghassemlooy, Northumbria University, Newcastle upon Tyne, UK

G Swift, Sheffield Hallam University, Sheffield, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The ever-increasing aggregate demand of electrically based time division multiplexing systems should have coped with the steady growth rate of voice traffic. However, since 1990, the explosive growth of the Internet and other bandwidth-demanding multimedia applications, has meant that long-haul telecommunication traffic has been increasingly dominated by data, not voice traffic. Such systems suffer from a bandwidth bottleneck due to speed limitations of the electronics. This limits the maximum data rate to considerably less than the THz bandwidth offered by an optical fiber. Optical technology is proposed as the only viable option and is expected to play an ever increasing role in future ultrahigh-speed links/networks. There are a number of multiplexing techniques, such as space division multiplexing (SDM), wavelength division multiplexing (WDM), and optical time division multiplexing (OTDM), that are currently being applied to effectively utilize the bandwidth of optical fiber as a means to overcome the bandwidth bottleneck imposed by electrical time division multiplexing (TDM). In SDM a separate optical fiber is allocated to each channel, but this is the least preferred option for increasing channel numbers. In WDM, a number of different data channels are allocated to discrete optical wavelengths for transmission over a single fiber. Dense WDM technology has been improving at a steady rate in recent years, with the latest systems capable of operating at a data rate of >1 T bps, using a large number of wavelengths over a single fiber link. However, there are a number of problems associated with the WDM systems such as:

- Performance of WDM is highly dependent on the nonlinearities associated with fiber, i.e.:
  - *Stimulated Raman scattering*: degrades the signal-to-noise (SNR) as the number of channels increases;
  - *Four-wave mixing*: limits the channel spacing;
  - *Cross-phase modulation*: limits the number of channels.
- Relatively static optical paths, thus offering no fast switching with high performance within the network;

- Switching is normally carried out by separating each wavelength of each fiber onto different physical outputs. Space switches are then used to spatially switch the separated wavelengths, an extremely inefficient way of utilizing network resources;
- The need for amplifiers with high gain and flat spectra.

In order to overcome these problems, OTDM was introduced that offers the following:

- Flexible high bandwidth on demand (>1 Tbit/s compared to the bit rates of 2.5–40 Gbit/s per wavelength channel in WDM systems);
- The total bandwidth offered by a single channel network is equal to DWDM;
- In a network environment, OTDM provides potential improvements in:
  - Network user access time, delay and throughput, depending on the user rates and statistics.
  - Less complex end node equipment (single-channel versus multichannels).
- Self-routing and self-clocking characteristics;
- Can operate at second- and third-transmission windows:
  - 1500 nm (like WDM) due to Erbium doped fiber amplifier (EDFA);
  - 1300 nm wavelengths.
- Offers both broadcast and switched based networks.

Principle of OTDM

Figure 1 show the generic block diagram of a point-to-point OTDM transmission link, where N optical data channels, each of capacity M Gbps, are multiplexed to give an aggregate rate of $N \times M$ Gbps. The fundamental components are a pulsed light source, an optical modulator, a multiplexer, channel, add/drop unit, and a demultiplexer. The light source needs to have good stabilities and be capable of generating ultrashort pulses (<1 ps). Direct modulation of the laser source is possible but the preferred method is based on external modulation where the optical signal is gated by the electronic data. The combination of these techniques allows the time division multiplexed data to be encoded inside a subnanosecond time slot, which is subsequently interleaved into a frame format. Add/drop units provide added versatility (see Figure 2) allowing the 'adding' and 'dropping' of selected OTDM channels to interchange data at chosen points on the link. At the receiving end, the OTDM pulse stream is demultiplexed down to the individual channels at the initial M Gbps data rate. Data retrieval is then within the realm of electronic devices and the distinction between electronic and optical methods is no longer relevant. Demultiplexing requires high-speed all optical switching and can be achieved using a number of methods, which will be discussed in more detail below.

The optical multiplexing (or interleaving) can be carried out at the bit level (known as bit interleaving) or at the packet level (known as packet interleaving), where blocks of bits are interleaved sequentially. This is in accord with the popular conception of packet switched networks. The high data rates required and consequently narrow time slots necessitate the need for strict tolerances at processing nodes, e.g., switches. As such, it is important that the duration

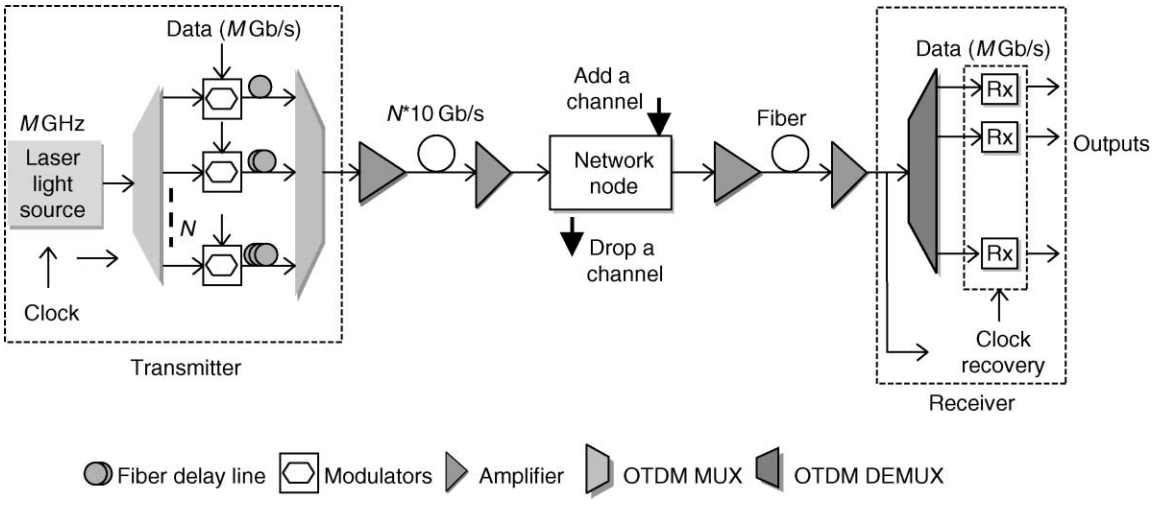


Figure 1 Block diagram of a typical OTDM transmission system.

of the optical pulses is chosen to be significantly shorter than the bit period of the highest multiplexed line rate, in order to reduce the crosstalk between channels.

Bit Interleaved OTDM

A simple conceptual description of a bit interleaved multiplexer is shown in Figure 3. It uses a number of different length optical fiber delay lines (FDL) to interleave the channels. The propagation delay of each FDL is chosen to position the optical channel in its corresponding time slot in relation to the aggregate OTDM signal. Prior to this, each optical pulse train is modulated by the data stream. The output of the modulators and an undelayed pulse train, labeled the framing signal, are combined, using a star coupler or combiner, to produce the high bit rate OTDM signal (see Figure 3b). As shown in Figure 3b, the framing pulse has a higher intensity for clock recovery purpose. At the demultiplexer, the incoming OTDM pulse train is split into two paths (see Figure 4a). The lower path is used to recover the framing pulse by means of thresholding, this is then delayed by an amount corresponding to the position of the *i*th (wanted) channel (see Figure 4b). The delayed framing pulse and the OTDM pulse stream are then passed through an AND gate to recover the *i*th channel. The AND operation can be carried out all optically using, for example, a nonlinear loop mirror, a terahertz optical asymmetric demultiplexer, or a soliton-trapping gate.

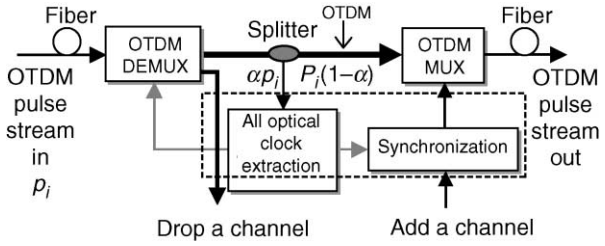


Figure 2 Add/drop unit in an OTDM network node.

Packet Interleaved OTDM

Figure 5a shows the block diagram of a system used to demonstrate packet interleaving. Note that the time interval between successive pulses now needs to be much less than the bit interval T. This is achieved by passing the low bit rate output packet of the modulator into a compressor, which is based on a feed forward delay line structure. The feed forward delay lines are based on a cascade of passive M–Z interferometers with unequal arm lengths. This configuration generates the delay times required, i.e., $T-\tau, 2(T-\tau)\ldots(2^{n-1})(T-\tau)$, etc., where $n = \log_2 k$ is the number of stages, T and τ are the

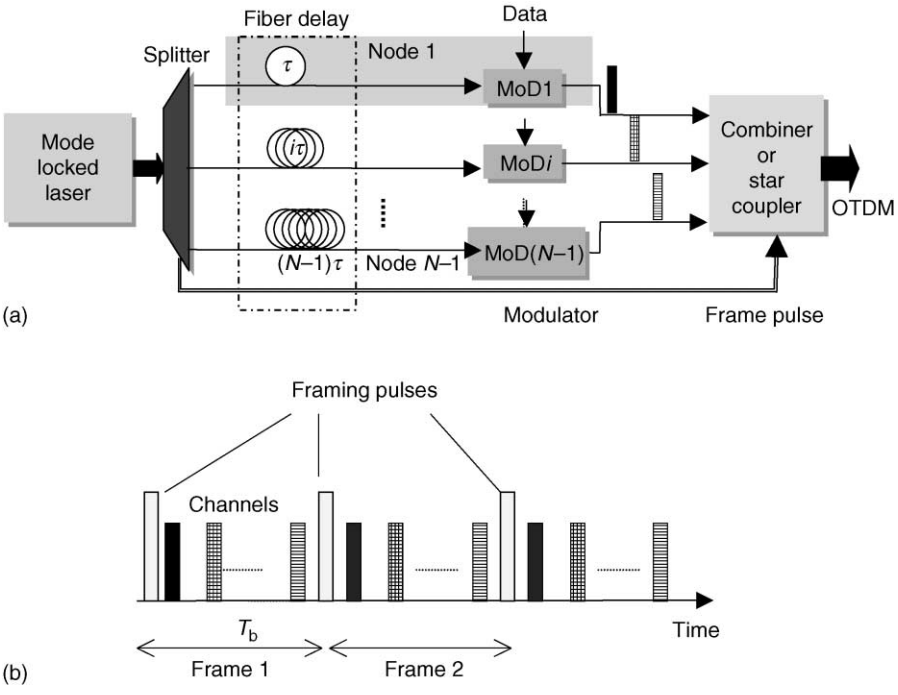


Figure 3 Bit interleaved OTDM; (a) block diagram and (b) timing waveforms.

incoming bit duration and the outgoing bit duration, respectively, and k is the packet length in bits. Pulse i's location at the output is given by $(2^n - 1)(T - t) + (i - 1)\tau$. The signal at the input of the compressor is:

$$I_{\text{in}}(t) = \sum_{i=0}^{k-1} \delta(t - iT)A_i \qquad [1]$$

Where $\delta(\cdot)$ is the optical carrier pulse shape, and A_i is the ith data bit in the packet.

As shown in Figure 5b, each bit is split, delayed, and combined to produce the four bit packets O_i:

$$O_i(t) = \frac{1}{2^{n+1}} \sum_{i=0}^{k-1} I_i[t - i(T - \tau)] \qquad [2]$$

Note the factor of $1/(2^{n-1})$ is due to the signal splitting at the 2×2 3-dB coupler at the input of each stage.

The combined signal is shown in Figure 5b, and is defined as:

$$I_{\text{out}} = \sum_{i=0}^{k-1} O_i = \frac{1}{2^{n+1}} \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} I_i[t - (i+j)T + j\tau]A_i \qquad [3]$$

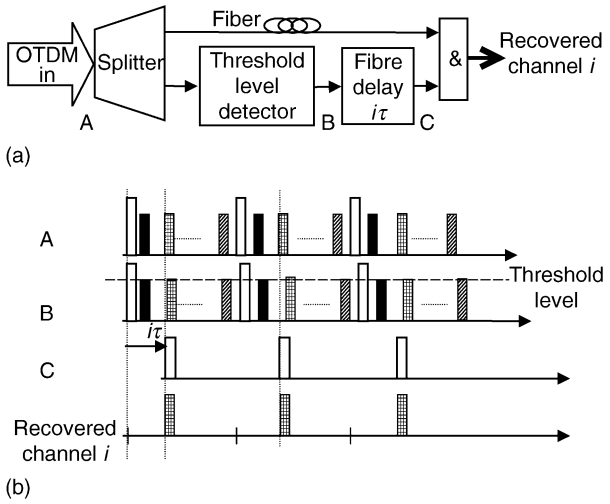


Figure 4 Bit interleaving demultiplexer: (a) block diagram and (b) typical waveforms.

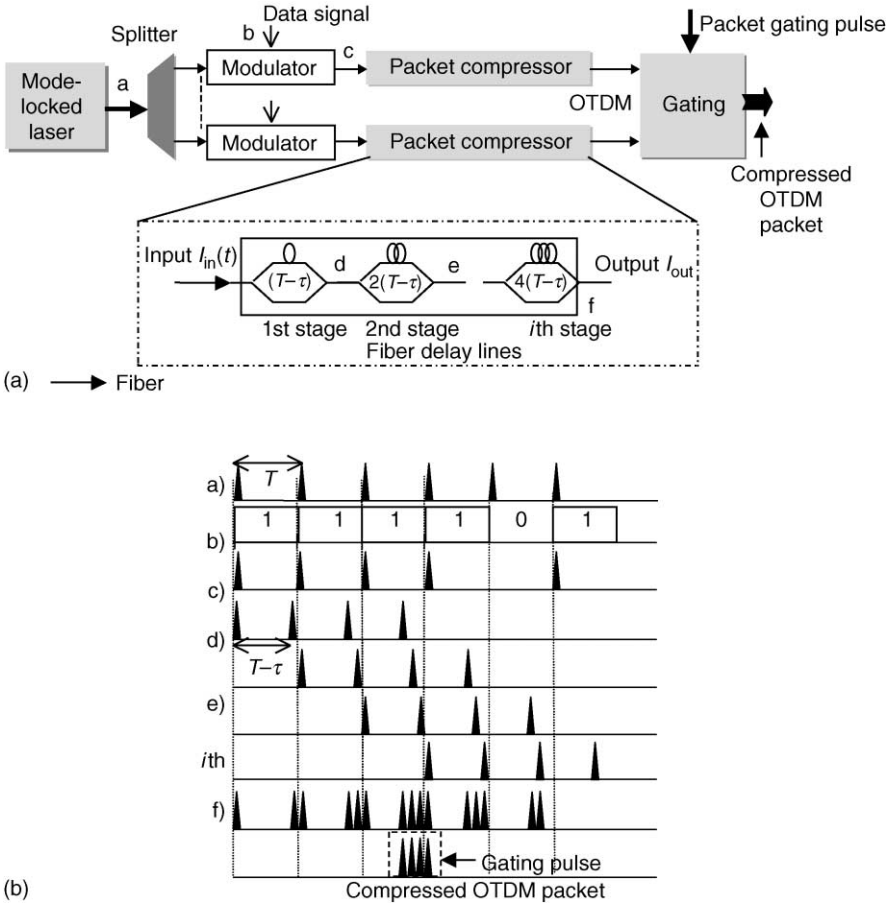


Figure 5 Packet interleaving multiplexer: (a) block diagram and (b) typical waveforms.

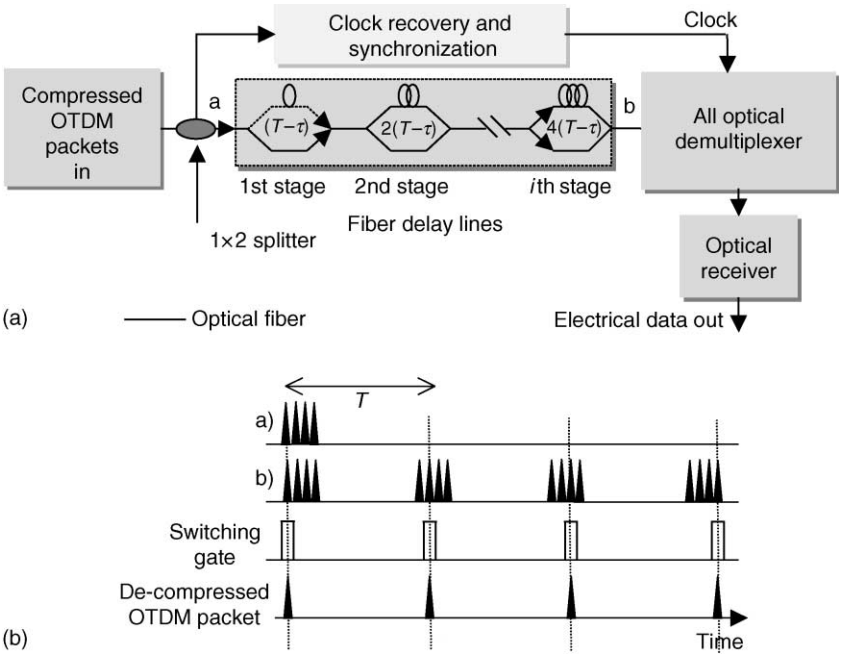


Figure 6 Packet interleaving demultiplexer: (a) block diagram and (b) typical waveforms.

For a packet delay of $(i + j)T$ packets are being built up. For the case $(i+j) = 3$, four bits are compressed into a packet (see Figure 5b). An optical gating device is used to select the only complete usable compressed copy of the incoming packets from the unwanted copies.

At the receiving end, demultiplexing is carried out by decompressing the OTDM packet stream using the same delay line structure (see Figure 6). Multiple copies of the compressed packet are generated, and each is delayed by $(T - \tau)$. An optical demultiplexer placed at the output of the delay lines generates a sequence of switching window at the rate of $1/T$. By positioning the switching pulses at the appropriate position, a series of adjacent packet (channel) bits from each of the copied compressed packets is extracted (see Figure 6b). The output of the demultiplexer is the decompressed packet signal, which can now be processed at a much slower rate using electronic circuitry.

Components of an OTDM System

Optical Sources

In an ultrahigh speed OTDM system, it is essential that the optical sources are capable of generating transform-limited subnanosecond pulses having low duty cycles, tuneability, and a controllable repetition rate for synchronization. A number of suitable light sources are: gain-switch distribute feedback laser (DFB), active mode locked lasers (MLL) (capable of generating repetitive optical pulses), and harmonic mode-locking Erbium-doped fiber (EDF) lasers. Alternative techniques include super-continuum pulse generation in a dispersion shifted fiber with EDF pumping, adiabatic soliton compression using dispersion-flattened dispersion decreasing fiber, and pedestal reduction of compressed pulses using a dispersion-imbalanced nonlinear optical loop mirror. As shown in Figure 1, the laser light source is power split to form the pulse source for each channel. These are subsequently modulated with an electrical data signal (external modulation). External modulation is preferred in an OTDM system as it can achieve narrow carrier linewidth, thus reducing the timing jitter of the transmitted pulse. For ultrahigh bit rate OTDM systems, the optical pulses emerging from the external modulators may also need compressing. One option is to frequency-chirp the pulses and pass them through an anomalous dispersive medium. Using this approach, it is essential that the frequency chirp be linear throughout the duration of the pulse, in addition the midpoint of the linear-frequency chirp should coincide with the center of the pulse. When a frequency-chirped pulse passes through a dispersive medium, different parts of the pulse travel at different speeds, due to a temporal variation of frequency. If the trailing edge travels faster than the leading edge, the result would be pulse compression. An optical fiber cable or a semiconductor laser amplifier (SLA) can be used to create the frequency chirp.

Multiplexers

Multiplexing of the pulses generated by the optical sources can be implemented either passively or actively. The former method is commonly implemented using a mono-mode optical fiber. This method has the advantage of being simple and cost-effective. The latter method uses devices more complex in nature, for example, electro-optic sampling switches, semiconductor optical amplifiers, and integrated optics.

An integrated active multiplexer can be made by integrating semiconductor laser amplifiers (SOA) into a hybrid planar light-wave circuit (PLC). If the optical path length is kept short then temperature control is made easier. An alternative to this approach is to use lithium niobate (LiNb) technology. Periodically poled LiNb (PPLN) based OTDM multiplexers offer compact size and low noise (due to the absence of amplifier spontaneous emission noise and pattern effect, which is a feature of SOA based devices). Figure 7 is a schematic of a PLC-based OTDM multiplexer composed of two PLCs (R and L), composed of 1×8 and 8×1 couplers, eight 2×1 couplers, and eight different path length PPLN waveguides. The input clock pulse is split into eight by the 1×8 coupler in PLC-L, and are combined with the modulated optical pulse trains using the 2×1 couplers. The outputs of the 2×1 couplers are then passed through a PPLN waveguide to generate a return-to-zero (RZ) optically modulated signal. These are then combined, using the 8×1 coupler in PLC-R. When the path-length difference between the waveguides is set to one time slot (equal to 1/8 of the base data rate), then the output of the 8×1 coupler is the required high-bit rate OTDM pulse.

Demultiplexers

In contrast to multiplexing, demultiplexing must be performed as an active function. It can be implemented electro-optically or optically. The former method needs to complete demultiplexing of all channels in order to extract a single channel (see Figure 8). The demultiplexer in Figure 8 uses two LiNb Mach–Zehender (M–Z) modulators in tandem. The first and second modulators are driven with a sinusoidal signal of amplitudes $2V_\pi$ and V_π, respectively, to down-covert the N-bit rate to $N/2$ and $N/4$, respectively. Channels can be selected by changing either the DC-bias V + to the M–Zs, or the electrical phase delay. At ultrahigh speed implementation of an electro-optics demultiplexer becomes increasingly difficult due to the higher drive voltage requirement by the M–Z. An alternative is to use all optical demultiplexing based on the nonlinear effect in a fiber and optical active devices offering switching resolution in the order of picoseconds. There are a number of methods available to implement all optical demultiplexing. The most popular methods that use fast phase modulation of an optical signal are based on M–Z and Sagnac interferometers. Four-wave mixing is another popular method.

Mach–Zehnder (M–Z) Interferometers

The key to interferometric switching is the selection of an appropriate material for the phase modulation. Semiconductor materials, often in the form of an SLA, are suitable devices for this purpose.

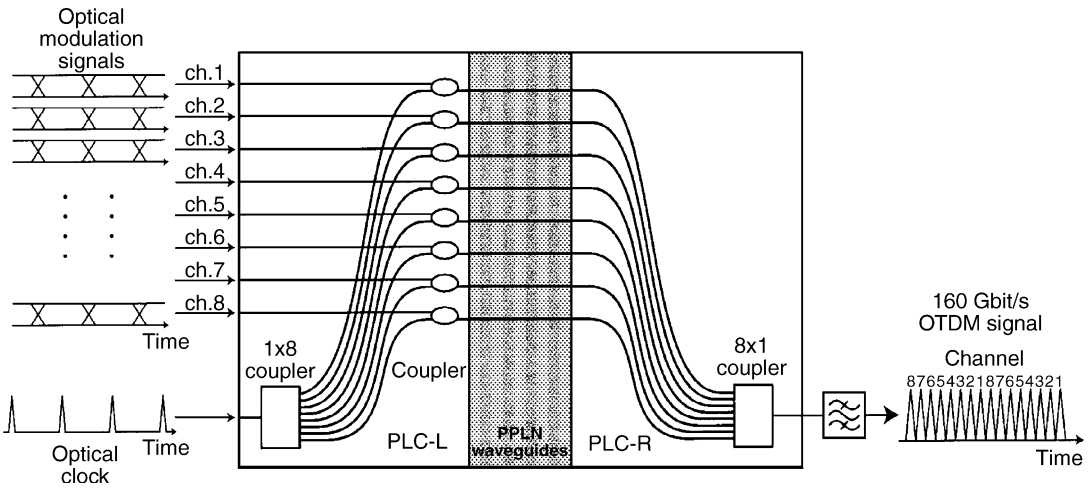


Figure 7 Configuration of a PLC-OTDM-MUX. Reproduced with permission from Ohara T, Takara H and Shake I, *et al.* (2003). 160-Gb/s optica-time-division multiplexing with PPLN hybrid integrated planar lightwave circuit. *IEEE Photonics Letters* 15(2): 302–304. © 2003 IEEE.

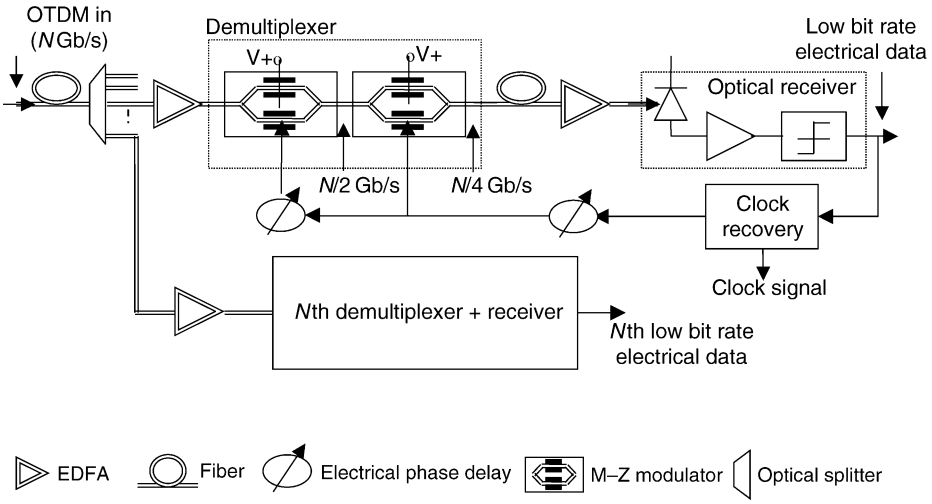


Figure 8 Electro-optics demultiplexer and receiver system block diagram.

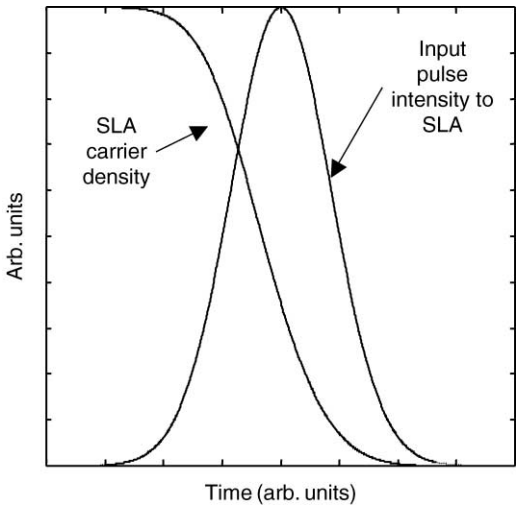


Figure 9 SLA carrier density response to input pulse.

The refractive index of an SLA is a function of the semiconductor carrier density, which can be modulated optically resulting in fast modulation. If a high-intensity pulse is input to an SLA the carrier density changes nonlinearly, as shown in Figure 9. Phase modulation is affected via the material chirp index (dN/dn), which represents the gradient of the refractive index carrier density curve for the material. The phase modulation is quite strong and, in contrast to the intensity-based nonlinearity in an optical fiber, (see Kerr effect below), is a consequence of a resonant interaction between the optical signal and the material. The nonlinearity is relatively long-lived and would be expected to limit the switching speed when using semiconductors. Semiconductors, when in excitation from an optical field, tend to experience the effect quite quickly (ps) with a slow release (hundreds of picoseconds) time. Advantage can be taken of this property by allowing the slow recovery to occur during the time between channels of an OTDM signal, as this time may be of the order of hundreds of pico-seconds or more. A Mach–Zehnder configuration using two SLAs, one in each arm of the interferometer placed asymmetrically, is shown in Figure 10. An optical control pulse entering the device via port 3 initiates the nonlinearity. The transmission equation relating the input signal (port 1) to the output port is given by

$$\frac{I_{\text{out}}(t)}{I_{\text{in}}(t)} = 0.25(G_{\text{SLA1}}(t) + G_{\text{SLA2}}(t) \pm 2\sqrt{G_{\text{SLA1}}(t)G_{\text{SLA2}}(t)}\cos\Delta\phi(t)) \quad [4]$$

where $G_{\text{SLA1}}(t)$ and $G_{\text{SLA2}}(t)$ refer to the gain profiles of the respective SLAs, and $\Delta\phi(t)$ is the time-dependent phase difference between them. Assuming that the gain and phase profiles of an excited amplifier are given by $G(t)$ and $\phi(t)$, respectively, then the signals passing through the upper and lower arms experience optical properties of the material given by:

$$G(t),\ G(t - T_{\text{d}}),\ \phi(t) \text{ and } \phi(t - T_{\text{d}}) \quad [5]$$

where T_{d} is given by $2L_{\text{d}}N_{\text{SLA}}/c$, L_{d} is the distance between SLA1 and SLA2, N_{SLA} is the SLA index, and c is the speed of light in a vacuum. As the width of the $\Delta\phi(t)$ profile is dependent on the distance between the SLAs, placing them in close proximity allows high-resolution switching. Less emphasis has been placed on the SLA gain as this is considered to be less effective when phase differences of π are reached. It only remains for the gain and phase

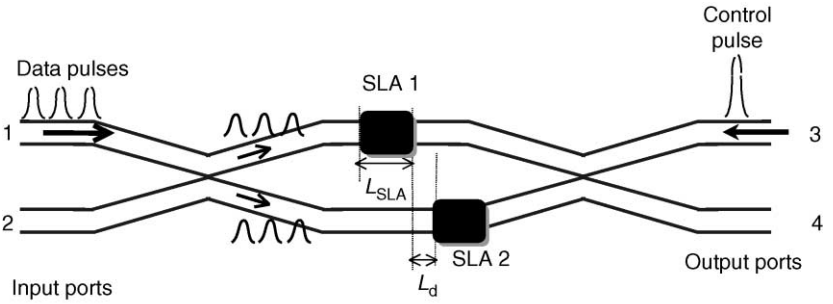


Figure 10 Asymmetric TWSLA Mach–Zehnder devices.

modulation to recover, and as indicated previously, this is allowable over a time-scale commensurate with tributary rates.

Sagnac Interferometers

There are two main types; the nonlinear optical loop mirror (NOLM), and the terahertz optical asymmetric demultiplexer (TOAD).

NOLM

In this method of switching the inherent nonlinearity of an optical fiber known as the Kerr effect is used. The phase velocity of any light beam passing through the fiber will be affected by its own intensity and the intensity of any other beams present. When the intrinsic third-order nonlinearity of silica fibers is considered via the intensity-dependent nonlinear refraction component, then the signal phase shift is

$$\Delta\phi_{\text{signal}} = \frac{2\pi}{\lambda_{\text{ds}}} n_2 L I_{\text{ds}} + 2\frac{2\pi}{\lambda_{\text{ds}}} n_2 L I_{\text{c}} \qquad [6]$$

where n_2 is the Kerr coefficient, I_{ds} is the intensity of the data signal to be switched, I_{c} is the intensity of the control signal used to switch the data signal, λ_{ds} is the data signal wavelength, and L is the fiber length. The optical loop mirror consists of a long length of mono-mode fiber formed into a fiber coupler at its free ends (see Figure 11). The input to the loop comprises the high-frequency data stream plus a control pulse at the frame rate. The data split at the coupler and propagate around the loop in contra directions (clockwise E_{CW} and counter-clockwise E_{CCW}) recombining back at the coupler. In the absence of a control pulse, the pulse exits via port 1. If a particular pulse in the loop (in this example E_{CW}) is straddled by the control pulse (see Figure 12), then that pulse experiences cross-phase modulation, according to the second term on the right-hand side of eqn [6], and undergoes a phase

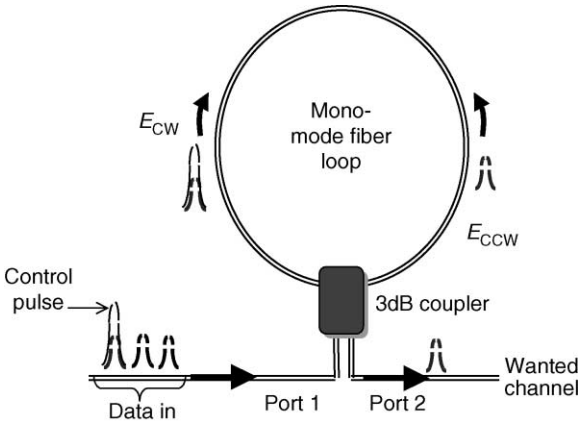


Figure 11 Nonlinear optical loop mirror.

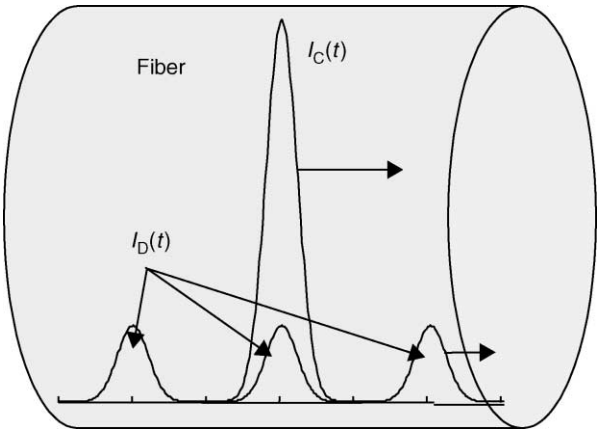


Figure 12 Control and data pulses propagation in the fiber.

change relative to E_{CW}. The difference in the phase between E_{CW} and E_{CCW} causes the pulse to exit via port 2. The phase shift profile experienced by the co-propagating pulse is

$$\Delta\phi(t) = \frac{4\pi}{\lambda_{\text{ds}}} n_2 \int_0^L I_{\text{c}}(t)\, \mathrm{d}x \qquad [7]$$

Assuming unity gain around the loop, the transmittance of the NOLM is

$$T_x = 1 - \cos^2\left(\frac{\Delta\phi}{2}\right) = \Delta\phi(t)$$

$$= 1 - \cos^2\left(\frac{2\pi}{\lambda_{ds}} n_2 \int_0^L I_C(t)\, dx\right) \qquad [8]$$

As it stands, the switching resolution is determined by the width of the control pulse; however, it is possible to allow the signal and control to 'walk off' each other, allowing the window width to be increased by an amount determined by the 'walk off'. The phase shift is now

$$\Delta\phi(t) = 2\frac{2\pi}{\lambda_{ds}} n_2 \int_0^L I_C(t - T_w x)dx \qquad [9]$$

where the parameter T_w, is the walk off time per unit length between control and signal pulses. The increased window width is accompanied by a lower peak transmission (see **Figure 13**).

TOAD

The TOAD uses a loop mirror architecture that incorporates an SLA and only needs a short length of fiber loop (see **Figure 14a**). The SLA carrier density is modulated by a high-intensity control pulse, as in the M–Z-based demultiplexer. The operation of the TOAD is similar to the NOLM, where the loop transmittance is determined by the phase difference between CW and CCW traveling pulses. Strictly speaking, the gain of the SLA must also be taken into account; however, without any loss of generality the effect is adequately described by considering only the phase property. **Figure 14a** shows the timing diagram associated with a TOAD demultiplexer. The control pulse, shown in **Figure 14b(1)**, is incident at the SLA at a time t_1; the phase profiles for the CCW and CCW data pulses are shown in **Figures 14b(2)** and **(3)**, respectively. The resulting transmission window is shown in **Figure 14b(4)**. The window width is given by $t_{asy} = 2\Delta x/c_f$, where Δx is the SLA offset from the loop center and c_f is the speed of light in the fiber. As in NOLM, if the phase difference is of sufficient magnitude, then data can be switched to port 2. The switch definition is, in principle, determined by how close the SLA is placed to the loop center when the asymmetry is relatively large. However, for small asymmetries the switching window is asymmetric, which is due to the CW and CCW gain profiles being different (see **Figure 15**). Assuming the phase modulation dominates the transmission then the normalized transmission for a small asymmetry loop is as depicted in **Figure 16**. The gain response (not shown) would have a similar shape and temporal position as the phase response.

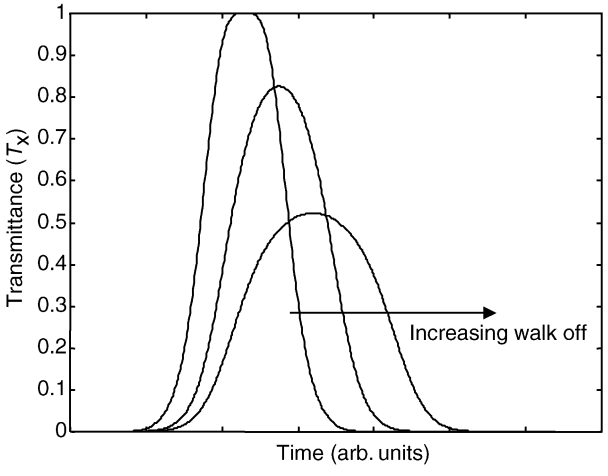


Figure 13 Transmittance profiles with the walk-off time as a parameter.

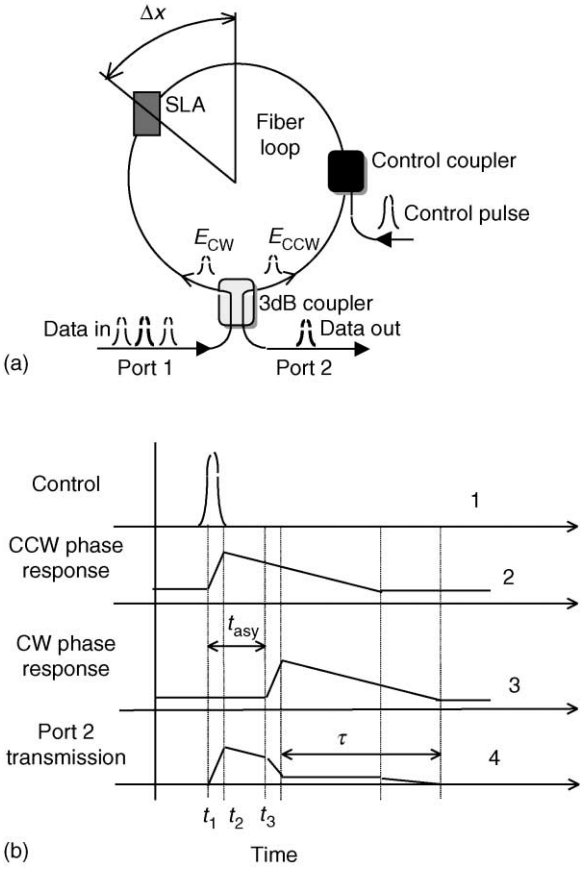


Figure 14 TOAD: (a) architecture and (b) timing diagrams.

Four-Wave Mixing (FWM)

FWM demultiplexing uses a concept whereby two optical signals of different wavelengths are mixed together in a nonlinear medium to produce harmonic components. The nonlinearity in this case arises from the third-order nonlinear susceptibility $\chi^{(3)}$ of an optical fiber, such that the polarization P induced on a pair of electric fields propagating through the fiber is

$$P = \varepsilon_0 \chi^{(3)} (E_{ds} + E_c)^3 \qquad [10]$$

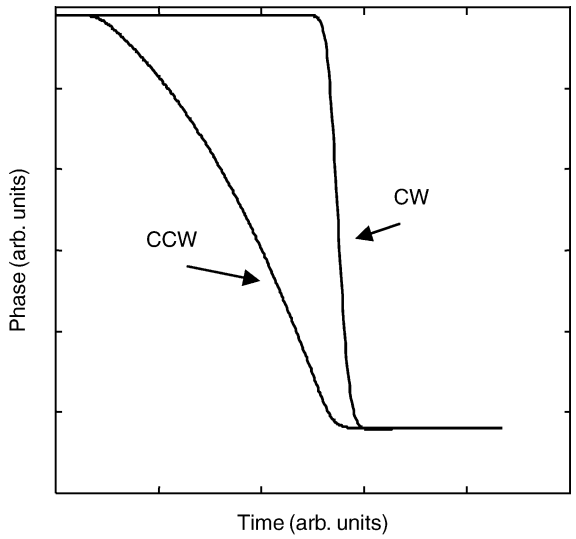


Figure 15 Phase responses for CW and CCW components of TOAD.

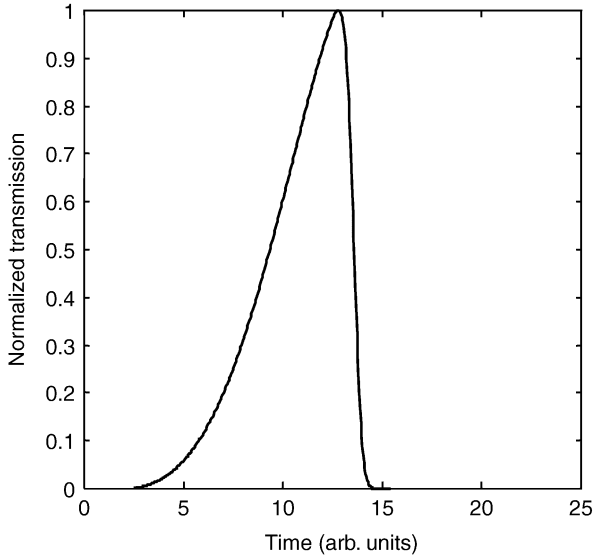


Figure 16 TOAD transmission window profile for small asymmetry loop.

where ε_0 is the vacuum permittivity of the medium, E_{ds} and E_c are the electric fields of data and control signals, respectively. The mixing of the two signals takes place in a long length of fiber in which the control signal propagates alongside the channel to be demultiplexed (see Figure 17). The control signal is of sufficient intensity to cause the fiber to operate in the nonlinear regime. The nonlinear relationship causes a number of frequencies to be generated, with the ones of interest having a frequency given by

$$f_{FWM} = 2f_c - f_{ds} \qquad [11]$$

An optical filter is then used to demultiplex the required channel from the composite signal. The nonlinearity is detuned from the resonant frequency of the fiber glass and as such tends to be weak, requiring long lengths of fiber to give a measurable effect. The power in the demultiplexed signal, for given data and control signal wavelengths and fiber material, depends on the input power and the fiber length according to

$$P_{FWM} = kP_{ds}P_cL_E^2 \qquad [12]$$

where k is a constant, P_{ds} is the signal launch power, P_c is the control signal launch power, and L_E is the fiber effective length. FWM is essentially an inefficient method as power is wasted in the unused frequency components of the four-wave signal. More in line with integrated structures, the nonlinear properties of a semiconductor laser amplifier can be used. Here a relatively high-power optical signal is input to the SLA (see Figure 18). This enables saturation and operation in the nonlinear regime

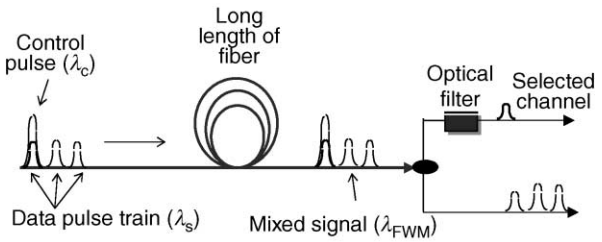


Figure 17 Block diagram of FWM demultiplexer.

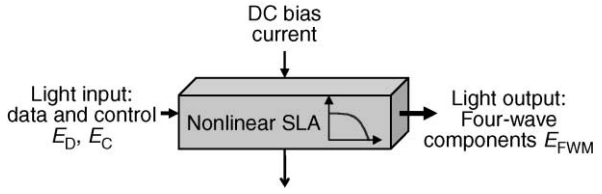


Figure 18 Four-wave mixing in SOA.

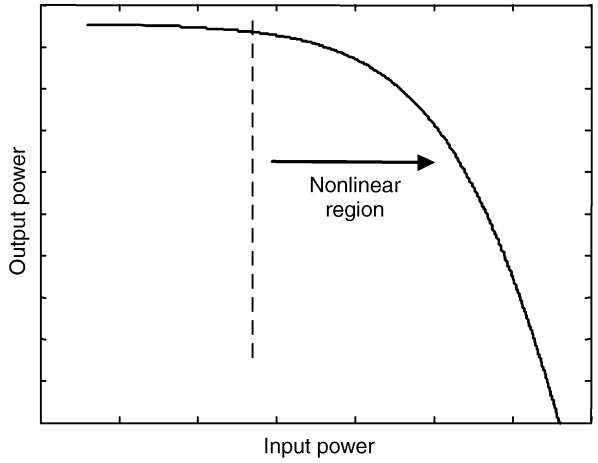


Figure 19 SLA output–input power curve.

(see Figure 19). Operating the SLA in saturation allows the nonlinear effects to produce the FWM components as in the fiber method.

Clock and Data Synchronization in OTDM

In common with electronically based TDM systems, clock recovery is fundamental to the recovery of data in ultrahigh-speed OTDM systems. Two main methods are proposed: (i) clock signal transmitted with the OTDM signal (i.e., multiplexed); and (ii) extraction of clock signal from the incoming OTDM pulse stream. In a packet-based system, synchronization between the clock and the data packet is normally achieved by sending a synch pulse with each packet. However, techniques based on optical-phased locked loops are proving popular and remove the need for a separate clocking signal.

Clock Multiplexing

(i) *Space division multiplexing*: This is conceptually the simplest to implement, where the clock signal is carried on a separate fiber from the data. However, it is susceptible to any differential delay between the different paths taken by clock and data due to temperature variation. It is difficult to justify in systems where the installed fiber base is at a premium.

(ii) *Wavelength division multiplexing*: Here different wavelengths are allocated to the clock, and payload. It is only really practical for predetermined path lengths between nodes in single-hop networks such as point–point links or broadcast-and-select star networks. It also suffers from random delays between the clock and the payload, which is problematic in an asynchronous packet switched based network, where the optical path length a packet may take is nondeterministic.

(iii) *Orthogonal polarization*: This is suitable for small links, where separate polarizations are used for the clock and data. However, in large networks it is quite difficult to maintain the polarization throughout the transmission link due to polarization mode dispersion and other nonlinear effects.

(iv) *Intensity division multiplexing*: This uses higher-intensity optical clock pulses to differentiate it from the data pulses as discussed above. However, in long-distance transmission links, it is difficult to maintain both the clock intensity and its position, due to the fiber nonlinearity.

(v) *Time division multiplexing*: In this scheme a single clock pulse, which has the same wavelength, polarization, and amplitude as the payload pulses, is separated in time, usually ahead of the data pulses.

Synchronization – Optical Phased Locked Loops (PLL)

The PLL is a common technique used for clock recovery in electronic TDM systems. However, the speed of conventional electronic PLLs tends to be limited by the response of the phase comparators used. There are a number of approaches based on opto-electronic PLL. Opto-electronic PLLs based on four-wave mixing in a traveling wave laser amplifier are complex and can suffer from frequency modulation in the recovered clock. However, others based on balanced photo-detectors, result in low timing jitter and good phase stability. In contrast, a number of all optical methods clock recovery scheme exist, one technique based on the TOAD (see above) is depicted in Figure 20. The high-speed data stream enters the TOAD at a rate $n \times R$ where n is the number of channels per OTDM frame, and R is the frame rate. A pulse generator, such as a mode locked fiber laser (MLFL) clocked by a local oscillator (LO), is used as the TOAD control input (MLFL-C). The OTDM data is switched by the TOAD at a frequency of say, $R + \Delta f$ Hz. Thus, the switching window samples data pulses at a rate higher or lower than $n \times R$ Hz and uses this signal for cross correlation in the PLL unit. The output of the phase comparator is used to regulate a voltage controlled oscillator (VCO)

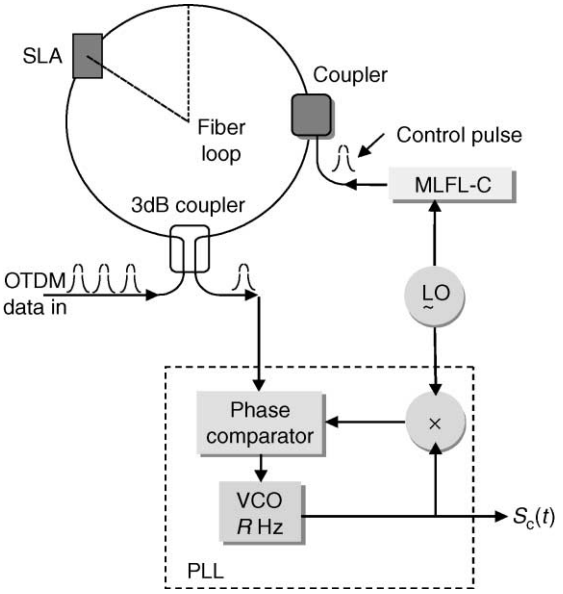


Figure 20 Ultrafast clock recovery using TOAD.

running at R Hz, which in turn feeds the control signal. The PLL circuit locks into the clock frequency, generating the clock signal $S_c(t)$.

OTDM Bit-Error Rate (BER) Performance

Figure 21 shows a typical block diagram of a high-speed optical receiver for an OTDM system, composed of a NOLM or a TOAD demultiplexer, an optical pre-amplifier, an optical bandpass filter, and a conventional optical receiver using a PIN photodiode. Due to the crosstalk introduced in the demultiplexing process, the demultiplexed optical signal contains not only the target channel but may also contain nontarget channels with reduced amplitude (see the inset in Figure 21). The intensity of the demultiplexed optical signal is boosted by the optical pre-amplifier (EDFA). Amplified spontaneous emission (ASE) from the optical preamplifier adds a wide spectrum of optical fields onto the demultiplexed optical signal. Although an optical bandpass filter (BPF) can reduce the ASE, it still remains one of the major noise sources. The function of the optical filter is to reduce the excess ASE to within the range of the signal spectrum. The PIN photodiode converts the received optical power into an equivalent electrical current, which is then converted to a voltage signal by an electrical amplifier. Finally, the output voltage is sampled periodically by a decision circuit for estimating the correct state (mark/space) of each bit.

The BER performance of an OTDM system deteriorates because of the noise and crosstalk

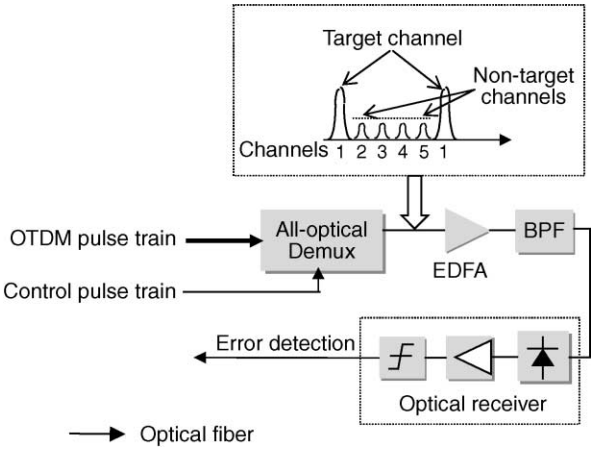


Figure 21 Block diagram of OTDM receiver.

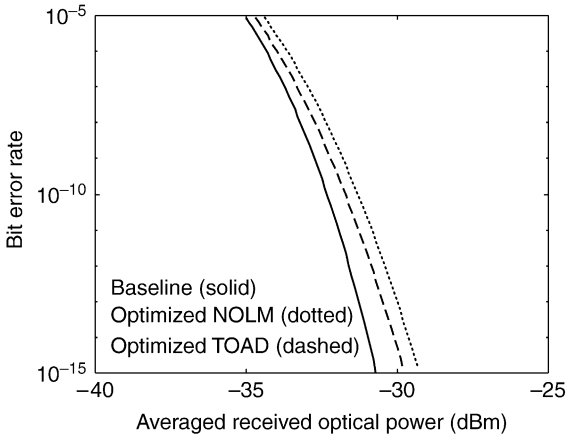


Figure 22 BER versus the average received optical power for 100 Gb/s (10 channels) OTDM system.

introduced by the demultiplexer and is:

$$\mathrm{BER} = \frac{1}{\sqrt{2\pi}} \frac{\exp\left(-Q^2/2\right)}{Q} \qquad [13]$$

where Q is defined as:

$$Q = \frac{\bar{I}_m - \bar{I}_s}{\sqrt{\sigma^2_{\mathrm{RIN,m}} + \sigma^2_{\mathrm{RIN,s}} + \sigma^2_{\mathrm{amp,m}} + \sigma^2_{\mathrm{amp,s}} + \sigma^2_{\mathrm{rec,m}} + \sigma^2_{\mathrm{rec,s}}}} \qquad [14]$$

and $\bar{I}_m$ and $\bar{I}_s$ are the average photocurrents for a mark and a space, respectively. $\sigma_{x,i}$ are the variances of the relative intensity noise (RIN), further optical pre-amplifier and receiver for a mark and a space.

For 100 Gb/s (10 channels) OTDM system, the BER against average received optical power for optimized NOLM and TOAD demultiplexers, is shown in Figure 22.

List of Units and Nomenclature

A_i	ith data bit in a packet
c	Speed of light in a vacuum
c_f	Speed of light within the fiber
E_c	Electric fields of control signals
E_{ds}	Electric fields of data signals
f_{FWM}	Frequency due to four-wave mixing
$G_{SLA}(t)$	Gain profiles of SLA
I_c	Intensity of the control signal
I_D	Intensity of the data signal
$\overline{I}_m$	Average photo current for a mark
$\overline{I}_s$	Average photo current for a space
k	Packet length in bits
L	Fiber length
L_d	Distance between two SLAs
L_E	Fiber effective length
n	Number of stages
n_2	Kerr coefficient
N_{SLA}	SLA index
P	Polarization
P_c	Control signal launch power
P_{ds}	Data signal launch power
t_{asy}	The window width is given by $t_{asy} = 2\Delta x/c_f$
T	Incoming bit interval
T_w	Walk off time per unit length between control and data pulses
T_x	Transmittance of the NOLM
$\delta(\cdot)$	Optical carrier pulse shape
$\Delta\phi(t)$	Time-dependent phase difference
Δx	SLA offset from the fiber loop center
ε_0	Vacuum permittivity of the medium
λ_s	Data signal wavelength
$\sigma_{amp,m}$	Variances for optical pre-amplifier for a mark
$\sigma_{amp,s}$	Variances for optical pre-amplifier for a space
$\sigma_{rec,m}$	Variances for receiver for a mark
$\sigma_{rec,s}$	Variances for receiver for a space
$\sigma_{RIN,m}$	Variances of RIN for a mark
$\sigma_{RIN,s}$	Variances of RIN for a space
τ	Outgoing bit interval
$\phi(t)$	Phase profile of SLA
$\chi^{(3)}$	Third-order non-linear susceptibility of an optical fiber

See also

Interferometry: Overview. **Nonlinear Optics, Basics**: Four-Wave Mixing. **Optical Communication Systems**: Optical Time Division Multiplexing; Wavelength Division Multiplexing.

Further Reading

Agrawal GP (2002) *Fiber-Optic Communication Systems*, 3rd edn. New York: J Wiley and Sons Inc.

Chang K (ed.) (2003) *Handbook of Optical Components and Engineering*, 2nd edn. New York: Wiley and Sons Inc.

De Marchis G and Sabella R (1999) *Optical Networks Design and Modelling*. Dordrecht: Kluwer Academic.

Hamilton SA, Robinson BS, Murphy TE, Savage SJ and Ippen EP (2002) 100 Gbit/s optical time-division multiplexed networks. *Journal of Lightwave Technology* 20(12): 2086–2100.

Jhon YM, Ki HJ and Kim SH (2003) Clock recovery from 40 gbps optical signal with optical phase-locked loop based on a terahertz optical asymmetric demultiplexer. *Optics Communications* 220: 315–319.

Kamatani O and Kawanishi S (1996) Ultrahigh-speed clock recovery with phase locked loop based on four wave mixing in a traveling-wave laser diode maplifier. *Journal of Lightwave Technology* 14: 1757–1767.

Nakamura S, Ueno Y and Tajima K (2001) Femtosecond switching with semiconductor-optical-amplifier-based symmetric-Mach–Zehnder-type all-optical switch. *Applied Physics Letters* 78(25): 3929–3931.

Ohara T, Takara H, Shake I, *et al.* (2003) 160-Gb/s optica-time-division multiplexing with PPLN hybrid integrated planar lightwave circuit. *IEEE Photonics Letters* 15(2): 302–304.

Ramaswami R and Sivarajan KN (2002) *Optical Network; a Practical Perspective*. Morgan Kaufman.

Reman E (2001) Trends and evolution of optical networks and technology. *Alcatel Telecommun. Rev. 3rd Quarter* 173–176.

Sabella R and Lugli P (1999) *High Speed Optical Communications*. Dordrecht: Kluwer Academic.

Seo SW, Bergman K and Prucnal PR (1996) Transparent optical networks with time division multiplexing. *Journal of Lightwave Technology* 14(5): 1039–1051.

Sokolloff JO, Prucnal PR, Glesk I and Kane M (1993) A terahertz optical asymmetric demultiplexer (TOAD). *IEEE Photonics Technology Letters* 5(7): 787–790.

Stavdas A (ed.) (2001) *New Trends in Optical Network Design and Modelling*. Dordrecht: Kluwer Academic.

Ueno Y, Takahashi M, Nakamura S, Suzuki K, Shimizu T, Furukawa A, Tamanuki T, Mori K, Ae S, Sasaki T and Tajima K (2003) Control scheme for optimizing the interferometer phase bias in the symmetric-Mach–Zehnder all-optical switch (invited paper). Joint special issue on recent progress in optoelectronics and communications. *IEICE Trans. Electron.* E86-C(5): 731–740.

Coherent Analog Optical Processors

H H Arsenault, S Roy and D Lefebvre, University of Laval, Quebec, PQ, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Linear filtering is a useful tool commonly used in image processing. The correlation operation can be defined as:

$$f(x,y)*h(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)h(x-x',y-y')\mathrm{d}x'\mathrm{d}y' \quad [1]$$

where $*$ denotes the correlation. The operation of correlation and the Fourier transform

$$G(\mu,\nu)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x,y)\mathrm{e}^{-j2\pi(\mu x+\nu y)}\mathrm{d}x\mathrm{d}y \quad [2]$$

are intimately related through the correlation theorem of Fourier theory, which states that the correlation between two functions $f(x,y)$ and $h(x,y)$ is equal to the inverse Fourier transform of the product of the Fourier transforms:

$$f(x,y)*h(x,y)=FT^{-1}[F(\mu,\nu)H^{*}(\mu,\nu)] \quad [3]$$

where $F(\mu,\nu)$ and $H(\mu,\nu)$ are the Fourier transforms of $f(x,y)$ and of $h(x,y)$ respectively, and where H^{*} designates the complex conjugate of H. The ability to carry out and display Fourier transforms in two dimensions, by means of optical systems, is at the root of optical information processing.

One of the main advantages of correlation is that the operations do not require segmentation of the regions of interest. This means that the locations of objects do not have to be segmented from the scene and that the detection is shift-invariant, that is the correlation peaks appear at locations in the scene corresponding to locations of the targets. These advantages are clearer for detection in uncontrolled conditions, for example for detecting vehicles in natural scenes. But they are also useful in security applications, such as identity verification or for detecting defects in assembly lines.

The Optical Fourier Transform

The Fourier transform of eqn [2] is widely used in the field of information processing. The inverse Fourier transform is simply defined as:

$$g(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(\mu,\nu)\mathrm{e}^{i2\pi(\mu x+\nu y)}\mathrm{d}\mu\mathrm{d}\nu \quad [4]$$

We will now discuss how to perform Fourier transforms by means of four different optical setups.

FT Without a Lens

If we illuminate a transparency $U(x_0,y_0)$ with a coherent plane wave, the diffraction pattern we observe at a point d from the input plane is given by the following relation, known as the Fresnel approximation for diffraction (Figure 1):

$$U(x_1,y_1)=\frac{\exp\left(i\dfrac{2\pi d}{\lambda}\right)}{j\lambda d}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}U(x_0,y_0)\times\exp\left\{i\frac{\pi}{\lambda d}[(x_0-x_1)^2+(y_0-y_1)^2]\right\}\mathrm{d}x_0\mathrm{d}y_0 \quad [5]$$

where λ is the wavelength of the light and d is the distance between the diffracting screen and the observed pattern.

In order to use the Fresnel approximation we need only to satisfy the geometrical optics paraxial conditions. If we add an additional approximation, we can further simplify this equation to yield the Fourier transform of $U(x_0, y_0)$. The extra condition is known as the Fraunhofer or far-field approximation, which is only satisfied when the diffracted image is observed far away from the diffracting screen or when the extent of $U(x_0, y_0)$ is very small. Mathematically this condition is expressed by:

$$\frac{\pi}{\lambda d}(x_0^2+y_0^2)\ll 1 \quad [6]$$

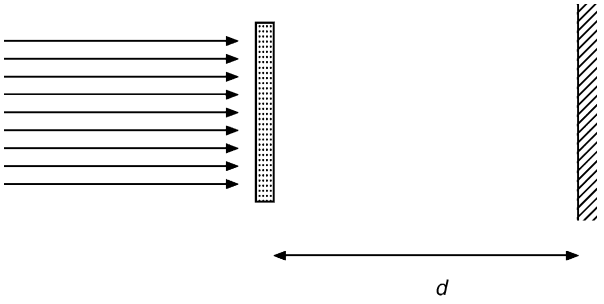


Figure 1 Fresnel diffraction without a lens.

After expanding the squares in the exponentials, there will be three terms: one can go outside the integral and one will be approximately equal to unity so the diffracted pattern becomes

$$U(x_1, y_1) = \frac{\exp\left(i\frac{2\pi d}{\lambda}\right)}{i\lambda d} \times \exp\left[i\frac{\pi}{\lambda d}(x_1^2 + y_1^2)\right] \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_0, y_0) \times \exp\left[-i\frac{2\pi}{\lambda d}(x_0 x_1 + y_0 y_1)\right] dx_0 dy_0 \quad [7]$$

This is the Fourier transform of $U(x_0, y_0)$, scaled by a factor λd, with an extra quadratic phase factor

$$\exp\left[\frac{i\pi}{\lambda d}(x_1^2 + y_1^2)\right]$$

The factor $\exp(i2\pi d/\lambda)/i$ is usually of no importance since it is a phase factor distributed across the output plane. In some applications we only want to know the power spectrum (squared modulus) of the Fourier transform. In such cases the quadratic phase factor disappears because the squared modulus of any complex exponential is equal to 1. Performing a Fourier transform without a lens is possible within the Fraunhofer approximation. But if it is required to obtain the Fourier transform of an image which extends about 1 cm from the optical axis using a coherent He–Ne laser ($\lambda = 632.8$ nm), the condition to be satisfied is $d \gg 496$ m: the required observation distance d is at least a few kilometers. This kind of optical setup is usually impractical, but can fortunately be alleviated by means of lenses.

Object on the Lens

Instead of having only the transparency in the input plane, we place a lens immediately after it as shown in Figure 2. For mathematical simplicity the lens and input plane are superimposed (more realistic cases are considered below). The lens acts as a phase factor

$$\exp\left[-i\frac{\pi}{\lambda f}(x_0^2 + y_0^2)\right]$$

where f is the focal length of the lens. Now in the input plane the object $U(x_0, y_0)$ is multiplied by the phase factor of the lens:

$$U'(x_0, y_0) = U(x_0, y_0)\exp\left[-i\frac{\pi}{\lambda f}(x_0^2 + y_0^2)\right] \quad [8]$$

Inserting this expression into the integral formula for Fresnel diffraction yields the diffraction pattern observed in the focal plane of the lens:

$$U(x_1, y_1) = \frac{\exp\left(i\frac{2\pi f}{\lambda}\right)}{i\lambda f} \times \exp\left[i\frac{\pi}{\lambda f}(x_1^2 + y_1^2)\right] \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_0, y_0) \times \exp\left[-i\frac{2\pi}{\lambda f}(x_0 x_1 + y_0 y_1)\right] dx_0 dy_0 \quad [9]$$

The lens exactly compensates for the other phase factor that was in the integral. This expression for the Fraunhofer approximation is almost the same as before, but the focal length f of the lens replaces the distance d. So by placing a lens immediately after the input plane we can have a much more compact setup since lenses are available with a wide range of focal lengths, usually from a few centimeters to a few meters. Another way to look at this is to consider that the effect of a lens is to bring the pattern at infinity to the focal plane of the lens – except for the phase factor.

Object Before the Lens

Another setup used to perform the Fourier transform is shown in Figure 3, where the lens is placed somewhere between the input and the output planes. Using the Fresnel formula for diffraction it can be

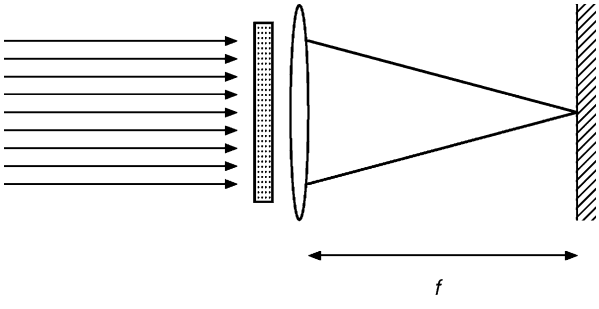


Figure 2 Diffraction with the object on the lens.

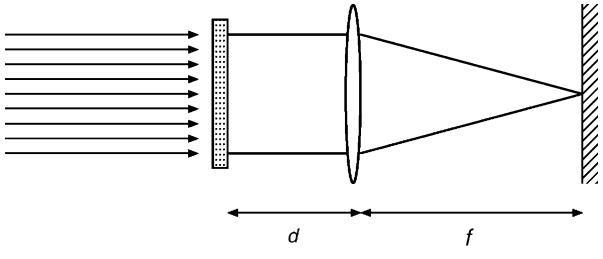


Figure 3 Diffraction with the object before the lens.

shown that the diffraction pattern in the focal plane is:

$$U(x_1, y_1) = \frac{1}{i\lambda f} \times \exp\left[i\frac{\pi}{\lambda f}\left(1 - \frac{d}{f}\right)(x_1^2 + y_1^2) \right] \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_0, y_0) \times \exp\left[-i\frac{2\pi}{\lambda f}(x_0 x_1 + y_0 y_1) \right] dx_0 dy_0 \quad [10]$$

where d is the distance between the input plane and the lens. If we set d equal to the focal length f of the lens, the quadratic phase factor in front in the integral vanishes and we have an exact Fourier transform. This setup with the object and filter in the front and back focal planes of the lens respectively is the one most frequently seen in scientific articles and is known as the *f-f* Fourier transform system.

Object After the Lens

Still another configuration is the one shown in Figure 4, where the lens is placed at a distance d before the input plane. For this configuration the Fresnel diffraction yields the following pattern in the focal plane of the lens:

$$U(x_1, y_1) = \frac{\exp\left[i\frac{\pi}{\lambda d}(x_1^2 + y_1^2) \right]}{i\lambda d} \frac{f}{d} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_0, y_0) \times \exp\left[-i\frac{2\pi}{\lambda d}(x_0 x_1 + y_0 y_1) \right] dx_0 dy_0 \quad [11]$$

Once again the Fourier transform of $U(x_0, y_0)$ is multiplied by a quadratic phase factor. It is interesting to note that, in this case, the Fourier transform is scaled by a factor λd. This can be used to control the scale of the Fourier transform by changing the distance between the lens and the object.

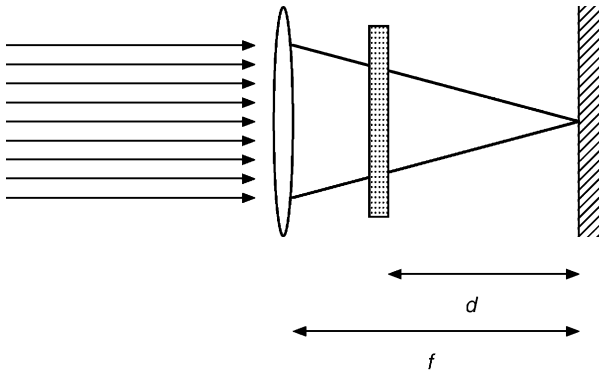


Figure 4 Diffraction with the object after the lens.

Note that although two of the three methods yield Fourier transforms that are accompanied by a quadratic phase factor, the latter is not necessarily an impediment to correlation, contrary to popular belief. Indeed all three configurations have been used for correlation pattern recognition, and each configuration has its advantages.

Optical Correlation

The last section showed how to obtain Fourier transforms by means of an optical setup. The proposed optical setups cannot perform the inverse Fourier transform. Because the Fourier transform and its inverse are very similar, the Fourier transform is performed instead, yielding the mirror image of the correlation:

$$(f * h)(-x, -y) = \mathrm{FT}[F(\mu, \nu)H^*(\mu, \nu)] \quad [12]$$

The correlation can be considered as the result of three operations: a Fourier transform; a product; and an additional Fourier transform. The Fourier transform of a 2D function can be obtained by means of coherent optics. The equivalent of multiplication in optics is transmissivity, so multiplication is accomplished by sending light successively through two superimposed transparencies.

There are two main categories of optical correlators, serial correlators and joint transform correlators.

Serial Correlators

A serial correlator functions in two stages. The first stage performs the Fourier transform of the input image $f(x, y)$, which is multiplied in the Fourier plane with a transparency $H(\mu, \nu)$. The filter $H(\mu, \nu)$ can be a hologram, a spatial light modulator (SLM) or a simple transparency. A second Fourier transform displays in the output plane the correlation of the input image with the impulse response $h(x, y)$, which is the Fourier transform of the filter function $H(\mu, \nu)$.

Serial correlators directly implement the two Fourier transforms and the multiplication operation in a pipelined manner. The input plane and the correlation plane are conjugate planes of the system, that is in the absence of a filter, the output plane displays an inverted image of the input $f(x, y)$. Between those conjugate planes, at the conjugate plane of the source with respect to the first lens, is the Fourier plane where the spatial frequency filter is placed. When the input is illuminated with parallel

light, the frequency plane is at the back focal plane of the first lens. A SLM can be placed in the input plane to feed the images to be correlated; a second SLM can be placed in the Fourier plane to display the filter. The filter is often calculated by a digital computer, and usually takes the form of a hologram if phase information must be used in the filter.

The 4-*f* correlator

The 4-*f* correlator is the most intuitive optical correlator system. We showed above that an *f*-*f* system can be used to optically obtain a Fourier transform (Figure 5). If $f(x, y)$ is in the front focal plane, its Fourier transform $F(\mu, \nu)$ appears at the back focal plane of the first lens. If a filter function $H^*(\mu, \nu)$ is inserted at this point, by means of a SLM or a hologram, then a second *f*-*f* system can perform a second Fourier transform to yield the correlation plane at the back focal plane of the second lens.

Convergent beam correlator

This kind of correlator consists of two conjugate systems. This time, the Fourier transform is obtained by placing the input object in a convergent beam after the first lens. The second lens is placed so that the output plane is conjugate to the input plane, and as before, the filter is placed at the point of convergence of the light after the first lens. It is important to note that this is not the focal plane of the lens, but the conjugate plane of the source of light. In order to have an imaging system between the input and output planes, the distances must satisfy the conjugate relation:

$$\frac{1}{s_0} + \frac{1}{s_i} = \frac{1}{f_2} \qquad [13]$$

where s_i and s_0 are defined in Figure 6. As previously discussed for convergent light, the Fourier transform will have an additional parabolic phase factor. This must be taken into account when constructing the filter H.

This configuration has some advantages over the 4-*f* correlator. By moving the object between the first lens and the filter, it is possible to control the scale of the Fourier transform. And because the first lens is imaging only a single point on the optical axis, only spherical aberration needs to be compensated for. In 4-*f* systems, the first lens must be corrected for all aberrations for the whole input plane, and for a given extent of the input object, the first lens must be larger than that for the convergent system in order to avoid space variance effects. Unfortunately, specially designed expensive lenses are required to correct for those aberrations, so well-corrected Fourier transform lenses for 4-*f* systems are expensive, whereas a simple doublet is sufficient to correct for on-axis spherical aberration, which is all that is required for the convergent light system. In addition, the convergent light system does not require an expensive collimator that is required in the 4-*f* system to obtain the parallel illuminating beam. Aberration correction requirements for the second lens are less stringent

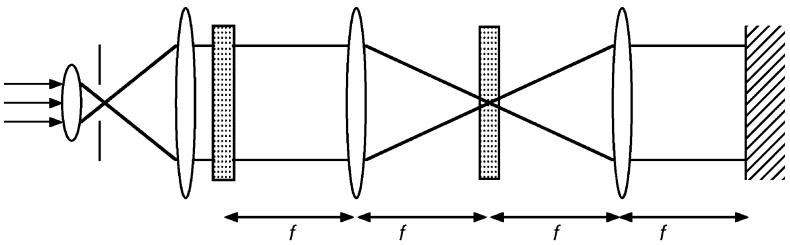


Figure 5 4-*f* correlator setup.

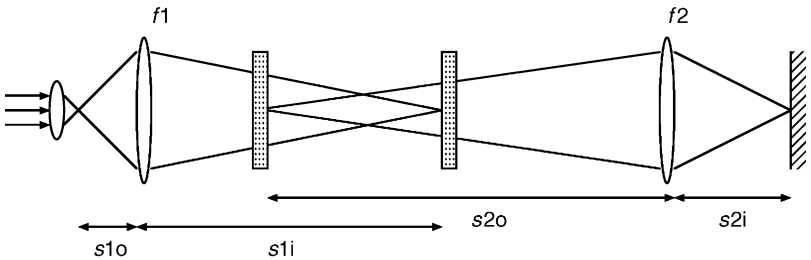


Figure 6 Convergent beam correlator setup.

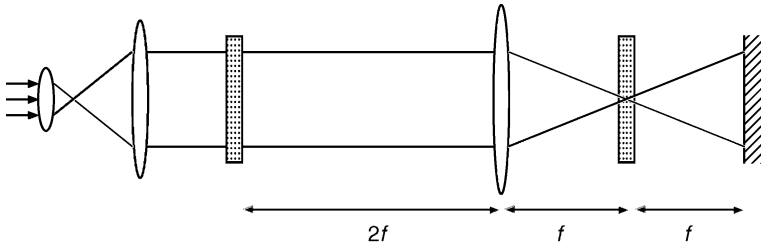


Figure 7 Single lens correlator setup.

than for the first, because in most cases, a slight deformation of the output plane can be acceptable, whereas comparable distortions in the Fourier plane can cause misregistrations between the Fourier transform and the filter that can seriously degrade the correlation.

Divergent beam correlators

Instead of placing the input object in the convergent beam, as in the previous case, another configuration that has been used places the input object in the divergent beam before the first lens. This configuration is rarely seen today, because it exacerbates the space variance and aberration problems of the 4-f system, and the control it allows over the scale of the Fourier transform can be achieved with fewer penalties by means of the convergent light system.

Single lens correlator

Consider an imaging system using only one lens illuminated with a parallel beam of coherent light (Figure 7). The input and output planes are located at distances $2f$ on either side of the lens – the shortest conjugate distances possible. If a filter $H(\mu, \nu)$ is placed at the focal plane of the lens on the object's Fourier transform, the correlation will appear on the output plane. This system requires satisfying impractical conditions: aberration corrections for both the Fourier transform and for the imaging operation, so it is rarely seen.

Filters

Many types of filters can be used for different applications. Table 1 lists some commonly used filters for pattern recognition. The classical matched filter yields the optimum signal to noise ratio in the presence of additive noise. The phase-only filter (POF) is frequently used because of its high discrimination capability. The inverse filter is sometimes used because it theoretically yields very sharp correlation peaks, but is very sensitive to noise, and when the spectrum $F(\mu, \nu)$ has some zeros, the realization of $1/F(\mu, \nu)$ can be impossible. A close relative of the inverse filter is the Wiener filter, which avoids the problem of the inverse filter by introducing a quantity $S(\mu, \nu)$ in the denominator of the filter function. The function $S(\mu, \nu)$ can sometimes be a constant. $H(\mu, \nu)$ is usually chosen to minimize the effect of noise or to optimize some parameter such as the peak to sidelobe ratio, or the discrimination ability of the system.

Table 1 Commonly used filters for image processing

Name	*Filter*
Matched filter	$H_{\mathrm{CMF}}(\mu, \nu) = F(\mu, \nu)$
Phase-only filter	$H_{\mathrm{POF}}(\mu, \nu) = \dfrac{F(\mu, \nu)}{\lvert F(\mu, \nu)\rvert}$
Inverse filter	$H_{\mathrm{INV}}(\mu, \nu) = \dfrac{1}{F(\mu, \nu)}$
Wiener filter	$H_{\mathrm{W}}(\mu, \nu) = \dfrac{F^{*}(\mu, \nu)}{\lvert F(\mu, \nu)\rvert^{2} + S(\mu, \nu)}$

Joint Transform Correlators

In serial transform correlators, the filter is generally a complex function. Spatial light modulators are usually conceived with only one degree of freedom for each pixel, which can be optimized for amplitude modulation or for phase modulation. In order to achieve a fully complex filter, two modulators are required, or some spatial bandwidth of a modulator is sacrificed to encode a computer-generated hologram.

The joint-transform correlator (JTC) does not require a fully complex filter. The filter is stored in the spatial domain so there is usually no need for the computation of a filter in the spectral domain. In the JTC, the correlation of a scene with a reference image is contained in the power spectrum of an input plane consisting of the scene placed side by side with the reference. The correlation terms are revealed when the Fourier transform of the power spectrum is

computed. The joint transform correlation can be summarized in three steps:

1. The scene and the reference are located side by side in the input plane.
2. The Fourier transform of the input plane is computed or carried out optically by one of the methods previously described. The transform intensity is measured and constitutes the joint power spectrum. Some optional enhancements, based on spatial filtering and/or nonlinear processing on the power spectrum, can also be performed.
3. The Fourier transform of this joint power spectrum is carried out, yielding the correlation plane containing the correlation of the reference with the scene.

A mathematical analysis is required to understand the location of the correlation peaks. The side-by-side input scene $s(x, y)$ and reference image $r(x, y)$ are

$$i(x, y) = r(x - x_r, y - y_r) + s(x - x_s, y - y_s) \qquad [14]$$

The joint power spectrum $E(\mu, \nu)$ is equal to the modulus squared of the Fourier transform of the input plane:

$$\begin{aligned} I(\mu, \nu) &= R(\mu, \nu)e^{-i2\pi(\mu x_r + \nu y_r)} + S(\mu, \nu)e^{-i2\pi(\mu x_s + \nu y_s)} \\ E(\mu, \nu) &= \|I(\mu, \nu)\|^2 = \\ &\quad \|R(\mu, \nu)\|^2 + \|S(\mu, \nu)\|^2 \\ &\quad + R(\mu, \nu)S^*(\mu, \nu)e^{-i2\pi[\mu(x_r - x_s) + \nu(y_r - y_s)]} \\ &\quad + R^*(\mu, \nu)S(\mu, \nu)e^{-i2\pi[\mu(x_s - x_r) + \nu(y_s - y_r)]} \end{aligned} \qquad [15]$$

The joint power spectrum contains the Fourier transforms of the autocorrelation of the input scene, the autocorrelation of the reference, the correlation of the scene with the reference and its complex conjugate. The correlation peaks are obtained from the Fourier transform of the joint power spectrum:

$$\begin{aligned} C(x', y') = {} & R_{ss}(x', y') + R_{rr}(x', y') \\ & + R_{sr}(x' - (x_r - x_s),\ y' - (y_r - y_s)) \\ & + R_{sr}(x' + (x_r - x_s),\ y' + (y_r - y_s)) \end{aligned} \qquad [16]$$

where

$$R_{ab}(x', y') = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a(\xi, \zeta)b^*(\xi - x', \zeta - y')\mathrm{d}\xi\mathrm{d}\zeta \qquad [17]$$

Multiple targets and multiple references

The input scene of the JTC may contain multiple targets. Each target in the scene will correlate with the other targets. The additional correlation peaks can clutter the correlation plane, so some care must be taken when positioning the scene with respect to the reference. Let $s(x, y)$ contain multiple targets $s_i(x, y)$:

$$s(x, y) = \sum_{i=1}^{N} s_i(x - x_i, y - y_i) \qquad [18]$$

The joint power spectrum becomes:

$$\begin{aligned} E(\mu, \nu) = {} & \|R(\mu, \nu)\|^2 + \sum_{i=1}^{N} \|S_i(\mu, \nu)\|^2 + \sum_{i=1}^{N} \sum_{\substack{k=1 \\ i \neq k}}^{N} \\ & \times S_i(\mu, \nu)S_k^*(\mu, \nu)e^{-i2\pi[\mu(x_i - x_k) + \nu(y_i - y_k)]} + \text{C.C.} \\ & + \sum_{i=1}^{N} R(\mu, \nu)S_i^*(\mu, \nu)e^{-i2\pi[\mu(x_r - x_i) + \nu(y_r - y_i)]} + \text{C.C.} \end{aligned} \qquad [19]$$

where C.C. denotes the complex conjugate of the preceding term. The correlation plane will now be:

$$\begin{aligned} C(x', y') = {} & R_{rr}(x', y') + \sum_{i=1}^{N} R_{s_i s_i}(x', y') \\ & + \sum_{i=1}^{N} \sum_{\substack{k=1 \\ i \neq k}}^{N} R_{s_i s_k}(x' - (x_i - x_k),\ y' - (y_i - y_k)) \\ & + \sum_{i=1}^{N} R_{rs_i}(x' - (x_r - x_i),\ y' - (y_{ri} - y_i)) \\ & + \sum_{i=1}^{N} \sum_{\substack{k=1 \\ i \neq k}}^{N} R_{s_k s_i}(x' + (x_i - x_k),\ y' + (y_i - y_k)) \\ & + \sum_{i=1}^{N} R_{s_i r}(x' + (x_r - x_i),\ y' + (y_{ri} - y_i)) \end{aligned} \qquad [20]$$

The second lines of eqns [9] and [20] are the cross-correlation terms between the targets. In order to avoid confusion between reference–target correlations and target–target correlations, special care must be taken in positioning the reference with respect to the scene. All the target–target correlations will be inside a centered box that has twice the length and twice the width of the input scene, so the reference should be placed at least at twice the length or width of the scene.

Multiple targets in the input scene can be viewed as multiple references, and the correlation of multiple

references with multiple targets are all computed simultaneously. This property has led to applications in multiclass pattern recognition problems and in optical neural networks, where each neuron of a layer is coded as a reference object. This is one of the advantages of the JTC compared to serial correlators, since serial correlators are limited to one filter at a time.

Example

Here is an example of joint transform correlation. The input plane is shown in Figure 8a. The plane is composed of the input scene in the top third of the input plane and the reference is placed in the center. This input scene contains two targets. The joint power spectrum is shown in Figure 8b with its pixels mapped to a logarithmic intensity scale. The correlation plane is shown in Figure 8c and d with the central peak truncated. The central peak contains the autocorrelation terms R_{ss} and R_{rr}. On each side of the autocorrelation peak are the cross-correlations between the different targets of the input scene. The correlation between the reference image and the input scene is located in the top part of the correlation plane. The correlation plane is symmetrical with respect to the origin, so the correlation peaks are reproduced on the bottom half of the output plane.

Joint transform correlator architectures

Different architectures have been used for JTCs. Three setups are presented here. This is not an exhaustive list. The simplest optical setup shown in Figure 9 is an optoelectronic implementation of a power spectrum machine. A convergent beam

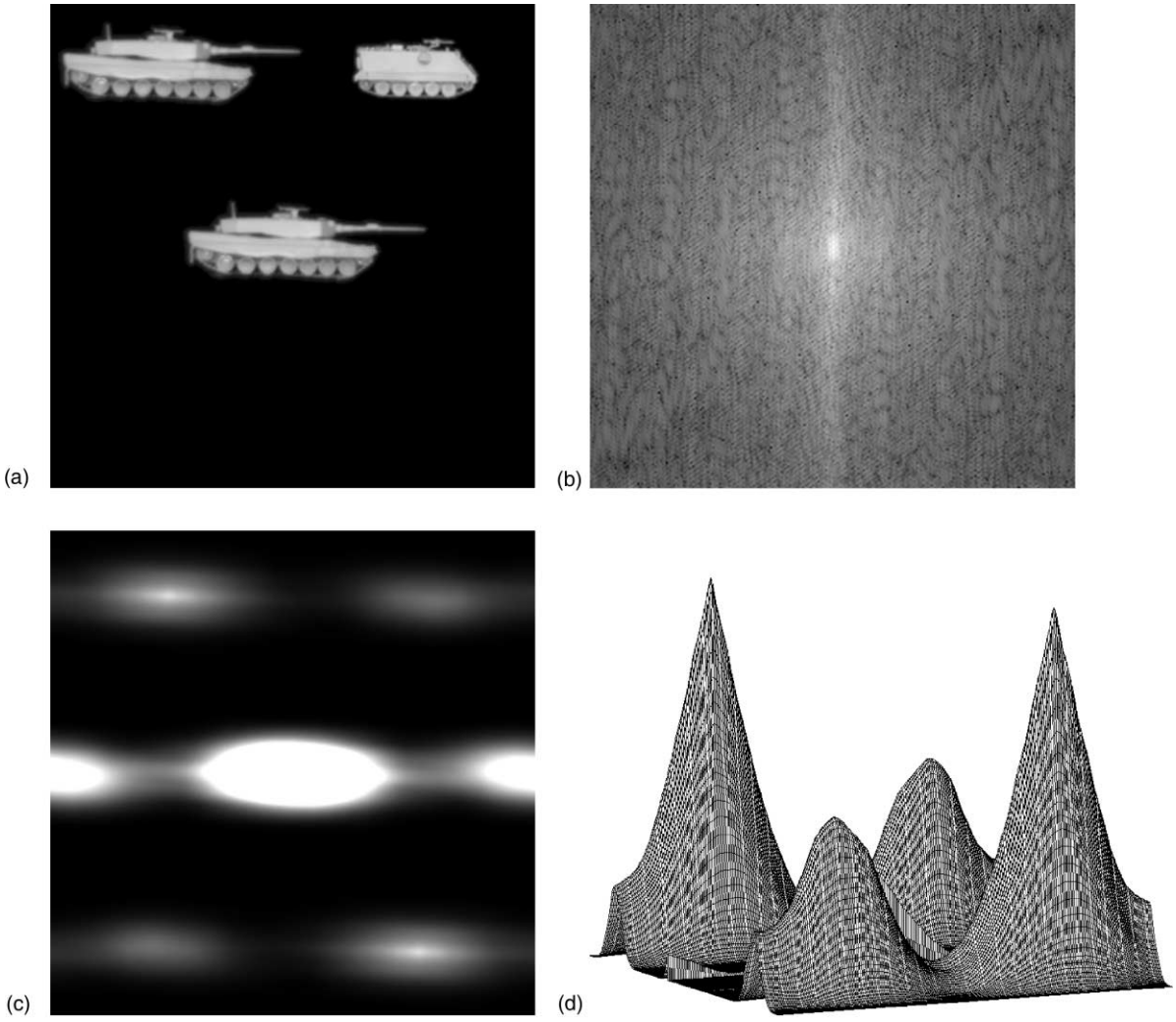


Figure 8 Joint transform correlation example. (a) Input scene; (b) joint power spectrum; (c) correlation plane (top view); (d) correlation plane (mesh view).

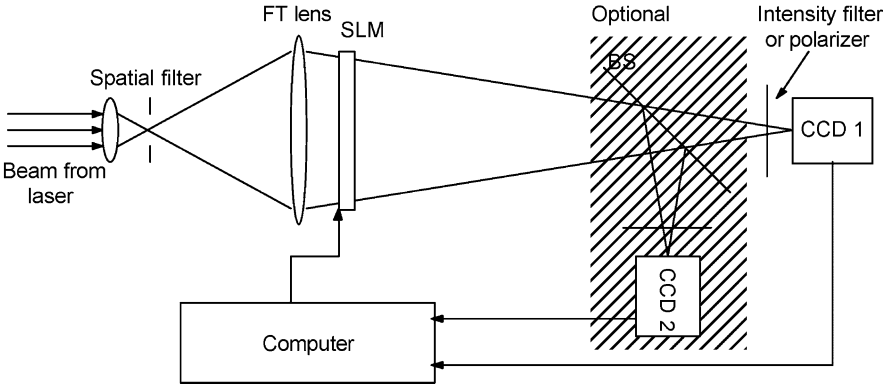


Figure 9 Joint transform correlator implementation as an optical power spectrum machine.

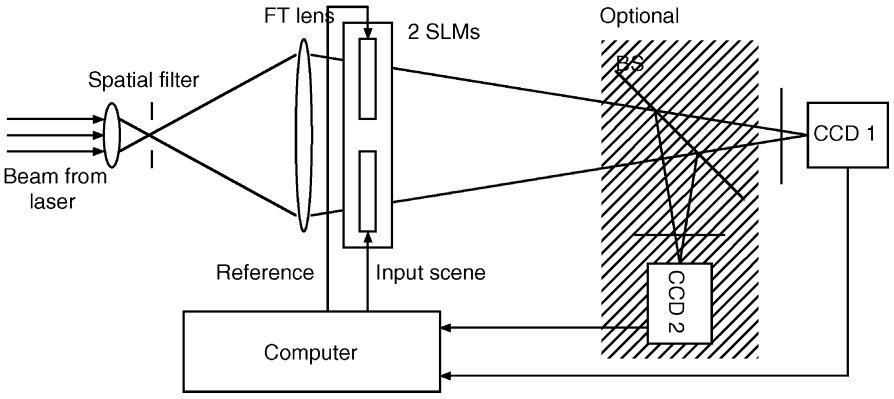


Figure 10 Optical JTC using two SLMs in the input plane.

optical Fourier transform system is used to produce the Fourier transform of an image displayed on a single spatial light modulator (SLM). A CCD camera is placed in the Fourier plane to measure the joint power spectrum. The joint transform correlation is a two-step process: the input plane containing the scene and the reference is first displayed on the SLM, and the camera simultaneously captures the joint power spectrum. In the second step, the joint power spectrum is displayed on the SLM, so the correlation plane is now on the camera. An optional second camera with intensity filters can be added to this setup. In this case, one camera serves to digitize the joint power spectrum during step 1 and the second camera captures the correlation plane at step 2 and transmits it to a display device. This optional configuration is useful to keep both the correlation plane and the joint power spectrum within the dynamic range of their specific cameras.

The setup shown in Figure 9 has two main disadvantages: the correlation is performed in two steps and requires some processing to construct the input plane, and only half of the spatial light modulator can be used to display the input scene. There are alternative setups to alleviate those drawbacks, but they require the use of two or three SLMs. Two modulators can be used in the input plane to simulate a full modulator for the scene, as shown in Figure 10. The original setup can also be doubled with one optical Fourier transform system for the joint power spectrum and one for the correlation plane (Figure 11). With three modulators, it is possible to perform the joint transform correlation in a single step with a full modulator for the input scene.

Comparison to serial transform correlators

The joint transform correlator has some advantages over the serial transform correlator. Since the phase of the Fourier transform is destroyed in the JTC, there are much fewer constraints on the design. So it is advantageous to use an optical convergent beam setup to obtain the Fourier transform. As mentioned before, there is no need for a fully complex spatial light modulator in the joint transform correlator.

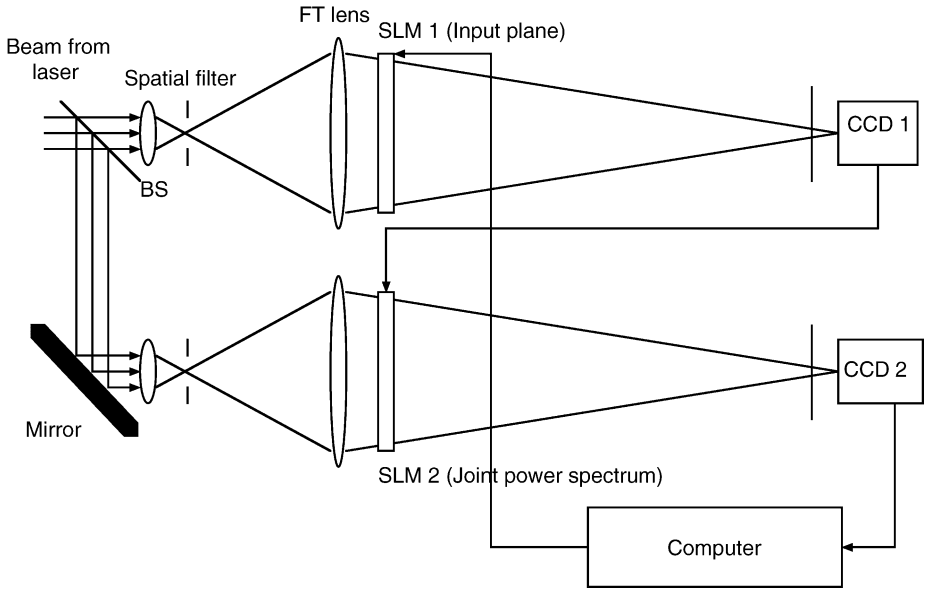


Figure 11 JTC implementation using two power spectrum machines.

This allows the use of commercially available spatial light modulators primarily designed for projectors or for displays, where phase modulation is not required. Since there is a wider market for such devices, they are cheaper.

For pattern recognition, the joint transform correlator has some properties that a serial transform correlator cannot reproduce. Some pattern recognition techniques require a linear combination of multiple amplitude correlations of the scene with different references. This is impossible with serial correlators because only the intensity of the correlation plane is detected. In a JTC, the joint power spectrum is linked to the correlation plane by the Fourier transform. So it is possible to obtain linear combinations of correlation planes by performing the corresponding operations on the joint power spectra. Some recent pattern recognition techniques such as morphological correlation and the sliced orthogonal nonlinear generalized (SONG) correlation, have been optically implemented using a JTC, but the required operations could not be carried out by means of serial optical correlators.

Whereas optical serial transform correlators always carry out linear filtering operations, optical joint transform correlators can be made to carry out nonlinear filtering. This is achieved by imposing a nonlinear characteristic curve on the joint power spectrum, instead of the usual squaring operation inherent to the detection process. The nonlinearity can be intrinsic to the components (saturation, dynamic range) or obtained by introducing electronics between the CCD camera and the second SLM,

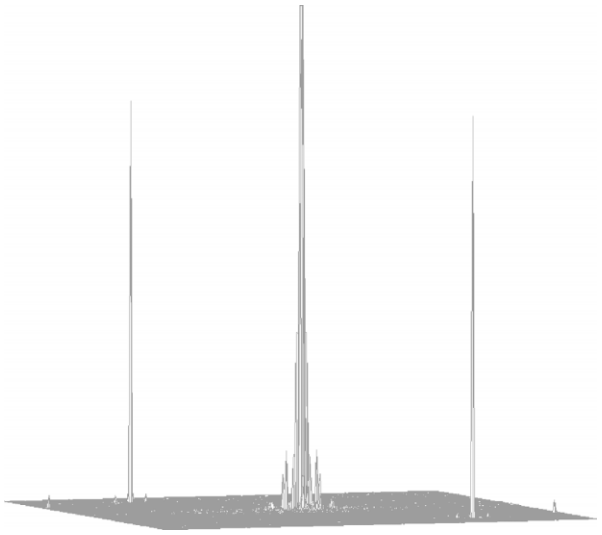


Figure 12 Correlation plane with a logarithmic transformation of the joint power spectrum.

or by using an optically addressed spatial light modulator. Nonlinear processing can enhance the weights of the high frequencies in the correlation, resulting in sharper correlation peaks and improved discrimination. To illustrate this effect, the joint power spectrum of the example of Figure 8 has been modified with a logarithmic transformation:

$$E(\mu, \nu) = \ln\left(1 + \beta \|I(\mu, \nu)\|^2\right) \qquad [21]$$

with $\beta = 0.25$. The correlation plane for the transformed joint power spectrum is shown in Figure 12.

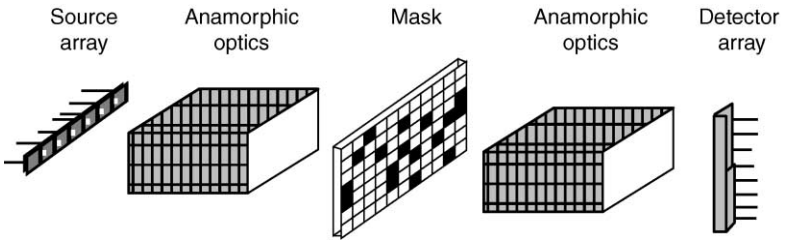


Figure 13 Optical vector–matrix multiplier.

It is clear that the discrimination has improved over that of the previous case.

Optical Vector–Matrix Multiplier

An operation that is frequently required in information processing is the product of a matrix and a vector. This can be accomplished by means on anamorphic optics as shown in Figure 13.

Here the input vector is a set of vertical fans of light from the source array, that are spread out, each over one row of the matrix mask by means of anamorphic optics. The inverse operation is carried out in the second half, so that the vertical array of spots in the output contains the vector–matrix product between the intensities of the source array and the transmissivities of the mask array.

A vector–matrix product can also be carried out by means of a conventional optical correlator as shown in Figure 14.

The input vector **a** is placed in a parallel beam at the focal plane of a conventional 2D correlator. A hologram in the Fourier plane contains a vector array **H**. The vector–matrix product **aH** appears along the axis in the correlation plane.

The above figure shows an input vector $f(x,b)\delta(y-b)$ containing the vector components $f(x1,x2,\ldots,xN)$ arrayed along the horizontal axis x, and shifted a distance a from the horizontal axis. If the Fourier transform of the mask in the filter plane is $h(x,y)$, then the output plane has the distribution:

$$g(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\mu,\upsilon)\delta(\mu-b)h(x+\mu,y+\upsilon)\mathrm{d}\mu\mathrm{d}\upsilon \qquad [22]$$

$$g(x,y)=\int_{-\infty}^{\infty} f(\mu,a)h(x+\mu,b-y)\mathrm{d}\mu \qquad [23]$$

$$g(x,b)=\int_{-\infty}^{\infty} f(\mu,b)h(x+\mu,b)\mathrm{d}\mu \qquad [24]$$

$$g(0,b)=\int_{-\infty}^{\infty} f(\mu,b)h(\mu,b)\mathrm{d}\mu \qquad [25]$$

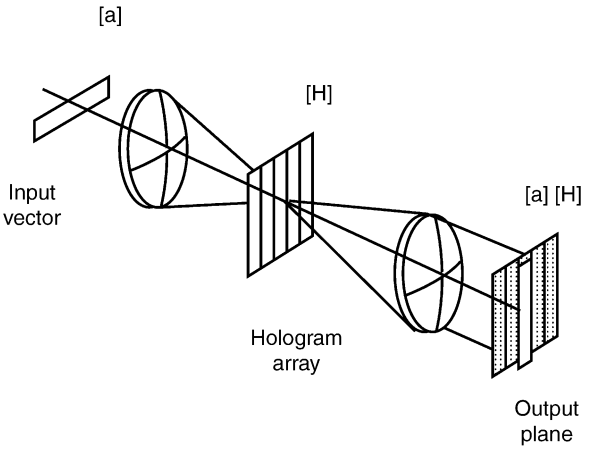


Figure 14 A vector–matrix product using a conventional 2D optical correlator.

So the output value along the y-axis at a distance a from the horizontal axis is equal to the scalar product of the vector $f(x,b)$ with the vector $h(x,b)$. Now if input vectors are stacked up one over the other in the input plane, each one will yield along the output y-axis a value that is the scalar product of the corresponding vector with the vector **h**. The set of values along the y-axis is therefore the vector–matrix product of the matrix constituted by the set of vectors in the input plane, with the vector **h** that is the impulse response of the filter. The cost for this operation is some considerable loss of light, since only the light along the y-axis is used in the output.

Applications of the optical vector–matrix product are optical crossbar, numerical matrix operations such as the solving of linear equations, optical neural networks, spectrum analysis and discrete Fourier transforms.

Conclusion

It has been estimated that serial coherent optical correlators should be able to attain data rates of up to 10^{12} operations per second, using the fastest spatial

light modulators available. Joint transform correlators are somewhat slower, due to the input plane being divided between the object and the reference, to geometrical constraints in placement of targets, and due to the need to use electronics to carry out some of the operations. Until now, speeds have been limited by the data rates at the input, but as the speeds of optically addressed SLMs increase, the problem could be shifted to the output, where the lack of optical high spatial bandwidth threshold devices could cause a bottleneck.

List of Units and Nomenclature

A	Vector or matrix
CCD	Charge coupled device
$F(\mu, \nu)$	Fourier transform of $f(x, y)l$
FT	Fourier transform
FT^{-1}	Inverse Fourier transform
H^*	Complex conjugate of H
JTC	Joint-transform correlator
POF	Phase-only filter
R_{ab}	Correlation between functions a and b
(x, y)	Spatial coordinates
λ	Wavelength of light
(μ, ν)	Spatial frequencies
$*$	Correlation
SLM	Spatial light modulator
SONG	Sliced orthogonal nonlinear generalized correlation

See also

Coherence: Coherence and Imaging. **Diffraction:** Fraunhofer Diffraction; Fresnel Diffraction. **Fourier Optics**. **Geometrical Optics:** Aberrations. **Holography, Techniques:** Computer-Generated Holograms. **Imaging:** Information Theory in Imaging. **Information Processing:** Incoherent Analog Optical Processors; Optical Neural Networks.

Further Reading

Arsenault HH, Szoplik T and Macukow B (eds) (1989) *Optical Processing and Computing*. New York: Academic Press.

Bartelt H (1987) Unconventional correlators. In: Horner JL (ed.) *Optical Signal Processing*, pp. 97–128. New York: Academic Press.

Garcia-Martinez P, Arsenault HH and Ferreira C (2001) Time sequential nonlinear correlations for optical pattern recognition. Optoelectronic Information processing: optics for information systems. *Proceedings of SPIE* CR81-12: 186–207.

Goodman JW (1988) *Introduction to Fourier Optics*. New York: McGraw-Hill.

Javidi B (1994) Nonlinear joint transform correlators. In: Javidi B and Réfrégier P (eds) *Workshop on Optical Pattern Recognition*. Bellingham, WA: SPIE – The International Society for Optical Engineering, PM-12.

Javidi B and Horner JL (eds) (1994) *Real-Time Optical Information Processing*. New York: Academic Press.

Javidi B and Johnson KM (eds) (1996) *Optoelectronic Devices and Systems for Processing*. Bellingham, WA: SPIE – The International Society for Optical Engineering, CR65.

Lu XL, Yu FTS and Gregory D (1990) Comparison of Vander Lugt and joint transform correlators. *Applied Physics B* 51: 153–164.

McAulay AD (1991) *Optical Computer Architectures*. New York: Wiley & Sons.

Preston K (1972) *Coherent Optical Computers*. New York: McGraw-Hill.

Vander Lugt AB (1992) *Optical Signal Processing*. New York: John Wiley & Sons.

Free-Space Optical Computing

A A S Awwal, Lawrence Livermore National Laboratory, Livermore, CA, USA

M Arif, Infineon Technologies, Longmont, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Free space optical computing encompasses a vast field, from digital and discrete to analog and hybrid optical computing. This article gives an overview of various algorithms, such as shadow-casting, symbolic substitution to nonlinear optics-based processors. Some of the core devices, such as modulators and nonlinear logic elements, are discussed. The optical computing concepts, originated with nonlinear optics and analog optical computing, are more commonly known as optical signal processing. Now they have culminated into digital optical computing, opening the door to highly specialized parallel computing architecture.

Parallel computation is a direct approach to enhance the speed of digital computations which are otherwise inhibited by the sequential processing bottleneck. Physical limitations – such as limited interconnections, interactions amongst electrons, with the consequent inability to share a data bus simultaneously, and constant propagation delay of electronic circuits – place an upper bound on the speed of the fastest operation possible with an electronic computer, without a modification to the computing architecture. Optics provides an architectural advantage over electronics, due to the fact that the most natural way to use optics lends itself to a parallel implementation. Optical signals propagate in parallel, cross each other without interference (whereas charged electrons interact heavily with each other), have the lowest possible propagation delay of all signals, and can provide million channel free-space interconnections with simple lenses (whereas an unrealistic number of connections are necessary for electrons). Thus, optics is an excellent candidate for implementing parallel computing machines using the 3D space connectivity. While electronic architecture is naturally 2D, optical free space architecture is 3D, exploiting the third dimension of space. The communication bottleneck forces electronic computers to update the computation state space sequentially, whereas the use of optics will empower one to change the entire computation space in parallel.

Historical Perspective

Historically, digital optical computing was proposed as a natural extension of nonlinear optics. Just as transistors, which are nonlinear electronic devices, led to the realization of the computing logic using transistors, it was envisioned that nonlinear optical effects would give rise to optical transistors which would lead to development of optical logic gates. One of the limitations of nonlinear optical device-based computation is the high laser power requirement to demonstrate any nonlinear optical effect.

One of the effects used to demonstrate nonlinearity is to change the refractive index with the incident intensity. In other words, most materials will behave linearly with low optical energy. The refractive index expressed as $n = n_0 + n_2 I$ is nominally linear because the nonlinear co-efficient n_2 is usually very small. In the presence of high power, the $n_2 I$ is significant enough to cause an optical path change by altering the refractive index. When such a medium is placed inside a Fabry–Perot resonator, it will allow one to control the transmission of the filter by changing the incident intensity. Thus, at high power, it shows a nonlinear behavior or bistability. The term optical bistability was coined, which meant the presence of two stable states of output for the same intensity value.

Another development took place at the University of Michigan by the development of a precision optical processor; this was an optical processor which could process radar signals using optics, this is the example of an analog optical computing processor.

A third area which is in between the digital and the analog are the discrete level devices which most often take the form of matrix-based processors. Free space optical matrix vector multipliers are examples of such processors.

The latest field of optical computing research is in the algorithms and architecture area. The motivation here is to develop new algorithms which take advantage of optics. The thing that is lacking is the presence of appropriate device technology that can implement the algorithm in a given architecture. In terms of practical realization, analog optical computing is the most successful one. Adaptive optic processors have found success both in astronomy as well as vision sciences, and more recently in optical imaging through the atmosphere.

Development of free space optical computing has two main driving forces: the technology (device) base and the algorithm base. In this classification, since the field is in its infancy, both device-based classes and architectural (algorithm) classes are listed side-by-side as separate entities, ignoring the natural overlap that may be present in some instances. For example, nonlinear optical devices could be implemented into a variety of architecture that is listed below. Holographic processor is another technology-based class while some other architectures, such as, symbolic substitution may involve use of holograms, but it does not exclude it being a class by itself. The following list gives a classification of optical free space computing:

1. Digital Optical Computing:
   - (a) Pattern coded computing or cellular logic:
     - (i) Optical shadow-casting processor;
     - (ii) Symbolic substitution;
     - (iii) Theta logic based computation.
   - (b) Parallel arithmetic circuits:
   - (c) Hybrid optical computing:
   - (d) Nonlinear optical device based:
   - (e) Threshold logic.
2. Discrete optical computing:
   - (a) Optical fuzzy logic:
   - (b) Optical matrix processors:
     - (i) Vector matrix multiplier;
     - (ii) Differential equation solver;

(iii) Linear algebra processor;
(iv) DFT processor;
(v) Programmable logic array;
(vi) Associative memory processor.
(c) Optical neural network.
3. Analog optical computing:
(a) Adaptive optics processor;
(b) Partially coherent processing (white light processing):
(c) Fourier transform based processor:
(d) Nonlinear contact screen based:
(e) Holographic processors:
(f) Linear optical processors (image differentiation).

We start by discussing some of the device based classes and their underlying concepts: (a) nonlinear optics; (b) SLMs; and (c) hologram based.

Core Device Technologies Used for Free-Space Optical Computing

One of the most difficult challenges of optical computing (OC) is inventing devices that can realize aforementioned key advantages of optics. There exist, however, several key technology areas, where a breakthrough in any one of these will pave the way to real-world optical computing. We discuss here some of the major device technologies that relate to OC.

Spatial light modulator (SLM) converts information from different forms to the optical domain. As a result, SLM is used as an input device to an optical computing system. High contrast ratio, very high speed, and parallel addressing scheme are desirable properties of an SLM as an input device. For example, the SLM could be used to input a 2-dimensional signal or an image into a Fourier transform-based processor. Several applications of SLM are possible:

1. *Optical encoding*: Pattern coded logic can perform logic operations if inputs can be encoded using certain pattern;
2. *Encoding complex filter*: It could encode phase, amplitude or both; and
3. *As phase modulator*: for adaptive optics.

Two types of SLMs are liquid-crystal based and micro-electromechanical (MEM) based. The former uses liquid crystals to change the phase of transmitted or reflected light leading to modulation of phase. Although the most successful commercial application of MEMs is its use in consumer electronics, such as projection TVs or video projectors, in free-space optical computing it can be used as:

(a) an interconnection device;
(b) an SLM for matrix; or
(c) a phase modulating device in AO close loop control of optical phase aberrations or phase only filter.

The heart of the MEMs is tiny electrically controllable mirrors which can steer the incoming beam in a different direction or bend to create a relative optical path difference for the reflected beam.

Algorithms, where an SLM plays a major role, are pattern coded computing, optical matrix processors, adaptive optic processors, partially coherent processing, Fourier transform based processors, and linear optical processors.

Nonlinear Optical Element

A nonlinear optical element is an architecture or device structure which allows for utilization of the nonlinearity of the optical material that produces detectable changes of optical intensity. Figure 1 shows the basic block diagram for an optical computing basic element. To achieve an optical logic gate, one needs a scheme where light can be controlled by light, as current is controlled by current in a transistor. An optical device showing optical bistability can be used for this purpose. Optical bistability is defined as the situation where two different output intensities are possible for a given input intensity.

For example, as shown in Figure 2 below, at the intensity level of C, there are two stable states, one HIGH and one LOW. Just as the transistor has two stable states ON and OFF, the optical bistable device exhibits two states ON and OFF. The fact that such nonlinear behavior could implement logic behavior

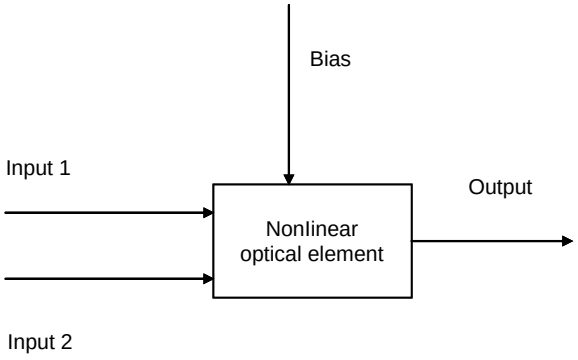


Figure 1 Block diagram of an optical logic gate using nonlinear optical element.

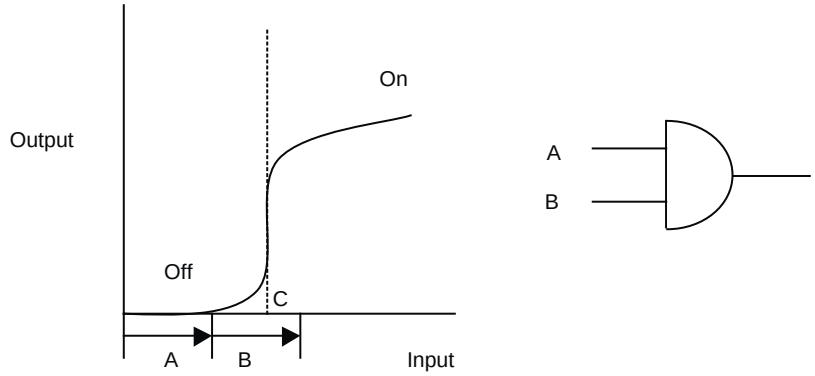


Figure 2 Bistable behavior, a typical AND behavior with inputs A and B.

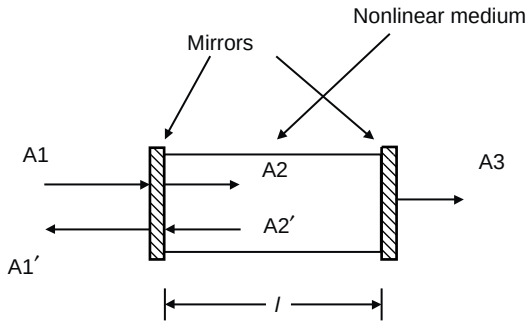


Figure 3 Bistable optical device, nonlinear medium in a Fabry–Perot cavity.

can be explained as follows. Assume that a two input device possesses a nonlinear characteristic as shown in Figure 2. Assume that the input intensity value of 1 is indicated by an intensity level equal to A. When both A and B are equal to 1, then the combined intensity switches the device to HIGH output state. However, if only one input is present, the output will be LOW. Thus, this bistable device acts as a logical AND gate.

To achieve optical bistability, two features are required, namely, nonlinearity and feedback. One way to achieve this is by placing a nonlinear medium in a Fabry–Perot cavity. Figure 3 shows a schematic of these devices. In Figure 3, A_1 is the incident field, A_1' is the reflected field, A_2 and A_2' are the forward and backward fields inside the cavity, respectively, and l is the cavity length.

The coupled wave equation for this system is given by

$$A_2' = \rho A_2 e^{2ikl-\alpha l} \qquad [1a]$$

$$A_2 = \tau A_1 + \rho A_2' \qquad [1b]$$

where, ρ is the amplitude reflectivity and τ is the amplitude transmissivity. The wave vector $k = n\omega/c$ and α is the absorption coefficient. In deriving the above equations, it was assumed that k and α are spatially invariant and real. Solving for A_2, from eqns [1a] and [1b] yields

$$A_2 = \frac{\tau A_1}{1 - \rho^2 e^{2ikl-\alpha l}} \qquad [2]$$

Equation [2] is the so-called Airy's equation which describes the properties of the Fabry–Perot cavity. From eqn [2] it can be seen that if k or α is a sufficiently nonlinear function, this device will show bistability. In general, both k and α can show nonlinearity. Hence optical bistability can be either absorptive or dispersive.

Absorptive Bistability

If the absorption coefficient α depends nonlinearly on the intensity, assuming k is constant, after some manipulation, eqn [2] can be written as

$$I_2 = \frac{TI_1}{[1 - R(1 - \alpha l)]^2} \qquad [3]$$

where, $I_i \propto |A_i|^2$, $T = |\tau|^2$ and $R = |\rho|^2$. In deriving eqn [3], it was assumed that the mirror separation is such that the system is in resonance and $\alpha l \ll 1$. If α depends on the intensity, it can be assumed that it follows the relationship of two level saturable absorbers, for simplicity, which is

$$\alpha = \frac{\alpha_0}{1 + I/I_s} \qquad [4]$$

where, α_0 denotes the unsaturated absorption coefficient, I is the local value of intensity, and I_s is the saturation intensity. Using eqn [4] in eqn [3] yields

$$I_1 = TI_2\left(1 + \frac{C_0}{1 + 2I_2/I_s}\right)^2 \qquad [5]$$

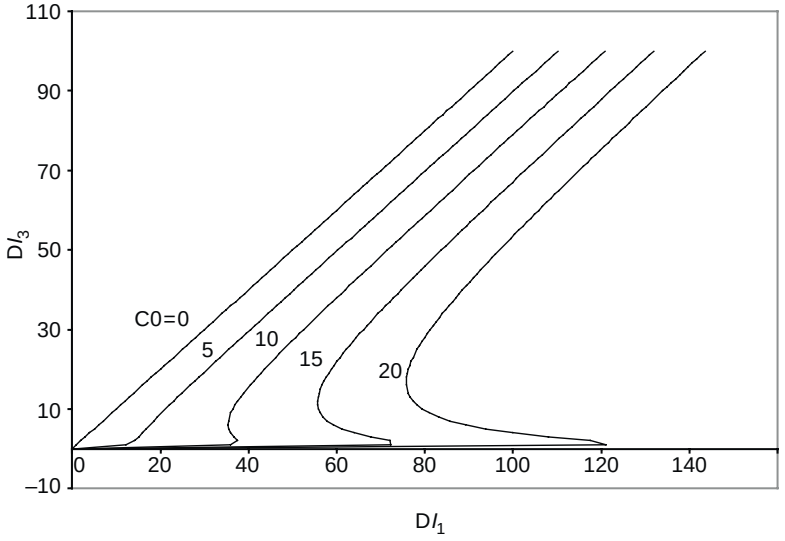


Figure 4 The input–output relationship of absorptive bistability predicted by eqns [5] and [6].

where, $C_0 = R\alpha_0 l/(1-R)$. Also the output intensity I_3 is related to I_2 by

$$I_3 = TI_2 \qquad [6]$$

Using eqns [5] and [6], it can be shown that there can be multiple solutions to the above input–output relationship which predicts the optical bistability. Figure 4 shows a plot of DI_1 versus DI_3, where $D = 2/TI_s$, as a function of various C_0. As can be seen from Figure 4, for a high enough value of C_0, the system possesses multiple solutions.

Dispersive Bistability

The intensity dependence of the refractive index can be written as

$$n = n_0 + n_2 I \qquad [7]$$

If the refractive index n varies nonlinearly with intensity, assuming $\alpha = 0$ and using eqn [7], eqn [2] can be written as

$$A_2 = \frac{\tau A_1}{1-\rho^2 e^{2ikl}} = \frac{\tau A_1}{1-Re^{i\delta}} \qquad [8]$$

where, $\delta = \delta_0 + \delta_2$. δ_0 and δ_2 are given by

$$\delta_0 = \varphi + 2n_0\frac{\omega c}{l}$$
$$\delta_2 = 2n_2 I\frac{\omega c}{l} \qquad [9]$$

where, φ is the phase associated with ρ and the fact that $k = n\omega/c$ is used.

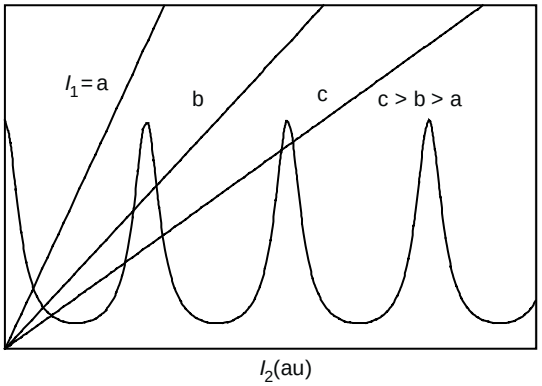


Figure 5 Plots of both sides of eqn [10] as a function of I_2 and for increasing values of input intensity I_1. The oscillatory curve represents the right-hand side of eqn [10] and the straight lines represent the left-hand side with increasing values.

After some manipulation, eqn [8] can be rewritten in terms of intensity as

$$\frac{I_2}{I_1} = \frac{1/T}{1+(4R/T^2)\sin^2(\delta/2)} \qquad [10]$$

and

$$\delta = \delta_0 + \left(4n_2\frac{\omega}{c}l\right)I_2 \qquad [11]$$

Equation [10] can be solved graphically to show the introduction of bistable behavior as a function of input intensity. Figure 5 shows the plots of both sides of eqn [10] as a function of I_2 and for increasing values of input intensity I_1. The oscillatory curve represents the right-hand side of eqn [10] and the

straight lines represent the left-hand side with increasing values of I_1. It can be readily seen from Figure 5 that at sufficiently high values of I_1, multiple solutions are possible for eqn [10]. This again will give rise to the bistable behavior.

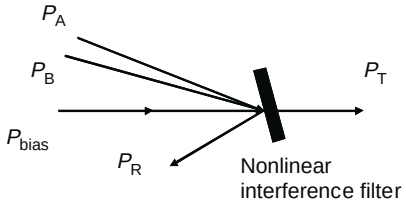
Nonlinear Interference Filters

The optical bistability from the nonlinear interference filters (or etalons) can be used to achieve optical logic gates. The transmission property of these etalons can be modified by another optical beam, known as a bias beam. A typical logic gate and the transmission property are shown in the following figure. If the bias beam is such that, $P_{\text{switch}} - (P_a + P_{\text{bias}}) < P_b$, then the transmitted beam shown in Figure 6b readily follows an AND gate. The truth table for such configurations is shown in Figure 6d.

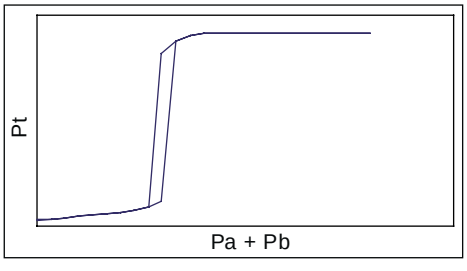
Four-Wave Mixing

Four Wave Mixing (FWM) can be used to implement various computing functions. Using the photorefractive effect and FWM in a third-order nonlinear optical material, optical phase conjugate beams can be obtained. A schematic of FWM and phase conjugation is shown in Figure 7. The forward pump beam interferes with the probe beam to produce a grating in the photorefractive material. The backward pump beam is then diffracted to produce a beam which retraces the path of the probe beam but has an opposite phase. It can be seen from Figure 7, that the optical phase conjugation can be readily used as an AND logical operator. By using the four-wave mixing, an optical counterpart of transistors has been introduced. FWM have also been used in AND-based optical symbolic substitution (OSS) operations. FWM is also used to produce matched filter-based optical correlators.

Another device showing the optical bistability is the self electro-optic effect device (SEED). SEEDs were used to implement optical logic circuits. The SEEDs are produced by putting multiple quantum wells (MQW) in the intrinsic region of a p-i-n structure. The MQWs are created by placing alternating thin layers of high (barriers) and low (wells) bandgap materials. An electric field is applied to the SEEDs externally. The absorption coefficient is a function of applied bias voltage. When a light beam is incident on these devices, it creates a photocurrent, consequently lowering the bias voltage across the MQW. This, in turn, increases the absorption and when the input optical power is high enough, the system sees a peak absorption and suddenly switches to a lower state and thus shows the bistable behavior. A schematic of SEED is shown in Figure 8.



(a)

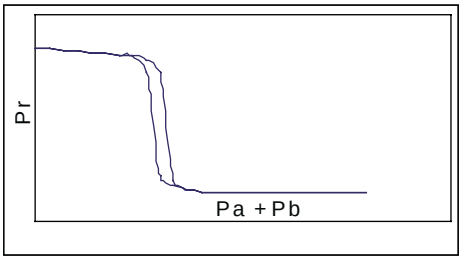


(b)

Truth table for an AND function from the above setup from the transmitted beam

P_A= input 1	P_B = input 2	P_T = output
0	0	0
0	1	0
1	0	0
1	1	1

(d)



(c)

Figure 6 A typical optical logic gate configuration using nonlinear interference filter: (a) schematic; (b) the transmitted power output; (c) reflected power output (both as functions of inputs); and (d) the truth table for AND operation with beams A and B as inputs.

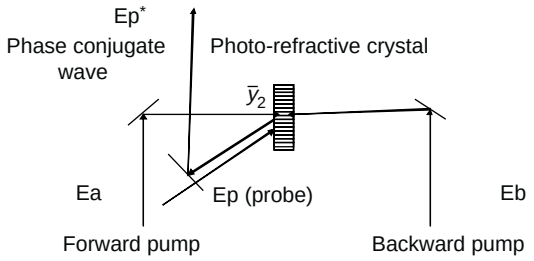


Figure 7 Four-wave mixing setup.

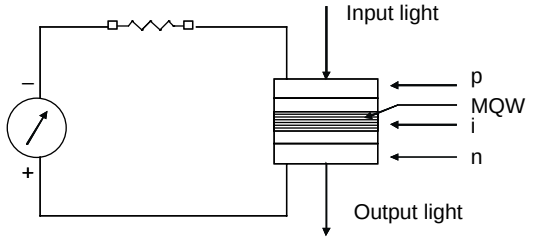


Figure 8 Schematic of self electro-optic effect devices.

Replacing the resistive load by a photodiode, a diode-based SEED (D-SEED) can be constructed, which is analogous to the way diode transistor logic was developed from resistor transistor logic in electronics. Electrically connecting two MQW p-i-n structure symmetric SEEDs (S-SEED) are constructed, which are less sensitive to optical power fluctuations and effectively provide isolation between input and output.

Pattern Coded Optical Computing: Symbolic Substitution or Cellular Logic

This is a keyword which is equivalent to simple digital logic operation. In digital logic, a truth table can describe the relationship between input and output variable. In symbolic substitution logic, this relationship is expressed as a 2D pattern rule. In any specific rule (equivalent to a row in the truth table) of SS logic, a specific 2D pattern is identified and replaced by another pattern. For example, in the case of binary logic, we can express the addition of two binary bits with the following truth table:

Rule number	*Input*		*Output*	
	A	*B*	*C*	*S*
1	0	0	0	0
2	0	1	0	1
3	1	0	0	1
4	1	1	1	0

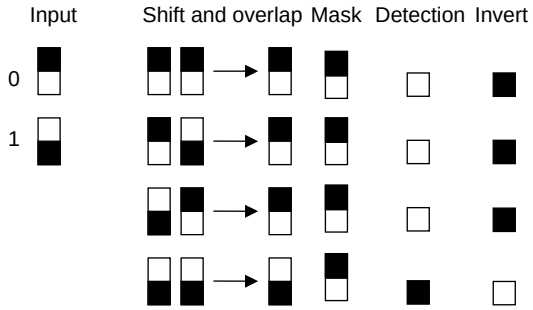


Figure 9 Symbolic substitution example, the two inputs are coded using dual rails. The second column shows the four combinations of inputs (00, 01, 10, and 11). They are shifted and overlapped creating the fourth column. A mask detects the output as a dark pixel appearing through the transparent opening. The dark is inverted to create a detection output.

Here A and B are the two digits to be added, C and S are the sum and the carry bit, respectively. To convert this operation to an SS operation, we can say that a two bit binary pattern is identified and replaced by a sum bit and a carry bit. If we now express the above in terms of symbolic substitution, we will say let 0 and 1 be represented by the following 2D symbol. Then the rule for the carry bit generation is shown in Figure 9.

Note that each of these rules has to be implemented individually. However, one can implement these by only generating the 1 output or the zero output, whichever leads to the minimum rule. Here implementing the last rule is convenient, one can make a copy of the whole pattern and then overlap it with its left shift version. This will result in a dark pixel at the origin. This is inverted and then produces a detection output. Natural extension of SS is in signed digit computing, where carry free addition can be achieved, allowing parallel implementation of numerical algorithm using free space optics.

Optical Shadow Casting (OSC)

All digital logic is nonlinear. The nonlinearity of the digital logic is converted into nonlinearity of coding. The OSC system originally proposed by Tanida and Ichioka uses two-dimensional spatially encoded patterns as data input and light emitting diodes (LEDs) as light sources. In the original shadow-casting system, the inputs were represented by vertical and horizontal stripes of opaque and transparent bars. With this system, programmability was attained by changing the LED pattern and it was possible to realize 16 logical operations between two

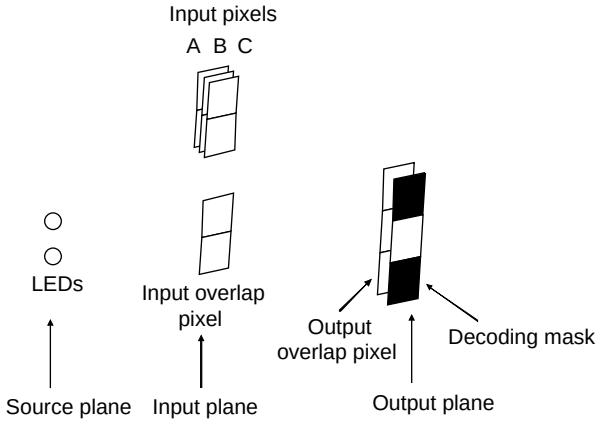


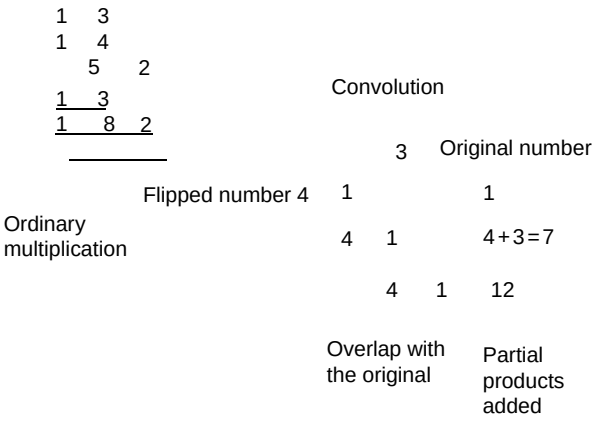
Figure 10 An optical shadow-casting system.

variables as well as a half addition as an example of arithmetic operation. Later, polarized codes were introduced into shadow-casting systems. This improved technique, called the polarization-encoded optical shadow casting (POSC), enabled complex combinational logic units like binary and trinary full adders, as well as image processing applications, to be realized.

The lensless OSC system, as shown in Figure 10, consists of spatially encoded 2D binary pixel patterns as the inputs. The input patterns are kept in close contact at the input plane and the resulting input overlapped pixel pattern is illuminated by a set of LEDs from the source plane. The light originating from each of the LEDs located at the source plane produces a shadow of the input overlap pixel pattern at the output plane. The overlap of these relatively displaced shadows results in an output overlap pixel pattern at the output plane. A decoding mask placed at the output plane is then used to spatially filter and detect the logical output. To prevent crosstalk, specific spacing among the different processing elements must be maintained.

Discrete Processors: Optical Matrix Processor

Most of the discrete optical processing schemes revolve around optical matrix–vector multiplication. Many numerically useful operations such as: (a) linear algebraic operations; (b) signal processing algorithms; (c) Boolean logic; (d) digital multiplication by analog convolution; and (e) neural networks, can be formulated as matrix–vector multiplication. An example of the digital multiplication by analog convolution (DMAC) can be understood by the following numerical example:



The two numbers are convolved as shown above. This is done by reversing one of the numbers (4 1), while keeping the other number in the same order (1 3), then sliding the reversed number to the right, and collecting the element by element products. The second partial products are 4×1 and 1×3 which are then added to give the convolution results. Surprisingly, the weighted number in mixed radix (since numbers that have digits higher than 9, are no longer decimal) has the same value as the product shown in the upper left. It is possible to have fast parallel convolution optically, resulting in fast parallel multiplier. However, an efficient electronic post processor is necessary to convert the mixed radix numbers to the radix of the to-be-multiplied numbers.

A basic optical matrix–vector multiplier is shown in Figure 11. Here the vector, represented by an expanding column of light, is projected onto the 2D matrix, possibly represented by a glass plate or electronically addressable spatial light modulator. As the light representing the vector passes through the transmitting medium, each element of the vector (**I**) gets multiplied by a column of matrix represented by the transmittance of the matrix (T). The product ($\mathbf{I}T$)

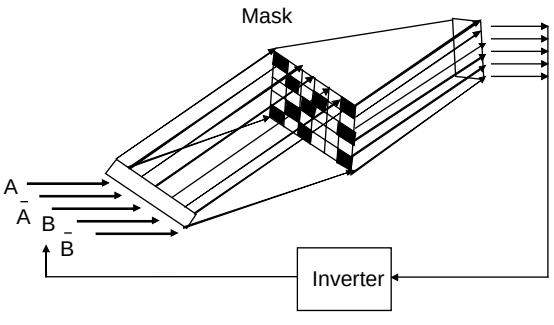


Figure 11 A matrix–vector multiplier.

is integrated horizontally using a cylindrical lens. This leads to an addition operation of the partial products, resulting in the generation of the output product vector. Mathematically:

$$y = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}b_1 + a_{12}b_2 + a_{13}b_3 + a_{14}b_4 \\ a_{21}b_1 + a_{22}b_2 + a_{23}b_3 + a_{24}b_4 \\ a_{31}b_1 + a_{32}b_2 + a_{33}b_3 + a_{34}b_4 \\ a_{41}b_1 + a_{42}b_2 + a_{43}b_3 + a_{44}b_4 \end{bmatrix} \quad [12]$$

The above can be achieved if the vector b is expanded as

$$\begin{bmatrix} b_1 & b_2 & b_3 & b_4 \\ b_1 & b_2 & b_3 & b_4 \\ b_1 & b_2 & b_3 & b_4 \\ b_1 & b_2 & b_3 & b_4 \end{bmatrix}$$

and a point-by-point multiplication is performed between the expanded vector and the original matrix. Then the addition is performed by the cylindrical lenses.

Another simple application of vector matrix multiplier is in digital logic, known as programmable logic array (PLA). PLA is an electronic device that consists of a set of AND gates followed by a set of OR gates, which can be organized in AND–OR format suitable for realizing any arbitrary general purpose logic operation. AND logic is performed when light representing a logic level is passed through a transparency representing the other variable; OR is achieved by converging light to a common detector.

$$\text{Assume a function } F = AC + \bar{A}BD + A\bar{B}C \quad [13]$$

where $\bar{A}$ represents the logical inverse of A. Using DeMorgans' theorem in digital logic, this can be expressed as

$$F = \overline{(\bar{A} + \bar{C})} + \overline{(A + \bar{B} + \bar{C})} + \overline{(\bar{A} + B + \bar{C})}$$

$$= \bar{y}_1 + \bar{y}_2 + \bar{y}_2 \quad [14]$$

Now the y_is can be generated by the vector matrix product, as shown below:

$$\text{where } \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} A \\ \bar{A} \\ B \\ \bar{B} \\ C \\ \bar{C} \\ D \\ \bar{D} \end{pmatrix} \quad [15]$$

Note here that the AND has been converted to OR, which needs to be inverted and then ORed to generate the function F. A setup shown in Figure 11 can be used to accomplish this. In step 1, the y_is are generated, and then in step 2 they are inverted and summed by a second pass through the same system with a different mask. Other applications of vector–matrix multipliers are in neural networks, where the interconnection weights are represented by the matrix and the inputs are the vector.

Analog Optical Computing

Correlation is a way to detect the presence of signal in additive noise. Optics offer a fast way of performing correlations. The practical motivation of this area comes from the fact that a lens performs a Fourier transform of a two-dimensional signal at its focal plane. The most famous analog optical computing is known as optical matched filter. The theoretical basis of an optical matched filter is in the Fourier theorem of matched filtering which states that the cross correlation between two signals can be formed by multiplying the Fourier transform of the signal and the complex conjugate of the second signal and followed by a subsequent Fourier transform operation. Mathematically:

$$\text{Correlation}(f, g) = \mathrm{F}^{-1}[\mathrm{F}\{f\}^* \text{conjugate}(\mathrm{F}\{g\})] \quad [16]$$

where F represents a Fourier transform operation and F^{-1} represents an inverse transform. Since the lens can easily perform the Fourier transform, and multiplication can be performed by light passing through a medium, the only problem that has to be

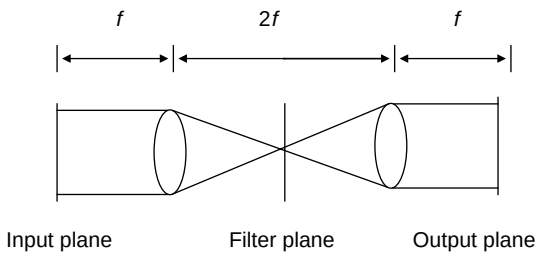


Figure 12 An optical 4-*f* set up. Lenses perform Fourier transform at their focal plane.

solved is to represent the signal and the filter in the optical domain. Since an SLM can represent both the real signal and the complex filter, thus the whole operation can be performed in a 4-*f* setup as shown in Figure 12. The input is displayed in the input plane by an SLM, the first lens performs a Fourier transform. This filter (conjugate F{*g*}) is encoded using another SLM and placed at the filter plane. When light from the F{*f*} passes through the filter, it performs the required multiplication. The second lens performs a second Fourier transform and produces the correlation output at the output plane. Variations of the basic setup exist as a joint transform correlator (JTC) or binary phase-only filter (BPOF). While the JTC achieves the product of the complex domain by adding the two signals and squaring, the BPOF is achieved by simplifying the complex domain filter by numerically approximating it with the binarized phase. Other applications of linear optical processing can be performed in the basic 4-*f* optical setup where filters such as low-pass or high-pass filters, can be set up in the correlation plane.

The hologram base computing treats holograms as a legitimate way of signal transformation from one form to another, just as a lens transforms from one domain to another. Holograms are complex representations that can act on a signal. A lens can be approximated by a hologram, in reality functions of many optical elements, such as lens, grating, prism, and other transformation such as beamsplitting, it can be combined into single computing element known as computer-generated hologram (CGH). The smartness of this type of system depends on how many operations have been replaced by a single CGH.

It is envisioned that optical computing is better suited for special purpose computing rather than general purpose computing. However, the future of optical computing lies in the marriage of appropriate device technology with the right architecture at the right applications. The development of quantum computing may hold new promises for optical computing.

List of Units and Nomenclature

DMAC	Digital multiplication by analog convolution
JTC	Joint transform correlator
OC	Optical computing
PLA	Programmable logic array
QC	Quantum computing
SLM	Spatial light modulator

See also

Nonlinear Optics, Basics: Four-Wave Mixing; Nomenclature and Units. **Quantum Optics:** Quantum Computing with Atoms.

Further Reading

Arsenault H, Szoplik T and Macukow B (eds) (1989) *Optical Processing and Computing*. San Diego, CA: Academic Press.

Boyd RW (1992) *Nonlinear Optics*. San Diego, CA: Academic Press.

Gaskill J (1978) *Linear Systems Fourier Transform and Optics*. New York: Wiley.

Gibbs HM (ed.) (1982) *Optical Bistabiligy II*. New York: Plenum Press.

Goodman JW (1996) *Introduction to Fourier Optics*, 2nd edn. New York: McGraw Hill.

Ichioka Y and Tanida J (1984) Optical parallel logic gates using a shadow-casting system for optical digital computing. *Proceedings of the IEEE* 72: 787.

Karim MA and Awwal AAS (1992) *Optical Computing: An Introduction*. New York: Wiley.

Koren I (1993) *Computer Arithmetic Algorithms*. New Jersey: Prentice Hall.

McAulay A (1992) *Optical Computing Architecture*. New York: Wiley.

Reynolds GO, DeVelis JB, Parrent GB and Thompson BJ (eds) (1989) *Physical Optics Notebook: Tutorials in Fourier Optics*. Bellingham, CA: SPIE Optical Engineering Press.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: John Wiley & Sons.

VanderLugt A (1992) *Optical Signal Processing*. New York: Wiley.

Yu FTS and Jutamulia S (1992) *Optical Signal Processing, Computing and Neural Networks*. New York: Wiley.

Incoherent Analog Optical Processors

S Jutamulia, University of Northern California, Petaluma, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

After low-cost lasers became widely available, it seemed that incoherent analog optical processing was no longer important. In general, incoherent optical processing has no apparent significant advantages, as compared with its coherent optical counterpart or digital computer processing. However, although nobody now attempts to build an incoherent optical processor, such as an incoherent correlator, most instruments are still operating using incoherent or natural light. These incoherent instruments relate to imaging, including cameras, microscopes, telescopes, projection displays, and lithographic equipment. From this point of view, it is still very important to study the principles of incoherent analog optical processing, in order to further progress or improve the incoherent image.

Incoherent Image Formation

The image formation in a coherent optical system can be explained using the linear system. A perfect point $\delta(x, y)$ in the input plane is imaged in the output plane as the impulse response of the system $h(x, y)$. When the input is only a single point, the output amplitude is simply $h(x, y)$ and the output intensity is $|h(x, y)|^2$. Accordingly, the function of $|h(x, y)|^2$ is called point spread function (PSF).

If we consider that the input object is the collection of a large number of very fine points, under coherent light illumination, the output image will be the collection of the same number of blurred spots. The shape of each blurred spot is $h(x, y)$. If the amplitude function of input object is $f(x, y)$, the amplitude function of output image is

$$o(x, y) = \int\int_{-\infty}^{\infty} f(p, q)h(x - p, y - q)\mathrm{d}p\ \mathrm{d}q \qquad [1]$$

which is a convolution of $f(x, y)$ and $h(x, y)$. The intensity function of output image is

$$|o(x, y)|^2 = \left| \int\int_{-\infty}^{\infty} f(p, q)h(x - p, y - q)\mathrm{d}p\ \mathrm{d}q \right|^2 \qquad [2]$$

According to the convolution theorem, in the Fourier transform domain or frequency domain, we have

$$O(u, v) = F(u, v)H(u, v) \qquad [3]$$

where $O(u, v)$, $F(u, v)$, and $H(u, v)$ are the Fourier transforms of $o(x, y)$, $f(x, y)$, and $h(x, y)$, respectively. $H(u, v)$ is the transfer function of the coherent imaging system. To distinguish it from the incoherent imaging system, it is also called coherent transfer function (CTF).

The same optical imaging system is now illuminated with incoherent light instead of coherent light. The amplitude impulse response is still the same $h(x, y)$, and the point spread function is also the same $|h(x, y)|^2$. Similarly, the input object is the collection of perfect points, and the output image is the collection of blurred images of each point. However, since the illuminating light is incoherent, light from any point in the input object is not coherent with light from any other points. Thus, there is no interference with light from different points. The output image is simply the addition of intensity patterns generated by each point in the input object. The amplitude function of the output image is not computable. However, the intensity function of the output image is

$$|o(x,y)|^2 = \int\int_{-\infty}^{\infty} |f(p,q)|^2 |h(x-p,y-q)|^2 \mathrm{d}p\,\mathrm{d}q \qquad [4]$$

In the frequency domain, we now have

$$O_I(u,v) = F_I(u,v)H_I(u,v) \qquad [5]$$

where $O_I(u,v)$, $F_I(u,v)$, $H_I(u,v)$ are Fourier transforms of $|o(x,y)|^2$, $|f(x,y)|^2$, and $|h(x,y)|^2$, respectively. The intensity impulse response $|h(x,y)|^2$ is the PSF. Note that the index I of O_I, F_I, and H_I denotes intensity. The function $H_I(u,v)$ is now the transfer function of the incoherent imaging system, which is called optical transfer function (OTF).

Referring to Fourier analysis, if $H(u, v)$ is the Fourier transform of $h(x, y)$, the Fourier transform of $|h(x, y)|^2$ is the autocorrelation of $H(u, v)$ as follows:

$$H_I(u, v) = \int\int_{-\infty}^{\infty} H(p, q)H^*(p - u, q - v)\mathrm{d}p\ \mathrm{d}q \qquad [6]$$

The OTF (the function $H_I(u, v)$) is the autocorrelation of CTF (the function $H(u, v)$), which is a complex

function in general. It can be written as

$$\text{OTF} = |\text{OTF}| \exp(i\phi) \quad [7]$$

We can further write

$$\text{MTF} = |\text{OTF}| \quad [8]$$

where MTF stands for modulation transfer function, and

$$\text{PTF} = \phi \quad [9]$$

where PTF stands for phase transfer function.

Measurement of MTF

The MTF of an incoherent imaging system can be measured empirically using an input that is a cosine grating having intensity transmittance function:

$$|f(x)|^2 = 1 + \cos(2\pi a x) \quad [10]$$

where a is the frequency of grating. The output or the image intensity function will be

$$|o(x)|^2 = 1 + m\cos(2\pi a x + \varphi) \quad [11]$$

where m, the contrast, can be measured by

$$m = \frac{|o(x)|^2_{\max} - |o(x)|^2_{\min}}{|o(x)|^2_{\max} + |o(x)|^2_{\min}} \quad [12]$$

In the frequency domain, the Fourier transform of the input function is

$$F_I(u) = \delta(u) + \tfrac{1}{2}\delta(u-a) + \tfrac{1}{2}\delta(u+a) \quad [13]$$

The Fourier transform of the output function will be

$$O_I(u) = F_I(u)\text{OTF}(u) \quad [14]$$

Substitution of eqn [13] yields

$$O_I(u) = \delta(u)\text{OTF}(0) + \tfrac{1}{2}\delta(u-a)\text{OTF}(a) + \tfrac{1}{2}\delta(u+a)\text{OTF}(-a) \quad [15]$$

Notice that

$$\text{OTF}(0) = 1 \quad [16]$$

due to normalization, and

$$\text{OTF}(a) = \text{OTF}(-a) \quad [17]$$

Because OTF is an autocorrelation function, it must be a symmetric function. Substitution of eqns [16] and [17] into eqn [15] yields

$$O_I(u) = \delta(u) + \tfrac{1}{2}\delta(u-a)\text{OTF}(a) + \tfrac{1}{2}\delta(u+a)\text{OTF}(a) \quad [18]$$

Remembering that OTF is a complex function given in eqn [7], we can write

$$\text{OTF}(a) = \text{MTF}(a)\exp(i\phi(a)) \quad [19]$$

Substitution of eqn [19] into eqn [18] yields

$$O_I(u) = \delta(u) + \tfrac{1}{2}\delta(u-a)\text{MTF}(a)\exp(i\phi(a)) + \tfrac{1}{2}\delta(u+a)\text{MTF}(a)\exp(i\phi(a)) \quad [20]$$

The inverse Fourier transform of eqn [20] is

$$|o(x)|^2 = 1 + \text{MTF}(a)\cos(2\pi a x + \phi(a)) \quad [21]$$

By comparing eqn [21] and eqn [11], we find

$$\text{MTF}(a) = m \quad [22]$$

and the PTF

$$\phi(a) = \varphi \quad [23]$$

Therefore, the MTF at frequency $u = a$ of the incoherent imaging system can be measured using eqn [12]. To get the complete MTF(u), the measurement is repeated using cosine gratings with different frequencies. The PTF can also be measured at the same time.

Incoherent Spatial Filtering

In the 4f coherent optical processor, the first lens performs the Fourier transform, and the second lens performs the inverse Fourier transform, as shown in Figure 1. The Fourier transform domain or frequency domain is materialized in the frequency plane that is the back focal plane of the first lens.

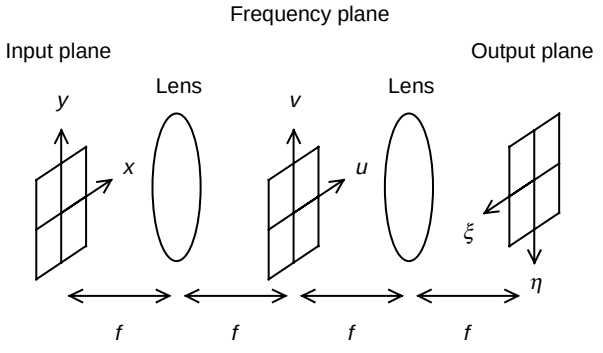


Figure 1 4f coherent optical processor consists of two Fourier transform lenses.

Thus, the Fourier transform of the object can be visualized in the frequency plane. For example, if the object is a grating, we will see its frequency components in the frequency plane. However, if the coherent illumination (e.g., from a laser) is replaced with incoherent illumination (e.g., from a lightbulb), the pattern of Fourier transform in the frequency plane disappears. Does it mean that we can no longer perform spatial filtering? No, we still can perform spatial filtering, although the pattern of Fourier transform is no longer visualized.

Consider that we have a spatial filter in the frequency plane whose physical form can be expressed by the function $H(u, v)$. For illustration, this spatial filter is simply a small one-dimensional window $H(u)$, as shown in Figure 2a. If the 4f optical processor is illuminated with coherent light, the CTF of the system is $H(u)$ itself as shown in Figure 2b. However, if the 4f optical processor is illuminated with incoherent light, we will use OTF instead of CTF in the analysis of linear system. The OTF is the autocorrelation of CTF, which is graphically depicted in Figure 2c. In this example, since the OTF is real and positive, the MTF is identical to the OTF.

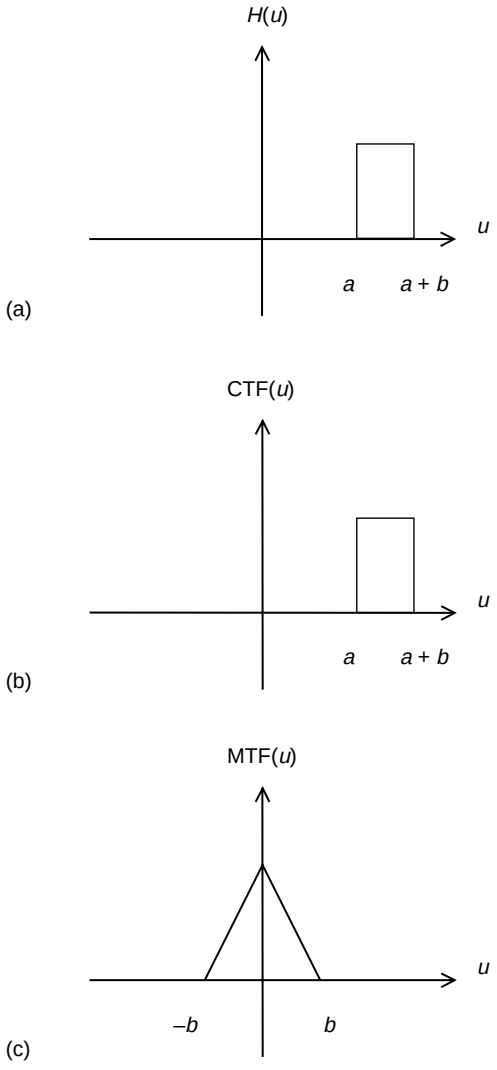


Figure 2 (a) One-dimensional window in frequency plane functions as a spatial filter. (b) CTF and (c) OTF or MTF produced by spatial filter (a).

For simplicity, we analyze and show a one-dimensional filter only in a two-dimensional graph. It certainly can be extended to the two-dimensional filter. And we need a three-dimensional graph to show the two-dimensional autocorrelation.

It is important to note that, although the CTF is a high pass filter that passes frequencies in the band of $u = a$ to $u = a + b$, the MTF is a low pass filter that passes frequencies from $u = -b$ to $u = b$, independent of a. This indicates that an incoherent processor cannot enhance high frequency. It is always a low pass processor. It is interesting to note that the OTF is independent of the location of the filter $u = a$, but it is determined by the width of the filter $\Delta u = b$ only.

Accordingly, an aperture made up of randomly distributed pinholes behaves as a low pass filter in an incoherent imaging system. The cut-off frequency of such a filter can be derived from the diameter of the pinhole, which is equivalent to Figure 2b. The procedure for using this filter is simple. If a camera is used to take a picture, the random pinhole filter can be simply attached to the camera lens. The picture taken will not consist of high-frequency components, because the filter acts like a low pass filter. The low pass filtering can also be done by reducing the aperture of camera. However, by doing so, the light entering the camera is also reduced. The use of a random pinhole filter removes high frequencies without reducing the light intensity. When the random pinhole filter is placed in the frequency plane of a 4f coherent optical processor, it can be used to reduce the speckle noise. Note that the speckle noise is the main drawback in coherent optical processing.

Incoherent Complex Matched Spatial Filtering

Lohmann and Werlich pointed out that the Vander Lugt correlator can also be operated with incoherent light. In the coherent Vander Lugt correlator the correlation term is:

$$o_3(\xi, \eta) = \int\int_{-\infty}^{\infty} f(\alpha, \beta) g^*(\alpha - \xi, \beta - \eta) \mathrm{d}\alpha \mathrm{d}\beta \qquad [24]$$

Its intensity distribution is

$$|o_3(\xi,\eta)|^2 = \left| \int\int_{-\infty}^{\infty} f(\alpha,\beta) g*(\alpha-\xi,\beta-\eta) \mathrm{d}\alpha \mathrm{d}\beta \right|^2 \quad [25]$$

However, when the input object is illuminated with incoherent light instead of coherent light, the intensity distribution of the correlation term becomes

$$|o_3(\xi,\eta)|^2 = \int\int_{-\infty}^{\infty} |f(\alpha,\beta)|^2 |g(\alpha-\xi,\beta-\eta)|^2 \mathrm{d}\alpha \mathrm{d}\beta \quad [26]$$

Equation [26] shows the correlation of $|f(x,y)|^2$ and $|g(x,y)|^2$. Therefore, when $f(x,y)$ is identical to $g(x,y)$, the correlation peak always appears regardless of whether the illumination is coherent or incoherent. However, the detailed structures of the coherent and incoherent correlation outputs are different. Note that the joint transform correlator cannot be operated using incoherent light, since no joint Fourier spectra will be formed.

Depixelization of Projection Display

Recently, a technique to remove the pixel structure from an image projected by liquid crystal projection display has been reported. A liquid crystal projection display must have a pixel structure to form an image. Can we remove this pixel structure from the projected image to obtain a movie-like image? It is well known that a single hole in the frequency plane can perform a coherent spatial filtering, such that the pixel structure will disappear from the output image.

Because of the pixel structure, we get multiple Fourier spectra in the frequency plane. By passing only one spectrum order through the hole in the frequency plane, the pixel structure can be removed as shown in Figure 3.

It is interesting to note that selecting any one of the spectra will produce the image at the same location.

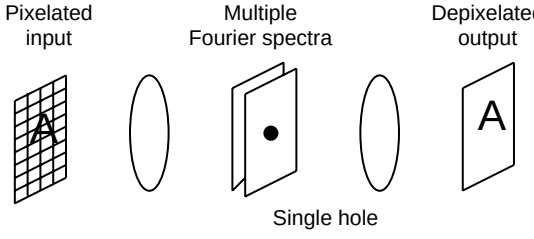


Figure 3 Removal of pixel structure using a single hole in frequency plane. Most of the energy is wasted and the output intensity is dim.

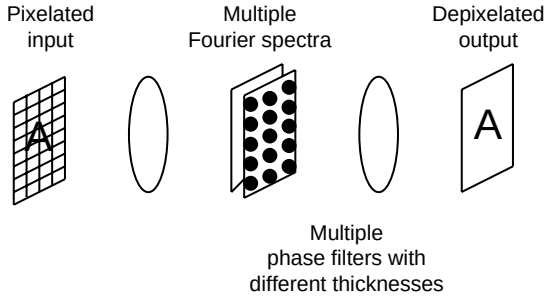


Figure 4 Removal of pixel structure using multiple phase filters with different thickness to cover each Fourier spectrum order. No energy is wasted and the output intensity is very bright.

However, if two spectra are passed, on the top of the produced image, Young interference fringes are also produced. If all spectra are passed, the produced interference fringes are in fact the pixel structure itself. By passing only one spectrum order and blocking most spectrum orders, we lose most of the energy. The projected image will be very dim, although the pixel structure is removed.

To overcome this problem, we may cover each spectrum order with a transparent material having different thickness as shown in Figure 4. Thus, every spectrum order is passed and delayed by a phase filter with a different thickness. If the delay produced by the phase filter is larger than the coherent length of light, every spectrum order is no longer coherent to each other. In other words, the resultant image is the sum of the intensities of all images. As a result, the pixel structure does not appear, and no intensity is lost. For a white light source, since the frequency bandwidth is large, the coherent length is typically on the order of several tens of μm.

This technique can significantly improve the quality of the liquid crystal projection display. Figure 5 shows the experimental result of a depixelated projection image. Figure 5a shows the projected image of an input with pixel structure when no phase filter is applied in the frequency plane. Figure 5b shows the projected image of the same input with pixel structure when phase filters are applied in the frequency plane. The pixel structure is successfully removed as shown in Figure 5b.

Computed Tomography

The computed tomography or CT using X-ray is usually considered beyond optical information processing. However, it is interesting to review the principle of the X-ray CT, since the X-ray source is incoherent, and the CT image reconstruction utilizes Fourier transformation that is commonly used in

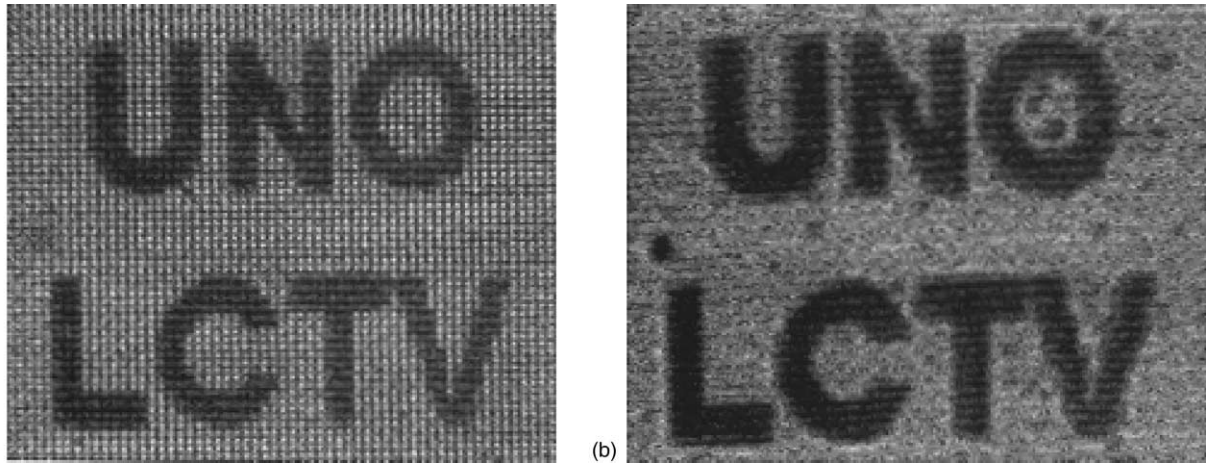


Figure 5 (a) Projected image consists of pixel structure when no phase filter is applied in the frequency plane. (b) Depixelated image is produced when phase filters are applied in the frequency plane.

optical information processing. The X-ray CT is used for taking the cross-section picture of the human body. Note that a conventional X-ray picture only provides a projection instead of a cross-section picture.

Consider the case when an X-ray penetrates through an object characterized by its attenuation coefficient or density function $f(x, y)$ as shown in Figure 6. The detected X-ray intensity is

$$I = I_0 \exp - \left[\int f(x, y) \mathrm{d}s \right] \qquad [27]$$

where I_0 is the original intensity. Equation [27] can be written as

$$\int f(x, y) \mathrm{d}s = -\ln \frac{I}{I_0} \qquad [28]$$

Apparently, we cannot directly measure $f(x, y)$, although we can obtain $\int f(x, y) \mathrm{d}s$ from the measurement of I and I_0. However, we want to get the cross-section picture which is the density function $f(x, y)$. The basic idea is to compute $f(x, y)$ from its Fourier transform $F(u, v)$, and $F(u, v)$ is derived from eqn [28]. $F(u, v)$ can be expressed as

$$F(u, v) = \int \int_{-\infty}^{\infty} f(x, y) \exp[-\mathrm{i}2\pi(ux + vy)] \mathrm{d}x \, \mathrm{d}y \qquad [29]$$

For $v = 0$, eqn (29) becomes

$$F(u, 0) = \int \int_{-\infty}^{\infty} f(x, y) \exp[-\mathrm{i}2\pi ux] \mathrm{d}x \, \mathrm{d}y \qquad [30]$$

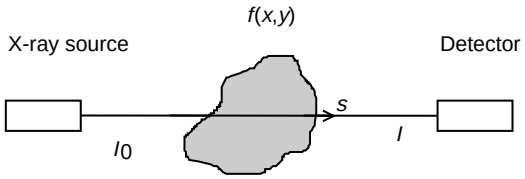


Figure 6 X-ray passes through object having density function $f(x, y)$.

which can be written

$$F(u, 0) = \int_{-\infty}^{\infty} p(x) \exp[-\mathrm{i}2\pi ux] \mathrm{d}x \qquad [31]$$

where

$$p(x) = \int_{-\infty}^{\infty} f(x, y) \mathrm{d}y \qquad [32]$$

The projection function $p(x)$ given by eqn [32] can be obtained using an X-ray parallel beam as shown in Figure 7. Then $F(u, 0)$ can be computed using computer based on eqn [31]. $F(u, 0)$ is the Fourier transform along the u-axis. To get other data, we rotate the coordinate (u, v) to (u', v') by a small angle α. Correspondingly, the coordinate (x, y) is also rotated to (x', y') by the same angle α while the object is not rotated as shown in Figure 8.

From the measurement shown in Figure 8, we can get $p(x')$, which, in turn, provides $F(u', v' = 0)$. By increasing the rotation angle α, we will get other data of $F(u'', v'' = 0)$. The completion of rotation of 180 degrees will give us the data in the frequency domain (u, v) as shown in Figure 9.

After the data $F(u, v)$ is collected, as shown in Figure 9, $F(u, v)$ can be inverse Fourier transformed to

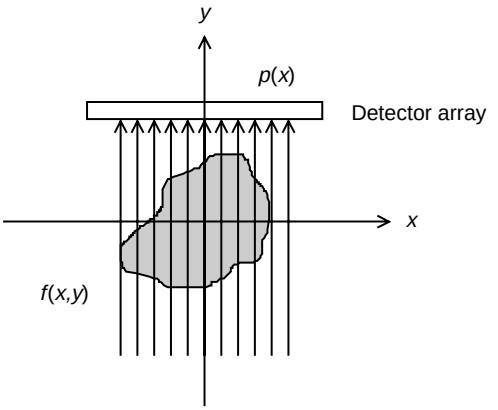


Figure 7 X-ray parallel beam passes through the object to produce projection function $p(x)$.

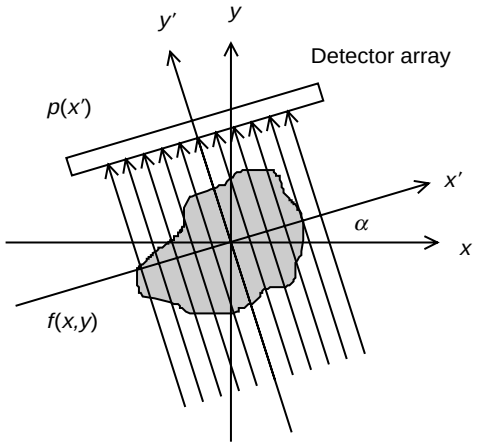


Figure 8 Rotated X-ray parallel beam produces the rotated projection function $p(x')$.

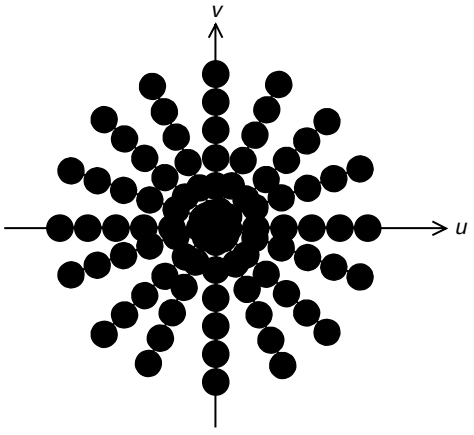


Figure 9 $F(u, v)$ computed from a set of rotated projection functions.

produce $f(x, y)$. There are two approaches in the inverse Fourier transform. The first approach is to interpolate the computed value $F(u, v)$ such that the value of $F(u, v)$, on a regular grid structure, can be defined before an inverse Fourier transform can be taken.

The second approach is to apply polar coordinates, instead of Cartesian coordinates, to the inverse Fourier transform, such that no interpolation in frequency domain (u, v) is required. The inverse Fourier transform in Cartesian coordinates is as follows:

$$f(x, y) = \int\int_{-\infty}^{\infty} F(u, v) \exp[i2\pi(ux + vy)]du\, dv \quad [33]$$

It can be written in polar coordinates as:

$$f(x,y) = \int_0^{2\pi}\int_0^{\infty} F(\rho\cos\phi, \rho\sin\phi) \times \exp[i2\pi\rho(x\cos\phi + y\sin\phi)]|\rho| d\rho\, d\phi \quad [34]$$

where

$$u = \rho\cos\phi \quad [35]$$

$$v = \rho\sin\phi \quad [36]$$

Therefore $f(x,y)$ can be obtained from the collected $F(u,v)$, as shown in **Figure 9**, without further interpolation using eqn [34]. Note that eqns [31], [33], and [34] are computed using a computer program.

See also

Spectroscopy: Fourier Transform Spectroscopy.

Further Reading

Bracewell RN (1965) *The Fourier Transform and its Applications*. New York: McGraw-Hill.

Hounsfield GN (1973) Computerized transverse axial scanning (tomography): Part I. *British Journal of Radiology* 46: 1016–1022.

Iizuka K (1985) *Engineering Optics*. New York: Springer-Verlag.

Jutamulia S, Toyoda S and Ichihashi Y (1995) Removal of pixel structure in liquid crystal projection display. *Proceedings of SPIE* 2407: 168–176.

Klein MV (1970) *Optics*. New York: Wiley.

Lohmann AW and Werlich HW (1971) Incoherent matched filtering with Fourier holograms. *Applied Optics* 10: 670–672.

Reynolds GO, DeVelis JB, Parrent GB Jr and Thompson BJ (1989) *The New Physical Optics Notebook: Tutorials in Fourier Optics*. Bellingham, WA: SPIE.

Vander Lugt A (1964) Signal detection by complex spatial filtering. *IEEE Trans. Inf. Theo.* IT-10: 139–145.

Yang X, Jutamulia S and Li N (1996) Liquid-crystal projection image depixelization by spatial phase scrambling. *Applied Optics* 35: 4577–4580.

Yu FTS and Wang EY (1973) Speckle reduction in holography by means of random spatial sampling. *Applied Optics* 12: 1656–1659.

Optical Bit-Serial Computing

A D McAulay, Lehigh University, Bethlehem, PA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Why Optical Bit-Serial Computing?

Optical bit-serial computing will lower cost and improve dataflow rates in optical telecommunications and associated computing. Moreover, the electromagnetic immunity of optics will enhance security for computing as well as for communications. Also, absence of sparking with optics enables safe operation, even in hazardous environments.

Currently, optical bit-serial communication involves passing terabits of data per second (10^{12} bps) – thousands of encyclopedias per second – along an optical fiber by switching light of multiple colors on and off, representing '1' and '0' bits respectively. Replacing electronic computing functions, constrained to tens of gigabits per second (10^9 bps), with integrated optics, ones of much higher speed, enables faster flow rates – allowing continuation of the annual doubling of capacity-distance that has occurred since the 1970s. Cost is reduced by eliminating the demultiplexing and multiplexing for optic-electronic-optic (OEO) conversions. Two illustrative areas of optical bit-serial computing for replacing electronic computing functions are described: computing at up to 40 Gbps with semiconductor optical amplifiers (SOAs) and computing at higher bit rates with integrated optic components. Several books address the relevant technical background: free-space optical information processing, optical computer architectures, optical fiber communications, and integrated optics.

Computing with Semiconductor Optical Amplifiers

Semiconductor optical amplifiers, SOAs, allow nonlinear operations (switching, logic) at low power (mW) because of amplification – typically up to 1000. An SOA is a laser diode with antireflection coatings in place of reflecting facets. Light pulses at bit rates in excess of 40 Gbps are amplified by absorbing power from an electronic pump. SOAs are used in the following, at bit rates up to 40 Gbps, for nonlinear operations in cross-gain, cross-phase, or cross-polarization.

SOA-XGM for Logic

Figure 1 shows an SOA in cross-gain modulation (XGM) performing a NOR function. Either A or B inputs at a '1' level, driving the SOA through couplers into saturation, lowering its gain. A continuous wave at C, having a different frequency λ_3, from that of A, λ_1 and B, λ_2, experiences the decrease in gain, resulting in an output '0' level at frequency λ_3. The filter in Figure 1 can be avoided by counter-propagating A and B through the SOA from right to left. XOR gates use interference between two incoming streams of bits (locked to identical frequency, phase, polarization) to correlate headers on internet protocol (IP) address headers – similar to matching zip codes in postal mail routing machines. Bit-serial-adders are constructed from combinations of XOR and NOR gates; subsequently the adders are used to construct ripple-carry-adders for word addition. In addition, NOR gates are used to construct flip-flops; hence registers which form short-term memory.

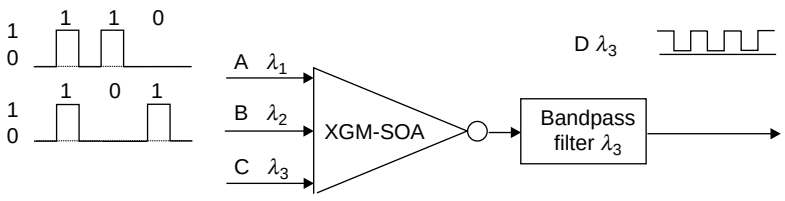


Figure 1 Optical XGM-SOA NOR gate.

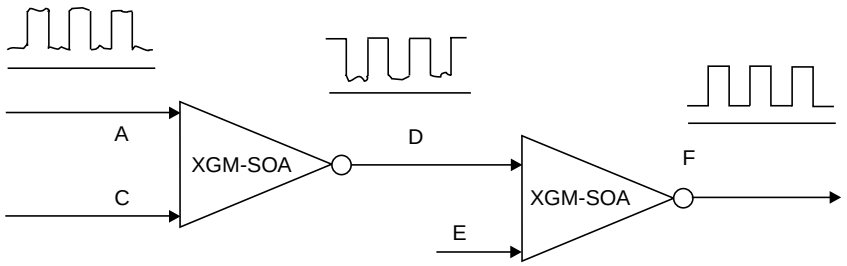


Figure 2 Optical XGM-SOA 2R signal regeneration.

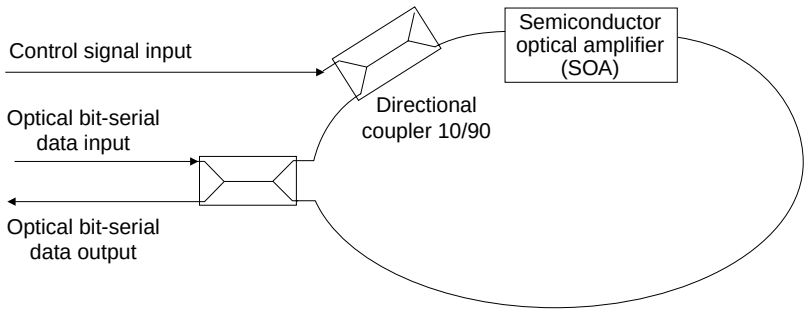


Figure 3 XPM-SOA in Sagnac interferometer for optical manipulation of bits at high data rates.

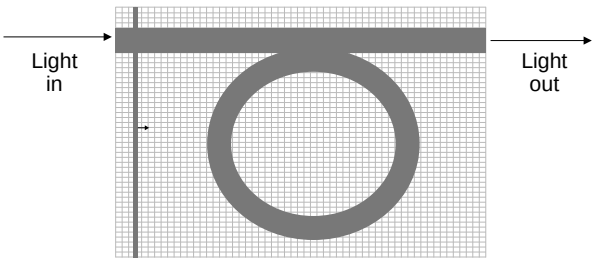


Figure 4 Integrated optic Optiwave layout for microring resonator.

SOA-XGM for Frequency Conversion

Ignoring input B, the bit information in Figure 1 has been transferred from carrier frequency A onto the carrier frequency C, with inversion. This enables switching of signals in wavelength division multiplexing (WDM), for example, color reuse is simplified and a color red is routed to the location where blue was previously routed; this is equivalent to time slot interchangers for switching in time division multiplexing (TDM).

SOA-XGM for 2-R Signal Regeneration

Figure 2 show how two SOAs may be used in sequence to restore signal levels distorted by propagation or computation; reamplification and reshaping (2-R). The '1' level at the output of the first SOA is clipped by gain saturation, removing noise, this becoming the '0' level at the output of the second SOA. The second SOA clips the '1' level at the output.

SOA-XPM with Interferometer

In cross-phase modulation (XPM), the signal at A, Figure 1, changes the phase of carrier for signal C passing through the SOA (signal levels may be less than for XGM). Inclusion of an SOA in an interferometer, Mach–Zehnder or Sagnac, converts phase to intensity for frequency conversion, switching or high-speed bit manipulation. Figure 3 shows a control pulse entering a coupler at the left and providing a control input to the SOA at one frequency. The bits circulating round the loop at a different frequency can be switched on and off by the control pulse, to synchronize or manipulate high bit rate signals or for frequency conversion.

Computing with Integrated Optics

Combining several discrete optical components into a single optical chip is more beneficial than in electronic very large-scale integration (VLSI) because optical connections require greater precision; moreover bit rates are higher because photons, unlike electrons, have no mass or charge (only spin).

Integrated Optic Microring Resonator

Figure 4 shows a microring resonator in an integrated optic chip for optical filtering, light entering at the left couples into the loop. For loop lengths that are a multiple of wavelength, light cycling the loop is in phase and resonance occurs. At resonant frequencies,

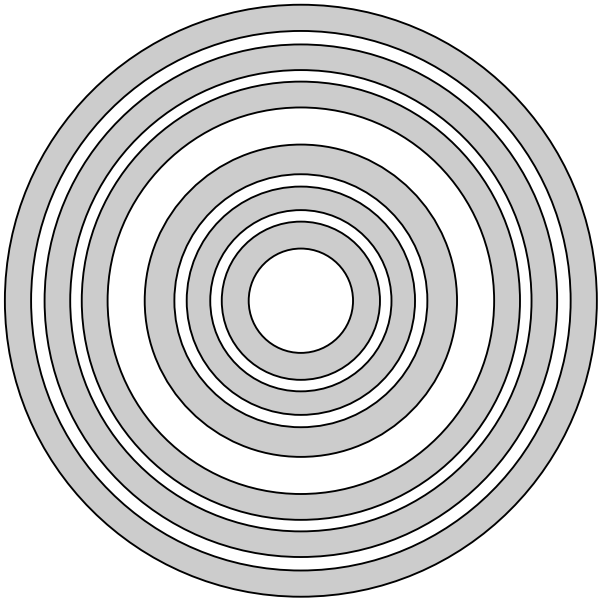


Figure 5 Integrated optic Optiwave layout for Bragg microring resonator.

light exiting the loop is out of phase with that in the linear guide, causing cancellation of the input. As frequency varies away from resonance, light passes and the device acts as a notch filter to block light at the resonant frequencies. With nonlinear Kerr material, low-level light intensity may be used to change a resonant frequency to make a switch.

Bragg Integrated Optic Microring Resonator

Performance is improved by using a Bragg microring resonator, Figure 5, which confines light in the central light colored ring by dielectric mirror reflection from Bragg gratings on either side. The refractive index is less in the ring than in the surrounding regions, in contrast to a conventional waveguide (used in Figure 4) that guides by total internal reflection. The higher reflectivity, smaller ring radius and lower refractive index – these help to reduce power loss, volume, and travel time around the ring. This increases filter selectivity, drives up peak power for enhanced nonlinearity for switching, decreases stored energy for fast response, and increases frequency range (free spectral range, *FSR*). In this case, an equivalent photonic crystal may be more expensive because it requires higher resolution lithography and greater refractive index contrast.

Photonic Crystal Mach–Zehnder Switch

Recently developed photonic crystal technology allows a reduction in integrated optic chip size from centimeters to submillimeters. The 2D or 3D periodic

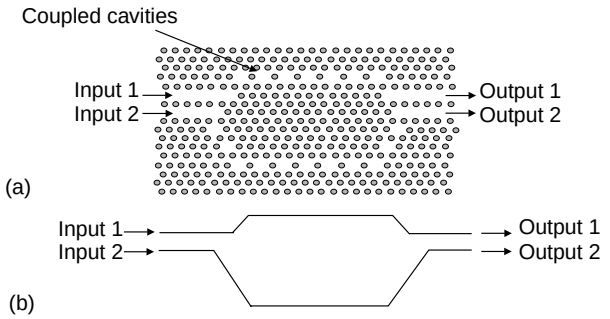


Figure 6 Proposed photonic crystal Mach–Zehnder interferometer.

structure of dielectric variations emulates crystals and provides dielectric mirror reflections in multiple dimensions for lossless waveguide bends, high *Q* resonators, and slowing of wave propagation to enhance effective nonlinearity of materials. Figure 6 shows a proposed 2D photonic crystal Mach–Zehnder interferometer switch in which light is split into two paths by a coupler and propagation velocity is slowed for enhancing nonlinearity by using coupled-resonator optical waveguides (CROWs), in which alternate posts of different dielectric constant remain in the waveguides to form coupled cavities.

Conclusion

Optical bit-serial computing can reduce cost and increase flow-rate for expanding telecommunication network capacity, while providing enhanced security from electromagnetic sabotage and tampering. Advances in integrated optics (microring resonators), progress in optoelectronic components (SOAs), improved materials and fabrication techniques – these technologies enable cost-effective optical bit-serial computing solutions for future telecommunication evolution.

See also

All-Optical Signal Regeneration. **Information Processing:** All-Optical Multiplexing/Demultiplexing.

Further Reading

Agrawal G (2002) *Fiber-Optic Communication Systems*. New York: Wiley.

Connelly MJ (2002) *Semiconductor Optical Amplifiers*. Boston, MA: Kluwer Academic Publishers.

Goodman JW (1996) *Introduction to Fourier Optics*, 2nd edn. New York: McGraw-Hill.

Heuring VP, Jordan HF and Pratt JP (1992) A bit serial architecture for optical computing. *Applied Optics* 31(17): 3213–3224.

Hunsperger RG (2002) *Integrated Optics*, 5th edn. New York: Springer.

Joannopoulios JD, Meade RD and Winn JN (1995) *Photonic Crystals*. Princeton, NJ: Princeton University Press.

McAulay AD (1991) *Optical Computer Architectures*. New York: Wiley.

McAulay AD (2004) Novel all-optical flip-flop using semiconductor optical amplifiers in innovating frequency-shifting inverse-threshold pairs. *Optical Engineering* 43(5): 1115–1120.

Scheuer J and Yariv A (2003) Two-dimensional optical ring resonators based on radial Bragg resonance. *Optics Letters* 28(17): 1528–1530.

Xu L, Glesk I, Baby V and Prucnal PR (2004) All-optical wavelength conversion using SOA at nearly symmetric position in a fiber-based Sagnac interferometric loop. *IEEE Photonics Technical Letters* 16(2): 539–541.

Yariv A (1997) *Optical Electronics in Modern Communications*, 5th edn. New York: Oxford University Press.

Optical Digital Image Processing

B L Shoop, United States Military Academy, West Point, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Optical approaches to image processing provide a natural method by which to obtain the inherent advantages associated with optics including massive parallelism, high-speed processing, and the inherent compatibility with image formats. Early applications of optics employed analog processing techniques with the most common being the optical Fourier transform, matrix-vector processors, and correlators. Later, advances in digital signal processing (DSP), digital storage, and digital communication systems demonstrated the potential for higher resolution, improved flexibility and functionality, and increased noise immunity over analog techniques and quickly became the preferred method for accurate signal processing. Consequently, optical processing architectures, that incorporated digital processing techniques, were explored. Advances in very large-scale integration (VLSI) electronic circuitry and optoelectronic devices later enabled smart pixel technology, which integrates the programmability and processing of electronic circuitry with the two-dimensional (2D) nature of an optical architecture and made possible a particular realization of optical digital image processing. This chapter describes an application of optical digital image processing, based on smart pixel technology, called digital image halftoning.

Natural images are, by definition, continuous in intensity and color. Halftoning is the process by which a continuous-tone, gray-scale image is converted to one containing only binary output pixels. Halftoning can be thought of as an image compression technique whereby a high-resolution image is transformed to a low-resolution image containing only black-and-white pixels. The transformation from a continuous-tone, gray-scale image to one containing only binary-valued pixels is also similar, in principle, to low-resolution analog-to-digital (A/D) conversion.

An early manual approach to printing continuous-tone images, using only the presence or absence of ink, is the Mezzotint. Ludwig von Siegen (1609–1680), an amateur Dutch printmaker, first invented the mezzotint, or 'half tone process', in 1640. The process later came into prominence in England during the early eighteenth century. Mezzotints are produced on copper plates where the surface of the plate is roughed with a tool called the Mezzotint Rocker, shaped like a wide chisel with a curved and serrated edge. By rocking the toothed edge backwards and forwards over the plate, a rough burr is produced which holds the ink. The dark regions of an image were roughed in a random fashion, while the areas to be lightened were scraped and burnished. This process was found to be especially useful for the reproduction of paintings, due to its ability to capture the subtlest gradations of tone from rich, velvety blacks to glowing highlights. The mezzotint is, therefore, an early artistic predecessor to the modern day halftone.

Optical halftoning has been in use by commercial printers for over 100 years. Commercial halftone screens are based on a discovery made by William Henry Fox Talbot (1800–1877), in 1852. He demonstrated the feasibility of optical halftoning by photographing an image through a loosely woven fabric or 'screen'. In the 1890s, this process came into practical use when the halftone screen, consisting of two ruled glass plates cemented together, became commercially available. Commercial halftone screens produce an effect of variably sized dots on the photographic plate that gives the illusion of a continuous tone image.

Digital image halftoning, sometimes referred to as spatial dithering, is the process of converting an

electronic continuous-tone, gray-scale image to one containing only binary-valued pixels, for the purpose of displaying, printing, or storing the image. The underlying concept is to provide the viewer of the image with the illusion of viewing a continuous-tone image when, in fact, only black and white pixel values are used in the rendering. This process is particularly important in applications such as laser printing, bilevel displays, xerography, and more recently, facsimile.

There are a number of different methods by which this digital image halftoning can be accomplished, which are generally classified as either point or neighborhood processes. A point process is one which computes the output pixel value based strictly on some characteristic of the corresponding input pixel. In contrast, a neighborhood process computes a single output pixel based on a number of pixels in a neighborhood or region of the input image. Ordered dither, which is considered a point process, produces an output by comparing a single continuous-tone input value against a deterministic periodic array of threshold values. If the value of the input pixel under consideration is greater than the corresponding threshold value, the corresponding output pixel is rendered white. If the intensity is less that the threshold, it is rendered black. Dispersed-dot ordered dither is a subset of ordered dither in which the halftone dots are of a fixed size while clustered-dot ordered dither uses variable-sized dots and simulates the variable-sized dots of a printer's halftone screen in the rendering. Among the advantages of point process halftoning, in general and ordered dither halftoning specifically, are simplicity and speed of implementation. The primary disadvantage is that ordered dither produces patterns in the halftoned image, which are visually undesirable.

In contrast, halftoning using the error diffusion algorithm, first introduced by Floyd and Steinberg in 1975, employs neighborhood operations and is currently the most popular neighborhood process. In this approach to halftoning, the output pixel value is determined, not solely by the value of the input pixel and some deterministic pattern, but instead by a weighted average of values from pixels in a neighborhood surrounding the specific pixel being computed. Here, the error of the quantization process is computed and spatially distributed or diffused within a local neighborhood in order to influence future pixel quantization decisions within that neighborhood and thereby improve the overall quality of the halftoned image. In classical unidirectional error diffusion, the image is processed sequentially, proceeding from the upper-left to the lower-right. Starting in the corner of the image, the first pixel is thresholded and the quantizer error is calculated. The error is then diffused to neighboring pixels that have not yet been processed according to the scaling and interconnect defined by the error diffusion filter. The remainder of the pixels in the image are subsequently processed with each quantizer input, now being the algebraic sum of the corresponding original input pixel intensity and the weighted error from previously processed pixels. While error diffusion produces halftone images of superior quality, the unidirectional processing of the algorithm continues to introduce visual artifacts that are directly attributable to the algorithm itself.

In an effort to overcome the implementation constraints of serial processing and improve overall halftone image quality, a number of different parallel architectures have been investigated, including several based on neural network algorithms. While neural network approaches can provide distinct advantages in improved halftone image quality, they also present challenges to hardware implementations, including speed of convergence and the physical interconnect requirements. These challenges, in addition to the natural compatibility with imaging media, provide the motivation for developing optical image processing architectures for these types of applications.

The Error Diffusion Algorithm

Figure 1 shows a block diagram of the error diffusion architecture used in digital image halftoning. The input image is represented as $x_{m,n}$, $w_{m,n}$ is the impulse response of a 2D causal, unity gain error diffusion filter, and the output quantized image is described by $y_{m,n} \in \{-1, 1\}$. The quantizer $q[u_{m,n}]$ provides the one-bit thresholding functionality necessary to convert each analog input pixel to a low-resolution digital output pixel. The quantizer error $\varepsilon_{m,n}$ is computed as the difference between the output and input to the quantizer and distributed to adjacent pixels according to the weighting and interconnection specified by the error diffusion filter. The unity gain constraint ensures

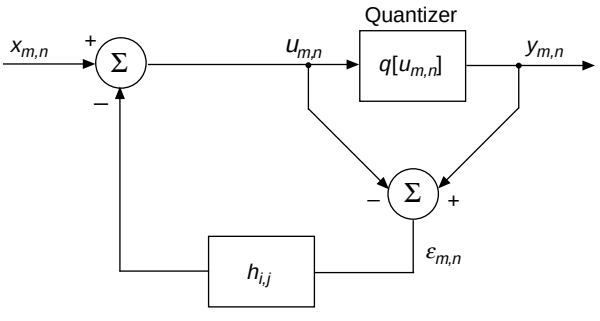


Figure 1 Block diagram of recursive error diffusion architecture.

that no amplification or attenuation of the error signal occurs during the error diffusion process and preserves the average intensity of the image.

In this architecture, the error associated with the quantizer decision at spatial coordinates (m, n) is diffused within a local 2D region to influence adjacent quantization decisions. All of the state variables in this architecture are scalar values since this architecture represents a recursive algorithm in which each pixel in the image is processed sequentially. In the general case, where $w_{m,n}$ is assumed to be a 2D finite impulse response (FIR) filter with coefficients that span a region of support defined by $R_{m,n}$, the quantizer input state $u_{m,n}$ can be written as

$$u_{m,n} = x_{m,n} - \sum_{i,j \in R_{m,n}} w_{i,j} \varepsilon_{m-i,n-j} \qquad [1]$$

Here, $w_{i,j} = 0$, for all $i = j$, since the error signal is diffused spatially within the region of support $R_{m,n}$ and does not influence the pixel under consideration. In this unidirectional error diffusion, the 2D error diffusion filter is therefore noncausal. Figure 2 shows the impulse response of several popular error diffusion filters that have been used in rectangular grid architectures. Figure 2a shows the original error diffusion filter weights developed by Floyd and Steinberg. They argued that the four weights were the smallest number of weights that would produce good halftone images. Figure 2b is a filter that contains a larger region of support developed by Jarvis, Judice, and Ninke in 1976. Finally, the filter in Figure 2c, developed by Stucki, contains the same region of support with coefficients that are a multiple of 2 for digital compatibility and computational efficiency. Here, '•' represents the pixel being processed and the weights describe the local region over which the error is diffused. The normalization factors ensure that the filter coefficients sum to one and therefore meet the unity gain criterion.

To understand the effect of error diffusion and its impact on the quantizer error, it is instructive to describe the effect of error diffusion mathematically. Consider the relationship between the quantizer error $\varepsilon_{m,n} \equiv y_{m,n} - u_{m,n}$ and the overall quantization error $\hat{\varepsilon}_{m,n} \equiv y_{m,n} - x_{m,n}$ in the frequency domain. If we assume that the error is uncorrelated with the input and has statistical properties consistent with a white process, z-transform techniques can be applied to show that the feedback architecture provides the following noise-shaping characteristics:

$$H_{\text{ns}}(z_1, z_2) \equiv \frac{\hat{E}(z_1, z_2)}{E(z_1, z_2)} = 1 - W(z_1, z_2) \qquad [2]$$

Here, $\hat{E}(z_1, z_2)$, $E(z_1, z_2)$, and $W(z_1, z_2)$ represent the z-transforms of $\hat{\varepsilon}_{m,n}$, $\varepsilon_{m,n}$, and $w_{m,n}$, respectively. Equation [2] shows that the noise-shaping characteristic of the error diffusion architecture is directly related to the spectral characteristics of $W(z_1, z_2)$. Appropriate selection of $W(z_1, z_2)$ can spectrally shape the quantizer noise in such a way that minimizes the effect of the low-resolution quantization process on the overall halftoning process. An error diffusion filter with low-pass spectral characteristics produces an overall noise-shaping characteristic $1 - W(z_1, z_2)$, which has the desired high-pass frequency characteristics. This noise-shaping function suppresses the quantization noise within those frequencies occupied by the image and spectrally shapes the noise to high frequencies, which are less objectionable to the human visual system. Circular symmetry is another important characteristic in the error diffusion filter since the human visual system is particularly sensitive to directional artifacts in the image.

In the following qualitative description of digital image halftoning, a 348×348 natural image of the Cadet Chapel at West Point was selected as the test image. This image provides a variety of important image characteristics, with regions of uniform gray-scale, edges, and areas with fine detail. This image was scanned from a high-resolution black and white photograph at 150 dots per inch (dpi) and then printed using a 300 dpi laser printer.

Figure 3 shows the test image of the Cadet Chapel rendered using dispersed-dot ordered dither and results in 256 gray-levels at 300 dpi. Figure 4 shows the halftone image of the Cadet Chapel using the error diffusion algorithm and the Floyd–Steinberg filter coefficients shown in Figure 2a. The unidirectionality of the processing and the causality of the diffusion filter result in undesirable visual artifacts in the halftone image. These include directional hysteresis, which is manifested as 'snakes' running from upper-left to lower-right, and transient behavior near boundaries, which appears as 'shadows' below and to the right of sharp-intensity changes. The directional

$$\left(\frac{1}{16}\right)\times \begin{array}{ccc} & \bullet & 7 \\ 3 & 5 & 1 \end{array}$$

(a)

$$\left(\frac{1}{48}\right)\times \begin{array}{ccccc} & & \bullet & 7 & 5 \\ 3 & 5 & 7 & 5 & 3 \\ 1 & 3 & 5 & 3 & 1 \end{array}$$

(b)

$$\left(\frac{1}{42}\right)\times \begin{array}{ccccc} & & \bullet & 8 & 4 \\ 2 & 4 & 8 & 4 & 2 \\ 1 & 2 & 4 & 2 & 1 \end{array}$$

(c)

Figure 2 Three common error diffusion filters used in rectangular grid architectures. (The '•' represents the origin.)

hysteresis is particularly objectionable in uniform gray intensity areas such as the cloud structure in the upper-left of Figure 4. Similar artifacts are also present in images halftoned using the error diffusion algorithm and other causal diffusion kernels. The logical conclusion to draw from this limited, qualitative analysis is that if we could diffuse the error symmetrically and simultaneously process the entire image in parallel, we could reduce some of these visual artifacts and thereby improve overall halftone image quality.



Figure 3 Original gray-scale image of the Cadet Chapel rendered using dispersed-dot ordered dither.



Figure 4 Halftone image of the Cadet Chapel produced using the error diffusion algorithm and the Floyd–Steinberg weights.

The Error Diffusion Algorithm and Neural Networks

The popularity of the neural network-based approach to signal and image processing lies in the ability to minimize a particular performance metric associated with a highly nonlinear system of equations. Specifically, the problem of creating a halftone image can be cast in terms of a nonlinear quadratic optimization problem where the performance measure to be minimized is the difference between the original and the halftone images.

The Hopfield-Type Neural Network

Tank and Hopfield first proposed a mathematical description of the functionality of the neural processing network for signal processing applications. Figure 5 shows an electronic implementation of a four-neuron architecture. Here, a single neuron is comprised of both a standard and inverting amplifier and the synapses or neural interconnections are represented by the physical connections between the input and output of the amplifiers. If the input to amplifier i is connected to the output of amplifier j by a resistor with resistance $R_{i,j}$, the amplitude of the connection $T_{i,j}$ is the conductance $1/R_{i,j}$.

The dynamic behavior of an N-neuron Hopfield-type neural network can be described by the following system of N nonlinear differential equations:

$$c\frac{\mathrm{d}u_i(t)}{\mathrm{d}t} = -u_i(t) + \sum_j T_{i,j}\Im[u_j(t)] + x_i \qquad [3]$$

where $i = 1, 2, \ldots N$, $\Im[\cdot]$ is a monotonically increasing sigmoid function, x_i is an input vector containing N

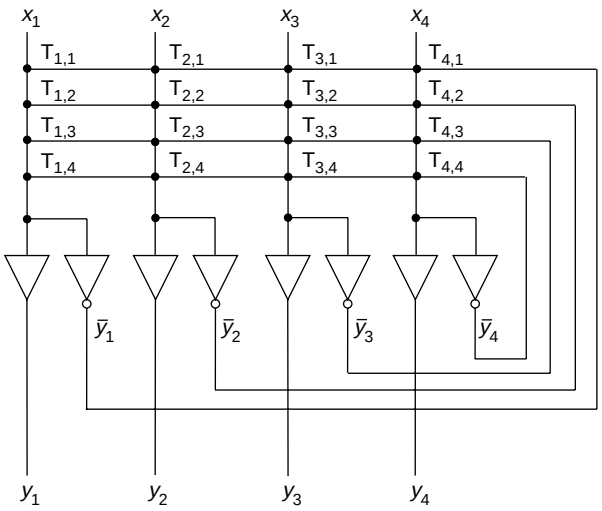


Figure 5 Electronic realization of a four-neuron Hopfield-type neural network.

elements, and c is a scaling factor. In equilibrium, eqn [3] implies that

$$u_i = x_i + \sum_j T_{i,j}\Im[u_j] \qquad [4]$$

Hopfield showed that when the matrix of interconnection weights $\mathbf{T}$ is symmetric with zero diagonal elements and the high-gain limit of the sigmoid $\Im[\cdot]$ is used, the stable states of the N functions $y_i(t) = \Im[u_j]$ are the local minima of the energy function:

$$E = -\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{T}\mathbf{y} - \mathbf{x}^{\mathrm{T}}\mathbf{y} \qquad [5]$$

where $\mathbf{y} \in \{-1, 1\}$ is the N-vector of quantized states. As a result, if the neural network can be shown to be stable, as the network converges the energy function is minimized. In most neural network applications, this energy function is designed to be proportional to a performance metric of interest and therefore as the network converges, the energy function and consequently the performance metric is also minimized.

The Error Diffusion Neural Network

An understanding of the operation and characteristics of the Hopfield-type neural network can be applied to the development and understanding of a 2D extension of the error diffusion algorithm. Figure 6 shows an electronic implementation of a four-neuron error diffusion-type neural network, where the individual neurons are represented as amplifiers and the synapses by the physical connections between the input and output of the amplifiers.

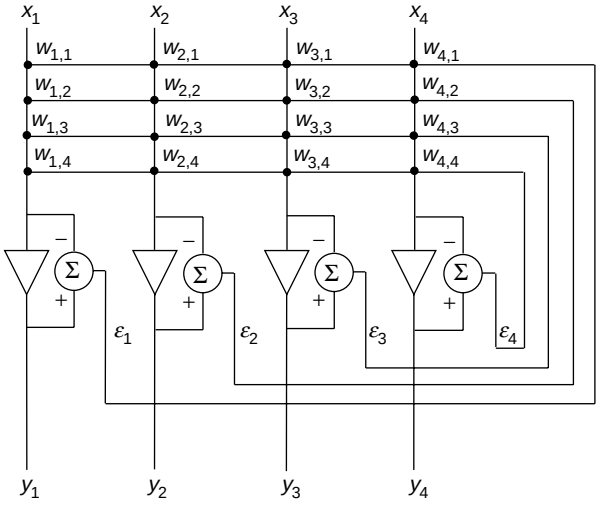


Figure 6 Four-neuron electronic implementation of the error diffusion neural network.

In equilibrium, the error diffusion neural network can be shown to satisfy

$$\mathbf{u} = \mathbf{W}(\mathbf{y} - \mathbf{u}) + \mathbf{x} \qquad [6]$$

Here, $\mathbf{u}$ is the state vector of neuron inputs, $\mathbf{W}$ is a matrix containing the interconnect weights, $\mathbf{x}$ is the input state vector, and $\mathbf{y}$ is the output state vector. For an $N \times N$ image, $\mathbf{W}$ is an $N^2 \times N^2$ sparse, circulant matrix derived from the original error diffusion weights $w_{m,n}$. If we define the coordinate system such that the central element of the error diffusion kernel is $(i, j) = (0, 0)$, then the matrix $\mathbf{W}$ is defined as $W(i, j) = -w[(j - i) \text{ div } N, (j - i) \bmod N]$ where:

$$x \text{ div } y = \left\lfloor \frac{x}{y} \right\rfloor \text{ if } xy \geq 0 \text{ or } \left\lceil \frac{x}{y} \right\rceil \text{ if } xy < 0 \qquad [7]$$

An equivalence to the Hopfield network can therefore be described by

$$\mathbf{u} = \mathbf{A}(\mathbf{W}\mathbf{y} + \mathbf{x}) \qquad [8]$$

where $\mathbf{A} = (\mathbf{I} + \mathbf{W})^{-1}$. Effectively, the error diffusion network includes a pre-filtering of the input image $\mathbf{x}$ by the matrix $\mathbf{A}$ while still filtering the output image $\mathbf{y}$ but now with a new matrix, $\mathbf{AW}$. Recognizing that $\mathbf{AW} = \mathbf{I} - \mathbf{A}$ and adding the arbitrary constant $k = \mathbf{y}^{\mathrm{T}}\mathbf{y} + \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, we can write the energy function of the error diffusion neural network as

$$\hat{E}(x, y) = \mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} \qquad [9]$$

This energy function is a quadratic function, which can be factored into

$$\hat{E}(x, y) = [\mathbf{B}\underbrace{(\mathbf{y} - \mathbf{x})}_{\text{error}}]^{\mathrm{T}}[\mathbf{B}\underbrace{(\mathbf{y} - \mathbf{x})}_{\text{error}}] \qquad [10]$$

where $\mathbf{A} = \mathbf{B}^{\mathrm{T}}\mathbf{B}$. From eqn [10] we find that as the error diffusion neural network converges and the energy function is minimized, so too is the error between the output and input images.

If the neurons update independently, the convergence of the error diffusion network is guaranteed if

$$\forall k : [\mathbf{AW}]_{k,k} \geq 0 \qquad [11]$$

We find in practice, that even in a synchronous implementation, the halftoned images converge to a solution, which results in significantly improved halftone image quality over other similar halftoning algorithms.

The Error Diffusion Filter

The purpose of the error diffusion filter is to spectrally shape the quantization noise in such a way that the

quantizer error is distributed to higher spatial frequencies, which are less objectionable to the human visual system. In this application, a feedback filter for the error diffusion neural network was designed using conventional 2D filter design techniques resulting in the 7×7 impulse response shown in Figure 7. Figure 8 shows the same image of the Cadet Chapel halftoned using the error diffusion neural network algorithm and the 2D error diffusion filter shown in Figure 7. There is clear improvement in halftone image quality over the image produced using sequential error diffusion in Figure 4. Notice particularly the uniform distribution of pixels in the cloud formation in the upper-left of Figure 8. Also noteworthy is the improvement around the fine detail portions of the tree branches and next to the vertical edges of the chapel.

0.0003	0.0019	0.0051	0.0068	0.0051	0.0019	0.0003
0.0019	0.0103	0.0248	0.0328	0.0248	0.0103	0.0019
0.0051	0.0248	0.0583	0.0766	0.0583	0.0248	0.0051
0.0068	0.0328	0.0766	•	0.0766	0.0328	0.0068
0.0051	0.0248	0.0583	0.0766	0.0583	0.0248	0.0051
0.0019	0.0103	0.0248	0.0328	0.0248	0.0103	0.0019
0.0003	0.0019	0.0051	0.0068	0.0051	0.0019	0.0003

Figure 7 Impulse response of a 7×7 symmetric error diffusion filter. ('•' represents the origin.)

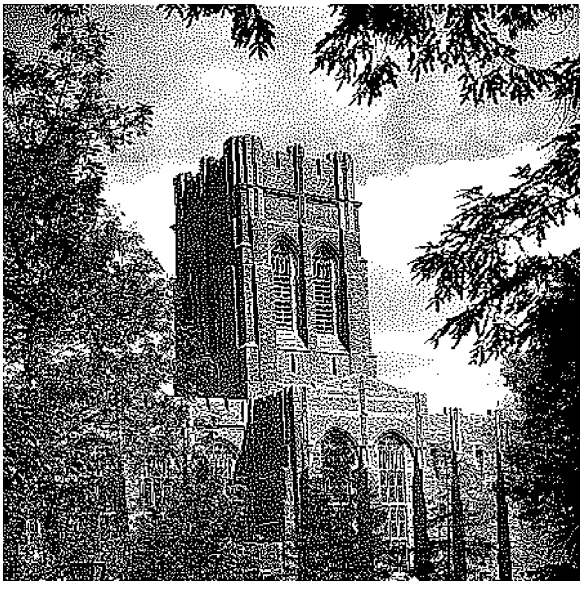


Figure 8 Halftone image of the Cadet Chapel using the error diffusion neural network and the 7×7 symmetric error diffusion filter shown in Figure 7.

Smart Pixel Technology and Optical Image Processing

The concept underlying smart pixel technology is to integrate electronic processing and individual optical devices on a common chip, to take advantage of the complexity of electronic processing circuits and the speed of optical devices. Arrays of these smart pixels then allow the advantage of parallelism that optics provides. There are a number of different approaches to smart pixel technology that differ primarily in the kind of opto-electronic devices, the type of electronic circuitry, and the method of integration of the two. Common opto-electronic devices found in smart pixels include: light-emitting diodes (LEDs); laser diodes (LDs); vertical cavity surface emitting lasers (VCSELs); multiple quantum well (MQW) modulators; liquid crystal devices (LCDs); and photodetectors (PDs). The most common types of electronic circuitry are silicon-based semiconductors such as complementary metal oxide semiconductors (CMOS) and compound semiconductors such as gallium arsinide (GaAs)-based circuitry. There are a number of different approaches to smart pixels, which generally differ in the way in which the electronic and optical devices are integrated. Monolithic integration, direct epitaxy, and hybrid integration are the three most common approaches in use today.

Smart pixel technology provides a natural methodology by which to implement optical image processing architectures. The opto-electronic devices provide a natural optical interface while the electronic circuitry provides the ability to perform either analog or digital computation. It is important to understand that any functionality that can be implemented in electronic circuitry can be integrated into a smart pixel architecture. The only limitations arise from physical space constraints imposed by the integration of the opto-electronic devices with the electronic circuitry. While smart pixels can be fabricated with digital or analog circuitry, the smart pixel architecture described subsequently uses mixed-signal circuitry.

A Smart Pixel Implementation of the Error Diffusion Neural Network

A smart pixel hardware implementation of the error diffusion neural network provides the potential to simultaneously achieve the computational complexity of electronic circuitry and the parallelism and high-speed switching of optics. The specific smart pixel architecture described here integrates MQW modulators called self electro-optic effect devices (SEEDs) with silicon CMOS

VLSI circuitry, using a hybrid integration approach called flip-chip bonding. This type of smart pixel is commonly referred to as CMOS-SEED smart pixel technology.

To provide an example of the application of smart pixel technology to digital image halftoning, a 5×5 CMOS-SEED smart pixel array was designed and fabricated. The CMOS circuitry was produced using a 0.5 μm silicon process and the SEED modulators were subsequently flip-chip bonded to silicon circuitry using a hybrid integration technique. The central neuron of this new smart pixel array consists of approximately 160 transistors while the complete 5×5 array accounts for over 3,600 transistors. A total of 50 optical input/output channels are provided in this implementation.

Figure 9 shows the circuitry associated with a single neuron of the smart pixel error diffusion neural network. All state variables in each circuit of this architecture are represented as currents. Beginning in the upper-left of the circuit and proceeding clockwise, the input optical signal incident on the SEED is continuous in intensity and represents the individual input analog pixel intensity. The input SEED at each neuron converts the optical signal to a photocurrent and subsequently, current mirrors are used to buffer and amplify the signal. The width-to-length ratio of the metal oxide semiconductor field effect transistors (MOSFETs) used in the CMOS-SEED circuitry provide current gain to amplify the photocurrent. The first circuit produces two output signals: $+I_u$ which represents the state variable $u_{m,n}$ as the input to the quantizer and $-I_e$ which represents the state variable $-\varepsilon_{m,n}$ as the input to the feedback differencing node. The function of the quantizer is to provide a smooth, continuous threshold functionality for the neuron that produces the output signal I_{out}, corresponding to the output state variable $y_{m,n}$. This second electronic circuit is called a modified wide-range transconductance amplifier and produces a hyperbolic tangent sigmoidal function when operated in the sub-threshold regime. The third circuit takes as its input I_{out}, produces a replica of the original signal, and drives

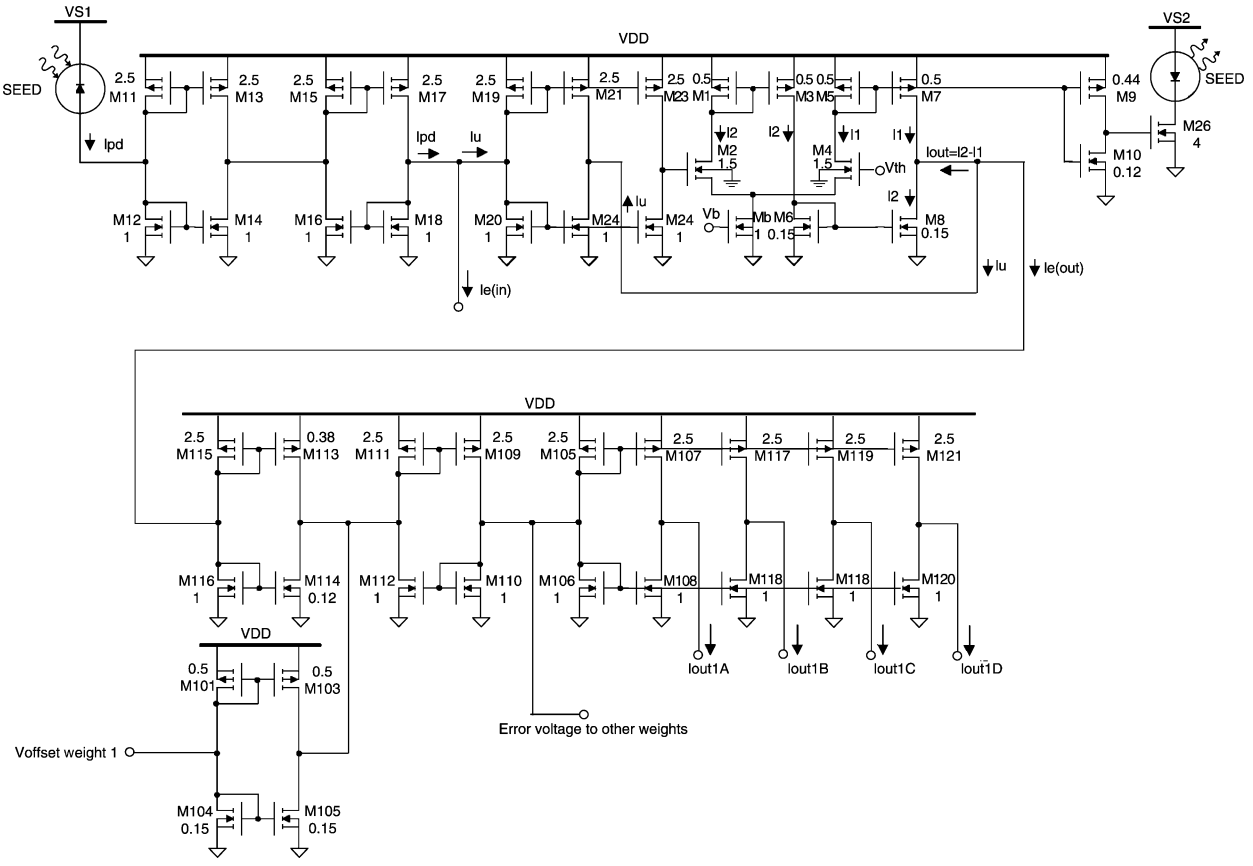


Figure 9 Circuit diagram of a single neuron and a single error weight of the 5×5 error diffusion neural network based on a CMOS-SEED-type smart pixel architecture. (Reprinted with permission from Shoop BL, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 8.6, p. 222. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

the output optical SEED. In this case, the output optical signal is a binary quantity represented as the presence or absence of light. When the SEED is forward-biased, light is generated through electroluminescence. The last circuit at the bottom of the schematic implements a portion of the error weighting and distribution function of the error diffusion filter. The individual weights are implemented by scaling the width-to-length ratio of the MOSFETs to achieve the desired weighting coefficients. The neuron-to-neuron interconnections are accomplished using the four metalization layers of the 0.5 μm silicon CMOS process. In this design, the error diffusion filter was limited to a 5×5 region of support because of the physical constraints of the circuitry necessary to implement the complete error diffusion neural network in silicon circuitry. The impulse response of the 5×5 filter used in this particular smart pixel architecture is shown in Figure 10. The error weighting circuitry at the bottom of Figure 9 represents only the largest weight (0.1124) with interconnects to its four local neighbors ($I_{\text{outA}} - I_{\text{outD}}$).

0.0023	0.0185	0.0758	0.0185	0.0023
0.0185	0.0254	0.1124	0.0254	0.0185
0.0758	0.1124	•	0.1124	0.0758
0.0185	0.0254	0.1124	0.0254	0.0185
0.0023	0.0185	0.0758	0.0185	0.0023

Figure 10 Error diffusion filter coefficients used in the 5×5 CMOS-SEED error diffusion architecture. ('•' represents the origin.)

Figure 11 shows a photomicrograph of a 5×5 error diffusion neural network, which was fabricated using CMOS-SEED smart pixels. Figure 12 shows a photomicrograph of a single neuron of the 5×5 CMOS-SEED neural network. The rectangular features are the MQW modulators while the silicon circuits are visible between and beneath the modulators. The MQW modulators are approximately 70 μm × 30 μm and have optical windows, which are 18 μm × 18 μm.

Simulations of this network predict individual neuron switching speeds of less than 1 μs, which corresponds to network convergence speeds capable of providing real-time digital image halftoning. Individual component functionality and dynamic operation of the full 5×5 neural array were both experimentally characterized. Figure 13 shows CCD images of the operational 5×5 CMOS-SEED smart pixel array. The white spots show SEED MQW modulators emitting light. Figure 13a shows the fully functioning array while Figure 13b shows the 5×5 CMOS-SEED neural array under 50% gray-scale input. Here 50% of the SEED modulators are shown to be in the *on*-state demonstrating correct operation of the network and the halftoning architecture.

Both the simulations and the experimental results demonstrate that this approach to a smart pixel implementation of the error diffusion neural network provides sufficient accuracy for the digital halftoning application. The individual neuron switching speeds also demonstrate the capability for this smart pixel hardware implementation to provide real-time halftoning of video images.

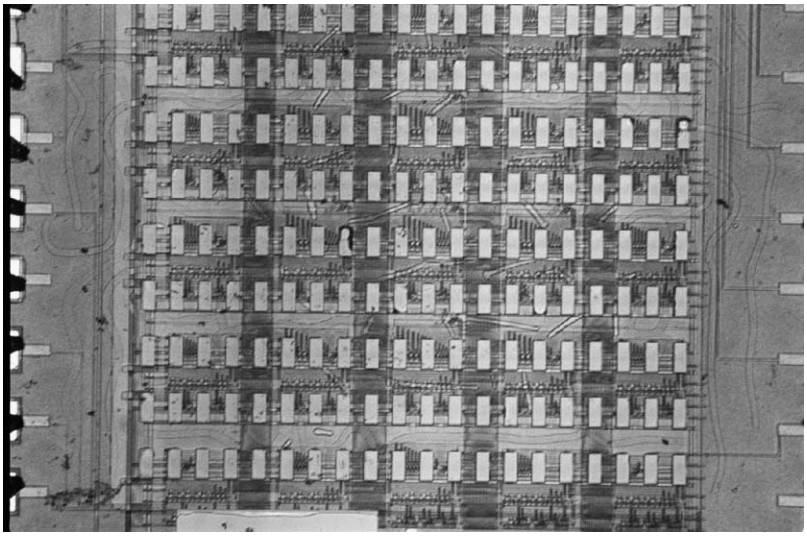


Figure 11 Photomicrograph of a smart pixel implementation of a 5 × 5 CMOS-SEED error diffusion neural network for digital image halftoning. (Reprinted with permission from Shoop BL, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 8.5, p. 221. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

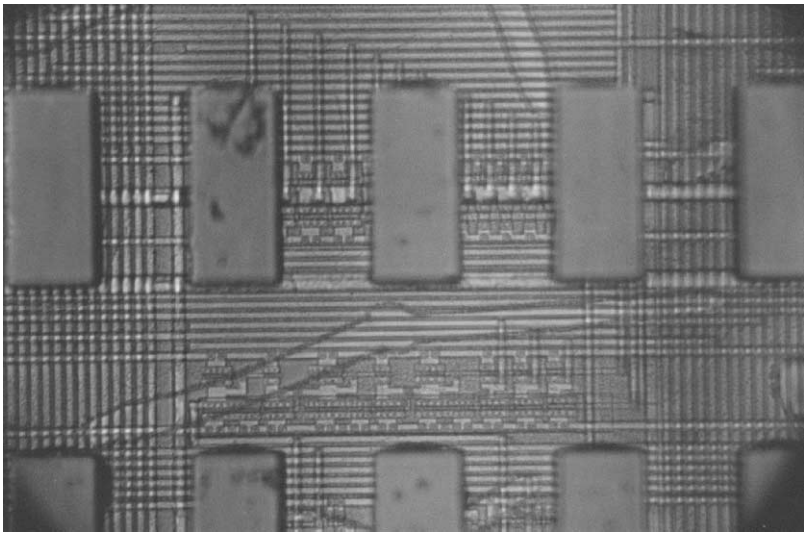
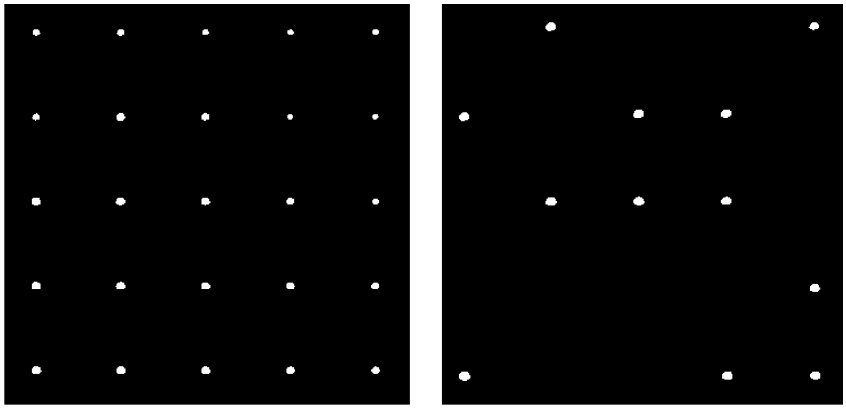


Figure 12 Photomicrograph of a single neuron of the 5 × 5 error diffusion neural network. (Reprinted with permission from Shoop BL, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 8.4, p. 220. Copyright 2001, Springer-Verlag GmbH & Co. KG.)



(a) (b)

Figure 13 CCD images of the 5 × 5 CMOS-SEED array. (a) Fully-operational array, and (b) under 50% gray scale illumination. (Reprinted with permission from Shoop BL, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Figs. 8.19 and 8.20, p. 236. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

Image Processing Extensions

Other image processing functionality is also possible by extending the fundamental concepts of 2D, symmetric error diffusion to other promising image processing applications. One important application includes color halftoning while other extensions include edge enhancement and feature extraction. In the basic error diffusion neural network, the error diffusion filter was specifically designed to produce visually pleasing halftoned images. Care was taken to ensure that the frequency response of the filter was circularly symmetric and that the cutoff frequency was chosen in such a way as to preserve image content. An analysis of the error diffusion neural network shows that the frequency response of this filter directly shapes the frequency spectrum of the output halftone image and therefore directly controls halftone image content. Other filter designs with different spectral responses can provide other image processing features such as edge enhancement, which could lead to feature extraction and automatic target recognition applications.

See also

Detection: Smart Pixel Arrays. **Information Processing:** Optical Neural Networks. **Optical Processing Systems**.

Further Reading

Dudgeon DE and Mersereau RM (1984) *Multidimensional Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Lau DL and Arce GR (2001) *Modern Digital Halftoning*. New York: Marcel Dekker.

Lim JS (1990) *Two-Dimensional Signal and Image Processing*. Englewood Cliffs, NJ: Prentice Hall.

Mead CA (1989) *Analog VLSI and Neural Systems*. New York: Addison-Wesley.

Poon T-C and Banerjee PP (2001) *Contemporary Optical Image Processing with Matlab*. New York: Elsevier.

Russ JC (1998) *The Image Processing Handbook*. Boca Raton, FL: CRC Press.

Shoop BL (2001) *Photonic Analog-to-Digital Conversion*. Springer Series in Applied Optics, vol. 81. Berlin: Springer-Verlag.

Shoop BL, Sayles AH and Litynski DM (2002) New devices for optoelectronics: smart pixels. In: De Cusatis (ed.) *Fiber Optic Data Communication: Technological Trends and Advances*, pp. 352–421. San Diego, CA: Academic Press.

Ulichney R (1987) *Digital Halftoning*. Cambridge, MA: The MIT Press.

Wax C (1990) *The Mezzotint: History and Technique*. New York: Harry N. Abrahms, Inc.

Optical Neural Networks

H J Caulfield, Fisk University, Nashville, TN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

An artificial neural network (ANN) is a nonlinear signal processing system based on the neural processes observed in animals. Usually they have multiple inputs and often multiple outputs also. Conventionally, each input sends its signal to many neurons, and each neuron receives signals from many inputs. The neuron forms an intermediate sum of weighted inputs and transforms that sum, according to some nonlinearity, to form an output signal. Often the output signal from one neuron is used as the input signal for another in feed forward, feedback, or mixed mode complex neural networks.

ANNs are of value in helping us understand how biological neural networks perform and in a variety of technical applications, most of which involve pattern association. An auto-associative neural network, that associates patterns with themselves, can have useful noise cleanup or pattern completion properties. A hetero-associative neural network is often used in pattern recognition to associate input patterns with preselected class labels.

Neural Networks – Natural and Artificial

The most primitive one-celled organisms have no nerve cells, but they do exhibit various chemical means for interacting with their environments, within their own cells, and among individuals of the same or different types. As organisms increased in size, a common chemical environment became impractical, and nerves arose as a means for long-distance communication within larger organisms. One thing seen in evolution is that widespread useful adaptations are often modified but seldom discarded. Thus, it should not be surprising to find that chemical interactions play a major role in the most complex thing we know of – the human brain. Chemicals play a huge role in making us who we are, setting our moods, and so forth. Realizing this, in the late twentieth century, psychiatrists started to make serious progress in treating mental diseases through the use of chemistry.

There are a variety of processes that we use as humans, but the primary means is neural. Only some animals and no plants have neurons, but soon after neurons evolved, they began to concentrate in special regions of the body that developed into primitive brains. Not only were there many neurons, but also they were densely and complexly interconnected to form neural networks.

Neural networks did not replace chemical interactions, they used and supplemented them. Chemical communication across the synaptic cleft dominates communication between networks, but the neural networks also control chemical production and distribution within the brain, and those chemicals change the behavior of the neurons that cause their secretion. The functioning of real brains is incredibly complex, possibly irreducibly complex in terms of current mathematics and logic.

These things are noted to avoid a confusion that is distressingly common. Technologists produce simplified ANNs and then try to interpret the behavior of brains in those terms. Quite often, ANNs yield useful insight into brain behavior but seldom offer a detailed account of them. Reality is far more complex than any ANN. Current mathematics provides no means

to describe continuous sets of events, each of which both causes and is caused by the others.

ANNs themselves have grown much more complicated as well. Starting out as simple, feed-forward, single-layer systems, they rapidly became bidirectional and even multilayered as well. Higher-order neural networks processed the input signals before inserting them into the neural network. Pulse coupled neural networks (PCNNs) arose that more closely modeled the pulsations in real brains. Learning was broken into learning with instruction, following example, and self-organizing, the latter leading to self organized maps and adaptive resonance theory, and many related subjects. These are but crude pointers to the wealth and variety of concepts now part of the ANN field.

The history of ANNs is both complicated and somewhat sordid. This is not the occasion to retrace the unpleasant things various members of this community have done to others in their quests for stature in the field. One result of that history is a dramatically up-and-down pattern of interest in and support for ANNs. The first boom followed the development of a powerful single-layer neural network called the Perceptron by its inventor Rosenblatt. Perceptrons were biologically motivated, fast, and reasonably effective; so interest in the new field of ANNs was high. That interest was destroyed by a book on the field arguing that Perceptrons were only linear discriminants and most interesting problems are not linearly discriminable. Ironically, the same argument has been made about optical Fourier correlators, and it is true there as well. But optical processing folk largely ignored the problem and kept working on it. Sociologists of science might enjoy asking why reactions differed so much in the two fields. Funding and interest collapsed and the field went into hibernation.

It was widely recognized that multilayer ANNs would comprise nonlinear discriminants, but training them presented a real problem. The simple rewards and punishment of weights depending on performance was easy when each output signal came from a known set of weighted signals that could be rewarded or punished jointly, according to how the output differed from the sought-after output. With a multilayer perceptron, credit assignment of inner layers is no longer simple. It was not until Werbos invented 'backward error propagation' that a solution to this problem was available and the field started to boom again. Also key to the second boom were enormously popular articles by Hopfield on rather simple auto- and hetero-associative neural networks. Basically, those networks are almost never used, but they were of great value in restarting the field. In terms of research and applications, we are still in that second boom, that began during the 1980s.

> An ANN is a signal processor comprised of a multiplicity of nodes which receive one or more input signals and generate one or more output signals in a nonlinear fashion wherein those nodes are connected with one or two way signaling channels.

Note also that biological inspiration is no longer a part of the definition. Most ANNs used for association are only vaguely related to their biological counterparts.

Optical Processors

Optics is well-suited for massive parallel interconnections, so it seems logical to explore optical ANNs. The first question to address is "Why bother to use optical processors at all?" as electronic digital computers seem to have everything:

- Between the writing of this article (on an electronic computer) and its publication, the speed of electronic computers will almost double and they will become even cheaper and faster.
- They can have arbitrary dynamic range and accuracy.
- Massively improved algorithms also improve their speed. For instance, the wonderful fast Fourier transform (FFT) that opticists have used to simulate optics for decades has been replaced by an even faster fastest Fourier transform in the West (FFTW). And for wavelets, the fast wavelet transform (FWT) is even faster.

Optical processors, on the other hand, have severely limited analog accuracy and few analog algotectures or archirithms. They are specialized systems (equivalent in electronics, not to computers but to application specific integrated circuits (ASICs)). They are usually larger and clumsier than electronic systems and always cost more.

There are several reasons why optics may still have a role in some computations. First, many things on which we wish to perform computations are inherently optical. If we can do the computations optically before detection, we can avoid some time and energy and sensitivity penalties inflicted, if we first detect the signals then process them electronically, and then (sometimes) convert them back to optics. Examples include spectroscopy and optical communication.

Second, pure optics is pure quantum mechanics. Input of data and setup of the apparatus is experiment preparation. Readout of data is classical measurement. Nature does the rest free of charge. There is a unitary transformation from input to output that

requires no human intervention and is, in fact, destroyed by human intervention. Those intermediate (virtual) computations require no energy and no time, beyond the time taken for light to propagate through the system. Speed and power consumption are limited only by the input and output. This applies for everything from Fourier optics to digital optics to logic gates. Thus, optics can have speed and power consumption advantages over electronics.

Third, if quantum computers involving entanglement ever emerge, they may well be optical. Such computers would be used for nondeterministic solutions of hard problems. That is, they would explore all possible paths at once and see which one does best. This is something we can already do electronically or optically for maze running (using classical, not quantum entanglement) means.

Fourth, most optical processors are analog, which is sometimes superior to digital. In those cases, we would still have to choose between analog optics and analog electronics, of course.

So, optical processing (like all other processing) is a niche market. But some of the niches may be quite large and important. It is not a universal solution, but neither is digital electronics.

The Major Branches of Optical Processing and Their Corresponding Neural Networks

Every branch of optical processing leads to its own special purpose optical neural networks, so those branches become a good place to start an article such as this one.

Optical Linear Algebra

Many neural network paradigms use linear algebra followed by a point-by-point nonlinearity in each of multiple stages. The linear operation is a matrix–vector multiplication, a matrix–matrix multiplication, or even a matrix–tensor multiplication. In any case, the key operations are multiplication and addition – operations very amenable to performance with analog optics. With effort, it is possible to combine simple electronics with analog optics operating on non-negative signals (the amount of light emitted from a source, the fraction of light passing through a modulator, etc.) to expand to real or even complex numbers. With even more effort, we can use multiple analog non-negative signals to encrypt a digital signal, allowing digital optical processors. Almost all optical neural networks using linear algebra use only real numbers. In most of these, the real number is encoded as a positive number and a negative number. Data are read in parallel using spatial light modulators (SLMs), source arrays, acousto-optic modulators, etc. They are operated upon by other SLMs, acousto-optic systems, and so forth. Finally, they are detected by individual detectors or detector arrays. The required nonlinearities are then done electronically. It is a theory that holds for all linear algebraic processes up to computational complexity $O(N^4)$ that they can be performed with temporal complexity $O(N^0)$ if the extra dimensions of complexity are absorbed in space, up to $O(N^2)$ through 2D arraying and in fanin/fanout, the ability of multiple beams of light to be operated upon in parallel by the same physical modulator – this has the same complexity as the 2D array of modulators, namely $O(N^2)$. Thus, we can invert a matrix, an order $O(N^4)$ for an $N \times N$ matrix, $O(N^0)$ in time, that is, independently of the size of N, as long as that matrix is accommodated by the apparatus. This offers a speed electronics cannot match.

Most of the optical neural networks of the mid-1980s had optical vector–matrix multipliers at their heart. Most were a few layers of feed forward systems, but by the 1990s feedback had been incorporated as well. These are typically N inputs connected to N outputs through an $N \times N$ matrix, with N varying from 100 to 1000. With diffuser interconnection and a pair of SLMs, it is possible to connect an $N \times N$ input with an $N \times N$ output using N^4 arbitrary weights, but that requires integrating over a time up to N^2 intervals. Most of the N^4-dimensional matrices can be approximated well by fewer than N^4 terms using singular value decomposition (SVD) and (as this version of the interconnection uses outer products), that reduces the number of integration times considerably in most cases. We have shown, for instance, that recognizing M target images can be accomplished well with only M terms in the SVD.

Coherent Optical Fourier Transformation

It took opticists several years, during the early 1960s, to realize that if Fourier mathematics was useful in describing optics, optics might be useful in performing Fourier mathematics. The first applications of optical Fourier transforms were in image filtering and particularly in pattern recognition. Such processors are attractive not only for their speed but also for their ability to locate the recognized pattern in 1D, 2D, or even 3D. It was soon obvious that a Fourier optical correlator was a special case of a single layer perceptron. Of course, it inherits the weaknesses of a single layer perceptron too. Soon, workers had added

feedback to create a winner-takes-all type of auto-associative optical neural network, a system that associated images with themselves. This proved useful in restoring partial or noisy images. It was not until 2003 that we learned how to make powerful hetero-associative networks with Fourier optical systems. This creates a nonlinear discrimination system that maintains the target location and generalizes well from a few samples of multiple classes to classify new data quite accurately.

Imaging

Imaging is the central act of optics and might seem to be of little value in optical processing. But multiple imaging (with lenslets, self-focusing fibers, holograms, and the like) has been the basis of numerous optical neural networks. Multiple imaging is a nonrandom way of distributing information from one plane to the next. Diffusers have also been used to produce fully interconnected optical neural networks. If each broad image point contains the whole image (something easily arranged with a diffuse hologram), then we have an auto-associative memory. Different such auto-associative networks can be superimposed in a single hologram. Then feedback after nonlinear image filtering can be used to make a general auto-associative memory.

Even deliberately poor imaging has been used. A blurred or out-of-focus image can accomplish the lateral signal spreading needed to facilitate a PCNN. These networks are substantially more like real brain networks than those discussed so far. It starts with a 2D array of integrate-and-fire units. The firing rate of each depends only in the incident power (how long integration to the thresholding firing takes). But if the fired signal is spread to its neighbors, the firing of a neighbor can hasten any unit's firing. This causes firings to synchronize and results in 'autowaves' of firing patterns moving across the 2D plane. Thus an incident 2D image is converted to a 3D signal – the two spatial dimensions and time. This results, for all but some very simple cases, in a chaotic signal occupying a strange attractor. It is interesting but not very revealing in itself. But suppose we 'integrate out' either the two spatial dimensions or the time dimension. If we integrate out the spatial dimensions, we are left with a time signal. Its strange attractors are remarkably useful in pattern recognition, because they are more syntactic than statistical (in pattern recognition terms). They describe the shape as similar, independently of the size, rotation, perspective, or even shading of the pattern! If we integrate out the time dimension, we restore the original image – almost. In fact, we can use this aspect of PCNNs for optimum blind segmentation of images. The autowaves can be made to 'leave tracks' so they can be reversed, which allows them to solve maze problems nondeterministically, because they can take all the paths. When the first wave exits, we simply retrace it to find the shortest path through the maze – a prototypical nondeterministic polynomial operation.

Holograms

Holograms are the most general, complex, and flexible means of light manipulation so far invented. Naturally, they have many applications in optical neural networks. One such application is the interconnection of each of an $N \times N$ array of inputs to each of an $N \times N$ array of outputs with its own unique connection strength (weight in the language of neural networks). The practical limit seems to be around $N = 1000$, yielding 10^{12} weights, roughly the same number as in the human brain! The brain, however, is not fully interconnected, so this produces an entirely different type of neural network. The resulting network would have fixed weights, but it could perform a variety of useful operations and be both forward and backward connected.

Special on-axis holograms, called diffractive optical elements, can be mass manufactured very effectively and inexpensively to perform tasks such as multiple imaging.

Photorefractive Neural Networks

A very complex thing happens when a pattern of fringes, such as those that form an interference pattern for hologram recording, strike materials such as lithium niobate, strontium barium niobate, gallium arsenide, and many others that fall into this category of 'photorefractive' materials. Of course, 'photo' refers to light and 'refractive' refers to the speed of light in a material – the speed of light in a vacuum divided by the index of refraction. These materials are photoconductive, so the light releases charge carriers in what would otherwise be an electrical insulator. The periodic charge pattern sets up a periodic space charge within the material. But the material is also electro-optic, that is, its index of refraction changes with applied electric field, thus generating a periodic index of refraction grating in the material. Finally, that index of refraction pattern (a phase hologram) diffracts the incoming light. Usually, but not always, a steady-state situation is reached after time that depends on the material and the photon density, although chaos can arise under some conditions. This makes for a dynamic

hologram, one that can change or adapt over time as the input pattern changes. Once a satisfactory photorefractive hologram is formed, it can be stabilized (fixed) if desired. Most optical neural network uses envision a dynamic situation with changing interconnections.

What has not been discussed above is that the index of refraction varies with the light's electric field vector cubed, so photorefractives provide a convenient way to accomplish such nonlinear operations as winner-takes-all.

Sometimes, we can record holograms with photorefractives that cannot be recorded by conventional holograms in conventional materials. To explain this, a hologram is a transducer between two optical wavefronts – usually called the reference and object beams. If the reference beam is incident on the hologram, it is (fully or partially) converted into the object beam. If the precisely reversed (phase conjugated) reference beam is incident on it, the precisely reversed (phase conjugated) object beam is derived from it. Likewise, the object wavefront (or its phase conjugate) produces the reference wavefront (or its phase conjugate). The hologram is made by recording the interference pattern between the two wavefronts: the reference and object beams. The ability of two wavefronts to form a temporally stable interference pattern at some point in space is called the mutual coherence of those wavefronts. Usually, both wavefronts are derived from the same laser beam to make achieving high mutual coherence possible. And, of course, both beams must be present to form an interference pattern. In some cases of photorefractive holograms, all of those conditions can be violated. The two wavefronts can derive from two sources that are different in wavelength, polarization, etc., and need not be simultaneously present. This phenomenon, called mutual phase conjugation, converts each wavefront into the phase conjugate of the other, but for effects due to any differences in wavelength and polarization. This allows new freedoms that have been exploited in optical neural networks.

Photorefractive materials also tend to be piezoelectric. It has been found that up to about 20 totally independent reflective holograms can be stored with different applied electric field in lithium niobate. Changing electric fields so changes the fringe spacings and indices of refraction, that only one of those holograms has any appreciable diffraction efficiency at any one electric field value. The result is an electrically selectable interconnection pattern. The same hologram can contain the information needed for many neural networks or multiple layers of fewer of them. There is no need, in general, to have more than an input layer, an output layer, and two intermediate or hidden layers.

Others have used photorefractive material to implement learning in neural networks. In biological neural networks the synaptic strengths (the weights between neurons) continue to adapt. With photorefractive holograms, we can achieve this in an optical ANN.

As previously discussed, a key development in the history of ANNs was a way to train multilayer neural networks and thus accomplish nonlinear discrimination. The first, and still most popular, way to do this is called backward error propagation. Several groups have been implemented in photorefractive optical neural networks.

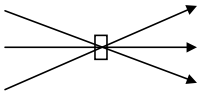
Conclusions

Essentially every major type of neural network has been implemented optically. Every major tool of optics has found its way into these optical neural networks. Yet, it remains the case that almost every neural network used for practical applications is electronic. The mere existence proof of a use for optical neural networks may be what the field needs to move forward from the demonstration stage to the application stage. The tools exist and there are niches where optics seems more appropriate than electronics. Hopefully, the next review of the field will include a number of commercial successes. Nothing spurs science better than the funds that follow successful applications.

List of Units and Nomenclature

Adaptive resonance theory (ART)	Another self-organized neural net categorizer that works by mutual adjustment between bottom-up and top-down neural connections.
Computational complexity	The measure of how the number of required calculations scales with the problem size, usually denoted by N. The scaling is represented by the Big O form of (N), where $f(n)$ shows how it scales asymptotically (large N). Problems that scale as $O(N^P)$ are said to be of polynomial complexity. Unfortunately, some of the most important problems have exponential complexity.

Fanin/Fanout
: Beams of light (unlike beams of electrons or holes) do not interfere with one another, so many of them can be modulated by the same physical elements quite independently, so long as they can be identified. This is usually done by sending the different beams in at different angles, giving rise to the words fanin and fanout, as suggested by the drawing below:



Self organized maps (Kohonen networks)
: A means of self-organization of inputs into (normally 2D) maps wherein every concept falls into some category and the categories themselves develop out of application of the Kohonen's algorithm and user-selected parameters applied to the given data.

Singular value decomposition (SVD)
: Any matrix can be broken into (decomposed) a weighted sum of simpler, rank-one matrices. The simplest such decomposition (in the sense of fewest number of terms) is the SVD. The weights are the singular values and are indicators of the importance of that term in the decomposition.

Spatial light modulator (SLM)
: A 1D or 2D array of elements under electronic or optical control that change some property of the light incident on them. Usually, it is the index of refraction in some direction that is changed. That change can be used to modulate the amplitude, the polarization, or the phase of the light.

See also

Fourier Optics. **Holography, Techniques:** Overview. **Nonlinear Optics, Basics:** Photorefraction. **Optical Communication Systems:** Local Area Networks. **Optical Processing Systems**.

Further Reading

Abu-Mostafa YS and Psaltis D (1987) Optical neural computers. *Scientific American* 256(3): 88–95.

Anderson DZ and Erie MC (1987) Resonator memories and optical novelty filters. *Optical Engineering* 26: 434–444.

Brady D, Psaltis D and Wagner K (1988) Adaptive optical networks using photorefractive crystals. *Applied Optics* 27: 1752–1759.

Caulfield HJ, Kinser J and Rogers SJ (1987) Optical neural networks. *Proc. IEEE* 77: 1573–1583.

Chavel P, Lalanne P and Taboury J (1989) Optical inner-product implementation of neural network models. *Applied Optics* 28: 377–385.

Collins SA, Ahalt SC, Krishnamurthy AK and Stewart DF (1993) Optical implementation and applications of closest-vector selection in neural networks. *Applied Optics* 32: 1297–1303.

Farhat NH, Psaltis D, Prata A and Paek E (1985) Optical implementation of the Hopfield model. *Applied Optics* 24: 1469–1475.

Galstyan T, Frauel Y, Pauliat G, Villing A and Roosen G (1997) Topological map from a photorefractive self-organizing neural network. *Optics Communications* 135: 179–188.

Goodman JW, Dias AR and Woody JM (1978) Fully parallel, high-speed incoherent optical method for performing discrete Fourier transforms. *Optics Letters* 2: 1–3.

Johnson JL (1994) Pulse-coupled neural nets: Translation, rotation, scale, distortion, and intensity signal invariance for images. *Applied Optics* 33: 6239–6253.

Jutamulia S (1994) *Selected Papers on Optical Neural Networks*. Bellingham: SPIE Press.

Kinser JM, Caulfield HJ and Shamir J (1988) Design for a massive all-optical bidirectional associative memory: the big BAM. *Applied Optics* 27: 3442–3444.

Neifeld MA and Psaltis D (1993) Optical implementations of radial basis classifiers. *Applied Optics* 32: 1370–1379.

Owechko Y, Dunning GJ, Marom E and Soffer BH (1987) Holographic associative memory with nonlinearities in the correlation domain. *Applied Optics* 26: 1900–1910.

Paek EG and Psaltis D (1987) Optical associative memory using Fourier-transform holograms. *Optical Engineering* 26: 428–433.

Psaltis D and Farhat NH (1985) Optical information processing based on associative-memory model of neural nets with thresholding and feedback. *Optics Letters* 10: 98–100.

Soffer BH, Dunning GJ, Owechko Y and Marom E (1986) Associative holographic memory with feedback using phase-conjugate mirrors. *Optics Letters* 11(2): 118–120.

Yu FTS (1993) Optical neural networks – architecture, design and models. In: Wolf E (ed.) *Progress in Optics*, II, pp. 63–144. Amsterdam: North-Holland.

INSTRUMENTATION

Contents

Astronomical Instrumentation
Ellipsometry
Photometry
Scatterometry
Spectrometers
Telescopes

Astronomical Instrumentation

J Allington-Smith, University of Durham, Durham, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article describes the requirements placed on astronomical instruments and the methods used to achieve them, making reference to generic techniques described elsewhere in this encyclopedia. It is limited in scope to electromagnetic radiation with wavelengths from the atmospheric cutoff in the ultraviolet (0.3 μm) to the point in the near-infrared where the technology changes radically (5 μm), taking in the transition from uncooled to cooled instruments at ~1.7 μm. Thus, it excludes the extreme UV and X-ray regimes, where standard optical techniques become ineffective (e.g., requiring grazing incidence optics) and all mid-infrared, sub-mm, THz, and radio-frequency instruments. Terrestrial telescopes are the main focus but space-borne telescopes are subject to similar design considerations, except that the diffraction limit of the optics may be achieved at short wavelengths. Adaptive optics (AO), of increasing importance in astronomy and almost essential for the next generation of extremely large telescopes (ELTs), is covered elsewhere in this encyclopedia.

Astronomical Data

The purpose of astronomical instruments is to measure the following properties of photons arriving on or near to the Earth and brought to a focus by an astronomical telescope:

(i) direction of arrival;
(ii) energy;
(iii) polarization state; and
(iv) time of arrival.

In practice, these properties cannot be measured simultaneously, requiring the following generic classes of instrument:

- *Imagers* to measure (i) and, if equipped with ultra-low-noise detectors, (iv);
- *Spectrographs* to measure (ii), although some can be configured to do a good job on (i) as well. *Integral field spectrographs* can measure (i) and (ii) simultaneously; and
- *Polarimeters* for (iii) plus (i) and *spectropolarimeters for* (iii) plus (ii).

In principle, energy-resolving detectors, such as arrays of superconducting tunnel-junction devices (STJs) can replace (i), (ii), and (iv).

These must interface to the telescope which provides a focal surface with an image scale which is approximately constant over the field. Currently, almost all telescopes are based on the Cassegrain or Ritchy-Chretien reflecting configurations, comprising a large curved primary mirror (M1) which defines the collecting aperture (diameter D_T) and a curved secondary mirror (M2). This feeds the Cassegrain focus via a hole in M1 in line with the central obstruction caused by M2, or the Nasmyth focus if a tertiary (usually plane) mirror is added to direct light from M2 along the elevation axis to a location in which the gravity vector changes only in rotation about the optical axis. For the Cassegrain focus, the full range of change in the gravity vector is experienced as the telescope tracks objects during long exposures or slews between targets. It is also possible to dispense with M2 and mount the

Table 1 Optical interfaces to representative types of telescope and achievable angular resolution

Aperture diameter D_T (m)	*Primary focus*			*Final focus*			*Diffraction limit $2.44\lambda/D_T$(mas)*	
	F/#	*Image scale ($\mu m/''$)*	*FOV ($'$)*	*F/#*	*Image scale ($\mu m/''$)*	*FOV ($'$)*	*$\lambda = 0.5\ \mu m$*	*$\lambda = 5.0\ \mu m$*
2.4[i]	2.3			24	279	25	101	1007
4.2[ii]	2.8	60	40	11	221	15	60	600
8.0[iii]	1.8	(70)	(45)	16	621	10	32	315
30[iv]	1–1.5	~180	~20	15–19	~2500	~20	8.4	84
100[v]	~1.4			~6.5	~3000	~10	2.5	25
			For enclosed telescopes				500–1500	
All	Obtainable angular resolution for terrestrial telescopes		Free-atmosphere seeing at good site			Best	250	160
						Median	420	270
			With adaptive optics (approx)				50–200	*Diff. limit*[a]

[a]The diffraction limit listed in the column above for $\lambda = 5.0\ \mu m$ applies.
Exemplars: (i) Hubble Space Telescope; (ii) Herschel Telescope, La Palma; (iii) Gemini telescopes, Hawaii/Chile; (iv) CELT/GSMT (USA); (v) OWL (Europe).

instrument directly at the focus of M1 (prime focus), although this imposes severe restrictions on the mass and size of the instrument. An alternative mounting for instruments is via optical fibers. The input, with any fore-optics required to couple the telescope beam into the fiber, is mounted at the appropriate focus while the output, possibly several tens of meters distant, is mounted where the instrument is not subjected to variations in the gravity vector.

The useful field of view (FOV) of the telescope is determined by how rapidly the image quality degrades with radius in the field; restrictions on the size of the Cassegrain hole and the space set aside for the star trackers required for accurate guiding and auxiliary optics involved in target acquisition, AO, etc.

Typical telescope parameters are shown in Table 1, of which the last two entries are speculative examples of ELTs. The table also lists the diffraction limit of the telescopes and the limits imposed by seeing, with and without the effects of the telescope structure and enclosure, and approximate limits for performance with AO which approaches the diffraction limit in the infrared.

Instrument Requirements

Efficiency

Almost all astronomical observations are of very faint objects or require very high signal-to-noise (SNR) to extract the desired subtle information. The former, for example, might reveal the initial phase of galaxy or star formation, while the latter could reveal the mix of elements in the big bang or uncover changes in fundamental constants on cosmological timescales ($\sim 10^{10}$ years). The term 'efficiency' has several components:

(i) Throughput of the optical system, including the efficiency of the optical components (lenses or mirrors), dispersers, filters, and also detectors and also includes losses due to transmission/reflection, diffraction, scattering, and vignetting by internal obstructions;
(ii) Noise from all sources: the detector, including time-independent (read-noise) and time-dependent (dark-current) components; thermal emission from the instrument, telescope, enclosure and sky; straylight (see above), and electronic interference;
(iii) Efficient use of the instrument and telescope to maximize the fraction of the time spent accumulating useful photons instead of, say, calibrating; reconfiguring the instrument; or acquiring targets;
(iv) Parallelizing the observations so that the same instrument records data from many objects at the same time. This is discussed further below.

Angular Resolution

The required angular resolution is summarized in Table 2 for key features in galaxies. The resolutions currently achievable with AO systems (with severe restrictions on sky coverage pending routine availability of laser beacons), or without, are shown in bold. Finer resolution is possible using interferometric techniques, but is limited to a handful of bright stars.

Table 2 Typical requirements for angular resolution and single-object field of view

Target	*Distance*[a]	*Active nucleus*		*Star cluster (10 pc)*	*Visible extent*
		Central engine (10 AU)	*Dynamical core (0.5 pc)*		
Milky Way	8 kpc	1 mas	10″	4′	
Local group galaxy (M31)	800 kpc	0.01 mas	100 mas	3″	100″
Distant galaxy	>1 Gpc	10^{-5} mas	0.1 mas	2 mas	0.5–2″
Cluster of galaxies	>1 Gpc				3′
Cosmological structure	>1 Gpc				~1°

[a]1 pc = 3.1×10^{16} m.

Field of View and Multiplexing

The FOV of an individual observation must be sufficient to cover the angular extent of the object with the minimum number of separate pointings of the telescope. As well as minimizing the total observation time, this reduces photometric errors caused by changes in the ambient conditions (e.g., flexure, thermal effects, and changes in the sky background). (Photometry is the measurement of the flux density received from an astronomical object, usually done with the aid of an imager employing filters with standardized broad passbands. Spectrophotometry is the same but in the much narrower passbands obtained by dispersing the light.) The last column of Table 2 indicates the required size of FOV. It is also highly desirable to observe many objects simultaneously since most studies require large statistically homogenous samples. The precise requirement depends on the surface number density of potential targets and the sampling strategy adopted to create the sample. Cosmological studies of the large-scale distribution of galaxies and clusters emphasize the size of the field ($\geq 0.5°$), since the targets may be sparsely sampled, while studies of clustered objects (galaxies or stars) aim to maximize the surface density of targets (≤ 1000 arcmin^{-2}) within a modest FOV (~5′).

Spectral Resolution

The required energy resolution is specified in terms of the spectral resolving power, $R \equiv \lambda/\delta\lambda$, where $\delta\lambda$ is the resolution in wavelength defined usually in terms of the Rayleigh criterion. One of the main data products required in astrophysics is the velocity along the line of sight (radial velocity) obtained through measurement of the wavelength of a known spectral feature (emission or absorption line). This may be determined to an accuracy of:

$$\delta v = \frac{Qc}{R} \qquad [1]$$

where Q is the fraction of the width of the spectral resolution element, $\delta\lambda$, to which the centroid can be determined. This is limited by either the SNR of the observation of the line via $Q \approx 1/\mathrm{SNR}$, or by the stability of the spectrograph via:

$$Q = \frac{\Delta x}{\chi}\left(\frac{\mathrm{d}\chi}{\mathrm{d}x}\right) \qquad [2]$$

where χ is the angular width of the slit measured on the sky and $\delta\chi/\mathrm{d}x$ is the image scale at the detector. Δx is the amount of uncorrectable flexure in units of the detector pixel size. This is the part of the motion of the centroid of the line which is not susceptible to modeling with the aid of frequent wavelength calibration exposures. Typically, $\Delta x \approx 0.2$ pixels is achievable (roughly 3 μm at the detector or 10 μm at the slit for an 8 m telescope) implying $Q \approx 1/30$ for a typical $(\mathrm{d}\chi/\mathrm{d}x) = 0.1–0.2''$/pixel and $\chi = 0.5–1''$. Table 3 gives typical requirement for R and the value of Q required to achieve it, given the spectral resolution obtainable for different types of current spectrograph. Extrasolar planet studies require very high stability.

Some spectral observations require very accurate measurements of the shape of the spectral features to infer the kinematics of ensembles of objects which are individually not resolvable (e.g., stars in distant galaxies, gas in blackhole accretion disks). Other applications require accurate measurements of the relative flux of various spectral features within the same spectrum (e.g., to reveal abundances of elements or different types of star).

Optical Principles

Imaging

The simplest type of imager consists of a detector, comprising a 2D array of light-sensitive pixels (Table 4) placed at a telescope focal surface without any intervening optics except for the filters required for photometry. Although maximizing throughput, this arrangement removes flexibility in changing the image scale since the physical size of pixels (typically 15–25 μm) is fixed. Thus the Gemini telescopes

Table 3 Typical requirements for spectral resolution and stability

Target	*δv or required data product*	*Required λ/δλ*	*Spectral regime*	*Available R*	*Stability required = 1/Q*	*Simultaneous wavelength range*
Cosmological parameters	1000 km/s (line centroid)	300	*Low dispersion*	300–3000	1	≳1 octave
Galaxies						
Intra-cluster	100 km/s (line centroid)	3000			1–10	100 nm
Element abundance	Ratio of line fluxes	1000–3000				1 octave
Stellar orbits	10 km/s (line shapes)	3×10^4	*Medium*	$1–3 \times 10^4$	(1)	10 nm
Individual stars	1 km/s (line centroid)	3×10^5			10–30	10 nm
Extrasolar planets	<1 m/s (line centroid)	$>3 \times 10^8$	*High*	$\sim 10^5$	>3000	<1 nm

Table 4 Detectors in current use for astronomy

Wavelength (μm)	*Material/ type*	*Pixel size (μm)*	*Format*		*Read noise (e-/pix)*	*Dark current (e-/pix/hr)*	*QE range (%)*
			Single	*Mosaic*			
		Typical	*State-of-the-art*				
0.3–1.0	CCD	10–20	4k × 4k	12k × 8k	2	Negligible	10–90[a]
	CMOS	5–10	4k × 4k		5–30		
1–2.5 +	HgCdTe	15–20					
0.6–5	InSb	20–30	2k × 2k	4k × 4k	10	0.01–0.1	70–90

[a]Quantum efficiency varies strongly over quoted wavelength range.

($D_T = 8$ m) provides an uncorrected image scale of 25–40 mas/pixel at its sole F/16 focus. Although well-suited to high-resolution observations with AO, this is inappropriate for most observations of faint galaxies where 100–200 mas/pixel is required. This arrangement also provides limited facilities for defining the wavelength range of the observations since there is no intermediate focal surface at which to place narrowband filters, whose performance is sensitive to field angle and are, therefore, best placed at an image conjugate. There is also no image of the telescope pupil at which to place a cold stop to reject the external and instrumental thermal radiation which is the dominant noise source at $\lambda > 1.5$ μm.

The alternative is to use a focal-reducer consisting of a collimator and camera. The auxiliary components (filters, cold-stops, Fabry–Perot etalons, etc.) then have an additional image and pupil conjugate available to them and the image scale may be chosen by varying the ratio of the camera and collimator focal lengths. The extra complexity of optics inevitably reduces throughput but, through the use of modern optical materials, coatings such as Sol–Gel and the avoidance of internally obstructed reflective optics, losses may be restricted to 10–20% except at wavelength shorter than ~0.35 μm. An imaging focal reducer is a special case of the generic spectrograph described in the next section.

Spectroscopy

Basic principles

Astronomical spectrographs generally employ plane diffraction gratings. The interference between adjacent ray paths as they propagate through the medium is illustrated in Figure 1. From this, we obtain the grating equation:

$$m\rho\lambda = n_1 \sin\alpha + n_2 \sin\beta \equiv G \qquad [3]$$

where α and β are the angles of incidence and diffraction respectively, $\rho = 1/a$ is the ruling density, λ is the wavelength, and m is the spectral order. The refractive indices in which the incident and diffracted rays propagate are n_1 and n_2 respectively.

The layout of a generic spectrograph employing a plane diffraction grating in a focal reducer arrangement is shown in Figure 2. The spectrograph re-images the slit onto the detector via a collimator and camera. The disperser is placed in the collimated beam close to a conjugate of the telescope pupil.

A field lens placed near the focus is often incorporated in the collimator for this purpose.

The geometrical factor, G, defined in the equation is constrained by the basic angle of the spectrograph:

$$\psi = \alpha - \beta \qquad [4]$$

which is the (normally fixed) angle between the optical axes of the collimator and camera. $\psi = n\pi$, corresponds to the Littrow configuration (for integer n).

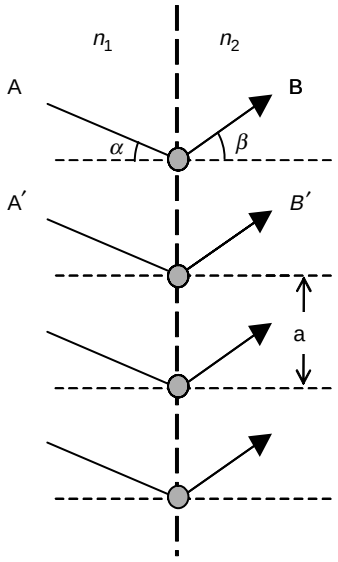


Figure 1 Derivation of the grating equation by consideration of the optical path difference between neighboring ray paths AB and A′B′.

Considering the case of a diffraction grating in vacuo, differentiation of the grating equation with respect to the diffracted angle yields the angular dispersion:

$$\frac{d\lambda}{d\beta} = \frac{\cos\beta}{m\rho} \qquad [5]$$

from which the linear dispersion is obtained by considering the projection of $d\beta$ on the detector, dx:

$$\frac{d\lambda}{dx} = \frac{d\lambda}{d\beta}\frac{d\beta}{dx} = \frac{\cos\beta}{m\rho f_2} \qquad [6]$$

where f_2 is the focal length of the camera (see Figure 2).

In the diffraction-limited case, where the slit is arbitrarily narrow, the resolving power is given simply by the total number of rulings multiplied by the spectral order:

$$R^* = m\rho W \qquad [7]$$

where w is the length of the disperser as defined below. But in astronomy, it is usually determined by the width of the image of the slit, s, projected on the detector, s'. By conservation of Etendue:

$$s' = s\frac{F_2}{F_1} \qquad [8]$$

where F_1 and F_2 are the collimator and camera focal ratios, respectively. The spectral resolution of the spectrograph is simply the width of this expressed in

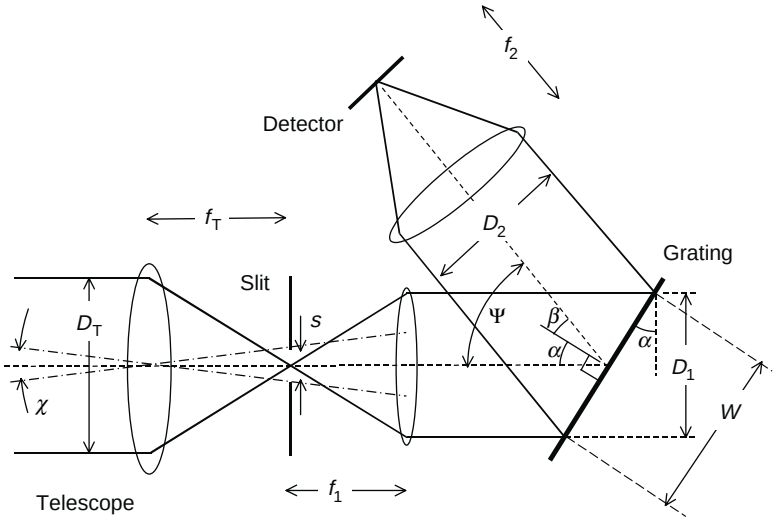


Figure 2 Generic spectrograph employing a plane diffraction grating. Note that according to the sign convention implicit in Figure 1, the diffraction angle, β, is negative in this configuration.

wavelength units:

$$\delta\lambda = \left(\frac{d\lambda}{dx}\right)s' = \frac{\cos\beta}{m\rho f_2}s\frac{F_2}{F_1} = \frac{sD_1\cos\beta}{m\rho D_2 f_1} \quad [9]$$

The length of the intersection between the collimated beam and the plane of the grating (not necessarily the actual physical length of the grating) is

$$W = \frac{D_2}{\cos\beta} = \frac{D_1}{\cos\alpha} \quad [10]$$

Using this to eliminate $\cos\beta$, the spectral resolution becomes

$$\delta\lambda = \frac{s}{m\rho F_1 W} \quad [11]$$

and the resolving power is

$$R \equiv \frac{\lambda}{\delta\lambda} = \frac{m\rho\lambda F_1 W}{s} = G\frac{F_1 W}{s} \quad [12]$$

where G is the quantity defined in eqn [3]. Note that R is now independent of the details of the camera. This expression is useful for the laboratory since it is expressed in terms of the parameters of the experimental apparatus; but, for astronomy, it is more useful to express the resolving power in terms of the angular slit width, χ, and the telescope aperture diameter, D_T, via:

$$s = \chi f_T = \chi F_T D_T = \chi F_1 D_T \quad [13]$$

since the collimator and telescope focal ratios are the same if the spectrograph is directly beam-fed from the telescope. Note that even if the telescope focal surface is re-imaged onto the slit, the expression for the resolving power still holds due to the conservation of Etendue in the re-imaging optics. Thus the resolving power is

$$R = \frac{m\rho\lambda W}{\chi D_T} = \frac{GW}{\chi D_T} \quad [14]$$

Note that the resolving power obtained with a nonzero slit width is always less than the theoretical maximum, $R \leq R_*$, for wavelengths:

$$\lambda < \lambda_* = \chi D_T \quad [15]$$

Thus, long-wavelength applications may approach the theoretical limit; in which case they are said to be diffraction-limited. If so, the resolving power is independent of the slit width, simplifying the interface to different telescopes. For a nondiffraction-limited spectrograph, the resolving power obtained depends on the aperture of the telescope to which it is fitted.

Optical configuration

The non-Littrow configuration shown in Figure 2 introduces anamorphism into the recorded spectra such that a monochromatic image of the slit mask is compressed in the dispersion direction by a factor:

$$A = \frac{D_2}{D_1} = \frac{\cos\beta}{\cos\alpha} \quad [16]$$

which varies, depending on the choice of angles required to place the chosen wavelength on the detector using eqns [3] and [4].

Practical configurations that maximize R, do so by increasing $W = D_1/\cos\alpha$. This implies maximizing $\cos\beta$, since $\psi = \alpha - \beta$, which results in $A > 1$. This is known as the grating-to-camera configuration since the grating normal points more towards the camera than to the collimator. The grating-to-collimator configuration may also be used at the penalty of lower R but does permit a smaller camera. Thus the camera aperture must be oversized in the dispersion direction since $D_2 > D_1$. The Littrow configuration is rarely used with reflective designs when large FOV is required because of internal vignetting and consequent light loss. However, it can be used effectively with transmissive dispersers.

If the disperser is replaced by a plane mirror (or if the disperser is removed in a transmissive system), the system acts as a versatile imager, with $A = 1$. This also facilitates target acquisition which requires that the telescope attitude be adjusted until the targets are aligned with apertures cut in an opaque mask placed at the input of the spectrograph. These apertures (normally slits) are usually oversized in the direction perpendicular to the dispersion to allow the spectrum of the sky to be recorded directly adjacent to that of the target to permit accurate subtraction of the background.

The slit width is chosen by trading off the desire to maximize the amount of light entering the instrument from the target (but not so much as to admit excessive background light) and the need to maximize spectral resolving power, R, by reducing the slit width, χ, according to eqn [14].

Since $|G| \leq 2$, the maximum attainable resolving power, according to eqn [14], is determined chiefly by the telescope aperture and angular slitwidth. The only important parameter of the spectrograph is the illuminated grating length, W. Therefore, maintaining the same resolving power as the telescope aperture increases, requires W to increase in direct proportion. Because of limits placed on the geometry by the need to avoid obstructions, etc., this means that both the collimated beam diameter and the size of the disperser must increase. This is one of the major

problems in devising instruments for ELTs. Current instruments on 8 m telescopes feature collimated beam diameters of $D_1 \approx 150$ mm. For ELTs, the implied sizes are 0.5–2 m, requiring huge optics, accurately co-aligned mosaics of dispersing elements, and enormous support structures to provide the required stability. A simple calculation indicates that the ruling stylus for a classically ruled diffraction grating would have to cover a thousand kilometers without significant wear unless the disperser was made from smaller pieces.

Choice of dispersing element

Astronomical spectrographs mostly use surface-relief diffraction (SL) gratings. These are ruled with groove densities $50 < \rho < 3000\ \text{mm}^{-1}$.

Blazing diffraction gratings. The intensity of the multibeam interference pattern from N rulings of finite width is given by

$$I = \left(\frac{\sin^2 N\phi}{\sin^2 \phi}\right)\left(\frac{\sin^2 \theta}{\theta^2}\right) \qquad [17]$$

where 2ϕ is the phase difference between the center of adjacent rulings and θ is the phase difference between the center and edge of a single ruling. To be useful for astronomy, the peak in this pattern, determined by the second term, must coincide with a useful order, such as $m = 1$, rather than zero order. This can be done by introducing a phase shift into the second term, which represents the diffraction of a single ruling, such that:

$$\theta = \frac{\pi \cos\gamma}{\rho\lambda}(\sin i - \sin r) \qquad [18]$$

where the groove angle, γ, is the angle by which the facets are tilted with respect to the plane of the grating. With the aid of Figure 3, it can be seen that $i = \alpha - \gamma$ and $r = \gamma - \beta$ so that the maximum efficiency, which occurs when $\theta = 0$, corresponds to $i = r$, which is equivalent to simple reflection from the facets, and $\alpha + \beta = 2\gamma$.

Thus, through this process of blazing, the grating eqn [3], at the blaze wavelength, λ_B, becomes:

$$\rho m \lambda_B = 2\sin\gamma\cos\frac{\psi}{2} \qquad [19]$$

where ψ, is given by eqn [4] using the identity: $\sin x + \sin y \equiv 2\sin\left(\frac{x+y}{2}\right)\cos\left(\frac{x+y}{2}\right)$. Thus, the resolving power at blaze is obtained from eqns [10], [14], and [19] as:

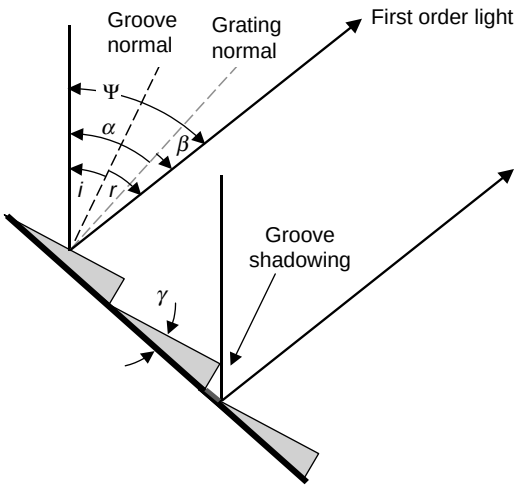


Figure 3 Blazing of ruled diffraction grating.

$$R_B = \frac{2D_1}{\chi D_T}\frac{\sin\gamma\cos(\psi/2)}{\cos(\gamma + \psi/2)} \qquad [20]$$

Use of grisms. An alternative configuration uses transmission gratings. The advantage here is that the collimator and camera can share the same optical axis in a straight-through configuration with unit anamorphism, $A = 1$, which does not require oversizing of the camera. The same phase-shift can be applied as before by making the facets into prisms with the required groove angle. However, this would only allow zero-order to propagate undeviated into the camera so an additional prism is required with vertex angle $\phi = \gamma$ to allow first-order light to pass undeviated into the camera. This composite of blazed transmission grating and prism is called a grism (Figure 4).

The grating equation (eqn [3]) is modified for a grism in the blaze condition to $m\rho\lambda_B = (n-1)\sin\phi$ since $\alpha = -\beta = \phi$ and $n_1 \equiv n, n_2 = 1$. Noting that $\psi = 0$, the resolving power at blaze is then:

$$R_B = \frac{D_1}{\chi D_T}(n-1)\tan\phi \qquad [21]$$

where n is the refractive index of the prism and grating material. Due to problems such as groove shadowing (illustrated for reflection gratings in Figure 3), these lack efficiency when $\phi \gtrsim 30°$ restricting their use to low-resolution spectroscopy.

Use of volume phase holographic gratings. A newer alternative to surface-relief (SR) dispersers are volume phase holographic (VPH) gratings in which

the interference condition is provided by a variation of refractive index of a material such as dichromated gelatine, n_g, which depends harmonically on position inside the grating material as:

$$n_g(x,z) = \bar{n}_g + \Delta n_g \cos[2\pi\rho_g(x\sin\gamma + z\cos\gamma)] \qquad [22]$$

where ρ_g is the density of the lines of constant n_g and $\bar{n}_g$ and Δn_g are the mean refractive index and its modulation amplitude, respectively (Figure 5). These are described by the same grating (eqn [3]), as before leading to an identical expression for the resolving power at blaze, where $m\rho\lambda_B = 2\sin\alpha$, to that of a blazed plane diffraction grating in the Littrow configuration (eqn [20]):

$$R_B = \frac{D_1}{\chi D_T} 2\tan\alpha \qquad [23]$$

Unlike SR gratings, the blaze condition of VPH gratings may be varied by changing the basic angle of the spectrograph, $\psi = 2(\gamma+\alpha)$, although this is often mechanically inconvenient. They may also be sandwiched between prisms to provide the necessary incident and diffraction angles while retaining the advantage of a straight-through configuration, to form an analogue of a grism, sometimes termed a vrism. Unlike grisms, these are not constrained by groove shadowing and so can be efficient at relatively high dispersion, subject to the constraint imposed by the size of the device.

Use of echelle gratings. Another option to increase the resolving power is to use a coarse grating in high order via the use of an echelle format. The resolving power in the near-Littrow configuration in which it is usually operated, is given by eqn [20] with $\psi = 0$.

$$R_B = \frac{D_1}{\chi D_T} 2\tan\gamma \qquad [24]$$

where γ is the groove angle.

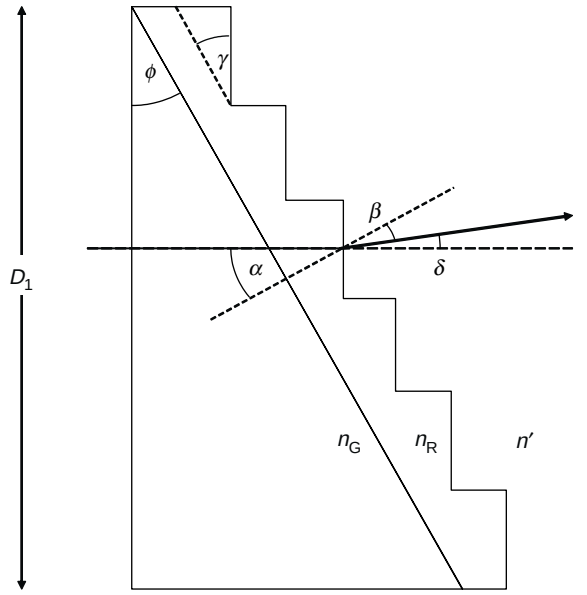


Figure 4 Typical configuration of a grism. The prism vertex and groove angles are chosen for maximum efficiency on the assumption that refractive indices, $n_G = n_R = n$ and $n' = 1$. The blaze condition occurs when $\delta = 0$. In this configuration, $\phi = \gamma = \alpha$.

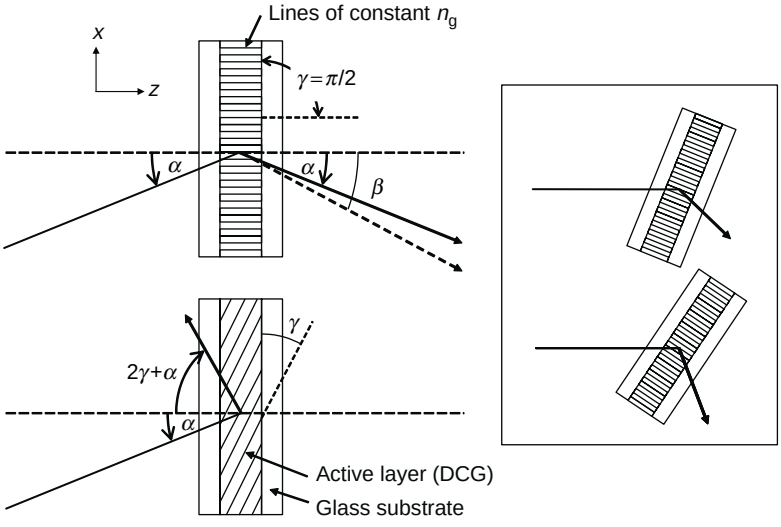


Figure 5 Various configurations of volume phase holographic gratings. The blaze condition is when $\beta = \alpha$. The inset shows how the blaze condition may be altered by changing the basic angle of the spectrograph.

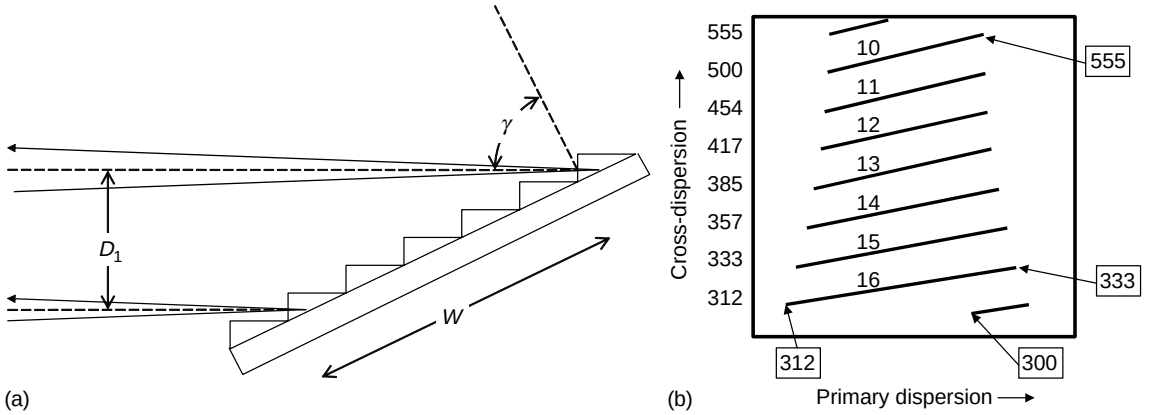


Figure 6 (a) Basic layout of an Echelle grating used in a near-Littrow configuration. (b) Illustrative cross-dispersed spectrogram showing a simplified layout on the detector. The numbers 10–16 label the different spectral orders. The numbers labeling the vertical axis are the wavelength (nm) at the lowest end of each complete order. Other wavelengths are labeled for clarity. For simplicity the orders are shown evenly spaced in cross-dispersion.

In the configuration shown in Figure 6, R is maximized by increasing W. This also means that $G = m\rho\lambda$ is large. In order to avoid excessively high ruling densities, the grating may be used in high orders. However, this presents the problem of order overlap since the wavelength in order m occurs in the same place as wavelength $\lambda_n = (m/n)\lambda_m$ in order n. Limiting the observation to a single order via the use of order-sorting filters is one solution while another is to use cross-dispersion via a chain of prisms or a grating with dispersion perpendicular to that of the echelle to separate the orders out. This option severely curtails the use of such spectral formats in multi-object spectroscopy.

Increasing resolving power using immersed gratings. All the methods of using diffraction gratings discussed so far are subject to various geometric constraints which ultimately limit the maximum obtainable resolving power by placing limits on W (e.g., in eqn [14]). These may be partly overcome by optically coupling the dispersers to immersing prisms. This can lead to a doubling of the resolving power in some cases.

Use of prisms. The final method of dispersion is to use a prism. This has the advantage that, since it does not work by multibeam interference, the light is not split into separate orders which removes the problem of order-overlap and improves efficiency. However, it does not produce the high spectral resolution required for some applications and the dependency of dispersion on λ is often markedly nonlinear.

From a consideration of Fermat's principle, it can be shown that the resolving power of a prism is:

$$R = \frac{\lambda t}{\chi D_{\mathrm{T}}}\left(\frac{\mathrm{d}n}{\mathrm{d}\lambda}\right) \qquad [25]$$

where t is the baselength of the prism. Using modern materials, $R \lesssim 300$ may be obtained. The problem of the nonlinearity in dispersion can be alleviated by the use of composite materials with opposite signs of dispersion.

Multiobject spectroscopy (MOS)

As discussed above, there is great value in increasing the multiplex gain. This can be done in two ways:

(i) *Increasing the number of apertures in the slit mask (Multislits)*. This requires that each slit is carefully cut in the mask because, although the telescope attitude (translation and rotation) may be adjusted to suit one aperture, errors in the relative positions of slits cannot be compensated since the image scale is fixed. Each aperture produces a spectrum on the detector and the mask designer must ensure that the targets and their matching slits are chosen to avoid overlaps between spectra and orders. This has the effect of limiting the surface density of targets which can be addressed in one observation. Passband filters can be used to limit the wavelength range and reduce the overlap problem. This option requires superior optics able to accommodate both a wide FOV and large disperser.

(ii) *Using optical fibers (Multifibers)*. A number of fibers may be deployed at the telescope focus to

direct light from the various targets to a remote spectrograph whose input focus consists of one or more continuous pseudoslits. These are made up of the fiber outputs arranged in a line. The continuous nature of the slit means that spectral overlap is avoided without restricting the surface density of addressable targets; although there is a limit imposed by the distance of closest approach of the fibers. The method of deploying the fiber inputs may be a plugplate consisting of pre-cut holes into which encapsulated fibers are manually plugged, or a pick-and-place robot which serially positions the fiber inputs at the correct location on a magnetized field plate. The latter is highly versatile but mechanically complex with a significant configuration time that may erode the actual on-target integration time.

The two systems have contrasting capabilities (Figure 7). The multislit approach provides generally better SNR since the light feeds into the spectrograph directly, but is compromised in terms of addressable surface density of targets, by the need to avoid spectral overlap and in FOV by the difficulty of the wide-field optics. A current example of this type of instrument, GMOS, is shown in Figure 8 and described in Table 5. The multifiber approach is limited in SNR by modal noise in the fibers and attendant calibration uncertainties and the lack of contiguous estimates of the sky background. However, it is easier to adapt to large fields since the spectrograph field can be much smaller than the field over which the fiber inputs are distributed. In summary, multislit systems are best for narrow-but-deep surveys while the multifiber systems excels at wide-but-shallow surveys. Fiber-fed instruments may further prove their worth in ELTs where technical problems may require that these bulky instruments are mounted off the telescope (see below for further discussion).

Of paramount importance in MOS is the quality of the background subtraction as discussed above. Traditionally, this requires slits which sample the sky background directly adjacent to the object. An alternative is nod-and-shuffle (va-et-vient) in which nearby blank regions are observed alternately with the main field using the same slit mask by moving the telescope (the nod). In the case of CCDs, the photo-generated charge from the interleaved exposures is stored temporarily on the detector by moving it to an unilluminated portion of the device (the shuffle). After many repeats on a timescale less than that of variations in the sky background (a few minutes), the accumulated charge is read out, incurring the read-noise penalty only once. Although requiring an effective doubling of the exposure time and an

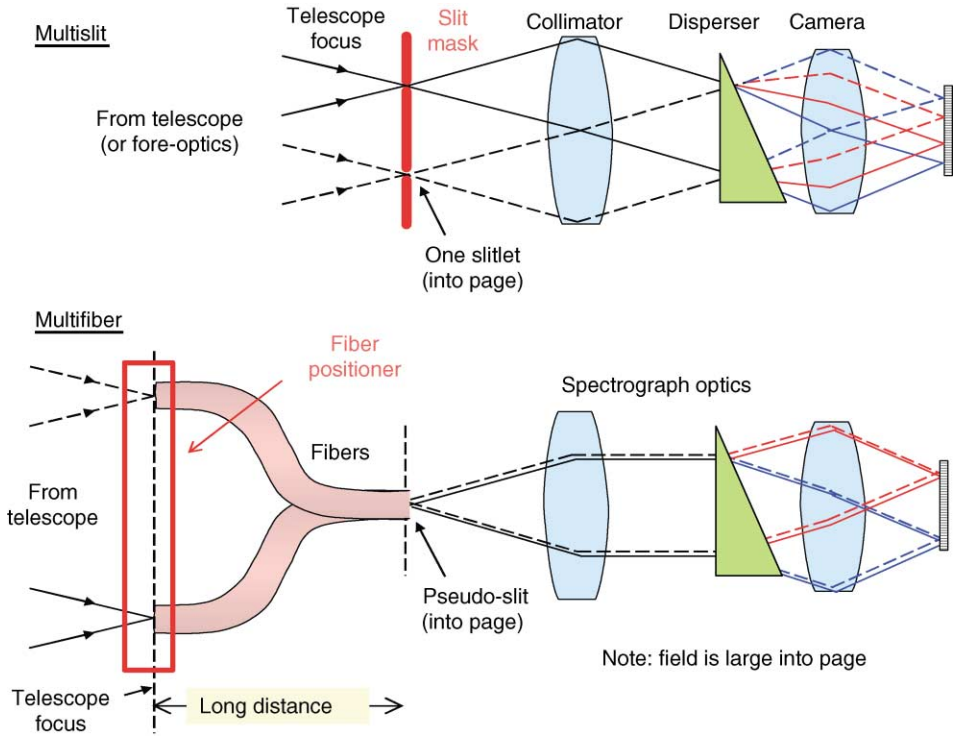


Figure 7 Illustration of the difference between the multislit and multifiber approaches to multi-object spectroscopy.

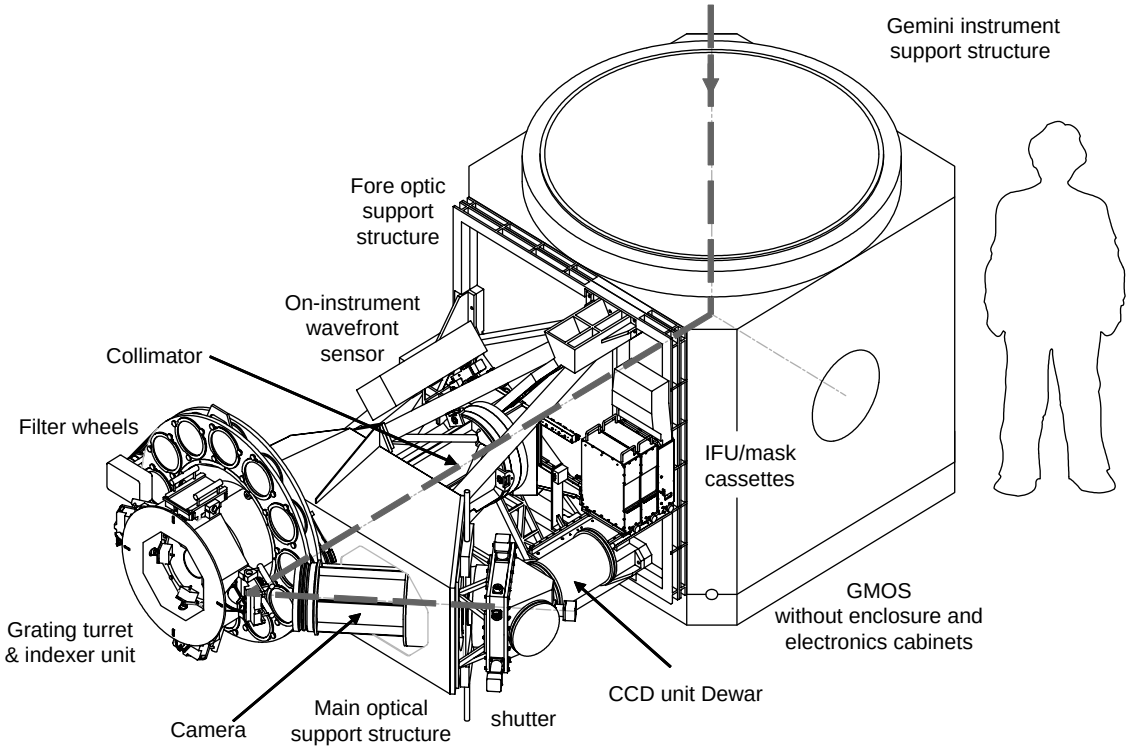


Figure 8 The Gemini Multiobject Spectrograph (GMOS). One of two built for the two Gemini 8-m telescopes by a UK–Canadian consortium. It includes an integral field capability provided by a fiber-lenslet module built by the University of Durham. The optical configuration is the same as shown in **Figure 2**. It is shown mounted on an instrument support unit which includes a mirror to direct light from the telescope into different instruments. The light path is shown by the red dashed line. The slit masks are cut by a Nd-YAG laser in 3-ply carbon fiber sheets. See **Table 5** for the specification.

Table 5 Main characteristics of the Gemini multiobject spectrographs

Image scale	72 mas/pixel	FOV	$5.5' \times 5.5'$
Detector	CCD with 13.5 μm pixels Format: $3 \times (4608 \times 2048)$	Slit/mask configuration	$\leq$ few $\times$ 100 slits of width $\geq 0.2''$
Wavelength range	Total: 0.4–1.0 μm Simultaneous: $\leq$ 1 octave	Integral field unit	$1500 \times 0.2''$ samples in 50 sq.$''$
Spectral resolving power	$R \leq 5000$ with $0.5''$ slit	Dispersion options	3 gratings + mirror for imaging

increase in the size of the detector, this technique allows the length of the slits to be reduced since no contiguous sky sample is required, thereby greatly increasing the attainable multiplex gain if the number density of potential targets is sufficiently high.

Integral field spectroscopy (IFS)

IFS provides a spectrum of each spatial sample simultaneously within a contiguous FOV. Other approaches provide the same combination of imaging and spectroscopy but require a series of nonsimultaneous observations. Examples include imaging through a number of narrow passband filters with different effective wavelength and spectroscopy with a single slit which is stepped across the object. Other such techniques are Fabry–Perot, Fourier-transform and Hadamard spectroscopy. Since the temporal variation of the sky background is a major limitation in astronomy of faint objects, IFS is the preferred technique for terrestrial observations of faint objects, but nonsimultaneous techniques are preferable in certain niche areas, and relatively more important in space where the sky background is reduced.

The main techniques (**Figure 9**) are as follows:

(i) *Lenslets*. The field is subdivided by placing an array of lenslets at the telescope focus (or its conjugate following re-imaging fore-optics) to form a corresponding array of micropupils. These are re-imaged on the detector by a focal

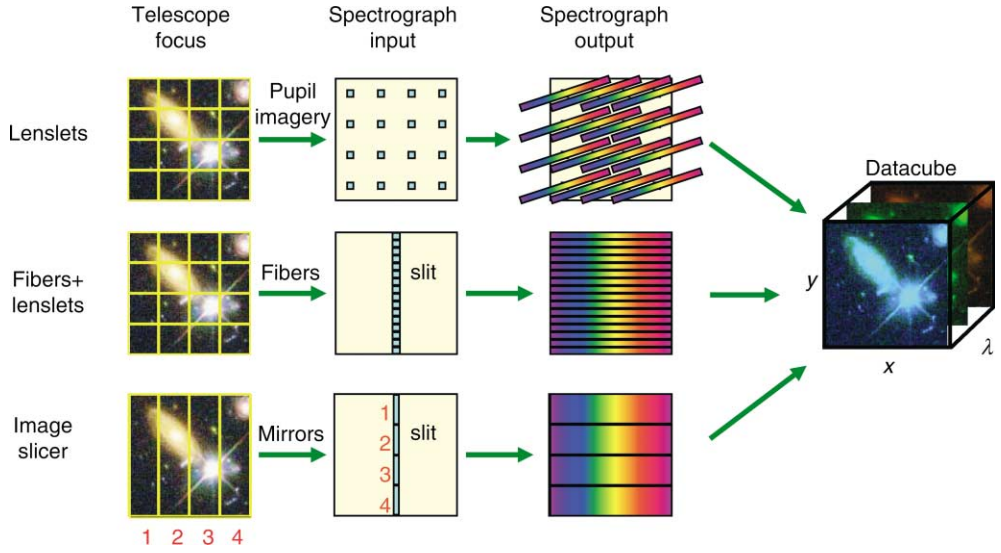


Figure 9 Main techniques of integral field spectroscopy.

reducer using a conventional disperser. The spectra are, therefore, arranged in the same pattern as that of the lenslet array. Spectral overlap is reduced by angling the dispersion direction away from the symmetry axes of the lenslet array and by the fact that the pupil images are smaller than the aperture of the corresponding lenslet.

(ii) *Fibers + lenslets*. The field is subdivided as in (i) but the pupil images are relayed to a remote spectrograph using optical fibers which reformat the field into linear pseudoslits. This avoids the problem of overlap, allowing a greater length of spectrum than in (i), but is more complex since the arrays of fibers and lenslets must be precisely co-aligned and are subject to modal noise in the fibers.

(iii) *Image slicer*. The field is subdivided in only one dimension by a stack of thin slicing mirrors placed at a conjugate of the telescope focus (Figure 10). Each mirror is angled so as to direct light to its own pupil mirror which re-images the field to form part of a linear pseudoslit. Thus the slices into which the image is divided are rearranged end-to-end to form a continuous slit at the spectrograph input. Unlike the other techniques, this retains spatial information along the length of each slice. This is dispersed by conventional means via a focal-reducer. An additional optic is often required at the image of each slice at the slit to reimage the micropupil images produced by the slicing mirrors onto a common pupil inside the spectrograph. In principle this is the most efficient of the three methods since the one-dimensional slicing produces fewer diffraction losses in the spectrograph than the two-dimensional division used by the others, and minimizes the fraction of the detector surface which must be left unilluminated in order to avoid cross-talk between noncontiguous parts of the field. However the micromirrors are difficult to make since they require diamond-turning or grinding in metal or glass with a very fine surface finish (typically with RMS ~ 1 nm for the optical and ~ 10 nm in the infrared). This sort of system is, however, well-matched to cold temperatures, since the optical surfaces and mounts may be fabricated from the same material (e.g., Al) or from materials with similar thermal properties.

The data are processed into a datacube whose dimensions are given by the two spatial coordinates, x, y, plus wavelength. The datacube may be sliced up in ways analogous to tomography to understand the physical process operating in the object.

Spectropolarimetry and Polarimetry

Spectropolarimetry and polarimetry are analogous to spectroscopy and imaging, where the polarization state of the light is measured instead of the total intensity. Like IFS, spectropolarimetry is a photon-starved (i.e., limited in SNR by photon noise) area which benefits from the large aperture of current and projected telescopes.

The Stokes parameters of interest are I, Q, and U. V is generally very small for astronomical objects.

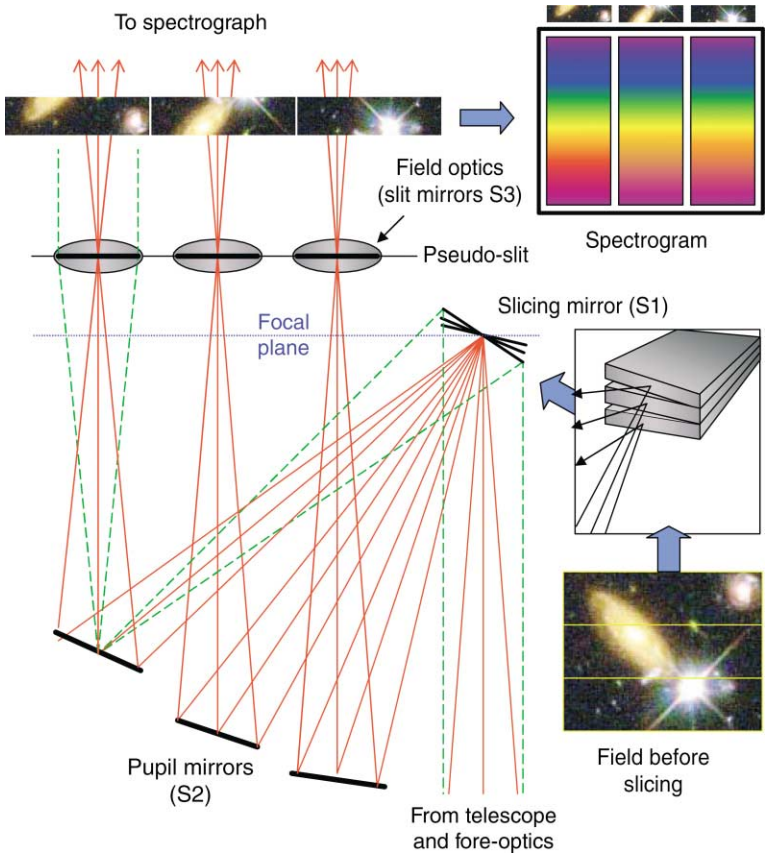


Figure 10 Principle of the advanced image slicer. Only three slices are shown here. Real devices have many more, e.g., the IFU for the *Gemini Near-Infrared Spectrograph* has 21 to provide a $5'' \times 3''$ field with $0.15'' \times 0.15''$ sampling. Reproduced from Content R (1997) A new design for integral field spectroscopy with 8 m telescopes. *Proceedings SPIE* 2871: 1295–1305.

From these, the degree and azimuthal angle of linear polarization are obtained as:

$$\pi_{\mathrm{L}} = \left(\frac{Q^2 + U^2}{I^2}\right) \qquad \psi = \frac{1}{2}\arctan\left(\frac{U}{Q}\right) \qquad [26]$$

The Stokes parameters may be estimated using a modified spectrograph with a rotatable achromatic half-wave retarder, characterized by the angle of rotation about the optical axis, θ, placed before the instrument and a polarizing beamsplitter which separates the incident beam into orthogonal polarization states (Figure 11). The two states are recorded simultaneously on different regions of the detector. The separation is achieved usually through the angular divergence produced by a Wollaston prism placed before the disperser or through the linear offset produced by a calcite block placed before the slit. The intensity as a function of the waveplate rotation angle, $S(\theta)$, is recorded for each wavelength and position in the spectrum or image. The Stokes parameters for each sample are then given by

$$\begin{aligned}
\frac{Q}{I} &= \frac{S(0) - S(\pi/2)}{S(0) + S(\pi/2)} \\
&= \left\langle \frac{A_1 - \omega_{AB}B_1}{A_1 + \omega_{AB}B_1}, \frac{A_2 - \omega_{AB}B_2}{A_2 + \omega_{AB}B_2} \right\rangle \\
\frac{U}{I} &= \frac{S(\pi/4) - S(3\pi/4)}{S(\pi/4) + S(3\pi/4)} \\
&= \left\langle \frac{C_1 - \omega_{CD}D_1}{C_1 + \omega_{CD}D_1}, \frac{C_2 - \omega_{CD}D_2}{C_2 + \omega_{CD}D_2} \right\rangle \\
I &= S(0) + S(\pi/2) \\
&= \left\langle \frac{A_1 + \omega_{AB}B_1}{\eta_{A1}}, \frac{A_2 + \omega_{AB}B_2}{\eta_{A2}} \right\rangle \text{ etc.}
\end{aligned} \qquad [27]$$

where A and B are the signals (functions of position and/or λ) recorded with the waveplate at $\theta = 0$ and $\theta = \pi/2$ and the subscripts 1 and 2 refer to the two polarization states produced by the beamsplitter.

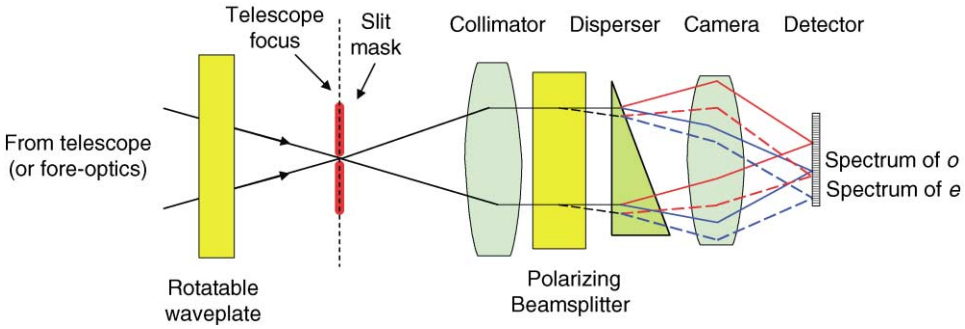


Figure 11 Schematic layout of spectropolarimeter using a polarizing beamsplitter in the collimated beam (e.g., a Wollaston prism producing *e*- and *o*-states) to produce separate spectra for orthogonal polarization states on the detector. Alternative configurations in which the two polarizations have a linear offset (e.g., a calcite block) may be used in which case the beamsplitter is placed before the slit.

Likewise, C and D are a pair of observations taken with $\theta = \pi/4$ and $\theta = 3\pi/4$. The angled brackets indicate a suitable method of averaging the quantities inside. The factors ω_{AB} and ω_{CD} are estimated as

$$\omega_{AB} = \sqrt{\frac{A_1 A_2}{B_1 B_2}} \qquad \omega_{CD} = \sqrt{\frac{C_1 C_2}{D_1 D_2}} \qquad [28]$$

and η_{A1} is a single calibration factor for polarization state 1 of observation A, etc. One problem is instrumental polarization caused by oblique reflections from telescope mirrors upstream of the polarizing waveplate, such as the fold mirror external to GMOS (see Figure 8). This must be carefully accounted for by taking observations of stars with accurately known polarization. The dispersing element is also likely to be very strongly dependent on the polarization state of the incident light with widely different efficiency as a function of wavelength. Although this effect is cancelled out using the procedure described above, it introduces a heavy toll in terms of SNR reduction.

Technology Issues

Use of Optical Fibers in Astronomy

As discussed above, optical fibers are commonly used in astronomy for coupling spectrographs to wide, sparsely sampled FOVs in the case of MOS, or small contiguously sampled FOVs in the case of IFS. For both applications, the most important characteristics are:

(i) *Throughput*. Near-perfect transmission is required for wavelength intervals of ~1 octave within the overall range of 0.3–5 μm. For MOS, a prime focus fiber feed to an off-telescope spectrograph implies fiber runs of several tens of meters for 8 m telescopes, increasing proportionally for ELTs. For IFS, runs of only ~1 m are required for self-contained par-focal IFUs. For $\lambda < 2.5$ μm, fused silica is a suitable material, although it is not currently possible to go below ~0.35 μm, depending on fiber length. Newer techniques and materials may improve the situation. For $\lambda > 2.5$ μm, alternative materials are required, such as ZrF4 and, at longer wavelengths, chalcogenide glasses, but these are much more expensive and difficult to use than fused silica.

(ii) *Conservation of Etendue*. Fibers suffer from focal ratio degradation, a form of modal diffusion, which results in a speeding up of the output beam with respect to the input. Viewed as an increase in entropy, it results in a loss of information. In practice, this implies either reduced throughput, as the fiber output is vignetted by fixed-size spectrograph optics, or a loss in spectral resolution if the spectrograph is oversized to account for the faster beam from the fibers. This effect may be severe at the slow focal ratios pr oduced by many telescopes ($F > 10$) but is relatively small for fast beams ($F < 5$). Thus, its effect may be mitigated by speeding up the input beam by attaching lenslets directly to the fibers, either individually (for MOS) or in a close-packed array (for IFS). If the spectrograph operates in both fiber- and beam-fed modes (e.g., Figure 8), it is also necessary to use lenslets at the slit to convert the beam back to the original speed.

(iii) *Efficient coupling to the telescope*. This requires that the fibers are relatively thick to match the physical size of the sampling at the telescope focus which is typically $\chi_s = 0.1''$ for IFS and $\chi_s = 2''$ for MOS. By conservation of Etendue, the physical size of the fiber aperture is

$$d_f = \chi_s D_T F_f \qquad [29]$$

where F_f is the focal ratio of the light entering the fiber and D_T is the diameter of the telescope aperture. Using the above values with $F_f = 4$, as required to maintain Etendue, implies $15 < d_f < 300$ μm for an 8 m or $60 < d_f < 1200$ μm for a 30 m telescope. Except at the lower end, this implies the use of multimode (step-index) fibers. However, the need in IFS to oversize the fibers to account for manufacturing errors and current requirements of $\chi_s \geq 0.2''$ to produce sufficient SNR in a single spatial sample gives a practical limit of $d_f > 50$ μm which is in the multimode regime.

Recently, the use of photonic crystal fibers in astronomy has been investigated. These, together with other improvements in material technology may improve the performance of optical fibers, which will be of special relevance to ELTs where, for example, there will be a need to couple very fast beams into fibers.

Cooling for the Near-Infrared

As shown in Figure 12, it is necessary to cool instruments for use in the infrared. This is to prevent thermal emission from the instrument becoming the dominant noise source. The minimum requirement is to place a cold stop at a conjugate of the telescope pupil. However, much of the rest of the instrument requires cooling because of the nonzero emissivity of the optics. The telescope also requires careful design. For example, the Gemini telescopes are optimized for the NIR by undersizing M2 so that the only stray light entering the final focal surface is from the cold night sky.

Irrespective of the thermal background and wavelength, all detectors (see Table 4), require cooling to reduce internally generated noise to acceptable levels. For the CCDs used in the optical, the cryogen is liquid nitrogen, but for the infrared, cooling with liquid helium is required.

Structure

As discussed above, the necessity for instruments to scale in size with the telescope aperture cannot be achieved by simply devising a rigid structure, since the required materials are generally not affordable (instruments on 8 m telescopes cost roughly €/$ 5–15 M, including labor). Furthermore, the stability also depends critically on the mounting of the individual optical elements, where the choice of material is constrained by the need for compliance with the optical materials. The solution adopted for GMOS was to design the collimator and camera mounting so that variations in the gravity vector induce a translation in each component orthogonal to the optical axis without inducing tilts or defocus. Therefore, the only effect of instability is a translation of the image of the slit on the detector. Since the slit area is rigidly attached to the telescope interface, the image of a celestial object does not move with respect

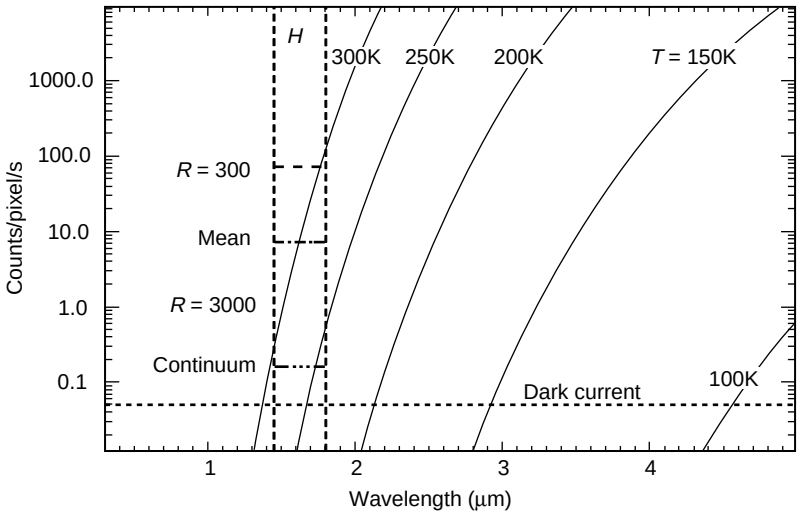


Figure 12 Illustration of the need for cooling in NIR instruments. This is modeled for the case of a spectrograph on an 8 m telescope with $0.2'' \times 0.2''$ sampling with both throughput and total emissivity of 50%. The curves indicate the cumulative thermal background for wavelengths shortward of that shown on the scale. For comparison, a typical dark current is shown for NIR detectors (see Table 4). Also shown is the background from the night sky for the case of the *H*-band, for two values of spectral resolution, *R*. The mean sky background level is shown for both values of *R*. For $R = 3000$, the continuum level found between the strong, narrow OH-lines which make up most of the signal is also shown. To observe in the *H*-band, cooling to $-40\,^{\circ}$C is required for high spectral resolution, but is unnecessary at low resolution. At longer wavelengths, fully cryogenic cooling with liquid nitrogen is generally required.

to the slit providing that the telescope correctly tracks the target. The motion of the image of the slit on the detector is then corrected by moving the detector orthogonal to the optical axis to compensate. Taking account of nonrepeatable motion in the stucture and the detector mounting, it proved possible to attain the desired stability in open-loop by measuring the flexure experienced with the control system turned off, making a look-up table to predict the required detector position for each attitude setting (elevation and attitude) and applying the correction by counting pulses sent to the stepper motors which control the detector translation.

Some instruments require much greater stability than GMOS does. The flexure-control outlined above may be augmented by operation in closed-loop so that nonrepeatable movements may be accounted for. However, this ideally requires a reference light source which illuminates the optics in the same way as the light from the astronomical target and which is recorded by the science detector without adversely impacting the observation. Alternatively, partial feedback may be supplied by repeated metrology of key optical components such as fold-mirrors.

One strategy to measure radial velocities to very small uncertainities (a few meters per second) is to introduce a material (e.g., iodine) into the optical path, which is present throughout the observations. This absorbs light from the observed object's continuum at one very well-defined wavelength. Thus, instrument instability can be removed by measuring the centroid of the desired spectral line in the object relative to the fixed reference produced by the absorption cell. However, flexure must still be carefully controlled to avoid blurring the line during the course of a single exposure.

Mounting of instruments via a fiber-feed remotely from the telescope, at a location where they are not subjected to variations in the gravity vector, is a solution for applications where great stability is required. But, even here, care must be taken with the fore-optics and pickoff system attached to the telescope, and to account for modal noise induced by changes in the fiber configuration as the telescope tracks.

Finally, the instrument structure has other functions. For cooled instruments, a cryostat is required inside which most of the optical components are mounted. For uncooled instruments, careful control of temperature is needed to avoid instrument motion due to thermal expansion. This requires the use of an enclosure which not only blocks out extraneous light, but provides a thermal buffer against changes in the ambient temperature during observations and reduces thermal gradients by forced circulation of air.

See also

Diffraction: Diffraction Gratings. **Fiber and Guided Wave Optics**: Fabrication of Optical Fiber. **Imaging**: Adaptive Optics. **Instrumentation**: Telescopes; Spectrometers. **Spectroscopy**: Fourier Transform Spectroscopy; Hadamard Spectroscopy and Imaging.

Further Reading

Allington-Smith JR, Murray G, Content C, *et al.* (2002) Integral field spectroscopy with the Gemini Multiobject Spectrograph. *Publications of the Astronomical Society of the Pacific* 114: 892.

Bjarklev A, Broeng J and Sanchez Bjarklev AS (2003) *Photonic Crystal Fibers*. Kluwer.

Bowen IS (1938) The Image-Slicer: a device for reducing loss of light at slit of stellar spectrograph. *Astrophysical Journal* 88: 113.

Carrasco E and Parry I (1994) A method for determining the focal ratio degradation of optical fibres for astronomy. *Monthly Notices of the Royal Astronomical Society* 271: 1.

Courtes G (1982) *Instrumentation for Astronomy with Large Optical Telescope, IAU Colloq, 67*. In: Humphries CM (ed.) Astrophysics and Space Science Library, vol. 92, pp. 123. Dordrecht: Reidel.

Glazebrook K and Bland-Hawthorn J (2001) Microslit nod-shuffle spectrocopy: a technique for achieving very high densities of spectra. *Publications of the Astronomical Society of the Pacific* 113: 197.

Lee D, Haynes R, Ren D and Allington-Smith JR (2001) Characterisation of microlens arrays for astronomical spectroscopy. *PASP* 113: 1406.

McLean I (1997) *Electronic Imaging in Astronomy*. Chichester, UK: John Wiley and Sons Ltd.

Palmer C (2000) *Diffraction Grating Handbook*. Richardson Grating Laboratory, 4th edn. Rochester, NY.

Ridgway SH and Brault JW (1984) Astronomical Fourier transform spectroscopy revisited. *Ann. Rev. Astron. Astrophys.* 22.

Saha SK (2002) Modern optical astronomy: technology and impact of interferometry. *Review of Modern Physics* 74: 551.

Schroeder D (2000) *Astronomical Optics*, 2nd edn. San Diego, CA: Academic Press.

van Breugel W and Bland-Hawthorn J (eds) *Imaging the Universe in Three Dimensions*, ASP Conference Series, vol. 195.

Weitzel L, Krabbe A, Kroker H, *et al.* (1996) 3D: The next generation near-infrared imaging spectrometer. *Astron. & Astrophys. Suppl.* 119: 531.

Ellipsometry

J N Hilfiker, J A Woollam & Co., Inc., Lincoln, NE, USA

J A Woollam, University of Nebraska, Lincoln, NE, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Ellipsometry measures a change in polarization as light reflects from or transmits through a material structure. The polarization-change is represented as an amplitude ratio, Ψ, and a phase difference, Δ. The measured response is dependent on optical properties and thickness of each material. Thus, ellipsometry is primarily used to determine film thickness and optical constants. However, it is also applied to the characterization of composition, crystallinity, roughness, doping concentration, and other material properties associated with a change in optical response.

Interest in ellipsometry has grown steadily since the 1960s as it provided the sensitivity necessary to measure nanometer-scale layers used in microelectronics. Today, the range of applications has spread to basic research in physical sciences, semiconductor, data storage, flat panel display, communication, biosensor, and optical coating industries. This widespread use is due to increased dependence on thin films in many areas and the flexibility of ellipsometry to measure most material types: dielectrics, semiconductors, metals, superconductors, organics, biological coatings, and composites of materials.

This article provides a fundamental description of ellipsometry measurements along with the typical data analysis procedures. The primary applications of ellipsometry are also surveyed.

Light and Materials

Ellipsometry measurements involve the interaction between light and material.

Light

Light can be described as an electromagnetic wave traveling through space. For ellipsometry, it is adequate to discuss the electric field behavior in space and time, also known as polarization. The electric field of a wave is always orthogonal to the propagation direction. Therefore, a wave traveling along the z-direction can be described by its x- and y-components. If the light has completely random orientation and phase, it is considered unpolarized. For ellipsometry, we are interested in the case where the electric field follows a specific path and traces out a distinct shape at any point. This is known as polarized light. When two orthogonal light waves are in-phase, the resulting light will be linearly polarized (Figure 1a). The relative amplitudes determine the resulting orientation. If the orthogonal waves are 90° out-of-phase and equal in amplitude, the resultant light is circularly polarized (Figure 1b). The most general polarization is 'elliptical', which combines orthogonal waves of arbitrary amplitude and phase (Figure 1c). This is how ellipsometry gets its name.

Materials

Two values are used to describe optical properties, which determine how light interacts with a material.

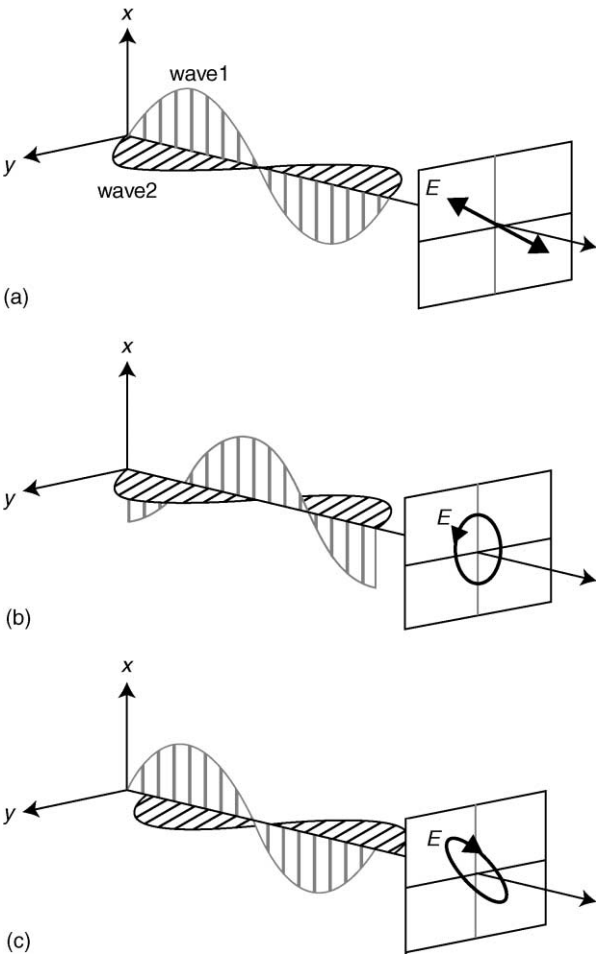


Figure 1 Orthogonal waves combine to demonstrate polarization: (a) linear; (b) circular; and (c) elliptical.

They are generally represented as a complex number. The complex refractive index ($\tilde{n}$) consists of the index (n) and extinction coefficient (k):

$$\tilde{n} = n + ik \quad [1]$$

Alternatively, the optical properties can be represented as the complex dielectric function:

$$\tilde{\varepsilon} = \varepsilon_1 + i\varepsilon_2 \quad [2]$$

with the following relation between conventions:

$$\tilde{\varepsilon} = \tilde{n}^2 \quad [3]$$

The index describes the phase velocity for light as it travels through a material compared to the speed of light in vacuum, c:

$$v = \frac{c}{n} \quad [4]$$

Light slows as it enters a material with higher index. Because frequency remains constant, the wavelength will shorten. The extinction coefficient describes the loss of wave energy to the material. It is related to the absorption coefficient, α, as:

$$\alpha = \frac{4\pi k}{\lambda} \quad [5]$$

Light loses intensity in an absorbing material, according to Beer's Law:

$$I(z) = I(0)\, e^{-i\alpha z} \quad [6]$$

Thus, the extinction coefficient relates how quickly light vanishes in a material. These concepts are demonstrated in Figure 2, where a light wave travels from air into two different materials of varying properties.

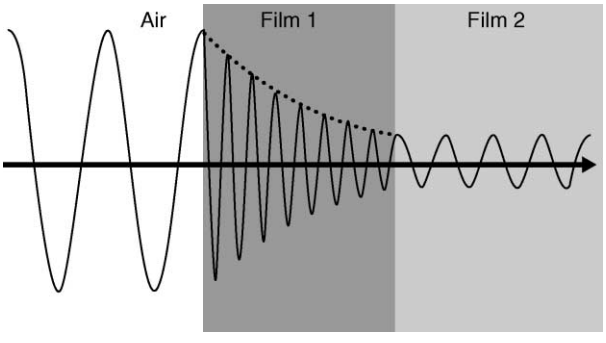


Figure 2 Wave travels from air into absorbing Film 1 and then transparent Film 2. The phase velocity and wavelength change in each material depending on index of refraction (Film 1: $n = 4$, Film 2: $n = 2$).

The optical constants for TiO_2 are shown in Figure 3 from the ultraviolet (UV) to the infrared (IR). The optical constants are wavelength dependent with absorption ($k > 0$) occurring in both UV and IR, due to different mechanisms that take energy from the light wave. IR absorption is commonly caused by molecular vibration, phonon vibration, or free-carriers. UV absorption is generally due to electronic transitions, where light provides energy to excite an electron to an elevated state. A closer look at the optical constants in Figure 3 shows that real and imaginary optical constants are not independent. Their shapes are mathematically coupled together through Kramers–Krönig consistency. Further details are covered later in this article.

Interaction Between Light and Materials

Maxwell's equations must remain satisfied when light interacts with a material, which leads to boundary conditions at the interface. Incident light will reflect and refract at the interface, as shown in Figure 4. The angle between the incident ray and sample normal

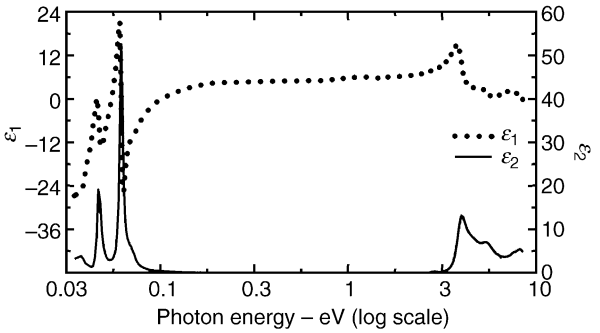


Figure 3 Complex dielectric function for TiO_2 film covering wavelengths from the infrared (small eV) to the ultraviolet (high eV).

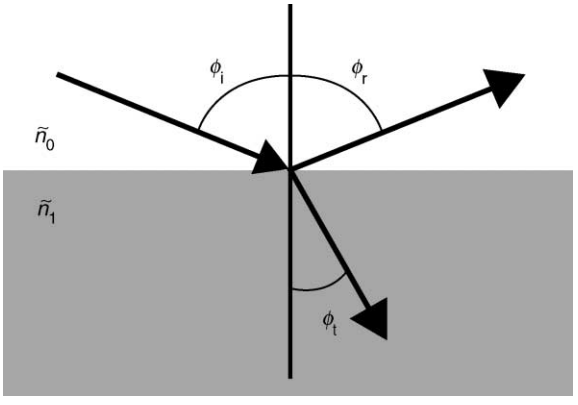


Figure 4 Light reflects and refracts according to Snell's law.

(ϕ_i) will be equal to the reflected angle, (ϕ_r). Light entering the material is refracted to an angle (ϕ_t) given by:

$$n_0 \sin(\phi_i) = n_1 \sin(\phi_t) \quad [7]$$

The same occurs at every interface where a portion reflects and the remainder transmits at the refracted angle. This is illustrated in Figure 5. The boundary conditions provide different solutions for electric fields parallel and perpendicular to the sample surface. Therefore, light can be separated into orthogonal components with relation to the plane of incidence. Electric fields parallel and perpendicular to the plane of incidence are considered p- and s-polarized, respectively. These two components are independent for isotropic materials and can be calculated separately. Fresnel described the amount of light reflected and transmitted at an interface between materials:

$$r_s = \left(\frac{E_{0r}}{E_{0i}}\right)_s = \frac{n_i \cos\theta_i - n_t \cos\theta_t}{n_i \cos\theta_i + n_t \cos\theta_t} \quad [8a]$$

$$r_p = \left(\frac{E_{0r}}{E_{0i}}\right)_p = \frac{n_t \cos\theta_i - n_i \cos\theta_t}{n_i \cos\theta_t + n_t \cos\theta_i} \quad [8b]$$

$$t_s = \left(\frac{E_{0t}}{E_{0i}}\right)_s = \frac{2n_i \cos\theta_i}{n_i \cos\theta_i + n_t \cos\theta_t} \quad [8c]$$

$$t_p = \left(\frac{E_{0t}}{E_{0i}}\right)_p = \frac{2n_i \cos\theta_i}{n_i \cos\theta_t + n_t \cos\theta_i} \quad [8d]$$

Thin film and multilayer structures involve multiple interfaces, with Fresnel reflection and transmission coefficients applicable at each. It is important to track the relative phase of each light component to correctly determine the overall reflected or transmitted beam. For this purpose, we define the film phase thickness as:

$$\beta = 2\pi\left(\frac{t_1}{\lambda}\right) n_1 \cos\theta_1 \quad [9]$$

The superposition of multiple light waves introduces interference that is dependent on the relative phase of each light wave. Figure 5 illustrates the combination of light waves in the reflected beam and their corresponding Fresnel calculations.

Ellipsometry Measurements

For ellipsometry, primary interest is measurement of how p- and s-components change upon reflection or transmission relative to each other. In this manner, the reference beam is part of the experiment. A known polarization is reflected or transmitted from the sample and the output polarization is measured. The change in polarization is the ellipsometry measurement, commonly written as:

$$\rho = \tan(\Psi)\, e^{i\Delta} \quad [10]$$

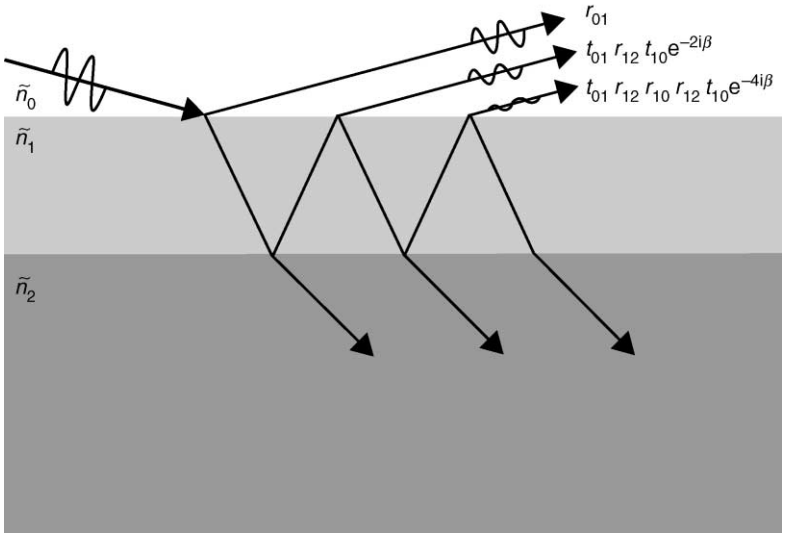


Figure 5 Light reflects and refracts at each interface, which leads to multiple beams in a thin film. Interference between beams depends on relative phase and amplitude of the electric fields. Fresnel reflection and transmission coefficients can be used to calculate the response from each contributing beam.

An example ellipsometry measurement is shown in Figure 6. The incident light is linear with both p- and s-components. The reflected light has undergone amplitude and phase changes for both p- and s-polarized light, and ellipsometry measures their changes.

The primary methods of measuring ellipsometry data all consist of the following components: light source, polarization generator, sample, polarization analyzer, and detector. The polarization generator and analyzer are constructed of optical components that manipulate the polarization: polarizers, compensators, and phase modulators. Common ellipsometer configurations include rotating analyzer (RAE), rotating polarizer (RPE), rotating compensator (RCE), and phase modulation (PME).

The RAE configuration is shown in Figure 7. A light source produces unpolarized light, which is sent through a polarizer. The polarizer passes light of a preferred electric field orientation. The polarizer axis is oriented between the p- and s-planes, such that both arrive at the sample surface. The linearly polarized light reflects from the sample surface, becoming elliptically polarized and travels through a continuously rotating polarizer (referred to as the analyzer). The amount of light allowed to pass will depend on the polarizer orientation relative to the electric field 'ellipse' coming from the sample. The detector converts light from photons to electronic signal to determine the reflected polarization. This information is used with the known input polarization to determine the polarization change caused by the sample reflection. This is the ellipsometry measurement of Ψ and Δ.

Single Wavelength Ellipsometry (SWE)

SWE uses a single frequency of light to probe the sample. This results in a pair of data, Ψ and Δ, used to determine up to two material properties. The optical design can be simple, low-cost, and accurate. Lasers are an ideal light source with well-known wavelength and significant intensity. The optical elements can be optimized to the single wavelength. However, there is relatively low information content (two measurement values). SWE is excellent for quick measurement of

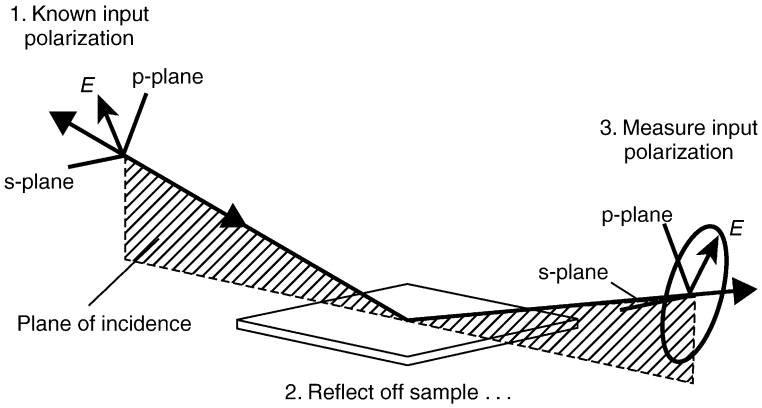


Figure 6 Typical ellipsometry configuration, where linearly polarized light is reflected from the sample surface and the polarization change is measured to determine the sample response.

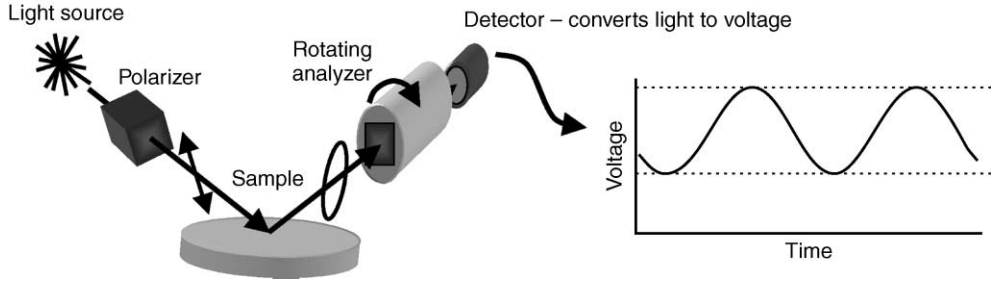


Figure 7 Rotating analyzer ellipsometer configuration uses a polarizer to define the incoming polarization and a rotating polarizer after the sample to analyze the outgoing light. The detector converts light to a voltage with the time-dependence leading to measurement of the reflected polarization.

nominally known films, like SiO_2 on Si. Care must be taken when interpreting unknown films, as multiple solutions occur for different film thickness. The data from a transparent film will cycle through the same values as the thickness increases. From Figure 5, it can be seen that this is related to interference between the multiple light beams. The refracted light may travel through different thicknesses to emerge in-phase with the first reflection, but is the returning light delayed by one wavelength or multiple wavelengths? Mathematically, the thickness cycle can be determined as:

$$D_\phi = \frac{\lambda}{2\sqrt{\tilde{n}_1^2 - \tilde{n}_0^2 \sin^2\phi}} \qquad [11]$$

This is demonstrated in Figure 8, where the complete thickness cycle is shown for SiO_2 on Si at 75° angle of incidence and 500 nm wavelength. A bare silicon substrate would produce data at the position of the star. As the film thickness increases, the data will move around this circle in the direction of the arrow. From eqn [11], the thickness cycle is calculated as 226 nm. Thus the cycle is completed and returns to the star when the film thickness reaches 226 nm. Any data point along the cycle represents a series of thicknesses which depend on how many times the cycle has been completely encircled. For example, the square represents a data point for 50 nm thickness. However, the same data point will represent 276 and 502 nm thicknesses. A second data point at a new wavelength or angle would help determine the correct thickness.

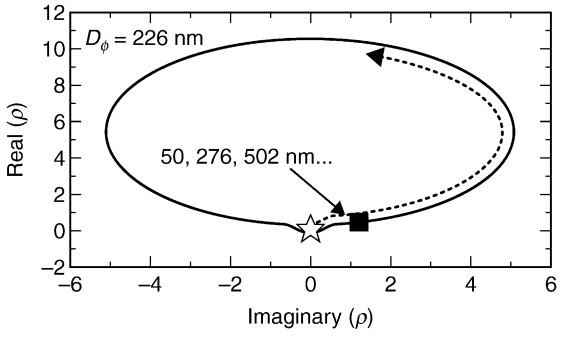


Figure 8 Data from a single wavelength will cycle through the same points as the film thickness increases. The star represents the starting point, with no SiO_2 film on Si. As the thickness increases, the data follow the cycle in the direction of the arrow. After the cycle is complete (thickness = D_ϕ), the data repeat the cycle. Thus, any point along the curve represents multiple possible thicknesses – example shown by a square at 50 nm thickness, which is also equal to 276 and 502 nm.

Spectroscopic Ellipsometry (SE)

Spectroscopic measurements solve the 'period' problem discussed for SWE. Data at surrounding wavelengths insure the correct thickness is determined, even as there remain multiple solutions at one wavelength in the spectrum. This is demonstrated in Figure 9, showing the spectroscopic data for the first three thickness results. The SE data oscillations clearly distinguish each thickness. Thus, a single thickness solution remains to match data at multiple wavelengths.

SE provides additional information content for each new wavelength. While the film thickness will remain constant regardless of wavelength, the optical constants will change across the spectrum. The optical constant shape contains information regarding the material microstructure, composition, and more. Different wavelength regions will provide the best information for each different material property. For this reason, SE systems have been developed to cover very wide spectral regions from the UV to the IR.

Spectroscopic measurements require an additional component, a method of wavelength selection. The most common methods include the use of monochromators, Fourier-transform spectrometers, and gratings or prisms with detection on a diode array. Monochromators are typically slow as they sequentially scan through wavelengths. Spectrometers and diode arrays allow simultaneous measurement at multiple wavelengths. This has become popular with the desire for high-speed SE measurements.

Variable Angle Ellipsometry

Ellipsometry measurements are typically performed at oblique angles, where the largest changes in polarization occur. The Brewster angle, defined as:

$$\tan(\phi_B) = \frac{n_1}{n_0} \qquad [12]$$

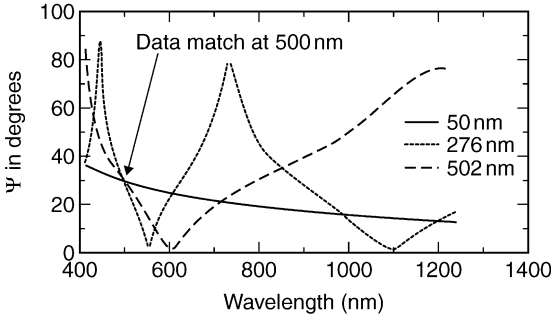


Figure 9 Three spectroscopic ellipsometry measurements that match at 500 nm wavelength, but are easily distinguishable as a function of wavelength.

gives the angle where reflection of p-polarized light goes through a minimum. Ellipsometry measurements are usually near this angle, which can vary from 55° for low-index dielectrics to 75° or 80° for semiconductors and metals.

It is also common to measure at multiple angles of incidence in ellipsometry. This allows additional data to be collected for the same material structure under different conditions. The most important change introduced by varying the angle is the change in path length through the film as the light refracts at a different angle. Multiple angles do not always add new information about a material structure, but the extra data builds confidence in the final answer.

In Situ

Within the last decade, it has become common to employ optical diagnostics during processing. *In situ* ellipsometry allows the optical response to be monitored in real-time. This has also led to feedback control for many processes. While *in situ* is generally restricted to a single angle of incidence, there is a distinct advantage to collecting data at different 'times' during processing to get unique 'glimpses' of the sample structure as it changes.

In situ SE is commonly applied to semiconductors. Conventional SE systems measure from the UV to NIR, which matches the spectral region where semiconductors absorb due to electronic transitions. The shape and position of the absorption can be very sensitive to temperature, composition, crystallinity, and surface quality, which allows SE to closely monitor these properties. *In situ* SE has also been used to monitor numerous material types, including metals and dielectrics. A promising *in situ* application is for multilayer optical coatings, where thickness and index can be determined in real-time to allow correction of future layers, to optimize optical performance.

Data Analysis

Ellipsometry measures changes in polarization. However, it is used to determine the material properties of interest, such as film thickness and optical constants. In the case of a bulk material, the equations derived for a single reflection can be directly inverted to provide the 'pseudo' optical constants from the ellipsometry measurement, ρ:

$$\langle\tilde{\varepsilon}\rangle = \sin^2(\phi)\left[1 + \tan^2(\phi)\left(\frac{1-\rho}{1+\rho}\right)\right] \qquad [13]$$

This equation assumes there are no surface layers of any type. However, there is typically a surface oxide or roughness for any bulk material and the direct inversion would incorporate these into the bulk optical constants. The more common procedure used to deduce material properties from ellipsometry measurements follows the flowchart in Figure 10. Regression analysis is required because an exact equation cannot be written. Often the answer is overdetermined with hundreds of experimental data points for a few unknowns. Regression analysis allows all of the measured data to be included when determining the solution.

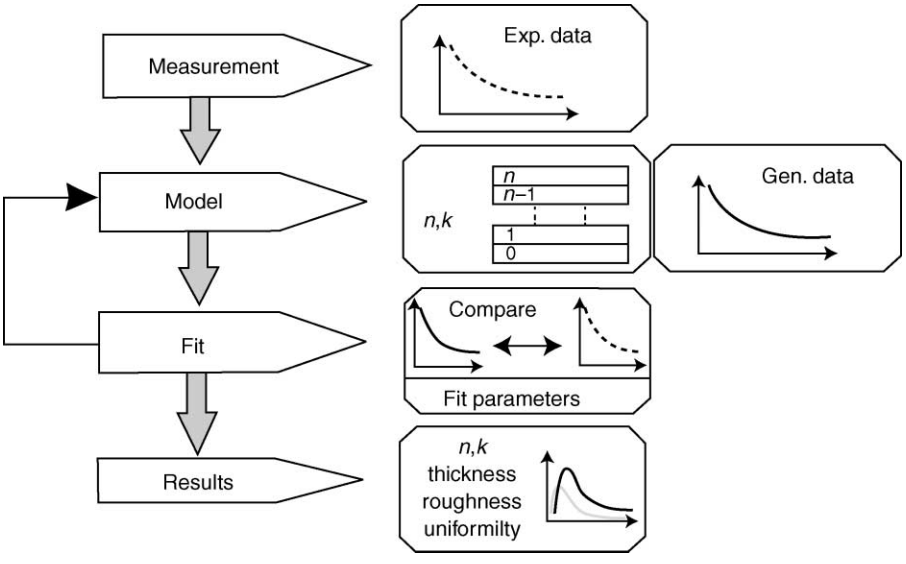


Figure 10 Flowchart for ellipsometry data analysis.

Data analysis proceeds as follows. After the sample is measured, a model is constructed to describe the sample. The model is used to calculate the predicted response from Fresnel's equations. Thus, each material must be described with a thickness and optical constants. If these values are not known, an estimate is given for the purpose of the preliminary calculation. The calculated values are compared to experimental data. Any unknown material properties can then be varied to improve the match between experiment and calculation. The number of unknown properties should not exceed the information content contained in the experimental data. For example, a single-wavelength ellipsometer produces two data points (Ψ,Δ) which allows a maximum of two material properties to be determined. Finding the best match between the model and experiment is typically done through regression. An estimator, like the mean squared error (MSE), is used to quantify the difference between curves. The unknown parameters are allowed to vary until the minimum MSE is reached.

Care must be taken to ensure the best answer is found, corresponding to the lowest MSE. For example, Figure 11a shows the MSE curve versus film thickness for a transparent film on silicon. There are multiple 'local' minima, but the lowest MSE value occurs at a thickness of 749 nm. This corresponds to the correct film thickness. It is possible that the regression algorithm will mistakenly fall into a 'local' minima, depending on the starting thickness and the MSE structural conditions. Comparing the results by eye for the lowest MSE and a local minima easily distinguish the true minima (see Figures 11b and c).

Ellipsometry Characterization

The two most common material properties studied by ellipsometry are film thickness and optical constants. In addition, ellipsometry can characterize material properties that affect the optical constants. Crystallinity, composition, resistivity, temperature, and molecular orientation can all affect the optical constants and in turn be measured by ellipsometry. This section details many of the primary applications important to ellipsometry.

Film Thickness

The film thickness is determined by interference between light reflecting from the surface and light traveling through the film. Depending on the relative phase of the rejoining light to the surface reflection, there can be constructive or destructive interference.

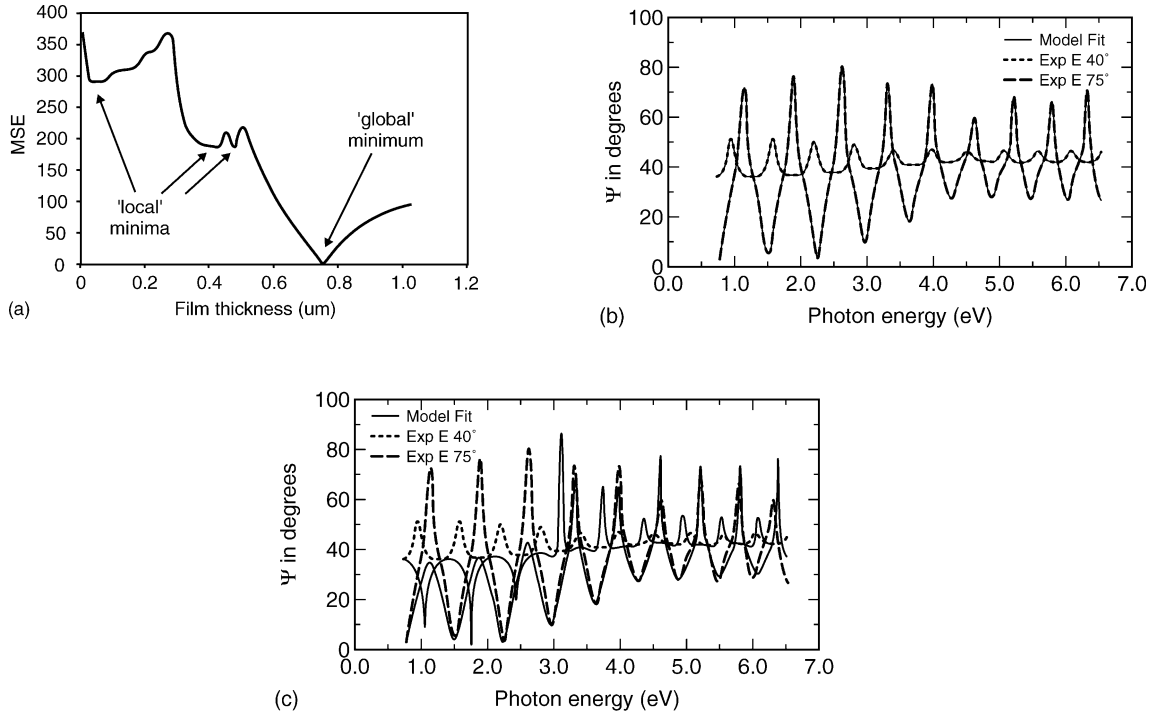


Figure 11 (a) MSE curve versus thickness shows the 'global' minimum where the best match between model and experiment occurs, and 'local' minima that may be found by the regression algorithm, but do not give the final result. (b) The experimental data and corresponding curves generated for the model at the 'global' minimum. (c) Similar curves at the 'local' minimum near 0.45 microns thickness is easily distinguishable as an incorrect result.

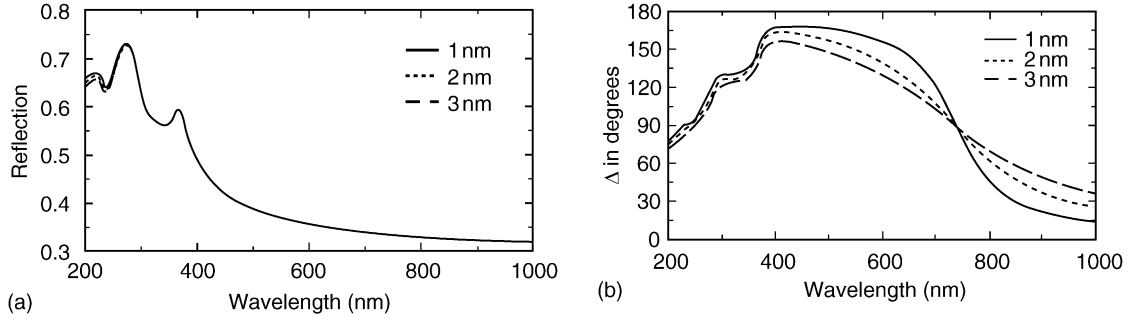


Figure 12 (a) Reflected intensity and (b) ellipsometric delta for three thin oxides on silicon show the high sensitivity of Δ to nanometer-scale films not available from the intensity measurement.

The interference involves both amplitude and phase information. The phase information from Δ is very sensitive to films down to submonolayer thickness. Figure 12 compares reflected intensity and ellipsometry for the same series of thin SiO_2 layers on Si. There are large variations in Δ, while the reflectance for each film is nearly the same.

Ellipsometry is typically used for films with thickness ranging from sub-nanometers to a few microns. As films become greater than several tens of microns thick, it becomes increasingly difficult to resolve the interference oscillations, except with longer infrared wavelengths, and other characterization techniques become preferred.

Thickness measurements also require a portion of the light to travel through the entire film and return to the surface. If the material is absorbing, thickness measurements by optical instruments will be limited to thin, semi-opaque layers. This limitation can be circumvented by measuring in a spectral region where there is lower absorption. For example, an organic film may strongly absorb UV and IR light, but remain transparent at mid-visible wavelengths. For metals, which strongly absorb at all wavelengths, the maximum layer for thickness determination is typically ~100 nm.

Optical Constants

Thickness measurements are not independent of the optical constants. The film thickness affects the path length of light traveling through the film, but the index determines the phase velocity and refracted angle. Thus, both contribute to the delay between surface reflection and light traveling through the film. Both n and k must be known or determined along with the thickness to get the correct results from an optical measurement.

The optical constants for a material will vary for different wavelengths and must be described at all wavelengths probed with the ellipsometer. A table of optical constants can be used to predict the response at each wavelength. However, it is less convenient to adjust unknown optical constants on a wavelength-by-wavelength basis. It is more advantageous to use all wavelengths simultaneously. A dispersion relationship often solves this problem, by describing the optical constant shape versus wavelength. The adjustable parameters of the dispersion relationship allow the overall optical constant shape to match the experimental results. This greatly reduces the number of unknown 'free' parameters compared to fitting individual n, k values at every wavelength.

For transparent materials, the index is often described using the Cauchy or Sellmeier relationship. The Cauchy relationship is typically given as:

$$n(\lambda) = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} \qquad [14]$$

where the three terms are adjusted to match the refractive index for the material. The Sellmeier relationship enforces Kramers–Krönig (KK) consistency, which ensures the optical dispersion retains a realistic shape. The Cauchy is not constrained by KK consistency and can produce unphysical dispersion. The Sellmeier relationship can be written as:

$$\varepsilon_1 = \frac{A\lambda^2\lambda_0^2}{(\lambda^2 - \lambda_0^2)} \qquad [15]$$

Absorbing materials will often have a transparent wavelength region that can be modeled with the Cauchy or Sellmeier relationships. However, the absorbing region must account for both real and imaginary optical constants. Many dispersion relationships use oscillator theory to describe the absorption for materials, which include the Lorentz, Harmonic, and Gaussian oscillators. They all share similar attributes, where the absorption features are described with an amplitude, broadening, and center energy (related to frequency of light).

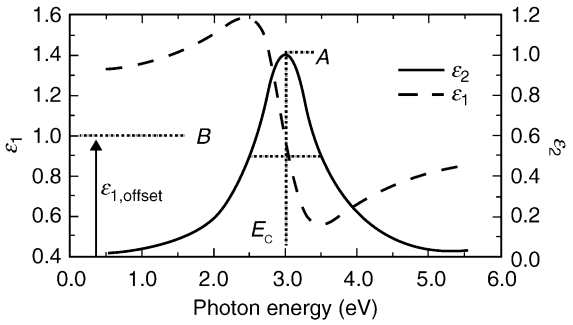


Figure 13 Lorentz oscillator illustrating the primary oscillator parameters: amplitude (A), broadening (B), and center energy (E_c) to describe the imaginary dielectric function shape and an offset ($\varepsilon_{1,\text{offset}}$) to help match the real component after KK transformation has defined its shape.

Kramers–Krönig consistency is used to calculate the shape of the real component after the imaginary behavior is described by the oscillator. An offset to the real component is added to account for extra absorption outside the measured spectral region. The Lorentz oscillator can be written as:

$$\tilde{\varepsilon} = \varepsilon_{1,\text{offset}} + \frac{AE_c}{E_c^2 - E^2 - iBE} \qquad [16]$$

where the parameters for amplitude (A), broadening (B), center energy (E_c), and offset ($\varepsilon_{1,\text{offset}}$) are also shown in Figure 13 for a typical Lorentz oscillator. The energy, E, is related to the frequency of a wave, ν:

$$E = h\nu \cong \frac{1240}{\lambda_{\text{nm}}} \qquad [17]$$

where h is Planck's constant and the wavelength, λ, is given in nanometers. More advanced dispersion models, like the Tauc–Lorentz and Cody–Lorentz, will include terms to describe the bandgap energy.

Mixing Materials

When two or more materials are mixed on a microscopic level, an effective medium approximation (EMA) may be used to determine the resulting optical constants. In the case of the Bruggemann EMA, which is commonly used, the mixture optical constants ($\tilde{\varepsilon}_{\text{eff}}$) relate to those of the individual materials, as:

$$f_a \frac{\varepsilon_a - \varepsilon_{\text{eff}}}{\varepsilon_a + 2\varepsilon_{\text{eff}}} + f_b \frac{\varepsilon_b - \varepsilon_{\text{eff}}}{\varepsilon_b + 2\varepsilon_{\text{eff}}} = 0 \qquad [18]$$

This can be interpreted as small particles of material A suspended in host material B. The length-scale for mixed particles must satisfy certain electromagnetic equations: typically smaller than one tenth the wavelength of light. In practice, EMA theory is useful for studying very thin surface roughness or interfacial intermixing of materials. These cases are both generally approximated by mixing the two surrounding materials in equal portions. In the case of surface roughness, the second material is void ($n = 1$). EMA theory has also been extended for application to porous materials where the directional dependence of the inclusions is handled mathematically.

Crystallinity

Semiconductors such as Si, are widely used materials, but their properties depend strongly on crystallinity. The UV absorption features in crystalline silicon are broadened and shifted as the material becomes more amorphous. This change in optical properties related to the degree of crystallinity has been used to advantage with ellipsometry measurements to monitor semiconductors and other films. Polysilicon films are used in both the display and semiconductor industry. The degree of crystallinity varies for different process conditions and can be monitored optically to ensure consistent material properties.

Composition

The composition of many alloys will affect the optical constants. The strongest changes often occur in the absorbing region, with shifts in position and amplitude of absorption features. For example, the electronic transitions in $Hg_{1-x}Cd_xTe$ move to higher energy with Cd concentration increase (Figure 14). This material is used for IR detectors that require precise control of composition. Spectroscopic ellipsometry performs this task in real-time with instant feedback to correct the composition during processing. Other applications include AlGaN and InGaN for optoelectronics.

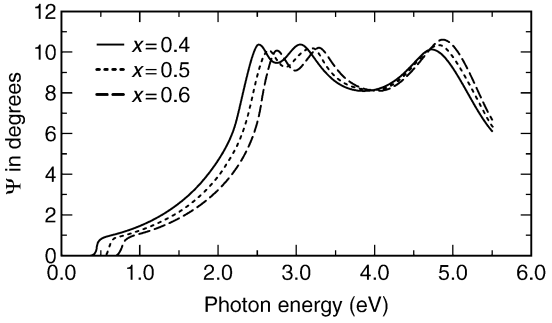


Figure 14 Optical properties for $Hg_{1-x}Cd_xTe$ vary with changes in composition.

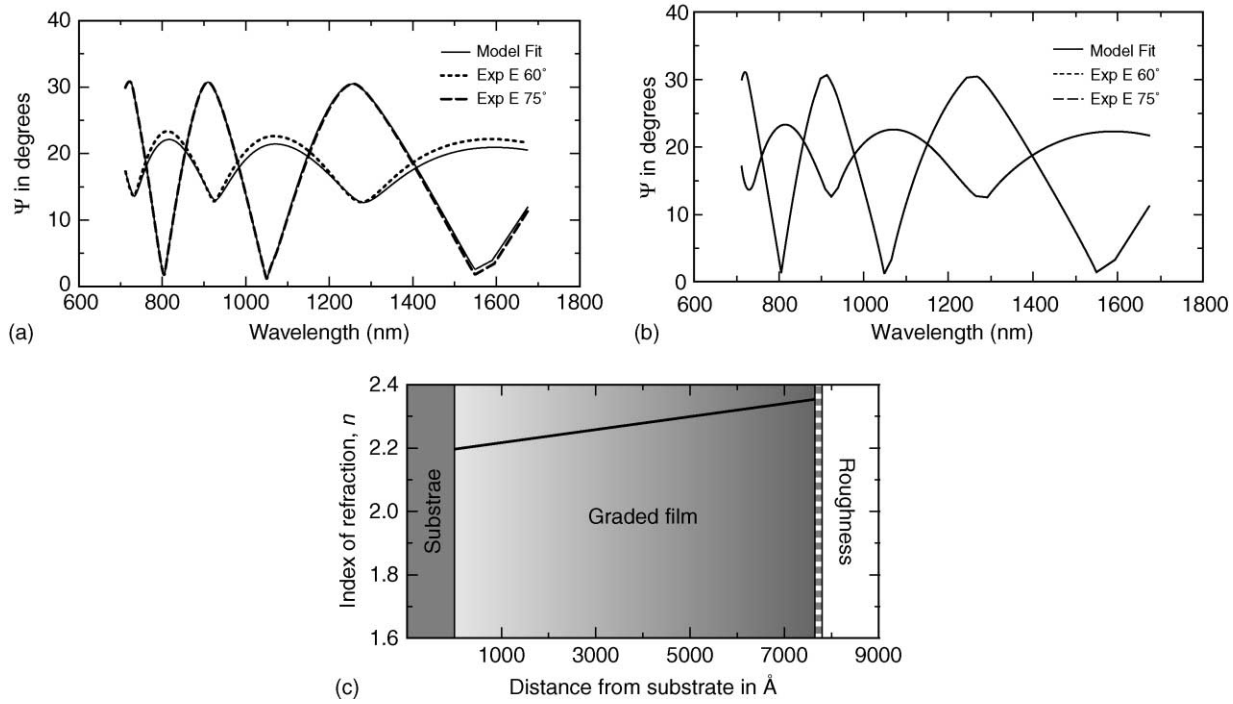


Figure 15 Experimental ellipsometry data and corresponding fit when the model is described as (a) a homogeneous single-layer and (b) a graded film with index variation through the film. (c) The best model to match subtleties in experimental data allows the index to increase toward the surface of the film, with a thin roughness layer on the surface.

Doping Concentration

Dopants in a material will introduce absorption in the infrared due to free-carriers. Spectroscopic ellipsometry measurements at long wavelengths can characterize this optical absorption, thus characterizing the doping concentration. This is common for highly doped semiconductors and transparent conductors such as indium tin oxide (ITO).

Optical Variation (Grading)

Many thin film properties change vertically throughout the film (along the direction perpendicular to the surface). This is most often induced by processing conditions, whether intentional or unintentional. Figure 15a shows the fit to experimental spectroscopic ellipsometry data taken from a single-layer film when it is modeled as a homogeneous layer. To improve the agreement between experimental and model-generated curves, the index was allowed to vary in a linear manner throughout the film. The best fit is shown in Figure 15b, where the model includes the index variation and a thin surface roughness layer. The sample structure is demonstrated in Figure 15c.

Optical Anisotropy

Many materials are optically anisotropic; i.e., their optical properties vary in different film directions. Typical ellipsometry measurements assume no cross-coupling between p- and s-polarizations. This cannot be assumed with anisotropic materials, which has led to 'generalized' ellipsometry measurements. Generalized ellipsometry measures additional information regarding the p- to s- and s- to p- conversion upon reflection. This allows characterization of anisotropic materials, which contain directionally dependent optical constants.

Conclusions

Ellipsometry is a common optical technique for measuring thin films and bulk materials. It relies on the polarization changes due to reflection or transmission from a material structure to deduce material properties, like film thickness and optical constants. This technique continues to develop as the requirement for thin film characterization increases.

See also

Geometrical Optics: Lenses and Mirrors; Prisms. **Optical Coatings:** Thin-Film Optical Coatings. **Optical Materials:** Measurement of Optical Properties of Solids.

Polarization: Introduction; Matrix Analysis. **Semiconductor Physics:** Band Structure and Optical Properties. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Aspnes DE (1985) The accurate determination of optical properties by ellipsometry. In: Palik ED (ed.) *Handbook of Optical Constants of Solids*, pp. 89–112. Orlando FL: Academic Press.

Azzam RMA and Bashara NM (1987) *Ellipsometry and Polarized Light*. Amsterdam, The Netherlands: Elsevier Science B.V.

Boccara AC, Pickering C and Rivory J (eds) (1993) *Spectroscopic Ellipsometry*. Amsterdam: Elsevier Publishing.

Collins RW, Aspnes DE and Irene EA (eds) (1998) *Proceedings from the Second International Conference on Spectroscopic Ellipsometry*. Thin Solid Films, vols. 313–314.

Gottesfeld S, Kim YT and Redondo A (1995) Recent applications of ellipsometry and spectroellipsometry in electrochemical systems. In: Rubinstein I (ed.) *Physical Electrochemistry: Principles, Methods, and Applications*. New York: Marcel Dekker.

Herman IP (1996) *Optical Diagnostics for Thin Film Processing*, pp. 425–479. San Diego, CA: Academic Press.

Johs B, Woollam JA, Herzinger CM, *et al.* (1999) Overview of variable angle spectroscopic ellipsometry (VASE), Part II: Advanced applications. *Optical Metrology*, vol. CR72, pp. 29–58. Bellingham, Washington: SPIE, Bellingham.

Johs B, Hale J, Ianno NJ, *et al.* (2001) Recent developments in spectroscopic ellipsometry for *in situ* applications. In: Duparré A and Singh B (eds) *Optical Metrology Roadmap for the Semiconductor, Optical, and Data Storage Industries II*, vol. 4449, pp. 41–57. Bellingham, Washington: SPIE.

Roseler A (1990) *Infrared Spectroscopic Ellipsometry*. Berlin: Akademie-Verlag.

Rossow U and Richter W (1996) Spectroscopic ellipsometry. In: Bauer G and Richter W (eds) *Optical Characterization of Epitaxial Semiconductor Layers*, pp. 68–128. Berlin: Springer-Verlag.

Tompkins HG (1993) *A User's Guide to Ellipsometry*. San Diego, CA: Academic Press.

Tompkins HG and Irene EA (eds) (in press) *Handbook of Ellipsometry*. New York: William Andrew Publishing.

Tompkins HG and McGahan WA (1999) *Spectroscopic Ellipsometry and Reflectometry*. New York: John Wiley & Sons, Inc.

Woollam JA (2000) Ellipsometry, variable angle spectroscopic. In: Webster JG (ed.) *Wiley Encyclopedia of Electrical and Electronics Engineering*, pp. 109–116. New York: John Wiley & Sons.

Woollam JA and Snyder PG (1992) Variable angle spectroscopic ellipsometry. In: Brundle CR, Evans CA and Wilson S (eds) *Encyclopedia of Materials Characterization: Surfaces, Interfaces, Thin Films*, pp. 401–411. Boston, MA: Butterworth-Heinemann.

Woollam JA, Johs B, Herzinger CM, *et al.* (1999) Overview of variable angle spectroscopic ellipsometry (VASE), Part I: Basic theory and typical applications. *Optical Metrology*, vol. CR72, pp. 3–28. Washington: SPIE, Bellingham.

Photometry

J Schanda, University of Veszprém, Veszprém, Hungary

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Fundamentals of Photometry

Photometry is defined by the International Commission on Illumination (internationally known as CIE from the abbreviation of its French name: Commission Internationale de l'Eclairage) as 'measurement of quantities referring to radiation as evaluated according to a given spectral efficiency function, e.g., $V(\lambda)$ or $V'(\lambda)$.' A note to the above definition states that in many languages it is used in a broader sense, covering the science of optical radiation measurement. We will restrict our treatise to the above fundamental meaning of photometry.

Under spectral efficiency function we understand the spectral luminous efficiency function of the human visual system. The internationally agreed symbols are $V(\lambda)$, for photopic vision (daylight conditions) and $V'(\lambda)$ for scotopic vision (nighttime conditions). CIE standardized the $V(\lambda)$ function in 1924 and the $V'(\lambda)$ function in 1951; their spectral distribution is shown in Figure 1. The $V(\lambda)$ function was determined using mainly the so-called flicker photometric technique, where the light of a reference stimulus and of a test stimulus of varying wavelengths are alternatively shown to the human observer at a frequency at which the observer is already unable to perceive the difference in color (due to the different wavelength of the two radiations), but still perceives a

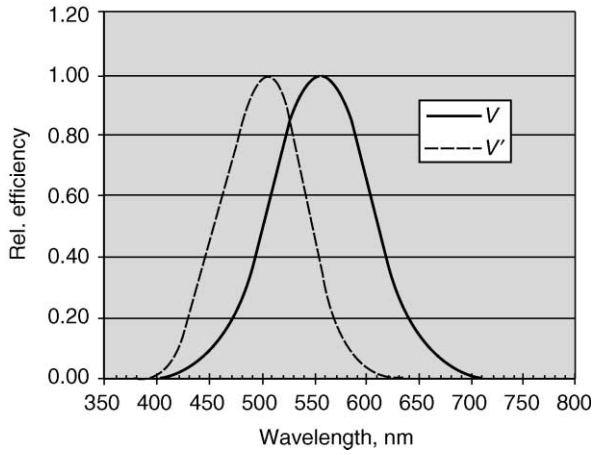


Figure 1 Spectral efficiency functions of the human eye under photopic $V(\lambda)$ and scotopic $V'(\lambda)$ conditions (see CIE The basis of physical photometry. CIE 18.2:1983).

flicker sensation if the luminances of the two stimuli are different. Adjusting the radiance of the test stimulus, we can reach the situation of minimum flicker sensation. In this case we state that we set the two stimuli to equal luminance. Luminance is the quantity that the human observer perceives as brightness (in the case of near white stimuli, see below); radiance is the physical counterpart measured in $\mathrm{W\,m^{-2}\,sr^{-1}}$. For the $V'(\lambda)$ function one can just project the two stimuli side by side, as under scotopic adaptation we cannot distinguish colors, only brightness differences are perceived, thus one can set equal scotopic luminance for the two stimuli by adjusting the brightness of the test stimulus until it agrees with that of the reference stimulus.

Illuminating engineering was mainly interested in describing the effect of near white light, and therefore a psychophysical correlation was selected that described the perceived brightness sensation reasonably well. For near white stimuli, additivity holds, i.e., if A, B, C, and D are four stimuli, and the A stimulus matches the B stimulus and the C stimulus matches the D stimulus, then the superposition of the A and C stimuli matches the superposition of the B and D stimuli. The 'match' word is used here to describe stimuli that produce the same perception. It could be shown that for these stimuli, if the radiance is weighted with the $V(\lambda)$ function, the constructed luminances will be equal, i.e:

$$\text{if} \qquad L_A = L_B \text{ and } L_C = L_D$$

$$\text{then} \quad L_A + L_C = L_B + L_D, \text{ or } L_A + L_D = L_B + L_C$$

where

$$L_x = K_m \int_{380\ \mathrm{nm}}^{780\ \mathrm{nm}} S_{\lambda x} V(\lambda) \mathrm{d}\lambda \qquad [1]$$

and x refers to A, or B, or C, or D, and $S_{\lambda x}$ is the spectral radiance distribution that produces the stimulus

$$S_\lambda = \frac{\mathrm{d}S(\lambda)}{\mathrm{d}\lambda}$$

where $S(\lambda)$ is the spectral radiance. K_m is the maximum value of the luminous efficacy of radiation, its value is $683\ \mathrm{lm\,W^{-1}}$ (see the discussion of luminous flux in the sub-section on photometric quantities). To be precise the integration should go from 0 nm to infinity, but it is usual to define the lower and upper wavelength limits of the visible spectrum as 380 nm and 780 nm. $V(\lambda)$ is defined between 360 nm and 830 nm, and $V'(\lambda)$ between 380 nm and 780 nm, see Figure 1.

We have to stress that the concept of luminance – and the entire system of the present-day photometry – is not to quantify brightness perception; it is only a reasonable approximation for near white stimuli. For colored lights a brightness–luminance discrepancy exists (called the Helmhotz–Kohlrausch effect: saturated colors look brighter than predicted by luminance). Luminance is, however, a good description for the visibility of fine details, thus it is a good concept for lighting calculations.

Photometry has a unique situation in the SI system of units: the candela (cd) is a base unit of the SI system, the only one that is connected to psychophysical phenomena. By 1979, in the definition of the 16th General Conference of Weights and Measures, the candela was traced back to radiation quantities; nevertheless it was kept as the base unit of photometry:

> The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} Hz and that has a radiant intensity in that direction of $1/683\ \mathrm{W\,sr^{-1}}$. ($540 \times 10^{12}$ Hz corresponds approximately to 555 nm.)

To be able to calculate a photometric quantity of radiation of other wavelengths, one of the psychophysical functions $V(\lambda)$ or $V'(\lambda)$ has to be used and eqn [1] applied. We will see later that besides $V(\lambda)$ and $V'(\lambda)$ in modern applications, a number of further spectral luminous efficiency functions might be used.

Photometric Quantities

As discussed in the previous section, from the point of view of vision, the most important quantity is

luminance. Looking, however, at the definition of the base unit, it is obvious that from the physical point of view the definition of a quantity corresponding to radiant power, measured in watts, can help to bridge the gap between photometry and radiometry: luminous flux, measured in lumens (lm), with the symbol Φ, is defined by eqn [1], with $S_\lambda(\lambda)$ inserted in W m^{-1}.

Based on this more practical quantity of luminous flux and its unit, lm, the different quantities used in photometry can be built up as follows:

$$\textit{Luminous flux}: \quad \Phi = K_m \int_0^\infty \frac{\mathrm{d}\Phi_e(\lambda)}{\mathrm{d}\lambda} V(\lambda)\mathrm{d}\lambda \qquad [2]$$

where $\Phi_e(\lambda)$ is the radiant flux measured in W,

$$\Phi_{e,\lambda} = \frac{\mathrm{d}\Phi_e(\lambda)}{\mathrm{d}\lambda}$$

is the spectral distribution (or spectral concentration) of the radiant flux (W m^{-1}). $K_m = 683$ lm W^{-1}, is the maximum value of the luminous efficacy of radiation for $\lambda_m \approx 555$ nm.

Similar equations can be written for scotopic luminous flux, with $V(\lambda)$ exchanged with $V'(\lambda)$, where $K'_m = 1700$ lm W^{-1} for $\lambda_m \approx 507$ nm. All further quantities can be defined both for photopic and scotopic conditions. Here we write them only as photopic quantities.

$$\textit{Luminous intensity}: \quad I = \frac{\mathrm{d}\Phi}{\mathrm{d}\Omega} \qquad [3]$$

where $\mathrm{d}\Phi$ is the luminous flux traveling in an elementary solid angle $\mathrm{d}\Omega$, assuming a point source (see Figure 2). The unit of luminous intensity is the candela (cd = lm sr^{-1}).

$$\textit{Luminance}: \quad L = \frac{\partial^2\Phi}{\partial A \cos\Theta\, \partial\Omega} \qquad [4]$$

where $\mathrm{d}\Phi$ is the luminous flux traveling in an elementary solid angle $\mathrm{d}\Omega$, $\mathrm{d}A$ is the elementary surface area emitting the radiation, and Θ is the angle between the normal of $\mathrm{d}A$ and the direction of luminance measurement (see Figure 3). The unit of luminance is cd m^{-2} (in some older literature called nit).

$$\textit{Illuminance}: \quad E = \frac{\mathrm{d}\Phi}{\mathrm{d}A} \qquad [5]$$

where $\mathrm{d}\Phi$ is the luminous flux incident on the $\mathrm{d}A$ element of a surface (see Figure 4). The unit of illuminance is the lux (lx = lm m^{-2}).

A remark on the use of SI units and related quantities
No prefixes can be added to the SI units, thus irrespective whether one measures luminance based

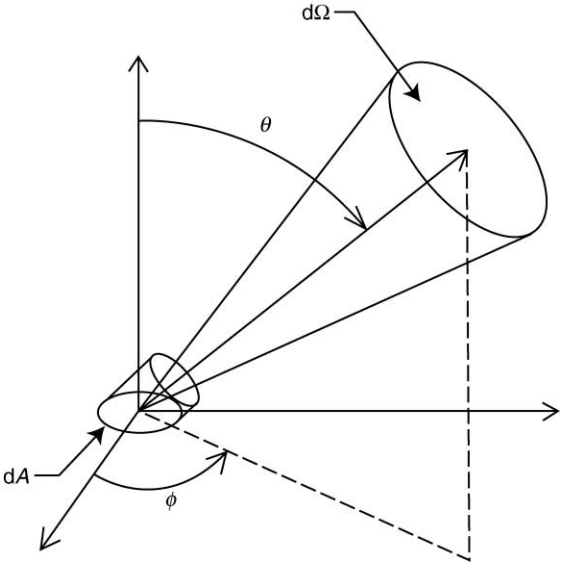


Figure 3 Geometry for the definition of luminance.

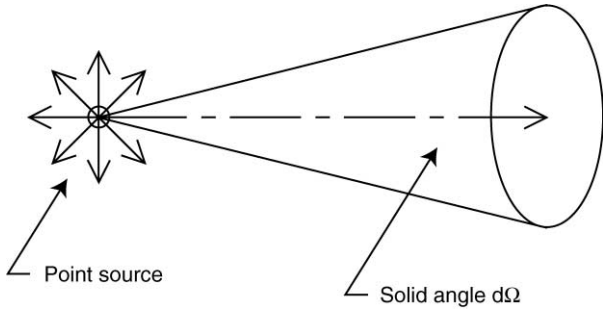


Figure 2 Concept of point source, solid angle, and luminous intensity.

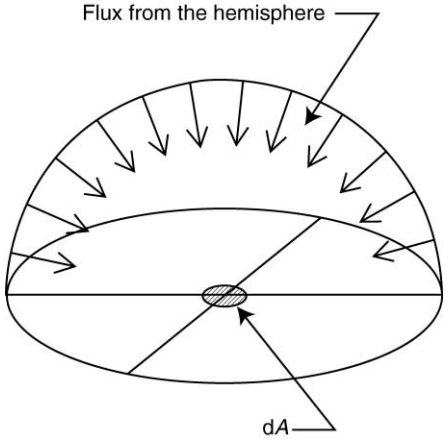


Figure 4 Illuminance is the total luminous flux per unit area incident at a point coming from a hemispherical solid angle.

on the photopic $V(\lambda)$ function, or the scotopic $V'(\lambda)$ function, the so-determined photopic luminance or scotopic luminance is measured in cd m^{-2} (no photopic or scotopic lumen, candela, etc., exist!). This often creates confusion, because only for a 555 nm monochromatic radiation will a 1 cd m^{-2} photopic or scotopic luminance produce equal visual sensation, for every other wavelength the two are not commensurable. One can only relate photopic measurement results to photopic ones, and scotopic measurement results to scotopic ones.

Concepts of Advanced Photometry

The quantities and units, as described in the previous section, are the quantities and units internationally agreed by the Meter Convention and the International Standards Organization (ISO). Modern photometric applications need, however, some further quantities and weighting functions. Thus CIE, the international organization for the development of standards in the field of optical radiation measurements, has defined a series of further weighting functions and quantities. The most important ones are the following:

$V_M(\lambda)$ function

In the blue part of the spectrum (below 460 nm) the values of the $V(\lambda)$ function turned out to be too low. For decades this was only of concern for the vision research community, but with the introduction of blue LEDs and other short wavelength emitting sources (e.g., blue channel of displays) this shortcoming of the $V(\lambda)$ function became of practical importance. The 'CIE 1988 2° spectral luminous efficiency function for photopic vision' corrects this anomaly (see Figure 5).

$V_{10}(\lambda)$ function

The official $V(\lambda)$ function is valid only for foveal vision (i.e., for targets that are smaller than 4° of visual angle). The foveal area of the retina is covered with a yellow pigmented layer (macula lutea) that absorbs in the blue part of the spectrum, at larger visual angles this screening is not effective anymore. The $V_{10}(\lambda)$ function was determined for a visual angle of 10° (see Figure 5). Its international recommendation is still under consideration at the time of writing this article and if accepted, it will be recommended for targets seen at approximately 10° off-axis, e.g., for measuring the photometric properties of traffic signs and signals, where the driver has to observe the information in the periphery of his or her eye.

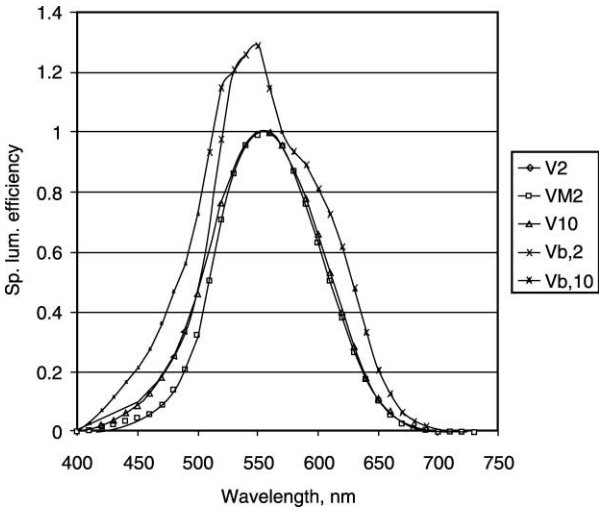


Figure 5 Spectral luminous efficiency (SPL) functions defined or under consideration for international adoption: V2: standard $V(\lambda)$ function; VM2: CIE 1988 $V_M(\lambda)$ function, this is equivalent to the brightness SPL for a point source; V10: 10° visual field SPL, Vb,2: 2° visual field brightness SPL, $V_{b,2}(\lambda)$; Vb,10: 10° visual field brightness SPL, $V_{b,10}(\lambda)$.

Brightness matching functions

As already mentioned in the Introduction, luminance is not a good correlate of brightness, which is a human perception. To find a better correlation of brightness the problem to be addressed is that brightness is a nonadditive phenomenon, i.e., in eqn [1] one cannot add (integrate) the monochromatic radiations to get a brightness correlation for a nonmonochromatic radiation. Brightness evaluating spectral luminous efficiency functions can be used only to compare monochromatic radiations. The CIE has compiled such functions for point sources, 2° and 10° visual field sizes. Figure 5 also shows these functions.

Due to the fact that the brightness perception is nonadditive in respect of the stimuli that produce it, no brightness photometry can be built that uses equations, as shown in eqn [1]. For brightness description we have to rely on the concept of equivalent luminance, a term the definition of which has recently been updated.

Equivalent luminance

Of a field of given size and shape, for a radiation of arbitrary relative spectral distribution L_{eq} : Luminance of a comparison field in which monochromatic radiation of frequency 540×10^{12} Hz has the same brightness as the field considered under the specified photometric conditions of measurement; the comparison field must have a specified size and shape

which may be different from that of the field considered.

To build an instrument that measures this quantity is a real challenge for the future.

Advanced Use of Photometry

Based on above newly defined quantities, several attempts are under way to extend the usefulness of photometry in designing the human visual environment. The eye is an optical system and as in every such system, the depth of focus and the different aberrations of the system will decrease with decreasing pupil size. Pupil size will decrease with increasing illumination, and in the case of constant luminance with higher content of short wavelength radiation. Thus, there exists a school of researchers who advocate that increased blue content in the light has beneficial effects on vision, and one should extend the classical photopic-based photometry with a scotopic-based one to properly describe the visual effect of lighting.

Other investigations are concerned about the visibility at low light levels, the range used in street lighting (from about a few thousands of a candela per square meter to about a few candelas per square meter, according to one definitions: 10^{-3}cd m^{-2}–3 cd m^{-2}). In this mesopic range both rods and cones are contributing to vision, and this changes with lighting level and direction of view (for foveal vision, i.e., looking straight ahead, photopic photometry seems to hold even at low light levels). For peripheral visual angles brightness perception and the perception of an object (a signal, sign or obstacle in a nighttime driving situation) seem to have different spectral responsivity. In driving situations the necessary reaction time of the driver is an important parameter, thus experiments are going on to define a photometric system based on reaction time investigations.

In indoor situations apart from the necessary level of illumination, the observed glare is a contributor whether an environment will be accepted as pleasing or annoying. Illuminating engineering distinguishes between two types of glare: disability glare reduces visibility, discomfort glare is just an annoying experience without influencing the short-term task performance. An interesting question is the spectral sensitivity to discomfort glare, as it can influence not only indoor but also outdoor activity. Preliminary experiments seem to show that luminance sensitivity and discomfort glare sensitivity have different spectral distribution; glare sensitivity seems to peak at shorter wavelengths.

The above might be related to a further question, lying already at the boundaries of photometry, but that has to be considered in photometric design and measurement: the human daily and yearly rhythm (circadian and seasonal rhythm) of human activity coupled to hormone levels. They are influenced by light, as, for example, the hormone melatonin production is influenced by light exposure. Physiological investigations showed that melatonin production suppression has a maximum around 460 nm and might be coupled to a radiation sensitive ganglion cell in the retina. Whether discomfort sensation is mediated via the same neural pathway or via a visual one has not yet been decided. But photometry has to take these also into consideration and in the future, measurement methods and instruments to determine them, will have to be developed.

Advances in Photometric Measurements

Primary Standards

The main concern in photometry is that the uncertainty of photometric measurements is still much higher than that in other branches of physics. This is partly due to the higher uncertainty in radiometry and partly to the increase in uncertainty within the chain of uncertainty propagation from the National Laboratory to the workshop floor measurement.

In National Standards Laboratories, very sophisticated systems are used to determine the power of the incoming radiation and then elaborated spectroradiometric techniques are used to evaluate the radiation in the form of light, i.e., perform photometric measurements (e.g., NIST CIRCUS equipment, where multiple laser sources are used as power sources, and highly sophisticated methods to produce a homogeneous nonpolarized radiation field for the calibration of secondary photometric detectors).

There is, however, also a method to supply end users with absolute detectors for the visible part of the spectrum. Modern high-end Si photovoltaic cells have internal quantum efficiencies in the visible part of the spectrum of over 99.9%. Reflection losses at the silicon surface are minimized by using three or more detectors arranged in a trap configuration, where the light reflected from one detector is fed to the second one, from there to the third one, and eventually to some further ones. In a three detector configuration, as shown in Figure 6, the light from the third detector is reflected back to the second and from there to the first one. As every detector reflects only a small amount of radiation, by the fifth reflection practically

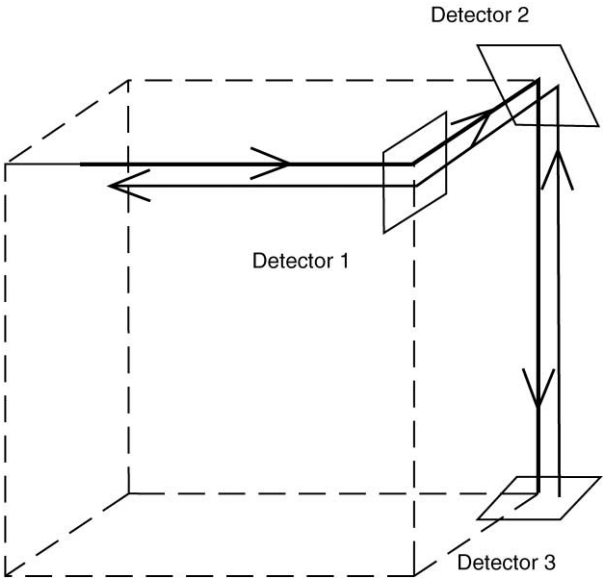


Figure 6 Schematic layout of a three Si-cell trap detector: light comes in from the left, is first partly absorbed, partly reflected on Detector 1, then on Detector 2, then on Detector 3, from where it is reflected back to Detector 2 and 1.

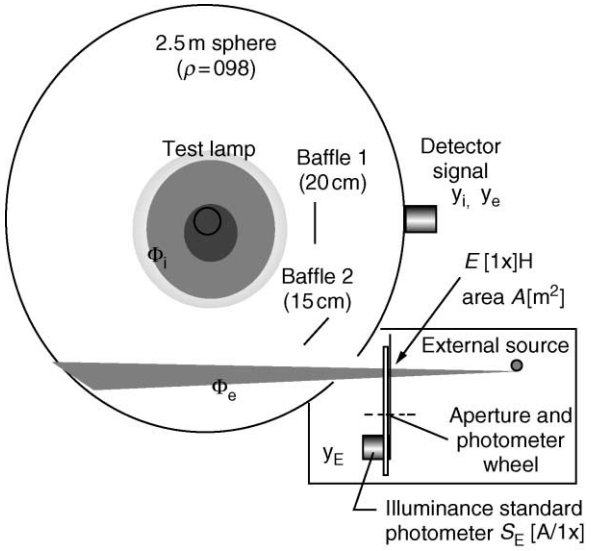


Figure 7 Arrangement of the absolute integrating sphere system developed at NIST for the detector-based total luminous flux calibration. By permission of IESNA from Ohno Y and Bergman R (2003) Detector-referenced integrating sphere photometry for industry. *J. IES Summer* 21–26.

all the radiation is absorbed and contributes to the electric signal. Such trap detectors have an almost 100% quantum efficiency in the visible part of the spectrum, and can be used as photometric detectors if a well designed color correcting filter is applied in front of the detector.

Secondary Type Measurements

In practical photometry the three most important quantities to be measured are the total luminous flux of different lamps, the illuminance in a plane and the luminance.

Light source measurement

Total luminous flux. The two methods to measure the total luminous flux is to use a goniophotometer or a photometer (Ulbicht) sphere. In goniophotometry, recent years have not brought major breakthroughs, the automation of the systems got better, but the principles are unchanged.

The integrating sphere photometer (a sphere with inner white diffuse coating, where the lamp is in the middle of the sphere) used to be a simple piece of equipment to compare total luminous flux lamps against flux standards. In recent years a new technique has been developed at NIST–USA. This enables the absolute measurement of luminous flux from illuminance measurement, the fundamentals of this new arrangement being shown in Figure 7: the test lamp is as usual in the middle of the sphere, but now light from an external source is introduced into the sphere. An illuminance meter measures the flux entering from this source. The sphere detector compares the two signals (y_i and y_e). Knowing the absolute characteristics of the sphere (a difficult measurement) one can determine the total luminous flux of the test lamp using the absolute illuminance value. As illuminance is easily determined from luminous intensity (from eqns [3] and [5] one gets with $d\Omega = dA/r^2$ that $E = I/r^2$, where r is the distance between the source and the illuminated surface, supposed to be perpendicular to the direction to the source), this technique permits us to derive the total luminous flux scale from illuminance or luminous intensity measurement using an integrating sphere.

Luminous intensity of LEDs. An other major break-through achieved during the past years was the unified measurement of LED intensity. Light-emitting diodes became, in recent years, important light sources for large scale signaling and signing, and it is foreseen that they will become important contributors in every field of light production (from car headlamps to general illumination). The most fundamental parameter of the light of an LED is its luminous intensity. The spatial power distribution of LEDs is usually collimated, but the LEDs often squint, as seen in Figure 8. In the past, some manufacturers measured the luminous intensity in

the direction of maximum emission, others used the direction of the optical axis for this quantity. The highly collimated character of the radiation made measurements in far field rather difficult. Therefore, CIE recommended a new term and measuring geometry: average LED intensity can be measured under two measuring conditions, as shown in Figure 9. The detector has to be set in the direction of the LED mechanical axis (discussions are still going on as to what the reference direction should be with modern surface mounted LEDs, as with those the mechanical axis is ill-defined, the normal to the base-plane could be a better reference direction). The detector has to have an exactly 1.00 cm^2 circular aperture, and the distance between this aperture and the tip of the LED is for condition A, $d = 0.316$ m, and for Condition B, $d = 0.100$ m (these two distances with the 1.00 cm^2 detector area provide 0.001 sr (steradians) and 0.01 sr opening angles). Recent international round-robins have shown that, based on the new recommendations, agreement between different laboratories decreased from the 10 to 20% level to 1 to 2%. The remaining difference is mainly due to the fact that the LEDs emit in narrow wavelength bands, and the transfer of the calibration value for the 100 mm^2 detector from a white (CIE Standard Illuminant A color temperature) incandescent lamp to the narrow band LED emission is still uncertain, mainly due to stray light effects in the spectral responsivity and emission measurement.

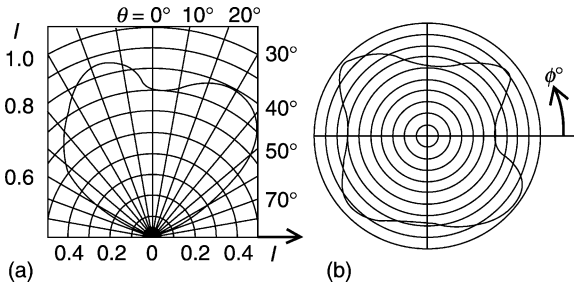


Figure 8 Spatial light distribution of an LED, Figure 8a shows the distribution of a 'squinting' LED in a plane including the optical axis, Figure 8b shows light distribution in a plane perpendicular to the optical axis. By permission of the Commission Internationale de l'Eclairage, from the Publication "Measurement of LEDs" CIE 127-1997; CIE Publications are obtainable from the CIE Central Bureau: Kegelgasse 27, A-1033 Wien, Austria.

Luminance distribution measurement

The human observer sees luminance (and color) differences. Thus for every illuminating engineering design task the luminance distribution in the environment is of utmost importance. Traditionally this was measured using a spot-luminance meter, aiming the device into a few critical directions. The recent development of charge coupled device (CCD) two-dimensionally sensitive arrays (and other, e.g., MOSFET, CMOS, Charge injection device (CID), charge imaging matrix (CIM) systems: as for the time being the CCD technology provides best performance, we will refer to two-dimensional electronic image capture devices as to CCD cameras) opened the possibility of using an image-capturing camera for luminance distribution measurements. Such measurements are badly needed in display calibration, near-field photometry (photometry in planes nearer than in which the inverse square law holds), testing of car headlamp light distribution, glare, and homogeneity studies indoors and outdoors, etc.

Solid-state cameras have the big advantage over older type vacuum-tube image capturing devices, that the geometric position and alignment of the single pixels is well-defined and stays constant. Nowadays, CCD image detector chips are mass produced, and one can get a variety of such devices and cameras starting with some small resolution

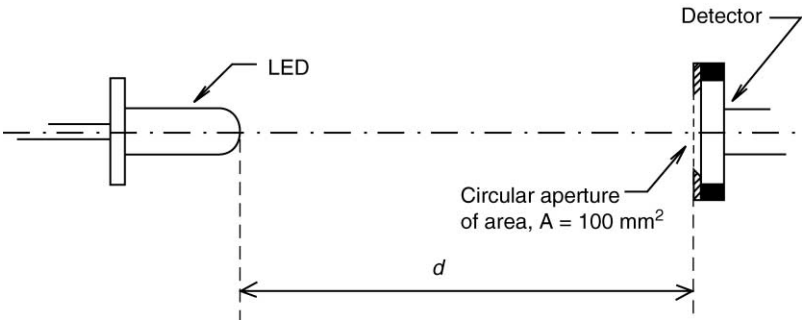


Figure 9 Schematic diagram of CIE Standard Conditions for the measurement of Average LED Intensity. Distance $d = 0.316$ m for Condition A and $d = 0.100$ m for Condition B. By permission of the Commission Internationale de l'Eclairage, from the Publication "Measurement of LEDs" CIE 127-1997; CIE Publications are obtainable from the CIE Central Bureau: Kegelgasse 27, A-1033 Wien, Austria.

(few thousand pixels) devices up to cameras with tens of mega pixel resolution. Detectors are now available with internal intensification enabling measurements down to a few photons per second intensity levels.

Main problems with these detectors are:

- spectral and absolute nonuniformities of the single pixels (see Figure 10), where a spatial homogeneity map of a CCD two-dimensional array detector is shown; the irregular 3% sensitivity change on the surface of the detector is negligible for imaging purposes, but has to be corrected in cases of photometric measurements. Even if the receptor chip would have an absolutely homogeneous sensitivity, there would be a drop in response from the middle of the imaging area to the boarders: Light reaching the edges of the detector reach the detector at an oblique angle and this produces a decrease of sensitivity with α^4, where α is the angle of incidence, measured from the middle of the lens to the given pixel of the detector and the surface normal of the detector.
- aliasing effects if the pixel resolution is not large enough to show straight lines as such when they are not in line with a pixel row or column.
- nonlinearity and cross-talk among the adjacent pixels. Figure 11 shows the so-called 'inverse gamma' characteristic of a CCD camera. The digital electronic output of the camera shows a $Y = E^{-\gamma}$ type function, where Y is the output DAC (digital-analog converter) values, E is the irradiance of the pixel, and γ is the exponent (this description comes from the film industry, where the film density depends exponentially on the irradiation; display devices have usually a nonlinear input (DAC value) – output (luminance) characteristic, and the camera inverse gamma value corrects for this output device characteristic. This is, however, not required if the camera is used for photometric measurements; this built-in nonlinearity has to be corrected in the evaluating software if the camera is intended for photometric measurements.
- to be able to perform photometric measurements the camera has to have a spectral responsivity corresponding to the CIE $V(\lambda)$-function. Many cameras have built-in filters to do this – eventually also red and blue filters to be able to capture color – but the color correction of cameras, where the correction is made by small adjacent filter chips, is usually very poor. Figure 12 shows the spectral

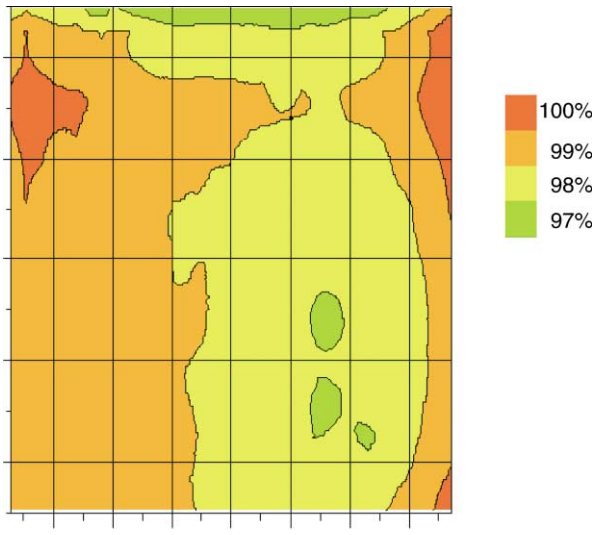


Figure 10 Spatial homogeneity of a two-dimensional CCD array.

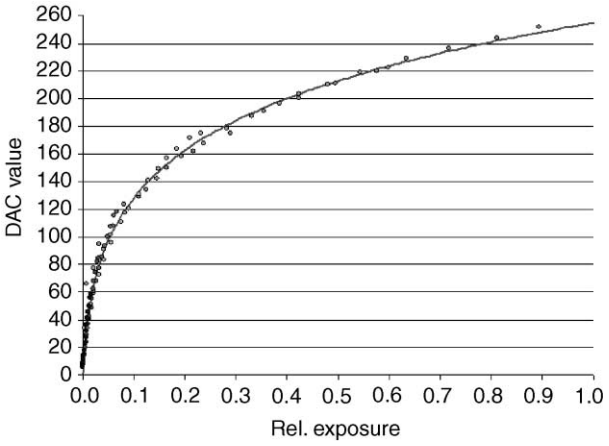


Figure 11 'Inverse gamma' characteristic of a commercial digital photographic camera, measurement points are shown at different speed settings, curve is a model function representative of the camera response.

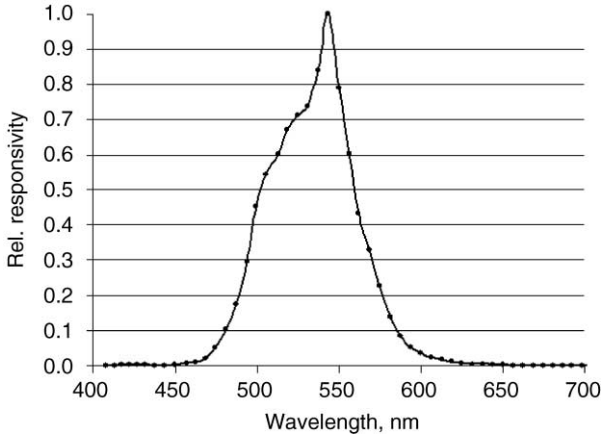


Figure 12 Spectral sensitivity of a commercial digital photographic camera.

sensitivity curve of a digital photographic camera; the output signal is produced by an internal matrix transformation of the signals produced by adjacent pixels equipped with different color filters. Figure 13 shows the spectral sensitivity of a CCD camera specially designed for photometric measurements. Naturally, meaningful photometric measurements can be made only with such a camera. Very often, however, an approximate luminance distribution is enough, and then a picture captured by a digital photographic camera, plus one luminance measurement of a representative object for absolute calibration, suffices.

Above nonspectral systematic errors can be corrected by appropriate soft-ware, so a CCD camera photometer, as shown schematically in Figure 14, is well suited to measure display characteristics, indoor and outdoor luminance distributions. The challenge for illuminating engineering is at present how the many millions of luminance values can be evaluated to get to meaningful light measurement data.

The real challenge will come if visual science provides better hints how the human visual system evaluates the illuminance distribution on the retina, and instrument manufacturers will be able to capture signals corresponding to those produced by the receptor cells and provide the necessary algorithms our brain uses to get to brightness, lightness, color, luminance contrast, and glare type of output information.

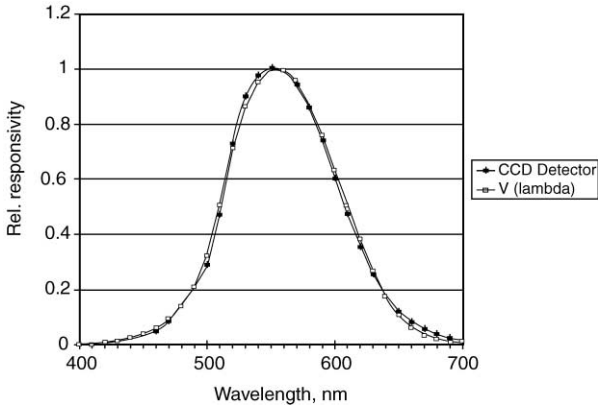


Figure 13 Spectral sensitivity of a CCD camera specially designed for photometric measurements. Kindly supplied by INPHORA Inc.

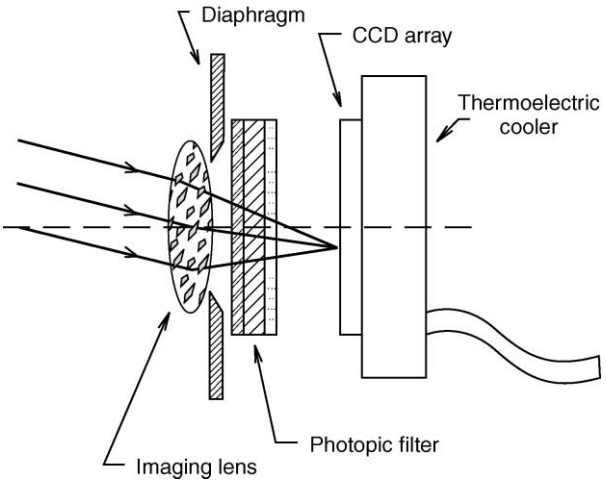


Figure 14 Cross-section of a photometric CCD camera.

Concluding Remarks

Advances in optical instrumentation, both in the field of light sources and detectors – coupled with the possibilities modern digital computation (eventually in the future increasing the use of neural networks) – provide already many new measurement technical solutions and further ones are certainly underway.

The use of LEDs needs the rethinking of many classical illuminating engineering concepts, from visibility and glare evaluation, evenness of illumination to color rendering. All of them are coupled with problems in basic photometry. Thus, there is a need to re-evaluate concepts used in design techniques. The new area-sensitive detectors provide methods of determining classical photometric quantities of entire visual fields in one shot, but already foreshadow the development of new quantities that correlate better with visual perceptions.

List of Units and Nomenclature

Terms with an $*$ refer to definitions published by the CIE in the International Lighting Vocabulary, CIE 17.4:1986, where further terms and definitions related to light and lighting are to be found.

Candela*: SI unit of luminous intensity: The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} hertz and that has a radiant intensity in that direction of 1/683 watt per steradian. (16th General Conference of Weights and Measures, 1979).

$$1\ \text{cd} = 1\ \text{lm sr}^{-1}$$

Equivalent luminance*: Luminance of a comparison field in which monochromatic radiation of frequency 540×10^{12} Hz has the same brightness as the field considered under the specified photometric conditions of measurement; the comparison field must have a specified size and shape which may be

different from that of the field considered.

$$\text{unit} : \text{cd m}^{-2}$$

Notes:

1. Radiation at a frequency of 540×10^{12} Hz has a wavelength in standard air of 555.016 nm.
2. A comparison field may also be used in which the radiation has any relative spectral distribution, if the equivalent luminance of this field is known under the same conditions of measurement.

Far field photometry: Photometry where the inverse square law is valid.

Flicker photometer*: Visual photometer in which the observer sees either an undivided field illuminated successively, or two adjacent fields illuminated alternately, by two sources to be compared, the frequency of alteration being conveniently chosen so that it is above the fusion frequency for colours but below the fusion frequency for brightnesses.

Fovea*: Central part of the retina, thin and depressed, which contains almost exclusively cones and forming the site of most distinct vision.

Note: The fovea subtends an angle of about 0.026 rad (1.5 degree) in the visual field.

Goniophotometer*: Photometer for measuring the directional light distribution characteristics of sources, luminaires, media or surfaces.

Illuminance*: Quotient of the luminous flux $d\Phi_v$ incident on an element of the surface containing the point, by the area dA of that element.

Equivalent definition. Integral, taken over the hemisphere visible from the given point, of the expression $L_v \cos\theta \, d\Omega$, where L_v is the luminance at the given point in the various directions of the incident elementary beams of solid angle $d\Omega$, and θ is the angle between any of these beams and the normal to the surface at the given point.

$$E_v = \frac{d\Phi_v}{dA} = \int_{2\pi sr} L_v \cos\theta \, d\Omega$$

$$\text{unit} : \text{lx} = \text{lm m}^{-2}$$

Inverse square law: The illumination at a point on a surface varies directly with the luminous intensity of the source, and inversely as the square of the distance between the source and the point if the source is seen as a point source.

Lumen*: SI unit of luminous flux: Luminous flux emitted in unit solid angle (steradian) by a uniform point source having a luminous intensity of 1 candela. (9th General Conference of Weights and Measures, 1948).

Equivalent definition. Luminous flux of a beam of monochromatic radiation whose frequency is 540×10^{12} hertz and whose radiant flux is 1/683 watt.

Luminance*: Quantity defined by the formula

$$L = \frac{\partial^2 \Phi}{\partial A \cos\theta \, \partial\Omega}$$

where $\partial\Phi$ is the luminous flux transmitted by an elementary beam passing through the given point and propagating in the solid angle $d\Omega$ containing the given direction; dA is the area of a section of that beam containing the given point, θ is the angle between the normal to that section and the direction of the beam.

$$\text{unit} : \text{cd m}^{-2}$$

Luminous intensity*: Quotient of the luminous flux $d\Phi_v$ leaving the source and propagated in the element of solid angle $d\Omega$ containing the given direction, by the element of solid angle:

$$I_v = \frac{d\Phi_v}{d\Omega}$$

$$\text{unit} : \text{cd} = \text{lm sr}^{-1}$$

Lux*: SI unit of illuminance: Illuminance produced on a surface of area 1 square meter by a luminous flux of 1 lumen uniformly distributed over that surface.

$$1 \text{ lx} = 1 \text{ lm m}^{-2}$$

Note: Non-metric unit: **lumen per square foot** (lm ft^{-2}) or **footcandle** (fc) (USA) = 10.764 lx.

Near-filed photometry: photometry made in the vicinity of an extended source, so that the inverse square law is not valid.

Photometer (or Ulbicht) sphere*: A hollow sphere, whitened inside. Owing to the internal reflexions in the sphere, the illumination on any part of the sphere's inside surface is proportional to the luminous flux entering the sphere, or produced inside the sphere by a lamp. The illuminance of the internal sphere wall is measured via a small window.

Pixel: The individual picture elements of an image or elements in a display that can be addressed individually.

Radiance*: Quantity defined by the formula

$$L = \frac{\partial^2 \Phi}{\partial A \cos\theta \, \partial\Omega}$$

where $\partial\Phi$ is the radiant flux transmitted by an elementary beam passing through the given point and propagating in the solid angle $d\Omega$ containing the given direction; dA is the area of a section of that beam containing the given point, θ is the angle between the normal to that section and the direction of the beam.

$$\text{unit} : \text{W m}^{-2} \text{ srad}$$

Retina*: Membrane situated inside the back of the eye that is sensitive to light stimuli; it contains photoreceptors, the cones and the rods, and nerve cells that interconnect and transmit to the optic nerve the signals resulting from stimulation of the photoreceptors.

Spectral*: An adjective that, when applied to a quantity X pertaining to electromagnetic radiation, indicates:

- either that X is a function of the wavelength λ, symbol: $X(\lambda)$,
- or that the quantity referred to is the spectral concentration of X, symbol: $X_\lambda \equiv dX/d\lambda$.

X_λ is also a function of λ and in order to stress this may be written $X_\lambda(\lambda)$ without any change of meaning.

Spectral luminous efficiency function: for photopic vision, $V(\lambda)$; for scotopic vision $V'(\lambda)$*: Ratio of the radiant flux at wavelength λ_m to that at wavelength λ such that both radiations produce equally intense luminous sensations under specified photometric conditions and λ_m is chosen so that the maximum value of this ratio is equal to 1.

Unless otherwise indicated, the values used for the spectral luminous efficiency in photopic vision are the values agreed internationally in 1924 by the CIE (*Compte Rendu* 6^e session, p.67), completed by interpolation and extrapolation (*Publications CIE* No.18 (1970), p.43 and No. 15 (1971), p.93), and recommended by the International Committee of Weights and Measures (CIPM) in 1972. For *scotopic vision*, the CIE in 1951 adopted, for young observers, the values published in *Compte Rendu* 12^e session, Vol. 3, p. 37, and ratified by the CIPM in 1976. These values define respectively the $V(\lambda)$ or $V'(\lambda)$ functions.

Total luminous flux: luminous flux of a source emitted into 4π steradians.

Trap detector: Detector array prepared from detectors of high internal quantum efficiency, where the reflected radiation of the first detector is directed to the second one, and so on so that practically all the radiation is absorbed by one of the detectors (the radiation is "trapped" in the detector system).

See also

Displays. **Incoherent sources:** Lamps.

Further Reading

Commission Internationale de l'Eclairage: *The basis of physical photometry* (1983) CIE 18.2.

Commission Internationale de l'Eclairage: Proceedings of the CIE Symposium 2002 on *Temporal and Spatial Aspects of Light and Colour Perception and Measurement*, August 2002: Publ.CIE x025:2003.

Cromer CL, Eppledauer G, Hardis JE, Larason TC and Parr AC (1993) National Institute of Standards and Technology detector-based photometric scale. *Applied Optics* 32: 2936–2948.

Decusatis C (ed.) (1997) *Handbook of Applied Photometry*. Woodbury, New York: American Institute of Physics.

Rea MS (ed.) (2000) Illuminating Engineering Society of North America. *The IESNA Lighting Handbook: Reference and Application*, 9th edn.

Kaiser PK and Boynton RM (1966) *Human Color Vision*, 2nd edn. Washington, DC: Optical Society of America.

McCluney WR (1994) *Introduction to radiometry and photometry*. Boston, London: Artech House.

Ohno Y and Bergman R (2003) Detector-referenced integrating sphere photometry for industry. *J IES Summer* 21–26.

Walsh JWT (1953) *Photometry*. London: Constable.

Scatterometry

J C Stover, The Scatter Works, Inc., Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Scattered light is a limiting source of optical noise in many advanced optical systems, but it can also be a sensitive indicator of optical component quality. Consider the simple case of a telescope successfully used to image a dim star against a dark background; however, if light from a bright source (such as the moon located well out of the field of view) enters the telescope, it will scatter from the interior walls and the imaging optics themselves. Some of this light eventually reaches the detector and creates a dim background haze that washes out the image of the distant star. A good telescope design accounts for these effects by limiting potential scatter propagation paths and by requiring that critical elements in the optical system meet scatter specifications. This means doing a careful system analysis and a means to quantify the scattering properties of the telescope components. This article discusses modern techniques for quantifying, measuring, and analyzing scattered light, and reviews their development.

Like many scientific advances moving scatterometry from an art to a reliable metrology was done in a series of small hops (not always in the same direction), rather than a single leap. It started in 1961 when a paper by Hal Bennett and Jim Porteous reported measurements made by gathering most of the light scattered from front surface mirrors and normalizing this signal by the much larger specular reflection. They defined this ratio as the total integrated scatter (TIS), and using a scalar diffraction theory result drawn from the radar literature, related it to the surface root mean square (rms) roughness. By the mid-1970s, several angle-resolved scatterometers had been built as research tools in university, government, and industry labs. Unfortunately, instrument operation and data manipulation were generally poor, and meaningful comparison measurements were virtually impossible due to instrument differences, sample contamination, and confusion over what parameters should be compared. Analysis of scatter data, to characterize sample surface roughness, was the subject of many publications. A derivation of what is commonly called 'BRDF' (bidirectional reflectance distribution function) was published by Nicodemus and co-workers at the National Bureau of Standards (now the National Institute of Science and Technology or NIST) in 1970, but did not gain common acceptance as a way to quantify scatter measurements until the late 1980s when the advent of small computers, combined with inspection requirements for defense-related optics, dramatically stimulated the development of scatter metrology. Commercial laboratory instrumentation became available that could measure and analyze as many as 50 to 100 samples a day, and the number (and sophistication) of measurement facilities increased dramatically. The first ASTM Standards were published (TIS in 1987 and BRDF in 1991), but it was still several years before most publications correctly used these quantifying terms. Government defense funding decreased dramatically in the early 1990s, following the end of the Cold War, but the economic advantages of using scatter metrology and analysis for space applications and in the rapidly advancing semiconductor industry, continued state-of-the-art advancements.

The following sections detail how scatter is quantified when related to area (roughness) and local (pit/particle) generating sources. Instrumentation and the use of scattering models are also briefly reviewed.

Quantifying Scattered Light

Scatter signals can be easily quantified as scattered light power per unit solid angle (in watts per steradian); however, in order to make the results more meaningful, these signals are usually normalized, in some fashion, by the light incident on the scatter source. The three ways commonly employed to do this are defined below.

If the scattering feature in question is uniformly distributed across the illuminated spot on the sample (such as surface roughness), then it makes sense to normalize the collected scattered power in watts/steradian by the incident power. This simple ratio, which has units of inverse steradians, was commonly referred to as 'the scattering function.' Although this term is occasionally still found in the literature, it has been generally replaced by the closely related BRDF, which is defined by the differential ratio of the sample radiance normalized by its irradiance. After some simplifying assumptions are made, this reduces to the original scattering function with a cosine of the polar scattering angle in the denominator. The BRDF, defined in this manner, has become the standard way to report angle-resolved scatter from features that uniformly fill the illuminated spot. The cosine term results from the fact that NIST used radiometric terms to define BRDF:

$$\mathrm{BRDF} = \frac{\mathrm{Ps}/\Omega}{\mathrm{Pi}\cos\theta_s} \qquad [1]$$

The scatter function is often referred to as the 'cosine corrected BRDF' and is simply equal to the BRDF multiplied by the cosine of the polar scattering angle. Figure 1 gives the geometry for the situation, and defines the polar and azimuthal angles (θ_s and ϕ_s), as well as the solid collection angle (Ω). Other common abbreviations are BSDF, for the more generic bidirectional scatter distribution function, and BTDF for quantifying transmissive scatter.

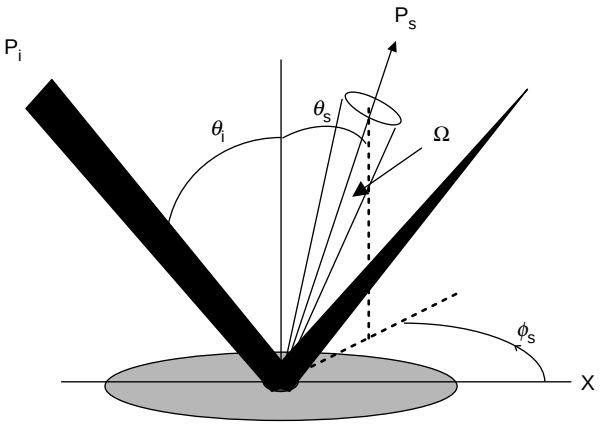


Figure 1 Scatter analysis uses standard spherical coordinates to define terms.

Integration of the scatter signal over much of the scattering hemisphere allows calculation of TIS, as the ratio of the scatter signal to the reflected specular power. This integration is usually carried out experimentally in such a way that both the incident beam and reflected specular beam are excluded. In the most common TIS situation, the beam is incident at a small angle near surface normal, and the integration is done from small values of θ_s to almost 90 degrees. If the fraction of light scattered from the specular reflection is small and if the scatter is caused by surface roughness, then it can be related to the rms surface roughness of the reflecting surface. As a ratio of powers, the TIS is a dimensionless quantity. The normalization is by Pr (instead of Pi) because reductions in scatter caused by low reflectance do not influence the roughness calculation. The pertinent relationships are given below, where σ is the rms roughness and λ is the light wavelength:

$$\mathrm{TIS} = \mathrm{Ps}/\mathrm{Pr} \cong (4\pi\sigma/\lambda)^2 \qquad [2]$$

Of course, all scatter measurements are integrations over a detector collection aperture, but the TIS designation is reserved for situations where the aim is to gather as much scattered light as possible, while 'angle resolved' designs are created to gain information from the distribution of the scattered light. Notice that TIS values become very large when measured from a diffuse surface, where the specular reflection is very small. Although TIS can be measured for any surface, the diffuse reflectance (equal to Ps/Pi) would often be more appropriate for diffuse surfaces. The various restrictions associated with relating TIS to rms roughness are detailed below.

Scatter from discrete features, such as particles and pits, which do not completely fill the illuminated spot, must be treated differently. This is because changes in spot size, with no corresponding change in total incident power, will change the incident intensity (watts/unit area) at the feature and thus also change the scatter signal (and BRDF) without any corresponding changes in the scattering feature. Clearly this is unacceptable if the object is to characterize the defect with scatter measurements. The solution is to define another quantification term, known as the differential scattering cross-section (DSC), where the normalization is the incident intensity at the feature (the units for DSC are area/steradian). Because this was not done in terms of radiometric units at the time it was defined, the cosine of the polar scattering angle is not in the definition. The same geometrical definitions, found in Figure 1, also apply for the DSC:

$$\mathrm{DSC} = \frac{\mathrm{Ps}/\Omega}{Ii} \qquad [3]$$

If the DSC is integrated over the solid angle associated with a collection aperture then the value has units of area. Because relatively small area focused laser beams are often used as a source, area is most commonly given in micrometers squared.

These three scatter parameters, the BRDF, the TIS, and the DSC, are obviously functions of system variables such as geometry, scatter direction (both in and out of the incident plane), incident wavelength and polarization, as well as feature characteristics. It is the dependence of the scatter signal on these system parameters that makes the scatter models useful for optimizing instrument designs. It is their dependence on feature characteristics that makes scatter measurement a useful metrology tool.

A key point needs to be stressed. When applied appropriately, TIS, BRDF, and DSC are absolute terms, not relative terms. The DSC of a 100 nm PSL in a given direction for a given source is a fixed value, which can be repeatedly measured and even accurately calculated from models. The same is true for TIS and BRDF values associated with surface roughness of known statistics. Scatter measuring instruments, such as particle scanners or lab scatterometers, can be calibrated in terms of these quantities. As has already been pointed out, the user of a scanner will almost always be more interested in characterizing defects than in the resulting scatter values, but the underlying instrument calibration can always be expressed in terms of these three quantities. This is true even though designers and users may find it convenient to use other metrics (such as polysterene latex (PSL) spheres) as a way to relate calibration.

Angle Resolved Scatterometers

The diagram in Figure 2 shows the most common scatterometer configuration. The source is fixed and the sample is rotated to the desired incident angle. The receiver is then rotated about the sample during scatter measurement. Most commonly, scatterometers operate just in the plane of incidence; however, instruments capable of measuring at virtually any location in either the reflective or transmissive hemispheres have been built. Although dozens of instruments have been built following this general design, other configurations are in use. For example, the source and receiver may be fixed and the sample rotated so that the scatter pattern moves past the receiver. This is easier mechanically than

moving the receiver at the end of an arm, but complicates analysis because the incident angle and the observation angle change simultaneously. Another combination is to fix the source and sample together, at constant incident angle, and rotate this unit (about the point of illumination on the sample) so that the scatter pattern moves past a fixed receiver. This has the advantage that a long receiver/sample distance can be used without motorizing a long (heavy) receiver arm. It has the disadvantage that heavy (or multiple) sources are difficult to deal with. Other configurations, with everything fixed, have been designed that employ several receivers to merely sample the BSDF and display a curve fit of the resulting data. This is an economical solution if the BSDF is relatively uniform without isolated diffraction peaks. The goniometer section of a real instrument, similar to that of Figure 2, is shown in Figure 3.

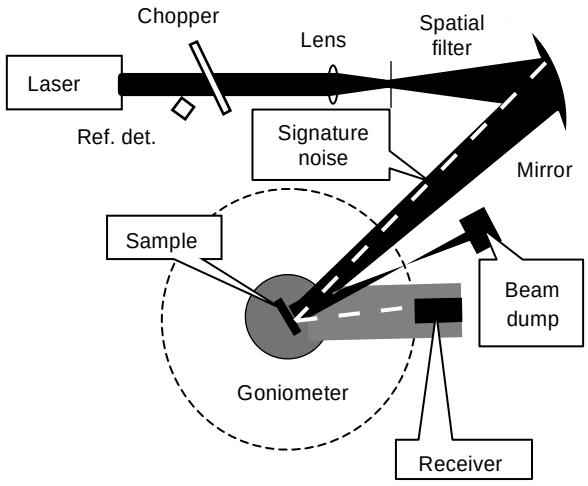


Figure 2 Basic elements of an incident plane scatterometer are shown.

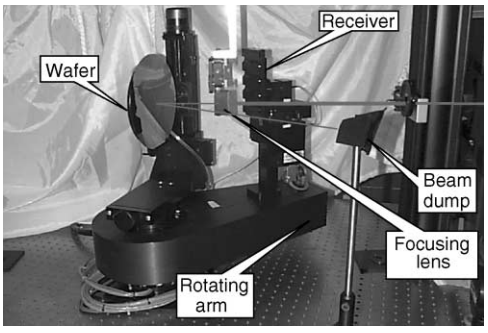


Figure 3 The author's scatterometer, which is similar to the diagram of Figure 2 is shown. In this case a final focusing lens is introduced to produce a very small illuminated spot on the silicon wafer sample. The white background was introduced to make the instrument easier to view.

Computer control of the measurement is essential to maximize versatility and minimize measurement time. The software required to control the measurement plus the display and analysis of the data can be expected to be a significant portion of total instrument development cost. The following reviews typical design features (and issues) associated with the source, sample mount and receiver components.

The source in Figure 2 is formed by a laser beam that is chopped, spatially filtered, expanded, and finally brought to a focus on the receiver path. The beam is chopped to reduce both optical and electronic noise. This is usually accomplished through the use of lock-in detection in the electronics package which suppresses all signals except those at the chopping frequency. Low noise, programmable gain electronics are essential to reducing system noise. The reference detector is used to allow the computer to ratio out laser power fluctuations and, in some cases, to provide the necessary timing signal to the lock-in electronics. Polarizers, wave plates, and neutral density filters are also commonly placed prior to the spatial filter. The spatial filter removes source scatter from the laser beam and presents a point source which is imaged by the final focusing element, in this case a mirror, to the detector zero position. Focusing the beam at this location allows near specular scatter to be more easily measured. Lasers are convenient sources, but are not necessary. Broadband sources are often required to meet a particular application or to simulate the environment where a sample will be used. Monochromators and filters can be used to provide scatterometer sources of arbitrary wavelength. The noise floors with these tunable incoherent sources increases as the spectral bandpass is narrowed, but they have the advantage that the scatter pattern does not contain laser speckle.

The sample mount can be very simple or very complex. In principal, six degrees of mechanical freedom are required to fully adjust the sample. The order in which these stages are mounted affects the ease of use (and cost) of the sample holder. In practice, it often proves convenient to either eliminate, or occasionally duplicate, some of these degrees of freedom. In addition, some of these axes may be motorized to allow the sample area to be raster-scanned to automate sample alignment or to measure reference samples. As a general rule, the scatter pattern is insensitive to small changes in incident angle but very sensitive to small angular deviations from specular. Instrumentation should be configured to allow location of the specular reflection (or transmission) very accurately. Receiver designs vary, but changeable entrance apertures, bandpass

filters, lenses, and field stops are generally positioned in front of the detector.

A serious measurement problem is getting light scattered by the instrument, called instrument signature, confused with light scattered by the sample. An example of instrument signature is shown by the dotted line in Figure 2, which represents scatter from the final mirror. The signature is often measured in the straight through (transmission) direction, multiplied by the measured specular reflectance and then compared to the measured sample BRDF. Another issue is the fact that the measured BRDF is really the convolution of the receiver aperture with the actual (incremental) BRDF. When the scatter signal varies slowly across the aperture the measurement is virtually identical to the true BRDF. Near the specular reflection, or at diffraction peaks, the differences between the measurement (or convolution) and the actual (incremental) BRDF can be huge. Measurements made using invisible sources and measurements of curved samples present additional problems. These problems and the issues of calibration and accuracy are covered in the Further Reading section at the end of this article.

TIS Instruments

The two common methods of making TIS measurements are shown in Figures 4 and 5. The first one is based on a hemispherical mirror (or Coblentz sphere) to gather scattered light from the sample and image it onto the scatter detector. The specular beam enters and leaves the hemisphere through a small circular hole. The diameter of that hole defines the near specular limit of the instrument. The reflected beam (not the incident beam) should be centered in the hole

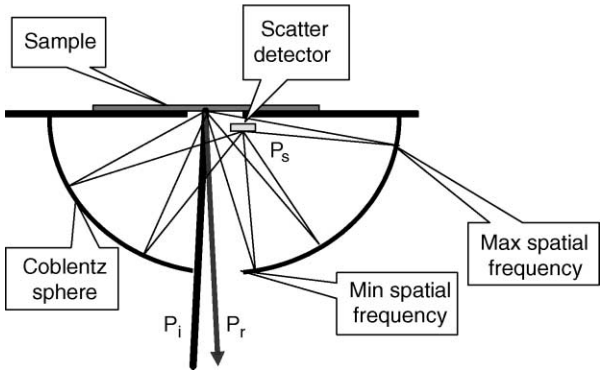


Figure 4 A diagram showing the Coblentz sphere approach to TIS measurements used in the early development of scatter instrumentation.

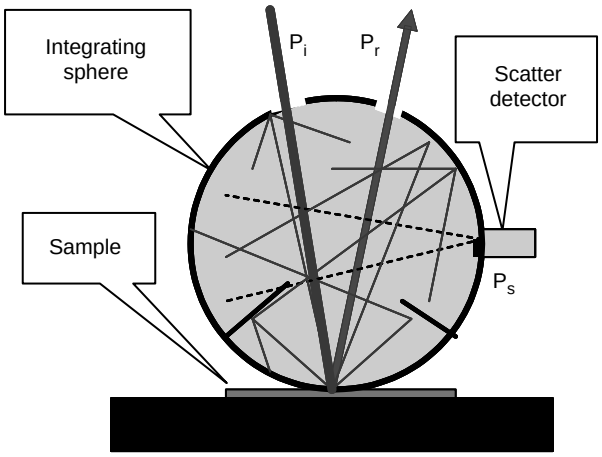


Figure 5 More modern TIS instruments make use of an integrating sphere approach which is easier to align and does not suffer from problems associated with measuring high-angle scatter from the sample.

because the BRDF will be symmetrical about it. Alignment of the hemispherical mirror is critical, and not trivial, in this approach. The second approach involves the use of an integrating sphere. A section of the sphere is viewed by a recessed detector. If the detector field of view (FOV) is limited to a section of the sphere that is not directly illuminated by scatter from the sample, then the signal will be proportional to total scatter from the sample. Again, the reflected beam should be centered on the exit hole. The Coblentz sphere method presents more signal to the detector; however, some of this signal is incident on the detector at very high angles. Thus, this approach tends to discriminate against high-angle scatter (which is not a problem for many samples). The integrating sphere is easier to align, but has a lower signal to noise ratio (less signal on the detector) and is more difficult to build in the IR where uniform diffuse surfaces are harder to obtain. A common mistake with TIS measurements is to assume that for near normal incidence, the orientation between source polarization and sample orientation is not an issue. TIS measurements made with a linearly polarized source on a grating at different orientations will quickly demonstrate this dependence.

TIS measurements can be made very near the specular reflection by utilizing a diffusely reflecting plate with a small hole in it. A converging beam is reflected off the sample and through the hole. Scatter is diffusely reflected from the diffuse plate to a receiver designed to uniformly view the plate. The reflected power is measured by moving the plate so the specular beam misses the hole and then taking that measurement. The ratio of the two scatter measurements gives the TIS. Measurements starting

closer than 0.1 degrees from specular can be made in this manner and it is an excellent way to check incoming optics or freshly coated optics for low scatter.

Analyzing Scatter from Surface Roughness

The preceding sections have concentrated on obtaining and quantifying accurate scatter data, but that leaves the question of what to do with it once you have it. In rare situations you may be given a scatter (BRDF) specification – such as, the BRDF from the mirror must be less than 10^{-4} sr^{-1} 10 degrees from specular when measured at a wavelength of 633 nm incident at 5 degrees with an S polarized source. Unfortunately this is very uncommon. If the issue is limiting scatter as a noise source, you will probably have to generate your own specification based on specific system requirements. More difficult, and often of more economic value, is the situation where scatter measurements are being used as a metrology to learn something about the sample characteristics – like roughness or defect size and/or type.

The relationship between the measured BRDF and reflector roughness statistics was a subject of intense interest from the mid-1970s through the early 1990s. Dozens of papers, and even some books, have been written on the subject, and it can only be outlined here. The relatively easy case of scatter from roughness on a clean, optically smooth, front surface reflector was first published in 1975; however it was several years later before confirming experiments were completed. The deceptively simple relationship, based on vector perturbation theory, is shown below.

$$\text{BRDF} = (16\pi^2/\lambda^4)\cos\theta_i \cos\theta_s \, QS(f_x, f_y) \qquad [4]$$

Q is the polarization factor and is determined by the material constants of the reflector, as well as the system geometry. In many cases, it is numerically about equal to the specular reflectance and this approximation is often justified. Exact expressions are available in the literature. $S(f_x, f_y)$ is the surface power spectral density function (or PSD). It may be thought of as roughness power (surface height variations squared) per unit spatial frequency (undulations per unit distance instead of per unit time). Integration of the PSD over spatial frequency space results in the mean square roughness over that band of frequencies. Taking the square root gives the root mean square roughness (or rms). Frequencies in both the x and y directions on the surface are involved and they are defined by the well-known grating equations as:

$$f_x = (\sin\theta_s \cos\phi_s - \sin\theta_i)/\lambda \qquad \text{and}$$

$$f_y = (\sin\theta_s \sin\phi_s)/\lambda \qquad [5]$$

Thus eqn [4] becomes a model for surface roughness that allows BRDF measurement to be used to find and/or verify surface roughness specifications. Of course, there are exceptions. If the 'clean, optically smooth, front surface reflector' limitations are violated, then the surface will have more than just roughness as a source of scatter and the PSD found from eqn [4] will be too large. Obtaining the same PSD from BRDF measurements made at different wavelengths, or different polarizations, is an indication that the surface is scattering 'topographically' and the PSD can be found using this technique. Because some scatter measurements can be made very rapidly, there are industry situations where scatter metrology offers a very fast means of monitoring surface quality.

A little study of eqns [4] and [5] makes it clear that ranges of spatial frequencies in the PSD correspond directly to angular ranges of the BRDF. Scatter from a single spatial frequency corresponds to a single scatter direction.

Following the above discussion, it becomes clear why the pioneering integrated (TIS) scatter measurements could be used to produce surface rms values as indicated by eqn [2]. Unfortunately, the result of eqn [4] was not available in 1961. Instead eqn [2] was derived for the special case of a smooth surface with Gaussian statistics. Nobody was thinking about spatial bandwidths and angular limits. When striking differences were found between roughness measurements made by TIS and profilometer, the Gaussian assumption became the 'whipping boy,' and TIS scatter measurements took an underserved hit. In fact, integration of eqn [4] results in the TIS result given in eqn [2] under the generally true small angle assumption that most of the scatter is close to the specular reflection. The differences make sense when the concept of spatial frequency bandwidths (or appropriate angular limits) is introduced. It becomes clear that different wavelengths, incident angles and scatter collection angles will also generate different rms values (for the same surface) and why so much confusion resulted following the definition of TIS in terms of rms roughness. There is no such thing as a unique rms roughness for a surface, anymore than there is a single spatial bandwidth for the PSD, or a single set of angles over which to integrate the BRDF. TIS measurements and rms measurements

should always be given with enough information to determine bandwidth limits.

Measuring and Analyzing Scatter from Isolated Surface Features

Understanding scatter from discrete surface features has led to big changes in the entertainment business (CDs, DVDs, digitized music, and films, etc.) as well as providing an important source of metrology for the semiconductor industry as they develop smaller faster chips for a variety of modern uses. Thus, just about everybody in the modern world utilizes our understanding of scatter from discrete surface features.

On the metrology side, the roughness signals described in the last section are a serious source of background noise that limits the size of the smallest defects that can be found. Discrete surface features come in a dazzling array of types, sizes, materials, and shapes – and they all scatter differently. A 100 nm silicon particle scatters a lot differently than a 100 nm silicon oxide particle (even if they have the same shape), and a 100 nm diameter surface pit will have a different scatter pattern. Models describing scatter from a variety of discrete surface features have been developed. Although many of these are kept confidential for competitive reasons, NIST offers some models publicly through a web site.

Confirming a model requires knowing exactly what is scattering the light. In order to accomplish this, depositions of PSLs of known size are made on the surface. Scatter is measured from one or more spheres. A second measurement of background scatter is then subtracted and the net BRDF is converted to DSC units using the known (measured) illuminated spot size. Measurements of this type have been used to confirm discrete feature scatter models. The model is then used to calculate scatter from different diameters and materials. Combined with the model for surface roughness to evaluate system noise, this capability allows signal to noise evaluation of different defect scanners designs.

Practical Industrial Instrumentation

Surface defects and particles, smaller than 100 nm are now routinely found on silicon wafers using scatter instruments known as particle scanners. Thousands of these instruments (with price tags approaching a million dollars each depending on type and use) are in daily use. These instruments are truly amazing; they inspect 200 mm wafers at a rate of one every thirty seconds reporting feature location, approximate size, and in some cases even type (pit or particle) by analyzing scatter signals that last only about 100 nanoseconds. Similar systems are now starting to be used in the computer disk industry and flat panel display inspection of surface features will follow.

Scanners are calibrated with PSLs of different sizes. Because it is impossible to identify all defect types and because diameter has little meaning for irregularly shaped objects, feature 'size' is reported in 'PSL equivalent diameters.' This leads to confusing situations for multiple detector systems, which (without some software help) would report different sizes for real defects that scatter much differently than PSLs. These difficulties have caused industry confusion similar to the 'Gaussian statistics/bandwidth limited' issues encountered before roughness scatter was understood. The publication of international standards relating to scanner calibration has reduced the level of confusion.

Scatter Related Standards

Early scatter related standards for BRDF, TIS, and PSD calculations were written for ASTM (the American Society for Testing Materials), but as the industrial need for these documents moved to the semiconductor industry, there was pressure to move the documents to SEMI (Semiconductor Equipment and Materials Inc.), which is an international organization. By 2004, the process of rewriting the ASTM documents in SEMI format was well underway in the Silicon Wafer Committee. Topics covered include: surface defect specification (M35), defect capture rate (M50), scanner specifications (M52), scanner calibration (M53), particle deposition testing (M58), BRDF measurement (ME1392), and PSD calculation (MF1811). This body of literature is probably the only place where all aspects of these related problems are brought together in one place.

Conclusion

The bottom line is that scatter measurement and analysis has moved from an art to a working metrology. Dozens of labs around the world can now take the same sample and get about the same measured BRDF from it. Thousands of industrial surface scanners employing scattered light signals are in use every day on every continent. In virtually every

house in the modern world there is at least one entertainment device that depends on scatter signals. In short – scatter works.

See also

Scattering: Scattering Theory.

Further Reading

Bennett JM and Mattsson L (1995) *Introduction to Surface Roughness and Scattering*, 2nd edn. Washington, DC: Optical Society of America.

Bennett HE and Porteus JO (1961) Relation between surface roughness and specular reflectance at normal incidence. *Journal of the Optical Society of America* 51: 123.

Cady FM, Bjork DR, Rifkin J and Stover JC (1989) BRDF error analysis. *Proceedings of SPIE* 1165: 154–164.

Church EL and Zavada JM (1975) Residual surface roughness of diamond-turned optics. *Applied Optics* 14: 1788.

Doicu A, Eremin Y and Wriedt T (2000) *Acoustic & Electromagnetic Scattering Analysis Using Discrete Sources*. San Diego, CA: Academic Press.

Gu ZH, Dummer RS, Maradudin AA and McGurn AR (1989) Experimental study of the opposition effect in the scattering of light from a randomly rough metal surface. *Applied Optics* 28(3): 537.

Nicodemus FE, Richmond JC, Hsia JJ, Ginsberg IW and Limperis T (1977) *Geometric Considerations and Nomenclature for Reflectance*. NBS Monograph 160, US Dept. of Commerce.

Scheer CA, Stover JC and Ivakhnenko VI (1998) Comparison of models and measurements of scatter from surface-bound particles. *SPIE Proceedings* 3275.

Schiff TF, Stover JC and Bjork DR (1992) Mueller matrix measurements of scattered light. *Proceedings of SPIE* 1753: 269–277.

Stover JC (1975) Roughness characterization of smooth machined surfaces by light scattering. *Applied Optics* 14(8): 1796.

Stover JC (1995) *Optical Scattering: Measurement and Analysis*, 2nd edn. Bellingham, WA: SPIE.

Stover JC (2001) Calibration of particle detection systems. In: Alain Diebold (ed.) *Handbook of Silicon Semiconductor Metrology*, chap. 21. New York: Marcel Dekker.

Wolfe WL and Bartell FO (1982) Description and limitations of an automated scatterometer. *Proceedings of SPIE* 362-30.

Spectrometers

K A More, SenSyTech, Inc. Imaging Group, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Sp7ectrometers were developed after the discovery that glass prisms disperse light. Later, it was discovered that diffraction from multiple, equally spaced, wires or fibers also dispersed light. About 100 years ago, Huygens proposed his wave theory of light and Fraunhofer developed diffraction theory, which allowed scientific development of diffraction gratings.

These discoveries then led to the development of spectrometers. Light theory was sufficiently developed such that spectrometer designs, developed over 100 years ago, are still being used today. These theories and designs are briefly described along with comments on how current technology has improved upon these designs. This is followed by some examples of imaging spectrometers which have wide spectral coverage from 450 nm to 14 μm and produce images of more than 256 spectral bands.

The basic elements of a spectroscopic instrument are shown in Figure 1. The source, or more usually an image of the source, fills an entrance slit and the radiation is collimated by either a lens or mirror. The radiation is then dispersed, by either a prism or a grating, so that the direction of propagation of the radiation depends upon its wavelength. It is then brought to a focus by a second lens or mirror and the spectrum consists of a series of monochromatic images of the entrance slit. The focused radiation is detected, either by an image detector such as a photographic plate, or by a flux detector such as a photomultiplier, in which case the area over which the flux is detected is limited by an exit slit. In some cases the radiation is not detected at this stage, but passes through the exit slit to be used in some other optical system. As the exit slit behaves as a monochromatic source, the instrument can be regarded as a wavelength filter and is then referred to as a monochromator.

Prisms

The wavelength dependence of the index of refraction is used in prism spectrometers. Such an optical

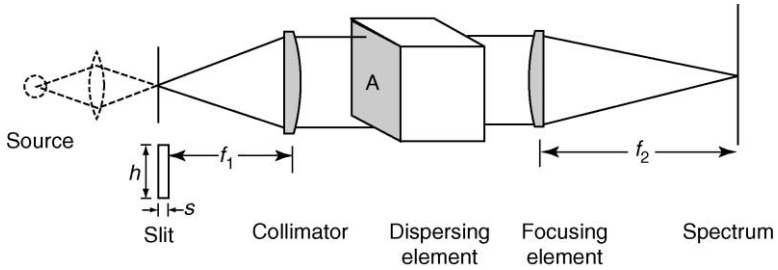


Figure 1 The basic elements of a spectroscopic instrument. With permission from Hutley MC (1982) *Diffraction Gratings*, pp. 57–232. London: Elsevier.

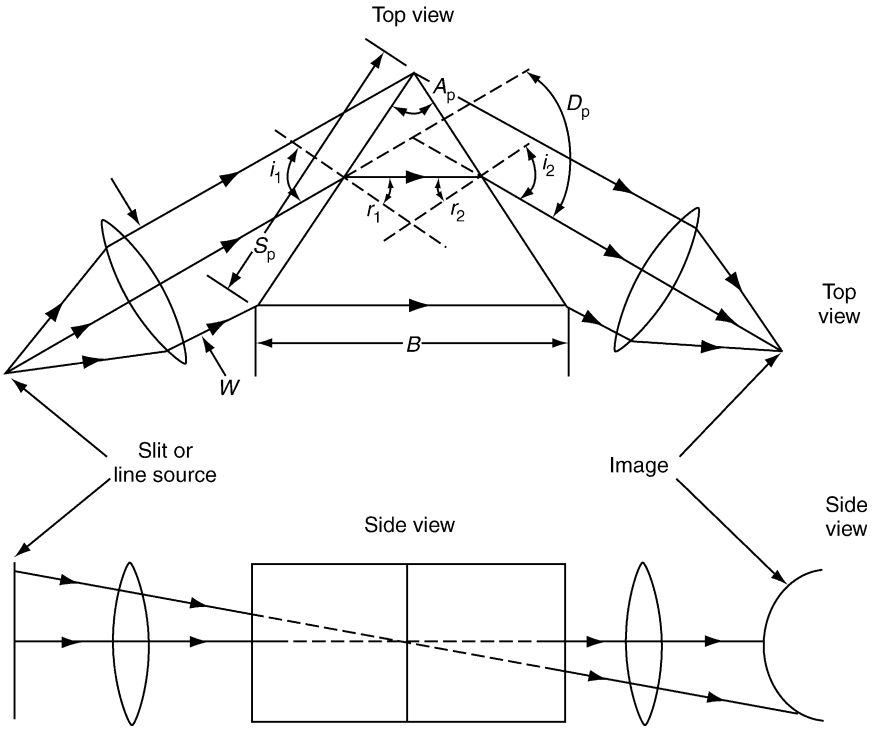


Figure 2 Elementary prism spectrometer schematic. W is the width of the entrance beam; S_p is the length of the prism face; and B is the prism base length. Reproduced with permission from *The Infrared Handbook* (1985). Ann Arbor, MI: Infrared Information Analysis Center.

element disperses parallel rays or collimated radiation into different angles from the prism according to wavelength. Distortion of the image of the entrance slit is minimized by the use of planewave illumination. Even with planewave illumination, the image of the slit is curved because not all of the rays from the entrance slit can traverse the prism in its principal plane. The prism is shown in the position of minimum angular deviation of the incoming rays in Figure 2. At minimum angular deviation, maximum power can pass through the prism. For a prism adjusted to the position of minimum deviation:

$$r_1 = r_2 = A_p/2 \qquad [1]$$

and

$$i_1 = i_2 = (D_p + A_p)/2 \qquad [2]$$

where

D_p = angle of deviation
A_p = angle of prism
r_1 and r_2 = internal angles of refraction
i_1 and i_2 = angles of entry and exit.

The angle of deviation, D_p, varies with wavelength. The resulting angular dispersion is defined as $dD_p/d\lambda$, while the linear dispersion is $dx/d\lambda = F\, dD_p/d\lambda$, where F is the focal length of the camera or imaging lens and x is the distance across the image plane.

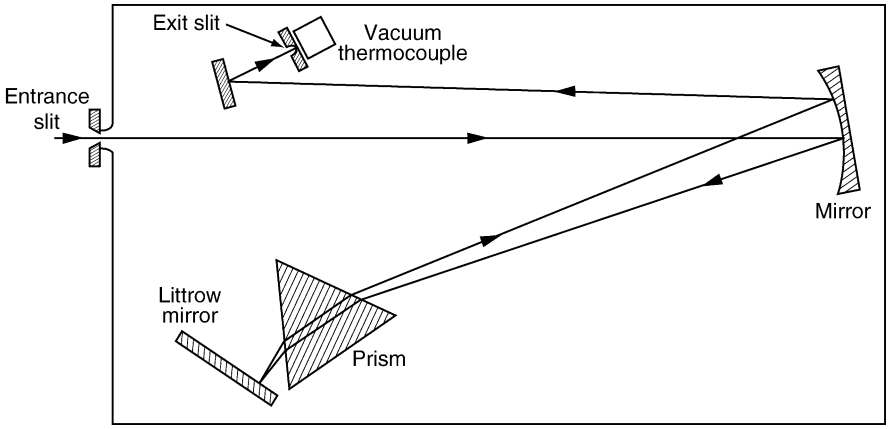


Figure 3 Infrared spectrograph of the Littrow-type mount with a rock salt prism. Reproduced with permission from *The Infrared Handbook* (1985) Ann Arbor, MI: Infrared Information Analysis Center.

It can be shown that:

$$dD_p/d\lambda = [B/W][dn/d\lambda] = [dD_p/dn][dn/d\lambda] \quad [3]$$

where

B = base length of the prism,
W = width of the illumination beam
n = index of refraction

while

$$dx/d\lambda = F[B/W][dn/d\lambda] \quad [4]$$

One may define the resolving power, RP, of an instrument as the smallest resolvable wavelength difference, according to the Rayleigh criterion, divided into the average wavelength in that spectral region. Thus:

$$RP = \lambda/\Delta\lambda = [\lambda/dD_p][dD_p/d\lambda]$$

$$= [\lambda/dD_p][B/W][dn/d\lambda] \quad [5]$$

The limiting resolution is set by diffraction due to the finite beamwidth, or effective aperture of the prism, which is rectangular. Thus:

$$RP = [\lambda/(\lambda/W)][B/W][dn/d\lambda] \quad [6]$$

so that:

$$RP = B[dn/d\lambda] \quad [7]$$

If the entire prism face is not illuminated, then only the illuminated base length must be used for B.

Littrow showed that aberrations would be minimized by making the angle of incidence equal to the angle of refraction β (also known as the Littrow Configuration). Littrow used a plane mirror behind the prism for a double pass through the prism, as shown in Figure 3.

Gratings

Rowland is credited with the development of a grating mount that reduced aberrations in the spectrogram. He found that a grating ruled on a concave surface of radius R and locating the entrance and exit slits on the same radius would give the least aberrations, as shown in Figure 4. $r = R\cos\alpha$ is the distance to the entrance slit *(r)* and $r_1 = R\cos\beta$ is the distance to the focal point for the exit slit (r_1). Rowland showed that:

$$\cos\alpha/R - \cos^2\alpha/r + \cos\beta/R - \cos^2\beta/r_1 = 0 \quad [8]$$

One solution to this equation is for α and β to each be zero and

$$r = R\cos\alpha$$

and

$$r_1 = R\cos\beta$$

This condition is met if r and r_1 lie on the Rowland circle.

There are various ways in which the Rowland circle condition may be satisfied and some of them are shown in Figure 5. The simplest mounting of all is that due to Paschen and Runge (Figure 5a) in which the entrance slit is positioned on the Rowland circle and a photographic plate (or plates) is constrained to fit the Rowland circle. Alternatively, for photoelectric detection, a series of exit slits is arranged around the Rowland circle each having its own detector. In the latter case, the whole spectrum is not recorded, only a series of predetermined wavelengths, but when used in this 'polychromator' form, it is very rugged and

convenient for applications such as the routine analysis of samples of metals and alloys. In this case, slits and detectors are set up to measure the light in various spectral lines, each characteristic of a particular component or trace element.

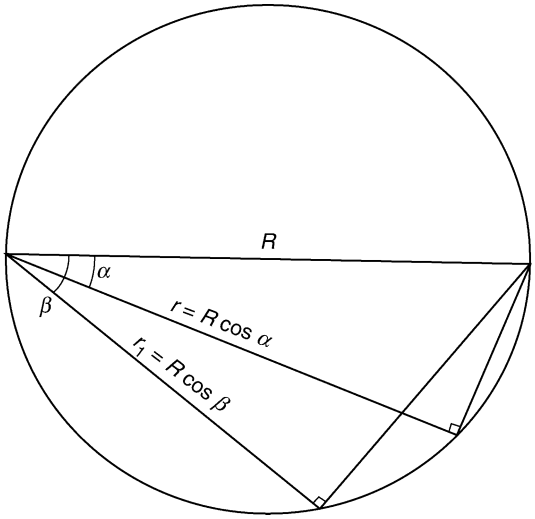


Figure 4 The construction of the Rowland Circle. With permission from Hutley MC (1982) *Diffraction Gratings*, pp. 57–232. London: Elsevier.

The oldest concave grating mount is that designed by Rowland himself and which bears his name (Figure 5b). In this case, the grating and photographic plates are fixed at opposite ends of the diameter of the Rowland circle by a moveable rigid beam. The entrance slit remains fixed above the intersection of two rails at right angles to each other and along which the grating plate holder (or the exit slit) is free to move. In this way, this entrance slit, grating, and plate holder are constrained always to lie on the Rowland circle and it has the advantage that the dispersion is linear, which is useful in the accurate determination of wavelengths. Unfortunately, this mounting is rather sensitive to small errors in the position of the entrance slit and in the orthogonality of the rails, and is now very rarely used.

A variation on this mounting was devised by Abney, who again mounted the grating and the plate holder on a rigid bar at opposite ends of the diameter of the Rowland circle (Figure 5c). The entrance slit is mounted on a bar of length equal to the radius of the Rowland circle. In this way, the slit always lays on the Rowland circle, but it had to rotate about its axis in order that the jaws should remain perpendicular to the line from the grating to

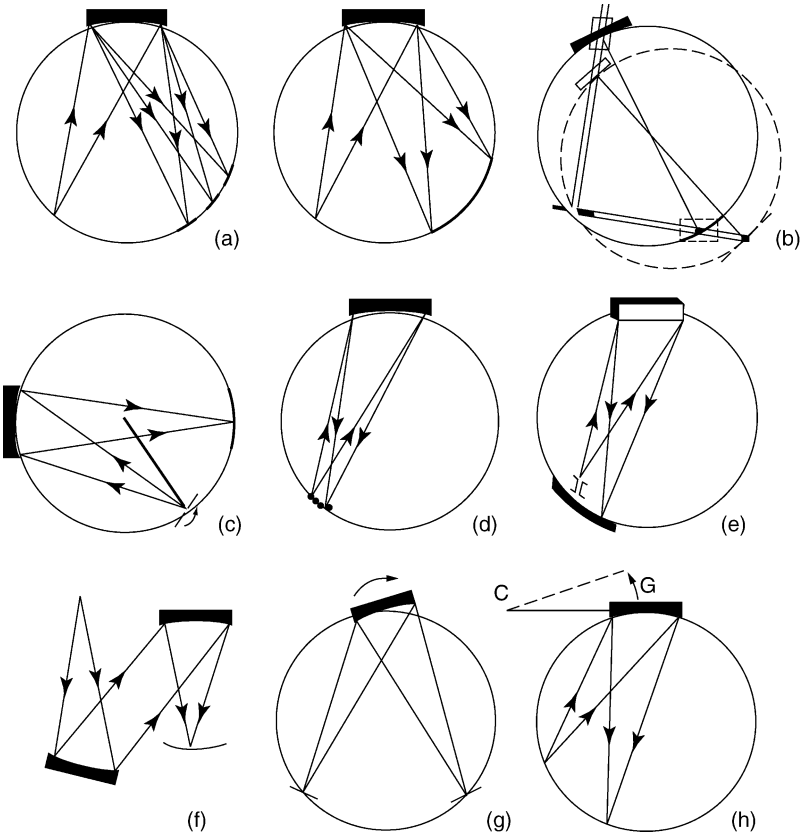


Figure 5 Mountings of the concave grating: (a) Paschen–Runge; (b) Rowland; (c) Abney; (d) and (e) Eagle; (f) Wadsworth; (g) Seya–Namioka; (h) Johnson–Onaka. With permission from Hutley MC (1982) *Diffraction Gratings*, pp. 57–232. London: Elsevier.

the slit. It also had the disadvantage that the source must move with the exit slit, which could be inconvenient.

In the Eagle mounting (Figures 5d and 5e), the angles of incidence and diffraction are made equal, or very nearly so, as with Littrow mounting for plane gratings. Optically, this system has the advantage that the astigmatism is generally less than that of the Paschen–Runge or Rowland mounting, but on the other hand, the dispersion is nonlinear. From a mechanical point of view, it has the disadvantage that it is necessary with great precision both to rotate the grating and to move it nearer to the slits in order to scan the spectrum. However, it does have the practical advantage that it is much more compact than other mountings, and this is of particular importance when we bear in mind the need to enclose the instrument in a vacuum tank. Ideally, the entrance slit and exit slit or photographic plate should be superimposed if we are to set $\alpha = \beta$. In practice, of course, the two are displaced either sideways, in the plane of incidence as shown in Figure 5, or out of the plane of incidence, in which case the entrance slit is positioned below the meridonal plane and the plate holder just above it, as shown in Figure 5. The out-of-plane configuration is generally used for spectrographs and the in-plane system for monochromators. The penalty incurred in going out of plane is that coma is introduced in the image, and slit curvature becomes more important. This limits the length of the ruling that can effectively be used.

The second well-known solution to the Rowland equation is the Wadsworth mounting (Figure 5f), in which the incident light is collimated, so r is set at infinity and the focal equation reduces to $r_1 = R\cos^2\beta/(\cos\alpha + \cos\beta)$. One feature of this mounting is that the astigmatism is zero when the image is formed at the center of the grating blank, i.e., when $\beta = 0$. It is a particularly useful mounting for applications in which the incident light is naturally collimated (for example, in rocket or satellite astronomy, spectroheliography and in work using synchrotron radiation). However, if the light is not naturally collimated, the Wadsworth mount requires a collimating mirror, so one has to pay the penalty of the extra losses of light at this mirror. The distance from the grating to the image is about half that for a Rowland circle mounting which makes the instrument more compact and, since the grating subtends approximately four times the solid angle, there is a corresponding increase in the brightness of the spectral image.

Not all concave grating mountings are solutions to the Rowland equation. In some cases, other advantages may compensate for a certain defect of focus. A particularly important example of this is the Seya–Namioka mounting (Figure 5g), in which the entrance slit and exit slit are kept fixed and the spectrum is scanned by a simple rotation of the grating (Figure 5g). In order to achieve the optimum conditions for this mounting, we set $\alpha = \varphi + \theta$ and $\beta = \theta - \varphi$, where 2φ is the angle subtended at the grating by the entrance and exit slit and θ is the angle through which the grating is turned. The amount of defocus is given by:

$$F(\theta, \varphi, r, r_1) = [\cos^2(\theta + \varphi)/r] + [\cos(\theta + \varphi)/R] + [\cos^2(\theta - \varphi)/r_1] + [\cos(\theta - \varphi)/R] \quad [9]$$

and the optimum conditions are those for which $F(\theta,\varphi,r,r_1)$ remains as small as possible as θ is varied over the required range. Seya set F and three derivatives of F with respect to θ equal to zero for $\theta = 0$ and obtained the result:

$$\varphi = \sin^{-1}(1/\sqrt{3}) = 35°15' \quad [10]$$

and

$$r = r_1 = R\cos\varphi \quad [11]$$

which corresponds to the Rowland circle either in zero order or for zero wavelength. In practice, it is usual to modify the angle slightly so that the best focus is achieved in the center of the range of interest rather than at zero wavelength.

The great advantage of the Seya–Namioka mounting is its simplicity. An instrument need consist only of a fixed entrance and exit slit, and a simple rotation of the grating is all that is required to scan the spectrum. It is, in fact, simpler than instruments using plane gratings. Despite the fact that at the ends of the useful wavelength range the resolution is limited by the defect of focus and the astigmatism is particularly bad, the Seya–Namioka mounting is very widely used, particularly for medium-resolution rather than high-resolution work.

A similar simplicity is a feature of the Johnson–Onaka mounting (Figure 5h). Here again, the entrance and exit slits remain fixed, but the grating is rotated about an axis which is displaced from its center; in this way, it is possible to reduce the change of focus that occurs in the Seya–Namioka mounting. The system is set up so that at the center of the desired wavelength range, the slits and grating lie on the Rowland circle, as shown in Figure 5h. The optimum radius of rotation, i.e., the distance GC, was found by

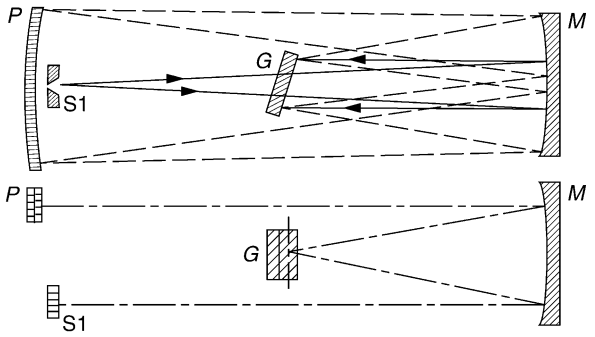


Figure 6 Ebert mounting of the plane grating designed by Fastie. 'SI' is the entrance slit; *G* is the grating; *M* is the concave mirror; and *P* is the photographic plate. The horizontal section is at the top and the vertical section is at the bottom. Reproduced with permission from *The Infrared Handbook* (1985) Ann Arbor, MI: Infrared Information Analysis Center.

Onaka to be:

$$GC_{opt} = [R \sin\tfrac{1}{2}(\alpha+\beta)]/[1 - \tfrac{1}{2}\tan(\alpha+\beta) \times (\tan\beta - \tan\alpha)] \quad [12]$$

Another non-Rowland spectrometer is the Ebert–Fastie mount, which mounts the slit, flat grating, and a detector array in the arrangement shown in Figure 6. Ebert first developed the design using two separate concave mirrors, one for collimating the incident beam and the second to focus the diffracted spectrum. Fastie used a single but larger concave mirror, which simplified the mounting structure and produced a rugged, compact spectrometer that has been used in rocket flights and space satellite observatories for astronomical and upper atmospheric applications. The Czerny–Turner mount is similar to the Ebert mount, except the flat grating is located in the same plane that contains the entrance and exit slits.

Advanced Spectrometers

While the above spectrometer designs are still used, major advances in implementation are now available. Ray tracing allows the designer to quantify the aberrations and determine solutions to remove them. Aspheric optic elements can now be fabricated to correct aberrations. Gratings can be ruled on aspherical surfaces to not only disperse the light beam, but also be an element in the optical design. Holography has been developed to etch gratings for use over a wide spectral range and to not only disperse the light, but also work as an optical element. Because holographic gratings are chemically etched, there are no machine burrs to scatter light and the hologram is free of periodic differences in groove widths that create ghosts in ruled gratings. Linear and array photodetectors have replaced film. An advantage of film was its ability to fit curved focal planes. Photodetectors are etched into flat wafers of the photodiode material. To adapt flat arrays to a curved focal plane, fiber optic face plate couplers have been ground on one side to match the focal plane curvature and the flat back side is either optically coupled to the detector or closely coupled for proximity focusing. The photodetector arrays are available in multiple materials to cover the spectral range, from soft X-rays to the thermal infrared. Often cooling is required to obtain low noise. They are now available with sensitivities never reached with film and the ease of coupling the output data to a computer for real-time analysis has taken much of the labor out of analyzing the spectrograph. Another advance is precision motor drives with precision motion sensors, some using laser interferometry, to provide feedback on automated movement of gratings, slits, and detectors as a spectrometer cycles through its wavelength range.

Imaging Spectrometers

Spectrometers that image the spectral characteristics in each pixel of the image to form a data cube, as shown in Figure 7, are named imaging spectrometers, and if the spectra have high resolution and blocks of consecutive neighbors, then the data are called hyperspectral. Imaging spectrometers have several applications that range from medical to remote sensing of land use and environment. This article will cover examples for remote sensing imaging spectrometers which require spectral coverage in all of the atmospheric windows from the UV to thermal IR.

Imaging spectrometers require a combination of spectrometers, light collecting optics, and scan mechanisms to scan the instantaneous field of view of the spectrometer over a scene. Remote sensing of the Earth applications requires an aerial platform. Either a helicopter, an aircraft, or orbital space satellites are used. The platform motion is used as part of the scanning process so that the optics image a single point on the ground so that a scanner (called a line scanner) scans a long line that is cross tracked to the platform motion. Or, the optics image a slit that is parallel to the platform track that covers many scan lines and a scanner moves the slit cross track to the platform motion. This scanner is called a whisk-broom scanner. Or, the optics image a large slit so no scan mechanism is needed other than the platform

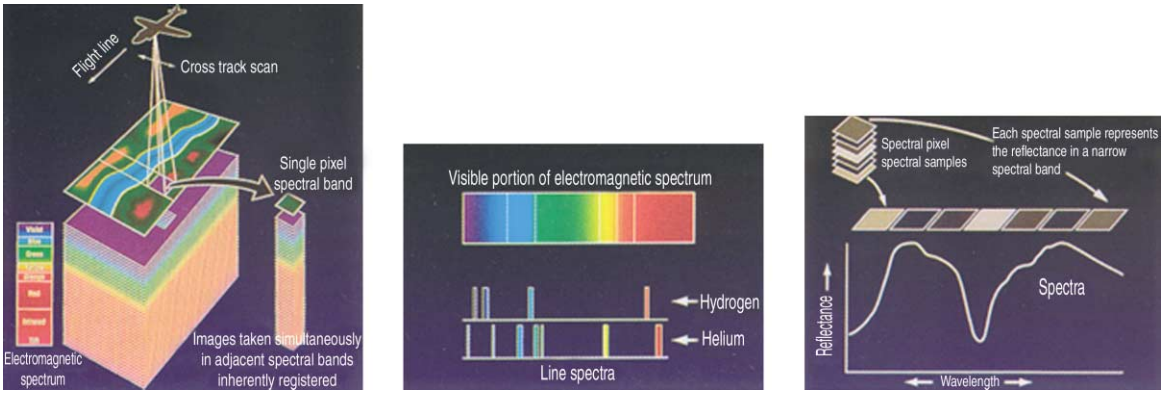


Figure 7 Hyperspectral Data Cube. Hyperspectral imagers divide the spectrum into many discrete narrow channels. This fine quantization ofn spectral information on a pixel by pixel basis enables researchers to discriminate the individual constituents in an area much more effectively. For example, the broad spectral bands of a multispectral sensor allow the user only to coarsely discriminate between areas of deciduous and coniferous forest, plowed fields, etc., whereas a hyperspectral imager provides characteristic signatures which can be correlated with specific spectral templates to help determine the individual constituents and possibly even reveal details of the natural processes which are affecting them.

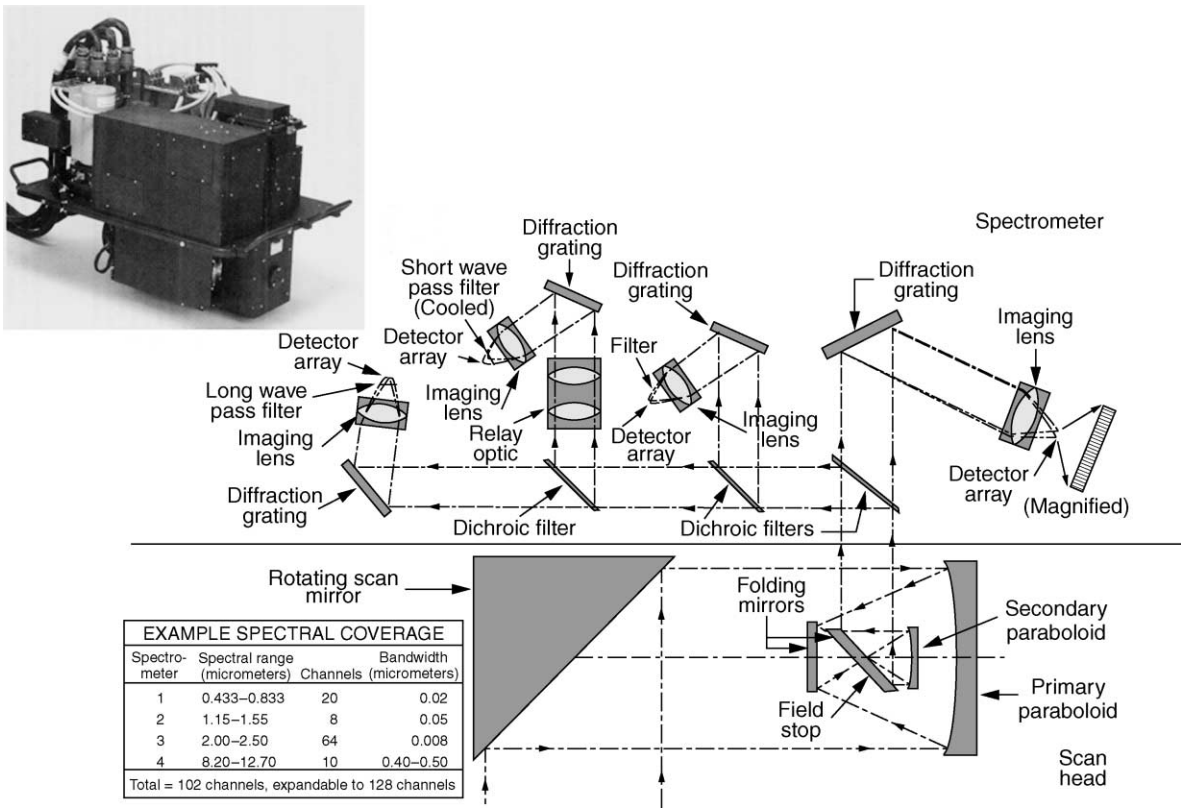


Figure 8 Multispectral infrared and visible imaging spectrometer optical schematic.

motion to form an image. This scanner is called a pushbroom scanner.

One important requirement is that all spectral measurements of a pixel be coregistered. Most airborne imaging spectrometers use a common aperture for a line scanner, or a slit for whisk and pushbroom scanners, so that platform instability from pitch roll and yaw and from inability to fly in a straight line do not compromise the coregistration of spectral data on each pixel. The image data may require geometric correction, but the spectral data are not compromised.

There is one other type of imaging spectrometer that uses a linear variable filter (also known as a wedge filter) over a 2D array of photodetectors. Each image contains a different spectral band over each row of pixels so that each frame images a scene with a different spectral band over each row of pixels. The array is oriented so the rows of spectral bands are perpendicular to the flight track. After the platform motion moves over one ground pixel, the whole array is read out and the frame is shifted one row of pixels so the second frame adds a second spectral band to each row imaged in the first frame. This frame stepping is carefully timed to the platform velocity and is repeated until each row of pixels is imaged in all spectral bands. This type of imaging spectrometer has been flown in aircraft and satellites. In aircraft, the platform motion corrupts the coregistration of the spectrum in each pixel. Extensive ground processing is required to geometrically correct each frame to improve spectral coregistration. This is a difficult task and this type of imaging spectrometer has lost favor for airborne use. Stabilized satellites have been a better platform and the wedge type of imaging spectrometer has been used successfully in space.

As examples of airborne and ground-based imaging spectrometers, a line scanner imaging

Table 1 MIVIS physical properties

	Height		*Width*		*Depth*[a]		*Weight*	
	in	*cm*	*in*	*cm*	*in*	*cm*	*lbs*	*kg*
Scan head	26.5	67.0	20.6	52.0	28.1	71.0	220	100
Electronics	40.0	102.0	19.0	48.3	24.0	61.0		
Total system weight (approx.)							460	209

[a]Not including connectors or cable bends.

Table 2 MIVIS spectral coverage (μm)

OPTICAL PORT 1			*OPTICAL PORT 3*						*OPTICAL PORT 4*		
CH #	*Band Edges*		*CH #*	*Band Edges*		*CH #*	*Band Edges*		*CH #*	*Band Edges*	
	Lower	*Upper*		*Lower*	*Upper*		*Lower*	*Upper*		*Lower*	*Upper*
1	0.43	0.45	29	2.000	2.008	61	2.250	2.258	93	8.20	8.60
2	0.45	0.47	30	2.008	2.016	62	2.258	2.266	94	8.60	9.00
3	0.47	0.49	31	2.016	2.023	63	2.266	2.273	95	9.00	9.40
4	0.49	0.51	32	2.023	2.031	64	2.273	2.281	96	9.40	9.80
5	0.51	0.53	33	2.031	2.039	65	2.281	2.289	97	9.80	10.20
6	0.53	0.55	34	2.039	2.047	66	2.289	2.297	98	10.20	10.70
7	0.55	0.57	35	2.047	2.055	67	2.297	2.305	99	10.70	11.20
8	0.57	0.59	36	2.055	2.063	68	2.305	2.313	100	11.20	11.70
9	0.59	0.61	37	2.063	2.070	69	2.313	2.320	101	11.70	12.20
10	0.61	0.63	38	2.070	2.078	70	2.320	2.328	102	12.20	12.70
11	0.63	0.65	39	2.078	2.086	71	2.328	2.336			
12	0.65	0.67	40	2.086	2.094	72	2.336	2.344			
13	0.67	0.69	41	2.094	2.102	73	2.344	2.352			
14	0.69	0.71	42	2.102	2.109	74	2.352	2.359			
15	0.71	0.73	43	2.109	2.117	75	2.359	2.367			
16	0.73	0.75	44	2.117	2.125	76	2.367	2.375			
17	0.75	0.77	45	2.125	2.133	77	2.375	2.383			
18	0.77	0.79	46	2.133	2.141	78	2.383	2.391			
19	0.79	0.81	47	2.141	2.148	79	2.391	2.398			
20	0.81	0.83	48	2.148	2.156	80	2.398	2.406			
OPTICAL PORT 2			49	2.156	2.164	81	2.406	2.414			
21	1.15	1.20	50	2.164	2.172	82	2.414	2.422			
22	1.20	1.25	51	2.172	2.180	83	2.422	2.430			
23	1.25	1.30	52	2.180	2.188	84	2.430	2.438			
24	1.30	1.35	53	2.188	2.195	85	2.438	2.445			
25	1.35	1.40	54	2.195	2.203	86	2.445	2.453			
26	1.40	1.45	55	2.203	2.211	87	2.453	2.461			
27	1.45	1.50	56	2.211	2.219	88	2.461	2.469			
28	1.50	1.55	57	2.219	2.227	89	2.469	2.477			
			58	2.227	2.234	90	2.477	2.484			
			59	2.234	2.242	91	2.484	2.492			
			60	2.242	2.250	92	2.492	2.500			

spectrometer and a ground-based scanned push broom scanner are described below.

MIVIS (multispectral infrared and visible imaging spectrometers) hyperspectral scanner developed by SenSyTech Imaging Group (formerly Daedalus Enterprises) for the CNR (Consiglio National Researche) of Italy is a line scanner imaging spectrometer. The optical schematic is shown in Figure 8. A rotating 45° mirror scans a line of pixels on the ground. Each pixel is collimated by a parabolic mirror in a Gregorian telescope mount. The collimated beam is reflected by a Pfund assembly to an optical bench that houses four spectrometers. An aperture pinhole in the Pfund assembly defines a common instantaneous field of view for each pixel. The collimated beam is then split off with either thin metallic coated or dielectric coated dichroics to four spectrometers. The thin metallic mirrors reflect long wavelengths and transmit short wavelengths. The multiple dielectric layers cause interference such that light is reflected at short wavelengths and transmitted at long wavelengths. After splitting off four wide bands (visible, near infrared, mid-wave infrared, and thermal infrared) each band is dispersed in its own spectrometers.

The wavelength coverage for each spectrometer is based on the wavelength sensitivity of different photodetector arrays. The visible spectrometer uses a silicon photodiode array; the near infrared spectrometer, an InGaAs array, the mid-infrared a InSb array and the thermal infrared, a MCT (mercury doped cadmium teluride) photo-conductor array. Note the beam expander in spectrometer 3 for the

Table 3 LAFS technical specifications

Spatial resolution	1.0 mrad IFOV vertical and horizontal (square pixels). Optical lens for 0.5 mrad IFOV
Spatial coverage	15° TFOV horizontal, nominal. A second, 7.5° TFOV, is an option
Focus range	5′ to infinity
Spectral resolution	5.0 nm per spectral channel, sampled at 2.5 nm interval
Spectral range	400–1100 nm
Dynamic range	8 bit (1 part in 256)
Illumination	Solar illumination ranging from 10:00 a.m. to 2:00 p.m. under overcast conditions to full noon sunshine
Acquisition time	1.5 sec for standard illuminations. Option for longer times for low light level conditions
Viewfinder	Near real time video view of the scene
Calibrations	Flat fielding to compensate for CCD variations in responsivity. Special calibration
Data displays	Single band imagery, selectable Waterfall Chart (one spatial line by spectral, as acquired by the CCD). Single pixel spectral display
Data format	Convertible to format compatible with image processors
Data storage	Replaceable hard disc, 14 data cubes/disk

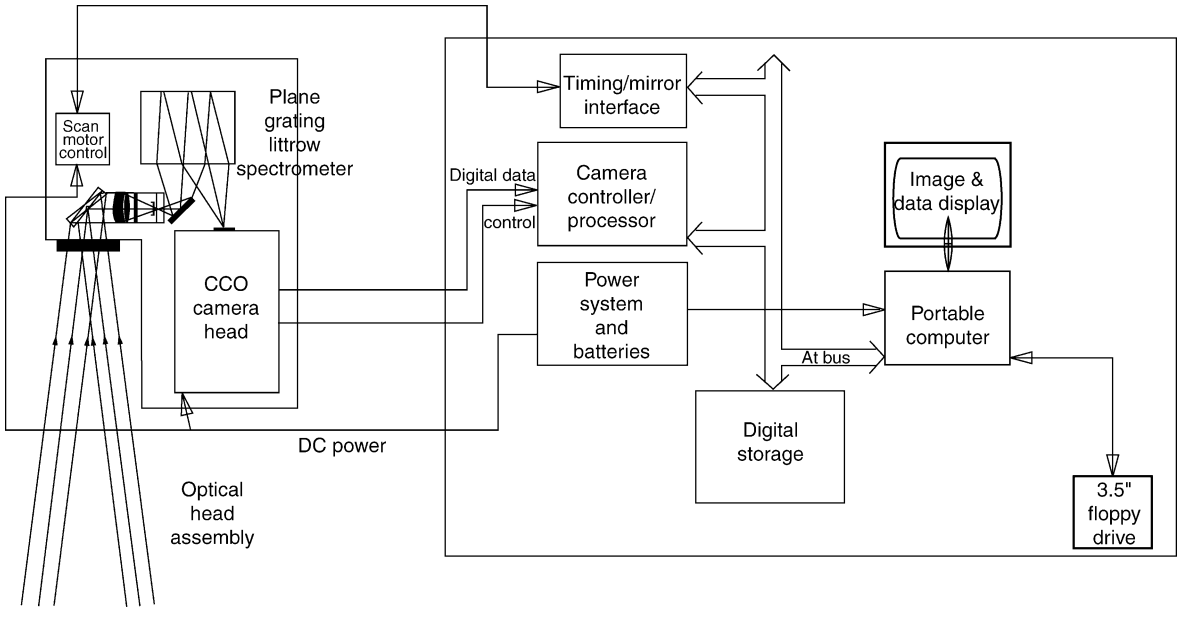


Figure 9 LAFS system block diagram.

mid-infrared (IR). Since wavelength resolution increases with the size of the illuminated area of a diffraction grating a beam expander was needed to meet the specified mid-IR resolution.

The physical characteristics of MIVIS are given in Table 1 and the spectral coverage in Table 2.

An example of a push broom imaging spectrometer is the large area fast spectrometer (LAFS), which is a field portable, tripod mounted, imaging spectrometer. LAFS uses a push broom spectrometer with a galvanometer driven mirror in front of the entrance slit. The mirror is stepped to acquire a 2D image $\times$ 256 spectral bands data cube.

LAFS was developed for the US Marine Corps under the direction of the US Navy Coastal Systems Station of the Dahlgren Division. The technical specifications are shown in Table 3. Various concepts for the imaging spectrometer were studied and the only concept that could meet the above specification was a Littrow configuration grating spectrometer that imaged one line in the scene and dispersed the spectra perpendicular to the line image onto a CCD array. A galvanometer mirror stepped the line image over the scene to generate a spectral image of the scene. A block diagram of LAFS is shown in Figure 9 and a photograph of the prototype in Figure 10. Figure 11 is a photograph of the optical head. Data from a single frame are stored in computer (RAM) memory then transferred to a replaceable hard disk for bulk storage. Data are collected in a series of 256 CCD frames and each CCD frame has a line of 256 pixels along one axis and 256 spectral samples along the second axis. The 256 CCD frames constitute a data cube as shown in Figure 12. This does not allow field viewing of an image in a single spectral band. Earth View™ software is used in the field portable computer



Figure 10 LAFS.

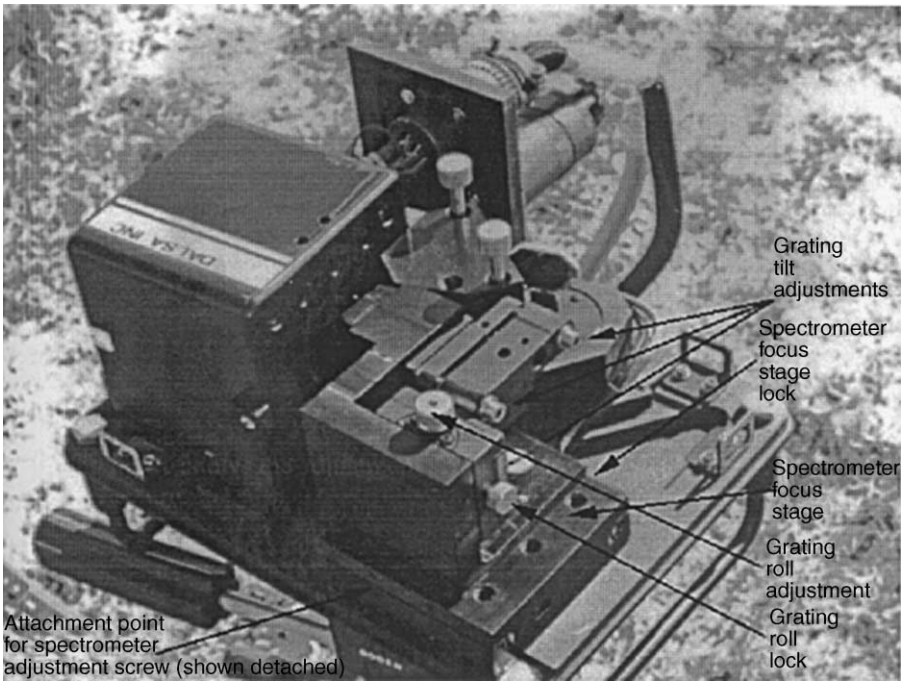


Figure 11 LAFS optical head.

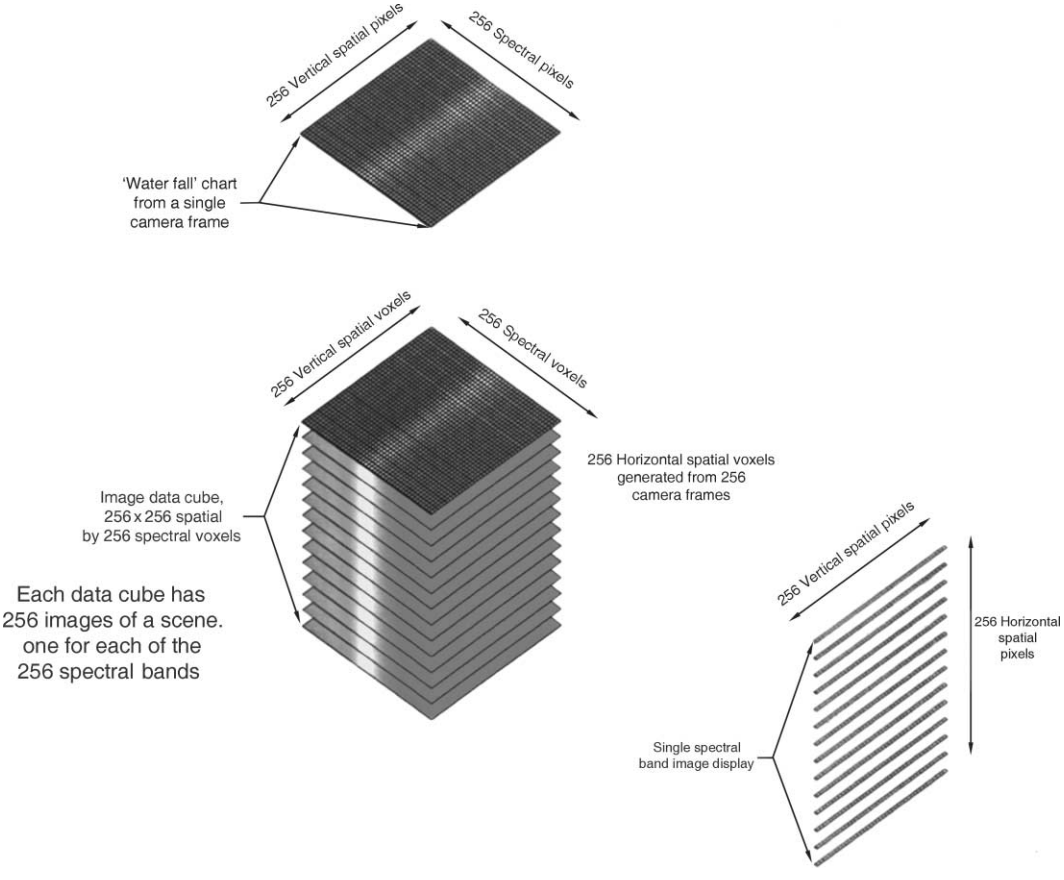


Figure 12 Arrangement of image memory and construction of image displays.

Table 4 LAFS physical properties

Dimensions and weight	*Size (inches)*	*weight (lbs)*
Optical head	7 × 9 × 8	11
Electronics	12 × 15.5 × 9	30
Battery	4.5 × 16.5 × 2	7
		48 + cables

Power: Less than 150 W at 12 VDC; Battery time – 45 min/charge.

to reorder the data cube into 256 spatial images, one for each spectral band as shown in Figure 12.

The prototype is packaged in two chassis, a compact optical head and a portable electronic chassis. The size, weight, and power are shown in Table 4.

LAFS was designed to also work as an airborne imaging spectrometer. The framing mirror can be locked so it views a scene directly below the aircraft (push broom mode) or the mirror can be programmed to sweep a 256 lines cross track to the aircraft flight path (whisk broom mode). In the whisk broom mode, the scan mirror rotation arc can be increased for a wider field of view than in the push broom mode. LAFS illustrates the decrease in size, weight, and power of a push broom imaging spectrometer that results from a 256 × 256 pixel array rather than single pixel line arrays of a line scanner. The array increases integration time of a line by a factor of 256, which allows longer detector integration dwell time on each pixel and thus smaller light collecting optics.

List of Units and Nomenclature

Data cube	Multiple images of a scene in many spectral bands. The spectra of each pixel can be obtained by plotting the spectral value on the same pixel in each spectral band image.
Pixel	An image point (or small area) defined by the instantaneous-field-of-view of either the optics, entrance aperture or slit, and the detector area, or by some combination of these components. A pixel is the smallest element in a scene that is resolved by the imaging system.
Instantaneous-field-of-view (IFOV)	The IFOV is used in describing scanners to define the size of the smallest field-of-view (usually in milliradians or microradians) that be resolved by a scanner system.
Field-of-view	The total size in angular dimensions of a scanner or imager.
Hyperspectral	A data cube with many (>48) spectral bands.
Line scanners	An optical/mechanical system that scans one pixel at a time along a line of a scene.
Whisk broom	An optical/mechanical system that scans two or more lines at a time.
Push broom	An optical system that images a complete line at a time.

See also

Diffraction: Diffraction Gratings; Fraunhofer Diffraction. **Fiber Gratings**. **Geometrical Optics:** Prisms. **Imaging:** Hyperspectral Imaging; Interferometric Imaging. **Interferometry:** Overview. **Modulators:** Acousto-Optics. **Optical Materials:** Color Filters and Absorption Glasses. **Semiconductor Materials:** Dilute Magnetic Semiconductors; Group IV Semiconductors, Si/SiGe; Large Gap II–VI Semiconductors; Modulation Spectroscopy of Semiconductors and Semiconductor Microstructures. **Spectroscopy:** Fourier Transform Spectroscopy; Raman Spectroscopy.

Further Reading

Baselow R, Silvergate P, Rappaport W, *et al.* (1992) *HYDICE (Hyperspectral Digital Imagery Collection Experiment Instrument Design*. International Symposium on Spectral Sensing Research.

Bianchi, R, Marino, CM and Pignatti, S *Airborne Hyperspectral Remote Sensing in Italy*. Pomezia (Roma), Italy: CNR, Progetto LARA, 00040.

Borough HC, Batterfield OD, Chase RP, Turner RW and Honey FR (1992) *A Modular Wide Field of View Airborne Imaging Spectrometer Preliminary Design Concept*. International Symposium on Spectral Sensing Research.

Carmer DC, Horvath R and Arnold CB (1992) *M7 Multispectral Sensor Testbed Enhancements*. International Symposium on Spectral Sensing Research.

Hackwell JA, Warren DW, Bongiovi RP, *et al.* (1996) LWIR/MWIR Imaging Hyperspectral Sensor for Airborne and Ground-Based Remote Sensing. In: Descour M and Mooney J (eds) *Imaging Spectrometry II*. Proc. SPIE 2819, pp. 102–107.

Hutley MC (1982) *Diffraction Gratings*. London: Academic Press.

James JF and Sternberg RS (1969) *The Design of Optical Spectrometers*. London: Chapman and Hall.

Kulgein NG, Richard SP, Rudolf WP and Winter EM (1992) *Airborne Chemical Vapor Detection Experiment*. International Symposium on Spectral Sensing Research.

Lucey PG, Williams T, Horton K, Hinck K and Buchney C (1994) *SMIFTS: A Cryogenically Cooled, Spatially Modulated, Imaging Fourier Transform Spectrometer for Remote Sensing Applications*. International Symposium on Spectral Sensing Research.

Mourales P and Thomas D (1998) Compact Low-Distortion Imaging Spectrometer for Remote Sensing. *SPIE Conference on Imaging Spectrometry IV* 3438: 31–37.

Pagano TS (1992) *Design of the Moderate Resolution Imaging Spectrometer – NADIR (MODIS-N)*. American Society of Photogrametry & Remote Sensing, pp. 374–385.

Stanich CG and Osterwisch FG (1994) *Advanced Operational Hyperspectral Scanners MIVIS and AHS*. First International Airborne Remote Sensing Conference and Exhibition, Strasbourg, France: Environmental Research Institute of Michigan.

Sun X and Baker JJ (2001) *A Hight Definition Hyperspectral Imaging System*. Proceedings, Fifth International Airborne Remote Sensing Conference and Exhibition, Veridian.

Torr MR, Baselow RW and Torr DG (1982) Imaging Spectrascopy of the Thermosphere from the Space Shuttle. *Applied Optics* 21: 4130.

Torr MR, Baselow RW and Mounti J (1983) An imaging spectrometric observatory for spacelab. *Astrophys. Space Science*. 92: 237.

Vane G, Green RO, Chrien TG, Enmark HT, Hansen EG and Porter WM (1993) The Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). *Remote Sensing of the Environment* 44: 127–143.

Warren CP, Anderson RD, Velasco A, Lin PP, Lucas JR and Speer BA (2001) *High Performance VIS and SWIR Hyperspectral Imagers for Airborne Applications*. Fifth International Conference and Exhibition, Veridian.

Wolfe WL (1997) *Introduction to Imaging Spectrometers*. SPIE Optical Engineering Press.

Wolfe WL and Zisfis GJ (eds) (1985) *The Infrared Handbook*, revised edition. Washington, DC: Office of Naval Research, Department of the Navy.

Zywicki RW, More KA, Holloway J and Witherspoon N (1994) *A Field Portable Hyperspectral Imager: The Large Area Fast Spectrometer*. Proceedings of the International Symposium on Spectral Sensing Research.

Telescopes

M M Roth, Astrophysikalisches Institut Potsdam, Potsdam, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In a broad sense, telescopes are optical instruments which provide an observer with an improved view of a distant object, where improved may be defined in terms of magnification, angular resolution, and light collecting power. Historically, the invention of the telescope was a breakthrough with an enormous impact on fundamental sciences (physics, astronomy), and, as a direct consequence, on philosophy; but also on other technological, economical, political, and military developments. It seems that the first refractive telescopes were built in the Netherlands near the end of the sixteenth century by Jan Lipperhey, Jakob Metius, and Zacharias Janssen. Based on reports of those first instruments, Galileo Galilei built his first telescope, which was later named after him, and made his famous astronomical observations of sunspots, the phases of Venus, Jupiter's moons, the rings of Saturn, and his discovery of the nature of the Milky Way as an assembly of very many stars. The so-called astronomical telescope was invented by Johannes Kepler, who in 1611 also published in *Dioptrice* the theory of this telescope. The disturbing effects of spherical and chromatic aberrations were soon realized by astronomers, opticians, and other users, but it was not before Moor Hall (1792) and John Dollond (1758) that at least the latter flaw was corrected for by introducing the achromat. Later, the achromat was significantly improved by Peter Dollond, Jesse Ramsden, and Josef Fraunhofer. The catadioptric telescope was probably introduced by Lenhard Digges (1571) or Nicolas Zucchius (1608). In 1671, Sir Isaac Newton was the first to use a reflector for astronomical observations. Wilhelm Herschel improved this technique and began in 1766 to build much larger telescopes with mirror diameters up to 1.22 m. When, in the nineteenth century, the conventional metal mirror was replaced by glass, it was Léon Foucauld (1819–1868) who applied a silver coating in order to improve the reflectivity of its surface.

Modern refractors can be built with superb apochromatic corrections, and astronomical reflectors have advanced to mirror diameters of up to 8.4 m (monolithic mirror), and 10 m (segmented mirror).

While the classical definition of a telescope involves an observer, i.e., the eye, in modern times often electronic detectors have replaced human vision. In the following, we will, therefore, use a more general definition of a telescope as an optical instrument, whose purpose is to image a distant object, either in a real focal plane, or in an afocal projection for the observation by eye.

Basic Imaging Theory

In the most simple case, a telescope may be constructed from a single lens or a single mirror,

creating a real focal plane. Let us, therefore, introduce some basic principles of imaging using optical elements with spherical surfaces.

Refraction at a Spherical Surface

Figure 1 shows an example of refraction at a spherical surface. A ray emerging from object O hits the surface between two media of refractive index n and n' at point P. After refraction, the ray continues at a deflected angle ε' towards the normal and intersects the optical axis at point O′. In the ideal case, all rays emerging from O at different angles σ are collected in O′, thus forming a real image of object point O in O′. In the paraxial approximation, we have $\sin\sigma = \tan\sigma = \sigma$, and $p \approx 0$. In triangle (O, O′, P): $\sigma' - \sigma = i - i'$. Using Snell's law $ni = n'i'$, we obtain:

$$\sigma' - \sigma = i' \frac{n' - n}{n} \qquad [1]$$

Since $\phi = \sigma' + i'$, and in the paraxial approximation:

$$\sigma' n' - \sigma n = \phi(n' - n) \qquad [2]$$

$$\frac{h}{s'} n' - \frac{h}{s} n = \frac{h}{r}(n' - n) \qquad [3]$$

we obtain finally:

$$\frac{n'}{s'} - \frac{n}{s} = \frac{n' - n}{r} \qquad [4]$$

For an object at infinity, where $s = \infty$, s' becomes the focal distance f':

$$f' = r\frac{n'}{n' - n} \quad \text{and} \quad \tilde{f} = -r\frac{n}{n' - n} \qquad [5]$$

Lenses

Combining two or more refractive surfaces, as in the preceding section, allows us to describe a single lens or more complex optical systems with several lenses in series. For a single lens in air, we just give the lens equation (without deriving it in detail):

$$s_2'(F') = \frac{1}{n-1}\frac{r_2[nr_1 - (n-1)d]}{(n-1)d + n(r_2 - r_1)} \qquad [6]$$

where d is the lens thickness, measured between the vertices of its surfaces, r_1, r_2 are the radii of the surfaces, n the index of refraction of the lens material, and $s_2'(F')$ the distance of the object image from the vertex of the last surface, facing the focal plane.

The back focal distance is given by:

$$f' = \frac{1}{n-1}\frac{nr_1r_2}{(n-1)d + n(r_2 - r_1)} \qquad [7]$$

An important special case is where the lens thickness is small compared to the radii of the lens surfaces, i.e., when $d \ll |r_2 - r_1|$. In this case we can neglect the term of $(n-1)d$ and write:

$$f' = \frac{1}{n-1}\frac{r_1r_2}{(r_2 - r_1)} \qquad [8]$$

Reflection at a Spherical Surface

Let us now consider the other simple case of imaging object O into O′ by reflection from the concave spherical surface, as depicted in Figure 2. In the paraxial case, we have:

$$\sigma = y/s, \quad i = \phi - \sigma, \quad \sigma' = y/s', \quad i' = \phi - \sigma', \quad \phi = y/r \qquad [9]$$

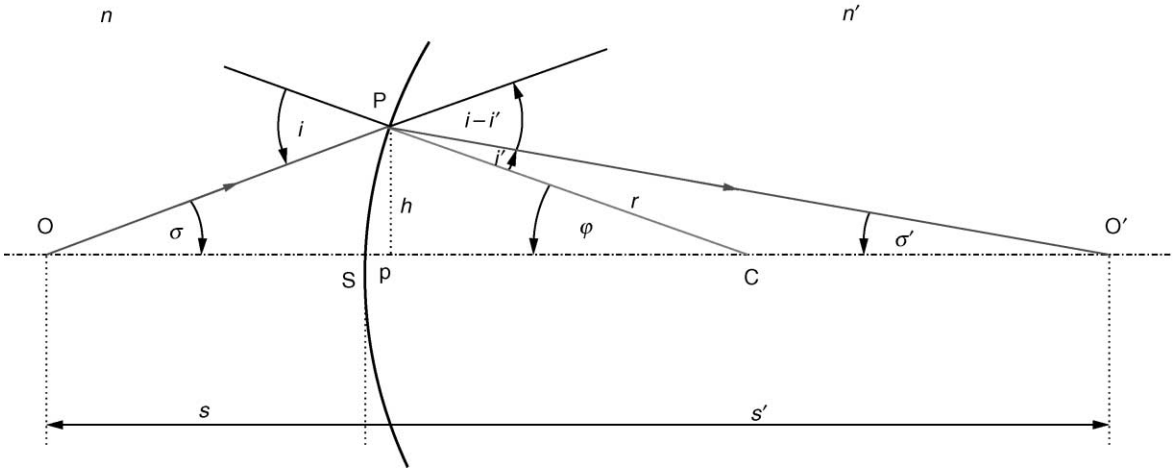


Figure 1 Refraction at a spherical surface between media with refractive indices n and n'.

With $i = -i'$, we obtain:

$$\frac{1}{s'} - \frac{1}{s} = -\frac{2}{r} \qquad [10]$$

Again, with $s = \infty$, the focal length becomes:

$$f' = -\frac{r}{2} \qquad [11]$$

In what follows, we will now assume that the distance of the object is always large (∞), which, in general, is the situation when a telescope is employed. The image is formed at a surface defined by O′, ideally a focal plane (paraxial case). We will consider deviations from the ideal case below.

Simple Telescopes

As we have observed above, a telescope is in principle nothing but an optical system, which images an object at infinity. In the most simple case, a lens with positive power or a concave mirror will satisfy this condition. The principle of a refractor and a reflecting telescope is shown in Figures 3 and 4. In both cases, parallel light coming from the distant object is entering through the entrance pupil, experiences refraction (reflection) at the lens (mirror), respectively, and is converging to form a real image in the focal plane.

Two point sources, for example two stars in the sky, separated by an angle δ, will be imaged in the focal plane as two spots with separation Δy. When f is the focal length of the telescope, we therefore have:

$$\tan \delta = \frac{\Delta y}{f} \qquad [12]$$

The so-called plate scale m of the telescope in units of arcsec/mm is given by:

$$m = 3600 \,\mathrm{atan}\frac{1}{f} \quad (f \text{ in mm}) \qquad [13]$$

In the case of the refractor, we have also shown an eyepiece, illustrating that historically the instrument was invented to enhance the human vision. The eye is located at the position of the exit pupil, thus receiving parallel light, i.e., observing an object at infinity, however (i) with a flux which is increased by a factor A, given by the ratio of the area of the telescope aperture with diameter D to the area of the eye's pupil d with $A = (D/d)^2$, and (ii) with a magnification M of the apparent angle between two separate objects at infinity, which is given by the ratio of the focal lengths of the objective and the eyepiece:

$$\Gamma = \frac{f_{\mathrm{obj}}}{f_{\mathrm{ocl}}} \qquad [14]$$

The magnification can also be expressed as the ratio of the diameters of the exit pupil and the entrance pupil:

$$\Gamma = \frac{D_{\mathrm{expup}}}{D_{\mathrm{enpup}}} \qquad [15]$$

The first astronomical telescopes with mirrors were equipped with an eyepiece for visual observation. The example in Figure 4 shows a configuration which is more common nowadays for professional astronomical observations, using a direct imaging detector in the focal plane (photographic plate, electronic camera).

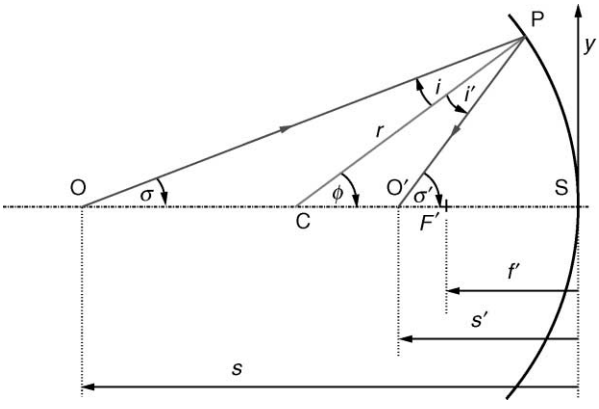


Figure 2 Reflection at a spherical surface with radius r.

Aberrations

Reflector Surfaces

We shall now study the behavior of a real telescope beyond the ideal in the case of a reflector. Figure 5 shows the situation of a mirror, imaging an on-axis

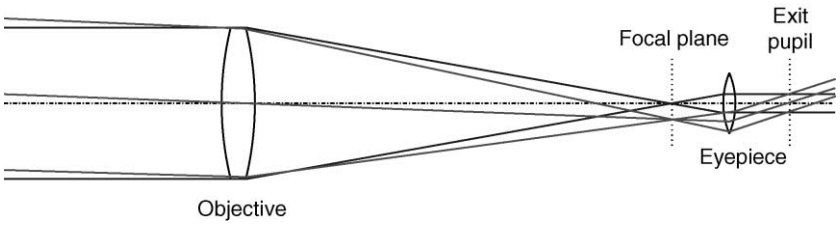


Figure 3 Refractor, astronomical telescope for visual observations with eyepiece (Kepler telescope).

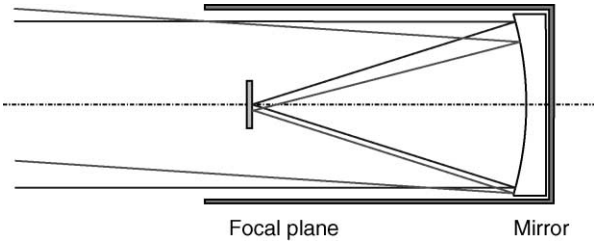


Figure 4 Reflector, astronomical telescope with prime focus imager.

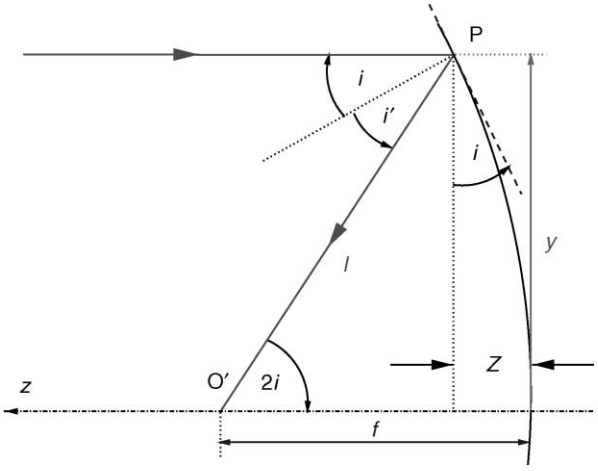


Figure 5 Imaging an object at infinity with reflector.

object at infinity into focus at point O′. With Fermat's principle, the optical paths of all rays forming the image must have identical length l. For all j, we must satisfy:

$$Z_j + f = l \qquad [16]$$

From the geometry in Figure 5, we find also:

$$y_i^2 + (f - Z_j)^2 = l^2 \qquad [17]$$

Combining eqns [16] and [17], we obtain the equation for a surface which satisfies the condition for perfect on-axis imaging:

$$y^2 = -4fz \qquad [18]$$

Obviously only a paraboloid is capable of fulfilling the condition for all rays parallel to the optical axis at whatever separation y. Also, we see that spherical mirrors are less than ideal imagers: rays at increasing distance y from the axis will suffer from increasing path length differences. The effect is called spherical aberration. We will investigate aberrations more quantitatively further below.

Elementary axisymmetrical surfaces can be expressed as follows, with parameter C being the conic constant:

$$y^2 - 2Rz + (1 + C)z^2 = 0 \qquad [19]$$

and the following characteristics:

$C > 0$	ellipsoidal, prolate
$C = 0$	spherical
$-1 < C < 0$	ellipsoidal, oblate
$C = -1$	paraboloidal
$C < -1$	hyperboloidal

Let us now leave the paraxial approximation and determine the focal length of rays which are parallel to, but further away at a distance y from the optical axis. Figure 5 shows the relevant geometrical relations. The intersection of the reflected ray with the optical axis at distance f from the vertex can be found from:

$$f = z + (f - z) \quad \text{where} \quad \frac{y}{f - z} = \tan(2i) \qquad [20]$$

We use the slope of the tangent in point P:

$$\frac{dz}{dy} = \tan i \qquad [21]$$

differentiate eqn [19] to get:

$$\frac{dz}{dy} = \frac{y}{R - (1 + C)z'} \qquad [22]$$

use the trigonometrical relation:

$$\tan 2\phi = \frac{2 \tan \phi}{1 - \tan^2 \phi} \qquad [23]$$

and substitute for $\tan(2i)$ in eqn [20] to obtain:

$$f = \frac{R}{2} + \frac{(1 - C)z}{2} - \frac{y^2}{2(R - (1 + C)z)} \qquad [24]$$

In order to express eqn [24] solely in terms of y, we rewrite eqn [19]:

$$z = \frac{R}{1 + C}\left(1 - \sqrt{1 - \frac{y^2}{R^2}(1 + C)}\right) \qquad [25]$$

Expanding z in a series and neglecting any orders higher than 6:

$$z = \frac{y^2}{2R} + (1 + C)\frac{y^4}{8R^3} + (1 - C)^3\frac{y^6}{16R^5} \qquad [26]$$

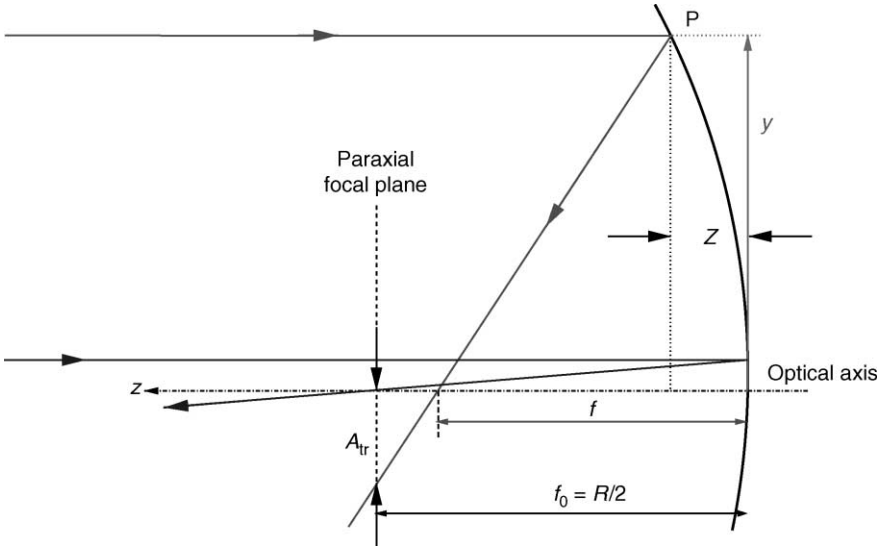


Figure 6 Transverse spherical aberration.

we obtain finally:

$$f = \frac{R}{2} - \frac{1+C}{4R}y^2 - \frac{(1+C)(3+C)}{16R^3}y^4 \qquad [27]$$

.

Transverse Spherical Aberration

Obviously, for any $C \neq -1$, rays of different distance y to the optical axis are focused at different focal lengths f (spherical aberration). The net effect for a distant point source is that the image is blurred, rather than forming a sharp spot. The lateral deviation of nonparaxial rays intersecting the nominal focal plane at $f_0 = R/2$ is described by the transverse spherical aberration A_{tr} (Figure 6):

$$A_{tr,sph} = \frac{y}{f-z}(f_0 - f) \qquad [28]$$

Using eqns [19] and [27], one obtains:

$$A_{tr,sph} = -\frac{1+C}{2R^2}y^3 + \frac{3(1+C)(3+C)}{8R^4}y^5 + \cdots \qquad [29]$$

Angular Spherical Aberration

It is sometimes more convenient to consider the angular spherical aberration, since a telescope is naturally measuring angles between objects at infinity. To this end, we compare an arbitrary reflector surface according to eqn [19] with a reference paraboloid, which is known to be free of on-axis spherical aberration (Figure 7). For any given ray, we observe how the reflected ray under study is tilted

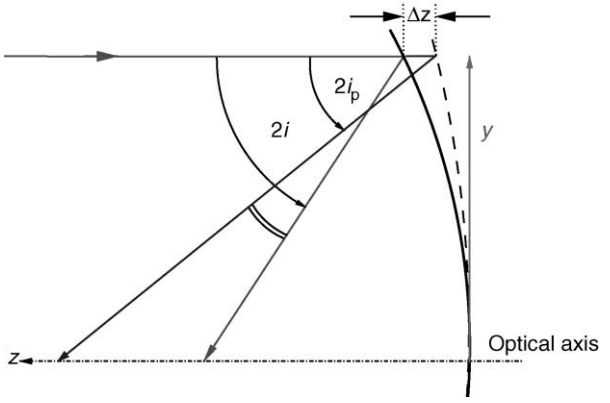


Figure 7 Angular aberration determined with reference to a paraboloid.

towards the reference ray, which would be reflected from the ideal surface of the paraboloid. While the former is measured under an exit angle of $2i$ towards the optical axis (see Figure 5), we designate the latter $2i_p$. The angular spherical aberration is defined as the difference of these two angles:

$$A_{ang,sph} = 2i - 2i_p = 2(i - i_p) \qquad [30]$$

Again, we can use the slope of the tangent to determine the change of angles $i \rightarrow i_p$:

$$A_{ang,sph} = 2\left(\frac{dz}{dy} - \frac{dz_p}{dy}\right) = 2\frac{d}{dy}\Delta z \qquad [31]$$

It is, therefore, sufficient to consider the difference in z between the surface under study and the

reference paraboloid. From eqn [26], we obtain as an approximation to third order:

$$A_{\text{ang,sph}} = -(1+C)\frac{y^3}{R^3} \qquad [32]$$

Aberrations in the Field

Since the use of a telescope exclusively on-axis would be somewhat limited, let us finally investigate the behavior under oblique illumination, i.e., when an object at infinity is observed under an angle Θ towards the optical axis of the system. This will allow us to assess aberrations in the field. For simplicity, we are considering a paraboloid, remembering that on-axis the system is free of spherical aberration. Again, we are employing an imaginary reference paraboloid which is pointing on-axis towards the object at angle Θ (Figure 8). Its coordinate system (O', z', y') is displaced and tilted with regard to the system under study (O, z, y). One can show that the offset Δz, which is used to derive the angular aberration in the same way as above, is given by:

$$2\Delta z = a_1\frac{y^3\Theta}{R^2} + a_2\frac{y^2\Theta^2}{R} + a_3 y\Theta^3 \qquad [33]$$

The angular aberration in the field is then:

$$A_{\text{ang,f}} = 3a_1\frac{y^2\Theta}{R^2} + 2a_2\frac{y\Theta^2}{R} + a_3\Theta^3 \qquad [34]$$

The coefficients are called:

a_1: coma

a_2: astigmatism

a_3: distortion

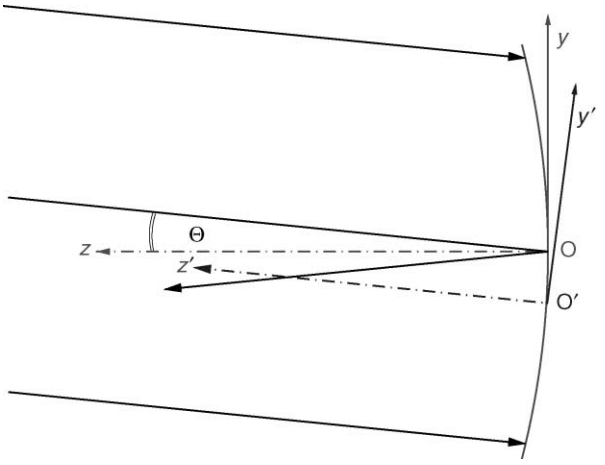


Figure 8 Aberration in the field, oblique rays.

Chromatic Aberration

So far we have only considered optical aberrations of reflecting telescopes. Spherical aberration in the case of a lens is treated in analogous ways and will not be discussed again. However, the wavelength dependence of the index of refraction in optical media (dispersion) is an important factor and gives rise to chromatic aberrations which are only encountered in refractive optical systems. We can distinguish two principal effects in the paraxial approximation: (i) longitudinal chromatic aberration, resulting in different axial image positions as a function of wavelength; and (ii) lateral chromatic aberration, which can be understood as an aberration of the principal ray as a function of wavelength, leading to a varying magnification as a function of wavelength (plate scale in the case of a telescope).

As an example, let us determine the wavelength dependence of the focal length of a single thin lens. Opticians are conventionally using discrete wavelengths of spectral line lamps of different chemical elements, e.g., the Mercury e line at 546.1 nm, Cadmium F' 480.0 nm, or Cadmium C' 643.8 nm, covering more or less the visual wavelength range of the human eye. Differentiation of eqn [8] yields for the wavelength of e:

$$\frac{f'}{\mathrm{d}n} = -\frac{1}{(n_e-1)^2}\frac{r_1 r_2}{r_2 - r_1} \qquad [35]$$

Substituting

$$\frac{r_1 r_2}{r_2 - r_1} = f'_e(n_e - 1) \qquad [36]$$

we get:

$$\mathrm{d}f' = -\frac{\mathrm{d}n}{n_e - 1}f'_e \qquad [37]$$

Replacing the differential by the differences $\Delta f' = f'_{F'} - f'_{C'}$ and $\Delta n = n_{F'} - n_{C'}$, then:

$$f'_{F'} - f'_{C'} = -\frac{n_{F'} - n_{C'}}{n_e - 1}f'_e = -\frac{f'_e}{\nu_e} \qquad [38]$$

ν_e is called Abbe's Number:

$$\nu_e = \frac{n_e - 1}{n_{F'} - n_{C'}} \qquad [39]$$

It is a tabulated characteristic value of dispersion for any kind of glass and allows one to quickly estimate the difference in focal length at the extreme wavelengths F' and C' when the nominal focal length of a lens is known as e.

Angular Resolution

In the previous sections, we have exclusively employed geometrical optics for the description of telescope properties. With geometrical optics and in the absence of aberrations, there is no limit to the angular resolution of a telescope, i.e., there is no limiting separation angle δ_{lim}, below which two distant point sources can no longer be resolved. However, real optical systems experience diffraction, leading to the finite width of the image of a point source (Airy disk), whose normalized intensity distribution in the focal plane as a function of radius r' from the centroid for the case of a circular aperture of diameter D can be conveniently described by a Bessell function (λ: wavelength, f': focal length of the objective):

$$\frac{I(r)}{I_0} = \left[\frac{2J_1(\nu)}{\nu}\right]^2 \quad \text{where} \quad \nu = \frac{\pi D r'}{\lambda f'} \qquad [40]$$

The intensity distribution is plotted in Figure 9. The first four minima of the diffraction ring pattern are listed in Table 1. Two point sources are defined to be at the resolution limit of the telescope when the maximum of one diffraction pattern happens to coincide with just the first diffraction minimum of the other one (Figure 9, right). As a rule of thumb, the angular resolution is of a telescope with diameter D is given by:

$$\delta_{\text{min}} = 1.22\frac{\lambda}{D} \rightarrow \frac{138}{D} \qquad [41]$$

the latter in units of arcsec for a wavelength of 550 nm and D in mm.

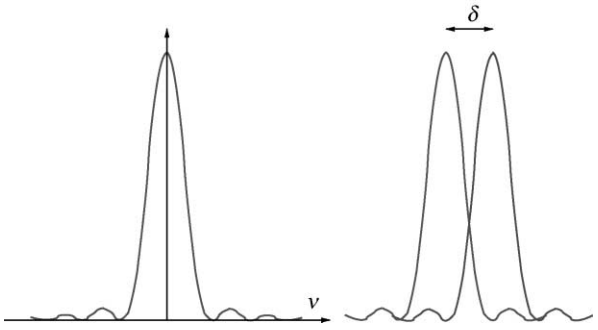


Figure 9 Left: radial intensity distribution of diffraction pattern (Airy disk). Right: two overlapping point source images with same intensity at limiting angular separation δ.

Table 1 Diffraction minima of airy disk

n	ν_n	$r'D/2f'\lambda$
1	3.83	0.61
2	7.02	1.12
3	10.2	1.62
4	13.3	2.12

Types of Telescopes

There are very many different types of telescopes which were developed for different purposes. In order to obtain an impression, we will only briefly introduce a nonexhaustive selection of refractors and reflectors, which by no means is representative of the whole. For a more complete overview, the reader is referred to the literature.

Refractor Objectives

Refractors are first of all characterized by their objectives. Figure 10 shows a sequence of lenses, indicating the progress of lens design developments from left to right. All lenses are groups 2 or 3, designed to compensate for chromatic aberration through optimized combinations of crown and flint glasses. The first lens is a classical C achromat, e.g., using BK7 and SF2 glass, which is essentially the design which was introduced by Hall and Dollond. It is still a useful objective for small telescopes, where a residual chromatic aberration is tolerable. The second lens is the Fraunhofer E type, which is characterized by an air gap between the two lenses (BK7, F2), which helps to optimize the image quality of the system, reducing spherical aberration and coma. As a disadvantage, the two additional glass–air interfaces lead to ghost images (reflected light) and Fresnell losses. The objective is critical with regard to the alignment of tip and tilt. The third example is an improved AS type achromat (KzF2, BK7), which is similar to the E type in

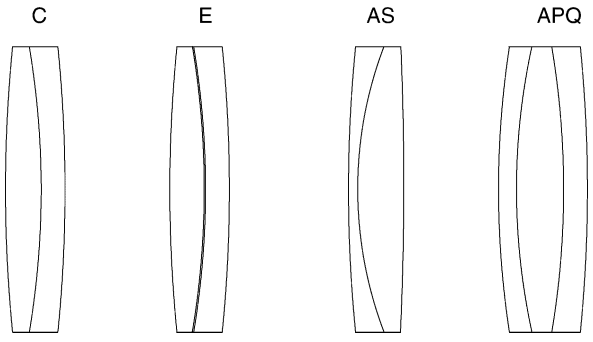


Figure 10 Different types of refractor objectives, from left to right: classical C achromat, Fraunhofer E achromat with air gap, AS achromat, APQ apochromat.

that it also possesses an air gap, however with an improved chromatic aberration correction in comparison to the Fraunhofer objective. The lenses show a more pronounced curvature and are more difficult to manufacture. The beneficial use of KzF2 glass is an example of the successful research for the development of new optical media, which was pursued near the end of the nineteenth century in a collaboration between Otto Schott and Ernst Abbé. The last example is a triplet, employing ZK2, CaF_2, and ZK2. This APQ type of objective is an apochromat, whose chromatic aberration vanishes at 3 wavelengths. It is peculiar for its use of a crystal (calcium fluoride, CaF_2) as the optical medium of the central lens. The manufacture of the CaF_2 lens is critical and expensive, but it provides for an excellent correction of chromatic and spherical aberration of this high quality objective. The triplet is also difficult to manufacture because of the coefficient of thermal expansion, which is roughly different by a factor of two between CaF_2 and the glass. In order to avoid disrupture of the cemented lens group, some manufacturers have used oil immersion to connect the triplet. Since CaF_2 not only has excellent dispersion properties to correct for chromatic aberration, but also exhibits an excellent transmission beyond the visual wavelength regime, this type of lens is particularly useful for applications in the blue and UV.

Finally, Figure 11 shows the complete layout of an astrograph, which is a telescope for use with photographic plates. The introduction of this type of instrument meant an important enhancement of astrophysical research by making available a significantly more efficient method, compared to the traditional visual observation of individual stars, one by one. Using photographic plates, it became possible to measure position and brightness of hundreds and thousands of stars, when previously one could investigate tens. Typical astrographs were built with apertures between 60 and 400 mm, and a field-of-view as large as 15 degrees. Despite the growing importance of space-based observatories for astrometry, some astrographs are still in use today.

Basic Oculars

While the general characteristics of a refractor are dominated by the objective, the eyepiece nevertheless plays an important role for the final appearance to the observer, with an effect on magnification, image quality, chromatism, field of view, and other properties of the entire system. From a large variety of different ocular types, let us briefly consider a few simple examples.

The first eyepiece in Figure 12 is the one of the Kepler telescope, which was already introduced in Figure 3. It has a quite simple layout with a single lens of positive power. The lens is located after the telescope focal plane at a distance which is equal to the ocular focal length f_oc. The telescope and the ocular focal planes are coinciding, and as a result, the outcoming beam is collimated, forming the system exit pupil after a distance f_oc. This is where ideally the observer's eye is located. The focal plane is creating a real image, which is a convenient location for a reticle. The Kepler telescope is, therefore, a good tool for measurement and alignment purposes. We note that the image is inverted, which is relatively unimportant for a measuring telescope and astronomy, but very annoying for terrestrial use.

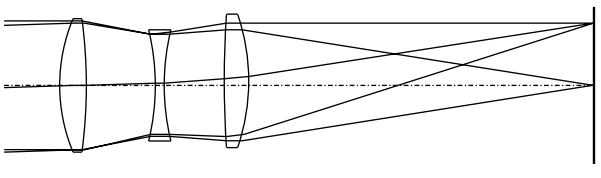


Figure 11 Astrograph.

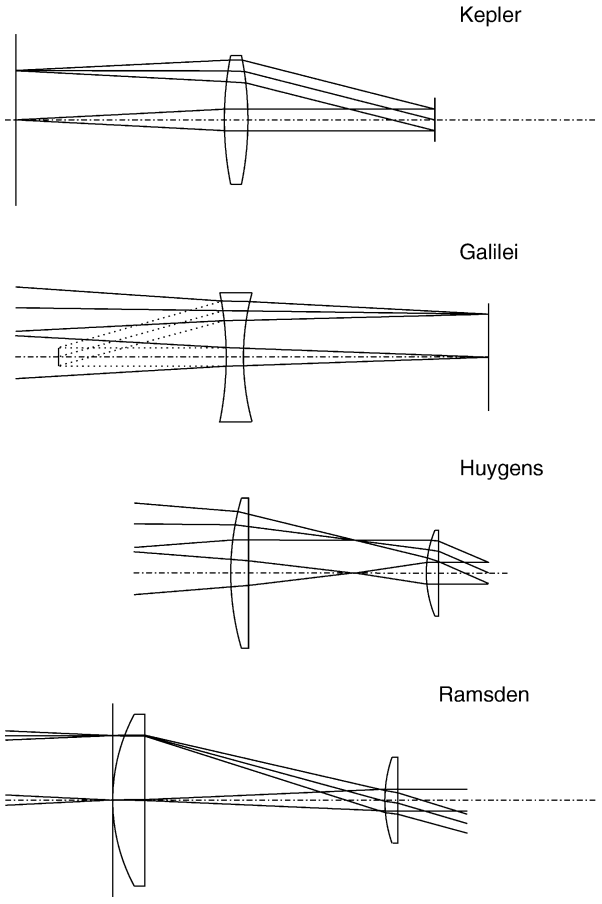


Figure 12 Elementary eyepieces after Kepler, Galilei, Huygens, and Ramdsen.

This latter property has been avoided in the next example, which is the ocular of the Galilean telescope, which was built by Galileo and used for his famous observations of planets and moons in the solar system. It uses a single lens, but contrary to the Kepler type with negative power. The lens is located in the telescope beam in front of the focal plane of the objective, which means that the exit pupil is virtual and not accessible to the observer (indicated by dotted lines in Figure 12). The overall length of the telescope is somewhat shorter than Kepler's, however at the expense of the lack of a real focal plane, so no reticle can be applied.

The third example is the 2-lens-ocular of Christian Huygens, who was able to prove with his design that it is possible to correct paraxially the lateral chromatic aberration. The eyepiece is still being built today.

Another elementary ocular is shown in Figure 12, which is the eyepiece after Ramsden. This system is also corrected for lateral chromatic aberration. The real image is formed on the front surface of the second lens, providing a location for a reticle or a scale for measuring purposes. The exit pupil is interior to the ocular. In addition to these selected basic types, there are many more advanced multilens oculars in use today, with excellent correction and large field of view (up to 60°).

Reflectors

Let us now describe several basic types of reflectors (see Figure 13). The first example is the Newton reflector, which uses a flat folding mirror to create an accessible exterior focus. It is conceptually simple and still in use today for amateur astronomy. When the main mirror is a paraboloid, the spherical aberration is corrected on the axis of the system.

The second example represents an important improvement over the simple 1-mirror design of Newton's telescope, in that the combination of a parabolic first mirror (main mirror, M1) with a hyperbolic secondary mirror (M2) eliminates spherical aberration, but not astigmatism and coma. The Cassegrain telescope has a curved focal surface which is located behind the primary mirror. Due to the onset of coma and astigmatism, the telescope has a modest useful field-of-view. Besides the correction properties, the system has significant advantages in terms of mechanical size and weight, playing an important role for the construction of large telescopes. The basic type of this design has practically dominated the construction of the classical large astronomical telescopes with M1 diameters of up to 5 m, and, more recently, the modern 8–10 m class telescopes. On the practical side, the external focus of the Cassegrain on the bottom of the mirror support cell is a convenient location for mounting focal plane instruments (e.g., CCD cameras, spectrographs, and others).

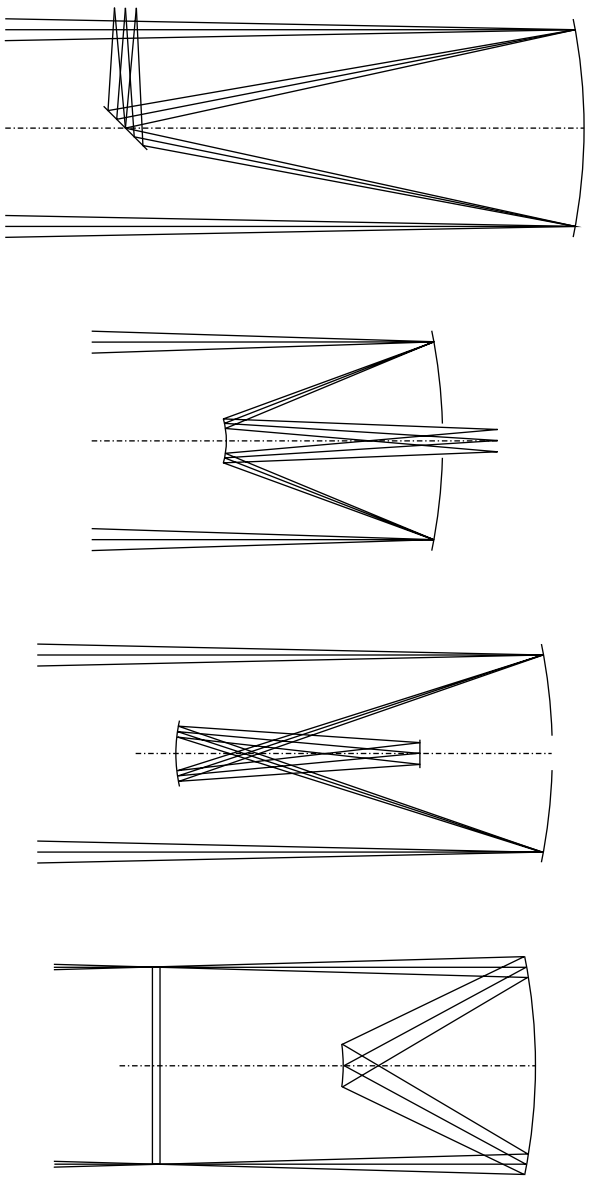


Figure 13 Reflectors, from top to bottom: Newton, Cassegrain, Gregory, Schmidt.

The more common variant of the Cassegrain type, which was actually used for the design of these telescopes, is the layout of Ritchey (1864–1945) and Cretien (1879–1956), which combines two hyperbolic mirrors for M1 and M2 (RCC), but otherwise has the same appearance as the Cassegrain. Besides spherical aberration, it also corrects for coma, giving a useful field-of-view of up to 1°. Like in the case of the Cassegrain telescope, the RCC exhibits some curvature of the focal plane.

Our third example is the Gregory reflector, which also is a 2-mirror system, however with a concave elliptical secondary mirror. The correction of this telescope is similar, but somewhat inferior to the one of the Cassegrain. The use of a concave M2 is advantageous for manufacture and testing. Nevertheless, the excessive length of the system has precluded the use of this layout for large astronomical telescopes, except for stationary beams, for example, in solar observatories.

The last example is the design put forward by Bernhard Schmidt (1879–1935) who wanted to create a telescope with an extremely wide field-of-view. The first telescope of his new type achieved a field-of-view of 16° – something which had not been feasible with previous techniques. The basic idea is to use a fast spherical mirror, and to correct for spherical aberration by means of an aspherical corrector plate in front of the mirror, which also forms the entrance pupil of the telescope. The focal surface is strongly curved, making it necessary to employ a detector with the same radius of curvature (bent photographic plates), or to compensate the curvature by means of a field lens immediately in front of the focus. The presence of the corrector plate and, if used, the field flattener, gives rise to significant chromatism. Nevertheless, the Schmidt telescope was a very successful design and dominated wide-angle survey work in astronomy for many decades.

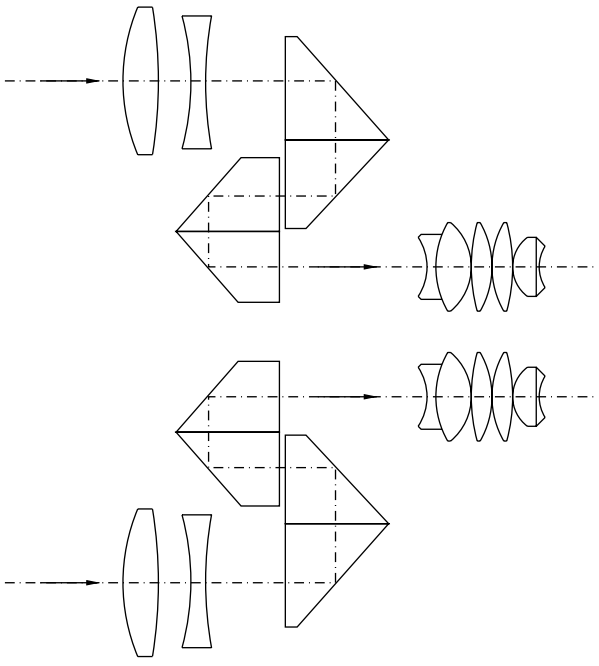


Figure 14 Principle of prism binoculars with folded beams.

Applications

General Purpose Telescopes

According to the original meaning of the Greek 'tele–skopein', the common conception of the telescope is the one of an optical instrument, which provides an enlarged view of distant objects. In addition, a large telescope aperture, compared to the eye, improves night vision and observations with poorly illuminated scenes. There are numerous applications in nautical, security, military, hunting, and other professional areas. For all of these applications, an intuitive vision is important, essentially ruling out the classical astronomical telescope (Kepler telescope, see Figure 3), which produces an inverted image.

Binoculars with Porro prisms are among the most popular instruments with upright vision, presenting a compact outline due to the folded beam geometry (Figure 14), but also single tube telescopes with re-imaging oculars are sometimes used.

Metrology

As we have realized in the introduction, telescopes are instruments which perform a transformation from angle of incidence in the aperture to distances from the origin in the focal plane. If the focal length of the objective is taken sufficiently long, and a high-quality eyepiece is used, very small angles down to the arcsec regime can be visually observed. Telescopes are, therefore, ideal tools to measure angles to very high precision.

In Figure 15 we see a pair of telescopes, facing each other, where one instrument is equipped with a light source, and the other one is used to observe the first one. The first telescope is also called a collimator. It possesses a lamp, a condensor system, a focal plane mask (usually a pinhole, a crosshair, or some other useful pattern imprinted on a transparent plate), and the telescope objective. The focal plane mask is projected to infinity. It can thus be observed with the

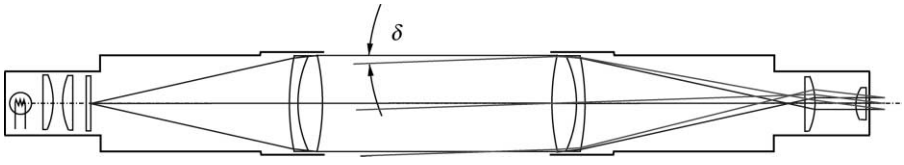


Figure 15 Collimator and telescope.

measuring telescope, such that an image of the collimator mask appears in the telescope focal plane. Also the telescope is equipped with a mask in its focal plane, usually with a precision scale or another pattern for convenient alignment with the collimator mask. If now the collimator is tilted by a small angle δ against the optical axis of the telescope, the angle of incidence in the measuring telescope changes by this same amount, giving rise to a shift of the image by $\Delta x = f \tan \delta$. Note that the measurement is completely insensitive to parallel shifts between the telescope and the collimator, since it is parallel light which is transmitted between the two instruments.

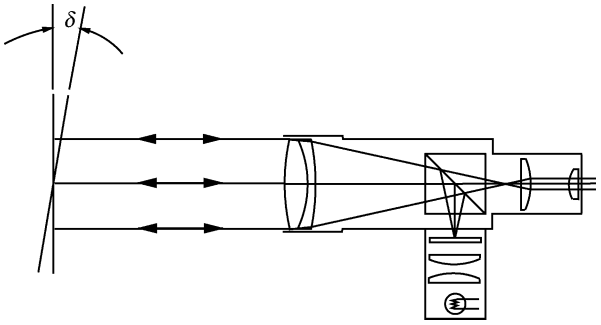


Figure 16 Autocollimator.

Another special device, which combines the former two instruments into one, is the Autocollimator. The optical principle is shown in Figure 16: in principle, the autocollimator is identical to a measuring telescope, except that a beamsplitter is inserted between the focal plane and the objective. The beamsplitter is used to inject light from the same kind of light source device as in the normal collimator. The axis of the folded light source beam and of the illuminated collimator mask are aligned to match the focal plane mask of the telescope section: the optical axis of the emerging beam is identical to the optical axis of the telescope. Under this condition, the reflected light from a test object will be exactly centered on the telescope focal plane mask, when it is perfectly aligned to normal incidence. Normally, a high-quality mirror is attached to the device under study, but sometimes also the test object itself possesses a plane surface of optical quality, which can be used for the measurement.

Figure 17 shows schematically an example for such a measurement, where a mirror is mounted on a linear stage carriage for the purpose of measuring the flexure of the stage to very high precision. With a typical length of ≈ 1 m, this task is not a trivial problem for a mechanical measurement.

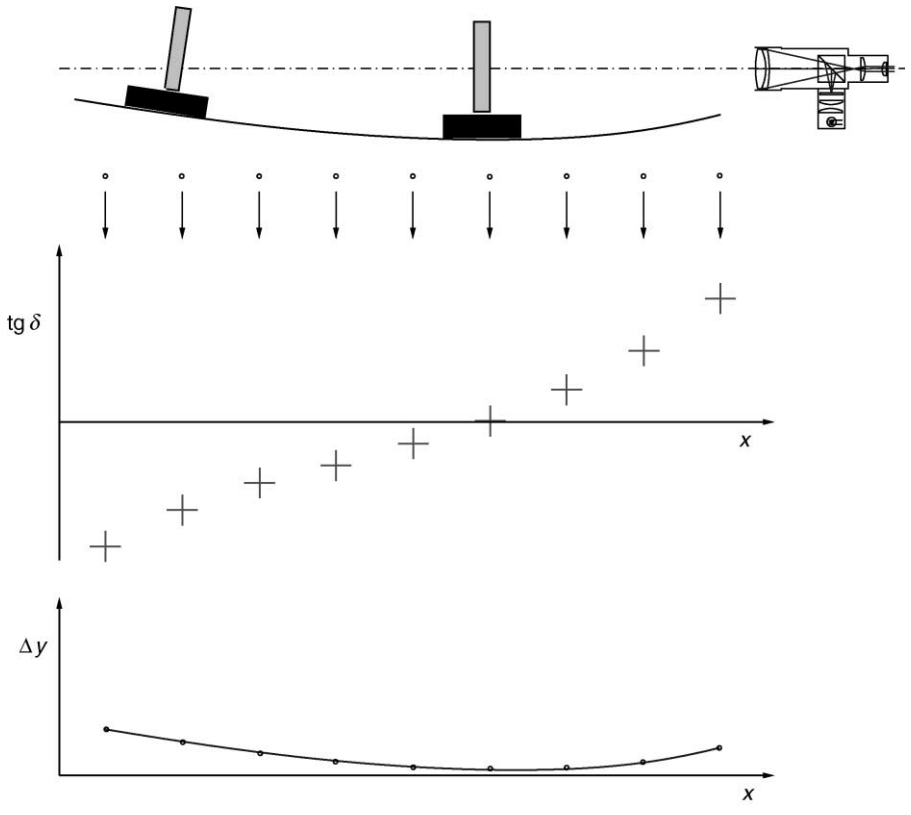


Figure 17 Using an autocollimator to measuring the flexure of a linear stage.

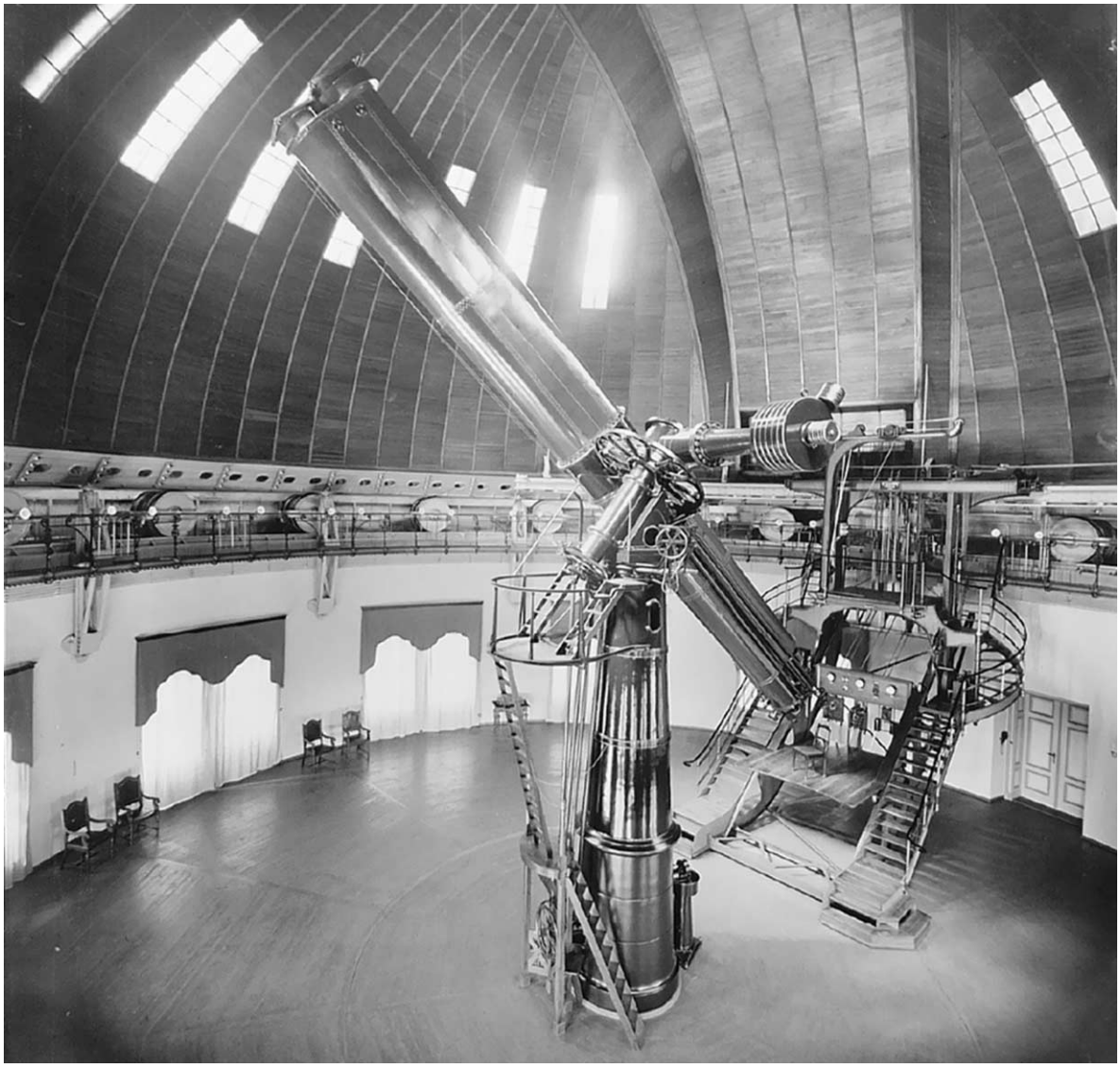


Figure 18 The 80 cm + 60 cm 'Great Refractor' at the Astrophysical Observatory Potsdam. The correction of the 80 cm objective of this telescope led Johannes Hartmann to the development of the 'Hartmann Test', which was the first quantitative method to measure aberrations. This method is still in use today, and has been further developed to become the 'Shack-Hartmann' technique of wavefront sensing. Courtesy of Astrophysikalisches Institut Potsdam.

By measurement with the autocollimator the tilt of the mirror consecutively for a sufficiently large number of points x_i along the rail, we can now plot the slope of the rail at each point x_i as tan d:

$$\frac{dy}{dx} = \tan \delta \qquad [42]$$

Integrating the series of slope values reproduces the shape of the rail $y = f(x)$.

Other applications include multiple autocollimation, wedge measurements, and right angle measurements.

Multiple autocollimation is achieved by inserting a fixed plane-parallel and semi-transparent mirror between the autocollimator and the mirror under study. Light coming from the test mirror will be partially transmitted through the additional mirror towards the autocollimator, and partially reflected back to the test mirror, and so forth. After k reflections, a deflection of $k\delta$ is measured, when δ is the tilt of the test mirror.

The wedge angle α of a plane-parallel plate with refractive index n_w is easily measured by observing the front and back surface reflections from the plate, keeping in mind that the beam bouncing back from the rear will also experience refraction in the plate. If δ is the measured angle difference between the two front and rear beams, then $\alpha = \delta/2n_w$.

Two surfaces of an object can be tested for orientation at right angles by attaching a mirror to surface 1 and observe the mirror with the fixed autocollimator. In a second step, an auxiliary pentaprism is inserted into the beam, deflecting the light towards surface 2, where the mirror is attached next. If a precision pentaprism is used, any deviation from right angles between the two surfaces is seen as an offset between step 1 and 2.

These are only a few examples of the very many applications for alignment and angle measurements in optics and precision mechanics.

Without going into much detail, we are just mentioning that also the theodolite and sextant make use of the same principle of angular measurements, however without the need of a collimator or autocollimation.

Astronomical Telescopes

Despite its importance as a measurement and testing device for optical technology and manufacture as described in the preceeding paragraph, it is the original invention and subsequent improvement of the telescope for *Astronomy* which is probably the most fascinating application of any optical system known to the general public. The historical development of the astronomical telescope spans from the 17th century Galileian refractor with an aperture of just a few centimeters in diameter, over the largest refractors built for the Lick and Yerkes observatories with objectives of up to ≈ 1 m in diameter (see also Figures 18–20), the 4 m-class reflectors of the second half of the 20th century like e.g. the Hale Observatory 5m, the Kitt Peak and Cerro Tololo 4 m, the La Silla 3.6 m, or the La Palma 4.2 m telescopes (to name only a few), to the current state-of-the-art of 8–10 m class telescopes, prominent examples being the two Keck Observatory 10 m telescopes on Mauna Kea (Hawaii), or the ESO Very Large Observatory (Paranal, Chile), consisting of four identical 8.2 m telescopes (Figures 21–24).

This development has always been driven by the discovery of fainter and fainter celestial objects, whose quantitative study is the subject of modern astrophysics, involving many specialized disciplines such as atom/quantum/nuclear physics, chemistry, general relativity, electrodynamics, hydrodynamics, magneto-hydrodynamics, and so forth. In fact, cosmic objects are often referred to as 'laboratories'

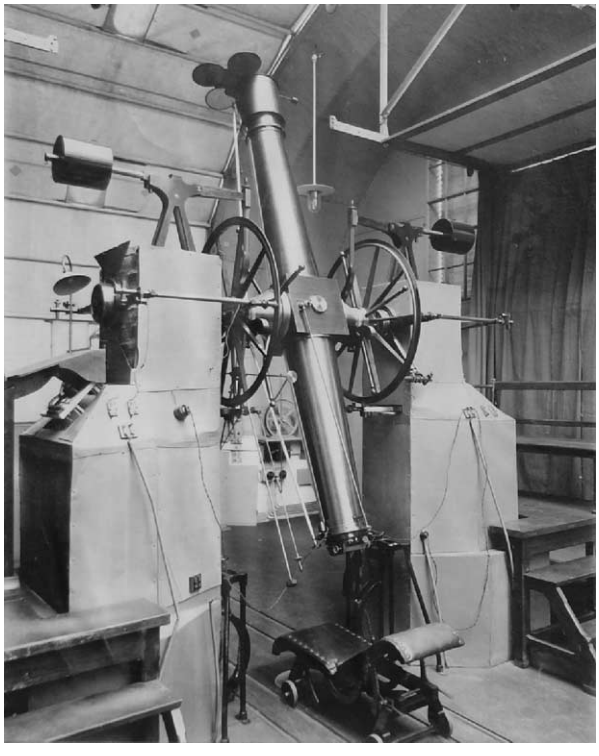


Figure 19 Pistor & Martin Meridian Circle at the former Berlin Observatory in Potsdam Babelsberg. The instrument was one of the finest art telescopes with ultra-high precision for astrometry, built in 1868. Courtesy of Astrophysikalisches Institut Potsdam.

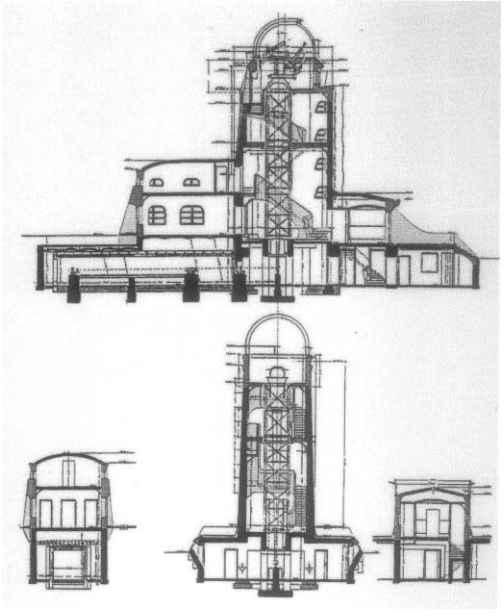


Figure 20 Architect's sketch of historical 'Einsteinturm' solar telescope, Astrophysical Observatory Potsdam. The telescope objective is mounted in the tower structure, facing upward a siderostat which is mounted inside the dome. The light is focused in the basement after deflection from a folding mirror on the slit of a bench-mounted high resolution spectrograph. The telescope represented state-of-the-art technology at the beginning of the twentieth century. Courtesy of Astrophysikalisches Institut Potsdam.



Figure 21 The Very Large Telescope Observatory (VLT) on Mt Paranal, Chile, with VLT Interferometer (VLTI). The four large enclosures are containing the active optics controlled VLT unit telescopes, which are operated either independently of each other, or in parallel to form the VLTI. The structures on the ground are part of the beam combination optics/delay lines. Courtesy of European Southern Observatory.

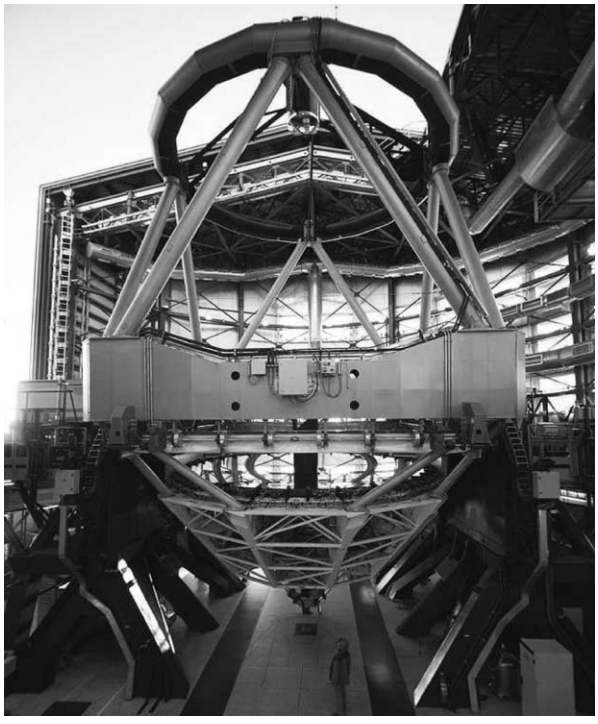


Figure 22 Total view of Kueyen, one of four unit telescopes of the VLT observatory. Courtesy of European Southern Observatory.

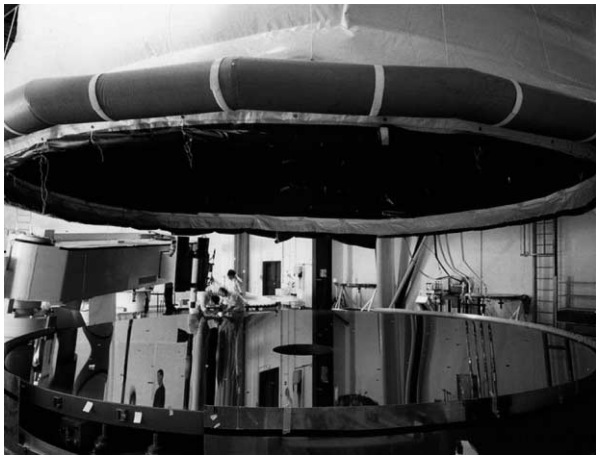


Figure 23 8.2 m thin primary mirror for the VLT. The VLT design is based on thin meniscus mirrors, made out of ZERODUR, which have low thermal inertia and which are always kept at ambient temperature, thus virtually removing the effect of thermal turbulence (mirror seeing). Seeing is the most limiting factor of ground-based optical astronomy. The optimized design of the VLT telescopes has introduced a significant leap forward to a largely improved image quality from the ground, and, consequently, improved sensitivity. The thin mirrors are constantly actively controlled in their mounting cells, using numerous piston actuators on the back surface of the mirror. Another key technology under development is 'adaptive optics', a method to compensate atmospheric wavefront deformations in real-time, thus obtaining diffraction limited images. Courtesy of European Southern Observatory.

with extreme conditions of density, temperature, magnetic field strength, etc., which would be impossible to achieve in any laboratory on Earth. Except for particles (neutrinos, cosmic rays), gravitational waves, and the physical investigation of meteorites and other solar system material collected by space probes, telescopes (for the entire wavelength range of the electromagnetic spectrum) are the *only* tool to accomplish quantitative measurements of these cosmic laboratories.

In an attempt to unravel the history of the birth and evolution of the universe, to discover and to measure the first generation of stars and galaxies, and to find evidence for traces of life outside of planet Earth, plans are currently underway to develop yet another new generation of giant optical telescopes with apertures of 30–100 m in diameter (ELT: extremely large telescope). The technological challenges in terms of precision and stability are outstanding.

Along with the impressive growth of light collecting area, we must stress the importance of angular resolution, which has also experienced a significant evolution over the history of astronomical telescopes: from ≈ 5 arcsec of the Galilei telescope, which already provided an order of magnitude improvement over the naked eye, over roughly 1 arcsec of conventional telescopes in the 20th century, a few hundredths of an arcsec for the orbiting Hubble Space

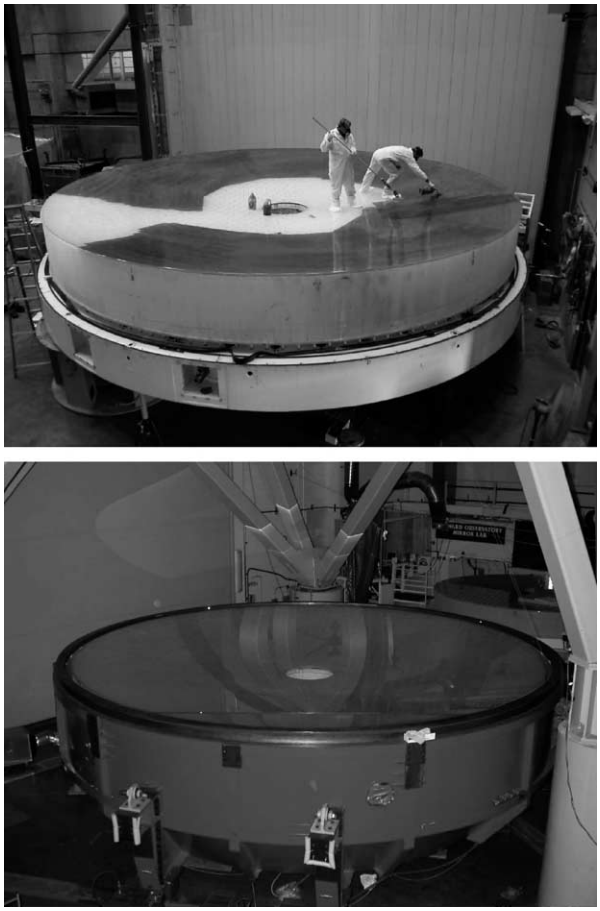


Figure 24 Top: spin-cast 8.4 m honeycomb mirror #1 for the Large Binocular Telescope, the largest monolithic mirror of optical quality in the world. Shown in the process of applying a protective plastic coating to the end-polished surface. Bottom: LBT mirror #1, finally mounted in mirror cell. Courtesy of Large Binocular Telescope Observatory.

Telescope (Figure 25), down to milli-arcsec resolution of the Keck and VLT Interferometers. The development of the aforementioned ELTs is thought to be only meaningful if operated in a diffraction-limited mode, using the emerging technique of *adaptive optics*, which is a method to correct for wavefront distortions caused by atmospheric turbulence (image blurr) in real-time.

The combination of large optical/infrared telescopes with powerful detectors and focal plane instruments (*see* **Instrumentation:** Astronomical Instrumentation) has been a prerequisite to make astrophysics one of the most fascinating disciplines of fundamental sciences.

See also

Geometrical Optics: Aberrations. **Imaging:** Adaptive Optics. **Instrumentation:** Astronomical Instrumentation.

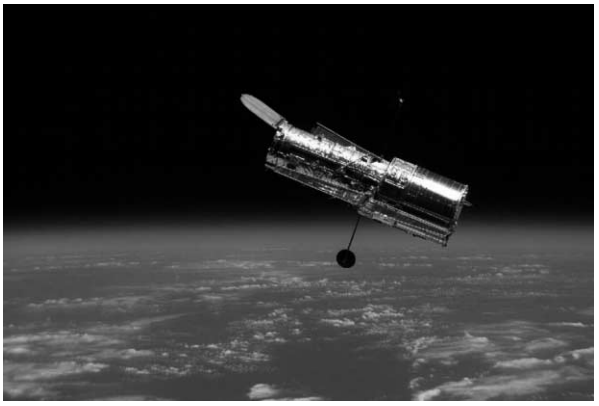


Figure 25 Hubble Space Telescope (HST) in orbit. This 2.4 m telescope is probably the most productive telescope which was ever built. Due to the absence of the atmosphere, the telescope delivers diffraction-limited, extremely sharp images, which have revealed unprecedented details of stars, star clusters, nebulae and other objects in the Milky Way and in other galaxies. Among the most spectacular results, HST has provided the deepest look into space ever, revealing light from faint galaxies that was emitted when the universe was only at 10% of its present age. Courtesy of Space Telescope Science Institute, Baltimore.

Further Reading

Bely PY (ed.) (2003) *The Design and Construction of Large Optical Telescopes*. New York: Springer.

Born M and Wolf E (1999) *Principles of Optics*, 7th ed. Cambridge: Cambridge University Press.

Haferkorn H (1994) *Optik*, 3rd ed. Leipzig: Barth Verlagsgesellschaft.

Korsch D (1991) *Reflective Optics*. San Diego: Academic Press.

Laux U (1999) *Astrooptik*, 2nd ed. Heidelberg: Verlag Sterne und Weltraum.

McCray WP (2004) *Giant Telescopes: Astronomical Ambition and the Promise of Technology*. Cambridge: Harvard University Press.

Osterbrock DE (1993) *Pauper and Prince: Richey, Hale, and Big American Telescopes*. Tucson: University of Arizona Press.

Riekher R (1990) *Fernrohre und ihre Meister*, 2nd ed. Berlin: Verlag Technik GmbH.

Rutten HGJ and van Venrooij MAM (1988) *Telescope Optics*. Richmond, VA: Willman-Bell.

Schroeder DJ (1987) *Astronomical Optics*. San Diego: Academic Press.

Welford WT (1974) *Aberrations of the Symmetrical Optical System*. London: Academic Press.

Wilson RN (1996) *Reflecting Telescope Optics I: Basic Design Theory and its Historical Development*. Berlin: Springer.

Wilson RN (1999) *Reflecting Telescope Optics II: Manufacture, Testing, Alignment, Modern Techniques*. Berlin: Springer.

INTERFEROMETRY

Contents

Overview
Gravity Wave Detection
Phase-Measurement Interferometry
White Light Interferometry

Overview

J C Wyant, University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Interferometers are a powerful tool used in numerous industrial, research, and development applications, including measuring the quality of a large variety of manufactured items such as optical components and systems, hard disk drives and magnetic recording heads, lasers and optics used in CDs and DVD drives, cameras, laser printers, machined parts, components for fiber optics systems, and so forth. Interferometers can also be used to measure distance, spectra, and rotations, etc. The applications are almost endless.

There are many varieties of interferometers. Some interferometers form interference fringes using two beams and some use multiple beams. Some interferometers are common path; some use lateral shear; some use radial shear; and some use phase-shifting techniques. Examples of the various types are given below, but first we must give the basic equation for two-beam interference.

Two-Beam Interference

When two quasi-monochromatic waves interfere, the irradiance of the interference pattern is given by

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos[\phi] \qquad [1]$$

where the I's represent the irradiance of the individual beams and ϕ is the phase difference between the two interfering beams. The maximums of the interference occur when $\phi = 2m\pi$, where m is an integer. In most applications involving interferometers, ϕ is the quantity of interest since it is related to the quantity being measured. Figure 1 shows typical two-beam interference fringes.

There are many two-beam interferometers and we will now look at a few of them.

Fresnel Mirrors

Fresnel Mirrors are a very simple two beam interferometer. A beam of light reflecting off two mirrors, set at a slight angle to one another, will produce an interference pattern. The fringes are straight and equally spaced. The fringes become more finely spaced as the angle between the two mirrors increases, d increases, and D becomes smaller (Figure 2).

Plane Parallel Plate

When a monochromatic source illuminates a plane parallel plate, as shown in Figure 3, the interference



Figure 1 Typical two-beam interference fringes.

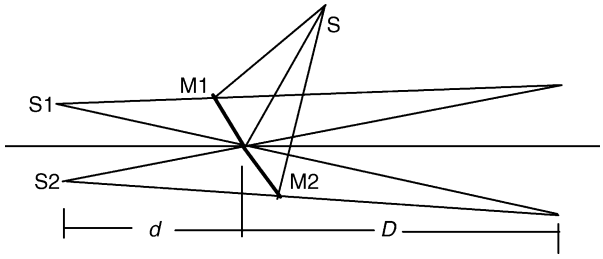


Figure 2 Fresnel mirrors.

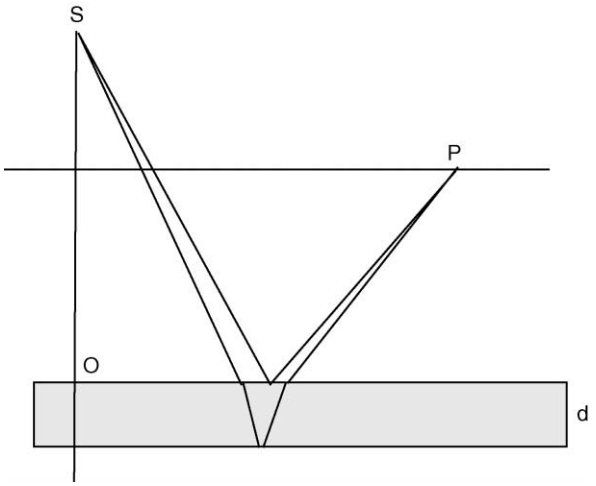


Figure 3 Plane parallel plate interferometer.

fringes are circles centered on the normal to plate. The fringes are called Haidinger fringes, or fringes of equal inclination since, for a given plate thickness, they depend on the angle of incidence. If the maximum or minimum occurs at the center, the radii of the fringes are proportional to the square root of integers. If the plate has a slight variation in thickness and it is illuminated with a collimated beam then the interference fringes are called Fizeau fringes of equal thickness and they give the thickness variations in the plate.

Michelson Interferometer

A Michelson interferometer is shown in Figure 4. Light from the extended source is split into two beams by the beamsplitter. The path difference can be viewed as the difference between the mirror M1 and the image of mirror M2, M2′. With the M1 and M2′ parallel, the fringes are circular and localized at infinity. With M1 and M2′ at a slight angle, the interference fringes are straight lines parallel to the equivalent intersection of the mirrors and localized approximately at the intersection. When the mirrors are only a few wavelengths apart, white light fringes

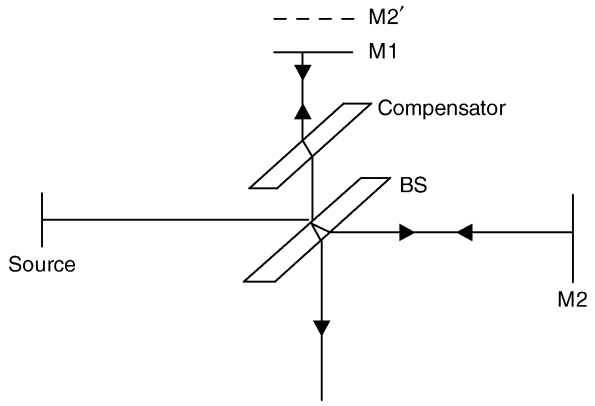


Figure 4 Michelson interferometer.

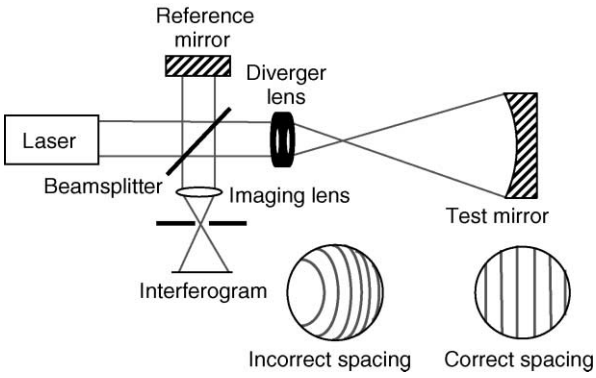


Figure 5 Twyman–Green interferometer for testing a concave spherical mirror.

appear, and can be used to determine their coincidence.

Twyman–Green Interferometer

If a point source is used in a Michelson interferometer it is generally called a Twyman–Green interferometer. Twyman–Green interferometers are often used to test optical components such as flat mirrors, curved mirrors, windows, lenses, and prisms. Figure 5 shows a Twyman–Green interferometer for the testing of a concave spherical mirror. The interferometer is aligned such that the focus of the diverger lens is at the center of curvature of the spherical mirror. If the mirror under test is perfect, straight equi-spaced fringes are obtained. Figure 6 shows two interference fringes and the relationship between surface height error for the mirror being tested and fringe deviation. λ is the wavelength of the light.

Fizeau Interferometer

The Fizeau interferometer is a simple device used for testing optical surfaces, especially flats and spheres

(Figure 7). The fringes are fringes of equal thickness. It is sometimes useful to tilt the reference surface a little to get several nearly straight interference fringes. The relationship between surface height error and fringe deviation from straightness is the same as for the Twyman–Green interferometer.

Mach–Zehnder Interferometer

A Mach–Zehnder interferometer is sometimes used to look at samples in transmission. Figure 8 shows a typical Mach–Zehnder interferometer for looking at a sample in transmission.

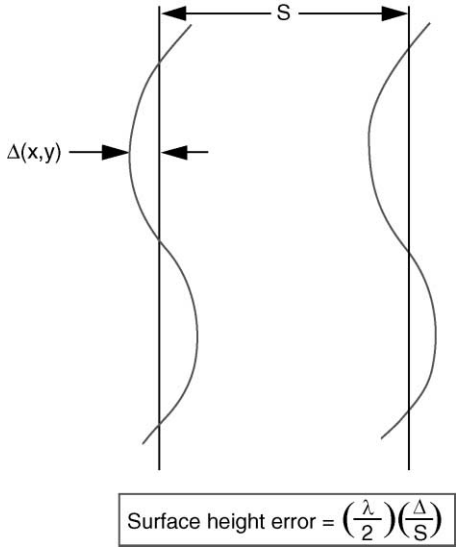


Figure 6 Relationship between surface height error and fringe deviation.

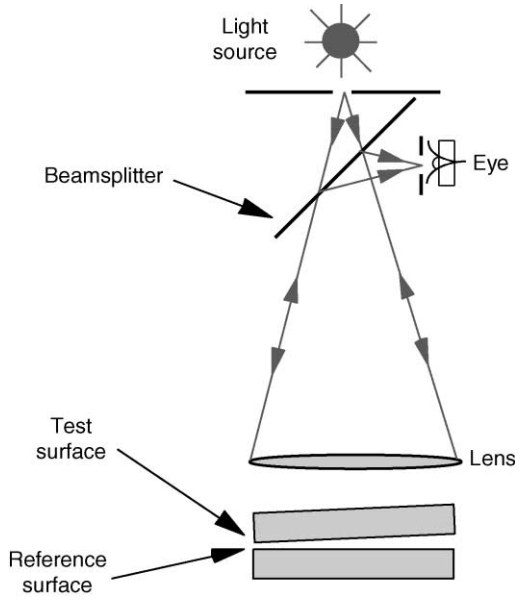


Figure 7 Fizeau interferometer.

Murty Lateral Shearing Interferometer

Lateral shearing interferometers compare a wavefront with a shifted version of itself. While there are many different lateral shear interferometers, one that works very well with a nearly collimated laser beam is a Murty shearing interferometer shown in Figure 9. The two laterally sheared beams are produced by reflecting a coherent laser beam off a plane parallel plate. The Murty interferometer can be used to measure the aberrations in the lens or it can be used to determine if the beam leaving the lens is collimated. If the beam incident upon the plane parallel plate is collimated a single fringe results, while if the beam is not perfectly collimated straight equi-spaced fringes result where the number of fringes gives the departure from collimation.

Radial Shear Interferometer

Radial shear interferometers compare a wavefront with a magnified or demagnified version of itself. While there are many different radial shear

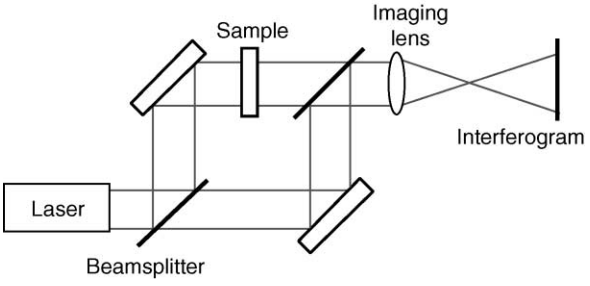


Figure 8 Mach–Zehnder interferometer.

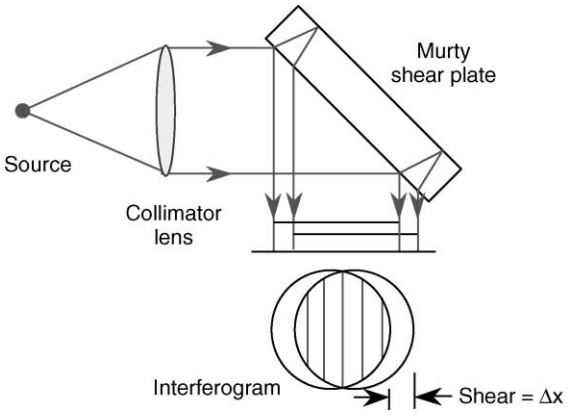


Figure 9 Murty lateral shear interferometer.

interferometers, one that works very well with a nearly collimated beam is shown in Figure 10. It is essentially a Mach–Zehnder interferometer where, in one arm, the beam is magnified and in the other arm the beam is demagnified. Interference fringes result in the region of overlap. The sensitivity of the interferometer depends upon the amount of radial shear. If the two interfering beams are approximately the same size there is little sensitivity, while if the two beams greatly differ in size there is large sensitivity. Sometimes radial shear interferometers are used test the quality of optical components.

Scatterplate Interferometer

A scatterplate interferometer, invented by Jim Burch in 1953, is one of the cleverest interferometers. Almost any light source can be used and no high-quality optics are required to do precision interferometry. The critical element in the interferometer is the scatterplate which looks like a piece of ground glass, but the important item is that the scattering points are arranged so the plate has inversion symmetry as shown in Figure 11. The light that is scattered the first time through the plate, and unscattered the second time through the plate, interferes with the light unscattered the first time and scattered the second time. The resulting interference fringes show errors in the mirror under test. This type of interferometer is insensitive to vibration because it is what is called a common path interferometer, in that the test beam (scattered–unscattered) and the reference beam (unscattered–scattered) travel along almost the same paths in the interferometer.

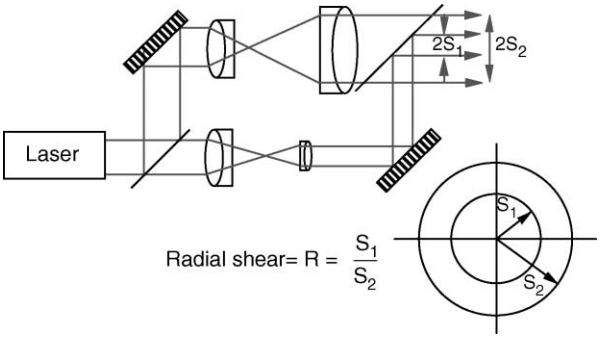


Figure 10 Radial shear interferometer.

Smartt Point Diffraction Interferometer

Another common path interferometer used for the testing of optics is the Smartt point diffraction interferometer (PDI) shown in Figure 12. In the PDI the beam of light being measured is focused onto a partially transmitting plate containing a pinhole. The pinhole removes the aberration from the light passing through it (reference beam), while the light passing through the plate (test beam) does not have the aberration removed. The interference of these two beams gives the aberration of the lens under test.

Sagnac Interferometer

The Sagnac interferometer has beams traveling in opposite directions, as shown in Figure 13. The interferometer is highly stable and easy to align. If the interferometer is rotated with an angular velocity there will be a delay between the transit times of the clockwise and the counterclockwise beams. For this reason, the Sagnac interferometer is used in laser gyros.

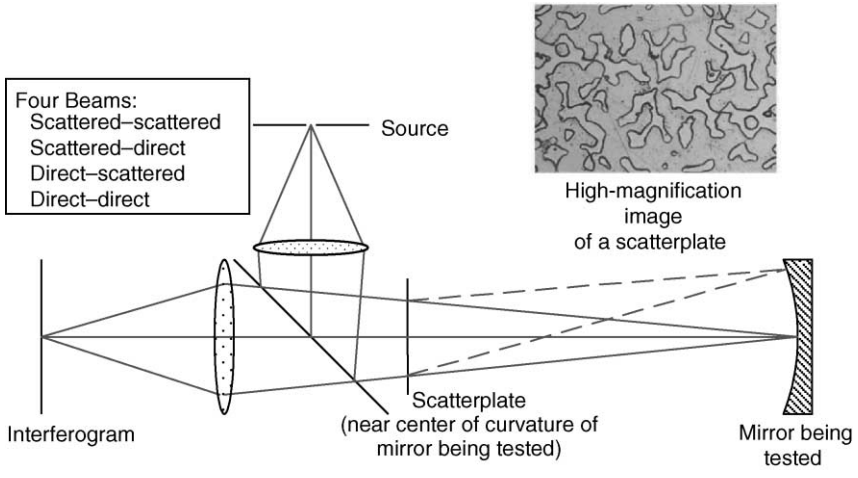


Figure 11 Scatterplate interferometer.

Fiber Interferometers

Fiber interferometers were first used for rotation sensing by replacing the ring cavity in the Sagnac interferometer with a multiloop made of a single-mode fiber. Fiber-interferometer rotation sensors are attractive as rotation sensors because they are small and low-cost.

Since the optical path in a fiber is affected by its temperature and it also changes when the fiber is stretched, fiber interferometers can be used as sensors for mechanical strains, temperature, and pressure. They can also be used for the measurement of magnetic fields by bonding the fiber to a magneto-restrictive element. Electric fields can be measured by bonding the fiber to a piezoelectric film.

Multiple Beam Interferometers

In general, multiple beam interferometers provide sharper interference fringes than a two beam interferometer. If light is reflected off a plane parallel plate there are multiple reflections. If the surfaces of the plane parallel plate have a low reflectivity, the higher-order reflections have a low intensity and the multiple reflections can be ignored and the resulting interference fringes have a sinusoidal intensity profile, but if the surfaces of the plane parallel plate have a high reflectivity, the fringes become sharper. Figure 14 shows the profile of the fringes for both the reflected and the transmitted light for different values of surface reflectivity. Multiple beam interference fringes are extremely useful for measuring the spectral content of a source. One such multiple beam interferometer, that is often used for measuring spectral distributions of a source, is a Fabry–Perot interferometer that consists of two plane-parallel plates separated by an air space, as shown in Figure 15.

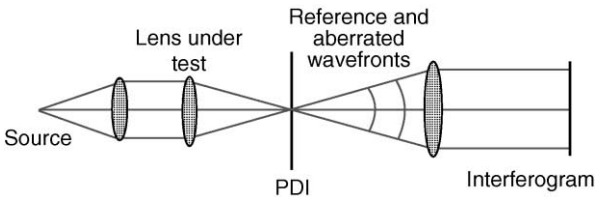


Figure 12 Smartt point diffraction interferometer.

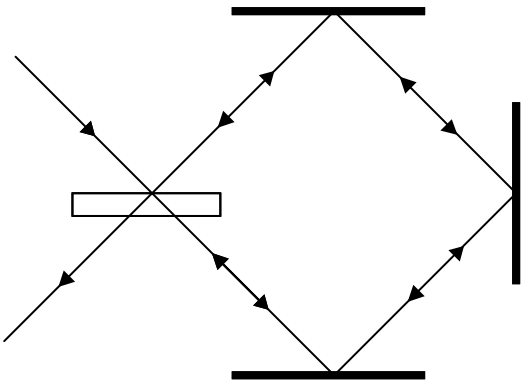


Figure 13 Sagnac interferometer.

Phase-Shifting Interferometry

The equation for the intensity resulting from two-beam interference, contains three unknowns, the two individual intensities and the phase difference between the two interfering beams. If three or more measurements are made of the intensity of the two-beam interference as the phase difference between the two beams is varied in a known manner, all three unknowns can be determined. This technique, called phase-shifting interferometry, is an extremely powerful tool for measuring phase distributions. Generally the phase difference is

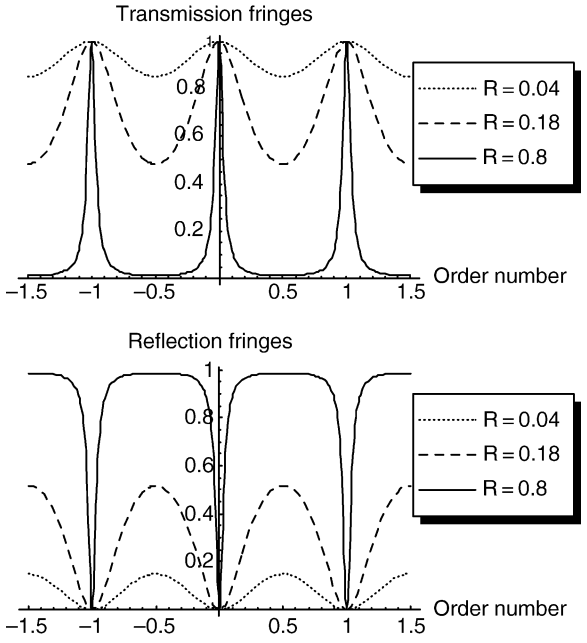


Figure 14 Multiple beam interference fringes for the transmitted and reflected light for a plane parallel plate.

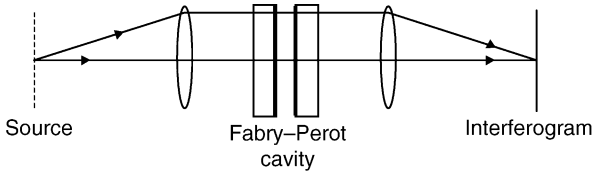


Figure 15 Fabry–Perot interferometer.

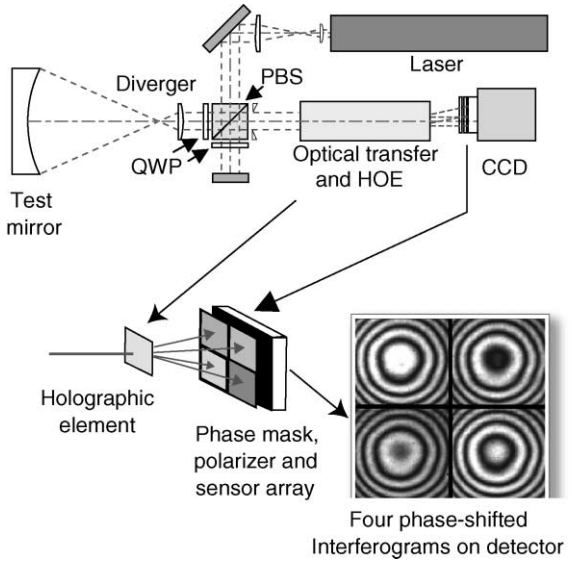


Figure 16 Vibration-insensitive phase-shifting interferometer.

changed by 90 degrees between consecutive measurements of the interference intensity in which case the three intensities can be written as

$$I_a(x,y) = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(\phi(x,y) - 90°)$$
$$I_b(x,y) = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(\phi(x,y))$$
$$I_c(x,y) = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(\phi(x,y) + 90°)$$

and the phase distribution is given by

$$\phi(x,y) = \mathrm{ArcTan}\left(\frac{I_a(x,y) - I_c(x,y)}{-I_a(x,y) + 2I_b(x,y) - I_c(x,y)}\right)$$

Phase-shifting interferometry has greatly enhanced the use of interferometry in metrology since it provides a fast low noise way of getting interferometric data into a computer.

Vibration-Insensitive Phase-Shifting Interferometer

While phase-shifting interferometry has greatly enhanced the use of interferometry in metrology, there are many applications where it cannot be used because of the environment, especially vibration. One recent phase-shifting interferometer that works well in the presence of vibration is shown in Figure 16. In this interferometer, four phase-shifted frames of data are captured simultaneously. The test and reference beams have orthogonal polarization. After the test and reference beams are combined they are passed through a holographic element to produce four identical beams. The four beams are then transmitted through wave retardation plates to cause 0, 90, 180, and 270 phase difference between the test and reference beams. After passing through a polarizer the four phase-shifted interferograms fall on the detector array. Not only can the effects of vibration be eliminated, but by making short exposures to freeze the vibration, the vibrational modes can be measured. Movies can be made showing the vibration. Likewise, flow fields can be measured.

Combining modern electronics, computers, and software with old interferometric techniques, provides for very powerful measurement capabilities.

List of Units and Nomenclature

I	irradiance
R	intensity reflectance
λ	wavelength
ϕ	phase

See also

Coherence: Speckle and Coherence. **Detection:** Fiber Sensors. **Holography, Techniques:** Computer–Generated Holograms; Digital Holography; Holographic Interferometry. **Imaging:** Adaptive Optics; Interferometric Imaging; Wavefront Sensors and Control (Imaging Through Turbulence). **Interferometry:** Phase Measurement Interferometry; White Light Interferometry; Gravity Wave Detection. **Microscopy:** Interference Microscopy. **Tomography:** Optical Coherence Tomography.

Further Reading

Candler C (1951) *Modern Interferometers*. London: Hilger and Watts.
Francon M (1966) *Optical Interferometry*. New York: Academic Press.
Hariharan P (1991) Selected papers on interferometry. *SPIE*, vol. 28.
Hariharan P (1992) *Basics of Interferometry*. San Diego: Academic Press.
Hariharan P (2003) *Optical Interferometry*. San Diego: Academic Press.
Hariharan P and Malacara D (1995) Selected papers on interference, interferometry, and interferometric metrology. *SPIE*, vol. 110.
Malacara D (1990) Selected papers on optical shop metrology. *SPIE*, vol. 18.
Malacara D (1992) *Optical Shop Testing*. New York: Wiley.
Malacara D (1998) *Interferogram Analysis for Optical Testing*. New York: Marcel Decker.
Robinson D and Reid G (1993) *Interferogram Analysis*. Brisol, UK: Institute of Physics.
Steel WH (1985) *Interferometry*. Cambridge: Cambridge University Press.
Tolansky S (1973) *An Introduction to Interferometry*. New York: Longman.

Gravity Wave Detection

N Christensen, Carleton College, Northfield, MN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In the later part of the nineteenth century Albert Michelson performed extraordinary experiments that shook the foundations of the physics world. Michelson's precise determination of the speed of light was an accomplishment that takes great skill to reproduce today. Edward Morley teamed up with Michelson to measure the velocity of the Earth with respect to the aether. The interferometer that they constructed was exquisite, and through amazing experimental techniques the existence of the aether was disproved. The results of Michelson and Morley led to a revolution in physics, and provided evidence that helped Albert Einstein to develop the general theory of relativity. Now the Michelson interferometer may soon provide dramatic confirmation of Einstein's theory through the direct detection of gravitational radiation.

An accelerating electric charge produces electromagnetic radiation – light. It should come as no surprise that an accelerating mass produces gravitational light, namely gravitational radiation. In 1888 Heinrich Hertz had the luxury to produce and detect electromagnetic radiation in his laboratory. There will be no such luck with gravitational radiation because gravity is an extremely weak force.

Albert Einstein postulated the existence of gravitational radiation in 1916, and in 1989 Joe Taylor and Joel Weisberg indirectly confirmed its existence through observations of the orbital decay of the binary pulsar 1913 + 16. Direct detection of gravity waves will be difficult. For humans to have any chance of detecting gravity waves there must be extremely massive objects accelerating up to relativistic velocities. The only practical sources are astrophysical: supernovae, pulsars, neutron star or black hole binary systems, black hole formation or even the Big Bang. The observation of these types of events would be extremely significant for contributing to knowledge in astrophysics and cosmology. Gravity waves from the Big Bang would provide information about the Universe at its earliest moments. Observations of supernovae will yield a gravitational snapshot of these extreme cataclysmic events. Pulsars are neutron stars that can spin on their axes at frequencies in the hundreds of hertz, and the signals from these objects will help to decipher their characteristics. Gravity waves from the final stages of coalescing binary neutron stars could help to accurately determine the size of these objects and the equation of state of nuclear matter. The observation of black hole formation from these binary systems would be the *coup de grace* for the debate on the existence of black holes, and the ultimate triumph for general relativity.

Electromagnetic radiation has an electric field transverse to the direction of propagation, and a charged particle interacting with the radiation will experience a force. Similarly, gravitational radiation will produce a transverse force on massive objects, a tidal force. Explained via general relativity it is more accurate to say that gravitational radiation will deform the fabric of space-time. Just like electromagnetic radiation there are two polarizations for gravity waves. Let us imagine a linearly polarized gravity wave propagating in the z-direction, $h(z,t) = h_{0+}e^{i(kz-\omega t)}$. The fabric of space is stretched due to the strain created by the gravity wave. Consider a length L_0 of space along the $\boldsymbol{x}$-axis. In the presence of the gravity wave the length oscillates like

$$L(t) = L_0 + \frac{h_{0+}L}{2}\cos\omega t$$

hence there is a change in its length of

$$\Delta L_x = \frac{h_{0+}L}{2}\cos\omega t$$

A similar length L_0 of the $\boldsymbol{y}$-axis oscillates, like

$$L(t) = L_0 - \frac{h_{0+}L}{2}\cos\omega t$$

or

$$\Delta L_y = \frac{-h_{0+}L}{2}\cos\omega t$$

One axis stretches while the perpendicular one contracts, and then vice versa, as the wave propagates through. The other polarization (h_{0x}) produces a strain on axes 45° from ($\boldsymbol{x}$,$\boldsymbol{y}$). Imagine some astrophysical event produces a gravity wave that has amplitude h_{0+} on Earth; in order to detect a small distance displacement ΔL one should have a detector that spans a large length L. A supernova within our own Galaxy might possibly produce a gravity wave of size $h \sim 10^{-18}$ with characteristic frequencies around

1 kHz, but the occurrence of an event such as this will be extremely rare. More events will come from novae in other galaxies, but due to the great distances and the fact that the magnitude of the gravity wave falls of as $1/r$, such events will be substantially diminished in magnitude.

A Michelson interferometer, with arms aligned along the x- and y- axes, can measure small phase differences between the light in the two arms. Therefore, this type of interferometer can turn the length variations of the arms produced by a gravity wave into changes in the interference pattern of the light exiting the system. This was the basis of the idea from which modern laser interferometric gravitational radiation detectors have evolved. Imagine a gravity wave of amplitude h is incident on an interferometer. The change in the arm length will be $h \sim \Delta L/L_0$, so in order to optimize the sensitivity it is advantageous to make the interferometer arm length L_0 as large as possible. Detectors coming on-line are attempting to measure distance displacements that are of order $\Delta L \sim 10^{-18}$ m or smaller, less that the size of an atomic nucleus! If the interferometers can detect such distance displacements, and directly detect gravitational radiation, it will be one of the most spectacular accomplishments in experimental physics. The first implementation of a laser interferometer to detect a gravity wave was by Forward who used earphones to listen to the motion of the interference signal. The engineering of signal extraction for modern interferometers is obviously far more complex.

Numerous collaborations are building and operating advanced interferometers in order to detect gravitational radiation. In the United States there is the Laser Interferometer Gravity Wave Observatory (LIGO), which consists of two 4-km interferometers located in Livingston, Louisiana and Hanford, Washington. In addition, there is an additional 2-km interferometer within the vacuum system at Hanford. An Italian–French collaboration (VIRGO) has a 3-km interferometer near Pisa, Italy. GEO, a German–British collaboration, is a 600-m detector near Hanover, Germany. TAMA is the Japanese 300-m interferometer in Tokyo. The Australians (AIGO) are constructing a 500-m interferometer in Western Australia. All of the detectors will be attempting to detect gravity waves with frequencies from about 50 Hz to 1 kHz.

As will be described below, there are a number of terrestrial noise sources that will inhibit the performance of the interferometric detectors. The sensitivity of detection increases linearly with interferometer arm length, which implies that there could be advantages to constructing a gravity wave detector in space. This is the goal of the Laser Interferometer Space Antenna (LISA) collaboration. The plan is to deploy three satellites in a heliocentric orbit with a separation of about 5×10^6 km. The launch for LISA is planned for around 2015, and the detector will be sensitive to gravity waves within the frequency range of 10^{-3} to 10^{-1} Hz. Due to the extremely long baseline, LISA is not strictly an interferometer, as most light will be lost as the laser beams expand while traveling such a great distance. Instead, the phase of the received light will be detected and used to lock the phase of the light that is re-emitted by another laser.

Joe Weber pioneered the field of gravitational radiation detection in the 1960s. He used a 1400-kg aluminum cylinder as an antenna. Gravitational waves would hopefully excite the lowest-order normal mode of the bar. Since the gravity wave is a strain on space-time, experimentalists win by making their detectors longer in length. This provides a long-term advantage for laser interferometers, which can scale in length relatively easily. However, as of 2002, interferometers and bars (now cooled to 4 K) have comparable sensitivities.

Interferometer Styles

The Michelson interferometer is the tool to be used to detect a gravitational wave. Figure 1 shows a basic system. The beamsplitter and the end mirrors would be suspended by wires, and effectively free to move in the plane of the interferometer. The arms have lengths L_1 and L_2 that are roughly equal on a kilometer scale. With a laser of power P and wavelength λ incident on the beamsplitter, the light

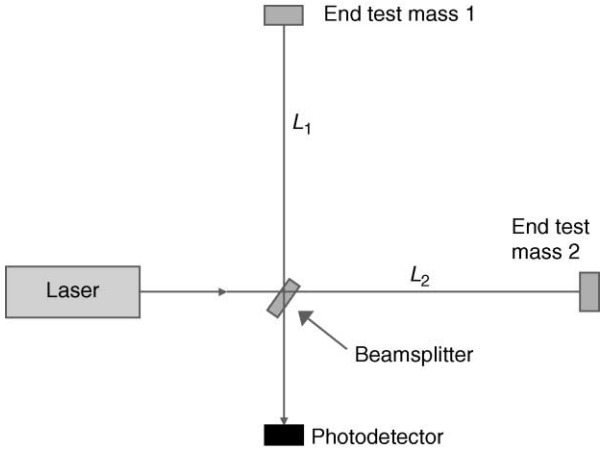


Figure 1 A basic Michelson interferometer. The photodetector receives light exiting the dark port of the interferometer and hence the signal.

exiting the dark port of the interferometer is

$$P_{\text{out}} = P \sin^2\left[\frac{2\pi}{\lambda}(L_1 - L_2)\right]$$

The interferometer operates with the condition that in the absence of excitation the light exiting the dark port is zero. This would be the case for a simple and basic interferometer. However, interferometers like those in LIGO will be more sophisticated, and will use a heterodyne detection strategy. If E_0 is the amplitude of the electric field from the laser, and assuming the use of a 50–50 beamsplitter, the electric field (neglecting unimportant common phase shifts) for the light incident on the photodetector would be

$$E_{\text{out}} = \frac{E_0}{2}\left(e^{i\delta\phi_1} - e^{i\delta\phi_2}\right) \approx i\frac{E_0}{2}(\phi_1 - \phi_2)$$

$$= iE_0\frac{2\pi}{L}(L_1 - L_2)$$

The laser light will be phase modulated at a frequency in the MHz regime. As such, the deconvolved current from the photodiode that detects light at the interferometer output will be proportional to the phase acquired by the light, namely $I \propto \frac{2\pi}{\lambda}(L_1 - L_2)$.

A gravity wave of optimal polarization normally incident upon the interferometer plane will cause one arm to decrease in length while the other increases. The Michelson interferometer acts as a gravity wave transducer; the change in arm lengths results in more light exiting the interferometer dark port. The mirrors in the interferometer are suspended via wires so that they are free to move under the influence of the gravity wave.

An interferometer's sensitivity increases with arm length, but geographical and financial constraints will limit the size of the arms. If there could be some way to bounce the light back and forth to increase the effective arm length it would increase the detector performance. Fabry–Perot cavities on resonance light have a storage time of $2L/[c(1 - \sqrt{R_1R_2})]$. Figure 2 shows the system of a Michelson interferometer with Fabry–Perot cavities. This gravity wave interferometer design was proposed and tested in the late 1970s by Ron Drever. The far mirror R_2 has a very high reflectivity ($R_2 \sim 1$) in order to ultimately direct the light back towards the beamsplitter. The front mirror reflectivity R_1 is such that LIGO's effective arm length increases to $L \sim 300$ km. The optical properties of the mirrors of the Fabry–Perot cavities must be exquisite in order to achieve success. LIGO's mirrors were tested, and the root mean squared surface uniformity is less than 1 nm, scattered light is less than 50 parts per million (ppm), absorption is less than 2 ppm, and the radii of curvature for the mirrors are matched to less than 3%. A LIGO test mass (and therefore a Fabry–Perot mirror) can be seen in Figure 3.

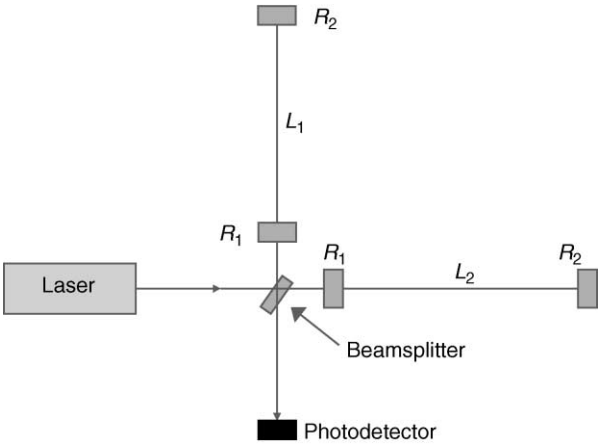


Figure 2 A Michelson interferometer with Fabry–Perot cavities in each arm. The front cavity mirrors have reflectivity R_1 while the end mirrors have $R_2 \sim 1$. By using Fabry–Perot cavities LIGO will increase the effective arm length by a factor of 75.

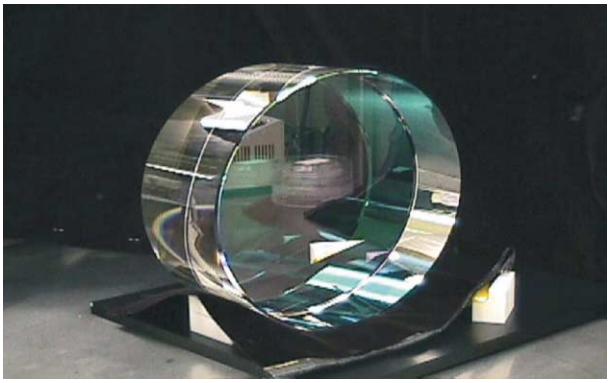


Figure 3 A picture of a mirror and test mass for LIGO. The fused silica component is 10.7 kg in mass and 25 cm in diameter. Photograph courtesy of Caltech/LIGO.

In 1888 Michelson and Morley, with their interferometer, had a sensitivity that allowed the measurement of 0.02 of a fringe, or about 0.126 radian. Prototype interferometers constructed by the LIGO science team have already demonstrated a phase noise spectral density of $\phi(f) = 10^{-10}$ radian/$\sqrt{\text{Hz}}$ for frequencies above 500 Hz. Assuming a 1 kHz signal with 1 kHz bandwidth this implies a phase sensitivity of $\Delta\phi = 3.2 \times 10^{-9}$ radian. This is about the phase sensitivity that LIGO hopes to accomplish in the 4-km Fabry–Perot system. There has been quite an evolution in interferometry since Michelson's time.

The noise sources that inhibit the interferometer performance are discussed below. However, let us

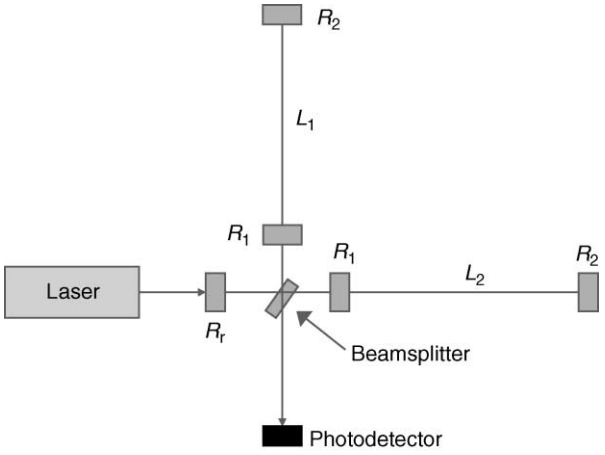


Figure 4 A power recycled Michelson interferometer with Fabry–Perot cavities in each arm. Normally light would exit the interferometer through the light port and head back to the laser. Installation of the recycling mirror with reflectivity R_r sends the light back into the system. A Fabry–Perot cavity is formed between the recycling mirror and the first mirror (R_1) of the arms. For LIGO this strategy will increase the power circulating in the interferometer by a factor of 50.

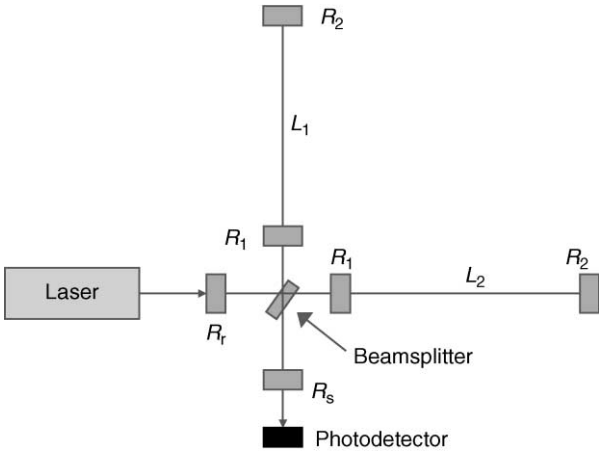


Figure 5 A signal recycled and power recycled Michelson interferometer with Fabry–Perot cavities in each arm. Normally light containing the gravity wave signal would exit the interferometer through the dark port and head to the photodetector. Installation of the signal recycling mirror with reflectivity R_s sends the light back into the system. The phase of the light acquired from the gravity wave will build up at a particular frequency determined by the reflectivity R_s.

consider one's ability to measure the relative phase between the light in the two arms. The Heisenberg uncertainty relation for light with phase ϕ and photon number N is $\Delta\phi\Delta N \sim 1$. For a measurement lasting time τ using laser power P and frequency f, the photon number is $N = P\lambda\tau/hc$, and with Poisson statistics describing the light $\Delta N = \sqrt{N} = \sqrt{P\lambda\tau/hc}$. Therefore

$$\Delta\phi\Delta N = \frac{2\pi}{\lambda}\Delta L\sqrt{P\lambda\tau/hc} = 1$$

implies that

$$\Delta L = \frac{1}{2\pi}\sqrt{\frac{hc\lambda}{P\tau}}$$

With more light power the interferometer can measure smaller distance displacements and achieve better sensitivity. LIGO will use about 10 W of laser power, and will eventually work towards 100 W. However, there is a nice trick one can use to produce more light circulating in the interferometer, namely power recycling. Figure 4 displays the power recycling interferometer design. The interferometer operates such that virtually none of the light exits the interferometer dark port, and the bulk of the light returns towards the laser. An additional mirror, R_r in Figure 4, recycles the light. For LIGO, recycling will increase the effective light power by another factor of 50. The higher circulating light power therefore improves the sensitivity.

There is one additional modification to the interferometer system that can further improve sensitivity, but only at a particular frequency. A further Fabry–Perot system can be made by installing what is called a signal recycling mirror; this would be mirror R_s in Figure 5. Imagine light in arm 1 on the interferometer that acquires phase as the arm expands due to the gravity wave. The traveling gravity wave's oscillation will subsequently cause arm 1 to contract while arm 2 expands. If the light that was in arm 1 could be sent to arm 2 while it is expanding, then the beam would acquire additional phase. This process could be repeated over and over. Mirror R_s serves this purpose, with its reflectivity defining the storage time for light in each interferometer arm. The storage time defined by the cavity formed by the signal recycling mirror, R_s, and the mirror at the front of the interferometer arm cavity, R_1, determines the resonance frequency. Signal recycling will give a substantial boost to interferometer sensitivity at a particular frequency, and will eventually be implemented in all the main ground-based interferometric detectors.

The LIGO interferometers are infinitely more complex than the relatively simple systems displayed in the figures of this paper. Figure 6 presents an aerial view of the LIGO site at Hanford, Washington. The magnitude of the 4-km system is apparent.

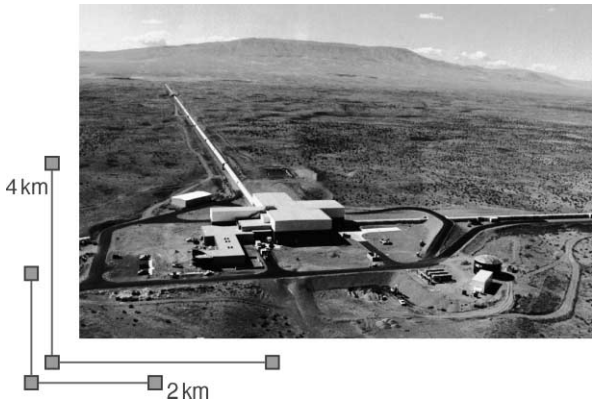


Figure 6 Arial view of the LIGO Hanford, Washington site. The vacuum enclosure at Hanford contains both 2-km and 4-km interferometers. Photograph courtesy of Caltech/ LIGO.

Noise Sources and Interferometer Sensitivity

If the interferometers are to detect distance displacements of order $\Delta L \sim 10^{-18}$ m then they must be isolated from a host of deleterious noise sources. Seismic disturbances should not shake the interferometers. Thermal excitation of components will affect the sensitivity of the detector and should be minimized. The entire interferometer must be in an adequate vacuum in order to avoid fluctuations in gas density that would cause changes in the index of refraction and hence modification of the optical path length. The laser intensity and frequency noise must be minimized. The counting statistics of photons influences accuracy. If ever there was a detector that must avoid Murphy's law this is it; little things going wrong cannot be permitted if such small distance displacements are to be detected. The expected noise sensitivity for the initial LIGO interferometers is displayed in Figure 7.

In the best of all worlds the interferometer sensitivity will be limited by the counting statistics of the photons. A proper functioning laser will have its photon number described by Poisson statistics, or shot noise; if the mean number of photons arriving per unit time is N then the uncertainty is $\Delta N = \sqrt{N}$, which as noted above implies an interferometer displacement sensitivity of

$$\Delta L = \frac{1}{2\pi}\sqrt{\frac{hc\lambda}{P\tau}}$$

or a noise spectral density of

$$\Delta L(f) = \frac{1}{2\pi}\sqrt{\frac{hc\lambda}{P}} \text{ (in units of m}/\sqrt{\text{Hz}})$$

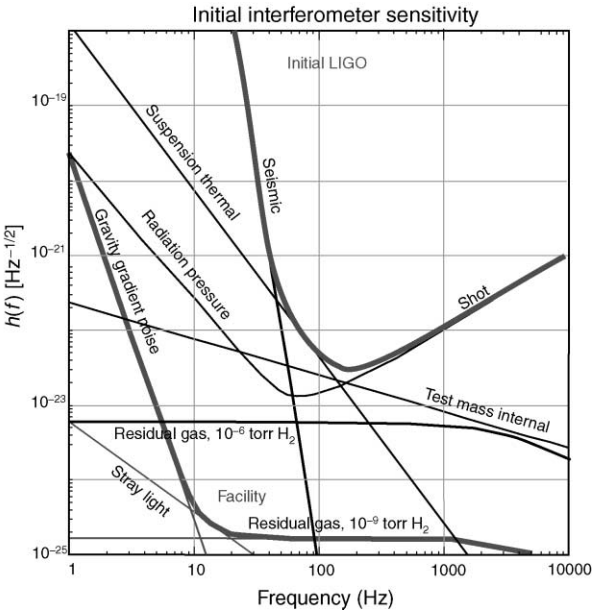


Figure 7 The target spectral density of the noise for the initial LIGO system. LIGO will be dominated by seismic noise at low frequencies (10–100 Hz), thermal noise (from the suspension system and internal modes within the mirrors) in the intermediate regime (100–300 Hz), and photon shot noise thereafter. Other sources of noise are also noted, specifically gravity gradients (gravitational excitation of masses from the seismic motion of the ground), radiation pressure of photons on the mirrors, stray (scattered) light, and index of refraction fluctuations from residual gas in the vacuum.

Note also that the sensitivity increases as the light power increases. The reason for this derives from the statistics of repeated measurements. The relative lengths of the interferometer arm lengths could be measured, once, by a photon. However, the relative positions are measured repeatedly with every photon from the laser, and the variance of the mean decreases as $\sqrt{N}$, where N is the number of measurements (or photons) involved. The uncertainty in the difference of the interferometer arm lengths is therefore inversely proportional to photon number, and hence the laser's power. In terms of strain sensitivity this would imply

$$h(f) = \frac{1}{2\pi L}\sqrt{\frac{hc\lambda}{P}} \text{ (in units of } 1/\sqrt{\text{Hz}})$$

This assumes the light just travels down the arm and back once. With Fabry–Perot cavities the light is stored, and the typical photon takes many trips back and forth before exiting the system. In order to maximize light power the end mirrors ($R_2 \sim 1$) and

the strain sensitivity is improved to

$$h(f) = \frac{1}{2\pi L'}\sqrt{\frac{hc\lambda}{P}}$$

where $L' = 4L/(1 - R_1)$.

As the frequency of gravity waves increases the detection sensitivity will decrease. If the gravity wave causes the interferometer arm length to increase, then decrease, while the photons are still in the arm cavity, then the phase acquired from the gravity wave will be washed away. This is the reason why interferometer sensitivity decreases as frequency increases, and explains the high-frequency behavior seen in Figure 7. Taking this into account, the strain sensitivity is

$$h(f) = \frac{1}{2\pi L'}\sqrt{\frac{hc\lambda}{P}}\left(1 + \left(\frac{2\pi L' f}{c}\right)^2\right)^{1/2}$$

where $L' = 4L/(1 - R_1)$ and f is the frequency of the gravity wave.

If the gravitational wave is to change the interferometer arm length then the mirrors that define the arm must be free to move. In systems like LIGO wires suspend the mirrors; each mirror is like a pendulum. While allowing the mirrors to move under the influence of the gravity wave is a necessary condition, the pendulum itself is the first component of an elaborate vibration isolation system. Seismic noise will be troublesome for the detector at low frequencies. The spectral density of the seismic noise is about $(10^{-7}/f^2)\mathrm{m}/\sqrt{\mathrm{Hz}}$ for frequencies above 1 Hz. A simple pendulum, by itself, acts as a motion filtering device. Above its resonance frequency the pendulum filters motion with a transfer function like $T(f) \propto (f_0/f)^2$. Detectors such as LIGO will have a pendulum with resonant frequencies of about $f_0 \approx 1$ Hz, thus providing an isolation of 10^4 when looking for signals at $f = 100$ Hz. The various gravity wave detector collaborations have different vibration isolation designs. The mirrors in these interferometers will be suspended in elaborate vibration isolation systems, which may include multiple pendulum and isolation stacks. Seismic noise will be the limiting factor for interferometers seeking to detect gravity waves in the tens of hertz range, as can be seen in the sensitivity curve presented in Figure 7.

Due to the extremely small distance displacements that these systems are trying to detect it should come as no surprise that thermal noise is a problem. This noise enters through a number of components in the system. The two most serious thermal noise sources are the wires suspending the mirrors in the pendulum, and the mirrors themselves. Consider the wires; there are a number of modes that can oscillate (i.e. violin modes). At temperature T each mode will have energy of k_BT, but distributed over a band of frequencies determined by the quality factor (or Q) of the material. Low-loss (or high-Q) materials work best; for the violin modes of the wires there will be much noise at particular frequencies (in the hundreds of hertz). The mirror is a cylindrical object, which will have normal modes of oscillation that can be thermally excited. The first generation of LIGO will have these masses composed of fused silica, which is typical for optical components. The Qs for the internal modes are greater than 2×10^6. A switch may eventually be made to sapphire mirrors, which have better thermal properties. The limitation to the interferometers' sensitivity due to the thermal noise internal to the mirrors can be seen in Figure 7, and will be a worrying noise source for the first-generation LIGO in the frequency band around 100–400 Hz.

The frequency noise of the laser can couple into the system to produce length displacement sensitivity noise in the interferometer. With arm lengths of 4 km, it will be impossible to hold the length of the two arms absolutely equal. The slightly differing arm spans will mean that the light sent back from each of the two Fabry–Perot cavities will have slightly differing phases. As a consequence, great effort is made to stabilize the frequency of the light entering the interferometer. The LIGO laser can be seen in Figure 8. The primary laser is a Lightwave Model 126 nonplanar ring oscillator. High power is generated from this stabilized laser through the use of optical amplifiers. The beam is sent through four optical amplifiers, and then retro-reflected back

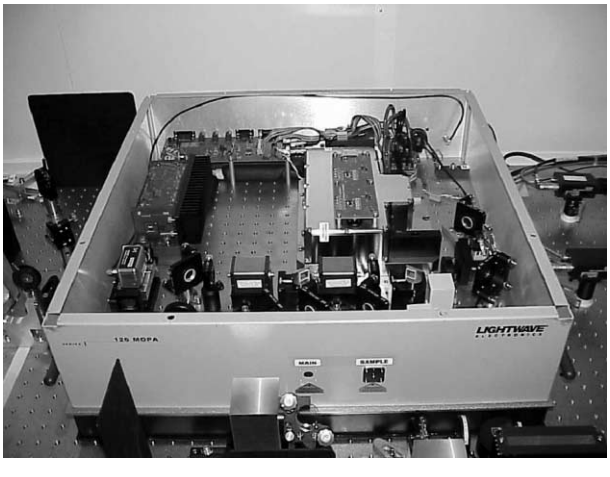


Figure 8 A picture of the Nd:YAG laser and amplifier system that produces 10 W of light for LIGO. Photograph courtesy of Caltech/LIGO.

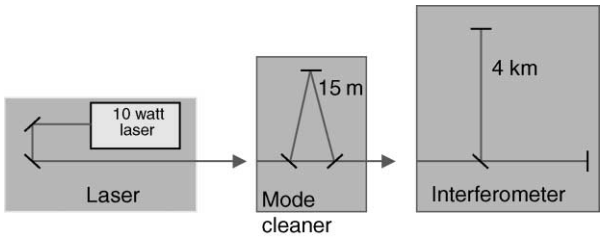


Figure 9 The optical system for the LIGO system. Light moves from the frequency-stabilized laser to a mode cleaner, and then to the interferometer. Presently 8 W of TEM_{00} laser power is delivered to the interferometer.

through the amplifiers again. For LIGO, the laser is locked and held to a specific frequency by use of signals from a reference cavity, a mode cleaner cavity, and the interferometer. For low-frequency stabilization the temperature of the ring oscillator is adjusted. At intermediate frequencies adjustment is made by signals to a piezo-electric transducer within the ring oscillator cavity. At high frequencies the noise is reduced with the use of an electro-optic crystal. The LIGO lasers currently have a frequency noise of $2 \times 10^{-2}\,\mathrm{Hz}/\sqrt{\mathrm{Hz}}$ at frequencies above 1 kHz.

It will prove important to worry about the stability of the laser power for the interferometric detectors. The hope is to be shot noise limited at frequencies above a few hundred hertz. The Nd:YAG power amplifiers used are pumped with an array of laser diodes, so the light power is controlled through feedback to the laser diodes. The LIGO requirements for the fluctuations on the power P are $\Delta P/P < 10^{-8}/\sqrt{\mathrm{Hz}}$. The spatial quality of the light is ensured through the use of a mode-cleaning cavity. LIGO uses a triangular array of mirrors separated by 15 m. LIGO's input optical system can be seen in Figure 9. The current LIGO optical system yields 8 W of 1.06 μm light in the TEM_{00} mode.

Conclusion

The attempt to measure gravitational radiation with laser interferometers could possibly be the most difficult optical experiment of our time. Over a hundred years ago Michelson succeeded in carrying off experiments of amazing difficulty as he measured the speed of light and disproved the existence of the aether. Gravity wave detection is an experiment worthy of Michelson, and there are hundreds of physicists striving to make it a reality.

Great success has already been achieved. The TAMA 300-m interferometer is operational and has achieved a sensitivity $h(f) = 5 \times 10^{-21}/\sqrt{\mathrm{Hz}}$ in the 700 Hz to 1.5 kHz frequency band. The LIGO interferometers are now operating, and scientists are presently de-bugging the system in order to achieve the target sensitivity. The first scientific data taking for LIGO commenced in 2002, and is continuing at the time of writing.

This will be a new telescope to peer into the heavens. With every new means of looking at the sky there has come unexpected discoveries. Physicists do know that there will be signals that they can predict. Binary systems of compact objects (neutron stars or black holes) will produce chirp signals that may be extracted by matched filtering techniques. A supernova will produce a burst that will hopefully rise above the noise. Pulsars, or neutron stars spinning about their axes at rates sometimes exceeding hundreds of revolutions per second, will produce continuous sinusoidal signals that can be seen by integrating for sufficient lengths of time. Gravity waves produced by the Big Bang will produce a background stochastic noise that can possibly be extracted by correlating the outputs from two or more detectors. These are exciting physics results that will come through tremendous experimental effort.

See also

Imaging: Interferometric Imaging. **Interferometry:** Overview.

Further Reading

Barish BC (1997) Gravitational wave detection. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 155–161. Tokyo: Universal Academic.

Blair DG, Taniwaki M, Zhao CN, *et al.* (1997) Progress in the development of technology for advanced laser interferometric gravitational wave detectors. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 75–94. Tokyo: Universal Academic.

Brillet A (1997) Virgo status report. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 163–173. Tokyo: Universal Academic.

Danzmann K (1997) LISA–A gravitational wave observatory in heliocentric orbit. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 111–116. Tokyo: Universal Academic.

Drever RWP, Ford GM, Hough J, *et al.* (1980) A gravity-wave detector using optical cavity sensing. In: Schmutzer E (ed.) *Proceedings of the Ninth International Conference on General Relativity and Gravitation*, pp. 265–267. Berlin: Springer.

Forward RL (1978) Wideband laser-interferometer gravitational-radiation experiment. *Physical Review* D17: 379.

Hough J and the GEO 600 Team (1997) GEO 600 Current status and some aspects of the design. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 175–182. Tokyo: Universal Academic.

Robertson NA (2000) Laser interferometric gravitational wave detectors. *Classical Quantum Gravity* 17: R19–R40.

Saulson PR (1994) Fundamentals of Interferometric Gravitational Wave Detectors. Singapore: World Scientific.

Taylor JH and Weisberg JM (1989) Further experimental tests of relativistic gravity using the binary pulsor PSR 1913 + 16. *Astrophysical Journal* 345: 434–450.

Thorne KS (1987) Gravitational radiation. In: Hawking S and Isreal W (eds) *300 Years of Gravitation*, pp. 330–458. Cambridge, UK: Cambridge University Press.

Tsubono K and the TAMA collaboration (1997) TAMA Project, Gravitational wave detection. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 183–191. Tokyo: Universal Academic.

Weiss R (1972) Electromagnetically coupled broadband gravitational antenna. *MIT Quarterly Progress Report* no. 105.

Phase-Measurement Interferometry

K Creath, Optineering and University of Arizona, Tucson, AZ, USA

J Schmit, Veeco Instruments, Inc., Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Phase-measurement interferometry is a way to measure information encoded in interference patterns generated by an interferometer. Fringe patterns (fringes) created by interfering beams of light are analyzed to extract quantitative information about an object or phenomenon. These fringes are localized somewhere in space and require a certain degree of spatial and temporal coherence of the source to be visible. Before these techniques were developed in the late 1970s and early 1980s, fringe analysis was done either by estimating fringe deviation and irregularity by eye or by manually digitizing the centers of interference fringes using a graphics tablet. Digital cameras and desktop computers have made it possible to easily obtain quantitative information from fringe patterns. The techniques described in this article are independent of the type of interferometer used.

Interferometric techniques using fringe analysis can measure features as small as a micron wide or as large as a few meters. Measurable height ranges vary from a few nanometers up to 10s of mm, depending upon the interferometric technique employed. Measurement repeatability is very consistent, and typically repeatability of 1/100 rms of a fringe is easily obtainable while 1/1000 rms is possible. Actual measurement precision depends on what is being measured and how the technique is implemented, while accuracy depends upon comparison to a sanctified standard.

There are many different types of applications for fringe analysis. For example, optical surface quality can be determined using a Twyman–Green or Fizeau interferometer. In addition, wavefront quality measurements of a source or an optical system can be made in transmission, and the index of refraction and homogeneity of optical materials can be mapped out. Many nonoptical surfaces can also be measured. Typically, surface topography information at some specific spatial frequency scale is extracted. These measurements are limited by the resolution of the optical system and the field of view of the imaging system. Lateral and vertical dimensions can also be measured. Applications of nonoptical surfaces include disk and wafer flatness, roughness measurement, distance and range sensing. Phase measurement techniques used in holographic interferometry, TV holography, speckle interferometry, moiré, grating interferometry, and fringe projection are used for nondestructive testing to measure surface structure as well as displacements due to stress and vibration. This article outlines the basics of phase measurement interferometry (PMI) techniques as well as the types of algorithms used. The bibliography lists a number of references for further reading.

Background

Basic Parts of a Phase-Measuring Interferometer

A phase-measuring interferometer consists of a light source, an illumination system (providing uniform illumination across the test surface), a beamsplitter (usually a cube or pellicle so both beams have the same optical path), a reference surface (needs to be good because these techniques measure the difference between the reference and test surface), a sample

fixture, an imaging system (images a plane in space where the surface is located onto the camera), a camera (usually a monochrome CCD), an image digitizer/frame grabber, a computer system, software (to control the measurement process and calculate the surface map), and often a spatial or temporal phase shifter to generate multiple interferograms.

Steps of the Measurement Process

To generate a phase map, a sample is placed on a sample fixture and aligned, illumination levels are adjusted, the sample image is focused onto the camera, the fringes are adjusted for maximum contrast, the phase shifter or fringe spacing is adjusted or calibrated as necessary, a number of images is obtained and stored with the appropriate phase differences, the optical path difference (OPD) is calculated as the modulo 2π phase and then is unwrapped at each pixel to determine the phase map.

To make consistent measurements, some interferometers need to be on vibration isolation systems and away from heavy airflows or possible acoustical coupling. It helps to cover any air paths that are longer than a few millimeters. Consideration needs to be made for consistency of temperature and humidity.

The human operating the interferometer is also a factor in the measurement. Does this person always follow the same procedure? Are the measurements sampled consistently? Are the test surfaces clean? Is the sample aligned the same and correctly? Many different factors can affect a measurement. To obtain repeatable measurements it is important to have a consistent procedure and regularly verify measurement consistency.

Common Interferometer Types

One of the most common interferometers used in optical testing is the Twyman–Green interferometer (Figure 1). Typically, a computer controls a mirror pushed by a piezo-electric transducer (PZT). The test surface is imaged onto the camera and the computer has a frame grabber that takes frames of fringe data.

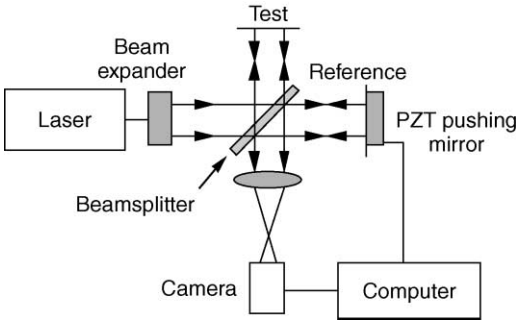


Figure 1 Twyman–Green interferometer.

The most common commercially available interferometer for optical testing is the Fizeau interferometer (Figure 2). This versatile instrument is very insensitive to vibrations due to the large common path for both interfering wavefronts. The reference surface (facing the test object) is moved by a PZT to provide the phase shift.

Interference microscopes (*see* **Microscopy:** Interference Microscopy. **Interferometry:** White Light Interferometry) are used for looking at surface roughness and small structures (see Figure 3). These instruments can employ Michelson, Mirau, and Linnik interference objectives with a laser or white light source, or Fizeau-type interference objectives that typically use a laser source because of the unequal paths in the arms of the interferometer. The phase shift is accomplished by moving the sample, the reference surface, or parts of the objective relative to the sample. Figure 3 shows a schematic of a Mirau-type interferometric microscope for phase measurement.

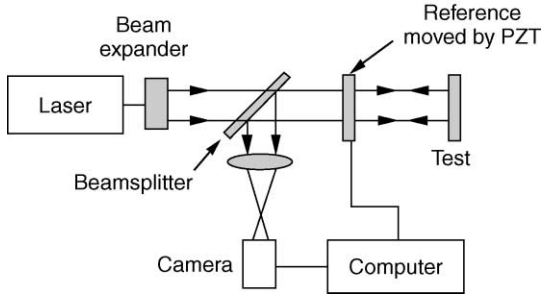


Figure 2 Fizeau interferometer.

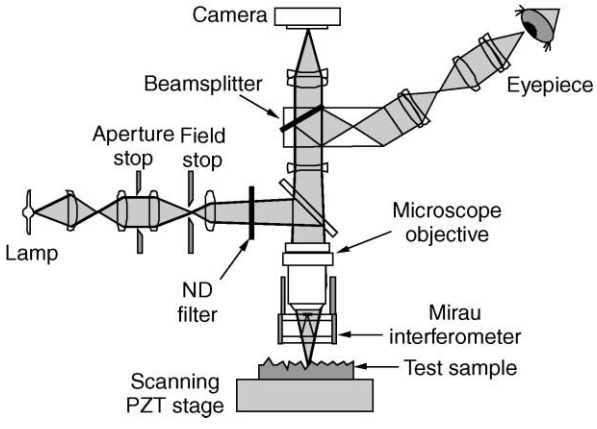


Figure 3 Mirau interference microscope.

Determination of Phase

The Interference Equation

Interference fringes from a coherent source (e.g., a laser) are theoretically sinusoidal (Figure 4a), while fringes from an incoherent source (e.g., white light) are localized in a wavepacket at a point in space (Figure 4b) where the optical paths of the arms of the interferometer are equal and marked on Figure 4 as scanner positions $z = 0$ (*see* **Interferometry:** White Light Interferometry). For generality, this analysis considers the determination of phase within the wave packet for fringes localized in space.

Interference fringes at any point in the wavepacket can be written in the following form:

$$I(x, y) = I_0\{1 + \gamma(x, y, z)\cos[\phi(x, y)]\} \qquad [1]$$

where I_0 is the dc irradiance, γ is the fringe visibility (or contrast), $2I_0\gamma$ is the modulation (irradiance amplitude or the ac part of the signal), and ϕ is the phase of the wavefront as shown in Figure 5. For simplicity, this drawing assumes that the interference fringe amplitude is constant.

Fringe visibility can be determined by calculating

$$\gamma = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \qquad [2]$$

where $I_{\max}$ is the maximum value of the irradiance for all phase values and $I_{\min}$ is the minimum value. The fringe visibility has a real value between 0 and 1 and varies with the position along the wavepacket. The fringe visibility, as defined here, is the real part of the complex degree of coherence.

Types of Phase Measurement Techniques

Phase can be determined from either a number of interference fringe patterns or from a single interferogram with appropriate fringe spacing. Temporal techniques require an applied phase shift between the test and reference beams as a function of time while multiple frames of interference fringe data are obtained. Spatial techniques can obtain data from a single interferogram that requires a carrier pattern of almost straight fringes to either compare phases of adjacent pixels or to separate orders while performing operations in the Fourier domain. Spatial techniques may also simultaneously record multiple interferograms with appropriate relative phase shift differences separated spatially in space. Multiple frame techniques require more data than single frame techniques. Temporal techniques require that the interference fringes be stable over the time period it takes to acquire the number of images. While single-frame spatial techniques require less data and can be done with a single image, they generally have a reduced amount of resolution and less precision than temporal techniques.

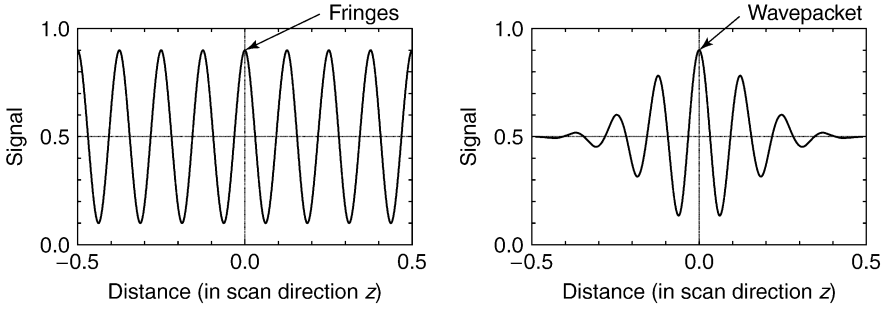
There are literally hundreds of algorithms and techniques for extracting phase data from interference fringe data. The references listed in the Further Reading offer more details of these techniques.

Temporal Phase Measurement

Temporal techniques use data taken as the relative phase between the test and reference beams is modulated (shifted). The phase (or OPD) is calculated at each measured point in the interferogram. As the phase shifter is moved, the phase at a single point in the interferogram changes. The effect looks like the fringes are moving across the interferogram, and because of these techniques are sometimes called fringe scanning or fringe shifting techniques. However, the fringes are not really moving; rather the irradiance at a single detector point is changing (hopefully sinusoidally) in time (see Figure 6). A 180° or π phase causes a bright fringe to become a dark fringe.

Phase Modulation Techniques

There are many ways to introduce a phase modulation (or shift). These include moving a mirror or the



(a) Coherent source (b) Incoherent source

Figure 4 Interference fringes for coherent and incoherent sources as observed at point $x, y, z = 0$ corresponds to equal optical path lengths of the reference and object beams.

sample, tilting a glass plate, moving a diffraction grating, rotating a polarizer, analyzer, or half-wave plate, using a two-frequency (Zeeman) laser source, modulating the source wavelength, and switching an acousto-optic Bragg cell, or magneto-optic/electro-optic cell. While any of these techniques can be used for coherent sources, special considerations need to be made for temporally incoherent sources (*see* **Interferometry:** White Light Interferometry). Figure 7 shows how moving a mirror introduces a relative phase shift between object and reference beams in a Twyman–Green interferometer.

Extracting Phase Information

Including the phase shift, the interference equation is written as

$$I(x, y) = I_0(x, y)\{1 + \gamma_0(x, y)\cos[\phi(x, y) + \alpha(t)]\} \quad [3]$$

where $I(x, y)$ is the irradiance at a single detector point, $I_0(x, y)$ is the average (dc) irradiance, $\gamma_0(x, y)$ is the fringe visibility before detection, $\phi(x, y)$ is the phase of the wavefront being measured, and $\alpha(t)$ is the phase shift as a function of time.

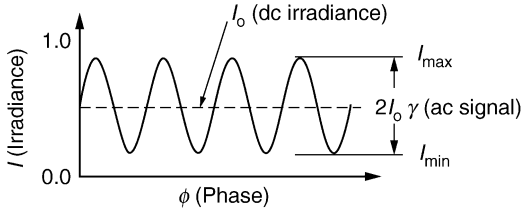


Figure 5 Variable definitions for interference fringes.

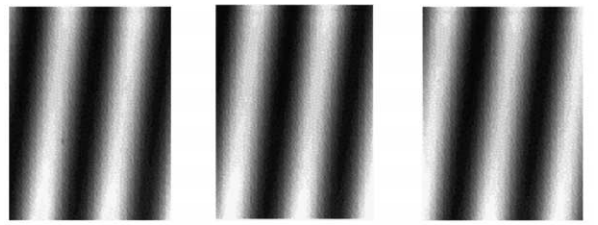


Figure 6 Interference fringe patterns corresponding to different relative phase shifts between test and reference beams.

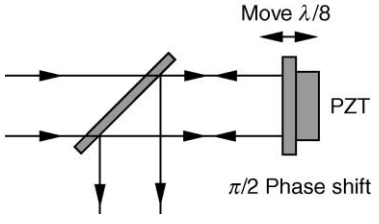


Figure 7 Moving a mirror to shift relative phase by λ/4.

Since the detector has to integrate for some finite time, the detected irradiance at a single point becomes an integral over the integration time Δ (Figure 8) where the average phase shift for the *i*th frame of data is α_i:

$$I_i(x, y) = \frac{1}{\Delta}\int_{\alpha_i-(\Delta/2)}^{\alpha_i+(\Delta/2)} I_0(x, y)\{1 + \gamma_0(x, y) \times \cos[\phi(x, y) + \alpha(t)]\}\mathrm{d}\alpha(t) \quad [4]$$

After integrating over $\alpha(t)$, the irradiance of the detected signal becomes

$$I_i(x, y) = I_0(x, y)\left\{1 + \gamma_0(x, y)\mathrm{sinc}\left(\frac{\Delta}{2}\right) \times \cos[\phi(x, y) + \alpha_i]\right\} \quad [5]$$

where $\mathrm{sinc}(\Delta/2) = \sin(\Delta/2)/(\Delta/2)$, which reduces the detected visibility.

Ramping Versus Stepping

There are two different ways of shifting the phase; either the phase shift can be changed in a constant and linear fashion (ramping) (see Figure 8) or it can be stepped in increments. Ramping provides a continuous smooth motion without any jerking motion. This option may be preferred if the motion does not wash out the interference fringes. However, ramping requires good synchronization of the camera/digitizer and the modulator to get the correct OPD (or phase) changes between data frames. Ramping allows faster data taking but requires electronics that are more sophisticated. In addition, it takes a finite time for a mass to move linearly. When ramping, the first frame or two of data usually needs to be discarded because the shift is not correct until the movement is linear.

The major difference between ramping and stepping the phase shift is a reduction in the modulation of the interference fringes after detection (the $\mathrm{sinc}(\Delta/2)$ term in eqn [5]). When the phase shift is stepped ($\Delta = 0$), the sinc term has a value of one. When the phase shift is ramped ($\Delta = \alpha$) for a phase

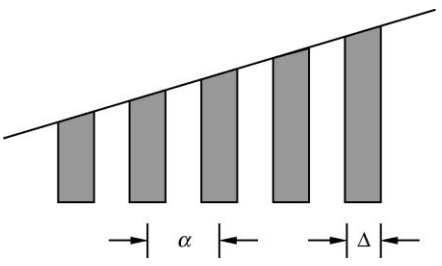


Figure 8 Change in OPD integrating over a Δ phase shift with sample separation α.

shift of $\alpha = 90° = \pi/2$, this term has a value of 0.9. Therefore, ramping slightly reduces the detected fringe visibility.

Signal Modulation

Phase measurement techniques assume that the irradiance of the interference fringes at the camera covers as much of the detector's dynamic range as possible and that the phase of the interference fringe pattern is modulating at each individual pixel as the phase between beams is modulated. If the irradiance at a single detector point does not modulate as the relative phase between beams is shifted, the height of the surface cannot be calculated. Besides the obvious reductions in irradiance modulation due to the detector sampling and pixel size or a bad detector element, scattered light within the interferometer, and defects or dirt on the test object, can also reduce signal modulation. Phase measurement techniques are designed to take into account the modulation of the signal at each pixel. If the signal does not modulate enough at a given pixel, then the data at that pixel are considered unusable, flagged as 'bad' and are often left blank. Phase values for these points may be interpolated from surrounding pixels if there are sufficient data.

Three-Frame Technique

The simplest method to determine phase uses three frames of data. With three unknowns, three sets of recorded fringe data are needed to reconstruct a wavefront providing a phase map. Using phase shifts of $\alpha_i = \pi/4$, $3\pi/4$, and $5\pi/4$, the three fringe measurements at a single point in the interferogram may be expressed as

$$I_1 = I_0\left[1 + \gamma\cos\left(\phi + \frac{\pi}{2}\right)\right] = I_0\left[1 + \frac{\sqrt{2}}{2}\gamma(\cos\phi - \sin\phi)\right] \quad [6]$$

$$I_2 = I_0\left[1 + \gamma\cos\left(\phi + \frac{3\pi}{2}\right)\right] = I_0\left[1 + \frac{\sqrt{2}}{2}\gamma(-\cos\phi - \sin\phi)\right] \quad [7]$$

$$I_3 = I_0\left[1 + \gamma\cos\left(\phi + \frac{5\pi}{2}\right)\right] = I_0\left[1 + \frac{\sqrt{2}}{2}\gamma(-\cos\phi + \sin\phi)\right] \quad [8]$$

Note that (x, y) dependencies are still implied. The choice of the specific phase shift values is to make the math simpler. The phase at each detector point is

$$\phi = \tan^{-1}\left(\frac{I_3 - I_2}{I_1 - I_2}\right) \quad [9]$$

In most fringe analysis techniques, we are basically trying to solve for terms such that we end up with the tangent function of the phase. The numerator and denominator shown above are respectively proportional to the sine and cosine of the phase. Note that the dc irradiance and fringe visibility appear in both the numerator and denominator. This means that variations in fringe visibility and average irradiance from pixel to pixel do not affect the results. As long as the fringe visibility and average irradiance at a single pixel is constant from frame to frame, the results will be good. If the different phase shifts are multiplexed onto multiple cameras, the results will be dependent upon the gain of corresponding pixels.

Bad data points with low signal modulation are determined by calculating the fringe visibility at each data point using:

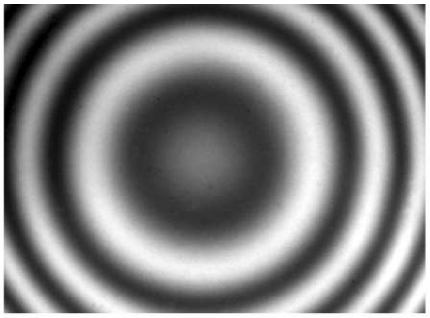
$$\gamma = \frac{\sqrt{(I_3 - I_2)^2 + (I_1 - I_2)^2}}{\sqrt{2}I_0} \quad [10]$$

It is simpler to calculate the ac signal modulation $(2I_0\gamma)$ and then set the threshold on the modulation to a typical value of about 5–10% of the dynamic range. If the modulation is less than this value at any given data point, the data point is flagged as bad. Bad points are usually caused by noisy pixels and can be due to scratches, pits, dust and scattered light.

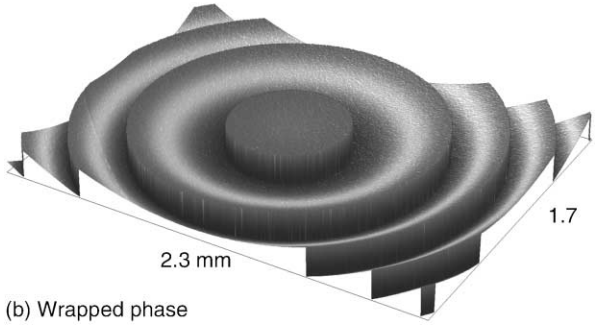
Although three frames of data are enough to determine the phase, this and other 3-frame algorithms are very sensitive to systematic errors due to nonsinusoidal fringes or nonlinear detection, phase shifter miscalibration, vibrations and noise. In general, the larger the number of frames of data used to determine the phase, the smaller the systematic errors.

Phase Unwrapping

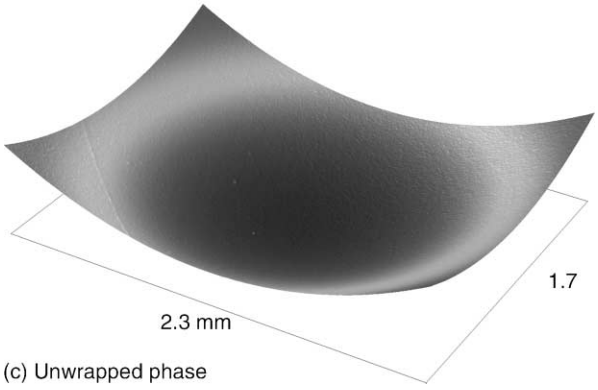
The removal of phase ambiguities is generally called phase unwrapping, and is sometimes known as integrating the phase. The phase ambiguities owing to the modulo 2π arctangent calculation can simply be removed by comparing the phase difference between adjacent pixels. When the phase difference between adjacent pixels is greater than π, a multiple of 2π is added or subtracted to make the difference less than π. For the reliable removal of



(a) Fringes for concave surface



(b) Wrapped phase



(c) Unwrapped phase

Figure 9 Fringes, wrapped, and unwrapped phase maps.

discontinuities, the phase must not change by more than π ($\lambda/2$ in optical path (OPD)) between adjacent pixels. Figure 9 shows an example of wrapped and unwrapped phase values.

Given a trouble-free wrapped phase map, it is enough to search row by row (or column by column) for phase differences of more than π between neighboring pixels. However, fringe patterns usually are not perfect and are affected by systematic errors. Some of the most frequently occurring error sources are noise, discontinuities in the phase map, violation of the sampling theorem and invalid data points (e.g., due to holes in the object or low modulation regions). Some phase maps are not easily unwrapped. In these cases different techniques like wave packet peak sensing in white light interferometry are used.

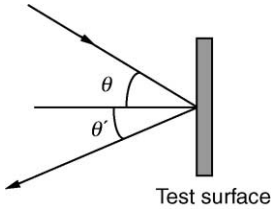


Figure 10 Angles of illumination and viewing.

From Wavefront to Surface

Once the phase of the wavefront is known, surface shape can be determined from the phase map. Surface height H of the test surface relative to the reference surface at a location (x, y) is given by

$$H(x, y) = \frac{\phi(x, y)\lambda}{2\pi(\cos\theta + \cos\theta')} \qquad [11]$$

where λ is the wavelength of illumination, and θ and θ' – the angles of illumination and viewing with respect to the surface normal – are shown in Figure 10. For interferometers (e.g., Twyman–Green or Fizeau) where the illumination and viewing angles are normal to the surface ($\theta = \theta' = 0$), the surface height is simply:

$$H(x, y) = \frac{\lambda}{4\pi}\phi(x, y) \qquad [12]$$

Since the wavefront measured represents the relative difference between the interfering reference and test wavefronts, this phase map only directly corresponds to the surface under test when the reference wavefront is perfectly flat. In practice, the shape of the reference surface needs to be accounted for by measuring it using a known test surface and subtracting this reference measurement from subsequent measurements of the test surface.

Phase Change on Reflection

Phase shifting interferometry measures the phase of reflected light to determine the shape of objects. The reflected wavefront will represent the object surface (within a scaling factor) if the object is made of a single material. If the object is comprised of multiple materials that exhibit different phase changes on reflection, the measured wavefront needs to be corrected for these phase differences (*see* **Interferometry:** White Light Interferometry).

Overview of Phase Measurement Algorithms and Techniques

There are literally hundreds of published algorithms and techniques. The optimal algorithm depends on

the application. Most users prefer fast algorithms using a minimal amount of data that are as accurate and repeatable as possible, immune to noise, adaptable, and easy to implement. In practice, there are trade-offs that must be considered when choosing a specific algorithm or technique. This section provides an overview of the types of algorithms to aid the reader in sifting through published algorithms.

Synchronous Detection

One of the first techniques for temporal phase measurement utilized methods of communication theory to perform synchronous detection. To synchronously detect the phase of a noisy sinusoidal signal, the signal is first correlated (or multiplied) with sinusoidal and cosinusoidal reference signals (signals in quadrature) of the same frequency and then averaged over many periods of oscillation. This method of synchronous detection as applied to phase measurement can be extracted from the least squares estimation result when the phase shifts are chosen such that N measurements are equally spaced over one modulation period. With phase shifts α_i such that:

$$\alpha_i = \frac{2\pi}{N} i, \qquad \text{with } i = 1, \ldots, N \qquad [13]$$

the phase can be calculated from

$$\phi(x, y) = \tan^{-1}\left[\frac{\sum I_i(x, y)\sin \alpha_i}{\sum I_i(x, y)\cos \alpha_i}\right] \qquad [14]$$

Note that N can be any number of frames (or samples). The more frames of data, the smaller the systematic errors. This technique does not take large amounts of memory for a large number of frames, because only the running sums of the fringes multiplied by the sine and cosine of the phase shift need to be remembered. The 4-frame algorithm from Table 1 is an example of a direct adaptation of synchronous detection where simple values of 1, -1 or 0 for every $\pi/2$ phase shift can be assigned to the sine and cosine functions.

Algorithm Design

In the last ten years, a lot of work has been done to generalize the derivation of fringe analysis algorithms. This work has enabled the design of algorithms for specific applications, which are insensitive to specific systematic error sources.

Most algorithms use polynomials for the numerator and denominator. Given fringe data:

$$I_i = I_0[1 + \gamma \cos(\phi + \alpha_i)] \qquad [15]$$

Table 1 Sampling function weights for a few selected algorithms

N frames	*Phase shift*	*Coefficients*
3	$\frac{\pi}{2}$	$\frac{1, -1, 0}{0, 1, -1}$
4	$\frac{\pi}{2}$	$\frac{0, -1, 0, 1}{1, 0, -1, 0}$
5	$\frac{\pi}{2}$	$\frac{0, -2, 0, 2, 0}{1, 0, -2, 0, 1}$
7	$\frac{\pi}{3}$	$\frac{\sqrt{3}[0, 1, 1, 0, -1, -1, 0]}{-1, -1, 1, 2, 1, -1, -1}$
8	$\frac{\pi}{2}$	$\frac{1, 5, -11, -15, 15, 11, -5, -1}{1, -5, -11, 15, 15, -11, -5, 1}$
12	$\frac{\pi}{3}$	$\frac{\sqrt{3}[0, -3, -3, 3, 9, 6, -6, -9, -3, 3, 3, 0]}{2, 1, -7, -11, -1, 16, 16, -1, -11, -7, 1, 2}$

the phase is calculated using

$$\phi = \tan^{-1}\left[\frac{\sum_i n_i I_i}{\sum_i d_i I_i}\right] \qquad [16]$$

The numerator and denominator of the arctangent argument are both polynomials. The numerator is a sum proportional to the sine (imaginary part) and the denominator is a sum proportional to the cosine (real part),

$$\text{num} = 2kI_0\gamma \sin \alpha_i = \sum_i n_i I_i \qquad [17]$$

$$\text{den} = 2kI_0\gamma \cos \alpha_i = \sum_i d_i I_i \qquad [18]$$

where the constant k depends on the values of coefficients. From this the fringe visibility is given by

$$\gamma = \frac{\sqrt{(\text{num})^2 + (\text{den})^2}}{2kI_0} \qquad [19]$$

The coefficient vectors for the numerator and denominator are window functions. For an algorithm such as the 4-frame technique, the weights of all samples are equal [1, 1, 1, 1]. This makes the coefficients all equal. For other algorithms, such as the 5-frame technique, the weights are larger on the middle frames than on the outer frames. The weights for the 5-frame technique are [1, 2, 2, 2, 1]. A property of the coefficient vectors is that the sum of the coefficients for each, the numerator and denominator, should be zero. Examples of coefficient vectors for a few selected algorithms are given in Table 1.

Heterodyne Interferometry

Historically, heterodyne techniques were developed and used before temporal phase-shifting techniques. These techniques generally determine the phase electronically at a single point by counting fringes and fractions of fringes. Areas are analyzed by scanning a detector.

Phase shifts are usually obtained using two slightly different frequencies in the reference and test beams. The beat frequency produced by the interference between the reference and test beams is compared to a reference sinusoidal signal, which may be produced either optically or electronically. The time delay (or distance traveled) between the crossing of the zero phase points of the test and reference sinusoidal signals is a measure of the phase. Every time the test signal passes through another zero in the same direction as the test surface is moved, another fringe can be counted. This is how fringe orders are counted. If the beam is interrupted as the detector is scanned across the interferogram, the fringe count is corrupted and the measurement needs to be started again. Frequency multiplication (harmonics) can also be used to determine fractions of fringes. Today, heterodyne techniques are used mainly in distance measuring interferometers. The precision and accuracy of distance measuring interferometers is at least on the order of 1 part in 10^6.

Fourier-Transform Technique

The Fourier-transform technique is a way to extract phase from a single interferogram. It is used a lot in nondestructive testing and stellar interferometry where it is difficult to get more than a single interferogram. The basic technique is shown schematically in Figure 11. The recorded interferogram distribution is Fourier transformed, and one order (usually the + or − first order) is either isolated and shifted to zero frequency or filtered out using a rectangular window. After an inverse Fourier transform, the result is the phase.

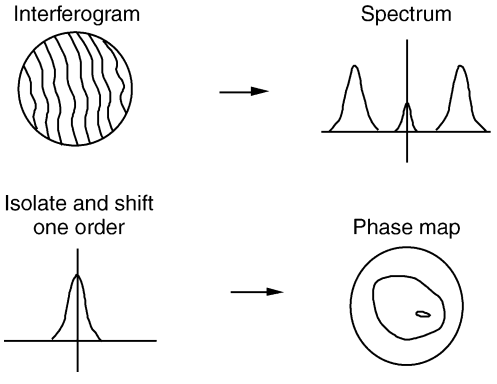


Figure 11 Fourier transform technique.

To illustrate this technique mathematically, the interference equation is rewritten as

$$I(x,y) = I_0(x,y) + c(x,y)\exp(\mathrm{i}2\pi f_0 x) + c^*(x,y)\exp(-\mathrm{i}2\pi f_0 x) \quad [20]$$

where $c(x,y) = I_0(x,y)\gamma(x,y)\exp[\mathrm{i}\phi(x,y)]$ and the $*$ indicates a complex conjugate. The term $c(x,y)$ contains the phase information we wish to extract. After performing a one-dimensional Fourier transform:

$$I(\xi,y) = I_0(\xi,y) + c(\xi - f_0, y) + c^*(\xi - f_0, y) \quad [21]$$

where ξ is the spatial frequency in the x direction, and italics indicate Fourier transforms. The next step is to filter out and isolate the second term, and then inverse Fourier transform to yield $c(x,y)$. The wavefront modulo 2π is then given by

$$\phi(x,y) = \tan^{-1}\left\{\frac{\mathrm{Im}[c(x,y)]}{\mathrm{Re}[c(x,y)]}\right\} \quad [22]$$

where Re and Im refer to the real and imaginary part of the function.

This technique has limitations. If the fringes are nonsinusoidal, there will not be a simple distribution in the frequency space; there will be many orders. Another problem is overlapping orders. There needs to be a carrier frequency present that ensures that the orders are separated in frequency space. This carrier frequency is produced by adding tilt fringes until the orders are separated. This means that the aberration (fringe deviation) has to be less than the fringe spacing. Another problem is aliasing. If the interferogram is not sampled sufficiently, there will be aliasing and it will not be possible to separate the orders in frequency space. Finally, large variations in average fringe irradiance and fringe visibility across the interferogram can also cause problems.

Spatial Carrier-Frequency Technique

This is essentially the equivalent of the Fourier transform technique but is performed in the spatial domain. It is also used when there is only one interferogram available and its major applications include nondestructive testing and measurement of large optics.

These techniques relate closely to the temporal phase-measurement methods; however, instead of using a number of interferograms they can obtain all the information from a single interferogram. As an example, let's assume that the fringes are vertical and parallel to the columns on the detector array.

The carrier frequency (i.e., the number of tilt fringes) is adjusted so that there is an α phase change from the center of one pixel to the next. As long as there is not much aberration (deviation), the phase change from pixel to pixel across the detector array will be approximately constant.

When the fringes are set up this way, the phase can be calculated using adjacent pixels (see Figure 12). If one fringe takes up 4 pixels, the phase shift α between pixels will be 90°. An algorithm such as the three-frame, four-frame, or five-algorithm can be used with adjacent pixels as the input. Therefore, 3, 4, or 5 pixels in a row will yield a single-phase point. Then, the analysis window is shifted sideways one pixel and phase is calculated at the next point. This technique assumes that the dc irradiance and fringe visibility do not change over the few pixels used to calculate each phase value.

Spatial Multichannel Phase-Shift Techniques

These techniques detect all phase maps simultaneously and multiplex the phase shift using static optical elements. This can be done by using either separate cameras as illustrated below or by using different detector areas to record each of the interferograms used to calculate the phase. The phase is usually calculated using the same techniques that are used for the temporal phase techniques.

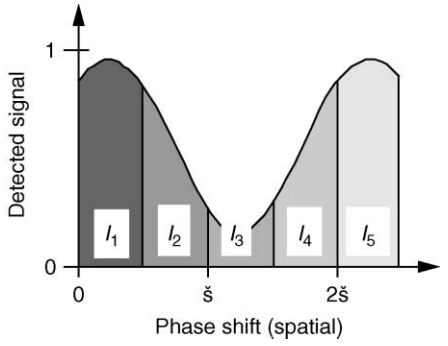


Figure 12 Spatial phase shifting. Different relative phase shifts are on adjacent detector elements.

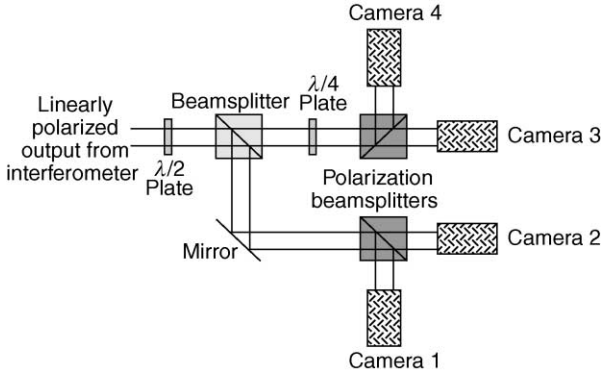


Figure 13 Spatial phase shifting using four cameras.

As an example, a four-channel interferometer can be made using the setup shown in Figure 13 to record four interferograms with 90° phase shifts between them. Camera 1 will yield fringes shifted 180° with respect to camera 2, and cameras 3 and 4 will have phase shifts of 90° and 270°. The optical system may also utilize a holographic optical element to split the beam to multiplex the four phase shifts on four quadrants of a single camera.

Signal Demodulation Techniques

The task of determining phase can be broadened by looking toward the field of signal processing. For communication via radio, radar, and optical fibers electrical engineers have developed a number of ways of compressing and encoding a signal as well as decompressing a signal and decoding it. An interference fringe pattern looks a lot like an am radio signal. Thus, it can be demodulated in similar ways. In recent years many new algorithms have been developed by drawing on techniques from communication theory and applying them to interferogram processing. Many of these use different types of transforms such as Hilbert transforms for straight fringes or circular transforms for closed fringes.

Extended Range Phase Measurement Techniques

A major limitation of phase measurement techniques is that they cannot determine surface discontinuities larger than $\lambda/4$ ($\lambda/2$ in optical path (OPD)) between adjacent pixels. One obvious solution is to use longer wavelength sources in the infrared where optically rough surfaces look smooth and their shape can be measured. An alternative is two-wavelength interferometry where two measurements at different wavelengths are taken and the measurable height limitation is now determined by the equivalent wavelength:

$$\lambda_{eq} = \frac{\lambda_1 \lambda_2}{|\lambda_1 - \lambda_2|} \qquad [23]$$

Another method that allows for measurement of smooth surfaces with large height discontinuities that is limited only by the working distance of the objective combines a white light and a phase measurement interferometric technique in one long scan. The position of the wave packet resolves 2π ambiguities that result from the arctangent function (see Phase unwrapping section above). The phase of the fringe close to the wave packet maximum is determined without ambiguity. Sometimes it is not the step height discontinuity that is a problem but rather the high slope of the measured surface. If we know that the surface is continuous, then the

unwrapping procedure can take advantage of this a priori information to look for a continuous.

Techniques for Deformation Measurement

Some of the techniques for deformation measurement have already been described earlier (see sections on Single frame techniques and Spatial multichannel phase-shift techniques). However, fringes in any interferometer can be analyzed not only in the x, y planes but also in the x, z, or y, z planes for which the carrier frequency of the fringes is introduced. Analysis of fringes in any plane has the same restrictions.

If the initial static object shape is measured using conventional phase measurement techniques, not only can the motion of each object point be determined but also the deformation of the whole object in time. If the motion of the object is periodic, then the object motion can be 'frozen' by using stroboscopic illumination of the same frequency as the object's motion. Once the motion is frozen, any temporal technique can be used to measure object shape at some phase of its motion. Changing the time offset between stroboscopic illumination and the periodic signal driving the motion, the object can be 'frozen' and thus measured at different phases of its periodic motion.

Systematic Errors

Noise Sensitivity

Measurement noise mostly arises from random fluctuations in the detector readout and electronics. This noise reduces precision and repeatability. Averaging multiple measurements can reduce effects of these random fluctuations.

Phase Shifter Errors

Phase shifter errors can be due to both miscalibration of the system and nonlinearities in the phase shift. It is possible to purchase very linear phase shifters. It is also possible to correct nonlinearities by determining the voltage signal making the phase shifter provide a linear phase shift.

Linear phase shifter errors (miscalibration) of the phase shift have an error signature that is at twice the frequency of the fringes. If there are two fringes across the field of view, the error signature will have four across the field of view. Figure 14 shows the difference between a calibrated and an uncalibrated phase shifter as well as the difference in error for two different phase measurement algorithms. Some algorithms are obviously more sensitive than others to this type of error.

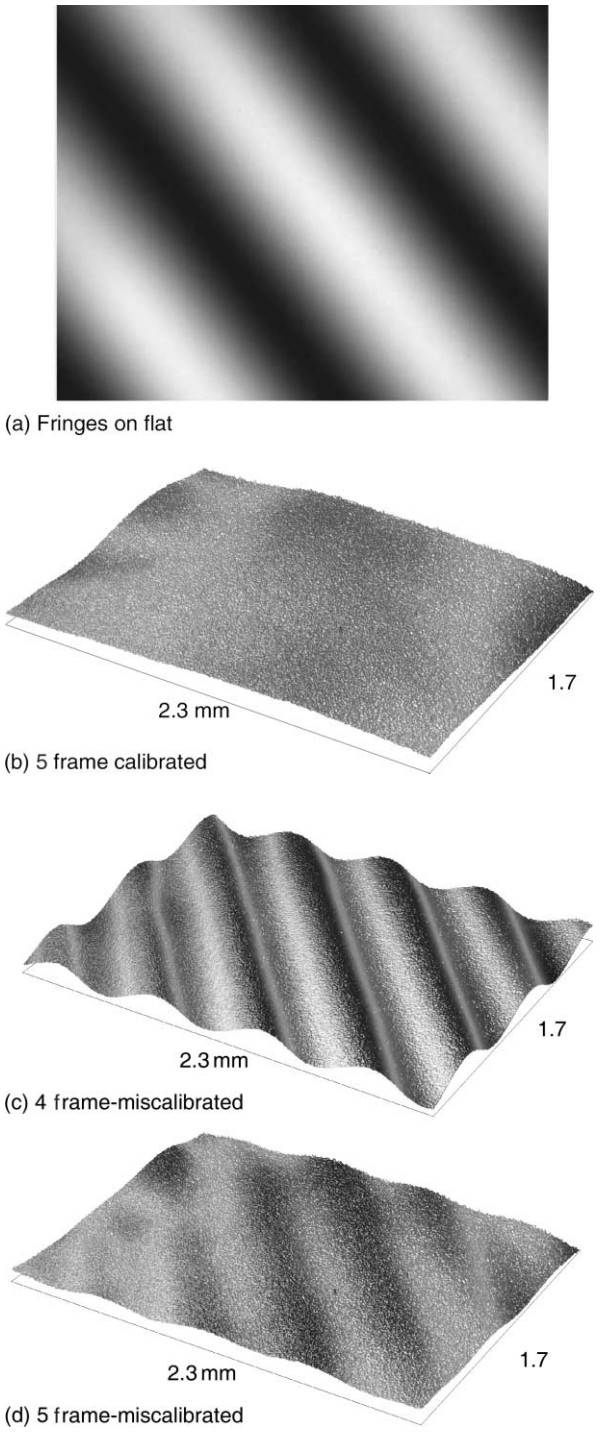


Figure 14 Comparison of phase maps for calibrated and uncalibrated phase shifts using two different phase measurement algorithms.

Other Types of Systematic Errors to Consider

Other types of errors to consider are detector nonlinearities, quantization errors due to analog-to-digital converters, and dissimilar materials (see Phase change upon reflection). For temporal phase measurement techniques errors due to vibration and

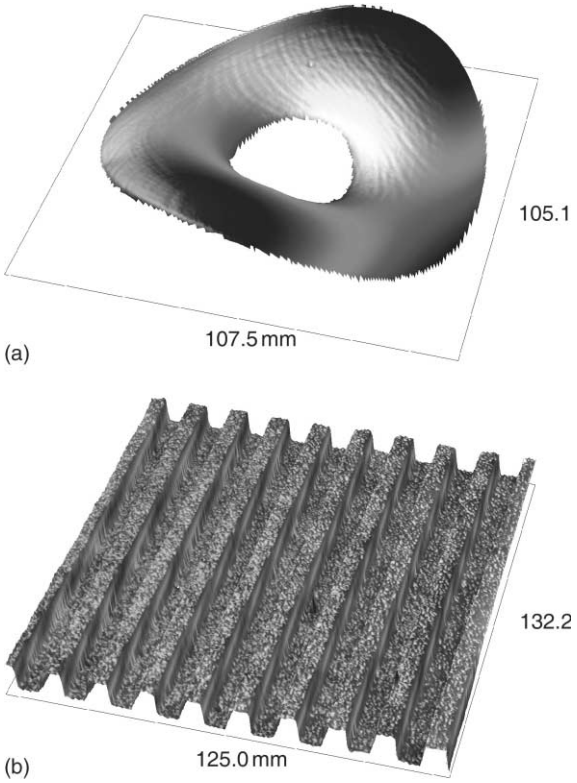


Figure 15 Measurements of (a) hard disc substrate with a vertical range is 2.2 micrometers and (b) binary grating roughness standard.

air turbulence need to be considered as well. Spatial phase measurement techniques are sensitive to miscalibrated tilt (wrong carrier frequency), unequally spaced fringes, and sampling and windowing in Fourier transform techniques.

Choosing an Algorithm or Technique

Each algorithm and type of measurement technique is sensitive to different types of systematic errors. Choosing a proper algorithm or technique for a particular type of measurement depends on the specific conditions of the test itself and reducing the systematic errors for a particular type of measurement. This is the reason that so many algorithms exist. The references in the bibliography will help the reader determine what type of algorithm will work best for a specific application.

Examples of Applications

Phase shifting interferometry can be used for measurements such as hard disk flatness, quality of optical elements, lens curvature, dimensions and quality of air-bearing surfaces of magnetic read/write heads, cantilevers, and semiconductor elements. Figure 15 shows results for measurements of a hard disk substrate and a roughness grating.

Conclusions

Phase measurement interferometry techniques have increased measurement range and precision enabling the production of more complex and more precise components. As work continues on development of interferometric techniques, phase measurement techniques will continue to become more robust and less sensitive to systematic errors. Anticipated advances will enable measurements of objects that were unimaginable 20 or 30 years ago.

See also

Interferometry: White Light Interferometry. **Microscopy:** Interference Microscopy.

Further Reading

Bruning JH (1978) Fringe scanning interferometers. In: Malacara D (ed.) *Optical Shop Testing*, pp. 409–438. New York: Wiley.

Creath K (1988) Phase-measurement interferometry techniques. In: Wolf E (ed.) *Progress in Optics*, vol. 26, pp. 349–393. Amsterdam: Elsevier.

Creath K (1993) Temporal phase measurement methods. In: Robinson DW and Reid GT (eds) *Interferogram Analysis*, pp. 94–140. Bristol: IOP Publishing.

Creath K and Schmit J (1996) N-point spatial phase-measurement techniques for non-destructive testing. *Optics and Lasers in Engineering* 24: 365–379.

Ghiglia DC and Pritt MD (1998) *Two-Dimensional Phase Unwrapping*. New York: Wiley.

Greivenkamp JE and Bruning JH (1992) Phase shifting interferometry. In: Malacara D (ed.) *Optical Shop Testing*, 2nd edn, pp. 501–598. New York: Wiley.

Hariharan P (1992) *Basics of Interferometry*. New York: Academic Press.

Kerr D (ed.) (2004) FASIG special issue on fringe analysis. *Optics and Lasers in Engineering* 41: 597–686.

Kujawinska M (1993) Spatial phase measurement methods. In: Robinson DW and Reid GT (eds). *Interferogram Analysis*, pp. 141–193. Bristol, UK: IOP Publishing.

Larkin KG (1996) Neat nonlinear algorithm for envelope detection in white light interferometry. *Journal of Optical Society of America A* 13: 832–842.

Malacara D (1992) *Optical Shop Testing*, 2nd edn. New York: Wiley.

Malacara D, Servin M and Malacara Z (1998) *Interferogram Analysis for Optical Testing*. New York: Marcel Dekker.

Robinson DW and Reid GT (1993) *Interferogram Analysis: Digital Processing Techniques for Fringe Pattern Measurement*. Bristol, UK: IOP Publishing.

Schmit J and Creath K (1996) Window function influence on phase error in phase-shifting algorithms. *Applied Optics* 35: 5642–5649.

Surrel Y (2000) Fringe analysis. *Photo-Mechanics: Topics in Applied Physics* 77: 55–102.

White Light Interferometry

J Schmit, Veeco Instruments, Inc., Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In 1960 the development of the laser injected new possibility into an old discipline, interferometry (*see* **Interferometry**: Overview). The excellent spatial and temporal coherence (ideally a single point and a single wavelength source) of the laser allowed for the formation of nonlocalized interference fringes. With lasers, researchers could easily generate good contrast fringes practically anywhere where the beams overlapped. These excellent-quality fringes from laser interferometers enabled high-resolution distance and displacement measurements on the order of meters and made possible noncontact surface probing with nanometer and subnanometer resolution. This kind of precision, which was supported by the advent of computers and the advancement of detectors, was previously unimaginable.

However, interference phenomena were used in metrology long before lasers and can be observed without any complicated device or source. On a rainy day we can observe the colorful interference patterns that have formed on a thin layer of gasoline in a puddle. The relationship between colors and the small distance between reflecting surfaces, observed as far back as 1665 by Hook, is still used to approximate layer thicknesses. The colors in interference patterns have been used to determine the spectral components of beams. By the end of the nineteenth century, Michelson and Benoit were using interference phenomena to determine distances, long before the invention of the computer and the laser.

While the application of lasers in interferometry has advanced the science, white light interferometry (WLI), which uses a spatially and temporally incoherent source and creates fringes localized over a few microns in space, has also benefited from technological advancements. The development of solid state detectors, fast computers, electronic signal processing, and precise scanning stages has allowed for incredibly fast analysis of white light interferograms. In fact, WLI is used in many applications from film thickness and surface measurement, through spectroscopy to astronomy.

This chapter focuses on the formation of white light fringes. We first examine the existence of white light fringes in the everyday world. A description of different interferometric setups and their uses follows. The bulk of this article details the formation of white light fringes and examines their different applications with an emphasis on the analysis of object structure. Suggestions for further reading, both within this encyclopedia and from outside sources, can be found at the end of this article. A basic theory of interferometry and coherence is to be found in corresponding articles of this encyclopedia.

White Light Interference in Everyday Life

The effects of interference display spectacular colors when observed in white light, such as the sun. We often see these effects in oil or gasoline spills, in soap bubbles, or between two smooth pieces of glass in contact (separated by a very thin layer of air film) – which are the simplest everyday white light interferometers. The colors we see in these displays are interference colors, and their origin is interference rather than dispersion as in the colors of a rainbow or light reflected from a CD. The picture in Figure 1 is of a gasoline spill under cloudy sky illumination; we see beautiful interference colors that are hard to find in a rainbow: iron-gray, magenta, grayish blue, whitish green, and brownish yellow. Interference colors will differ not only with the thickness of the layer but also with the layer's absorption and dispersion, the relative indices of refraction of the film, the surrounding media, and the illumination. Interference colors can be observed for layers from a fraction of a micron to a few microns thick.

Interferometers for White Light Observation

In a white light interferometer, either the colors or the intensity distribution of the fringes is typically analyzed to retrieve the necessary encoded information, such as the film thickness, birefringence, index of refraction, dispersion, spectral properties, and surface structure. White light interference can be observed only in interferometric designs where the optical paths in both arms of the interferometer are (nearly) equal and the system is (nearly) compensated for dispersion.

Interference in Thin Film

Beams of light of different wavelengths incident on a transparent thin film (such as the ones in the puddle) are partially reflected from the top air/film interface

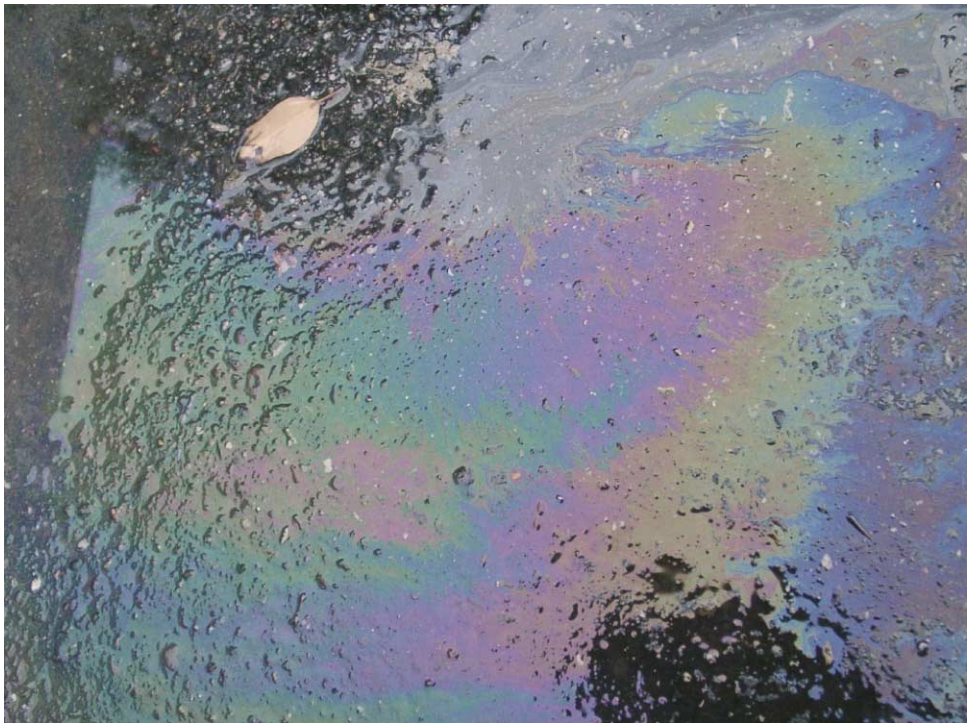


Figure 1 Interference colors in a gasoline spill in a puddle. This picture would closely resemble a picture from an interference microscope, if not for the leaf in the top left corner.

and partially from the bottom film/substrate interface and then beams of the same wavelength interfere with each other. The optical path difference (OPD) traveled by the interfering beams is related to the film thickness, the index of refraction for the given wavelength and the angle of incidence.

Fringes of equal thickness will be observed if the film thickness (distance between interfering wavefronts) varies. These fringes represent points of equal distance between the two surfaces of the film and are formed close to the film (see Figure 2). This situation is typical in any interferometer since rarely are wavefronts perfectly plane and parallel.

For any individual wavelength, λ_i, where the OPD is equal to $m\lambda_i$, and where m is an integer, a bright fringe of the color of the wavelength, λ_i, will be observed due to constructive interference. For white light illumination, the color at a given point will be dominated by the color of the wavelength for which the interference will be constructive. This color can be used to estimate the optical thickness of the film.

An additional factor needs to be taken into account when color is used for estimating optical thickness, namely the phase change on reflection. Different colors will be observed if the ratio of indices of refraction for both film interfaces is <1 (or >1) as opposed to when the ratio of indices is <1 for one interface and >1 for the other.

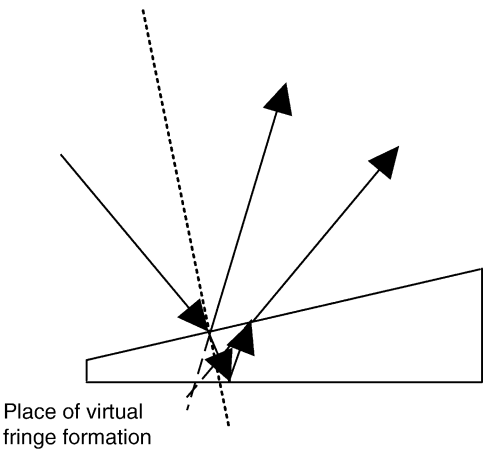


Figure 2 Formation of fringes of equal thickness for film with wedge.

Polarization Microscope

Interference colors that correspond to equal thickness fringes can represent the birefringence of an object; these colors are often observed using a polarization microscope (*see* **Microscopy:** Interference Microscopy). The polarization microscope is a conventional microscope with a set of polarizers with crossed polarization axes, one placed above and the other placed below the tested birefringent object.

Interference colors have been tabulated for many years and for many different purposes. Newton devised his color scale to describe interference fringes for two pieces of glass in contact with each other. Michel-Lévy and Lacroix in 1889, created a color scale to help recognize different rock forming minerals. For more information about colors in white light interference see sections on *White Light Interference* and *Spectral Interference* below.

Michelson Interferometer

A Michelson interferometric setup, shown in Figure 3, is often used to analyze white light interference, but the intensity of the fringes is observed as one mirror is scanned rather than the color of fringes that is usually observed.

Two plane mirrors (equivalent to the top and bottom surfaces of a thin film) return the beams to the beamsplitter, which recombines parts of the returned beams and directs them towards the detector where the interference is observed. Beam S1 travels through the parallel plate three times while beam S2 passes through the plate only once causing the system to be not well compensated (balanced) for dispersion. Thus, for observation of white light interference, an identical plate is placed in the path of beam S2. As one of the plane mirrors moves along the optical axis to change the OPD, the detector collects the irradiance, the intensity. Fringes will be observed if the system is well-compensated for dispersion and for optical path lengths. Any spectral changes or changes in optical path lengths in the interferometer affect the shape or position of the fringes, and the interferometer measures these changes.

Common-path interferometers, like the Mach–Zender or Jamin interferometers, are naturally compensated interferometers and can be used for measurement of dispersion in gases. Other interferometers, such as the Twyman–Green or Fizeau interferometers, make use of their unequal interferometer arms and their nonlocalized fringes from laser sources for testing different optical elements and systems in reflection and transition (*see* **Interferometry:** Phase Measurement Interferometry).

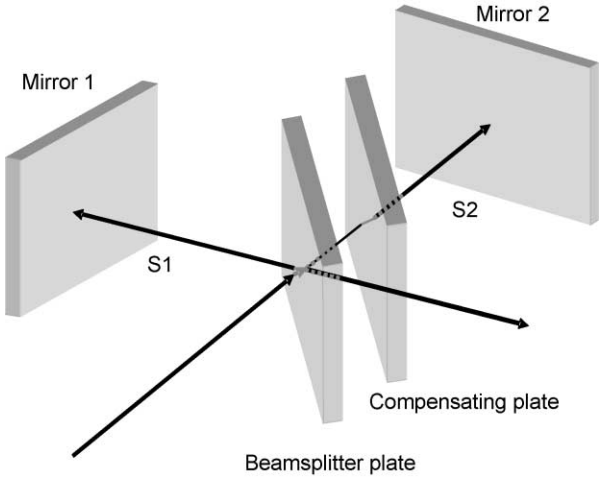


Figure 3 Michelson interferometer.

White Light Interference

A white light source consists of a wide spectrum of wavelengths in visible spectrum, from about 380 up to 750 (violet to red) nanometers. However, the principles of WLI described in this article basically apply when any low coherence source is used. Low coherence refers to a source of not only low temporal but also low spatial coherence; WLI can also be referred to as low coherence interferometry. We will be concerned mainly with temporal effect of the source, the source spectrum.

Since different wavelengths from the source spectrum are mutually incoherent, we will first look at interference between two waves of a selected monochromatic illumination with wave number $k = 2\pi/\lambda$, where λ is the wavelength. The intensity of the interference fringes at point x, y (these coordinates are omitted in all equations), as one of the mirrors is scanned along the optical axis, z can be described as

$$I(k,z) = I_1 + I_2 + 2\sqrt{I_1 I_2}|\gamma(z)|\cos(kz) \qquad [1]$$

or can be written in the form:

$$I(k,z) = I_0\left[1 + \frac{2\sqrt{I_1 I_2}}{I_1 + I_2}|\gamma(z)|\cos(kz)\right] \qquad [2]$$

where I_1 and I_2 are intensities of each of the beams and $I_0 = I_1 + I_2$, $|\gamma(z)|$ is the modulus of the complex mutual coherence function (*see* **Coherence:** Overview) assumed here to equal 1 (perfect coherence for each wavelength). The optical path difference z equals $z_1 - z_2$ where z_1 and z_2 are the optical path lengths that the interfering waves have traveled. The difference in the traveled paths, z, corresponds to the position of the scanning mirror.

White light interference (polychromatic interference) is the overlaying of all the monochromatic fringes created for each wavelength of the source spectrum (Figure 4a). A detector observes the sum of all the fringe intensities.

In mathematical form, this interference can be described as the integral of all the fringes $I(k,z)$ for all

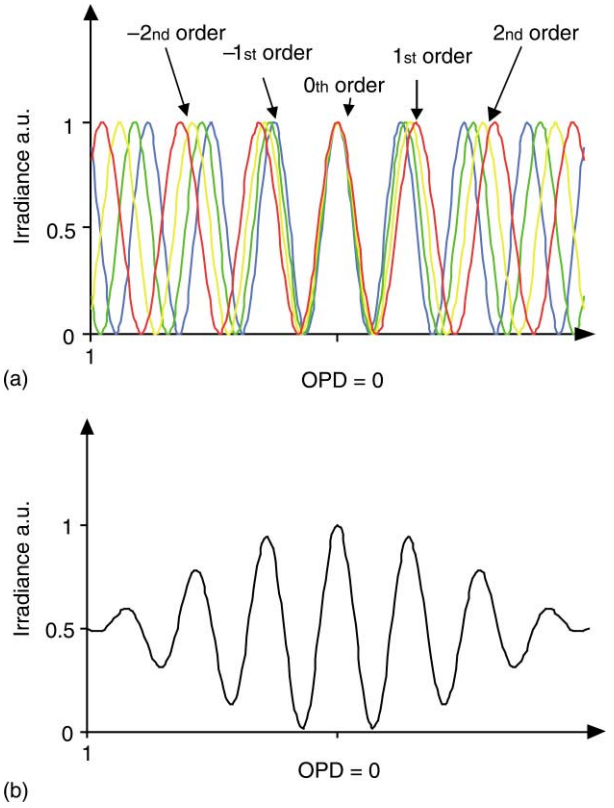


Figure 4 Formation of white light fringes: (a) fringes for individual wavelengths; (b) fringes for white light.

wavenumbers k:

$$I(z) = \int_{k_1}^{k_2} S(k)D(k)I(k,z)\mathrm{d}k \qquad [3]$$

where $S(k)$ is the spectral distribution of the light source, with $S(k) = 0$ outside of k_1 through k_2 range of wave numbers, and $D(k)$ is a spectral response of the detector. We assume that the detector's response $D(k)$ equals 1 over the whole spectrum. Because the spacing of the fringes for each wavelength emitted by the white light source is different, the maxima of fringes will align around only one point for a well-compensated interferometer, as shown in Figure 4a. This alignment around a single point, occurs because in a well-compensated interferometer there is one point for which the OPD is zero for all wavelengths. Away from this point, the observed sum of the intensities quickly falls off, as shown in Figure 4b. The maximum fringe, the fringe that marks the zero OPD, is called the zero-order fringe, and each next fringe of smaller amplitude on either side is called $+1$ and -1, $+2$ and -2 order fringe, and so on.

Because white light fringes are localized and can only be found within microns or tens of microns of the zero OPD, white light interferometers are excellent distance and 3D position sensors. However, this same characteristic makes them more difficult to align than interferometers with nonlocalized fringes.

Envelope of Fringes Due to Source Spectrum

The resultant intensity of interference from a broad spectrum source (see eqn [3]) can be described in general form as

$$I(z) = I_0[1 + V(z)\cos(k_0 z)] \qquad [4]$$

where $I_0 = I_1 + I_2$ is the background intensity, $V(z)$ is the fringe visibility function or coherence envelope, and $k_0 = 2\pi/\lambda_0$ is the central wave number for fringes under the envelope. The $V(z)$ is proportional to the modulus of the Fourier transform of the source spectrum $S(k)$. Generally, if the light source has a Gaussian spectrum $S(k)$, then the envelope of the fringes can be described also as a Gaussian function $V(z)$. If the spectrum of the source is rectangular, then the envelope of the fringes will be a sinc function. The wider that the spectrum of the source is the narrower the width of the envelope will be, as shown in Figure 5. The width of the fringe envelope determines the coherence length of the source (*see* **Coherence:** Overview); for a white light source this width is in the order of 1–2 microns.

Different white light sources, such as tungsten-halogen, incandescent, or arc lamps, have different spectra and thus create different coherence envelopes, as shown in Figure 5. The spectra of semiconductor light sources, such as LEDs and SLDs, are similar in shape to a Gaussian function.

The fact that the coherence envelope is related to the spectrum of the source by its Fourier transform is commonly used in Fourier transform spectroscopy, where the Fourier transform of the detected fringes is calculated to find the spectral components of the beams.

Position of Fringes Under Envelope Due to Reflection of Dielectric

Thus far we have assumed that in WLI the interferometer is compensated for all wavelengths, and thus, for this position the maximum of the fringes aligns with the maximum of the envelope, namely where there is a zero phase shift $\varphi_0 = 0$ between fringe and envelope maxima.

If there is an odd number of reflections from dielectric surfaces in one arm of the interferometer and an even number in the other, the fringes will be shifted by $\varphi_0 = 180°$ under the coherence envelope and the minimum of the fringe will align with the maximum of the coherence envelope. Thus eqn [4]

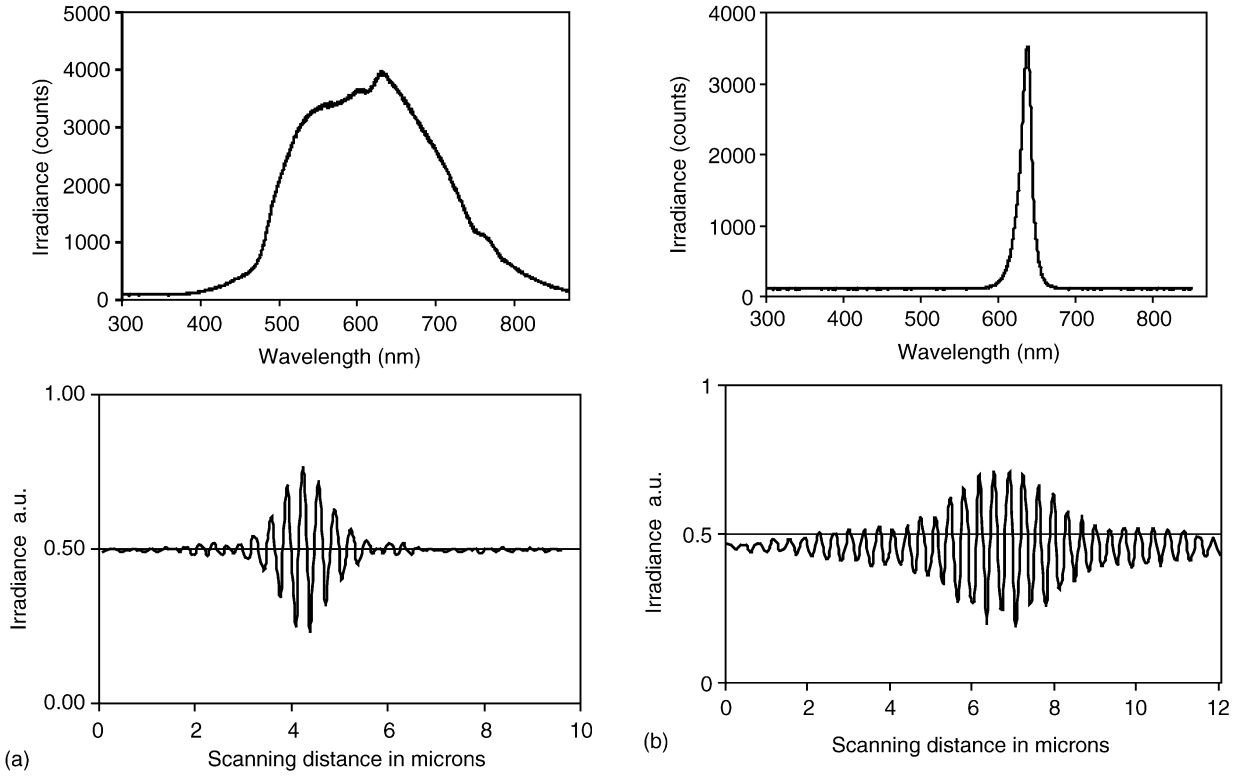


Figure 5 Spectrum and interferogram for (a) tungsten-halogen lamp and for (b) red LED sources.

can be expressed in a more general form:

$$I(z) = I_0[1 + V(z)\cos(k_0 z + \varphi_0)] \qquad [5]$$

Figure 6 shows color fringes for such a case; the dark fringe marks the zero OPD, and this fringe is surrounded by the greenish blue colors of shorter wavelengths. In contrast, the bright fringe at the maximum envelope position or marking the zero OPD would be surrounded by the reddish colors of longer wavelengths.

In a real system the fringes may be shifted with respect to the envelope by any amount of φ_0, and this shift may be due to any number of factors, such as the phase shift on reflection from nondielectric surfaces and dispersion, which we consider next.

Changes in Envelope and Fringes Due to Reflection of Nondielectric Materials

The relative position of the envelope maximum and fringe position will be different if the beam is reflected from different nondielectric materials. This difference exists because the phase change on reflection from a test surface, like metals or heavily doped semiconductors, varies with each wavelength. This variance in fringe and peak coherence position may be predicted and corrected for in surface height calculations. The linear dependence of the phase change on reflection versus wave number shifts the location of the coherence envelope peak position of fringes by different amounts. The constant phase change on reflection and higher-order terms only shift the fringes underneath the coherence envelope. As long as the change on reflection has a small dependence on the second- and higher-order of the wave number, the shape of the coherence envelope is preserved.

Changes in Envelope and Fringes Due to Dispersion

Dispersion in an interferometer that is not balanced, perhaps because a dispersive element was placed in one arm or the compensating plate has an incorrect thickness, will influence fringe formation. The phase delay between interference patterns for individual wavelengths is proportional to the product of the geometrical path and the index of refraction equal to $d \times n(k)$.

The intensity may be described as:

$$I(z) = \int_{k_1}^{k_2} \langle 1 + V(z)\cos\{kz - kd[n(k)]\}\rangle \, \mathrm{d}k \qquad [6]$$

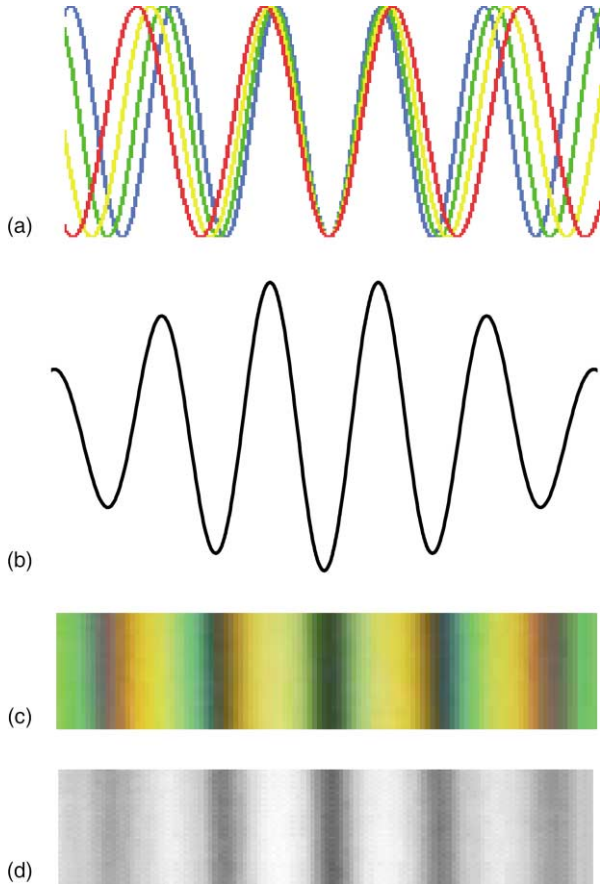


Figure 6 Formation of white light fringes with destructive interference for OPD = 0.

The dependence of the refractive index on the wave number k can be described as a linear expansion:

$$n(k) = n(k_0) + \frac{\mathrm{d}n}{\mathrm{d}k}(k - k_0) \qquad [7]$$

The linear dispersion shifts the envelope by the group index of refraction times the thickness of the dispersive element; this dispersion also shifts the fringes under the envelope slightly. In other words, the linear phase delay for a range of wavelengths causes a group delay of the whole envelope (wave packet). Higher-order dispersion, absorption of the elements, or effects due to thin films can cause the envelope to widen or even become asymmetrical, the position of the fringes under the envelope to shift, the fringes to lose contrast, and the period of the fringes to change or vary with the z position. Dispersion effects will be stronger for sources that have a wider spectrum; however, the observed changes will be different for different shapes of spectra.

The phase of the fringes under the envelope and the position of the envelope itself are parameters often used in astronomy. Highly accurate white light fringe estimation, using the optical path-length delay between the two arms of the interferometer, is a cornerstone of stellar interferometry.

Fourier Analysis of White Light Interferograms

Wavelength-dependent changes in a white light interferogram can be more easily analyzed in the spectrum rather than the spatial domain. The Fourier transform of a white light interferogram yields two symmetrical side lobes at the mean wavelength of the interferogram; analysis of just one of these side lobes is sufficient. The spectral amplitude of the side lobe contains information about the spectral components of the interfering beams, while the spectral phase in regions with appreciable spectral amplitude supplies information about any unbalanced dispersion in the interferometer, as shown in Figure 7. Fourier transform analysis is extensively used in Fourier transform spectrometry (*see* **Spectroscopy:** Fourier Transform Spectroscopy).

For a dispersion-balanced interferometer, the interferogram is symmetrical around the zero OPD position. For a symmetrical interferogram, the spectral phase will be zero if the zero OPD position of the sampled interferogram is in the middle of sampling range; otherwise a linear factor appears in the spectral phase. This linear term, while useful to surface profiling, because it determines the position of the coherence signal with respect to scanner sampling, is unwanted in Fourier spectroscopy and needs to be corrected for. Dispersion and possible thin-film effects will commonly introduce a nonlinearity in the spectral phase.

Controlled Phase Shift of Fringes Under the Envelope – Geometric Phase Shift

Many techniques in interferometry depend on shifting the phase of the interfering wavefronts. Mechanical shifters used in white light interferometry will introduce the same shift measured in nanometers for all wavelengths; however, when measured in degrees or radians, the shift will vary for different wavelengths. Geometric phase shifters (achromatic phase shifters operating on the principle of geometric phase) introduce for all wavelengths the same shift when measured in degrees and are based on polarization elements like a rotating wave plate in circularly polarized light or a rotating polarizer in a circularly polarized beam. Fringes underneath the coherence envelope shift, as shown in Figure 8, while the coherence envelope stays in place.

The advantage to these techniques is that only the phase of fringes changes, not the fringe contrast

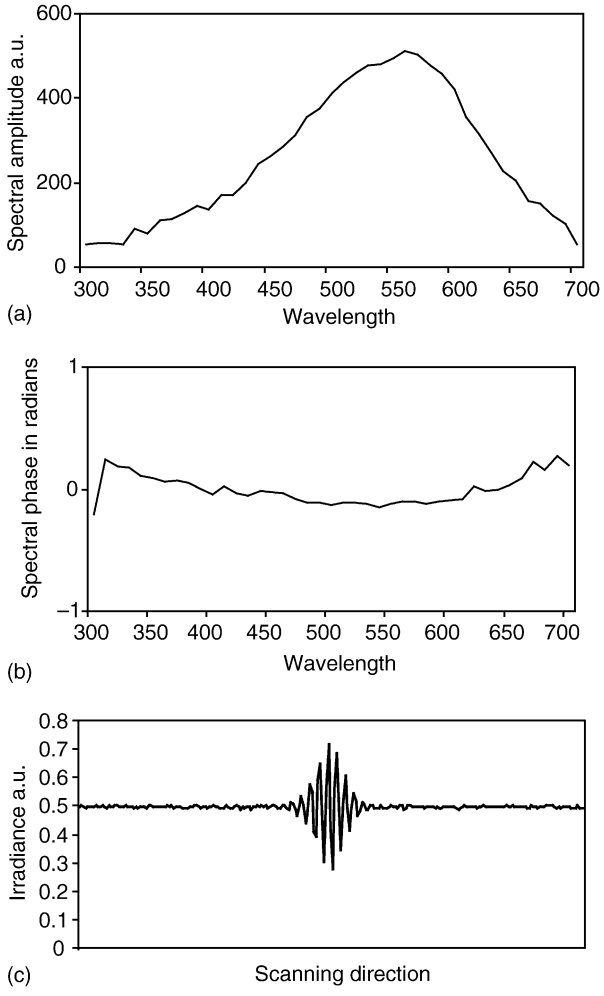


Figure 7 Spectral amplitude (a) and spectral phase (b) for input from unbalanced interferometer (c).

(*see* **Interferometry:** Phase Measurement Interferometry). A geometrical phase shifter can be very useful in polarization microscopes (Figure 9), white light shearing interferometers, or any system where the phase of the white light fringes needs to be measured.

Spectral Interference

White light fringes, because they are made up of fringes of many wavelengths, are observed only over very small path differences. If we filter only a narrow band from the white light spectrum, fringes would be visible over a much larger scan. For different wavelengths we would observe different color and frequency fringes; this is simply the reverse process of the description in the white light interference section.

If we place a spectrometer in the observation plane of the white light interferometer (Figure 10), we will observe fringes with continuously changing wavelengths in dispersed light. These fringes are called

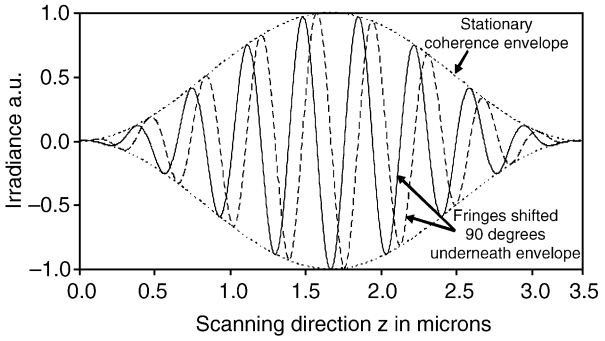


Figure 8 Fringes for two different geometric phase shifts.

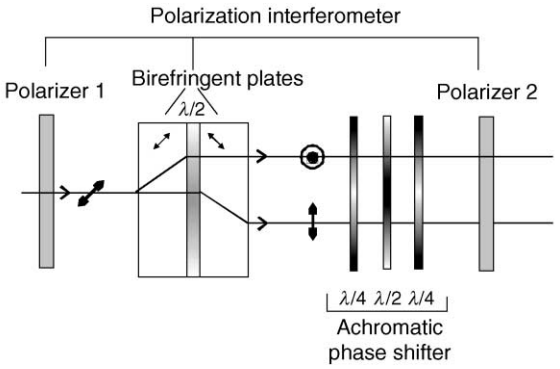


Figure 9 Polarization interferometer with geometric phase shift.

fringes of chromatic order or channeled spectrum fringes and find their application in film thickness measurement or absolute distance measurement in the range up to 1 mm. Channeled spectra were used for analysis of dispersion, thin film, and spectroscopic measurements. The number of observed fringes for a given wavelength is directly proportional to the measured optical path difference. The optical path difference can be determined if the difference in fringe numbers is determined for two well-known wavelengths (this is equivalent to two-wavelength interferometry). The optical path difference can be also quickly estimated from the frequency of fringes for a given wavelength; the larger the optical path difference, the denser the fringes.

Surface Topography and Object Structure Measurement

Although WLI has many applications, this section focuses on white light interference as applied to surface topography measurement. Interference microscopes that use white light illumination are often based on the Michelson setup shown in Figure 11.

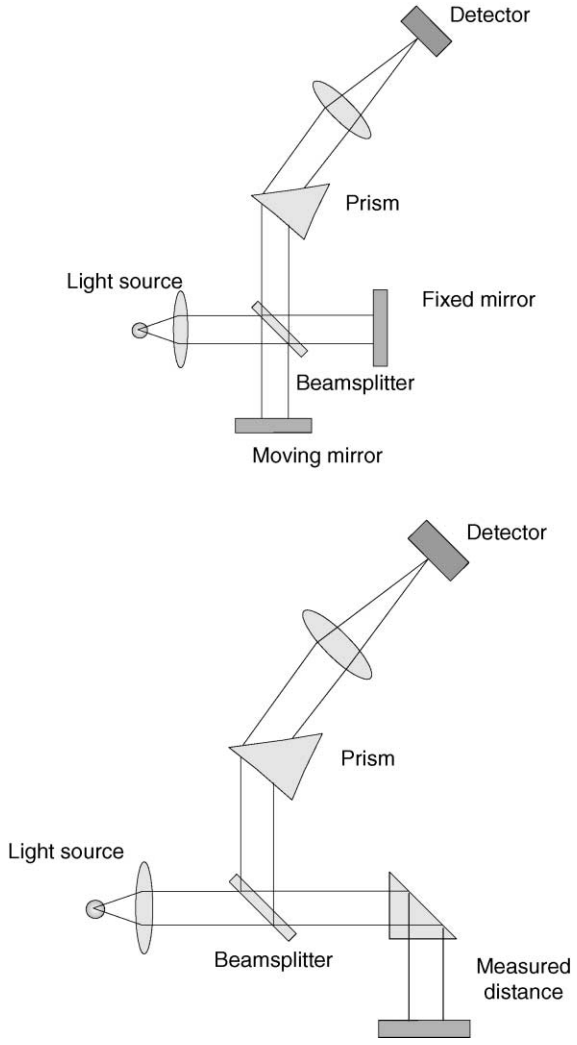


Figure 10 Interferometers with prism spectrometer to observe fringes of equal chromatic order for OPD measurement.

A beamsplitter cube is placed underneath the bright field objective, and one mirror is placed to the side at the focus plane of the objective while the other mirror is replaced by the measured surface with respect to which the whole interference objective is scanned. The narrowly localized fringes at the best focus position for each point of the surface corresponding to a pixel on the CCD camera is observed during the scan. Because of the design of the Michelson objective, in order to accommodate higher magnification objectives constraints, other interference setups like Mirau and Linnik were developed (*see* **Microscopy:** Interference Microscopy). However, these designs are still based on the Michelson interferometer, being also compensated interferometers with equal lengths and amounts of glass in each arm.

These interference microscopes typically have two complementary modes of operation; one mode uses

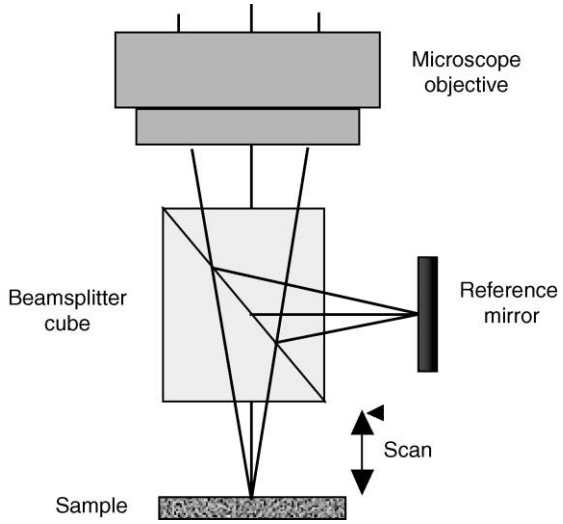


Figure 11 Michelson interferometric objective.

monochromatic illumination and the other employs a white light source. Monochromatic illumination provides excellent vertical resolution but is limited in its range. It cannot correctly assign fringe order to discontinuous surfaces larger than 160 nanometers. Figure 12a shows monochromatic fringes for the surface of the profile shown in Figure 13. We see that with the monochromatic illumination the height of the grooves remains unknown because it is impossible to assign order numbers to these fringes.

To resolve this ambiguity, white light illumination is employed because it allows for easy identification of the zero order fringe. Figure 12b shows white light fringes created for the same surface. We see the zero-order fringe for the zero OPD as well as the next orders due to the decreasing contrast of the fringes. The position of the zero-order fringe can be followed and approximate heights can be determined visually. Thus, white light illumination permits the measurement of a broader range of surfaces that are rough or have large height variations up to a few millimeters.

The white light interferometer is a great focus detector (Figure 14) at each point of the field of view. The principle behind WLI is finding these individual focus positions using localized fringes observed during the surface scan. WLI interferometers provide high lateral resolution and large vertical range in the order of hundreds of nanometers or micrometers (Figure 15). White light interferometers are commonly used to measure magnetic heads, MEMS devices, binary optic, and machined surfaces.

Signal Processing of White Light Interferograms

For WLI topographic measurements, each pixel registers an interferogram whose position varies

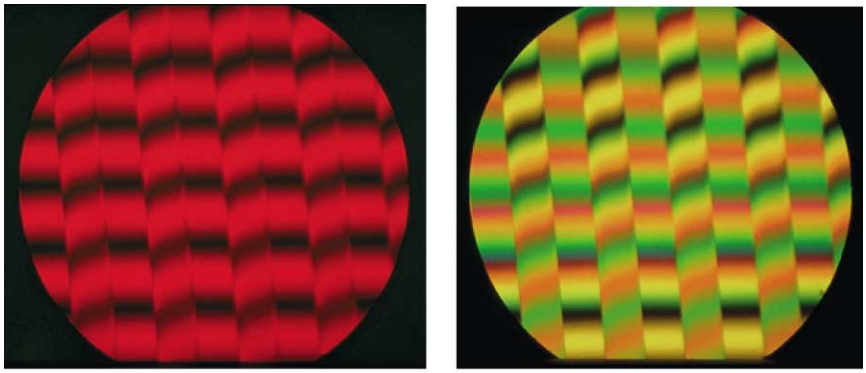


Figure 12 Fringes in quasimonochromatic and white light for object similar to one as presented in Figure 13.

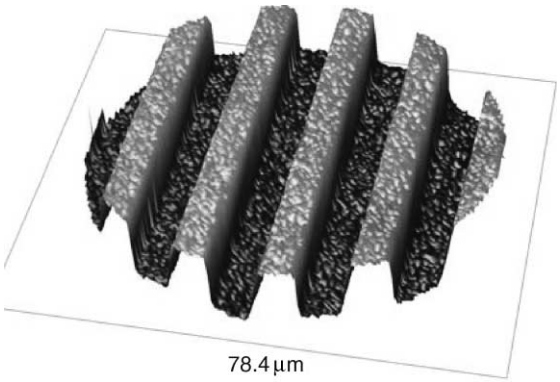


Figure 13 3D object profile – binary grating.

with the surface height. Figure 16 shows two interferograms for two positions in the field of view of a measured step height surface; the first shows the top surface of the step height and the second shows the bottom surface of the step height.

Algorithms have been developed to analyze the signal in different ways. Some focus on finding the fringe envelope position and others examine the position of the fringe underneath the envelope. These algorithms work under the assumption that white light fringes can be described by the same function over the whole field of view of the interferometer. Algorithms are applied to a signal so as to find the position of the coherence envelope for each pixel. This process is called coherence gating, and all algorithms perform coherence gating.

Some algorithms look for the position of the envelope's maximum in a process whereby first the coherence envelope of the fringes is found using signal filtering. Then, the curve is fit to a few points around the envelope's maximum, and finally the position of the envelope's maximum is found. Other algorithms calculate the center of mass of the coherence signal, as described in eqn [8]:

$$H(x,y) = \frac{\sum_{n=1}^{N-1} \left(I_i(x,y) - I_{i+1}(x,y)\right)^2 z_i}{\sum_{n=1}^{N-1} \left(I_i(x,y) - I_{i+1}(x,y)\right)^2} \qquad [8]$$

This method is very fast and computationally efficient. Center of mass calculations are equivalent to calculations of the maximum of the envelope position, but only for a symmetrical signal. For an asymmetrical signal a piston is introduced for each point; however, this piston does not affect the whole measurement. Still another method calculates the spectral phase slope in the Fourier domain to determine the position of the coherence envelope in a process that is equivalent to finding the center of mass. Finally, a different group of algorithms tracks the position of the bright or dark fringe close to the envelope's maximum.

Scanner Speed – Sampling Rate

A practical sampling rate of the white light interference signal is around four pixels per fringe. Because in most algorithms it is not the position of the fringes that we want to determine but rather the position of the envelope of the fringes, sampling the envelope of the fringes with four samples is usually sufficient. However, this sufficiency comes at the expense of lower repeatability (higher noise). The advantage to this sampling rate is short measurement time; higher sampling rates increase measurement time. Measurement speed can be increased by up to 25 times when the envelope of the fringes is widened by filtering light of a bandwidth about 20–40 nm from the white light source.

Scanner Nonlinearity

Scanner motion is assumed to be linear, but nonlinearity in this motion impairs measurement

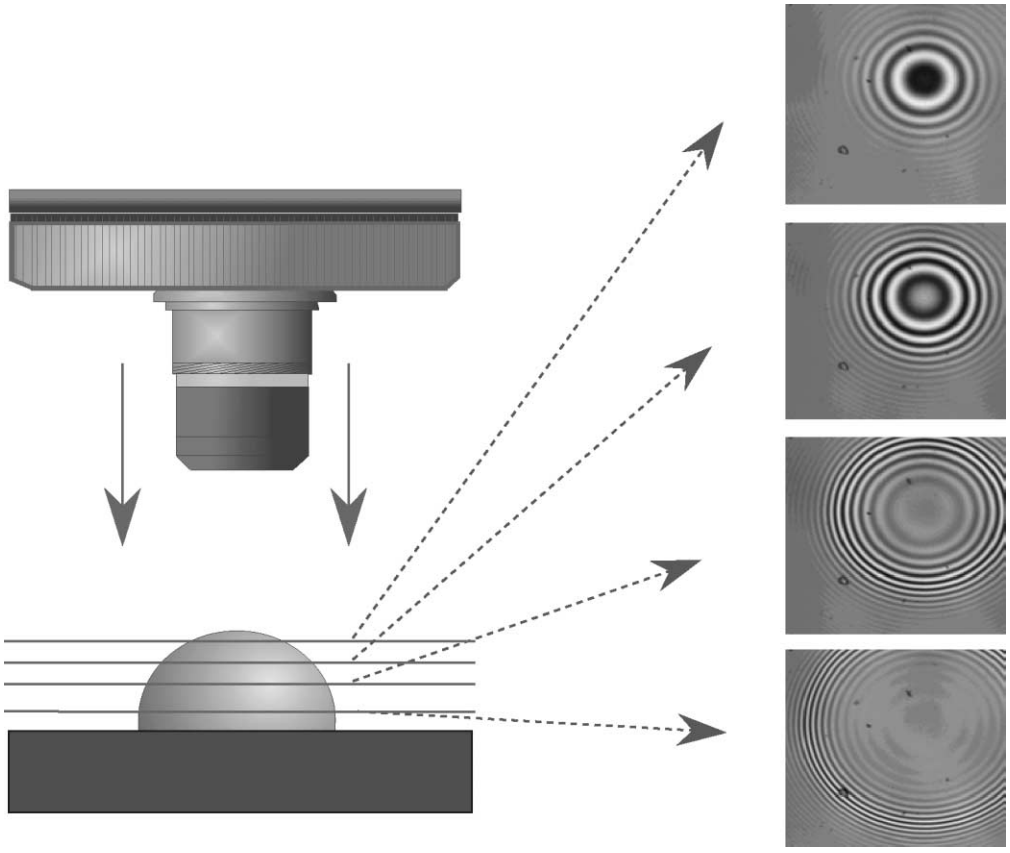


Figure 14 Operation of white light interferometer for surface topography measurement.

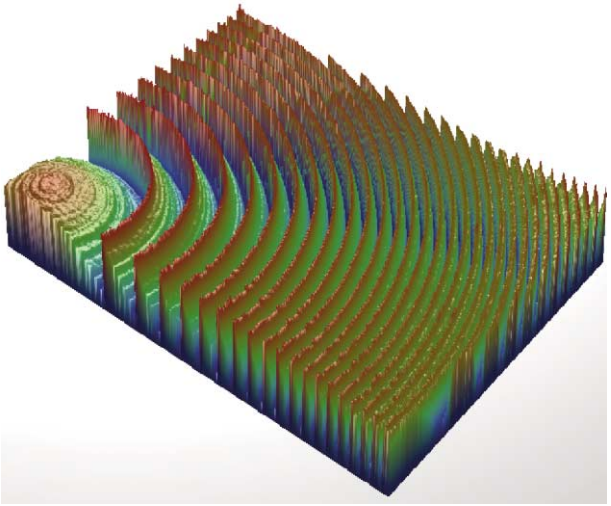


Figure 15 WLI measurement of Fresnel lens in a high-resolution mode. Courtesy of Glimmerglass.

accuracy. To account for this, simple algorithms can calculate the average steps along each interferogram in scanning direction (Figure 17) and measured steps can be used in coherence detection algorithms.

In order to have continuous information about the scan steps, the fringes should be visible somewhere in the field of view at each moment of the scan. This task may require introducing a large tilt for samples with large discontinuities such as stepped surfaces. For large discontinuities where introducing sample tilt may not be sufficient, the envelope of the fringes can be extended by increasing the spatial or temporal coherence of the light source. An alternate solution involves measuring the scanner motion with a distance-measuring interferometer or triggering the camera to collect intensity data at every equal scan step, which can be determined by using the zero-crossing technique, as is commonly done in Fourier transform spectroscopy. Equal scan steps are more suitable for techniques based on transform techniques that assume equal sampling. Other peak detection algorithms, such as center of mass calculations, make use of measured but not necessarily equal steps directly in the algorithm.

Observing the phase along the interferogram can also provide information that is important in optical fiber sensing. These observed changes, which assume that sampling rates of the interferogram are known,

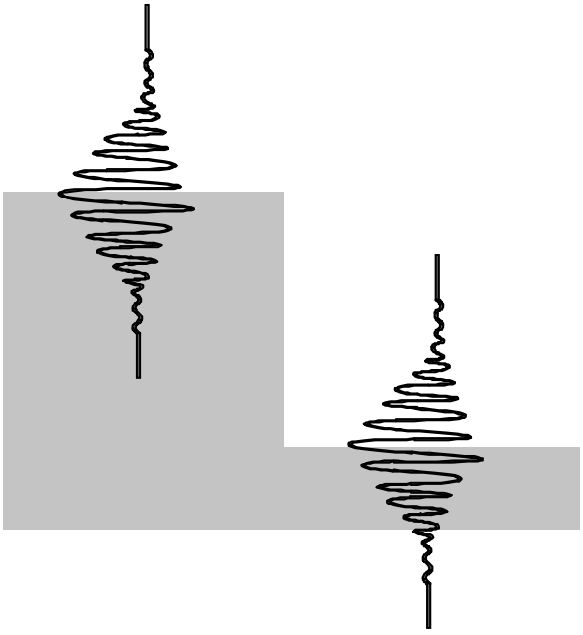


Figure 16 White light fringes as seen by two pixels during scan through focus for object in form of step.

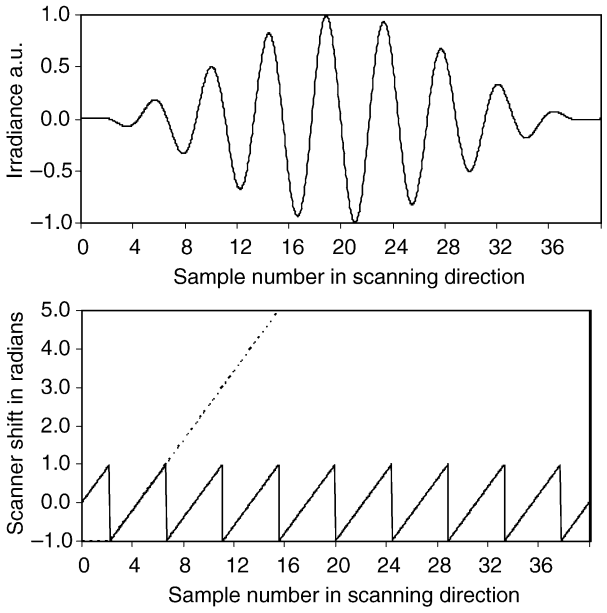


Figure 17 Phase calculated along the white light interferogram. This calculated phase can be used to determine scanner motion.

enable correction for changes in wave number value k_0 which can be due to such things as changes in the working voltage of bulb, an introduced higher-order dispersion, or a large tilt of the sample.

Increased Resolution White Light Interferometry

Interferometric methods that employ a monochromatic light source to detect the phase of the fringes

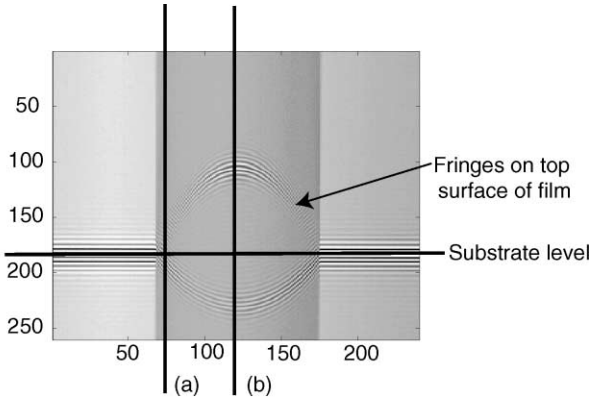


Figure 18 Fringes for (a) thin and (b) thick film.

(*see* **Interferometry:** Phase Measurement Interferometry) can achieve about 10 times better (0.3 nanometers versus 3 nanometers) vertical resolution than the WLI methods described so far. Combining coherence position detection from WLI to determine fringe order with the phase detection of phase shifting techniques allows for the measurement of samples with height discontinuities larger than 160 nanometers with the resolution and accuracy of phase shifting interferometry (PSI). This combination is particularly well-suited for determining the shape of smooth surfaces with large height differences such as binary diffractive optics or micro-electromechanical systems (MEMS) (see Figure 15).

Using this combined method we obtain both a lower-resolution map of the envelope position and a higher-resolution map of the phase (position) of the zero order fringe. These maps may differ slightly due to effects similar to those discussed in sections *Changes in Envelope and Fringes Due to Reflection of Nondielectric* and *Changes in Envelope and Fringes Due to Dispersion* above. In interference microscopes, shifts in the envelope and fringe position may be introduced by field-dependent and chromatic aberrations of the system and the system's effective numerical aperture. These shifts can vary for different points on the tested surface, but for simplicity are assumed to be constant over the field. Some correction for these effects can be applied.

Film Thickness Measurement

White light interferograms that are obtained in an optical profiler can be used to measure transparent film thicknesses because the character of the interferogram changes with the thickness of the film (Figure 18). Two different techniques, a thin or thick film technique, are used depending of the range of the film thickness. A thick film technique is used if two

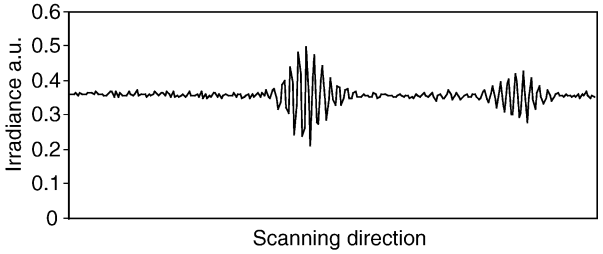


Figure 19 Thick film (see Figure 18b) irradiance.

sets of best contrast white light fringes from each interface are clearly separated, meaning that no interference occurs between the wavefronts reflected from the top and bottom surfaces of the film. A thin film technique is employed when the two sets of fringes overlap.

Thick film measurement

A simple technique for finding the relative position of the peaks of the fringe envelopes can be implemented to find the thickness of a film. Figure 19 shows two clearly separated sets of fringes formed for the air/film and film/substrate interfaces. Whether these two sets of fringes are separated or not depends mainly on the geometrical thickness of the film and its group index of refraction, which is determined by the dispersion of the film. The typical range of measurable film thicknesses runs from 3 to 150 microns, depending on the dispersion of the film. This measurement allows for the detection of flows on the surface and interface of the film. Similar principles based on finding the position of the coherence envelopes are used for distance sensing, thickness measurement of plates, and the cornea of the eye, in low coherence reflectometry and structure measurements in biological samples in optical coherence tomography.

Thin film measurement

For thicknesses of five microns down to tens of nanometers, the white light interferogram is created from the interference between the beams reflected from the reference mirror and the two beams reflected from the thin film layer. Once the interferogram is registered while the objective is scanned vertically with a constant velocity, the spectral phase is calculated by applying a Fourier transform to the measured signal at each pixel as described in the section *Fourier Analysis of White Light Interferograms* above. The phase slope is subtracted and the dispersion of the system needs to be known. The spectral phase for the thin film interference has the form of a polynomial; thus, the polynomial for the chosen film model (n and k) is fitted, and regression

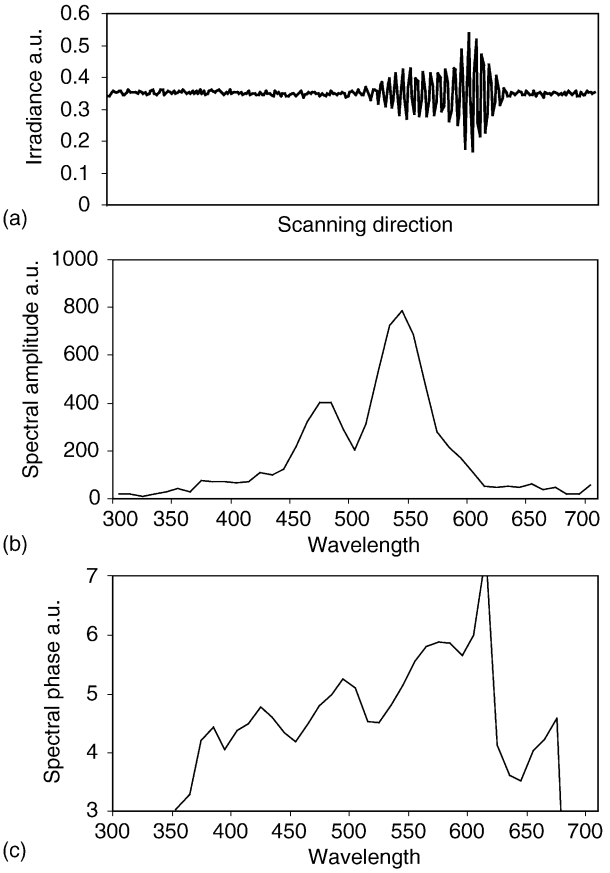


Figure 20 Thin film (see Figure 18a): (a) irradiance, (b) spectral amplitude and (c) spectral phase.

analysis is used to find the best fit and, therefore, the film thickness (Figure 20).

Spatial Coherence Effects in the Interference Microscope

So far we have been discussing temporal coherence effects, but everywhere that the source has a wavelength bandwidth, spatial coherence may play an important role. In an interference microscope for surface topography measurement, the size of the aperture of the condenser is on the order of the aperture of the objective that illuminates the surface with a wide range of angles. For large angles, which are determined by the numerical aperture (NA) of the objective, where NA = 0.5–0.95, the combination of the influence of the spatial and temporal coherence is clearly visible. The additional spatial effects include a reduction of the fringe envelope width and a decrease in fringe spacing corresponding to k_0. Calibration of fringe spacing is typically done on the system and accounts for spatial effects as well as for any uncertainty of k_0 due to the source spectra, the working temperature of the source, the spectral

response of the detector and the influence of other factors like the intensity distribution in the illuminating aperture. The spatial influences can be reduced by stopping down the condenser, which causes an increase in the contrast of the fringes.

Rough Surfaces

Rough surfaces are difficult to measure using interferometric techniques, but under certain coherence conditions white light interference can do the job. For rough surfaces, if the microstructure of the object is not resolved by the imaging system, speckles, rather than fringes, are observed. Each speckle has a random phase which is approximately constant in the whole speckle area. If the rough surface is scanned through focus, each individual speckle exhibits the intensity modulation that is typical for WLI. These speckles enable the measurement, but they also introduce noise proportional to the roughness of the measured surface.

Despite the noise that speckles introduce into the WLI measurement, WLI has an advantage because it rejects the light that has undergone scattering outside of a small sample volume, thus allowing precise noninvasive measurement of object structure, even in dense media.

Applications

WLI is used in many disciplines and instruments such as:

- Fourier transform spectroscopy – source and material properties;
- Michelson stellar interferometer – angular size of star, binary stars, delay measurement in optical paths of interferometer;
- Shearing interferometry – structure measurement;
- DIC Nomarski interferometry – structure measurement;
- Speckle interferometry – structure measurement;
- Holography – structure measurement;
- Optical sensors – temperature, pressure, distance;
- Optical coherence tomography – structure measurement.

See also

Coherence: Coherence and Imaging; Overview. **Holography, Techniques:** Holographic Interferometry. **Interferometry:** Phase Measurement Interferometry. **Microscopy:** Interference Microscopy. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Born M and Wolf E (1997) *Principles of Optics*. New York: Cambridge University Press.

Creath K (1997) *Sampling requirements for white light interferometry*. Fringe'97 Proceedings of the 3rd International Workshop on Automatic Processing of Fringe Patterns, pp. 52–59.

Danielson BL and Boisbert CY (1991) Absolute optical ranging using low coherence interferometry. *Applied Optics* 30: 2975–2979.

Davis SP, Abrams MC and Brault JW (2001) *Fourier Transform Spectrometry*. London: Academic Press.

de Groot P and Deck L (1995) Surface profiling by analysis of white light interferograms in the spatial frequency domain. *Journal of Modern Optics* 42: 389–401.

Dresdel T, Hausler G and Venzke H (1992) Three dimensional sensing of rough surfaces by coherence radar. *Applied Optics* 31(7): 919–925.

Harasaki A, Schmit J and Wyant JC (2001) Offset envelope position due to phase change on reflection. *Applied Optics* 40: 2102–2106.

Hariharan P (1992) *Basics of Interferometry*. New York: Academic Press.

Hariharan P, Larkin KG and Roy M (1994) The geometric phase: interferometric observation with white light. *Journal of Modern Optics* 41: 663–667.

Larkin KG (1996) Neat nonlinear algorithm for envelope detection in white light interferometry. *Journal of the Optical Society of America A* 13: 832–842.

Michelson AA and Benoit JR (1895) Détermination expérimentale de la valeur du métre en longueurs d'ondes lumineuses. *Travaux et Mémoires des Bureau International des Poids et Mesures* 11: 1.

Olszak AG and Schmit JG (2003) High stability white light interferometry with reference signal for real time correction of scanning errors. *Optical Engineering* 42(1): 54–59.

Park M-C and Kim S-W (2000) Direct quadratic polynomial fitting for fringe peak detection of white light scanning interferograms. *Optical Engineering* 39: 952–959.

Pavlicek P and Soubusta J (2004) Measurement of the influence of dispersion on white light interferometry. *Applied Optics* 43: 766–770.

Schmit JG and Olszak A (2002) High precision shape measurement by white light interferometry with real time scanner error correction. *Applied Optics* 41(28): 5943–5950.

Sheppard CJR and Larkin KG (1995) Effects of numerical aperture on interference fringe spacing. *Applied Optics* 34: 4731–4734.

Sheppard CJR and Roy M (2003) Low-coherence interference microscopy. In: Torok P and Kao F-J (eds) *Optical Imaging and Microscopy*, pp. 257–273. Berlin, Heidelberg, New York: Springer-Verlag.

Steel W (1987) *Interferometry*. New York: Cambridge University Press.

Turyshew SG (2003) Analytical modeling of the white light fringe. *Applied Optics* 42(1): 71–90.

L

LASERS

Contents

Carbon Dioxide Laser
Dye Lasers
Edge Emitters
Excimer Lasers
Free Electron Lasers
Metal Vapor Lasers
Noble Gas Ion Lasers
Optical Fiber Lasers
Organic Semiconductors and Polymers
Planar Waveguide Lasers
Semiconductor Lasers
Up-Conversion Lasers

Carbon Dioxide Laser

C R Chatwin, University of Sussex, Brighton, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article gives a brief history of the development of the laser and goes on to describe the characteristics of the carbon dioxide laser and the molecular dynamics that permit it to operate at comparatively high power and efficiency. It is these commercially attractive features and its low cost that has led to its adoption as one of most popular industrial power beams. This outline also describes the main types of carbon dioxide laser and briefly discusses their characteristics and uses.

Brief History

In 1917 Albert Einstein developed the concept of stimulated emission which is the phenomenon used in lasers. In 1954 the MASER (Microwave Amplification by Stimulated Emission of Radiation) was the first device to use stimulated emission. In that year Townes and Schawlow suggested that stimulated emission could be used in the infrared and optical portions of the electromagnetic spectrum. The device was originally termed the optical maser, this term being dropped in favor of LASER, standing for: Light Amplification by Stimulated Emission of Radiation. Working against the wishes of his manager at Hughes Research Laboratories, the electrical engineer Ted Maiman created the first laser on the 15 May 1960. Maiman's flash lamp pumped ruby laser produced pulsed red electromagnetic radiation at a wavelength of 694.3 nm. During the most active period of laser systems discovery Bell Labs made a very

significant contribution. In 1960, Ali Javan, William Bennet and Donald Herriot produced the first Helium Neon laser, which was the first continuous wave (CW) laser operating at 1.15 μm. In 1961, Boyle and Nelson developed a continuously operating Ruby laser and in 1962, Kumar Patel, Faust, McFarlane and Bennet discovered five noble gas lasers and lasers using oxygen mixtures. In 1964, C.K.N. Patel created the high-power carbon dioxide laser operating at 10.6 μm. In 1964, J.F. Geusic and R.G. Smith produced the first Nd:Yag laser using neodymium doped yttrium aluminum garnet crystals and operating at 1.06 μm.

Characteristics

Due to its operation between low lying vibrational energy states of the CO_2 molecule, the CO_2 laser has a high quantum efficiency, ~40%, which makes it extremely attractive as a high-power industrial materials processing laser (1 to 20 kW), where energy and running costs are a major consideration. Due to the requirement for cooling to retain the population inversion, the efficiency of electrical pumping and optical losses – commercial systems have an overall efficiency of approximately 10%. Whilst this may seem low, for lasers this is still a high efficiency. The CO_2 laser is widely used in other fields, for example, surgical applications, remote sensing, and measurement. It emits infrared radiation with a wavelength that can range from 9 μm up to 11 μm. The laser transition may occur on one of two transitions: $(00^01) \rightarrow (10^00)$, $\lambda = 10.6$ μm; $(00^01) \rightarrow (02^00)$, $\lambda = 9.6$ μm, see Figure 1. The 10.6 μm transition has the maximum probability of oscillation and gives the strongest output; hence, this is the usual wavelength of operation, although for specialist applications the laser can be forced to operate on the 9.6 μm line.

Figure 1 illustrates an energy level diagram with four vibrational energy groupings that include all the significantly populated energy levels. The internal relaxation rates within these groups are considered to be infinitely fast when compared with the rate of energy transfer between these groups. In reality the internal relaxation rates are at least an order of magnitude greater than the rates between groups.

Excitation of the upper laser level is usually provided by an electrical glow discharge. However, gas dynamic lasers have been built where expanding a hot gas through a supersonic nozzle creates the population inversion; this creates a nonequilibrium region in the downstream gas stream with a large population inversion, which produces a very high-power output beam (135 kW – Avco Everett Research Lab). For some time the gas dynamic laser was seriously considered for use in the space-based Strategic Defence Initiative (SDI-USA). The gas mixture used in a CO_2 laser is usually a mixture of carbon dioxide, nitrogen, and helium. The proportions of these gases varies from one laser system to another, however, a typical mixture is 10%-CO_2;

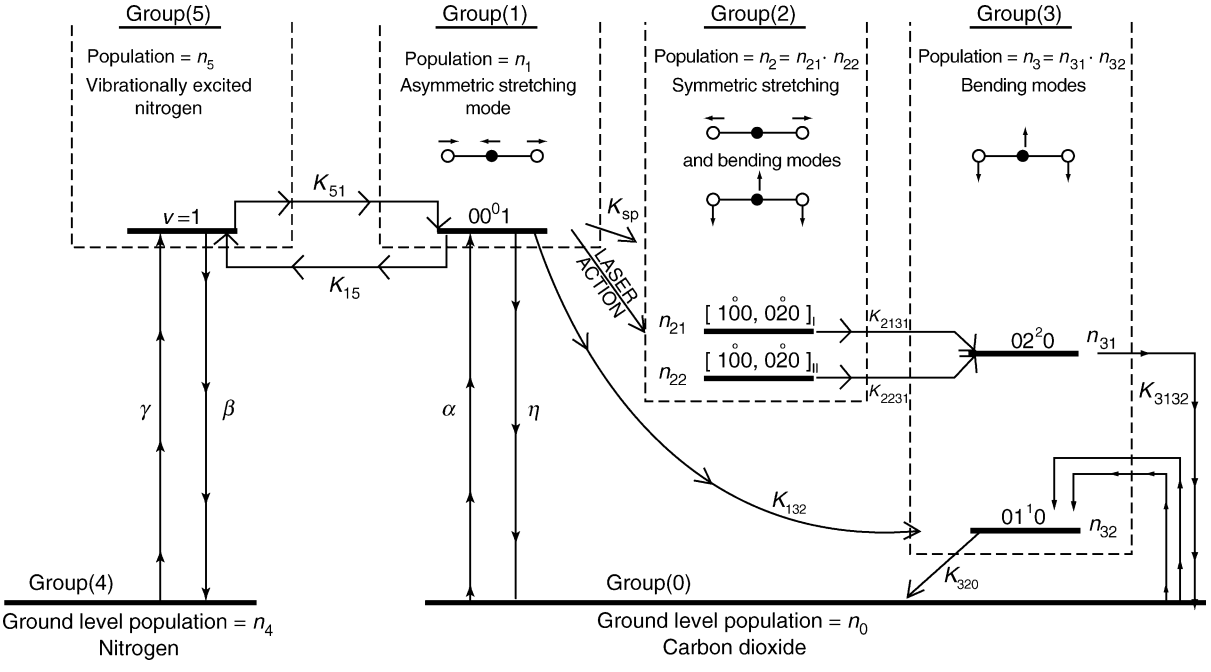


Figure 1 Six level model used for the theoretical description of CO_2 laser action.

10%-N_2; 80%-He. Helium plays a vital role in the operation of the CO_2 laser in that it maintains the population inversion by depopulating the lower laser level by nonradiative collision processes. Helium is also important for stabilization of the gas discharge; furthermore it greatly improves the thermal conductivity of the gas mixture, which assists in the removal of waste heat via heat exchangers.

Small quantities of other gases are often added to commercial systems in order to optimize particular performance characteristics or stabilize the gas discharge; for brevity we only concern ourselves here with this simple gas mixture.

Molecular Dynamics

Direct Excitation and De-excitation

It is usual for excitation to be provided by an electrical glow discharge. The direct excitation of carbon dioxide (CO_2) and nitrogen (N_2) ground state molecules proceeds via inelastic collisions with fast electrons. The rates of kinetic energy transfer are α and γ, respectively, and are given by eqns [1] and [2]:

$$\alpha = \frac{F_{CO_2} \times \mathrm{IP}}{E_{00^01} \times n_0} (\mathrm{sec}^{-1}) \qquad [1]$$

$$\gamma = \frac{F_{N_2} \times \mathrm{IP}}{E_{v=1} \times n_4} (\mathrm{sec}^{-1}) \qquad [2]$$

where:

F_{CO_2} = Fraction of the input power (IP) coupled into the excitation of the energy level E_{00^01}, n_0 is the CO_2 ground level population density;

F_{N_2} = Fraction of the input power (IP) coupled into the excitation of the energy level $E_{v=1}$; and n_4 is the N_2 ground level population density.

The reverse process of the above occurs when molecules lose energy to the electrons and the electrons gain an equal amount of kinetic energy; the direct de-excitation rates are given by η and β, eqns [3] and [4], respectively:

$$\eta = \alpha \times \exp\left(\frac{E_{00^01}}{E_e}\right) (\mathrm{sec}^{-1}) \qquad [3]$$

$$\beta = \gamma \times \exp\left(\frac{E_{v=1}}{E_e}\right) (\mathrm{sec}^{-1}) \qquad [4]$$

where E_e is the average electron energy in the discharge.

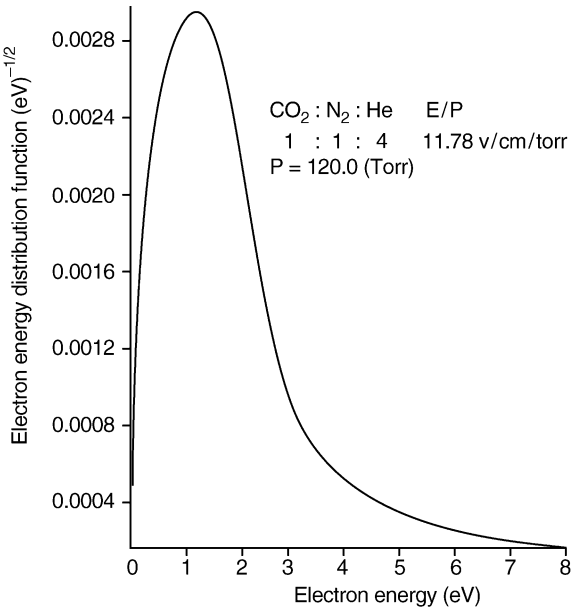


Figure 2 Electron energy distribution function.

E_e, F_{CO_2}, and F_{N_2} are obtained by solution of the Boltzmann transport equation (BTE); the average electron energy can be optimized to maximize the efficiency (F_{CO_2}, F_{N_2}) with which electrical energy is utilized to create a population inversion. Hence, the discharge conditions required to maximize efficiency can be predicted from the transport equation. Figure 2 shows one solution of the BTE for the electron energy distribution function.

Resonant Energy Transfer

Resonant energy transfer between the CO_2 (00^01) and N_2 ($v = 2$) energy levels (denoted 1 and 5 in Figure 1) proceeds via excited molecules colliding with ground state molecules. A large percentage of the excitation of the upper laser level takes place via collisions between excited N_2 molecules and ground state CO_2 molecules. The generally accepted rate of this energy transfer is given by eqns [5] and [6]:

$$K_{51} = 19\,000\, P_{CO_2} (\mathrm{sec}^{-1}) \qquad [5]$$

$$K_{15} = 19\,000\, P_{N_2} (\mathrm{sec}^{-1}) \qquad [6]$$

where P_{CO_2} and P_{N_2} are the respective gas partial pressures in Torr.

Hence, CO_2 molecules are excited into the upper laser level by both electron impact and impact with excited N_2 molecules. The contribution from N_2 molecules can be greater than 40% depending on the discharge conditions.

Collision–Induced Vibrational Relaxation of the Upper and Lower Laser Levels

The important vibrational relaxation processes are illustrated by Figure 1 and can be evaluated from eqns [7–10]; where the subscripts refer to the rate between energy levels 1 and 32; 21 and 31; 22 and 31; 32 and 0, respectively:

$$K_{132} = 367\,P_{CO_2} + 110\,P_{N_2} + 67\,P_{He}\ (\sec^{-1}) \qquad [7]$$

$$K_{2131} = 6\times10^5\,P_{CO_2}\ (\sec^{-1}) \qquad [8]$$

$$K_{2231} = 5.15\times10^5\,P_{CO_2}\ (\sec^{-1}) \qquad [9]$$

$$K_{320} = 200\,P_{CO_2} + 215\,P_{N_2} + 3270\,P_{He}\ (\sec^{-1}) \qquad [10]$$

K_{132}, K_{2131}, and K_{2231} are vibration/vibration transfer rates and K_{320} is a vibration/translation transfer rate. Note the important effect of helium on eqn [10]; helium plays a major role in depopulating the lower laser level, thus enhancing the population inversion. P_{He} is the partial pressure of helium in Torr.

Radiative Relaxation

Spontaneous radiative decay is not a major relaxation process in the CO_2 laser but it is responsible for starting laser action via spontaneous emission. The Einstein 'A' coefficient for the laser transition is given by eqn [11]:

$$A = (\tau_{sp})^{-1} = 0.213\ (\sec^{-1}) \qquad [11]$$

Gain

The gain (g) is evaluated from the product of the absorption coefficient (σ) and the population inversion, eqn [12]:

$$g = \sigma\left(n_{00^01} - \frac{g_1}{g_2}n_{10^00}\right)\ \mathrm{cm}^{-1} \qquad [12]$$

For most commercial laser systems the absorption coefficient is that for high-pressure collision-broadening ($P > 5.2$ Torr) where the intensity distribution function describing the line shape is Lorentzian. The following expression describes the absorption coefficient, eqn [13]:

$$\sigma = \frac{692.5}{Tn_{CO_2}\left(1 + 1.603\frac{n_{N_2}}{n_{CO_2}} + 1.4846\frac{n_{He}}{n_{CO_2}}\right)}\ (\mathrm{cm}^2) \qquad [13]$$

where T is the absolute temperature and n refers to the population density of the gas designated by the subscript.

This expression takes account of the different constituent molecular velocity distributions and different collision cross-sections for $CO_2 \rightarrow CO_2$, $N_2 \rightarrow CO_2$ and $He \rightarrow CO_2$ type collisions. Equation [13] also takes account of the significant line broadening effect of helium.

Neglecting the unit change in rotational quantum number, the energy level degeneracies g_1 and g_2 may be dropped. n_{10^00} is partitioned such that $n_{10^00} = 0.1452n_2$ and eqn [12] can be re-cast as eqn [14]:

$$g = \sigma(n_1 - 0.1452n_2)\ \mathrm{cm}^{-1} \qquad [14]$$

where n_1 and n_2 are the population densities of energy groups '1' and '2' respectively.

Stimulated Emission

Consider a laser oscillator with two plane mirrors, one placed at either end of the active gain medium, with one mirror partially transmitting (see Figure 4). Laser action is initiated by spontaneous emission that happens to produce radiation whose direction is normal to the end mirrors and falls within the resonant modes of the optical resonator. The rate of change of photon population density (I_p) within the laser cavity can be written as eqn [15]:

$$\frac{dI_p}{dt} = I_p cg - \frac{I_p}{T_0} \qquad [15]$$

where the first term on the right-hand side accounts for the effect of stimulated emission and the second term represents the number of photons that decay out of the laser cavity, T_0 is the photon decay time, given by eqn [16], and is defined as the average time a photon remains inside the laser cavity before being lost either through the laser output window or due to dispersion; if dispersion is ignored, I_p/T_0, is the laser output:

$$T_0 = \frac{2L}{c\log_e\left(\frac{1}{R_B R_F}\right)} \qquad [16]$$

where L is the distance between the back and the front mirrors, which have reflectivities of R_B and R_F, respectively. The dominant laser emission occurs on a rotational–vibrational P branch transition $P(22)$, that is $(J = 21) \rightarrow (J = 22)$ line of the $(00^01) \rightarrow (10^00)$, $\lambda = 10.6\ \mu$m transition, where J is the rotational quantum number. The rotational level relaxation rate is so rapid that equilibrium is maintained between rotational levels so that they

feed all their energy through the $P(22)$ transition. This model simply assumes constant intensity, basing laser performance on the performance of an average unit volume. By introducing the stimulated emission term into the molecular rate equations, which describe the rate of transfer of molecules between the various energy levels illustrated in Figure 1, a set of molecular rate eqns [17–21] can be written that permit simulation of the performance of a carbon dioxide laser:

$$\frac{dn_1}{dt} = \alpha n_0 - \eta n_1 + K_{51} n_5 - K_{15} n_1 - K_{sp} n_1 - K_{132}\left(n_1 - \left[\frac{n_1}{n_{32}}\right]_e n_{32}\right) - I_p c g \qquad [17]$$

$$\frac{dn_2}{dt} = K_{sp} n_1 + I_p c g - K_{2131}\left(n_{21} - \left[\frac{n_{21}}{n_{31}}\right]_e n_{31}\right) - K_{2231}\left(n_{22} - \left[\frac{n_{22}}{n_{31}}\right]_e n_{31}\right) \qquad [18]$$

$$\frac{dn_3}{dt} = 2K_{2131}\left(n_{21} - \left[\frac{n_{21}}{n_{31}}\right]_e n_{31}\right) - 2K_{2231}\left(n_{22} - \left[\frac{n_{22}}{n_{31}}\right]_e n_{31}\right) + K_{132}\left(n_1 - \left[\frac{n_1}{n_{32}}\right]_e n_{32}\right) - K_{320}\left(n_{32} - \left[\frac{n_{32}}{n_0}\right]_e n_0\right) \qquad [19]$$

$$\frac{dn_5}{dt} = \gamma n_4 - \beta n_5 - K_{51} n_5 + K_{15} n_1 \qquad [20]$$

$$\frac{dI_p}{dt} = I_p c g - \frac{I_p}{T_0} \qquad [21]$$

The terms in square brackets ensure that the system maintains thermodynamic equilibrium; subscript 'e' refers to the fact that the populations in the square brackets are the values for thermodynamic equilibrium. The set of five simultaneous differential equations can be solved using a Runge–Kutta method. They can provide valuable performance prediction data that is helpful in optimizing laser design, especially when operated in the pulsed mode. Figure 3a illustrates some simulation results for the transverse flow laser shown in Figure 9. The results illustrate the effect of altering the gas mixture and how this can be used to control the gain switched spike that would result in unwelcome work piece plasma generation if allowed to become too large. Figure 3b shows an experimental laser output pulse from the high pulse repetition frequency (prf–5kHz) laser illustrated in Figure 9. This illustrates that even a quite basic physical model can give a good prediction of laser output performance.

Optical Resonator

Figure 4 shows a simple schematic of an optical resonator. This simple optical system consists of two mirrors (the full reflector is often a water cooled gold coated copper mirror), which are aligned to be orthogonal to the optical axis that runs centrally along the length of the active gain medium in which there is a population inversion. The output coupler is a partial reflector (usually dielectrically coated zinc selenide – ZnSe–that may be edge cooled) so that some of the electromagnetic radiation can escape as an output beam. The ZnSe output coupler has a natural reflectivity of about 17% at each air–solid interface. For high power lasers (2 kW) 17% is sufficient for laser operation; however, depending on the required laser performance the inside surface is often given a reflective coating. The reflectivity of the inside face depends on the balance between the gain (eqn [14]), the output power and the power stability requirements. The outside face of the partial reflector must be anti-reflection (AR) coated.

Spontaneous emission occurs within the active gain medium and radiates randomly in all directions; a fraction of this spontaneous emission will be in the same direction as the optical axis, perpendicular to the end mirrors, and will also fall into a resonant mode of the optical resonator. Spontaneous emission photons interact with CO_2 molecules in the excited upper laser level, excited state (00^01), which stimulates these molecules to give up a quanta of vibrational energy as photons via the radiative transition $(00^01) \rightarrow (10^00)$, $\lambda = 10.6\mu m$. The radiation given up will have exactly the same phase and direction as the stimulating radiation and thus will be coherent with it. The reverse process of absorption also occurs, but so long as there is a population inversion there will be a net positive output. This process is called light amplification by stimulated emission of radiation (LASER). The mirrors continue to redirect the photons parallel to the optical axis and so long as the population inversion is not depleted, more and more photons are stimulated by stimulated emission which dominates the process and also dominates the spontaneous emission, which is important to initiate laser action.

Light emitted by lasers contains several optical frequencies, which are a function of the different modes of the optical resonator; these are simply the standing wave patterns that can exist within the resonator structure. There are two types of

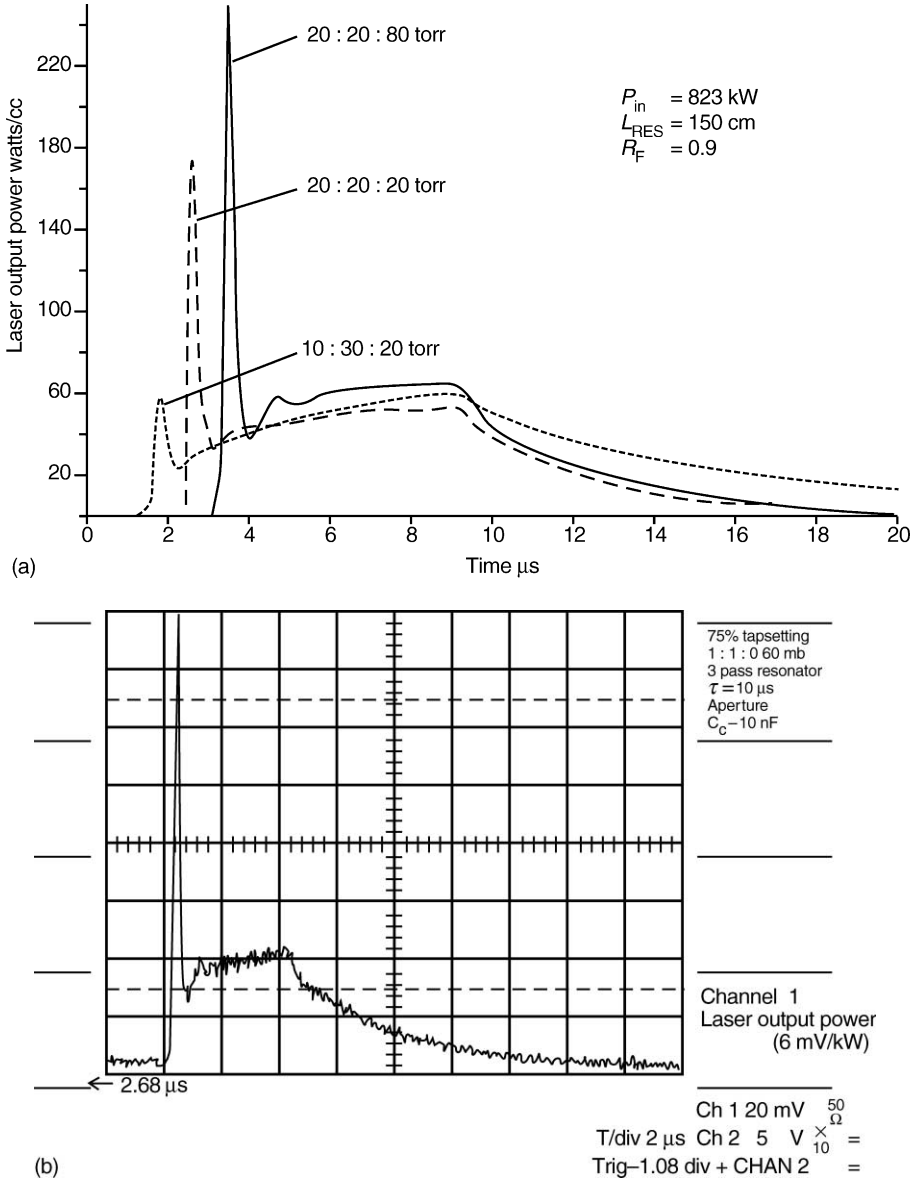


Figure 3 (a) Predicted output pulses for transverse flow CO_2 laser for different gas mixtures, (b) Experimental output pulse from transverse flow CO_2 laser.

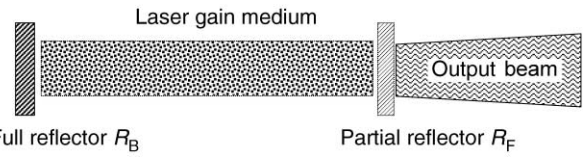


Figure 4 Optical resonator.

resonator modes: longitudinal and transverse. Longitudinal modes differ from one another in their frequency of oscillation whereas transverse modes differ from one another in their oscillation frequency and field distribution in a plane orthogonal to the direction of propagation. Typically CO_2 lasers have a large number of longitudinal modes; in CO_2 laser applications these are of less interest than the transverse modes, which determine the transverse beam intensity and the nature of the beam when focused. In cylindrical coordinates the transverse modes are labelled TEM_{pl}, where subscript '*p*' is the number of radial nodes and '*l*' is the number of angular nodes. The lowest order mode is the TEM_{00}, which has a Gaussian-like intensity profile with its maximum on the beam axis. A light beam emitted from an optical resonator with a Gaussian profile is said to be operating in the 'fundamental mode' or the TEM_{00} mode. The decrease in irradiance (I) with distance 'r' from the axis (I_0) of the Gaussian beam is

described by eqn [22]:

$$I(r) = I_0 \exp(2r^2/w^2) \qquad [22]$$

where w is the radial distance, where the power density is decreased to $1/e^2$ of its axial value. Ideally a commercial laser should be capable of operation in the fundamental mode as, with few exceptions, this results in the best performance in applications. Laser cutting benefits from operation in the fundamental mode; however, welding or heat treatment applications may benefit from operation with higher-order modes. Output beams are usually controlled to be linearly or circularly polarized, depending upon the requirements of the application. For materials processing applications the laser beam is usually focused via a water cooled ZnSe lens or, for very high power lasers, a parabolic gold coated mirror. Welding applications will generally use a long focal length lens and cutting applications will use a short focal length, which generates a higher irradiance at the work piece than that necessary for welding. The beam delivery optics are usually incorporated into a nozzle assembly that can deliver cooling water and assist gases for cutting and anti-oxidizing shroud gases for welding or surface engineering applications.

Laser Configuration

CO_2 lasers are available in many different configurations and tend to be classified on the basis of their physical form and the gas flow arrangement, both of which greatly affect the output power available and beam quality. The main categories are: sealed off lasers, waveguide lasers, slow axial flow, fast axial flow, diffusion cooled, transverse flow, transversely excited atmospheric lasers, and gas dynamic lasers.

Sealed-Off and Waveguide Lasers

Depopulation of the lower laser level is via collision with the walls of the discharge tube, so the attainable output power scales with the length of the discharge column and not its diameter. Output powers are in the range 5 W to 250 W. Devices may be constructed from concentric glass tubes with the inner tube providing the discharge cavity and the outer tube acting to contain water-cooling of the inner discharge tube. The inner tube walls act as a heat sink for the discharge thermal energy (see Figure 5). The DC electrical discharge is provided between a cathode and anode situated at either end of the discharge tube. A catalyst must be provided to ensure regeneration of CO_2 from CO. This may be accomplished by adding about 1% of H_2O to the gas mixture, or alternatively,

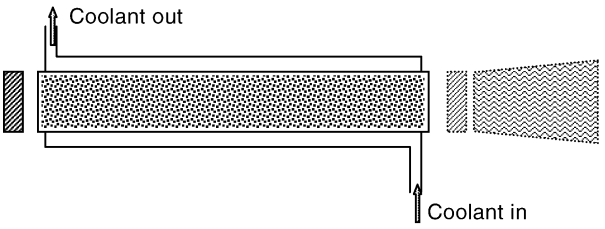


Figure 5 Schematic of sealed-off CO_2 laser, approximately 100 W per meter of gain length, gas cooled by diffusion to the wall.

recombination can be achieved via a hot (300 °C) Ni cathode, which acts as a catalyst. RF-excited all metal sealed-off tube systems can deliver lifetimes greater than 45 000 hours.

Diffusion-cooled slab laser technology will also deliver reliable sealed operation for 20 000 hrs. Excitation of the laser medium occurs via RF excitation between two water-cooled electrodes. The water-cooled electrodes dissipate (diffusion cooled) the heat generated in the gas discharge. An unstable optical resonator provides the output coupling for such a device (see Figure 6). Output powers are in the range 5 W to 300 W and can be pulsed from 0 to 100 kHz. These lasers are widely used for marking, rapid prototyping, and cutting of nonmetals (paper, glass, plastics, ceramics) and metals.

Waveguide CO_2 lasers use small bore tubes (2–4 mm) made of BeO or SiO_2 where the laser radiation is guided by the tube walls. Due to the small tube diameter, a gas total pressure of 100 to 200 Torr is necessary, hence the gain per unit length is high. This type of laser will deliver 30 W of output power from a relatively short (50 cm long) compact sealed-off design; such a system is useful for microsurgery and scientific applications. Excitation can be provided from a longitudinal DC discharge or from an RF source that is transverse to the optical axis; RF excitation avoids the requirement for an anode and cathode and results in a much lower electrode voltage.

Slow Axial Flow Lasers

In slow flow lasers the gas mixture flows slowly through the laser cavity. This is done to remove the products of dissociation that will reduce laser efficiency or prevent it from operating at all, and the main contaminant is CO. The dissociated gases (mainly CO and O_2) can be recombined using a catalyst pack and then reused via continuous recirculation. Heat is removed via diffusion through the walls of the tube containing the active gain medium. The tube is frequently made of Pyrex glass with a concentric outer tube to facilitate water-cooling of the

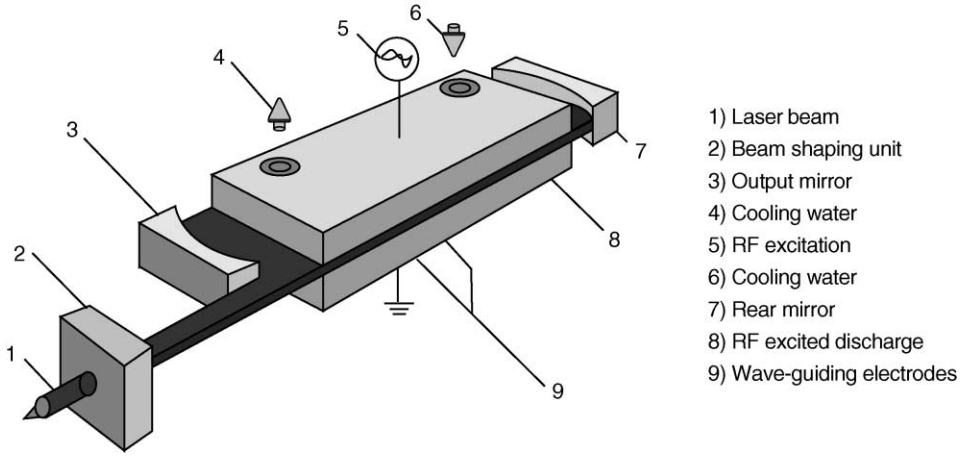


Figure 6 Schematic for sealed-off slab laser and diffusion cooled laser with RF excitation (courtesy of Rofin).

laser cavity (see Figure 7). Slow flow lasers operate in the power range 100 W to 1500 W, and tend to use a longitudinal DC electrical discharge which can be made to run continuously or pulsed if a thyratron switch is build into the power supply; alternatively, electrical power can be supplied via transverse RF excitation. The power scales with length, hence high power slow flow lasers have long cavities and require multiple cavity folds in order to reduce their physical size.

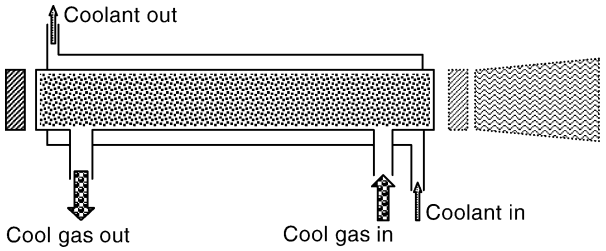


Figure 7 Slow flow CO_2 laser, approximately 100 W per meter of gain length, gas cooled by diffusion to the wall.

Fast Axial Flow Lasers

The fast axial flow laser, Figure 8, can provide output powers from 1 kW to 20 kW; it is this configuration that dominates the use of CO_2 lasers for industrial applications. Industrial lasers are usually in the power range 2–4 kW. The output power from these devices scales with mass flow, hence the gas mixture is recycled through the laser discharge region at sonic or supersonic velocities. Historically this was achieved using Rootes blowers to compress the gas upstream of the laser cavity. Rootes compressors are inherently inefficient and the more advanced laser systems utilize turbine compressors, which deliver greater efficiency and better laser stability. Rootes compressors can be a major source of vibration. With this arrangement heat exchangers are required to remove heat after the laser discharge region and also after the compressor stage, as the compression process heats up the laser gases. Catalyst packs are used to regenerate gases but some gas replacement is often required. These laser systems have short cavities and use folded stable resonator designs to achieve higher output powers with extremely high-quality beams that are particularly suitable for cutting applications. They also give excellent results when used for welding and surface treatments.

Fast axial flow lasers can be excited by a longitudinal DC discharge or a transverse RF discharge. Both types of electrical excitation are common. For materials processing applications it is often important to be able to run a laser in continuous wave (CW) mode or as a high pulse repetition rate (prf) pulsed laser and to be able to switch between CW and pulsed in real time; for instance, laser cutting of accurate internal corners is difficult using CW operation but very easy using the pulsed mode of operation. Both methods of discharge excitation can provide this facility.

Diffusion Cooled Laser

The diffusion-cooled slab laser is RF excited and gives an extremely compact design capable of delivering 4.5 kW pulsed from 8 Hz to 5 kHz prf or CW with good beam quality (see Figure 6). The optical resonator is formed by the front and rear mirrors and two parallel water cooled RF-electrodes. Diffusion cooling is provided by the RF-electrodes, removing the requirement for conventional gas recirculation via Rootes blowers or turbines.

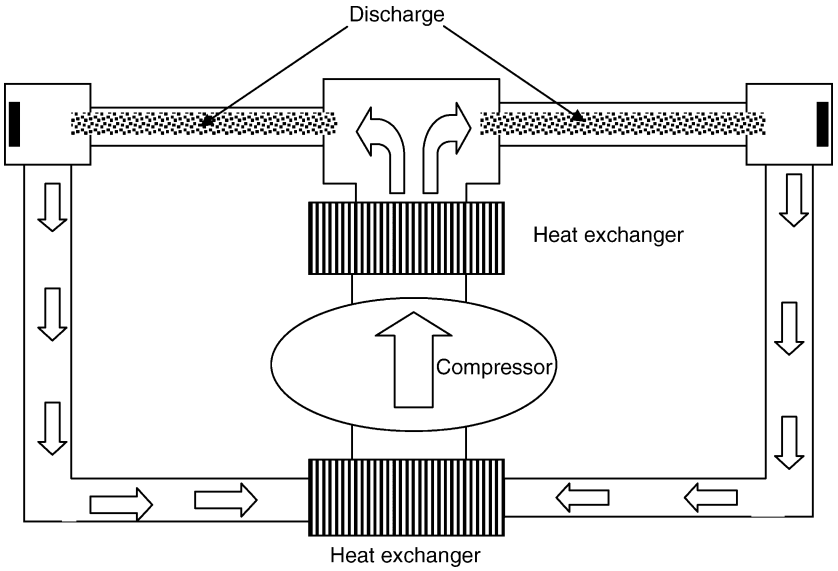


Figure 8 Fast axial flow carbon dioxide laser.

This design of laser results in a device with an extremely small footprint that has low maintenance and running costs. Applications include: cutting, welding, and surface engineering.

Fast Transverse Flow Laser

In the fast transverse flow laser (Figure 9a) the gas flow, electrical discharge, and the output beam are at right angles to each other (Figure 9b). The transverse discharge can be high voltage DC, RF, or pulsed up to 8 kHz (Figure 9c). Very high output power per unit discharge length can be obtained with an optimal total pressure (P) of ~100 Torr; systems are available delivering 10 kW of output power, CW or pulsed (see Figures 3a and b). The increase in total pressure requires a corresponding increase in the gas discharge electric field, E, as the ratio E/P must remain constant, since this determines the temperature of the discharge electrons, which have an optimum mean value (optimum energy distribution, Figure 2) for efficient excitation of the population inversion. With this high value of electric field, a longitudinal-discharge arrangement is impractical (500 kV for a 1 m discharge length); hence, the discharge is applied perpendicular to the optical axis. Fast transverse flow gas lasers provided the first multikilowatt outputs but tend to be expensive to maintain and operate. In order to obtain a reasonable beam quality, the output coupling is often obtained using a multipass unstable resonator. As the population inversion is available over a wide rectangular cross-section, this is a disadvantage of this arrangement and beam quality is not as good as that obtainable from fast axial flow designs. For this reason this type of laser is suitable for a wide range of welding and surface treatment applications.

Transversely Excited Atmospheric (TEA) Pressure

If the gas total pressure is increased above ~100 Torr it is difficult to sustain a stable glow discharge, because above this pressure instabilities degenerate into arcs within the discharge volume. This problem can be overcome by pulse excitation; using submicrosecond pulse duration, instabilities do not have sufficient time to develop; hence, the gas pressure can be increased above atmospheric pressure and the laser can be operated in a pulsed mode. In a mode locked format optical pulses shorter than 1 ns can be produced. This is called a TEA laser and with a transverse gas flow is capable of producing short high power pulses up to a few kHz repetition frequency. In order to prevent arc formation, TEA lasers usually employ ultraviolet or e-beam preionization of the gas discharge just prior to the main current pulse being applied via a thyratron switch. Output coupling is usually via an unstable resonator. TEA lasers are used for marking, remote sensing, range-finding, and scientific applications.

Conclusions

It is 40 years since Patel operated the first high power CO_2 laser. This led to the first generation of lasers which were quickly exploited for industrial laser materials processing, medical applications, defense,

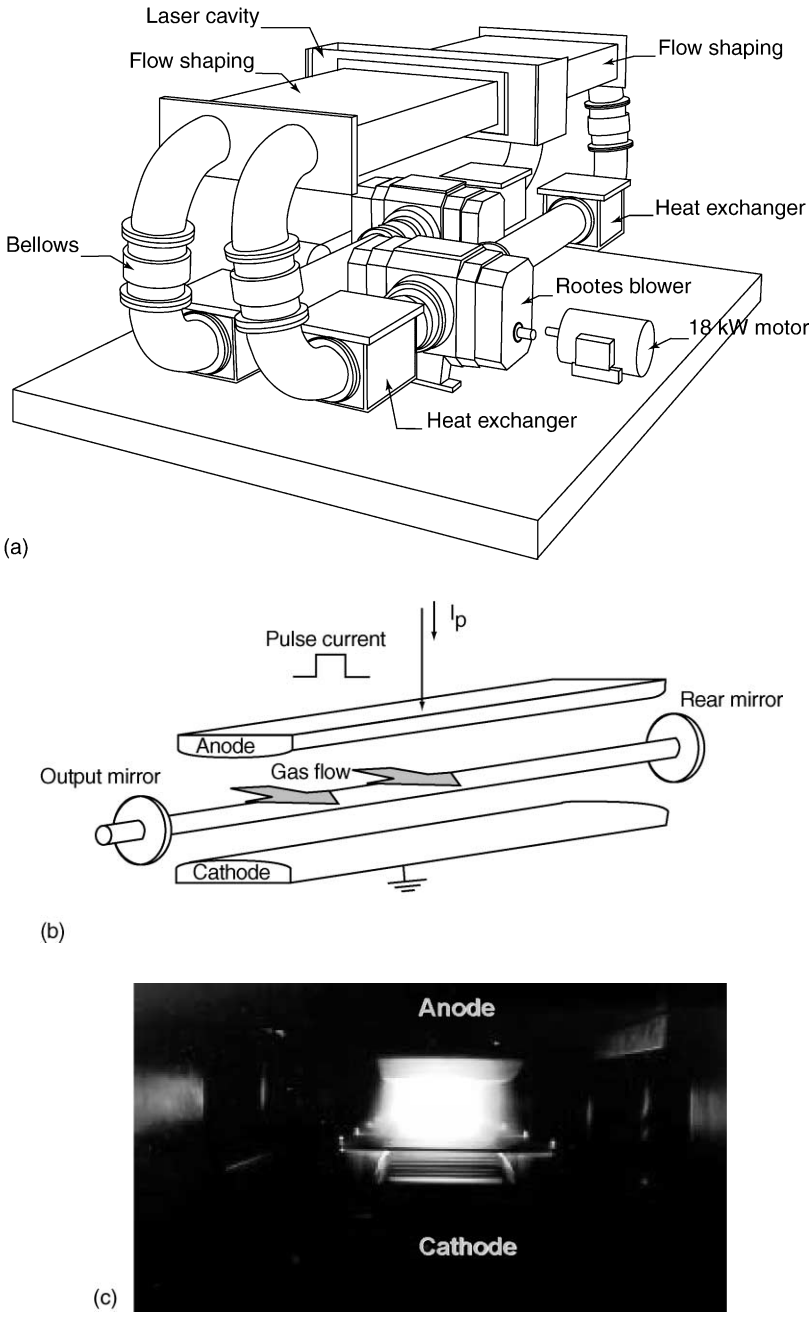


Figure 9 (a) Transverse flow carbon dioxide laser gas recirculator, (b) Transverse flow carbon dioxide electrodes, (c) Transverse flow carbon dioxide gas discharge as seen from the output window.

and scientific research applications; however, the first generation of lasers were quite unreliable and temperamental. After many design iterations, the CO_2 laser has now matured into a reliable, stable laser source available in many different geometries and power ranges. The low cost of ownership of the latest generation of CO_2 laser makes them a very attractive commercial proposition for many industrial and scientific applications. Commercial lasers incorporate many novel design features that are beyond the scope of this article and are often peculiar to the particular laser manufacturer. This includes gas additives and catalysts that may be required to stabilize the gas discharge of a particular laser design; it is this optimization of the laser design that has produced such reliable and controllable low-cost performance from the CO_2 laser.

List of Units and Nomenclature

α	The direct excitation of carbon dioxide (CO_2) ground state molecules (sec^{-1})
β	The direct de-excitation of nitrogen (N_2) (sec^{-1})
γ	The direct excitation of nitrogen (N_2) ground state molecules (sec^{-1})
η	The direct de-excitation of carbon dioxide (CO_2) (sec^{-1})
λ	wavelength (m)
σ	absorption coefficient (cm^2)
τ	Electrical current pulse length (μs)
τ_{sp}	Spontaneous emission life time of the upper laser level (sec)
$A = (\tau_{sp})^{-1} = 0.213(sec^{-1})$	The Einstein 'A' coefficient for the laser transition
c	velocity of light ($cm\ sec^{-1}$)
C_c	Coupling capacitance (nF)
CO_2	Carbon dioxide
e	subscript 'e' refers to the fact that the populations in the square brackets are the values for thermodynamic equilibrium
E	Electric Field ($V\ cm^{-1}$)
F_{CO_2}	Fraction of the input power (IP) coupled into the excitation of the energy level E_{00^01}
F_{N_2}	Fraction of the input power (IP) coupled into the excitation of the energy level $E_{v=1}$
g	gain (cm^{-1})
g_1 and g_2	energy level degeneracy's of levels 1 and 2
He	Helium
I	beam irradiance ($W\ cm^{-2}$)
I_0	beam irradiance at the center of a Gaussian laser beam ($W\ cm^{-2}$)
I_p	Photon population density ($photons\ cm^{-3}$)
I_p	Input current (A)
IP	Electrical input power ($W\ cm^{-3}$)
J	the rotational quantum number
K_{51}, K_{15}	Resonant energy transfer between the CO_2 (00^01) and $N_2(v = 2)$ energy levels proceeds via excited molecules colliding with ground state molecules (sec^{-1})
K_{132}, K_{2131}, K_{2231}	are vibration/vibration transfer rates (sec^{-1}) between energy levels 1 and 32; 21 and 31; 22 and 31, respectively (see Figure 1)
K_{320}	is a vibration/translation transfer rate (sec^{-1}) between energy levels 32 and 0 (see Figure 1)
K_{sp}, A	Spontaneous emission rate (sec^{-1})
L	The distance between the back and the front mirrors, which have reflectivity's of R_B and R_F (cm)
n	molecular population ($molecules\ cm^{-3}$)
N_2	Nitrogen
P	Pressure (Torr)
P_{CO_2}, P_{He} and P_{N_2}	The respective gas partial pressures (Torr)
P_{in}	Electrical input power (kW)
r	radius of laser beam (cm)
R_B	Back mirror reflectivity
R_F	Front mirror reflectivity
t	time (sec)
T	Temperature (deg K)
T_0	the photon decay time (sec)
w	the radial distance where the power density is decreased to $1/e^2$ of its axial value

See also

Fiber and Guided Wave Optics: Overview. **Lasers:** Noble Gas Ion Lasers.

Further Reading

Anderson JD (1976) *Gasdynamic Lasers: An Introduction*. New York: Academic Press.

Chatwin CR, McDonald DW and Scott BF (1991) Design of a High P.R.F. Carbon Dioxide Laser for Processing High Damage Threshold Materials. Selected Papers on Laser Design, SPIE Milestone Series, pp. 425–433. Washington: SPIE Optical Engineering Press.

Cool AC (1969) Power and gain characteristics of high speed flow lasers. *Journal of Applied Physics* 40(9): 3563.

Crafer RC, Gibson AF, Kent MJ and Kimmit MF (1969) Time-dependent processes on CO_2 laser amplifiers. *British Journal of Applied Physics* 2(2): 183.

Gerry ET and Leonard AD (1966) Measurement of 10.6-μ CO_2 laser transition probability and optical broadening cross sections. *Applied Physics Letters* 8(9): 227.

Gondhalekar A, Heckenberg NR and Holzhauer E (1975) The mechanism of single-frequency operation of the hybrid CO_2 laser. *IEEE Journal of Quantum Electronics* QE-11(3): 103.

Gordiets BF, Sobolev NN and Shelepin LA (1968) Kinetics of physical processes in CO_2 lasers. *Soviet Physics JETP* 26(5): 1039.

Herzberg G (1945) *Molecular Spectra & Molecular Structure*. Infra-red and Raman Spectra of Polyatomic Molecules, vol. 2. New York: Van Nostrand.

Hoffman AL and Vlases GC (1972) A simplified model for predicting gain, saturation and pulse length for gas dynamic lasers. *IEEE Journal of Quantum Electronics* 8(2): 46.
Johnson DC (1971) Excitation of an atmospheric pressure CO_2–N_2–He laser by capacitor discharges. *IEEE Journal of Quantum Electronics* Q.E.-7(5): 185.
Koechner W (1988) *Solid State Laser Engineering*. Berlin: Springer Verlag.
Kogelnik H and Li T (1966) Laser beams and resonators. *Applied Optics* 5: 1550–1567.
Levine AK and De Maria AJ (1971). *Lasers*, vol. 3. New York: Marcel Dekkar, Chapter 3.
Moeller G and Ridgen JD (1965) *Applied Physics Letters* 7: 274.
Moore CB, Wood RE, Bei-Lok Hu and Yardley JT (1967) Vibrational energy transfer in CO_2 lasers. *Journal of Chemical Physics* 11: 4222.
Patel CKN (1964) *Physics Review Letters* 12: 588.
Siegman A (1986) *Lasers*. Mill Valley, California: University Science.
Smith K and Thomson RM (1978) *Computer Modeling of Gas Lasers*. New York and London: Plenum Press.
Sobolev NN and Sokovikov VV (1967) CO_2 lasers. *Soviet Physics USPEKHI* 10(2): 153.
Svelto O (1998) *Principles of Lasers*, 4th edn. New York: Plenum.
Tychinskii VP (1967) Powerful lasers. *Soviet Physics USPEKHI* 10(2): 131.
Vlases GC and Money WM (1972) Numerical modelling of pulsed electric CO_2 lasers. *Journal of Applied Physics* 43(4): 1840.
Wagner WG, Haus HA and Gustafson KT (1968) High rate optical amplification. *IEEE Journal of Quantum Electronics* Q.E.-4: 287.
Witteman W (1987) *The CO_2 Laser*. Springer Series in Optical Sciences, vol. 53.

Dye Lasers

F J Duarte, Eastman Kodak Company, New York, NY, USA

A Costela, Consejo Superior de Investigaciones Cientificas, Madrid, Spain

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Background

Dye lasers are the original tunable lasers. Discovered in the mid-1960s these tunable sources of coherent radiation span the electromagnetic spectrum from the near-ultraviolet to the near-infrared (Figure 1). Dye lasers spearheaded and sustained the revolution in atomic and molecular spectroscopy and have found use in many and diverse fields from medical to military applications. In addition to their extraordinary spectral versatility, dye lasers have been shown to oscillate from the femtosecond pulse domain to the continuous wave (cw) regime. For microsecond pulse emission, energies of up to hundreds of joules per pulse have been demonstrated. Further, operation at high pulsed repetition frequencies (prfs), in the multi-kHz regime, has provided average powers at kW levels. This unrivaled operational versatility is summarized in Table 1.

Dye lasers are excited by coherent optical energy from an excitation, or pump, laser or by optical energy from specially designed lamps called flashlamps. Recent advances in semiconductor laser

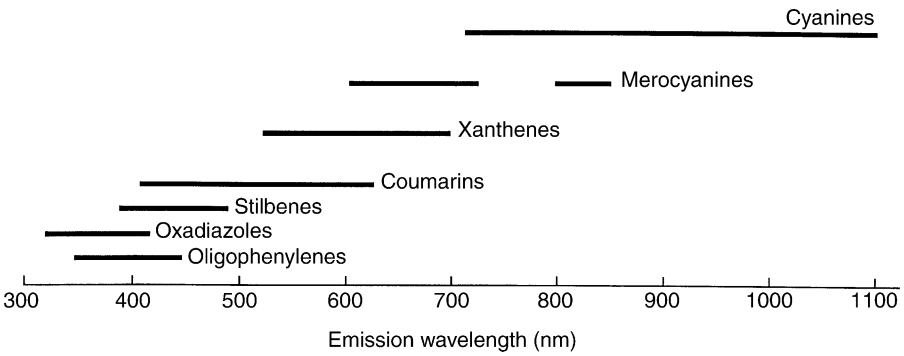


Figure 1 Approximate wavelength span from the various classes of laser dye molecules. Reproduced with permission from Duarte FJ (1995) *Tunable Laser Handbook*. New York: Academic Press.

Table 1 Emission characteristics of liquid dye lasers

Dye laser class	*Spectral coverage*[a]	*Energy per pulse*[b]	*Prf*[b]	*Power*[b]
Laser-pumped pulsed dye lasers	350–1100 nm	800 J[c]	13 kHz[d]	2.5 kW[d]
Flashlamp-pumped dye lasers	320–900 nm	400 J[e]	850 Hz[f]	1.2 kW[f]
CW dye lasers	370–1000 nm			43 W[g]

[a]Approximate range.
[b]Refers to maximum values for that particular emission parameter.
[c]Achieved with an excimer-laser pumped coumarin dye laser by Tang and colleagues, in 1987.
[d]Achieved with a multistage copper-vapor-laser pumped dye laser using rhodamine dye by Bass and colleagues, in 1992.
[e]Reported by Baltakov and colleagues, in 1974. Uses rhodamine 6G dye.
[f]Reported by Morton and Dragoo, in 1981. Uses coumarin 504 dye.
[g]Achieved with an Ar^+ laser pumped folded cavity dye laser using rhodamine 6G dye by Baving and colleagues, in 1982.

technology have made it possible to construct very compact all-solid-state excitation sources that, coupled with new solid-state dye laser materials, should bring the opportunity to build compact tunable laser systems for the visible spectrum. Further, direct diode-laser pumping of solid-state dye lasers should prove even more advantageous to enable the development of fairly inexpensive tunable narrow-linewidth solid-state dye laser systems for spectroscopy and other applications requiring low powers. Work on electrically excited organic gain media might also provide new avenues for further progress.

The literature of dye lasers is very rich and many review articles have been written describing and discussing traditional dye lasers utilizing liquid gain media. In particular, the books *Dye Lasers*, *Dye Laser Principles*, *High Power Dye Lasers*, and *Tunable Lasers Handbook* provide excellent sources of authoritative and detailed description of the physics and technology involved. In this article we offer only a survey of the operational capabilities of the dye lasers using liquid gain media in order to examine with more attention the field of solid-state dye lasers.

Brief History of Dye Lasers

1965: Quantum theory of dyes is discussed in the context of the maser (R. P. Feynman).
1966: Dye lasers are discovered (P. P. Sorokin and J. R. Lankard; F. P. Schäfer and colleagues).
1967: The flashlamp-pumped dye laser is discovered (P. P. Sorokin and J. R. Lankard; W. Schmidt and F. P. Schäfer).
1967–1968: Solid-state dye lasers are discovered (B. H. Soffer and B. B. McFarland; O. G. Peterson and B. B. Snavely).
1968: Mode-locking, using saturable absorbers, is demonstrated in dye lasers (W. Schmidt and F. P. Schäfer).
1970: The continuous-wave (cw) dye laser is discovered (O. G. Peterson, S. A. Tuccio, and B. B. Snavely).
1971: The distributed feedback dye laser is discovered (H. Kogelnik and C. V. Shank).
1971–1975: Prismatic beam expansion in dye lasers is introduced (S. A. Myers; E. D. Stokes and colleagues; D. C. Hanna and colleagues).
1972: Passive mode-locking is demonstrated in cw dye lasers (E. P. Ippen, C. V. Shank, and A. Dienes).
1972: The first pulsed narrow-linewidth tunable dye laser is introduced (T. W. Hänsch).
1973: Frequency stabilization of cw dye lasers is demonstrated (R. L. Barger, M. S. Sorem, and J. L. Hall).
1976: Colliding-pulse-mode locking is introduced (I. S. Ruddock and D. J. Bradley).
1977–1978: Grazing-incidence grating cavities are introduced (I. Shoshan and colleagues; M. G. Littman and H. J. Metcalf; S. Saikan).
1978–1980: Multiple-prism grating cavities are introduced (T. Kasuya and colleagues; G. Klauminzer; F. J. Duarte and J. A. Piper).
1981: Prism pre-expanded grazing-incidence grating oscillators are introduced (F. J. Duarte and J. A. Piper).
1982: Generalized multiple-prism dispersion theory is introduced (F. J. Duarte and J. A. Piper).
1983: Prismatic negative dispersion for pulse compression is introduced (W. Dietel, J. J. Fontaine, and J-C. Diels).
1987: Laser pulses as short as six femtoseconds are demonstrated (R. L. Fork, C. H. Brito Cruz, P. C. Becker, and C. V. Shank).
1994: First narrow-linewidth solid-state dye laser oscillator (F. J. Duarte).
1999–2000: Distributed feedback solid-state dye lasers are introduced (Wadsworth and colleagues; Zhu and colleagues).

Molecular Energy Levels

Dye molecules have large molecular weights and contain extended systems of conjugated double bonds. These molecules can be dissolved in an adequate organic solvent (such as ethanol, methanol, ethanol/water, and methanol/water) or incorporated into a solid matrix (organic, inorganic, or hybrid). These molecular gain media have a strong absorption generally in the visible and ultraviolet regions, and exhibit large fluorescence bandwidths covering the entire visible spectrum. The general energy level diagram of an organic dye is shown in Figure 2. It consists of electronic singlet and triplet states with each electronic state containing a multitude of overlapping vibrational–rotational levels giving rise to broad continuous energy bands. Absorption of visible or ultraviolet pump light excites the molecules from the ground state S_0 into some rotational–vibrational level belonging to an upper excited singlet state, from where the molecules decay nonradiatively to the lowest vibrational level of the first excited singlet state S_1 on a picosecond time-scale. From S_1 the molecules can decay radiatively, with a radiative lifetime on the nanosecond time-scale, to a higher-lying vibrational–rotational level of S_0. From this level they rapidly thermalize into the lowest vibrational–rotational levels of S_0. Alternatively, from S_1, the molecules can experience nonradiative relaxation either to the triplet state T_1 by an intersystem crossing process or to the ground state by an internal conversion process. If the intensity of the pumping radiation is high enough a population inversion between S_1 and S_0 may be attained and stimulated emission occurs. Internal conversion and intersystem crossing compete with the fluorescence decay mode of the molecule and therefore reduce the efficiency of the laser emission. The rate for internal conversion to the electronic ground state is usually negligibly small so that the most important loss process is intersystem crossing into T_1 that populates the lower metastable triplet state. Thus, absorption on the triplet–triplet allowed transitions could cause considerable losses if these absorption bands overlap the lasing band, inhibiting or even halting the lasing process. This triplet loss can be reduced by adding small quantities of appropriate chemicals that favor nonradiative transitions that shorten the effective lifetime of the T_1 level. For pulsed excitation with nanosecond pulses, the triplet–triplet absorption can be neglected because for a typical dye the intersystem crossing rate is not fast enough to build up an appreciable triplet population in the nanosecond time domain.

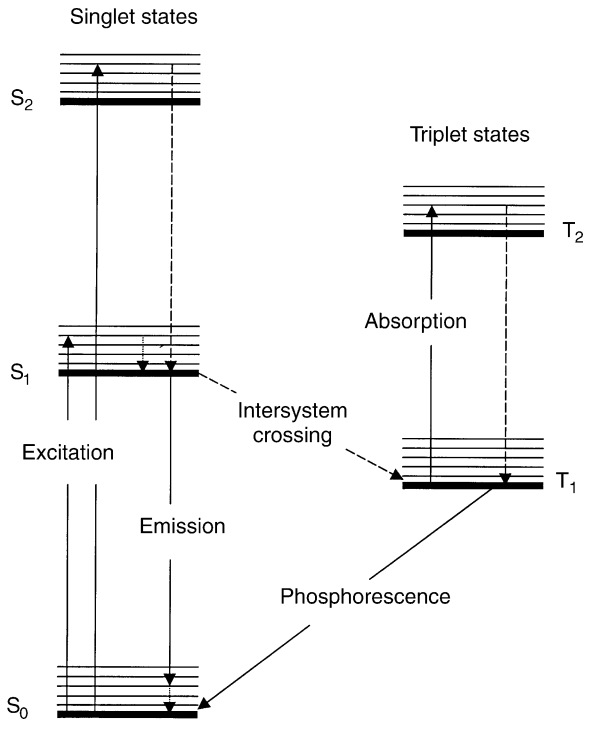


Figure 2 Schematic energy level diagram for a dye molecule. Full lines: radiative transitions; dashed lines: nonradiative transitions; dotted lines: vibrational relaxation.

Dye molecules are large (a typical molecule incorporates 50 or more atoms) and are grouped into families with similar chemical structures. A survey of the major classes of laser dyes is given later. Solid-state laser dye gain media are also considered later.

Liquid Dye Lasers

Laser-Pumped Pulsed Dye Lasers

Laser-pumped dye lasers use a shorter wavelength, or higher frequency, pulsed laser as the excitation or pump source. Typical pump lasers for dye lasers are gas lasers such as the excimer, nitrogen, or copper lasers. One of the most widely used solid-state laser pumps is the frequency doubled Nd:YAG laser which emits at 532 nm.

In a laser-pumped pulsed dye laser the active medium, or dye solution, is contained in an optical cell often made of quartz or fused silica, which provides an active region typically some 10 mm in length and a few mm in width. The active medium is then excited either longitudinally, or transversely, via a focusing lens using the pump laser. In the case of transverse excitation the pump laser is focused to a beam $\sim$10 mm in width and $\sim$0.1 mm in height. Longitudinal pumping requires focusing of the excitation beam to a diameter in the 0.1–0.15 mm range. For lasers operated at low prfs (a few pulses per

second), the dye solution might be static. However, for high-prf operation (a few thousand pulses per second) the dye solution must be flowed at speeds of up to a few meters per second in order to dissipate the heat. A simple broadband optically pumped dye laser can be constructed using just the pump laser, the active medium, and two mirrors to form a resonator. In order to achieve tunable, narrow-linewidth, emission, a more sophisticated resonator must be employed. This is called a dispersive tunable oscillator and is depicted in Figure 3. In a dispersive tunable oscillator the exit side of the cavity is comprised of a partial reflector, or an output coupler, and the other end of the resonator is composed of a multiple-prism grating assembly. It is the dispersive characteristics of this multiple-prism grating assembly and the dimensions of the emission beam produced at the gain medium that determine the tunability and the narrowness, or spectral purity, of the laser emission.

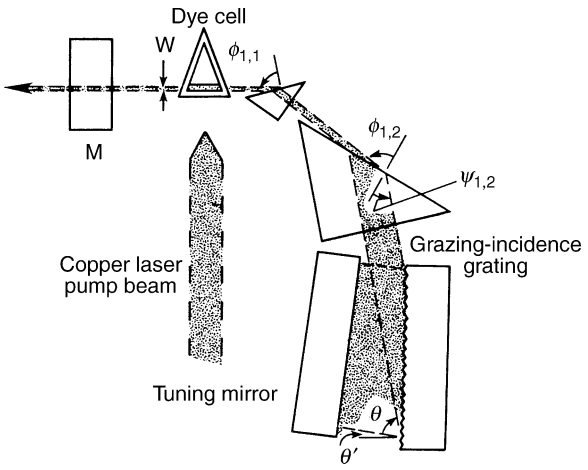


Figure 3 Copper-vapor-laser pumped hybrid multiple-prism near grazing incidence (HMPGI) grating dye laser oscillator. Adapted with permission from Duarte FJ and Piper JA (1984) Narrow linewidth high prf copper laser-pumped dye-laser oscillators. *Applied Optics* 23: 1391–1394.

In order to selectively excite a single vibrational–rotational level of a molecule such as iodine (I_2), at room temperature, one needs a laser linewidth of $\Delta\nu \approx 1.5$ GHz (or $\Delta\lambda \approx 0.0017$ nm at $\lambda = 590$ nm). The hybrid multiple-prism near-grazing-incidence (HMPGI) grating dye laser oscillator illustrated in Figure 4 yields laser linewidths in the 400 MHz $\leq \Delta\nu \leq$ 650 MHz range at 4–5% conversion efficiencies whilst excited by a copper-vapor laser operating at a prf of 10 kHz. Pulse lengths are ~10 ns at full-width half-maximum (FWHM). The narrow-linewidth emission from these oscillators is said to be single-longitudinal-mode lasing because only one electromagnetic mode is allowed to oscillate.

The emission from oscillators of this class can be amplified many times by propagating the tunable narrow-linewidth laser beam through single-pass amplifier dye cells under the excitation of pump lasers. Such amplified laser emission can reach enormous average powers. Indeed, a copper-vapor-laser pumped dye laser system at the Lawrence Livermore National Laboratory (USA), designed for the laser isotope separation program, was reported to

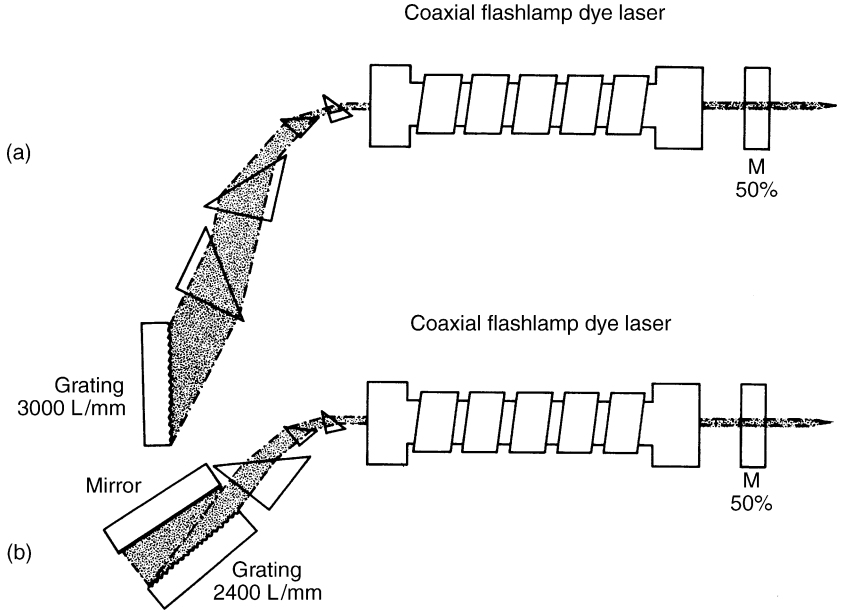


Figure 4 Flashlamp-pumped multiple-prism grating oscillators. From Duarte FJ, Davenport WE, Ehrlich JJ and Taylor TS (1991) Ruggedized narrow-linewidth dispersive dye laser oscillator. *Optics Communications* 84: 310–316. Reproduced with permission from Elsevier.

yield average powers in excess of 2.5 kW, at a prf of 13.2 kHz, at a better than 50% conversion efficiency as reported by Bass and colleagues in 1992. Besides high conversion efficiencies, excitation with copper-vapor lasers, at $\lambda = 510.554$ nm, has the advantage of inducing little photodegradation in the active medium thus allowing very long dye lifetimes.

Flashlamp-Pumped Dye Lasers

Flashlamps utilized in dye laser excitation emit at black-body temperatures in the 20 000 K range, thus yielding intense ultraviolet radiation centered around 200 nm. One further requirement for flashlamps, and their excitation circuits, is to deliver light pulses with a fast rise time. For some flashlamps this rise time can be less than a few nanoseconds.

Flashlamp-pumped dye lasers differ from laser-pumped pulsed dye lasers mainly in the pulse energies and pulse lengths attainable. This means that flashlamp-pumped dye lasers, using relatively large volumes of dye, can yield very large energy pulses. Excitation geometries use either coaxial lamps, with the dye flowing in a quartz cylinder at the center of the lamp, or two or more linear lamps arranged symmetrically around the quartz tube containing the dye solution. Using a relatively weak dye solution of rhodamine-6G (2.2×10^{-5} M), a coaxial lamp, and an active region defined by a quartz tube 6 cm in diameter and a length of 60 cm, Baltakov and colleagues, in 1974, reported energies of 400 J in pulses 25 μs long at FWHM.

Flashlamp-pumped tunable narrow-linewidth dye laser oscillators described by Duarte and colleagues, in1991, employ a cylindrical active region 6 mm in diameter and 17 cm in length. The dye solution is made of rhodamine 590 at a concentration of 1×10^{-5} M. This active region is excited by a coaxial flashlamp. Using multiple-prism grating architectures (see Figure 4) these authors achieve a diffraction limited TEM_{00} laser beam and laser linewidths of $\Delta\nu \approx 300$ MHz at pulsed energies in the 2–3 mJ range. The laser pulse duration is reported to be $\Delta t \approx 100$ ns. The laser emission from this class of multiple-prism grating oscillator is reported to be extremely stable. The tunable narrow-linewidth emission from these dispersive oscillators is either used directly in spectroscopic, or other scientific applications, or is utilized to inject large flashlamp-pumped dye laser amplifiers to obtain multi-joule pulse energies with the laser linewidth characteristics of the oscillator.

Continuous Wave Dye Lasers

CW dye lasers use dye flowing at linear speeds of up to 10 meters per second which are necessary to remove the excess heat and to quench the triplet states. In the original cavity reported by Peterson and colleagues, in 1970, a beam from an Ar^+ laser was focused on to an active region which is contained within the resonator. The resonator comprised dichroic mirrors that transmit the blue-green radiation of the pump laser and reflect the red emission from the dye molecules. Using a pump power of about 1 W, in a TEM_{00} laser beam, these authors reported a dye laser output of 30 mW. Subsequent designs replaced the dye cell with a dye jet, an introduced external mirror, and integrated dispersive elements in the cavity. Dispersive elements such as prisms and gratings are used to tune the wavelength output of the laser. Frequency-selective elements, such as etalons and other types of interferometers, are used to induce frequency narrowing of the tunable emission. Two typical cw dye laser cavity designs are described by Hollberg, in 1990, and are reproduced here in Figure 5. The first design is a linear three-mirror folded cavity. The second one is an eight-shaped ring dye laser cavity comprised of mirrors M_1, M_2, M_3, and M_4. Linear cavities exhibit the effect of *spatial hole burning* which allows the cavity to lase in more than one longitudinal mode. This problem can be overcome in ring cavities (Figure 5b) where the laser emission is in the form of a traveling wave.

Two aspects of cw dye lasers are worth emphasizing. One is the availability of relatively high powers in single longitudinal mode emission and the other is the

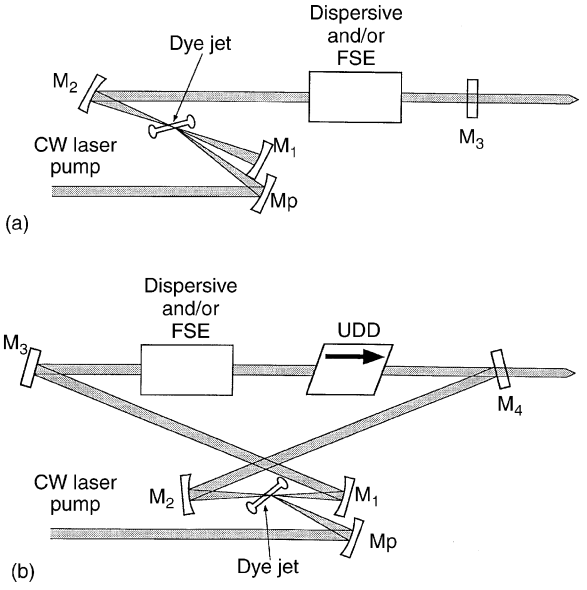


Figure 5 CW laser cavities: (a) linear cavity and (b) ring cavity. Adapted from Hollberg LW (1990) CW dye lasers. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 185–238. New York: Academic Press. Reproduced with permission from Elsevier.

demonstration of very stable laser oscillation. First, Johnston and colleagues, in 1982, reported 5.6 W of stabilized laser output in a single longitudinal mode at 593 nm at a conversion efficiency of 23%. In this work eleven dyes were used to span the spectrum continuously from ~400 nm to ~900 nm. In the area of laser stabilization and ultra narrow-linewidth oscillation it is worth mentioning the work of Hough and colleagues, in 1984, that achieved laser linewidths of less than 750 Hz employing an external reference cavity.

Ultrashort-Pulse Dye Lasers

Ultrashort-pulse, or femtosecond, dye lasers use the same type of technology as cw dye lasers configured to incorporate a saturable absorber region. One such configuration is the ring cavity depicted in Figure 6. In this cavity the gain region is established between mirrors M_1 and M_2 whilst the saturable absorber is deployed in a counter-propagating arrangement. This arrangement is necessary to establish a collision between two counter-propagating pulses at the saturable absorber thus yielding what is known as colliding-pulse mode locking (CPM) as reported by Ruddock and Bradley, in 1976. This has the effect of creating a transient grating, due to interference, at the absorber thus shortening the pulse. Intracavity prisms were incorporated in order to introduce negative dispersion, by Dietel and colleagues in 1983, and thus subtract dispersion from the cavity and ultimately provide the compensation needed to produce femtosecond pulses.

The shortest pulse obtained from a dye laser, 6 fs, has been reported by Fork and colleagues, in 1987, using extra-cavity compression. In that experiment, a dye laser incorporating CPM and prismatic compensation was used to generate pulses that were amplified by a copper-vapor laser at a prf of 8 kHz. The amplified pulses, of a duration of 50 fs, were then propagated through two grating pairs and a four-prism sequence for further compression.

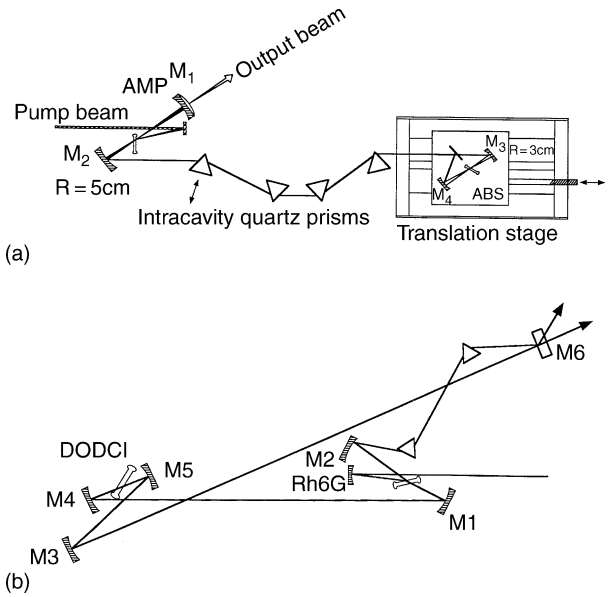


Figure 6 Femtosecond dye laser cavities: (a) linear femtosecond cavity and (b) ring femtosecond cavity. Adapted from Diels J-C (1990) Femtosecond dye lasers. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 41–132. New York: Academic Press. Reproduced with permission from Elsevier.

Solid-State Dye Laser Oscillators

In this section the principles of linewidth narrowing in dispersive resonators are outlined. Albeit the discussion focuses on multiple-prism grating solid-state dye laser oscillators, in particular, the physics is applicable to pulsed high-power dispersive tunable lasers in general.

Multiple-Prism Dispersion Grating Theory

The spectral linewidth in a dispersive optical system is given by

$$\Delta\lambda \approx \Delta\theta(\nabla_\lambda \theta)^{-1} \qquad [1]$$

where $\Delta\theta$ is the light beam divergence, $\nabla_\lambda = \partial/\partial\lambda$, and $\nabla_\lambda\theta$ is the overall dispersion of the optics. This identity can de derived either from the principles of geometrical optics or from the principles of generalized interferometry as described by Duarte, in 1992.

The cumulative single-pass generalized multiple-prism dispersion at the mth prism, of a multiple-prism array as illustrated in Figure 7, as given by Duarte and Piper, in 1982,

$$\nabla_\lambda \phi_{2,m} = H_{2,m}\nabla_\lambda n_m + (k_{1,m}k_{2,m})^{-1} \times \left(H_{1,m}\nabla_\lambda n_m \pm \nabla_\lambda \phi_{2,(m-1)}\right) \qquad [2]$$

In this equation

$$k_{1,m} = \cos\psi_{1,m}/\cos\phi_{1,m} \qquad [3a]$$

$$k_{2,m} = \cos\phi_{2,m}/\cos\psi_{2,m} \qquad [3b]$$

$$H_{1,m} = \tan\phi_{1,m}/n_m \qquad [4a]$$

$$H_{2,m} = \tan\phi_{2,m}/n_m \qquad [4b]$$

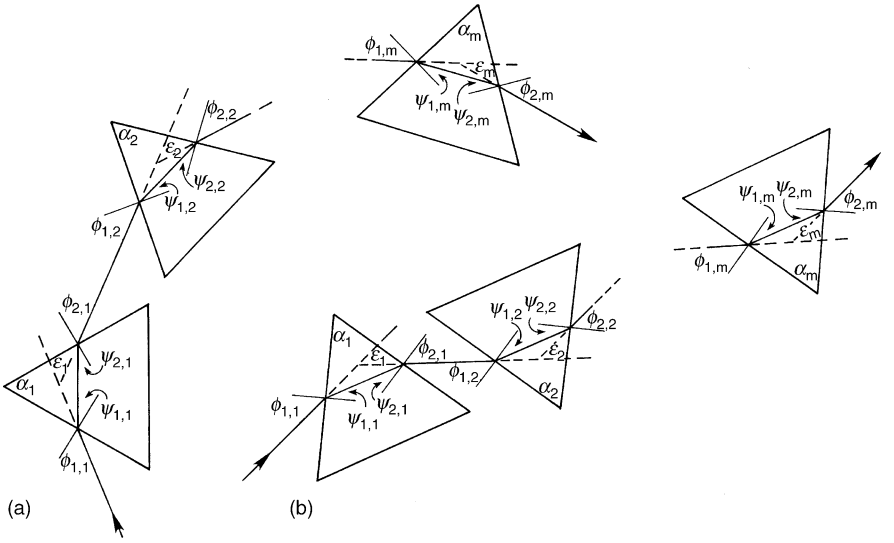


Figure 7 Generalized multiple-prism arrays in (a) additive and (b) compensating configurations. From Duarte FJ (1990) Narrow-linewidth pulsed dye laser oscillators. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 133–183. New York: Academic Press. Reproduced with permission from Elsevier.

Here, $k_{1,m}$ and $k_{2,m}$ represent the physical beam expansion experienced by the incident and the exit beams, respectively.

Equation [2] indicates that $\nabla_\lambda \phi_{2,m}$, the cumulative dispersion at the mth prism, is a function of the geometry of the mth prism, the position of the light beam relative to this prism, the refractive index of the prism, and the cumulative dispersion up to the previous prism $\nabla_\lambda \phi_{2,(m-1)}$.

For an array of r identical isosceles, or equilateral, prisms arranged symmetrically, in an additive configuration, so that the angles of incidence and emergence are the same, the cumulative dispersion reduces to

$$\nabla_\lambda \phi_{2,r} = r \nabla_\lambda \phi_{2,1} \qquad [5]$$

Under these circumstances the dispersions add in a simple and straightforward manner. For configurations incorporating right-angle prisms, the dispersions need to be handled mathematically in a more subtle form.

The generalized double-pass, or return-pass, dispersion for multiple-prism beam expanders was introduced by Duarte, in 1985:

$$\nabla_\lambda \Phi_P = 2M_1 M_2 \sum_{m=1}^{r} (\pm 1) H_{1,m} \left(\prod_{j=m}^{r} k_{1,j} \prod_{j=m}^{r} k_{2,j} \right)^{-1} \nabla_\lambda n_m + 2 \sum_{m=1}^{r} (\pm 1) H_{2,m} \left(\prod_{j=1}^{m} k_{1,j} \prod_{j=1}^{m} k_{2,j} \right) \nabla_\lambda n_m \qquad [6]$$

Here, M_1 and M_2 are the beam magnification factors given by

$$M_1 = \prod_{m=1}^{r} k_{1,m} \qquad [7a]$$

$$M_2 = \prod_{m=1}^{r} k_{2,m} \qquad [7b]$$

For a multiple prism expander designed for an orthogonal beam exit, and Brewster's angle of incidence, eqn [6] reduces to the succinct expression given by Duarte, in 1990:

$$\nabla_\lambda \Phi_P = 2 \sum_{m=1}^{r} (\pm 1)(n_m)^{m-1} \nabla_\lambda n_m \qquad [8]$$

Equation [6] can be used to either quantify the overall dispersion of a given multiple-prism beam expander or to design a prismatic expander yielding zero dispersion, that is, $\nabla_\lambda \Phi_P = 0$, at a given wavelength.

Physics and Architecture of Solid-State Dye-Laser Oscillators

The first high-performance narrow-linewidth tunable laser was introduced by Hänsch in 1972. This laser yielded a linewidth of $\Delta\nu \approx 2.5$ GHz (or $\Delta\lambda \approx 0.003$ nm at $\lambda \approx 600$ nm) in the absence of an intracavity etalon. Hänsch demonstrated that the laser linewidth from a tunable laser was narrowed significantly when the beam incident on the tuning grating was expanded using an astronomical

telescope. The linewidth equation including the intracavity beam magnification factor can be written as

$$\Delta\lambda \approx \Delta\theta(M\nabla_\lambda\Theta_G)^{-1} \qquad [9]$$

From this equation, it can be deduced that a narrow $\Delta\lambda$ is achieved by reducing $\Delta\theta$ and increasing the overall intracavity dispersion ($M\nabla_\lambda\Theta_G$). The intracavity dispersion is optimized by expanding the size of the intracavity beam incident on the diffractive surface of the tuning grating until it is totally illuminated.

Hänsch used a two-dimensional astronomical telescope to expand the intracavity beam incident on the diffraction grating. A simpler beam expansion method consists in the use of a single-prism beam expander as disclosed by several authors (Myers, 1971; Stokes and colleagues, 1972; Hanna and colleagues, 1975). An extension and improvement of this approach was the introduction of multiple-prism beam expanders as reported by Kasuya and colleagues, in 1978, Klauminzer, in 1978, and Duarte and Piper, in 1980. The main advantages of multiple-prism beam expanders, over two-dimensional telescopes, are simplicity, compactness, and the fact that the beam expansion is reduced from two dimensions to one dimension. Physically, as explained previously, prismatic beam expanders also introduce a dispersion component that is absent in the case of the astronomical telescope. Advantages of multiple-prism beam expanders over single-prism beam expansion are higher transmission efficiency, lower amplified spontaneous emission levels, and the flexibility to either augment or reduce the prismatic dispersion.

In general, for a pulsed multiple-prism grating oscillator, Duarte and Piper, in 1984, showed that the return-pass dispersive linewidth is given by

$$\Delta\lambda = \Delta\theta_R(MR\nabla_\lambda\Theta_G + R\nabla_\lambda\Phi_P)^{-1} \qquad [10]$$

where R is the number of return-cavity passes. The grating dispersion in this equation, $\nabla_\lambda\Theta_G$, can be either from a grating in Littrow or near grazing-incidence configuration. The multiple-return-pass equation for the beam divergence was given by Duarte, in 2001:

$$\Delta\theta_R = (\lambda/\pi w)\left(1 + (L_R/B_R)^2 + (A_R L_R/B_R)^2\right)^{1/2} \qquad [11]$$

Here, $L_R = (\pi w^2/\lambda)$ is the Rayleigh length of the cavity, w is the beam waist at the gain region, while A_R and B_R are the corresponding multiple-return-pass elements derived from propagation matrices.

At present, very compact and optimized multiple-prism grating tunable laser oscillators are found in two basic cavity architectures. These are the multiple-prism Littrow (MPL) grating laser oscillator, reported by Duarte in 1999 and illustrated in Figure 8, and the hybrid multiple-prism near grazing-incidence (HMPGI) grating laser oscillator, introduced by Duarte in 1997 and depicted in Figure 9. In early MPL grating oscillators the individual prisms integrating the multiple-prism expander were deployed in an additive configuration thus adding the cumulative dispersion to that of the grating and thus contributing to the overall dispersion of the cavity. In subsequent architectures the prisms were deployed in compensating configurations so as to yield zero dispersion and thus allow the tuning characteristics of the cavity to be

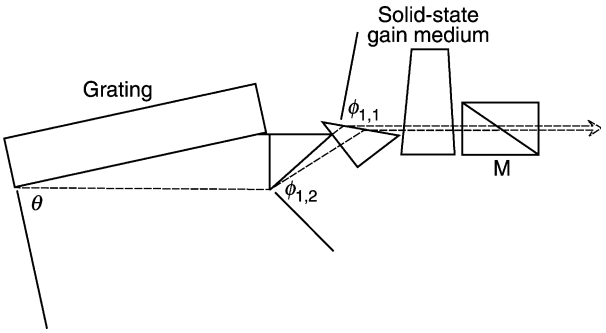


Figure 8 MPL grating solid-state dye laser oscillator: optimized architecture. The physical dimensions of the optical components of this cavity are shown to scale. The length, of the solid-state dye gain medium, along the optical axis, is 10 mm. Reproduced with permission from Duarte FJ (1999) Multiple-prism grating solid-state dye laser oscillator: optimized architecture. *Applied Optics* 38: 6347–6349.

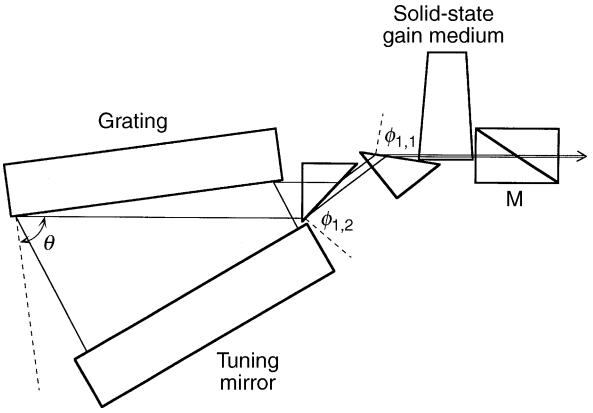


Figure 9 HMPGI grating solid-state dye laser. Schematics to scale. The length of the solid state dye gain medium, along the optical axis, is 10 mm. Reproduced with permission from Duarte FJ (1997) Multiple-prism near-grazing-incidence grating solid-state dye laser oscillator. *Optics and Laser Technology* 29: 513–516.

determined by the grating exclusively. In this approach the principal role of the multiple-prism array is to expand the beam incident on the grating thus augmenting significantly the overall dispersion of the cavity as described in eqns [9] or [10]. In this regard, it should be mentioned that beam magnification factors of up to 100, and beyond, have been reported in the literature. Using solid-state laser dye gain media, these MPL and HMPGI grating laser oscillators deliver tunable single-longitudinal-mode emission at laser linewidths 350 MHz $\leq \Delta\nu \leq$ 375 MHz and pulse lengths in the 3–7 ns (FWHM) range. Long-pulse operation of this class of multiple-prism grating oscillators has been reported by Duarte and colleagues, in 1998. In these experiments laser linewidths of $\Delta\nu \approx$ 650 MHz were achieved at pulse lengths in excess of 100 ns (FWHM) under flashlamp-pumped dye laser excitation.

The dispersive cavity architectures described here have been used with a variety of laser gain media in the gas, the liquid, and the solid state. Applications to tunable semiconductor lasers have also been reported by Zorabedian, in 1992, Duarte, in 1993, and Fox and colleagues, in 1997.

Concepts important to MPL and HMPGI grating tunable laser oscillators include the emission of a single-transverse-mode (TEM_{00}) laser beam in a compact cavity, the use of multiple-prism arrays, the expansion of the intracavity beam incident on the grating, the control of the intracavity dispersion, and the quantification of the overall dispersion of the multiple-prism grating assembly via generalized dispersion equations. Sufficiently high intracavity dispersion leads to the achievement of return-pass dispersive linewidths close to the free-spectral range of the cavity. Under these circumstances single-longitudinal-mode lasing is readily achieved as a result of multipass effects.

A Note on the Cavity Linewidth Equation

So far we have described how the cavity linewidth equation

$$\Delta\lambda \approx \Delta\theta(\nabla_\lambda \theta)^{-1}$$

and the dispersion equations can be applied to achieve highly coherent, or very narrow-linewidth, emission. It should be noted that the same physics can be applied to achieve ultrashort, or femtosecond, laser pulses. It turns out that intracavity prisms can be configured to yield negative dispersion as described by Duarte and Piper, in 1982, Dietel and colleagues, in 1983, and Fork and colleagues in 1984. This negative dispersion can reduce significantly the overall dispersion of the cavity. Under those circumstances, eqn [1] predicts broadband emission which according to the uncertainty principle, in the form of,

$$\Delta\nu\Delta t \approx 1 \qquad [12]$$

can lead to very short pulse emission since $\Delta\nu$ and $\Delta\lambda$ are related by the identities

$$\Delta\lambda \approx \lambda^2/\Delta x \qquad [13]$$

and

$$\Delta\nu \approx c/\Delta x \qquad [14]$$

Distributed-Feedback Solid-State Dye Lasers

Recently, narrow-linewidth laser emission from solid-state dye lasers has also been obtained using a distributed-feedback (DFB) configuration by Wadsworth and colleagues, in 1999, and Zhu and colleagues, in 2000. As is well known, in a DFB laser no external cavity is required, the feedback being provided by Bragg reflection from a permanently or dynamically written grating structure within the gain medium as reported by Kogelnik and Shank, in 1971. By using a DFB laser configuration where the interference of two pump beams induces the required periodic modulation in the gain medium, laser emission with linewidth in the 0.01–0.06 nm range have been reported. Specifically, Wadsworth and colleagues reported a laser linewidth of 12 GHz (0.016 nm at 616 nm) using Perylene Red doped PMMA at a conversion efficiency of 20%.

It should also be mentioned that the DFB laser is also a dye laser development that has found wide and extensive applicability in semiconductor lasers employed in the telecommunications industry.

Laser Dye Survey

Cyanines

These are red and near-infrared dyes which present long conjugated methine chains (!CH = CH!) and are useful in the spectral range longer than 800 nm, where no other dyes compete with them. An important laser dye belonging to this class is 4-dicyanomethylene-2-methyl-6-(*p*-dimethylamino-styryl)-4H-pyran (DCM, **Figure 10a**), with laser emission in the 600–700 nm range depending on the pump and solvent, liquid or solid, used.

Xanthenes

These dyes have a xanthene ring as the chromophore and are classified into rhodamines, incorporating

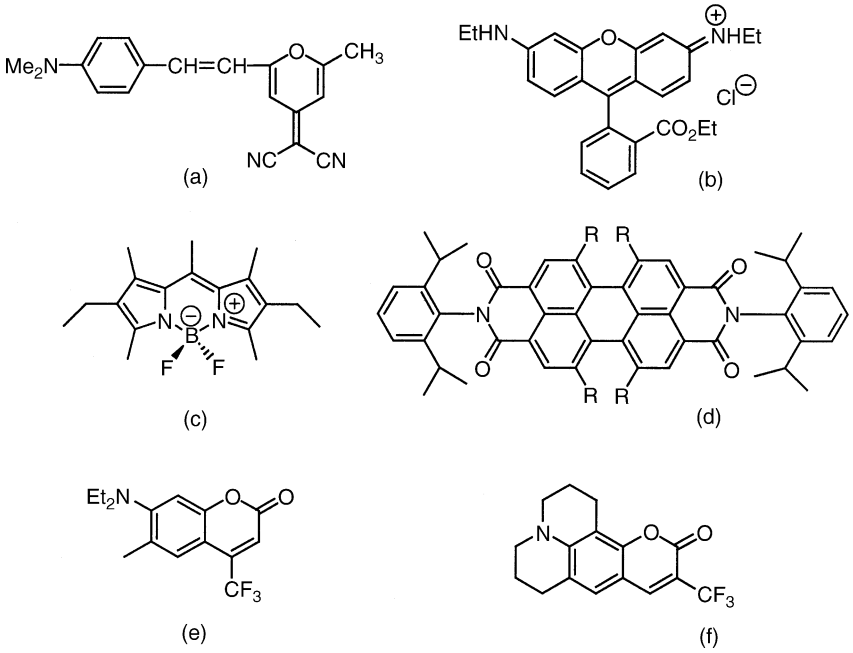


Figure 10 Molecular structures of some common laser dyes: (a) DCM; (b) Rh6G; (c) PM567; (d) Perylene Orange (R = H) and Perylene Red (R = C_4H_6O); (e) Coumarin 307 (also known as Coumarin 503); (f) Coumarin 153 (also known as Coumarin 540A).

amino radicals substituents, and fluoresceins, with hydroxyl (OH) radical substituents. They are generally very efficient and chemically stable and their emission covers the wavelength region from 500 to 700 nm.

Rhodamine dyes are the most important group of all laser materials, with rhodamine 6G (Rh6G, Figure 10b), also called rhodamine 590 chloride, being probably the best known of all laser dyes. Rh6G exhibits an absorption peak, in ethanol, at 530 nm and a fluorescence peak at 556 nm, with laser emission typically in the 550–620 nm region. It has been demonstrated to lase efficiently both in liquid and solid solutions.

Pyrromethenes

Pyrromethene.BF_2 complexes are a new class of laser dyes synthesized and characterized more recently as reported by Shah and colleagues, in 1990, Pavlopoulos and colleagues, in 1990, and Boyer and colleagues, in 1993. These laser dyes exhibit reduced triplet–triplet absorption over their fluorescence and lasing spectral region while retaining a high quantum fluorescence yield. Depending on the substituents on the chromophore, these dyes present laser emission over the spectral region from the green/yellow to the red, competing with the rhodamine dyes and have been demonstrated to lase with good performance when incorporated into solid hosts. Unfortunately, they are relatively unstable because of the aromatic amine groups in their structure, which render them vulnerable to photochemical reactions with oxygen as indicated by Rahn and colleagues, in 1997. The molecular structure of a representative and well-known dye of this class, 1,3,5,7,8-pentamethyl-2,6-diethylpyrromethene-difluoroborate complex (pyrromethene 567, PM567) is shown in Figure 10c. Recently, analogs of dye PM567 substituted at position 8 with an acetoxypolymethylene linear chain or a polymerizable methacryloyloxy polymethylene chain have been developed. These new dipyrromethene.BF_2 complexes have demonstrated improved efficiency and better photostability than the parent compound both in liquid and solid state.

Conjugated Hydrocarbons

This class of organic compounds includes perylenes, stilbenes, and *p*-terphenyl. Perylene and perylimide dyes are large, nonionic, nonpolar molecules (Figure 10d) characterized by their extreme photostability and negligible singlet–triplet transfer, as discussed by Seybold and Wagenblast and colleagues, in 1989, which have high quantum efficiency because of the absence of nonradiative relaxation. The perylene dyes exhibit limited solubility in conventional solvents but dissolve well in acetone, ethyl acetate, and methyl methacrylate, with emission wavelengths in the orange and red spectral regions.

Stilbene dyes are derivatives of uncyclic unsaturated hydrocarbons such as ethylene and butadiene, with emission wavelengths in the 400–500 nm range. Most of them end in phenyl radicals. Although they

are chemically stable, their laser performance is inferior to that of coumarins. *p*-Terphenyl is a valuable and efficient *p*-oligophenylene dye with emission in the ultraviolet.

Coumarins

Coumarin derivatives are a popular family of laser dyes with emission in the blue-green region of the spectrum. Their structure is based on the coumarin ring (**Figure 10e,f**) with different substituents that strongly affect its chemical characteristics and allow covering an emission spectral range between 420 and 580 nm. Some members of this class rank among the most efficient laser dyes known, but their chemical photobleaching is rapid compared with xanthene and pyrromethene dyes. An additional class of blue-green dyes is the tetramethyl derivatives of coumarin dyes introduced by Chen and colleagues, in 1988. The emission from these dyes spans the spectrum in the 453–588 nm region. These coumarin analogs exhibit improved efficiency, higher solubility, and better lifetime characteristics than the parent compounds.

Azaquinolone Derivatives

The quinolone and azaquinolone derivatives have a structure similar to that of coumarins but with their laser emission range extended toward the blue by 20–30 nm. The quinolone dyes are used when laser emission in the 400–430 nm region is required. Azaquinolone derivatives, such as 7-dimethylamino-1-methyl-4-metoxy-8-azaquinolone-2 (LD 390), exhibit laser action below 400 nm.

Oxadiazole Derivatives

These are compounds which incorporate an axodiazole ring with aryl radical substituents and which lase in the 330–450 nm spectral region. Dye 2-(4-biphenyl)-5-(4-*t*-butylphenyl)-1,3,4-oxadiazole (PBD), with laser emission in the 355–390 nm region, belongs to this family.

Solid-State Laser Dye Matrices

Although some first attempts to incorporate dye molecules into solid matrices were already made in the early days of development of dye lasers, it was not until the late 1980s that host materials with the required properties of optical quality and high damage threshold to laser radiation began to be developed. In subsequent years, the synthesis of new high-performance dyes and the implementation of new ways of incorporating the organic molecules into the solid matrix resulted in significant advances towards the development of practical tunable solid-state dye lasers. A recent detailed review of the work done in this field has been given by Costela *et al.*

Inorganic glasses, transparent polymers, and inorganic–organic hybrid matrices have been successfully used as host matrices for laser dyes. Inorganic glasses offer good thermal and optical properties but present the difficulty that the high melting temperature employed in the traditional process of glass making would destroy the organic dye molecules. It was not until the development of the low-temperature sol-gel technique for the synthesis of glasses that this limitation was overcome. The sol-gel process, based on inorganic polymerization reactions performed at about room temperature and starting with metallo-organic compounds such as alkoxides or salts, as reported by Brinker and Scherrer, in 1989, provides a safe route for the preparation of rigid, transparent, inorganic matrix materials incorporating laser dyes. Disadvantages of these materials are the possible presence of impurities embedded in the matrix or the occurrence of interactions between the dispersed dye molecules and the inorganic structure (hydrogen bonds, van der Waals forces) which, in certain cases, can have a deleterious effect on the lasing characteristics of the material.

The use of matrices based on polymeric materials to incorporate organic dyes offers some technical and economical advantages. Transparent polymers exhibit high optical homogeneity, which is extremely important for narrow-linewidth oscillators, as explained by Duarte, in 1994, good chemical compatibility with organic dyes, and allow control over structure and chemical composition making possible the modification, in a controlled way, of relevant properties of these materials such as polarity, free volume, molecular weight, or viscoelasticity. Furthermore, the polymeric materials are amenable to inexpensive fabrication techniques, which facilitate miniaturization and design of integrated optical systems. The dye molecules can be either dissolved in the polymer or linked covalently to the polymeric chains.

In the early investigations on the use of solid polymer matrices for laser dyes, the main problem to be solved was that of the low resistance to laser radiation exhibited by the then existing materials. In the late 1980s and early 1990s new modified polymeric organic materials began to be developed with a laser-radiation-damage threshold comparable to that of inorganic glasses and crystals as reported by Gromov and colleagues, in 1985, and Dyumaev and colleagues, in 1992. This, together with the above-mentioned advantages, made the use of polymers in solid-state dye lasers both attractive and competitive.

An approach that intends to bring about materials in which the advantages of both inorganic glasses and polymers are preserved but the difficulties are avoided is that of using inorganic–organic hybrid matrices. These are silicate-based materials, with an inorganic Si–O–Si backbone, prepared from organosilane precursors by sol-gel processing in combination with organic cross-linking of polymerizable monomers as reported by Novak, in 1993, Sanchez and Ribot, in 1994, and Schubert, in 1995. In one procedure, laser dyes are mixed with organic monomers, which are then incorporated into the porous structure of a sol-gel inorganic matrix by immersing the bulk in the solution containing monomer and catalyst or photoinitiator as indicated by Reisfeld and Jorgensen, in 1991, and Bosch and colleagues, in 1996. Alternatively, hybrids can be obtained from organically modified silicon alkoxides, producing the so-called ORMOCERS (organically modified ceramics) or ORMOSILS (organically modified silanes) as reported by Reisfeld and Jorgensen, in 1991. An inconvenient aspect of these materials is the appearance of optical inhomogeneities in the medium due to the difference in the refractive index between the organic and inorganic phases as well as to the difference in density between monomer and polymer which causes stresses and the formation of optical defects. As a result, spatial inhomogeneities appear in the laser beam, thereby decreasing its quality as discussed by Duarte, in 1994.

Organic Hosts

The first polymeric materials with improved resistance to damage by laser radiation were obtained by Russian workers, who incorporated rhodamine dyes into modified poly(methyl methacrylate) (MPMMA), obtained by doping PMMA with low-molecular weight additives. Some years later, in 1995, Maslyukov and colleagues demonstrated lasing efficiencies in the range 40–60% with matrices of MPMMA doped with rhodamine dyes pumped longitudinally at 532 nm, with a useful lifetime (measured as a 50% efficiency drop) of 15 000 pulses at a prf of 3.33 Hz.

In 1995, Costela and colleagues used an approach based on adjusting the viscoelastic properties of the material by modifying the internal plasticization of the polymeric medium by copolymerization with appropriate monomers. Using dye Rh6G dissolved in a copolymer of 2-hydroxyethyl methacrylate (HEMA) and methyl methacrylate (MMA) and transversal pumping at 337 nm, they demonstrated laser action with an efficiency of 21% and useful lifetime of 4500 pulses (20 GJ/mol in terms of total input energy per mole of dye molecule when the output energy is down to 50% of its initial value). The useful lifetime increased to 12 000 pulses when the Rh6G chromophore was linked covalently to the polymeric chains as reported by Costela and colleagues, in 1996.

Comparative studies on the laser performance of Rh6G incorporated either in copolymers of HEMA and MMA or in MPMMA were carried out by Giffin and colleagues, in 1999, demonstrating higher efficiency of the MPMMA materials but superior normalized photostability (up to 240 GJ/mol) of the copolymer formulation.

Laser conversion efficiencies in the range 60–70% for longitudinal pumping at 532 nm were reported by Ahmad and colleagues, in 1999, with PM567 dissolved in PMMA modified with 1,4-diazobicyclo[2,2,2] octane (DABCO) or perylene additives. When using DABCO as an additive, a useful lifetime of up to 550 000 pulses, corresponding to a normalized photostability of 270 GJ/mol, was demonstrated at a 2 Hz prf.

Much lower efficiencies and photostabilities have been obtained with dyes emitting in the blue-green spectral region. Costela and colleagues, in 1996, performed some detailed studies on dyes coumarin 540A and coumarin 503 incorporated into methacrylate homopolymers and copolymers, and demonstrated efficiencies of at most 19% with useful lifetimes of up to 1200 pulses, at a 2 Hz prf, for transversal pumping at 337 nm. In 2000, Somasundaram and Ramalingam obtained useful lifetimes of 5240 and 1120 pulses, respectively, for coumarin 1 and coumarin 490 dyes incorporated into PMMA rods modified with ethyl alcohol, under transversal pumping at 337 nm and a prf of 1 Hz.

A detailed study of photo-physical parameters of R6G-doped HEMA-MMA gain media was performed by Holzer and colleagues, in 2000. Among the parameters determined by these authors are quantum yields, lifetimes, absorption cross-sections, and emission cross-sections. A similar study on the photophysical parameters of PM567 and two PM567 analogs incorporated into solid matrices of different acrylic copolymers and in corresponding mimetic liquid solutions was performed by Bergmann and colleagues, in 2001.

Preliminary experiments with rhodamine 6G-doped MPMMA at high prfs were conducted using a frequency-doubled Nd:YAG laser at 5 kHz by Duarte and colleagues, in 1996. In those experiments, involving longitudinal excitation, the pump laser was allowed to fuse a region at the incidence window of the gain medium so that a lens was formed. This lens and the exit window of the gain medium comprised a

short unstable resonator that yielded broadband emission. In 2001, Costela and colleagues demonstrated lasing in rhodamine 6G- and pyrromethene-doped polymer matrices under copper-vapor-laser excitation at prfs in the 1.0–6.2 kHz range. In these experiments, the dye-doped solid-state matrix was rotated at 1200 rpm. Average powers of 290 mW were obtained at a prf of 1 kHz and a conversion efficiency of 37%. For a short period of time average powers of up to 1 W were recorded at a prf of 6.2 kHz. In subsequent experiments by Abedin and colleagues the rpf was increased to 10 kHz by using as pump laser a diode-pumped, *Q*-switched, frequency doubled, solid state Nd:YLF laser. Initial average powers of 560 mW and 430 mW were obtained for R6G-doped and PM567-doped polymer matrices, respectively. Lasing efficiency was 16% for R6G and the useful lifetime was 6.6 min (or about 4.0 million shots).

Inorganic and Hybrid Hosts

In 1990, Knobbe and colleagues and McKiernan and colleagues, from the University of California at Los Angeles, incorporated different organic dyes, via the sol-gel techniques, into silicate (SiO_2), aluminosilicate (Al_2O_3–SiO_2), and ORMOSIL host matrices, and investigated the lasing performance of the resulting materials under transversal pumping. Lasing efficiencies of 25% and useful lifetimes of 2700 pulses at 1 Hz repetition rate were obtained from Rh6G incorporated into aluminosilicate gel and ORMOSIL matrices, respectively. When the dye in the ORMOSIL matrices was coumarin 153, the useful lifetime was still 2250 pulses. Further studies by Altman and colleagues, in 1991, demonstrated useful lifetimes of 11 000 pulses at a 30 Hz prf when rhodamine 6G perchlorate was incorporated into ORMOSIL matrices. The useful lifetime increased to 14 500 pulses when the dye used was rhodamine B.

When incorporated into xerogel matrices, pyrromethene dyes lase with efficiencies as high as the highest obtained in organic hosts but with much higher photostability: a useful lifetime of 500 000 pulses at 20 Hz repetition rate was obtained by Faloss and colleagues, in 1997, with PM597 using a pump energy of 1 mJ. The efficiency of PM597 dropped to 40% but its useful lifetime increased to over 1 000 000 pulses when oxygen-free samples were prepared, in agreement with previous results that documented the strong dependence of laser parameters of pyrromethene dyes on the presence of oxygen. A similar effect was observed with perylene dyes: deoxygenated xerogel samples of Perylene Red exhibited a useful lifetime of 250 000 pulses at a prf of 5 Hz and 1 mJ pump energy, to be compared with a useful lifetime of only 10 000 pulses obtained with samples prepared in normal conditions. Perylene Orange incorporated into polycom glass and pumped longitudinally at 532 nm gave an efficiency of 72% and a useful lifetime of 40 000 pulses at a prf of 1 Hz as reported by Rhan and King, in 1998.

A direct comparison study of laser performance of Rh6G, PM567, Perylene Red, and Perylene Orange in organic, inorganic, and hybrid hosts was carried out by Rahn and King in 1995. They found that the nonpolar perylene dyes had better performance in partially organic hosts, whereas the ionic rhodamine and pyrromethene dyes performed best in the inorganic sol-gel glass host. The most promising combinations of dye and host for efficiency and photostability were found to be Perylene Orange in polycom glass and Rh6G in sol-gel glass.

An all-solid-state optical configuration, where the pump was a laser diode array side-pumped, Q-switched, frequency-doubled, Nd:YAG slab laser, was demonstrated by Hu and colleagues, in 1998, with dye DCM incorporated into ORMOSIL matrices. A lasing efficiency of 18% at the wavelength of 621 nm was obtained, and 27 000 pulses were needed for the output energy to decrease by 10% of its initial value at 30 Hz repetition rate. After 50 000 pulses the output energy decreased by 90% but it could be recovered after waiting for a few minutes without pumping.

Violet and ultraviolet laser dyes have also been incorporated into sol-gel silica matrices and their lasing properties evaluated under transversal pumping by a number of authors. Although reasonable efficiencies have been obtained with some of these laser dyes, the useful lifetimes are well below 1000 pulses.

More recently, Costela and colleagues have demonstrated improved conversion efficiencies in dye-doped inorganic–organic matrices using pyrromethene 567 dye. The solid matrix in this case is poly-trimethylsilyl-methacrylate cross-linked with ethylene glycol and copolymerized with methyl methacrylate. Also, Duarte and James have reported on dye-doped polymer-nanoparticle laser media where the polymer is PMMA and the nanoparticles are made of silica. These authors report TEM_{00} laser beam emission and improved dn/dT coefficients of the gain media that results in reduced $\Delta\theta$.

Dye Laser Applications

Dye lasers generated a renaissance in diverse applied fields such as isotope separation, medicine,

photochemistry, remote sensing, and spectroscopy. Dye lasers have also been used in many experiments that have advanced the frontiers of fundamental physics.

Central to most applications of dye lasers is their capability of providing coherent narrow-linewidth radiation that can be tuned continuously from the near ultraviolet, throughout the visible, to the near infrared. These unique coherent and spectral properties have made dye lasers particularly useful in the field of high-resolution atomic and molecular spectroscopy. Dye lasers have been successfully and extensively used in absorption and fluorescence spectroscopy, Raman spectroscopy, selective excitations, and nonlinear spectroscopic techniques. Well documented is their use in photochemistry, that is, the chemistry of optically excited atoms and molecules, and in biology, studying biochemical reaction kinetics of biological molecules. By using nonlinear optical techniques, such as harmonic and sum frequency and difference frequency generation, the properties of dye lasers can be extended to the ultraviolet.

Medical applications of dye lasers include cancer photodynamic therapy, treatment of vascular lesions, and lithotripsy as described by Goldman in 1990.

The ability of dye lasers to provide tunable sub-GHz linewidths in the orange-red portion of the spectrum, at kW average powers, made them particularly suited to atomic vapor laser isotope separation of uranium as reported by Bass and colleagues, in 1992, Singh and colleagues, in 1994, and Nemoto and colleagues, in 1995. In this particular case the isotopic shift between the two uranium isotopes is a few GHz. Using dye lasers yielding $\Delta\nu \leq 1\,\text{GHz}$ the ^{235}U can be selectively excited in a multistep process by tuning the dye laser(s) to specific wavelengths, in the orange-red spectral region, compatible with a transition sequence leading to photoionization. Also, high-prf narrow-linewidth dye lasers, as described by Duarte and Piper, in 1984, are ideally suited for guide star applications in astronomy. This particular application requires a high-average power diffraction-limited laser beam at $\lambda \approx 589\,\text{nm}$ to propagate for some 95 km, above the surface of the Earth, and illuminate a layer of sodium atoms. Ground detection of the fluorescence from the sodium atoms allows measurements on the turbulence of the atmosphere. These measurements are then used to control the adaptive optics of the telescope thus compensating for the atmospheric distortions.

A range of industrial applications is also amenable to the wavelength agility and high-average power output characteristics of dye lasers. Furthermore, tunability coupled with narrow-linewidth emission, at large pulsed energies, make dye lasers useful in military applications such as directed energy and damage of optical sensors.

The Future of Dye Lasers

Predicting the future is rather risky. In the early 1990s, articles and numerous advertisements in commercial laser magazines predicted the demise and even the oblivion of the dye laser in a few years. It did not turn out this way. The high power and wavelength agility available from dye lasers have ensured their continued use as scientific and research tools in physics, chemistry, and medicine. In addition, the advent of narrow-linewidth solid-state dye lasers has sustained a healthy level of research and development in the field. Today, some 20 laboratories around the world are engaged in this activity.

The future of solid-state dye lasers depends largely on synthetic and manufacturing improvements. There is a need for strict control of the conditions of fabrication and a careful purification of all the compounds involved. In the case of polymers, in particular, stringent control of thermal conditions during the polymerization step is obligatory to ensure that adequate optical uniformity of the polymer matrix is achieved, and that intrinsic anisotropy developed during polymerization is minimized. From an engineering perspective what are needed are dye-doped solid-state media with improved photostability characteristics and better thermal properties. By better thermal properties is meant better dn/dT factors. To a certain degree progress in this area has already been reported. As in the case of liquid dye lasers, the lifetime of the material can be improved by using pump lasers at compatible wavelengths. In this regard, direct diode laser pumping of dye-doped solid-state matrices should lead to very compact, long-lifetime, tunable lasers emitting throughout the visible spectrum. An example of such a device would be a green laser-diode-pumped rhodamine-6G doped MPMMA tunable laser configured in a multiple-prism grating architecture.

As far as liquid lasers are concerned, there is a need for efficient water-soluble dyes. Successful developments in this area were published in the literature with coumarin-analog dyes by Chen and colleagues, in 1988, and some encouraging results with rhodamine dyes have been reported by Ray and colleagues, in 2002. Efficient water-soluble dyes could lead to significant improvements and simplifications in high-power tunable lasers.

See also

Coherent Control: Experimental. **Nonlinear Optics, Applications:** Pulse Compression via Nonlinear Optics. **Quantum Optics:** Entanglement and Quantum Information. **Relativistic Nonlinear Optics**.

Further Reading

Costela A, García-Moreno I and Sastre R (2001) In: Nalwa HS (ed.) *Handbook of Advanced Electronic and Photonic Materials and Devices*, vol. 7, pp. 161–208. San Diego, CA: Academic Press.

Diels J-C and Rudolph W (1996) *Ultrashort Laser Pulse Phenomena*. New York: Academic Press.

Demtröder W (1995) *Laser Spectroscopy*, 2nd edn. Berlin: Springer-Verlag.

Duarte FJ (ed.) (1991) *High Power Dye Lasers*. Berlin: Springer-Verlag.

Duarte FJ (ed.) (1995) *Tunable Lasers Handbook*. New York: Academic Press.

Duarte FJ (ed.) (1995) *Tunable Laser Applications*. New York: Marcel Dekker.

Duarte FJ (2003) *Tunable Laser Optics*. New York: Elsevier Academic.

Duarte FJ and Hillman LW (eds) (1990) *Dye Laser Principles*. New York: Academic Press.

Maeda M (1984) *Laser Dyes*. New York: Academic Press.

Schäfer FP (ed.) (1990) *Dye Lasers*, 3rd edn. Berlin: Springer-Verlag.

Radziemski LJ, Solarz RW and Paisner JA (eds) (1987) *Laser Spectroscopy and its Applications*. New York: Marcel Dekker.

Weber MJ (2001) *Handbook of Lasers*. New York: CRC.

Edge Emitters

J J Coleman, University of Illinois, Urbana, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The semiconductor edge emitting diode laser is a critical component in a wide variety of applications, including fiber optics telecommunications, optical data storage, and optical remote sensing. In this section, we describe the basic structures of edge emitting diode lasers and the physical mechanisms for converting electrical current into light. Laser waveguides and cavity resonators are also outlined. The power, efficiency, gain, loss, and threshold characteristics of laser diodes are presented along with the effects of temperature and modulation. Quantum well lasers are outlined and a description of grating coupled lasers is provided.

Like all forms of laser, the edge emitting diode laser is an oscillator and has three principal components; a mechanism for converting energy into light, a medium that has positive optical gain, and a mechanism for obtaining optical feedback. The basic edge emitting laser diode is shown schematically in Figure 1. Current flow is vertical through a pn junction and light is emitted from the ends. The dimensions of the laser in Figure 1 are distorted, in proportion to typical real laser diodes, to reveal more detail. In practical laser diodes, the width of the stripe contact is actually much smaller than shown and the thickness is smaller still. Typical dimensions are (stripe width × thickness × length) $2 \times 100 \times 1000\ \mu m$. Light emission is generated only within a few micrometers of the surface, which means that 99.98% of the volume of this small device is inactive.

Energy conversion in diode lasers is provided by current flowing through a forward-biased pn junction. At the electrical junction, there is a high density of injected electrons and holes which can recombine in a direct energy gap semiconductor material to give an emitted photon. Charge neutrality requires equal numbers of injected electrons and holes. Figure 2 shows a simplified energy versus momentum (E–k) diagrams for (a) direct, and (b) indirect semiconductors. In a direct energy gap material, such as GaAs, the energy minima have the same momentum vector, so recombination of an electron in the conduction band and a hole in the valence band takes place directly. In an indirect material, such as silicon, momentum conservation requires the participation of a third particle – a phonon. This third-order process is far less efficient than direct recombination and, thus far, too inefficient to support laser action.

Optical gain in direct energy gap materials is obtained when population inversion is reached.

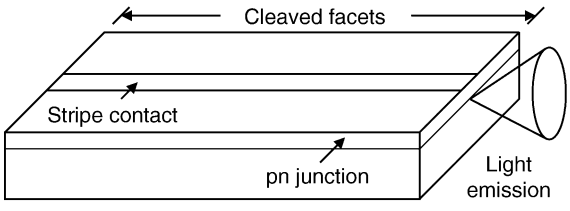


Figure 1 Schematic drawing of an edge emitting laser diode.

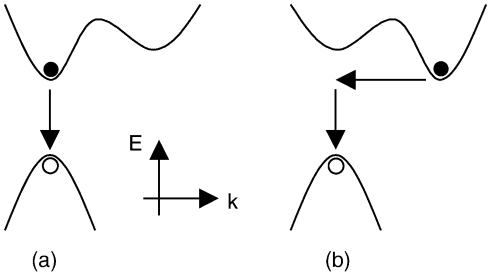


Figure 2 Energy versus momentum for (a) direct, and (b) indirect semiconductors.

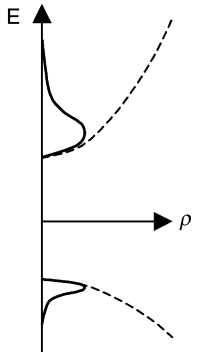


Figure 3 Density of states versus energy for conduction and valence bands.

Population inversion is the condition where the probability for stimulated emission exceeds that of absorption. The density of excited electrons n, in the conduction band is given by

$$n = \int_{E_c}^{\infty} \rho_n(E) f_n(E) \mathrm{d}E \ \mathrm{cm}^{-3} \qquad [1]$$

where $\rho_n(E)$ is the conduction band density of states function and $f_n(E)$ is the Fermi occupancy function. A similar equation can be written for holes in the valence band. Figure 3 shows the density of states ρ versus energy as a dashed line for conduction and valence bands. The solid lines in Figure 3 are the $\rho(E)f(E)$ products, and the areas under the solid curves are the carrier densities n and p.

The Fermi functions are occupation probabilities based on the quasi-Fermi levels E_{Fn} and E_{Fp}, which are, in turn, based on the injected carrier densities, δn and δp. Population inversion implies that the difference between the quasi-Fermi levels must exceed the emission energy which, for semiconductors, must be greater than the bandgap energy:

$$E_{Fn} - E_{Fp} \geq \hbar\omega \sim E_g \qquad [2]$$

Charge neutrality requires that $\delta n = \delta p$. Transparency is the point where these two conditions are just met and any additional injected current results in optical gain. The transparency current density J_o, is given by

$$J_o = \frac{q n_o L_z}{\tau_{spon}} \qquad [3]$$

where n_o is the transparency carrier density ($\delta n = \delta p = n_o$ at transparency), L_z is the thickness of the optically active layer, and τ_{spon} is the spontaneous carrier lifetime in the material (typically 5 nsec).

A resonant cavity for optical feedback is obtained simply by taking advantage of the natural crystal formation in single crystal semiconductors. Most III–V compound semiconductors form naturally into a zincblende lattice structure. With the appropriate choice of crystal planes and surfaces, this lattice structure can be cleaved such that nearly perfect plane–parallel facets can be formed at arbitrary distances from each other. Since the refractive index of these materials is much larger (~3.5) than that of air, there is internal optical reflection for normal incidence of approximately 30%. This type of optical cavity is called a Fabry–Perot resonator.

Effective lasers of all forms make use of optical waveguides to optimize the overlap of the optical field with material gain and cavity resonance. The optical waveguide in an edge emitting laser is best described by considering it in terms of a transverse waveguide component, defined by the epitaxial growth of multiple thin heterostructure layers, and a lateral waveguide component, defined by conventional semiconductor device processing methods. One of the simplest, and yet most common, heterostructure designs is shown in Figure 4. This five-layer separate confinement heterostructure (SCH) offers excellent carrier confinement in the active layer as well as a suitable transverse waveguide.

Clever choice of different materials for each layer results in the energy band structure, index of refraction

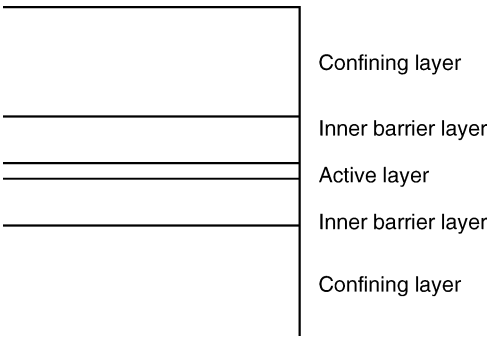


Figure 4 Schematic cross-section of a typical five-layer separate confinement heterostructure (SCH) laser.

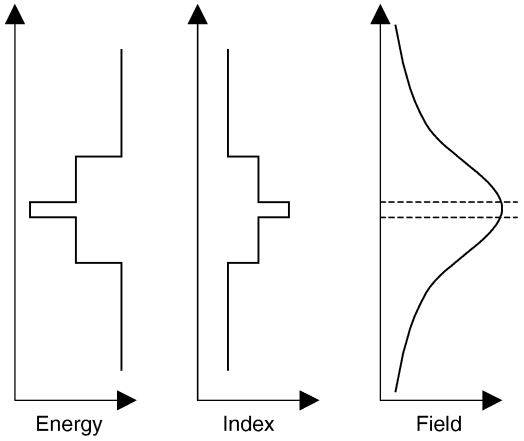


Figure 5 Energy band structure, index of refraction profile, and optical field mode profile of the SCH laser of **Figure 4**.

profile, and optical field mode profile shown in Figure 5. The outer confining layers are the thickest ($\sim 1\ \mu m$) and have the largest energy bandgap and lowest refractive index. The inner barrier layers are thinner ($\sim 0.1\ \mu m$), have intermediate bandgap energy and index of refraction, and serve both an optical role in defining the optical waveguide and an electronic role in confining carriers to the active layer. The active layer is the narrowest bandgap, highest index material, and is the only layer in the structure designed to provide optical gain. It may be a single layer or, as in the case of a multiple quantum well laser, a combination of layers. The bandgap energy, E, of this active layer plays a key role in determining the emission wavelength, λ, of the laser:

$$E = \frac{hc}{\lambda} \qquad [4]$$

where h is Planck's constant and c is the speed of light.

The bandgap energy of the SCH structure is shown in Figure 5. The lowest energy active layer collects injected electrons and holes and the carrier confinement provided by the inner barrier layers allows the high density of carriers necessary for population inversion and gain. The layers in the structure also have the refractive index profile shown in Figure 5 which forms a five-layer slab dielectric waveguide. With appropriate values for thicknesses and indices of refraction, this structure can be a very strong fundamental mode waveguide with the field profile shown. A key parameter of edge emitting diode lasers is Γ, the optical confinement factor. This parameter is the areal overlap of the optical field with the active layer, shown as dashed lines in the figure and can be as little as a few percent, even in very high performance lasers.

Lateral waveguiding in a diode laser can be accomplished in a variety of ways. A common example of an index guided laser is the ridge waveguide laser shown in Figure 6. After the appropriate transverse waveguide heterostructure sample is grown, conventional semiconductor processing methods are used to form parallel etched stripes from the surface near, but not through, the active layer. An oxide mask is patterned such that electrical contact is formed only to the center (core) region between the etched stripes, the core stripe being only a few microns wide. The lateral waveguide that results arises from the average index of refraction (effective index) in the etched regions outside the core being smaller than that of the core. Figure 7 shows the asymmetric near field emission profile of the ridge waveguide laser and cross-sectional profiles of the laser emission and effective index of refraction.

The optical spectrum of an edge emitting diode laser below and above threshold is the superposition of the material gain of the active layer and the resonances provided by the Fabry–Perot resonator. The spectral variation of material gain with injected carrier density (drive current) is shown in Figure 8. As the drive is increased, the intensity and energy of peak gain both increase.

The Fabry–Perot resonator has resonances every half wavelength, given by

$$L = \frac{m}{2}\frac{\lambda}{n} \qquad [5]$$

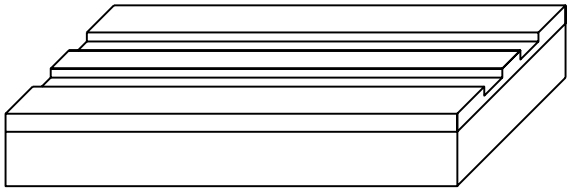


Figure 6 Schematic diagram of a real index guided ridge waveguide diode laser.

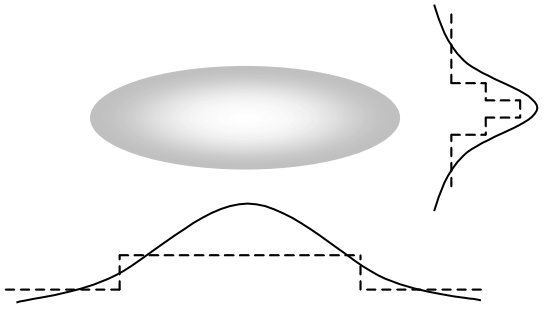


Figure 7 Near field emission pattern, cross-sectional emission profiles (solid lines), and refractive index profiles (dashed lines) from an index guided diode laser.

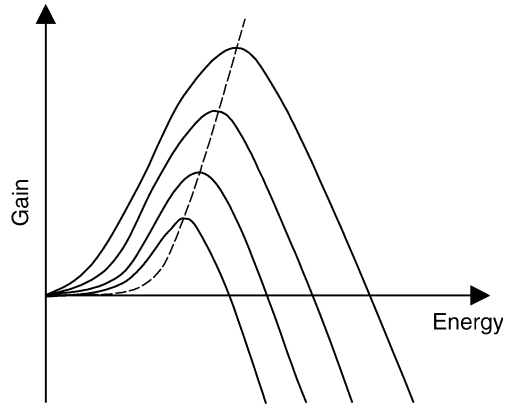


Figure 8 Gain versus emission energy for four increasing values of carrier (current) density. Peak gain is shown as a dashed line.

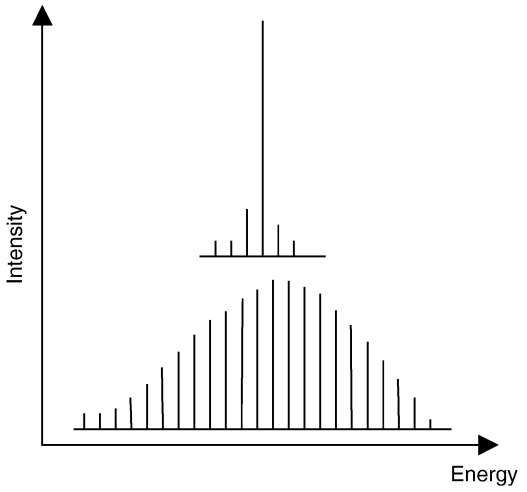


Figure 9 Emission spectra from a Fabry–Perot edge emitting laser structure at, and above, laser threshold.

where L is the cavity length, λ/n is the wavelength in the medium (n is the refractive index), and m is an integer. For normal length edge emitter cavities, these cavity modes are spaced only 1 or 2 Å apart. The result is optical spectra, near laser threshold and above, that look like the spectra of Figure 9. Many cavity modes resonate up to threshold, while above threshold, the superlinear increase in emission intensity with drive current tends to greatly favor one or two of the Fabry–Perot longitudinal cavity modes.

Laser threshold can be determined from a relatively simple analysis of the round trip gain and losses in the Fabry–Perot resonator. If the initial intensity is I_o, after one round trip the intensity will be given by

$$I = I_o R_1 R_2 e^{(g-\alpha_i)2L} \qquad [6]$$

where R_1 and R_2 are the reflectivities of the two facets, g is the optical gain, α_i are losses associated with optical absorption (typically 3–15 cm^{-1}), and L is the cavity length. At laser threshold, $g = g_{th}$ and $I = I_o$, i.e., the round trip gain exactly equals the losses. The result is

$$g_{th} = \alpha_i + \frac{1}{2L} \ln \frac{1}{R_1 R_2} \qquad [7]$$

where the second term represents the mirror losses α_m. Of course, mirror losses are desirable in the sense that they represent useable output power. Actually, the equation above does not take into account the incomplete overlap of the optical mode with the gain in the active layer. The overlap is defined by the confinement factor Γ, described above, and the equation must be modified to become

$$\Gamma g_{th} = \alpha_i + \frac{1}{2L} \ln \frac{1}{R_1 R_2} \qquad [8]$$

For quantum well lasers, the material peak gain, shown in Figure 8, as a function of drive current is given approximately by

$$g_t = \beta J_o \ln \frac{J}{J_o} \qquad [9]$$

where J_o is the transparency current density. The threshold current density J_{th}, which is the current density where the gain exactly equals all losses, is given by

$$J_{th} = \frac{J_o}{\eta_i} \exp \frac{\alpha + \frac{1}{2L} \ln \frac{1}{R_1 R_2}}{\Gamma \beta J_o} \qquad [10]$$

Note that an additional parameter η_i has appeared in this equation. Not all of the carriers injected by the drive current participate in optical processes; some are lost to other nonradiative processes. This is accounted for by including the internal quantum efficiency term η_i, which is typically at least 90%. The actual drive current, of course, must include the geometry of the device and is given by

$$I_{th} = wLJ_{th} \qquad [11]$$

where L is the cavity length, and w is the effective width of the active volume. The effective width is basically the stripe width of the core but also includes current spreading and carrier diffusion under the stripe.

The power–current characteristic (P–I) for a typical edge emitting laser diode is shown in Figure 10. As the drive current is increased

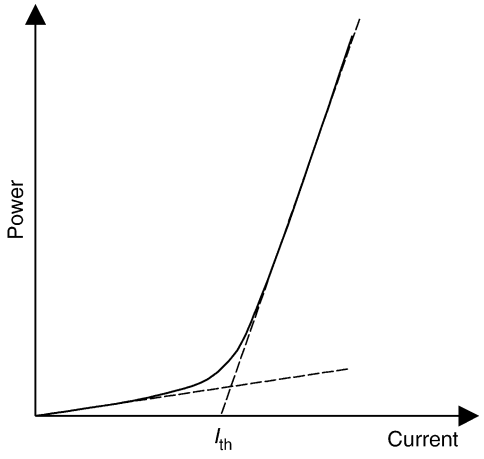


Figure 10 Power versus current (P–I) curve for a typical edge emitting laser diode.

from zero, the injected carrier densities increase and spontaneous recombination is observed. At some point the gain equals the internal absorption losses, and the material is transparent. Then, as the drive current is further increased, additional gain eventually equals the total losses, including mirror losses, and laser threshold is reached. Above threshold, nearly all additional injected carriers contribute to stimulated emission (laser output). The power generated internally is given by

$$P = wL(J - J_{th})\eta_i \frac{h\nu}{q} \qquad [12]$$

where wL is the area and $h\nu$ is the photon energy. Only a fraction of the power generated internally is extracted from the device

$$P_o = wL(J - J_{th})\eta_i \frac{h\nu}{q} \frac{\alpha_m}{\alpha_i + \alpha_m} \qquad [13]$$

Clearly a useful design goal is to minimize the internal optical loss α_i. The slope of the P–I curve of Figure 10 above threshold is described by an external differential quantum efficiency η_d, which is related to the internal quantum efficiency by

$$\eta_d = \eta_i \left[\frac{\alpha_m}{\alpha_i + \alpha_m} \right] \qquad [14]$$

Other common parameters used to characterize the efficiency of laser diodes include the power conversion efficiency η_p

$$\eta_p = \frac{P_o}{V_F I} \qquad [15]$$

and the wall plug efficiency η_w

$$\eta_w = \frac{P_o}{I} \qquad [16]$$

The power conversion efficiency η_p, recognizes that the forward bias voltage V_F is larger than the photon energy by an amount associated with the series resistance of the laser diode. The wall plug efficiency η_w, has the unusual units of WA^{-1} and is simply a shorthand term that relates the common input unit of current to the common output unit of power. In high performance low power applications, the ideal device has minimum threshold current since the current used to reach threshold contributes little to light emission. In high power applications, the threshold current becomes insignificant and the critical parameter is external quantum efficiency.

Temperature effects may be important for edge emitting lasers, depending on a particular application. Usually, concerns related to temperature arise under conditions of high temperature operation ($T > 50\,^{\circ}C$), where laser thresholds rise and efficiencies fall. It became common early in the development of laser diodes to define a characteristic temperature T_o for laser threshold, which is given by

$$I_{th} = I_{th_o} \exp\left(\frac{T}{T_o}\right) \qquad [17]$$

where a high T_o is desirable. Unfortunately, this simple expression does not describe the temperature dependence very well over a wide range of temperatures, so the appropriate temperature range must also be specified. For applications where the emission wavelength of the laser is important, the change in wavelength with temperature $d\lambda/dT$ may also be specified. For conventional Fabry–Perot cavity lasers, $d\lambda/dT$ is typically 3–5 $\text{Å}C^{-1}$.

In order for a diode laser to carry information, some form of modulation is required. One method for obtaining this is direct modulation of the drive current in a semiconductor laser. The transient and temporal behavior of lasers is governed by rate equations for carriers and photons which are necessarily coupled equations. The rate equation for carriers is given by

$$\frac{dn}{dt} = \frac{J}{qL_z} - \frac{n}{\tau_{sp}} - (c/n)\beta(n - n_o)\varphi(E) \qquad [18]$$

where J/qL_z is the supply, n/τ_{sp} is the spontaneous emission rate, and the third term is the stimulated emission rate $(c/n)\beta(n - n_o)\varphi(E)$ where β is the gain coefficient and $\varphi(E)$ is the photon density.

The rate equation for photons is given by

$$\frac{d\varphi}{dt} = (c/n)\beta(n - n_o)\varphi + \frac{\theta n}{\tau_{sp}} - \frac{\varphi}{\tau_p} \qquad [19]$$

where θ is the fraction of the spontaneous emission that couples into the mode (a small number), and τ_p is the photon lifetime (~1 psec) in the cavity. These rate equations can be solved for specific diode laser materials and structures to yield a modulation frequency response. A typical small signal frequency response, at two output power levels, is shown in Figure 11. The response is typically flat to frequencies above 1 GHz, rises to a peak value at some characteristic resonant frequency, and quickly rolls off. The characteristic small signal resonant frequency ω_r, is given by

$$\omega_r^2 = \frac{(c/n)\beta\bar{\varphi}}{\tau_p} \qquad [20]$$

where $\bar{\varphi}$ is the average photon density. Direct modulation of edge emitting laser diodes is limited by practicality to less than 10 GHz. This is in part because of the limits imposed by the resonant frequency and in part by chirp. Chirp is the frequency modulation that arises indirectly from direct current modulation. Modulation of the current modulates the carrier densities which, in turn, results in modulation of the quasi-Fermi levels. The separation of the quasi-Fermi levels is the emission energy (wavelength). Most higher-speed lasers make use of external modulation schemes.

The discussion thus far has addressed aspects of semiconductor diode edge emitting lasers that are common to all types of diode lasers irrespective of the choice of materials or the details of the active layer structure. The materials of choice for efficient laser devices include a wide variety of III–V binary compounds and ternary or quaternary alloys. The particular choice is based on such issues as emission wavelength, a suitable heterostructure lattice match, and availability of high-quality substrates. For example, common low-power lasers for CD players ($\lambda \sim 820$ nm) are likely to be made on a GaAs substrate using a GaAs–$Al_xGa_{1-x}As$ heterostructure. Lasers for fiberoptic telecommunications systems ($\lambda \sim 1.3$–1.5 μm) are likely to be made on an InP substrate using an InP–$In_xGa_{1-x}As_yP_{1-y}$ heterostructure. Red laser diodes ($\lambda \sim 650$ nm) are likely to be made on a GaAs substrate using a GaAs–$In_xGa_{1-x}As_yP_{1-y}$ heterostructure. Blue and ultraviolet lasers, a relatively new technology compared to the others, make use of $Al_xGa_{1-x}N$–GaN–$In_yGa_{1-y}N$ heterostructures formed on sapphire or silicon carbide substrates.

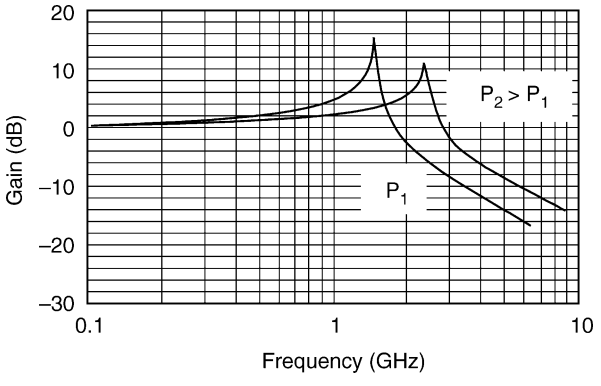


Figure 11 Direct current modulation frequency response for an edge emitting diode laser at two output power levels.

An important part of many diode laser structures is a strained layer. If a layer is thin enough, the strain arising from a modest amount of lattice mismatch can be accommodated elastically. In thicker layers, lattice mismatch results in an unacceptable number of dislocations that affect quantum efficiency, optical absorption losses, and, ultimately, long-term failure rates. Strained layer GaAs–$In_xGa_{1-x}As$ lasers are a critical component in rare-earth doped fiber amplifiers.

The structure of the active layer in edge emitting laser diodes was originally a simple double heterostructure configuration with a single active layer of 500–1000 Å in thickness. In the 1970s however, advances in the art of growing semiconductor heterostructure materials led to growth of high-quality quantum well active layers. These structures, having thicknesses that are comparable to the electron wavelength in a semiconductor (<200 Å), revolutionized semiconductor lasers, resulting in much lower threshold current densities, higher efficiencies, and a broader range of available emission wavelengths. The energy band diagram for a quantum well laser active region is shown in Figure 12.

This structure is a physical realization of the particle-in-a-box problem in elementary quantum mechanics. The quantum well yields quantum states in the conduction band at discrete energy levels, with odd and even electron wavefunctions, and breaks the degeneracy in the valence band, resulting in separate energy states for light holes and heavy holes. The primary transition for recombination is from the $n = 1$ electron state to the $h = 1$ heavy hole state and takes place at a higher energy than the bulk bandgap energy. In addition, the density of states for quantum wells becomes a step-like function and optical gain is enhanced. The advantages in practical edge

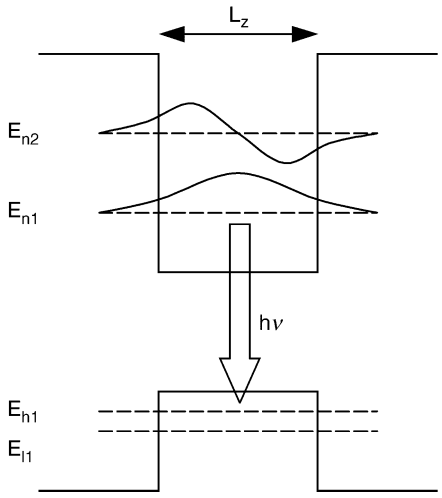


Figure 12 Energy band diagram for a quantum well heterostructure.

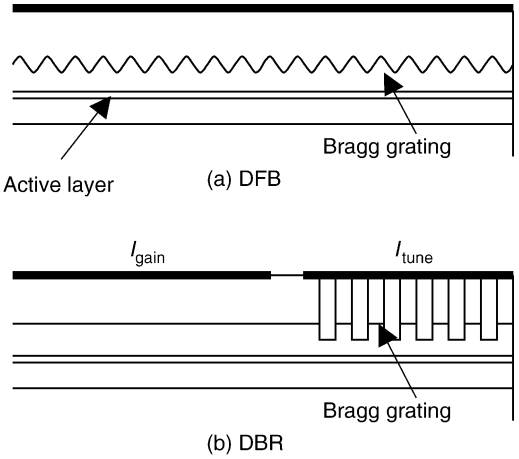


Figure 13 Schematic diagram of distributed feedback (DFB) and distributed Bragg reflector (DBR) edge emitting laser diodes.

emitting lasers that are the result of one-dimensional quantization are such that virtually all present commercial laser diodes are quantum well lasers. These kinds of advantages are driving development of laser diodes with additional degrees of quantization, including two dimensions (quantum wires) and three dimensions (quantum dots).

The Fabry–Perot cavity resonator described here is remarkably efficient and relatively easy to fabricate, which makes it the cavity of choice for many applications. There are many applications, however, where the requirements for linewidth of the laser emission or the temperature sensitivity of the emission wavelength make the Fabry–Perot resonator less desirable. An important laser technology that addresses both of these concerns involves the use of a wavelength selective grating as an integral part of the laser cavity. A notable example of a grating coupled laser is distributed feedback laser shown schematically in Figure 13a. This is similar to the SCH cross-section of Figure 4 with the addition of a Bragg grating above the active layer.

The period Λ, of the grating is chosen to fulfill the Bragg condition for mth-order coupling between forward- and backward-propagating waves, which is

$$\Lambda = \frac{m\lambda}{2n} \qquad [21]$$

where λ/n is the wavelength in the medium. The index step between the materials on either side of the grating is small and the amount of reflection is also small. Since the grating extends throughout the length of the cavity, however, the overall effective reflectivity can be large. Another variant is the distributed Bragg reflector (DBR) laser resonator, shown in Figure 13b. This DBR laser has a higher index contrast, deeply etched surface grating located at one or both ends of the otherwise conventional SCH laser heterostructure. One of the advantages of this structure is the opportunity to add a contact for tuning the Bragg grating by current injection.

Both the DBR and DFB laser structures are particularly well suited for telecommunications or other applications where a very narrow single laser emission line is required. In addition, the emission wavelength temperature dependence for this type of laser is much smaller, typically 0.5 $\text{Å}C^{-1}$.

See also

Semiconductor Physics: Quantum Wells and GaAs-based Structures.

Further Reading

Agrawal GP (ed.) (1995) *Semiconductor Lasers Past, Present, and Future*. Woodbury, NY: American Institute of Physics.

Agrawal GP and Dutta NK (1993) *Semiconductor Lasers*. New York: Van Nostrand Reinhold.

Bhattacharya P (1994) *Semiconductor Optoelectronic Devices*. Upper Saddle River, NJ: Prentice-Hall.

Botez D and Scifres DR (eds) (1994) *Diode Laser Arrays*. Cambridge, UK: Cambridge University Press.

Chuang SL (1995) *Physics of Optoelectronic Devices*. New York: John Wiley & Sons.

Coleman JJ (1992) *Selected Papers on Semiconductor Diode Lasers*. Bellingham, WA: SPIE Optical Engineering Press.
Einspruch NG and Frensley WR (eds) (1994) *Heterostructures and Quantum Devices*. San Diego, CA: Academic Press.
Iga K (1994) *Fundamentals of Laser Optics*. New York: Plenum Press.
Kapon E (1999) *Semiconductor Lasers I and II*. San Diego, CA: Academic Press.
Nakamura S and Fasol G (1997) *The Blue Laser Diode*. Berlin: Springer.
Verdeyen JT (1989) *Laser Electronics*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.
Zory PS Jr (ed.) (1993) *Quantum Well Lasers*. San Diego, CA: Academic Press.

Excimer Lasers

J J Ewing, Ewing Technology Associates, Inc., Bellevue, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

We present a summary of the fundamental operating principles of ultraviolet, excimer lasers. After a brief discussion of the economics and application motivation, the underlying physics and technology of these devices is described. Key issues and limitations in the scaling of these lasers are presented.

Background: Why Excimer Lasers?

Excimer lasers have become the most widely used source of moderate power pulsed ultraviolet sources in laser applications. The range of manufacturing applications is also large. Production of computer chips, using excimer lasers as the illumination source for lithography, has by far the largest commercial manufacturing impact. In the medical arena, excimer lasers are used extensively in what is becoming one of the world's most common surgical procedures, the reshaping of the lens to correct vision problems. Taken together, the annual production rate for these lasers is a relatively modest number compared to the production of semiconductor lasers. Current markets are in the range of \$0.4 billion per year (Figure 1). More importantly, the unique wavelength, power, and pulse energy properties of the excimer laser enable a systems and medical procedure market, based on the excimer laser, to exceed \$3 billion per year. The current average sales price for these lasers is in the range of \$250 000 per unit, driven primarily by the production costs and reliability requirements of the lasers for lithography for chip manufacture. Relative to semiconductor diode lasers, excimer lasers have always been, and will continue to be, more expensive per unit device by a factor of order 10^4. UV power and pulse energy, however, are considerably higher.

The fundamental properties of these lasers enable the photochemical and photophysical processes used in manufacturing and medicine. Short pulses of UV light can photo-ablate materials, being of use in medicine and micromachining. Short wavelengths enable photochemical processes. The ability to provide a narrow linewidth using appropriate resonators enables precise optical processes, such as lithography. We trace out some of the core underlying properties of these lasers that lead to the core utility. We also discuss the technological problems and

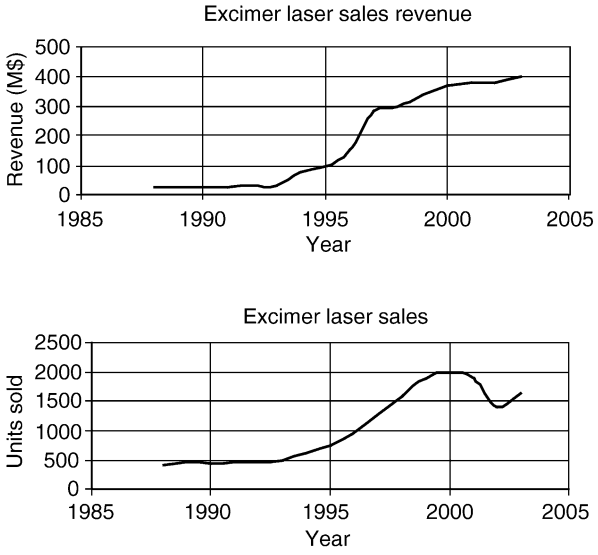


Figure 1 The sales of excimer lasers have always been measured in small numbers relative to other, less-expensive lasers. For many years since their commercial release in 1976 the sales were to R&D workers in chemistry, physics and ultimately biology and medicine. The market for these specialty but most useful R&D lasers was set to fit within a typical researcher's annual capital budget, i.e., less than \$100K. As applications and procedures were developed and certified or approved, sales and average unit sales price increased considerably. Reliance on cyclical markets like semiconductor fabrication led to significant variations in production rates and annual revenues as can be seen.

solutions that have evolved to make these lasers so useful.

The data in Figure 1 provide a current answer to the question of 'why excimer lasers?' At the dawn of the excimer era, the answer to this question was quite different. In the early 1970s, there were no powerful short wavelength lasers, although IR lasers were being scaled up to significant power levels. However, visible or UV lasers could in principle provide better propagation and focus to a very small spot size at large distances. Solid state lasers of the time offered very limited pulse repetition frequency and low efficiency. Diode pumping of solid state lasers, and diode arrays to excite such lasers, was a far removed development effort. As such, short-wavelength lasers were researched over a wide range of potential media and lasing wavelengths. The excimer concept was one of those candidates.

We provide, in Figure 2, a 'positioning' map showing the range of laser parameters that can be obtained with commercial or developmental excimer lasers. The map has coordinates of pulse energy and pulse rate, with diagonal lines expressing average power. We show where both development goals and current markets lay in a map. Current applications are shown by the boxes in the lower right-hand corner. Excimer lasers have been built with clear apertures in the range of a fraction of 0.1 to 10^4 cm^2, with single pulse energies covering a range of 10^7. Lasers with large apertures require electron beam excitation. Discharge lasers correspond to the more modest pulse energy used for current applications. In general, for energies lower than ~5 J, the excitation method of choice is a self-sustained discharge, clearly providing growth potential for future uses, if required.

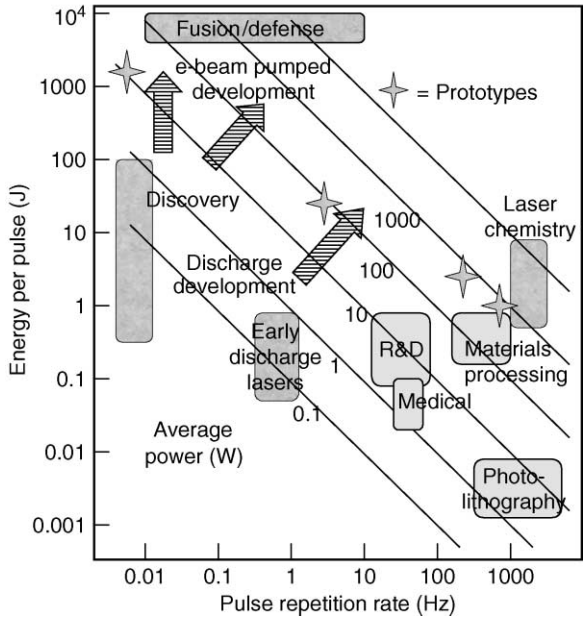


Figure 2 The typical energy and pulse rate for certain applications and technologies using excimer lasers. For energy under a few J, a discharge laser is used. The earliest excimer discharge lasers were derivatives of CO_2 TEA (transversely excited atmospheric pressure) and produced pulse energy in the range of 100 mJ per pulse. Pulse rates of order 30 Hz were all that the early pulsed power technology could provide. Early research focused on generating high energy. Current markets are at modest UV pulse energy and high pulse rates for lithography and low pulse rates for medical applications.

The applications envisaged in the early R&D days were in much higher energy uses, such as in laser weapons, laser fusion, and laser isotope separation. The perceived need for lasers for isotope separation, laser-induced chemistry, and blue-green laser-based communications from satellites, drove research in high pulse rate technology and reliability extension for discharge lasers. This research resulted in prototypes, with typical performance ranges as noted on the positioning map that well exceed current market requirements. However, the technology and underlying media physics developed in these early years made possible advances needed to serve the ultimate real market applications.

The very important application of UV lithography requires high pulse rates and high reliability. These two features differentiate the excimer laser used for lithography from the one used in the research laboratory. High pulse rates, over 2000 Hz in current practice, and the cost of stopping a computer chip production line for servicing, drive the reliability requirements toward 10^{10} shots per service interval. In contrast, the typical laboratory experiments are more often in the range of 100 Hz and, unlike a lithography production line, do not run 24 hours a day, 7 days a week. The other large market currently is in laser correction of myopia and other imperfections in the human eye. For this use the lasers are more akin to the commercial R&D style laser in terms of pulse rate and single pulse energy. Not that many shots are needed to ablate tissue to make the needed correction, and shot life is a straightforward requirement. However, the integration of these lasers into a certified and accepted medical procedure and an overall medical instrument drove the growth of this market.

The excimer concept, and the core technology used to excite these lasers, is applicable over a broad range of UV wavelengths, with utility at specific wavelength ranges from 351 nm in the near UV to 157 nm in the vacuum UV (VUV). There were many potential excimer candidates in the early R&D phase (Table 1). Some of these

Table 1 Key excimer wavelengths

Excimer emitter	*λ (nm)*	*Comment*
Ar_2	126	Very short emission wavelength, deep in the Vacuum UV; inefficient as laser due to absorption, see Xe_2.
Kr_2	146	Broad band emitter and early excimer laser with low laser efficiency. Very efficient converter of electric power into Vacuum UV light.
F_2	157	Molecule itself is not an excimer, but uses lower level dissociation; candidate for next generation lithography.
Xe_2	172	The first demonstrated excimer laser. Very efficient formation and emission but excited state absorption limits laser efficiency. Highly efficient as a lamp.
ArF	193	Workhorse for corneal surgery and lithography.
KrCl	222	Too weak compared to KrF and not as short a wavelength as ArF.
KrF	248	Best intrinsic laser efficiency; numerous early applications but found major market in lithography. Significant use in other materials processing.
XeI	254	Never a laser, but high formation efficiency and excellent for a lamp. Historically the first rare gas halide excimer whose emission was studied at high pressure.
XeBr	282	First rare gas halide to be shown as a laser; inefficient laser due to excited state absorption, but excellent fluorescent emitter; a choice for lamps.
Br_2	292	Another halogen molecule with excimer like transitions that has lased but had no practical use.
XeCl	308	Optimum medium for laser discharge excitation.
Hg_2	335	Hg vapor is very efficiently excited in discharges and forms excimers at high pressure. Subject of much early research, but as in the case of the rare gas excimers such as Xe_2 suffer from excited state absorption. Despite the high formation efficiency, this excimer was never made to lase.
I_2	342	Iodine UV molecular emission and lasing was first of the pure halogen excimer-like lasers demonstrated. This species served as kinetic prototype for the F_2 'honorary' excimer that has important use as a practical VUV source.
XeF	351, 353	This excimer laser transition terminates on a weakly bound lower level that dissociates rapidly, especially when heated. The focus for early defense-related laser development; as this wavelength propagates best of this class in the atmosphere.
XeF	480	This very broadband emission of XeF terminates on a different lower level than the UV band. Not as easily excited in a practical system, so has not had significant usage.
HgBr	502	A very efficient green laser that has many of the same kinetic features of the rare gas halide excimer lasers. But the need for moderately high temperature for the needed vapor pressure made this laser less than attractive for UV operation.
XeO	540	This is an excimer-like molecular transition on the side of the auroral lines of the O atom. The optical cross-section is quite low and as a result the laser never matured in practice.

candidates can be highly efficient, >50% relative to deposited energy, at converting electrical power into UV or visible emission, and have found a parallel utility as sources for lamps, though with differing power deposition rates and configurations. The key point in the table is that these lasers are powerful and useful sources at very short wavelengths. For lithography, the wavelength of interest has consistently shifted to shorter and shorter wavelengths as the printed feature size decreased. Though numerous candidates were pursued, as shown in Table 1, the most useful lasers are those using rare gas halide molecules. These lasers are augmented by dye laser and Raman shifting to provide a wealth of useful wavelengths in the visible and UV.

All excimer lasers utilize molecular dissociation to remove the lower laser level population. This lower level dissociation is typically from an unbound or very weakly bound ground-state molecule. However, halogen molecules also provide laser action on excimer-like transitions, that terminate on a molecular excited state that dissociates quickly. The vacuum UV transition in F_2 is the most practical method for making very short wavelength laser light at significant power. Historically, the rare gas dimer molecules, such as Xe_2, were the first to show excimer laser action, although self absorption, low gain, and poor optics in the VUV limited their utility. Other excimer-like species, such as metal halides, metal rare gas continua on the edge of metal atom resonance lines, and rare gas oxides were also studied. The key differentiators of rare gas halides from other excimer-like candidates are strong binding in the excited state, lack of self absorption, and relatively high optical cross-section for the laser transition.

Excimer Laser Fundamentals

Excimer lasers utilize lower-level dissociation to create and sustain a population inversion. For example, in the first excimer laser demonstrated, Xe_2, the lasing species is one which is comprised of two atoms that do not form a stable molecule in the ground state, as xenon dimers in the ground state are not a stable molecule. In the lowest energy state, that with neither of the atoms electronically excited, the interaction between the atoms in a collision is primarily one of repulsion, save for a very weak 'van der Waals' attraction at larger internuclear separations. This is shown in the potential energy diagram of Figure 3. If one of the atoms is excited, for example by an electric discharge, there can be a chemical binding in the electronically excited state of the molecule relative to the energy of one atom with an electron excited and one atom in the ground state. We use the shorthand notation * to denote an electronically excited atomic molecular species, e.g., Xe^*. The excited dimer, excimer for short, will recombine rapidly at high pressure from atoms into the Xe_2^* electronically excited molecule. Radiative lifetimes of the order 10 nsec to 1 μsec are typical before photon emission returns the molecule to the lower level. Collisional quenching often reduces the lifetime below the radiative rate. For Xe_2^* excimers, the emission is in the vacuum ultraviolet (VUV).

There are a number of variations on this theme, as noted in Table 1. Species, other than rare gas pairs, can exhibit broadband emission. The excited state binding energy can vary significantly, changing the fraction of excited states that will form molecules. The shape of the lower potential energy curve can vary as well. From a semantic point of view heteronuclear, diatomic molecules, such as XeO, are not excimers, as they are not made of two of the same atoms. However, the more formal name 'exciplex' did not stick in the laser world; excimer laser being preferred. Large binding energy is more efficient for capturing excited atoms into excited excimer-type molecules in the limited time they have before emission. Early measurements of the efficiency of converting electrical energy deposited into fluorescent radiation showed that up to 50% of the energy input could result in VUV light emission. Lasers were not this efficient due to absorption.

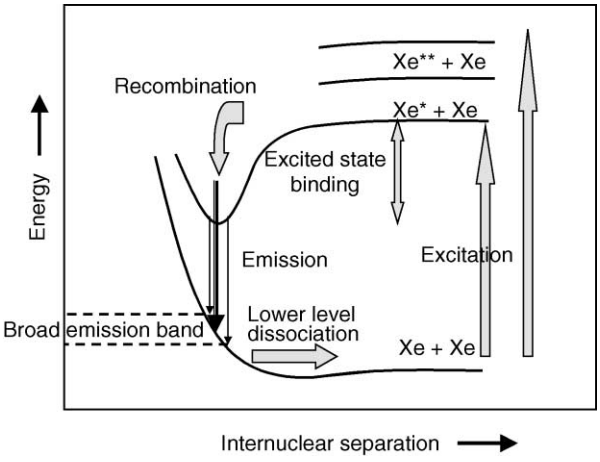


Figure 3 A schematic of the potential energy curves for an excimer species such as Xe_2. The excited molecule is bound relative to an excited atom and can radiate to a lower level that rapidly dissociates on psec timescales since the ground state is not bound, save for weak binding due to van der Waals forces. The emission band is intrinsically broad.

The diagram shown in Figure 3 is simplified because it does not show all of the excited states that exist for the excited molecule. There can be closely lying excited states that share the population and radiate at a lower rate, or perhaps absorb to higher levels, also not shown in Figure 3. Such excited state absorption may terminate on states derived from higher atomic excited states, noted as Xe^{**} in the figure, or yield photo-ionization, producing the diatomic ion-molecule plus an electron, such as:

$$Xe_2^* + h\nu \rightarrow Xe_2^+ + e \qquad [1]$$

Such excited state absorption limited the efficiency for the VUV rare gas excimers, even though they have very high formation and fluorescence efficiency.

Rare gas halide excimer lasers evolved from 'chemi-luminescence' chemical kinetics studies which were looking at the reactive quenching of rare gas metastable excited states in flowing after glow experiments. Broadband emission in reactions with halogen molecules was observed in these experiments, via reactions such as:

$$Xe^* + Cl_2 \rightarrow XeCl^* + Cl \qquad [2]$$

At low pressure the emissions from molecules, such as $XeCl^*$, are broad in the emissions bandwidth. Shortly after these initial observations, the laser community began examination of these species in high-pressure, electron-beam excited mixtures. The results differed remarkably from the low-pressure experiments and those for the earlier excimers such as Xe_2^*. The spectrum was shifted well away from the VUV emission. Moreover, the emission was much sharper at high pressure, though still a continuum with a bandwidth of the order of 4 nm. A basis for understanding is sketched in the potential energy curves of Figure 4. In this schematic we identify the rare gas atoms by Rg, excited rare gas atoms by

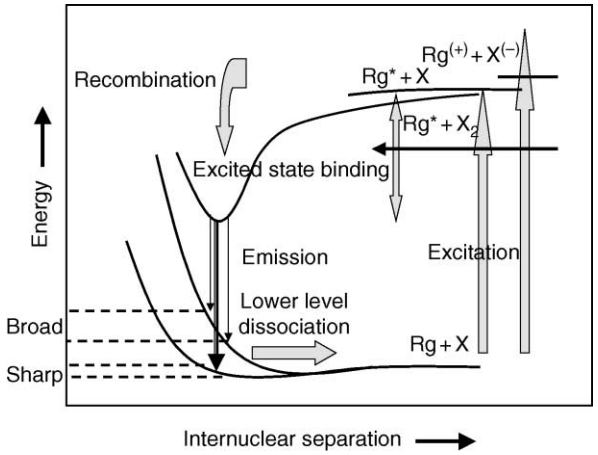


Figure 4 The potential energy curves for rare gas halides are sketched here. Both relatively sharp continuum emission is observed along with broad bands corresponding to different lower levels. The excited ion pair states consist of 3 distinct levels, two that are very close in energy and one that is shifted up from the upper laser level (the B state) by the spin orbit splitting of the rare gas ion.

Rg^*, and halogen atoms by X. Ions play a very important role in the binding and reaction chemistry, leading to excited laser molecules.

The large shift in wavelength from the VUV of rare gas dimer excimers, ~308 nm in $XeCl^*$ versus 172 nm in Xe_2^*, is due to the fact that the binding in the excited state of the rare gas halide is considerably stronger. The sharp and intense continuum at high pressure (> ~100 torr total pressure) is due to the fact that the 'sharp' transition terminates on a 'flat' portion of the lower-level potential energy curve. Indeed, in XeCl and XeF, the sharp band laser transition terminates on a slightly bound portion of the lower-level potential energy curve. Both the sharp emissions and broad bands are observed, due to the fact that some transitions from the excited levels terminate on a second, more repulsive potential energy curve.

An understanding of the spectroscopy and kinetic processes in rare gas plus halogen mixtures is based on the recognition that the excited state of a rare gas halide molecule is very similar to the ground state of the virtually isoelectronic alkali halide. The binding energy and mechanism of the excited state is very close to that of the ground state of the most similar alkali halide. Reactions of excited state rare gases are very similar to those of alkali atoms. For example, Xe^* and Cs are remarkably similar, from a chemistry and molecular physics point of view, as they have the same outer shell electron configuration and similar ionization potentials. The ionization potential of Xe^* is similar to that of cesium (Cs). Cs, and all the other alkali atoms, react rapidly with halogen molecules. Xenon metastable excited atoms react rapidly with halogen molecules. The alkali atoms all form ionically bonded ground states with halogens. The rare gas halides are effectively strongly bonded ion pairs, $Xe^{(+)}X^{(-)}$, that are very strongly bonded relative to the excited states that yield them. The difference between, for example, $XeCl^*$ and CsCl, is that $XeCl^*$ radiates in a few nanoseconds while CsCl is a stable ground-state molecule. The binding energy for these ionic bonded molecules is much greater than the more covalent type of bonding that is found in the rare gas excimer excited states. Thus XeI^*, which is only one electron different (in the I) than Xe_2^*, emits at 254 nm instead of 172 nm.

The first observed and most obvious connection to alkali atoms is in formation chemistry. Mostly, rare gas excited states react with halogen molecules rapidly. The rare gas halide laser analog reaction sequence is shown below, where some form of electric discharge excitation kicks the rare gas atom up to an excited state, so that it can react like an alkali atom, eqn [3]:

$$e_{fast} + Kr \rightarrow Kr^* + e_{slow} \qquad [3]$$

$$Kr^* + F_2 \rightarrow KrF^*(v, J)_{high} + F \qquad [4]$$

There are subtle and important cases where the neutral reaction, such as that shown in eqn [2] or [4], is not relevant as it is too slow compared to other processes. Note that the initial reaction does not yield the specific upper levels for the laser but states that are much higher up in the potential well of the excited rare gas halide excimer molecule. Further collisions with a buffer gas are needed to relax the states to the vibrational levels that are the upper levels for the laser transition. Such relaxation at high pressure can be fast, which leads the high-pressure spectrum to be much sharper than those first observed in flowing after-glow experiments.

The excited states of the rare gas halides are ion pairs bound together for their brief radiative lifetime. This opens up a totally different formation channel, indeed the one that often dominates the mechanism, ion–ion recombination. Long before excimer lasers were studied, it was well known that halogen molecules would react with 'cold' electrons (those with an effective temperature less than a few eV) in a process called attachment. The sequence leading to the upper laser level is shown below for the Kr/F_2 system, but the process is generically identical in other rare gas halide mixtures.

KrF^* formation kinetics via ion channel:

Ionization: $$e + Kr \rightarrow Kr^+ + 2e \quad [5]$$

Attachment: $$e + F_2 \rightarrow F^{(-)} + F \quad [6]$$

Ion–ion recombination:

$$Kr^+ + F^{(-)} + M \rightarrow KrF^*(v, J)_{high} + M \quad [7]$$

Relaxation:

$$KrF^*(v, J)_{high} + M \rightarrow KrF^*(v, J)_{low} + M \quad [8]$$

Finally, there is one other formation mechanism, displacement, that is unique to the rare gas halides. In this process, a heavier rare gas will displace a lighter rare gas ion in an excited state to form a lower-energy excited state:

$$ArF^* + Kr \rightarrow KrF^*(v, J)_{high} + Ar \quad [9]$$

In general, the most important excimer lasers tend to have all of these channels contributing to excited state formation for optimized mixtures.

These kinetic formation sequences express some important points regarding rare gas halide lasers. The source of the halogen atoms for the upper laser level disappears via both attachment reactions with electrons and by the reaction with excited states. The halogen atoms eventually recombine to make molecules, but at a rate too slow for sufficient continuous wave (cw) laser gain. For excimer-based lamps, however, laser gain is not a criterion and very efficient excimer lamps can be made. Another point to note is that forming the inversion requires transient species (and in some cases halogen fuels) that absorb the laser light. Net gain and extraction efficiency are a trade-off between pumping rate, an increase in which often forms more absorbers and absorption itself. Table 2 provides a listing of some of the key absorbers. Finally, the rare gas halide excited states can be rapidly quenched in a variety of processes. More often than not, the halogen molecule and the electrons in the discharge react with the excited states of the rare gas halide, removing the contributors to gain. We show, in Table 3, a sampling of some of the types of formation and quenching reactions with their kinetic rate constant parameters. Note that one category of reaction, three-body quenching, is unique to excimer lasers. Note also that the three-body recombination of ions has a rate coefficient that is effectively pressure dependent. At pressures above ~3 atm, the diffusion of ions toward each other is slowed and the effective recombination rate constant decreases.

Though the rare gas halide excited states can be made with near-unity quantum yield from the primary excitation, the intrinsic laser efficiency is not this high. The quantum efficiency is only ~25% (recall rare gas halides all emit at much longer wavelengths than rare gas excimers); there is inefficient extraction due to losses in the gas and windows. In practice there are also losses in coupling the power into the laser, and wasted excitation during the finite start-up time. The effect of losses is shown in Table 4 and Figure 5 where the pathways are charted and losses identified. In KrF, it takes ~20 eV to make a rare gas ion in an electron beam excited mixture,

Table 2 Absorbers in rare gas halide lasers

Species	*XeF*	*XeCl*	*KrF*	*ArF*
Laser λ (nm)	351, 353	308	248	193
F_2 absorption σ (cm^2)	8×10^{-21}	NA	1.5×10^{-20}	
Cl_2 absorption σ (cm^2)	NA	1.7×10^{-19}	NA	NA
HCl absorption σ (cm^2)	NA	Nil	NA	NA
$F^{(-)}$ absorption σ (cm^2)	2×10^{-18}	NA	5×10^{-18}	5×10^{-18}
$Cl^{(-)}$ absorption σ (cm^2)	NA	2×10^{-17}	NA	NA
$Rg_2^{(+)}$ diatomic ion absorption σ (cm^2)[a]	Ne: $\sim 10^{-18}$ Ar: $\sim 3 \times 10^{-17}$ Xe: $\sim 3 \times 10^{-17}$	Ne: $\sim 10^{-17}$ Ar: $\sim 4 \times 10^{-17}$ Xe: $\sim 3 \times 10^{-18}$	Ne: $\sim 2.5 \times 10^{-17}$ Ar: $\sim 2 \times 10^{-18}$ Kr $< 10^{-18}$	Ne: $\sim 10^{-18}$ Ar: NIL
Other transient absorbers	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$–10^{-17}	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$–10^{-17}	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$–10^{-17}	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$–10^{-17}
Other	Lower laser level absorbs unless heated	Weak lower level absorption $\sigma \sim 4 \times 10^{-16}$	Windows and dust can be an issue	Windows and dust can be an issue

[a]Note that the triatomic rare gas halides of the form Rg_2X will exhibit similar absorption as the diatomic rare gas ions as the triatomic rare gas halides of this form are essentially ion pairs of an $Rg_2^{(+)}$ ion and the halide ion.

Table 3 Typical formation and quenching reactions and rate coefficients

Formation	
Rare gas plus halogen molecule	$Kr^* + F_2 \rightarrow KrF^*(v, J)+F$ $k \sim 7\times10^{-10}$ cm^3/sec; Branching ratio to $KrF^* \sim 1$ $Ar^* + F_2 \rightarrow ArF^*(v, J)+F$ $k \sim 6\times10^{-10}$ cm^3/sec; Branching ratio to $ArF^* \sim 60\%$ $Xe^* + F_2 \rightarrow XeF^*(v, J)+F$ $k \sim 7\times10^{-10}$ cm^3/sec; Branching ratio to $XeF^* \sim 1$ $Xe^* + HCl(v = 0) \rightarrow Xe + H + Cl$ $k \sim 6\times10^{-10}$ cm^3/sec; Branching ratio to $XeCl^* \sim 0$ $Xe^* + HCl(v = 1) \rightarrow XeCl^* + H$ $k \sim 2\times10^{-10}$ cm^3/sec; Branching ratio to $XeCl^* \sim 1$
Ion–ion recombination	$Ar^{(+)} + F^{(-)} + Ar \rightarrow ArF^*(v,J)$ $k \sim 1\times10^{-25}$ cm^6/sec at 1 atm and below $k \sim 7.5\times10^{-26}$ cm^6/sec at 2 atm; Note effective 3-body rate constant decreases further as pressure goes over ~2 atm $Kr^{(+)} + F^{(-)} + Kr \rightarrow KrF^*(v, J)$ $k \sim 7\times10^{-26}$ cm^6/sec at 1 atm and below; rolls over at higher pressure
Displacement	$ArF^*(v, J)+Kr \rightarrow KrF^*(v, J)+Ar$ $k \sim 2\times10^{-10}$ cm^3/sec; Branching ratio to $KrF^* \sim 1$
Quenching	
Halogen molecule	$XeCl^* + HCl \rightarrow Xe + Cl + HCl$ $k \sim 5.6\times10^{-10}$ cm^3/sec; $KrF^* + F_2 \rightarrow Kr + F + F$ $k \sim 6\times10^{-10}$ cm^3/sec; All halogen molecule quenching have very high reaction rate constants
3-body inert gas	$ArF^* + Ar + Ar \rightarrow Ar_2F^* + Ar$ $k \sim 4\times10^{-31}$ cm^6/sec $KrF^* + Kr + Ar \rightarrow Kr_2F^* + Ar$ $k \sim 6\times10^{-31}$ cm^6/sec $XeCl^* + Xe + Ne \rightarrow Xe_2Cl^* + Ne$ $k \sim 1\times10^{-33}$ cm^6/sec; 4×10^{-31} cm^6/sec with Xe as 3rd body These reactions yield a triatomic excimer at lower energy that can not be recycled back to the desired laser states
Electrons	$XeCl^* + e \rightarrow Xe + Cl + e$ $k \sim 2\times10^{-7}$ cm^3/sec In 2-body quenching by electrons, the charge of the electron interacts with the dipole of the rare gas halide ion pair at long range; above value is typical of others
2-body inert gas	$KrF^* + Ar \rightarrow Ar + Kr + F$ $k \sim 2\times10^{-12}$ cm^3/sec $XeCl^* + Ne \rightarrow Ne + Xe + Cl$ $k \sim 3\times10^{-13}$ cm^3/sec 2-body quenching by rare gases is typically slow and less important than the reactions noted above

somewhat less in an electric discharge. The photon release is ~5 eV; the quantum efficiency is only 25%. Early laser efficiency measurements (using electron beam excitation) showed laser efficiency in the best cases slightly in excess of 10%, relative to the deposited energy. The discrepancy relative to the quantum efficiency is due to losses in the medium. For the sample estimate in Table 4 and Figure 5 we show approximate density of key species in the absorption chain, estimated on a steady-state approximation of the losses in a KrF laser for an excitation rate of order 1 MW/cc, typical of fast discharge lasers. The net small signal gain is of the order of 5%/cm. The key absorbers are the parent halogen molecule, F_2, the fluoride ion $F^{(-)}$, and the rare gas excited state atoms. A detailed pulsed kinetics code will provide slightly different results due to transient kinetic effects and the intimate coupling of laser extraction to quenching and absorber dynamics. The ratio of gain to loss is usually in the range of 7 to 20, depending on the actual mixture used and the rare gas halide wavelength. The corresponding extraction efficiency is then limited to the range of ~50%. The small signal gain, g_o, is related to the power deposition

Table 4 Typical ground and excited state densities shown for KrF at $P_{in} \sim 1$ MW/cm^3

Species	*Amount particles/cm^3*	*Absorption σ (cm)^2*	*Loss %/cm*
F_2	$\sim 1.2 \times 10^{17}$	1.5×10^{-20}	0.18
Kr	$\sim 3 \times 10^{18}$	–	–
Buffer	4×10^{19}	–	–
$F^{(-)}$	$\sim 10^{14}$	5×10^{-18}	0.05
Rg^*, Rg^{**}	1×10^{15}	$\langle 10^{-18} \rangle$ depends on ratio of Rg^{**}/Rg^*	0.1
KrF (before dissociating)	$\sim 10^{12}$	2×10^{-16}	0.02
Rg_2F^*	8×10^{14}, Note this species density decreases as flux grows	1×10^{-18}	0.08
KrF^* (B)	2.5×10^{14}	2×10^{-16}	5 (gain)

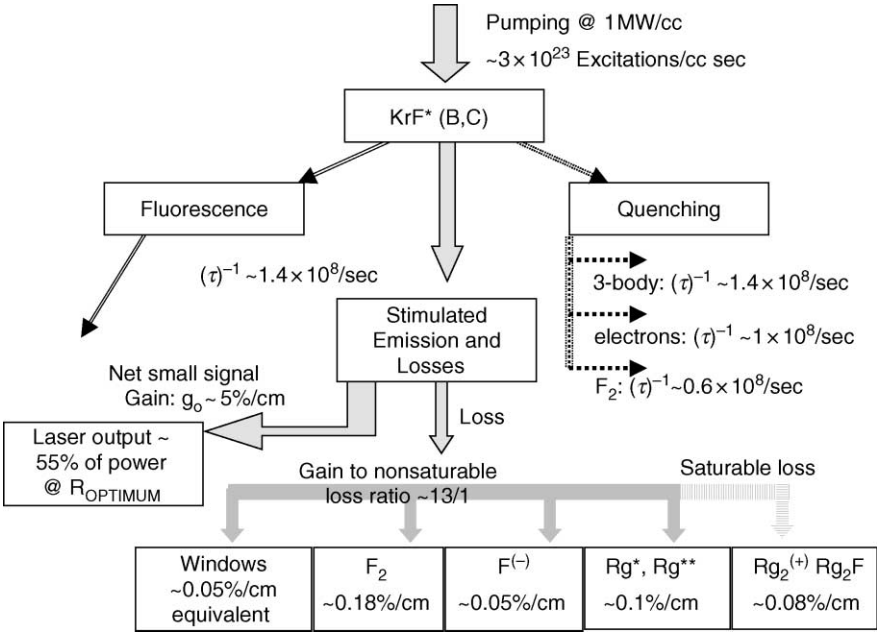


Figure 5 The major decay paths and absorbers that impact the extraction of excited states of rare gas halide excimers. Formation is via the ion–ion recombination, excited atom reaction with halogens and displacement reactions, not shown. We note for a specific mixture of KrF the approximate values of the absorption by the transient and other species along with the decay rates for the major quenching processes. Extraction efficiency greater than 50% is quite rare.

rate per unit volume, $P_{\text{deposited}}$, by eqn [10].

$$P_{\text{deposited}} = g_0 E^* (\tau_{\text{upper}} \eta_{\text{pumping}} \eta_{\text{branching}} f \sigma)^{-1} \qquad [10]$$

E^* is the energy per excitation, ~20 eV or 3.2×10^{-18} J/excitation, f is the fraction of excited molecules in the upper laser level, ~40% due to KrF^* in higher vibrational states and the slowly radiating 'C' state. The laser cross-section is σ, and the η terms are efficiencies for getting from the primary excitation down to the upper laser level. The upper state lifetime, τ_{upper}, is the sum of the inverses of radiative and quenching lifetimes. When the laser cavity flux is sufficiently high, stimulated emission decreases the flow of power into the quenching processes.

The example shown here gives a ratio of net gain to the nonsaturating component of loss of 13/1. Of the excitations that become KrF^*, only ~55% or so will be extracted in the best of cases. For other less ideal rare gas halide lasers, the gain to loss ratio, the extraction efficiency, and the intrinsic efficiency are all lower. Practical discharge lasers have lower practical wall plug efficiency due to a variety of factors, discussed below.

Discharge Technology

Practical discharge lasers use fast-pulse discharge excitation. In this scheme, an outgrowth of the CO_2 TEA laser, the gas is subjected to a rapid high-voltage pulse from a low-inductance pulse forming line or network, PFL or PFN. The rapid pulse serves to break down the gas, rendering it conductive with an electron density of $> \sim 10^{14}$ electrons/cm^3. As the

gas becomes conductive, the voltage drops and power is coupled into the medium for a short duration, typically less than 50 nsec. The discharge laser also requires some form of 'pre-ionization', usually provided by sparks or a corona-style discharge on the side of the discharge electrodes. The overall discharge process is intrinsically unstable, and pump pulse duration needs to be limited to avoid the discharge becoming an arc.

Moreover, problems with impedance matching of the PFN or PFL with the conductive gas lead to both inefficiency and extra energy that can go into post-discharge arcs, which both cause component failure and produce metal dust which can give rise to window losses. A typical discharge laser has an aperture of a few cm^2 and a gain length of order 80 cm and produces outputs in the range of 100 mJ to 1 J. The optical pulse duration tends to be ~20 nsec, though the electrical pulse is longer due to the finite time needed for the gain to build up and the stimulated emission process to turn spontaneously emitted photons into photons in the laser cavity.

The pre-ionizer provides an initial electron density on the order of 10^8 electrons/cm^3. With spark pre-ionization or corona pre-ionization, it is difficult to make a uniform discharge over an aperture greater than a few cm^2. For larger apertures, one uses X-rays for pre-ionization. X-ray pre-ionized discharge excimer lasers have been scaled to energy well over 10 J per pulse and can yield pulse durations over 100 nsec. The complexity and expense of the X-ray approach, along with the lack of market for lasers in this size range, resulted in the X-ray approach remaining in the laboratory.

Aside from the kinetic issues of the finite time needed for excitation to channel into the excimer upper levels and the time needed for the laser to 'start up' from fluorescent photons, these lasers have reduced efficiency due to poor matching of the pump source to the discharge. Ideally, one would like to have a discharge such that when it is conductive the voltage needed to sustain the discharge is less than half of that needed to break the gas down. Regrettably for the rare gas halides (and in marked distinction to the CO_2 TEA laser) the voltage that the discharge sustains in a quasi-stable mode is ~20% of the breakdown voltage, perhaps even less, depending on the specific gases and electrode shapes.

Novel circuits, which separate the breakdown phase from the fully conductive phase, can readily double the efficiency of an excimer discharge laser, though this approach is not necessary in many applications. The extra energy that is not dissipated in the useful laser discharge often results in arcing or 'hot' discharge areas after the main laser pulse, leading to electrode sputtering, dust that finds its way to windows, and the need to service the laser after ~10^8 pulses. When one considers the efficiency of the power conditioning system, along with the finite kinetics of the laser start-up process in short pulses, and the mismatch of the laser impedance to the PFN/PFL impedance, lasers that may have 10% intrinsic efficiency in the medium in e-beam excitation may only have 2% to 3% wall plug efficiency in a practical discharge laser.

A key aspect of excimer technology is the need for fast electrical circuits to drive the power into the discharge. This puts a significant strain on the switch that is used to start the process. While spark gaps, or multichannel spark gaps, can handle the required current and current rate of rise, they do not present a practical alternative for high-pulse rate operation. Thyratrons and other switching systems, such as solid state switches, are more desirable than a spark gap, but do not offer the desired needed current rate of rise. To provide the pumping rate of order 1 MW/cm^3 needed for enough gain to turn the laser on before the discharge becomes unstable, the current needs to rise to a value of the order of 20 kA in a time of ~20 nsec; the current rate of rise is ~10^{12} A/sec. This rate of rise will limit the lifetime of the switch that drives the power into the laser gas. As a response to this need, magnetic switching and magnetic compression circuits were developed so that the lifetime of the switch can be radically increased. In the simplest cases, one stage of magnetic compression is used to make the current rate of rise to be within the range of a conventional thyratron switch as used in radar systems. Improved thyratrons and a simple magnetic assist also work. For truly long operating lifetimes, multiple stages of magnetic switching and compression are used, and with a transformer one can use a solid state diode as the start switch for the discharge circuit. A schematic of an excimer pulsed power circuit, that uses a magnetic assist with a thyratron and one stage of magnetic compression, is shown in Figure 6. The magnetic switching approach can also be used to provide two separate circuits to gain high efficiency by having one circuit to break down the gas (the spiker) and a second, low-impedance 'sustainer' circuit, to provide the main power flow. In this approach, the laser itself can act as part of the circuit by holding off the sustaining voltage for a μ-sec or so during the charge cycle and before the spiker breaks down the gas. A schematic of such a circuit is shown in Figure 7. Using this type of circuit, the efficiency of a XeCl laser can be over 4% relative to the wall plug. A wide variety of technological twists on this approach have been researched.

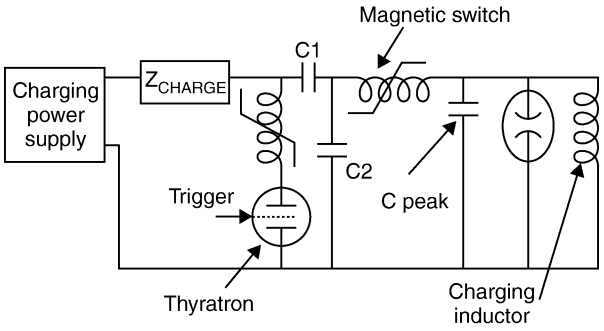


Figure 6 One of the variants on magnetic switching circuits that have been used to enhance the lifetime of the primary start switch. A magnetic switch in the form of a simple 1 turn inductor with a saturable core magnetic material such as a ferrite allows the thyratron to turn on before major current flows. Other magnetic switches may be used in addition to peak up the voltage rate of rise on the laser head or to provide significant compression. For example one may have the major current flow in the thyratron taking place on the 500 nsec time scale while the voltage pulse on the discharge is in the 50 nsec duration. A small current leaks through and is accommodated by the charging inductor.

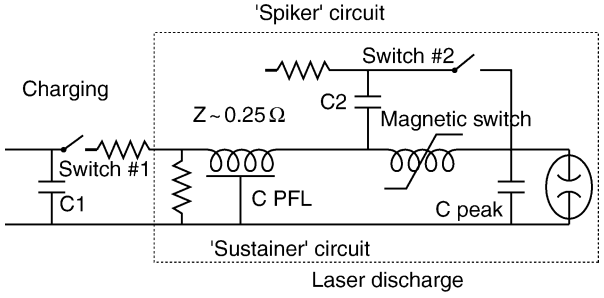


Figure 7 For the ultimate in efficiency the circuit can be arranged to provide a leading edge spike that breaks the gas down using a high impedance 'spiker' and a separate circuit that is matched to the discharge impedance when it is fully conductive. High coupling efficiency reduces waste energy that can erode circuit and laser head components via late pulse arcs. The figure shows a magnetic switch as the isolation but the scheme can be implemented with a second laser head as the isolation switch, a rail gap (not appropriate for long life) or a diode at low voltage.

Excimer lasers present some clear differences in resonator design compared to other lasers. The apertures are relatively large, so one does not build a TEM_{oo} style cavity for good beam quality and expect to extract any major portion of the energy available. The typical output is divergent and multimode. When excimer lasers are excited by a short pulse discharge, ~30 nsec, the resulting gain duration is also short, ~20 nsec, limiting the number of passes a photon in a resonator can have in the gain. The gain duration of ~20 nsec only provides ~3.5 round trips of a photon in the cavity during the laser lifetime, resulting in the typical high-order multimode beam. Even with an unstable resonator, the gain duration is not long enough to collapse the beam into the lowest-order diffraction-limited output. If line narrowing is needed, and reasonable efficiency is required, an oscillator amplifier configuration is often used with the oscillator having an appropriate tuning resonator in the cavity. If both narrow spectral linewidth and excellent beam quality are needed one uses a seeded oscillator approach where the narrowband oscillator is used to seed an unstable resonator on the main amplifier. By injecting the seed into the unstable resonator cavity, a few nsec before the gain is turned on, the output of the seeded resonator can be locked in wavelength, bandwidth, and provide good beam quality. Long pulse excimer lasers, such as e-beam excited lasers or long pulse X-ray pre-ionized discharge devices can use conventional unstable resonator technology with good results.

Excimer lasers are gas lasers that run at both high pressure and high instantaneous power deposition rates. During a discharge pulse, much of the halogen bearing 'fuel' is burned out by the attachment and reactive processes. The large energy deposition per pulse means that pressure waves are generated. All of these combine with the characteristic device dimensions to require gas flow to recharge the laser mixture between pulses. For low pulse rates, less than ~200 Hz, the flow system need not be very sophisticated. One simply needs to flush the gas from the discharge region with a flow having a velocity of the order of 5 m/sec for a typical discharge device. The flow is transverse to the optical and discharge directions. At high pulse rates, the laser designer needs to consider flow in much more detail to minimize the power required to push the gas through the laser head. Circulating pressure waves need to be damped out so that the density in the discharge region is controlled at the time of the next pulse. Devices called acoustic dampers are placed in the sides of the flow loop to remove pressure pulses. A final subtlety occurs in providing gas flow over the windows. The discharge, post pulse arcs, and reactions of the halogen source with the metal walls and impurities creates dust. When the dust coats the windows, the losses in the cavity increase, lowering efficiency. By providing a simple gas flow over the window, the dust problem can be ameliorated. For truly long life, high-reliability lasers, one needs to take special care in the selection of materials for electrodes and insulators and avoid contamination, by assembling in clean environments.

See also

Lasers: Carbon Dioxide Lasers. **Laser-Induced Damage of Optical Materials**. **Nonlinear Optics,**

Applications: Pulse Compression via Nonlinear Optics; Raman Lasers. **Scattering:** Raman Scattering.

Further Reading

Ballanti S, Di Lazzaro P, Flora F, *et al.* (1998) Ianus the 3-electrode laser. *Applied Physics B* 66: 401–406.

Brau CA (1978) Rare gas halogen lasers. In: Rhodes CK (ed.) *Excimer Lasers, Topics of Applied Physics*, vol. 30. Berlin: Springer Verlag.

Brau CA and Ewing JJ (1975) Emission spectra of XeBr, XeCl, XeF, and KrF. *Journal of Chemical Physics* 63: 4640–4647.

Ewing JJ (2000) Excimer laser technology development. *IEEE Journal of Selected Topics in Quantum Electronics* 6(6): 1061–1071.

Levatter J and Lin SC (1980) Necessary conditions for the homogeneous formation of pulsed avalanche discharges at high pressure. *Journal of Applied Physics* 51: 210–222.

Long WH, Plummer MJ and Stappaerts E (1983) Efficient discharge pumping of an XeCl laser using a high voltage prepulse. *Applied Physics Letters* 43: 735–737.

Rhodes CK (ed.) (1979) *Excimer Lasers, Topics of Applied Physics*, vol. 30. Berlin: Springer-Verlag.

Rokni M, Mangano J, Jacob J and Hsia J (1978) Rare gas fluoride lasers. *IEEE Journal of Quantum Electronics* QE-14: 464–481.

Smilanksi I, Byron S and Burkes T (1982) Electrical excitation of an XeCl laser using magnetic pulse compression. *Applied Physics Letters* 40: 547–548.

Taylor RS and Leopold KE (1994) Magnetic-spiker excitation of gas discharge lasers. *Applied Physics B, Lasers and Optics* 59: 479–509.

Free Electron Lasers

A Gover, Tel-Aviv University, Tel-Aviv, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The Free Electron Laser (FEL) is an exceptional kind of laser. Its active medium is not matter, but charged particles (electrons) accelerated to high energies, passing in vacuum through a periodic undulating magnetic field. This distinction is the main reason for the exceptional properties of FEL: operating at a wide range of wavelengths – from mm-wave to X-rays with tunability, high power, and high efficiency.

In this article we explain the physical principles of FEL operation, the underlying theory and technology of the device and various operating schemes, which have been developed to enhance performance of this device.

The term 'Free Electron Laser' was coined by John Madey in 1971, pointing out that the radiative transitions of the electrons in this device are between free space (more correctly – unbound) electron quantum states, which are therefore states of continuous energy. This is in contrast to conventional atomic and molecular lasers, in which the electron performs radiative transition between bound (and therefore of distinct energy) quantum states. Based on these theoretical observations, Madey and his colleagues in Stanford University demonstrated FEL operation first as an amplifier (at $\lambda = 10.6\ \mu m$) in 1976, and subsequently as an oscillator (at $\lambda = 3.4\ \mu m$) in 1980.

From the historical point of view, it turned out that Madey's invention was essentially an extension of a former invention in the field of microwave-tubes technology – the Ubitron. The Ubitron, a mm-wave electron tube amplifier based on a magnetic undulator, was invented and developed by Philips and Enderbry who operated it at high power levels in 1960. The early Ubitron development activity was not noticed by the FEL developers because of the disciplinary gap, and largely because its research was classified at the time. Renewed interest in high-power mm-wave radiation emission started in the 1970s, triggered by the development of pulsed-line generators of 'Intense Relativistic Beams' (IRB). This activity, led primarily by plasma physicists in the defense establishment laboratories of Russia (mostly IAP in Gorky – Nizhny Novgorod) and the US (mostly N.R.L. – DC) led to development of high-gain high-power mm-wave sources independently of the development of the optical FEL. The connection between these devices and between them to conventional microwave tubes (as Traveling Wave Tubes – TWT) and other electron beam radiation schemes, like Cenenkov and Smith-Purcell radiation that may also be considered FELs, was revealed in the mid-1970s, starting with the theoretical works of P. Spangle, A. Gover and A. Yariv who identified that all these devices satisfy the same dispersion equation as the TWT derived by John Pierce in the 1940s. Thus, the optical FEL could be conceived as a kind of immense electron tube, operating with a high-energy electron beam in the low gain regime of the Pierce TWT dispersion equation.

The extension of the low-gain FEL theory to the general 'electron-tube' theory is important because it led to development of new radiation schemes and new operating regimes of the optical FEL. This was exploited by physicists in the discipline of accelerator physics and synchrotron radiation, who identified, starting with the theoretical works of C. Pellegrini and R. Bonifacio in the early 1980s, that high-current, high-quality electron beams, attainable with further development of accelerators technology, could make it possible to operate FELs in the high-gain regime, even at short wavelengths (vacuum ultra-violet – VUV and soft X-ray) and that the high-gain FEL theory can be extended to include amplification of the incoherent synchrotron spontaneous emission (shot noise) emitted by the electrons in the undulator. These led to the important development of the 'self (synchrotron) amplified spontaneous emission (SASE) FEL', which promised to be an extremely high brightness radiation source, overcoming the fundamental obstacles of X-ray lasers development: lack of mirrors (for oscillators) and lack of high brightness radiation sources (for amplifiers).

A big boost to the development of FEL technology was given during the period of the American 'strategic defense initiative – SDI' (Star-Wars) program in the mid-1980s. The FEL was considered one of the main candidates for use in a ground-based or space-based 'directed energy weapon – DEW', that can deliver megawatts of optical power to hit attacking missiles. The program led to heavy involvement of major American defense establishment laboratories (Lawrence–Livermore National Lab, Los-Alamos National Lab) and contracting companies (TRW, Boeing). Some of the outstanding results of this effort were demonstration of the high-gain operation of an FEL amplifier in the mm-wavelength regime, utilizing an Induction Linac (Livermore, 1985), and demonstration of enhanced radiative energy extraction efficiency in FEL oscillator, using a 'tapered wiggler' in an RF-Linac driven FEL oscillator (Los-Alamos, 1983). The program has not been successful in demonstrating the potential of FELs to operate at the high average power levels needed for DEW applications. But after the cold-war period, a small part of the program continues to support research and development of medical FEL application.

Principles of FEL Operation

Figure 1 displays schematically an FEL oscillator. It is composed of three main parts: an electron accelerator, a magnetic wiggler (or undulator), and an optical resonator.

Without the mirrors, the system is simply a synchrotron undulator radiation source. The electrons in the injected beam oscillate transversely to their propagation direction z, because of the transverse magnetic Lorenz force:

$$\mathbf{F}_{\perp} = -ev_z\hat{\mathbf{e}}_z \times \mathbf{B}_{\perp} \qquad [1]$$

In a planar (linear) wiggler, the magnetic field on axis is approximately sinusoidal:

$$\mathbf{B}_{\perp} = B_{\mathrm{w}}\hat{e}_y \cos k_{\mathrm{w}}z \qquad [2]$$

In a helical wiggler:

$$\mathbf{B}_{\perp} = B_{\mathrm{w}}(\hat{\mathbf{e}}_y \cos k_{\mathrm{w}}z + \hat{\mathbf{e}}_x \sin k_{\mathrm{w}}z) \qquad [3]$$

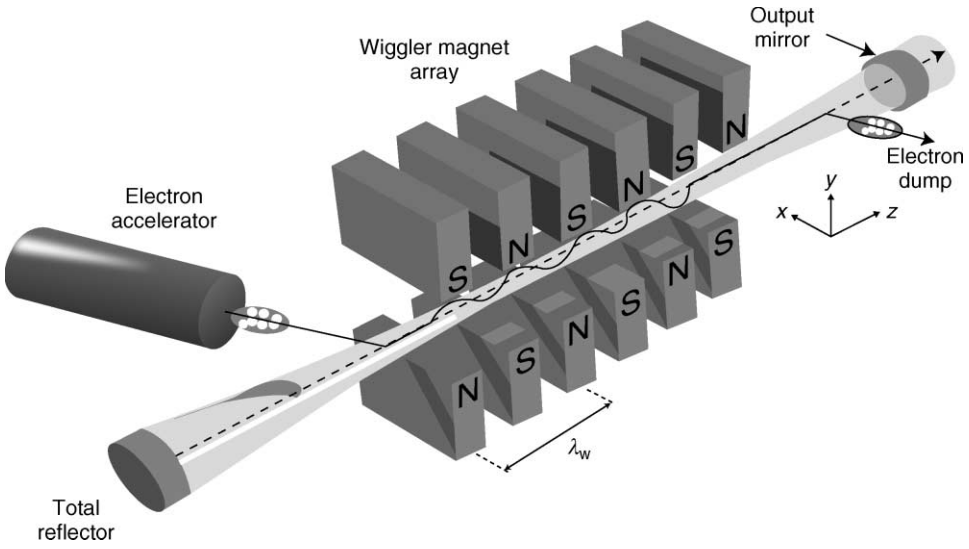


Figure 1 Components of a FEL-oscillator. (Reproduced from Benson SV (2003) Free electron lasers push into new frontiers. *Optics and Photonics News* 14: 20–25. Illustration by Jaynie Martz.)

In either case, if we assume constant (for the planar wiggler – only on the average) axial velocity, then $z = v_z t$. The frequency of the transverse force and the mechanical oscillation of the electrons, as viewed transversely in the laboratory frame of reference, is:

$$\omega_{0s} = k_w v_z = 2\pi \frac{v_z}{\lambda_w} \qquad [4]$$

where $\lambda_w = 2\pi/k_w$ is the wiggler period.

The oscillating charge emits an electromagentic radiation wavepacket. In a reference frame moving with the electrons, the angular radiation pattern looks exactly like dipole radiation, monochromatic in all directions (except for the frequency-line-broadening due to the finite oscillation time, i.e., the wiggler transit time). In the laboratory reference-frame the radiation pattern concentrates in the propagation ($+z$) direction, and the Doppler up-shifted radiation frequency depends on the observation angle Θ relative to the z-axis:

$$\omega_0 = \frac{\omega_{0s}}{1 - \beta_z \cos\Theta} \qquad [5]$$

On axis ($\Theta = 0$), the radiation frequency is:

$$\omega_0 = \frac{ck_w\beta_z}{1-\beta_z} = (1+\beta_z)\beta_z\gamma_z^2 ck_w \cong 2\gamma_z^2 ck_w \qquad [6]$$

where $\beta_z \equiv v_z/c$, $\gamma_z \equiv (1-\beta_z^2)^{-1/2}$ are the axial (average) velocity and the axial Lorenz factor, respectively, and the last part of the equation is valid only in the (common) highly relativistic limit $\gamma_z \gg 1$.

Using the relations $\beta_z^2 + \beta_\perp^2 = \beta^2$, $\beta_\perp = a_w/\gamma$, one can express γ_z:

$$\gamma_z = \frac{\gamma}{1 + a_w^2/2} \qquad [7]$$

(this is for a linear wiggler, in the case of a helical wiggler the denoninator is $1 + a_w^2$):

$$\gamma \equiv (1-\beta^2)^{-1/2} = 1 + \frac{\varepsilon_k}{mc^2} = 1 + \varepsilon_k\,[\text{MeV}]/0.511 \qquad [8]$$

and a_w – (also termed K) 'the wiggler parameter' is the normalized transverse momentum:

$$a_w = \frac{eB_w}{k_w mc} = 0.093 B_w\,[\text{KGauss}]\lambda_w\,[\text{cm}] \qquad [9]$$

Typical values of B_w in FEL wigglers (undulators) are of the order of Kgauss', and λ_w of the order of CMs, and consequently $a_w < 1$. Considering that electron beam accelerator energies are in the range of MeV to GeV, one can appreciate from eqns [6]–[8], that a significant relativistic Doppler shift factor $2\gamma_z^2$, in the range of tens to millions, is possible. It, therefore, provides incoherent synchrotron undulator radiation in the frequency range of microwave to hard X-rays.

Synchrotron undulator radiation was studied in 1951 and since then has been a common source of VUV radiation in synchrotron facilities. From the point of view of laser physics theory, this radiation can be viewed as 'spontaneous synchrotron radiation emission' in analogy to spontaneous radiation emission by electrons excited to higher bound-electron quantum levels in atoms or molecules. Alternatively, it can be regarded as the classical shot noise radiation, associated with the current fluctuations of the randomly injected discrete charges comprising the electron beam. Evidently this radiation is incoherent, and the fields it produces average in time to zero, because the wavepackets emitted by the randomly injected electrons interfere at the observation point with random phase. However, their energies sum up and can produce substantial power.

Based on fundamental quantum-electrodynamical principles or Einstein's relations, one would expect that any spontaneous emission scheme can be stimulated. This principle lies behind the concept of the FEL, which is nothing but stimulated undulator synchrotron radiation. By stimulating the electron beam to emit radiation, it is possible, as with any laser, to generate a coherent radiation wave and extract more power from the gain medium, which in this case is an electron beam, that carries an immense amount of power. There are two kinds of laser schemes which utilize stimulation of synchrotron undulator radiation:

(i) *A laser amplifier*. In this case the mirrors in the schematic configuration of **Figure 1** are not present, and an external radiation wave at frequencies within the emission range of the undulator is injected at the wiggler entrance. This requires, of course, an appropriate radiation source to be amplified and availability of sufficiently high gain in the FEL amplifier.

(ii) *A laser oscillator*. In this case an open cavity (as shown in **Figure 1**) or another (waveguide) cavity is included in the FEL configuration. As in any laser, the FEL oscillator starts building up its radiation from the spontaneous (synchrotron undulator) radiation which gets trapped in the resonator and amplified by stimulated emission along the wiggler. If the threshold condition is satisfied (having single path gain higher than the round trip losses), the oscillator arrives to saturation and steady state coherent operation after a short transient period of oscillation build-up.

Because the FEL can operate as a high-gain amplifier (with a long enough wiggler and an electron beam of high current and high quality), also a third mode of operation exists: self amplified spontaneous emission (SASE). In this case, the resonator mirrors in Figure 1 are not present and the undulator radiation generated spontaneously in the first sections of the long undulator is amplified along the wiggler and emitted at the wiggler exit at high power and high spatial coherence.

The Quantum-Theory Picture

A free electron, propagating in unlimited free space, can never emit a single photon. This can be proven by examining the conservation of energy and momentum conditions:

$$\varepsilon_{k_i} - \varepsilon_{k_f} = \hbar\omega \qquad [10]$$

$$\mathbf{k}_i - \mathbf{k}_f = \mathbf{q} \qquad [11]$$

that must be satisfied, when an electron in an initial free-space energy and momentum state $(\varepsilon_{k_i}, \hbar k_i)$ makes a transition to a final state $(\varepsilon_{k_f}, \hbar k_f)$, emitting a single photon of energy and momentum $(\hbar\omega, q)$. In free space:

$$\varepsilon_k = c\sqrt{(\hbar\mathbf{k})^2 + (mc^2)^2} \qquad [12]$$

$$\mathbf{q} = \frac{\omega}{c}\hat{\mathbf{e}}_q \qquad [13]$$

and eqns [10]–[13] have only one solution, $\omega = 0$, $\mathbf{q} = 0$. This observation is illustrated graphically in the energy–momentum diagram of Figure 2a in the framework of a one-dimensional model. It appears that if both eqns [10] and [11] can be satisfied, then the phase velocity of the emitted radiation wave $v_{ph} = \omega/q$ (the slope of the chord) will equal the electron wavepacket group velocity $v_g = v_z$ at some intermediate point $k^* = p^*/\hbar$:

$$v_{ph} = \frac{\omega}{q} = \frac{\varepsilon_{ki} - \varepsilon_{kf}}{\hbar(k_i - k_f)} = \left.\frac{\partial\varepsilon}{\partial p}\right|_{p^*} = v_g \qquad [14]$$

For a radiation wave in free space (eqn [13]), this results in $c = v_g$, which contradicts special relativity.

The reason for the failure to conserve both energy and momentum in the transition is that the photon momentum $\hbar q$ it too small to absorb the large momentum shift of the electron, as it recoils while releasing radiative energy $\hbar\omega$. This observation leads to ideas on how to make a radiative transition possible:

(i) *Limit the interaction length*. If the interaction length is L, the momentum conservation

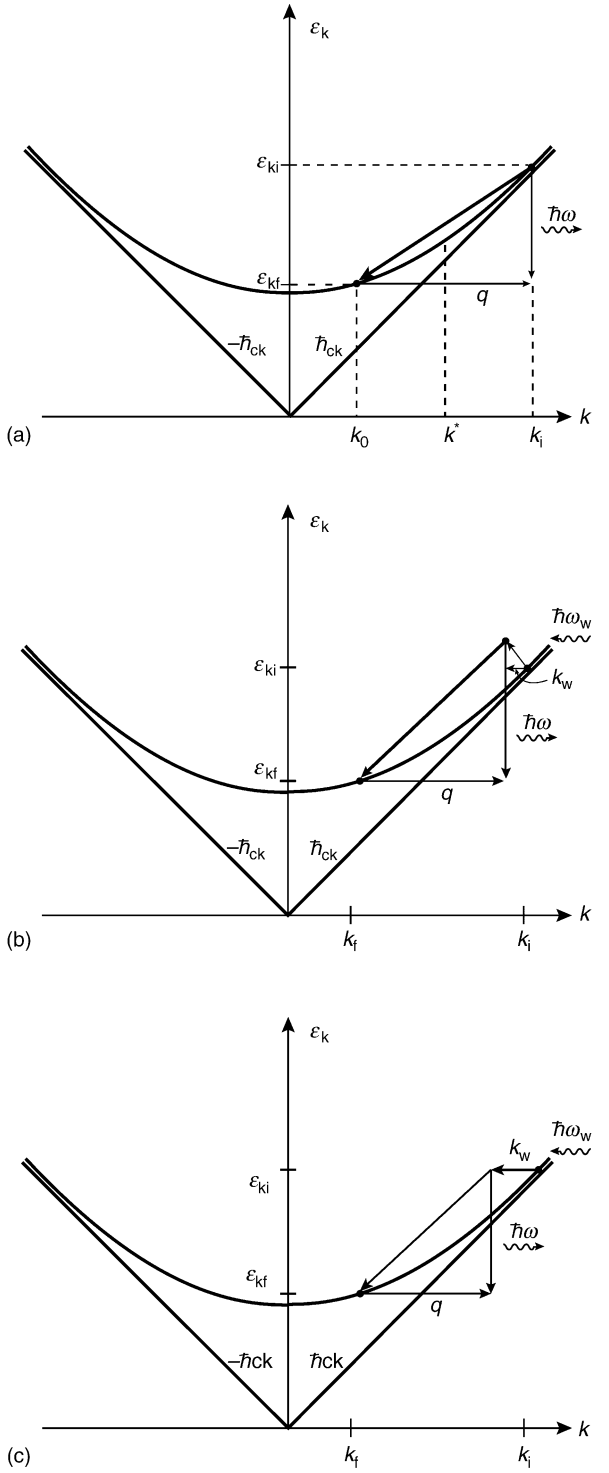


Figure 2 Conservation of energy and momentum in forward photon emission of a free electron: (a) The slope of the tangent to the curve at intermediate point k^*, $\partial\varepsilon_k(k^*)/\partial k$ may be equal to the slope of the chord $\hbar\omega/q$ which is impossible in free space. (b) electron radiative transition made possible with an electromagnetic pump (Compton Scattering). (c) The wiggler wavenumber – k_w conserves the momentum in electron radiative transition of FEL.

condition in eqn [11] must be satisfied only within an uncertainty range $\pm\pi/L$. This makes it possible to obtain radiative emission in free electron radiation effects like 'Transition Radiation' and in microwave tubes like the Klystron.

(ii) *Propagate the radiation wave* in a 'slow wave' structure, where the phase velocity of the radiation wave is smaller than the speed of light, and satisfaction of eqn [14] is possible. For example, in the Cerenkov effect, charged particles pass through a medium (gas) with index of refraction $n > 1$. Instead of eqn [13], $\mathbf{q} = n(\omega)(\omega/c)\hat{\mathbf{e}}_q$, and consequently $q_z = n(\omega)(\omega/c)\cos\Theta_q$, where we assume radiative emission at an angle Θ_q relative to the electron propagation axis z. Substitution in eqn [14] results in the Cerenkov radiation condition $v_g n(\omega)\cos\Theta_q = 1$.

Another example for radiation emission in a slow-wave structure is the Traveling Wave Tube (TWT). In this device, a periodic waveguide of periodicity λ_w permits (via the Floquet theorem) propagation of slow partial waves (space harmonics) with increased wavenumber $q_z + mk_w$ ($m = 1, 2, \ldots$), and again eqn [14] can be satisfied.

(iii) *Rely on a 'two-photon' radiative transition.* This can be 'real photon' Compton scattering of an intense radiation beam (electromagnetic pump) off an electron beam, or 'virtual photon' scattering of a static potential, as is the case in bremsstrahlung radiation and in synchrotron–undulator radiation. The latter radiation scheme may be considered as a 'magnetic brehmsstrahlung' effect or as 'zero frequency pump' Compton scattering, in which the wiggler contributes only 'crystal momentum' $\hbar k_w$, to help satisfy the momentum conservation condition in eqn [11]. The Compton scattering scheme is described schematically for the one-dimensional (back scattering) case in Figure 3, and its conservation of energy and momentum diagram is depicted in Figure 2b (a 'real photon' (ω_w, k_w) free-space pump wave is assumed with $k_w = \omega_w/c$). The analogous diagram of a static wiggler ($\omega_w = 0$, $k_w = 2\pi/\lambda_w$) is shown in Figure 2c. It is worth noting that the effect of the incident scattered wave or the wiggler is not necessarily a small perturbation. It may modify substantially the electron energy-dispersion diagram of the free electron and a more complete 'Brillouin diagram' should be used in Figure 2c. In this sense, the wiggler may be viewed as the analogue of a one-dimensional crystal, and its period λ_w analogous to the crystal lattice constant. The momentum conservation during a radiation transition, with the aid of the wiggler 'crystal momentum' $\hbar k_w$ is quite analogous to the occurrence of vertical radiative transitions in direct bandgap semiconductors, and thus the FEL has, curiously enough, some analogy to microscopic semiconductor lasers.

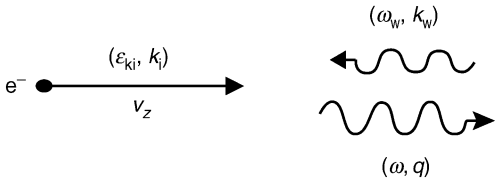


Figure 3 The scheme of backward scattering of an electromagnetic wave off an electron beam (Doppler shifted Compton scattering).

All the e-beam radiation schemes already mentioned can be turned into stimulated emission devices, and thus may be termed 'free electron lasers' in the wide sense. The theory of all of these devices is closely related, but most of the technological development was carried out on undulator radiation (Magnetic brehmsstrahlung) FELs, and the term FEL is usually reserved for this kind (though some developments of Cerenkov and Smith–Purcell FELs are still carried out).

When considering a stimulated emission device, namely enhanced generation of radiation in the presence of an external input radiation wave, one should be aware, that in addition to the emission process described by eqns [10] and [11] and made possible by one of the radiation schemes described above, there is also a stimulated absorption process. Also, this electronic transition process is governed by the conservation of energy and momentum conditions, and is described by eqns [10] and [11] with k_i and k_f exchanged.

Focusing on undulator-radiation FEL and assuming momentum conservation in the axial (z) dimension by means of the wiggler wavenumber k_w, the emission and absorption quantum transition levels and radiation frequencies are found from the solution of equations:

$$\varepsilon_{k_{zi}} - \varepsilon_{k_{zf}} = \hbar\omega_e \qquad [15a]$$

$$k_{zi} - k_{zf} = q_{ze} + k_w \qquad [15b]$$

$$\varepsilon_{k_{za}} - \varepsilon_{k_{zi}} = \hbar\omega_a \qquad [16a]$$

$$k_{za} - k_{zi} = q_{za} + k_w \qquad [16b]$$

For fixed k_w, fixed transverse momentum and given e-beam energy $\varepsilon_{k_{zi}}$ and radiation emission angle Θ_q ($q_z = (\omega/c)\cos\Theta_q$), eqns [15] and [16] have separately

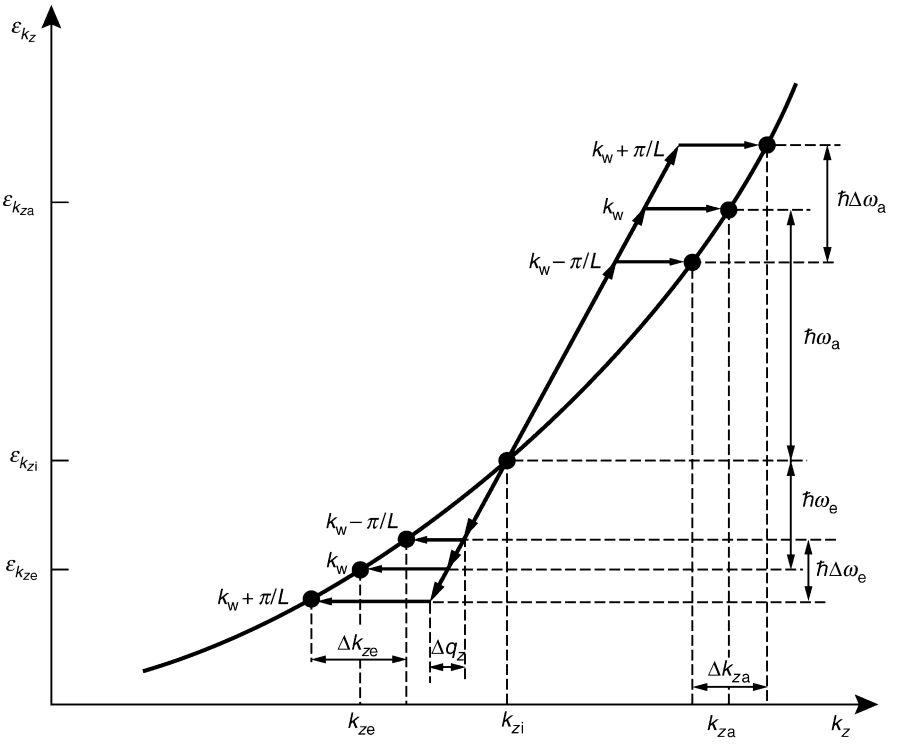


Figure 4 The figure illustrates that the origin of difference between the emission and absorption frequencies is the curvature of the energy dispersion line, and the origin of the homogeneous line broadening is momentum conservation uncertainty $\pm\pi/L$ in a finite interaction length. (Reproduced with permission from Friedman A, Gover A, Ruschin S, Kurizki G and Yariv A (1988) Spontaneous and stimulated emission from quasi-free electrons. *Reviews of Modern Physics* 60: 471–535. Copyright (1988) by the American Physical Society.)

distinct solutions, defining the electron upper and lower quantum levels for radiative emission and absorption respectively. The graphical solutions of these two set of equations are shown in Figure 4, which depicts also the 'homogeneous' frequency-line broadening $\hbar\Delta\omega_e$, $\hbar\Delta\omega_a$ of the emission and absorption lines due to the uncertainty in the momentum conservation $\pm\pi/L$ in a finite interaction length. In the quantum limit of a cold (monoenergetic) e-beam and a long interaction length L, the absorption line center ω_a is larger than the emission line center ω_e, and the linewidths $\Delta\omega_e \cong \Delta\omega_a = \Delta\omega_L$ are narrower than the emission and absorption lines spacing $\omega_a - \omega_e$, as shown in Figure 5a. The FEL then behaves as a 3-level quantum system, with electrons occupying only the central level, and the upper level is spaced apart from it more than the lower level (Figure 4).

In the classical limit $\hbar \to 0$, one can Taylor-expand ε_{kz} around k_{zi}. Using:

$$v_z = \frac{1}{\hbar}\frac{\partial\varepsilon_{kz}}{\partial k_z}, \qquad \frac{1}{\gamma_z^2\gamma m} = \frac{1}{\hbar^2}\frac{\partial^2\varepsilon_{kz}}{\partial k_z^2}$$

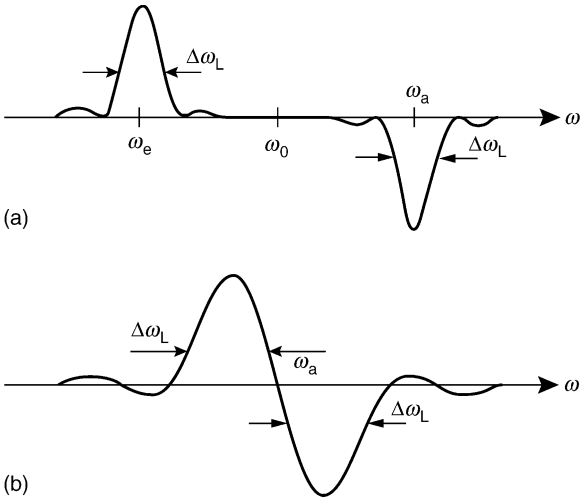


Figure 5 Net gain emission/absorption frequency lines of FEL: (a) in the quantum limit: $\omega_a - \omega_e \gg \Delta\omega_L$, (b) in the classical limit: $\omega_a - \omega_e \ll \Delta\omega_L$.

one obtains:

$$\omega_e \cong \omega_a \cong \omega_0 = v_z(q_{z0} + k_w) \qquad [17]$$

which for $q_{z0} = (\omega/c)\cos\Theta_q$ reproduces the classical synchronism condition in eqn [5]. The homogeneous

broadening linewidth is found to be:

$$\frac{\Delta\omega_L}{\omega_0} = \frac{1}{N_w} \qquad [18]$$

where $N_w = L/\lambda_w$ is the number of wiggler periods.

The classical limit condition requires that the difference between the emission and absorption line centers will be smaller than their width. This is expressed in terms of the 'recoil parameter ε' (for $\Theta_q = 0$):

$$\varepsilon = \frac{\omega_a - \omega_e}{\Delta\omega_L} = \frac{1+\beta}{\beta_z^2}\frac{\hbar\omega_0}{\gamma mc^2}N_w \ll 1 \qquad [19]$$

This condition is satisfied in all practical cases of realizable FELs. When this happens, the homogeneous line broadening dominates over the quantum-recoil effect, and the emission and absorption lines are nearly degenerate (Figure 5b). The total quantum-electrodynamic photonic emission rate expression:

$$\frac{d\nu_q}{dt} = \Gamma_{sp}[(\nu_q+1)F(\omega-\omega_e) - \nu_q F(\omega-\omega_a)] \qquad [20]$$

reduces then into:

$$\frac{d\nu_q}{dt} = \nu_q\Gamma_{sp}\varepsilon\Delta\omega_L\frac{d}{d\omega}F(\omega-\omega_e) + \Gamma_{sp}F(\omega-\omega_0) \qquad [21]$$

Here ν_q is the number of photons in radiation mode q, Γ_{sp} – the spontaneous emission rate, and $F(\omega-\omega_0)$ is the emission (absorption) lineshape function. Figure 5b depicts the transition of the net radiative emission/absorption rate into a gain curve which is proportional to the derivative of the spontaneous emission lineshape function (first term in eqn [21]). Equation [21] presents a fundamental relation between the spontaneous and stimulated emission of FELs, which was observed first by John Madey (Madey's theorem). It can be viewed as an extension of Einstein's relations to a classical radiation source.

The Classical Picture

The spontaneous emission process of FEL (synchrotron undulator radiation) is nothing but dipole radiation of the undulating electrons, which in the laboratory frame of reference is Doppler shifted to high frequency. The understanding of the stimulated emission process requires a different approach.

Consider a single electron, following a sinusoidal trajectory under the effect of a planar undulator magnetic field in eqn [2] (Figure 1):

$$v_x = v_w \cos k_w z_e(t) \qquad [22]$$

$$x = x_w \sin k_w z_e(t) \qquad [23]$$

where $v_w = ca_w/\gamma$ and $x_w = v_w/(v_z k_w)$. An electromagnetic wave $E_x(z,t) = E_0\cos(\omega t - k_z z)$ propagates collinearly with the electron. Figure 6 displays the electron and wave 'snap-shot' positions as they propagate along one wiggler period λ_w.

If the electron, moving at average axial velocity v_z, enters the interaction region $z = 0$ at $t = 0$, its axial position is $z_e(t) = v_z t$, and the electric force it experiences is $-eE_x(z_e(t),t) = -eE_0\cos(\omega - k_z v_z)t$. Clearly, this force is (at least initially at $t = 0$) opposite to the transverse velocity of the electron $v_x = v_w\cos(k_w v_z)t$ (imply in deceleration) and the power exchange rate $-ev_e\cdot E = -ev_x E_x$ corresponds to transfer of energy into the radiation field on account of the electron kinetic energy. Because the phase velocity of the radiation mode is larger than the electron velocity, $v_{ph} = \omega/k_z > v_z$, the electron phase $\varphi_e = (\omega - k_z v_z)t$ grows, and the power exchange rate $-ev_x E_x$ changes. However, if one synchronizes the electron velocity, so that while the electron traverses one wiggler period ($t = \lambda_w/v_z$), the electron phase advances by $2\pi: (\omega - k_z v_z)\cdot\lambda_w/v_z = 2\pi$, then the power exchange rate from the electron to the wave remains non-negative through the entire interaction length, because then the electron transverse velocity

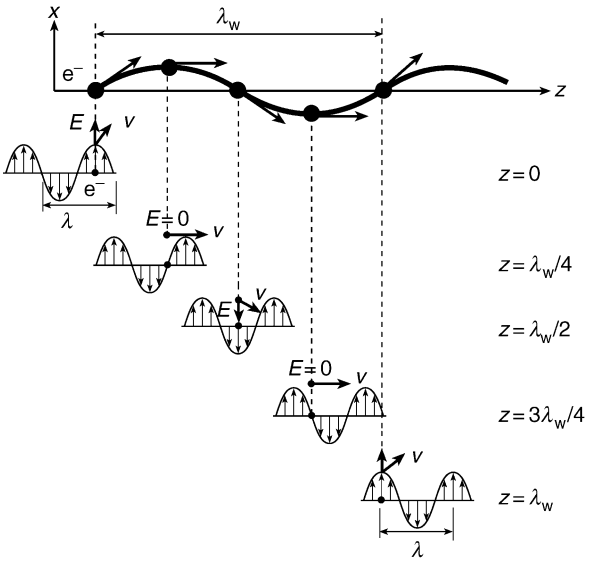


Figure 6 'Snapshots' of an electromagnetic wave period slipping relative to an undulating electron along one wiggler period λ_w. The energy transfer to the wave – $e\mathbf{v}\cdot\mathbf{E}$ remains non-negative all along.

v_x and the wave electric field E_x reverse sign, at each period exactly at the same points $(\lambda_w/4, 3\lambda_w/4)$. This situation is depicted in Figure 6, which shows the slippage of the wave crests relative to the electron at five points along one wiggler period. The figure describes the synchronism condition, in which the radiation wave slips one optical period ahead of the electron, while the electron goes through one wiggle motion. In all positions along this period, $v \cdot E \geq 0$ (in a helical wiggler and a circularly polarized wave this product is constant and positive $v \cdot E > 0$ along the entire period). Substituting $\lambda_w = 2\pi/k_w$, this phase synchronism condition may be written as:

$$\frac{\omega}{v_z} = k_z + k_w \qquad [24]$$

which is the same as eqns [17] and [5].

Figure 6 shows that a single electron (or a bunch of electrons of duration smaller than an optical period) would amplify a co-propagating radiation wave, along the entire wiggler, if it satisfies the synchronism condition in eqn [24] and enters the interaction region ($z = 0$) at the right (decelerating) phase relative to the radiation field. If the electron enters at the opposite phase it accelerates (on account of the radiation field energy which is then attenuated by 'stimulated absorption'). Thus, when an electron beam is injected into a wiggler at the synchronism condition with electrons entering at random times, no net amplification or absorption of the wave is expected on the averages. Hence, some more elaboration is required, in order to understand how stimulated emission gain is possible then.

Before proceeding, it is useful to define the 'pondermotive force' wave. This force originates from the nonlinearity of the Lorenz force equation:

$$\frac{d}{dt}(\gamma m \mathbf{v}) = -e(\mathbf{E} \times \mathbf{v} \times \mathbf{B}) \qquad [25]$$

At zero order (in terms of the radiation fields), the only field force on the right-hand side of eqn [25] is due to the strong wiggler field (eqns [2] and [3]), which results in the transverse wiggling velocity (eqn [22] for a linear wiggler). When solving next eqn [25] to first order in terms of the radiation fields:

$$\begin{aligned} \mathbf{E}_s(r, \mathbf{t}) &= \mathrm{Re}\left[\tilde{\mathbf{E}}_s \, e^{i(k_z z - \omega t)}\right] \\ \mathbf{B}_s(r, \mathbf{t}) &= \mathrm{Re}\left[\tilde{\mathbf{B}}_s \, e^{i(k_z z - \omega t)}\right] \end{aligned} \qquad [26]$$

the cross product $\mathbf{v} \times \mathbf{B}$ between the transverse components of the velocity and the magnetic field generates a longitudinal force component:

$$\mathbf{F}_{pm}(z, t) = \mathrm{Re}\left[\hat{\mathbf{e}}_z \tilde{F}_{pm} e^{i(k_z + k_w)z - i\omega t}\right] \qquad [27]$$

that varies with the beat wavenumber $k_s + k_w$ at slow phase velocity $(v_{ph} = \omega/(k_z + k_w) < c)$. This slow force-wave is called the pomdermotive (PM) wave. Assuming the signal radiation wave in eqn [26] is polarization-matched to the wiggler (linearly polarized or circularly polarized for a linear or helical wiggler respectively), the PM force amplitude is given by:

$$|\tilde{F}_{pm}| = e|\tilde{E}_{s\perp}|a_w/\gamma\beta_z \qquad [28]$$

With large enough k_w, it is always possible to slow down the phase velocity of the pondermotive wave until it is synchronized with the electron velocity:

$$v_{ph} = \frac{\omega}{k_z + k_w} = v_z \qquad [29]$$

and can apply along the interaction length a decelerating axial force, that will cause properly phased electrons to transfer energy to the wave on account of their longitudinal kinetic energy.

This observation is of great importance. It reveals that even though the main components of the wiggler and radiation fields are transverse, the interaction is basically longitudinal. This puts the FEL on an equal footing with the slow-wave structure devices as the TWT and the Cerenkov–Smith–Purcell FELs, in which the longitudinal interaction takes place with the longitudinal electric field component of a slow TM radiation mode. The synchronism condition in eqn [29] between the pondermotive wave and the electron, which is identical with the phase-matching condition in eqn [24], is also similar to the synchronism condition between an electron and a slow electromagnetic wave (eqn [14]).

Using the pondermotive wave concept, we can now explain the achievement of gain in the FEL with a random electron beam. Figure 7 illustrates the interaction between the pondermotive wave and electrons, distributed at the entrance ($z = 0$) randomly within the wave period. Figure 7a shows 'snap-shots' of the electrons in one period of the pondermotive wave $\lambda_{pm} = 2\pi/(k_z + k_w)$ at different points along the wiggler, when it is assumed that the electron beam is perfectly synchronous with the pondermotive wave $v_{ph} = v_0$. As explained before, some electrons are slowed down, acquiring negative velocity increment Δv. However, for each such electron, there is another one, entering the wiggler at an accelerating phase of the wave, acquiring the same positive velocity increment Δv. There is then no net change in the energy of the e-beam or the wave, however, there is clearly an effect of 'velocity-bunching' (modulation), which turns along the wiggler into 'density-bunching' at the same

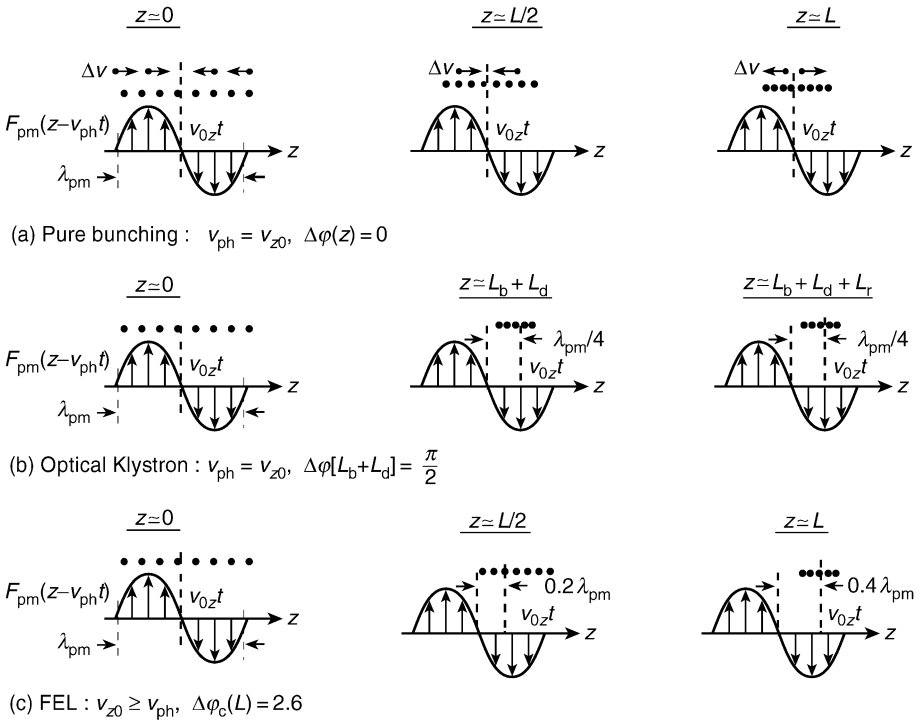


Figure 7 'Snapshots' of a pondermotive wave period interacting L with an initially uniformly distributed electron beam taking place respectively along the interaction length $0 < z < L_b$. (a) Exact synchronism in a uniform wiggler (bunching). (b) Energy bunching, density bunching and radiation are in the energy buncher, dispersive magnet $L_b < z < L_b + L_d$ and radiating wiggler $L_b + L_d < z < L_b + L_d + L_r$ sections of an Optical–Klystron. (c) Slippage from bunching phase to radiating phase at optimum detuning off synchronism in a uniform wiggler FEL.

frequency ω and wavenumber $k_z + k_w$ as the modulating pondermotive wave. The degree of density-bunching depends on the amplitude of the wave and the interaction length L. In the nonlinear limit the counter propagating (in the beam reference frame) velocity modulated electrons may over-bunch namely cross over, and debunch again.

Bunching is the principle of classical stimulated emission in electron beam radiation devices. If the e-beam had been prebunched in the first place, we would have injected it at a decelerating phase relative to the wave and obtained net radiation gain right away (super radiant emission). This is indeed the principle behind the 'Optical-Klystron' (OK) demonstrated in Figure 7b. The structure of the OK is described ahead in Figure 19. The electron beam is velocity (energy) modulated by an input electromagnetic radiation wave in the first 'bunching-wiggler section' of length L_b. It then passes through a drift-free 'energy-dispersive magnet section' (chicane) of length L_d, in which the velocity modulation turns into density bunching. The bunched electron beam is then injected back into a second 'radiating-wiggler section', where it co-propagates with the same electromagnetic wave but with a phase advance of $\pi/2 - m2\pi$, $m = 1, 2, \ldots$ (spatial lag of $\lambda_{pm}/4 - m\lambda_{pm}$ in real space) which places the entire bunch at a decelerating phase relative to the PM-wave and so amplifies the radiation wave.

The principle of stimulated-emission gain in FEL, illustrated in Figure 7c, is quite similar. Here the wiggler is uniform along the entire length L, and the displacement of the electron bunches into a decelerating phase position relative to the PM-wave is obtained by injecting the electron beam at a velocity v_{z0}, slightly higher than the wave v_{ph} (velocity detuning). The detailed calculation shows that detuning corresponding to a phase shift of $\Delta\Psi(L) = [(\omega/v_{z0}) - (k_z + k_w)]L = -2.6$ (corresponding to spatial bunch advance of $0.4\lambda_{pm}$ along the wiggler length), provides sufficient synchronism with the PM-wave in the first half of the wiggler to obtain bunching, and sufficient deceleration-phasing of the created bunches in the second part of the wiggler to obtain maximum gain.

Principles of FEL Theory

The 3D radiation field in the interaction region can be expanded in general in terms of a complete set of

free-space or waveguide modes $\{\varepsilon_q(x,y)\mathcal{H}_q(x,y)\}$:

$$E(\boldsymbol{r},t) = \mathrm{Re}\Big[\sum_q c_q(z)\varepsilon_q(x,y)\mathrm{e}^{\mathrm{i}(k_{zq}z-\omega t)}\Big] \qquad [30]$$

The mode amplitudes $C_q(z)$ may grow along the wiggler interaction length $0 < z < L$, according to the mode excitation equation:

$$\frac{\mathrm{d}}{\mathrm{d}z}C_q(z) = -\frac{1}{4\mathcal{P}}\mathrm{e}^{-\mathrm{i}k_{zq}}\int\int \mathbf{J}(x,y,z)\cdot\boldsymbol{\varepsilon}_q^*(x,y)\mathrm{d}x\,\mathrm{d}y \qquad [31]$$

where $\mathcal{P} = -\frac{1}{2}\mathrm{Re}\int\int \boldsymbol{\varepsilon}_q \times \mathcal{H}_q\cdot\hat{e}_z\,\mathrm{d}x\,\mathrm{d}y$ is the mode normalization power, and $\tilde{\mathbf{J}}$ is the bunching current component at frequency ω, that is phase matched to the radiation waves, and needs to be calculated consistently from the electron force equations.

The FEL Small Signal Regime

We first present the basic formulation of FEL gain in the linear (small signal) regime, namely the amplified radiation field is assumed to be proportional to the input signal radiation field, and the beam energy loss is negligible. This is done in the framework of a one-dimensional (single transverse radiation mode) model.

The electron beam charge density, current density, and velocity modulation are solved in the framework of a one-dimensional plasma equations model (kinetic or fluid equations). The longitudinal PM-force in eqn [27] modulates the electron beam velocity via the longitudinal part of the force eqn [25]. This brings about charge modulation $\rho(z,t) = \mathrm{Re}[\tilde{\rho}(k_z+k_\mathrm{w},\omega)\mathrm{e}^{\mathrm{i}(k_z+k_\mathrm{w})z-\mathrm{i}\omega t}]$ and consequently, also longitudinal space-charge field $\tilde{E}_\mathrm{sc}(k_{zq}+k_\mathrm{w},\omega)$ and longitudinal current density modulation $\tilde{J}_z(k_z+k_\mathrm{w},\omega)$, related through the Poison and continuity equations:

$$\mathrm{i}(k_z+k_\mathrm{w})\tilde{E}_\mathrm{sc}(k_z+k_\mathrm{w},\omega) = \tilde{\rho}(k_z+k_\mathrm{w},\omega)/\varepsilon \qquad [32]$$

$$(k_z+k_\mathrm{w})\tilde{J}_z(k_z+k_\mathrm{w},\omega) = \omega\tilde{\rho}(k_z+k_\mathrm{w},\omega) \qquad [33]$$

Solving the force eqn [25] for a general longitudinal force $F_z(z,t) = \mathrm{Re}[\tilde{F}_z(k_z,\omega)\mathrm{e}^{\mathrm{i}(k_z z-\omega t)}]$ results in a linear longitudinal current response relation:

$$\tilde{J}_z(k_z,\omega) = -\mathrm{i}\omega\chi_\mathrm{p}(k_z,\omega)\tilde{F}_z(k_z,\omega)/(-e) \qquad [34]$$

where $\chi_\mathrm{p}(k_z,\omega)$ is the longitudinal susceptibility of the electron beam 'plasma'.The beam charge density in the FEL may be quite high, and consequently the space charge field $\tilde{E}_\mathrm{sc}$, arising from the Poison eqn [32], may not be negligible. One should take into consideration then that the total longitudinal force $\tilde{F}_z$ is composed of both the PM-force of eqn [27] and an arising longitudinal space-charge electric force $-e\tilde{E}_\mathrm{sc}$. Thus, one should substitute in eqn [34]:

$$\tilde{F}_z(k_z+k_\mathrm{w},\omega) = -e[\tilde{E}_\mathrm{pm}(k_z+k_\mathrm{w},\omega) + \tilde{E}_\mathrm{sc}(k_z+k_\mathrm{w},\omega)] \qquad [35]$$

and solve it self-consistently with eqns [32] and [33] to obtain the 'external-force' response relation:

$$\tilde{J}_z(k_z+k_\mathrm{w},\omega) = \frac{-\mathrm{i}\omega\chi_\mathrm{p}(k_z+k_\mathrm{w},\omega)}{1+\chi_\mathrm{p}(k_z+k_\mathrm{w},\omega)/\varepsilon}\tilde{E}_\mathrm{pm}(k_z+k_\mathrm{w},\omega) \qquad [36]$$

where we defined the PM 'field': $\tilde{E}_\mathrm{pm} = \tilde{F}_\mathrm{pm}/(-e)$.

In the framework of a single-mode interaction model, we keep in the summation of eqn [30] only one mode q (usually the fundamental mode, and in free space – a Gaussian mode). The transverse current density components in eqn [31] $\tilde{\boldsymbol{J}}_\perp = \frac{1}{2}\tilde{\rho}\,\tilde{\boldsymbol{v}}_w$ are found using eqns [22], [33], and [36]. Finally, substituting $C_q(z) = \tilde{C}_q\,\mathrm{e}^{\mathrm{i}\delta k z}$ (where $\delta k \equiv k_z - k_{qz}$ and $k_z \cong k_{qz}$ is the wavenumber of the radiation wave modified by the interaction with the electrons) results in the general FEL dispersion relation:

$$(k_z - k_{zq})[1+\chi_\mathrm{p}(k_z+k_\mathrm{w},\omega)/\varepsilon] = \kappa\chi_\mathrm{p}(k_z+k_\mathrm{w},\omega)/\varepsilon_0 \qquad [37]$$

Equation [37] is a general expression, valid for a wide variety of FELs, including Cerenkov–Smith–Purcell and TWT. They differ only in the expression for κ. For the conventional (magnetic wiggler) FEL:

$$\kappa = \frac{1}{4}\frac{A_\mathrm{e}}{A_\mathrm{em}}\frac{a_\mathrm{w}^2}{\gamma^2\beta_z^2}\frac{\omega}{c}A_\mathrm{JJ}^2 \qquad [38]$$

where A_e is the cross-section area of the electron beam, and $A_\mathrm{em} \equiv \mathcal{P}_q/[\frac{1}{2}\sqrt{\varepsilon_0/\mu_0}|\varepsilon_{q\perp}(0,0)|^2]$ is the effective area of the interacting radiation-mode q, and it is assumed that the electron beam, passing on axis $(x_\mathrm{e},y_\mathrm{e}) = (0,0)$, is narrow relative to the transverse mode variation $A_\mathrm{e}/A_\mathrm{em} \ll 1$. The 'Bessel-functions coefficient' A_JJ is defined for a linear wiggler only, and is given by:

$$A_\mathrm{JJ} = J_0\left[\frac{a_\mathrm{w}^2}{2(a_\mathrm{w}^2+2)}\right] - J_1\left[\frac{a_\mathrm{w}^2}{2(a_\mathrm{w}^2+2)}\right] \qquad [39]$$

In a helical wiggler, $A_\mathrm{JJ} \equiv 1$. Usually $a_\mathrm{w} \ll 1$, and therefore $A_\mathrm{JJ} \cong 1$.

The Pierce Dispersion Equation

The longitudinal plasma response susceptibility function $\chi_\mathrm{p}(k_z,\omega)$ has been calculated, in any plasma formulation, including fluid model, kinetic model, or

even quantum-mechanical theory. If the electron beam axial velocity spread is small enough (cold beam), then the fluid plasma equations can be used. The small signal longitudinal force equation derived from eqn [25], together with eqn [33] and the small signal current modulation expression:

$$\tilde{J}_z \cong \rho_0 \tilde{v}_z + v_z \tilde{\rho} \qquad [40]$$

result in:

$$\chi_p(k_z, \omega) = -\frac{\omega_p'^2}{(\omega - k_z v_z)^2}\varepsilon \qquad [41]$$

where $\omega_p' = (e^2 n_0/\gamma\gamma_z^2 \varepsilon m)^{1/2}$ is the longitudinal plasma frequency, n_0 is the beam electrons density ($\rho_0 = -en_0$), v_z is the average axial velocity of the beam.

In this 'cold-beam' limit, the FEL dispersion eqn [37] reduces into the well-known 'cubic dispersion equation' derived first by John Pierce in the late 1940s for the TWT:

$$\delta k(\delta k - \theta - \theta_{pr})(\delta k - \theta + \theta_{pr}) = Q \qquad [42]$$

where $\delta k = k_z - k_{zq}$, θ is the detuning parameter (off the synchronism condition of eqn [24]):

$$\theta \equiv \frac{\omega}{v_z} - k_{zq} - k_w \qquad [43]$$

$$\theta_p = \frac{\omega_p}{v_z} \qquad [44]$$

$$Q = \kappa\theta_p^2 \qquad [45]$$

Here $\theta_{pr} = r_p\theta_p$, where $r_p < 1$ is the plasma reduction factor. It results from the reduction of the longitudinal space-charge field $\tilde{E}_{sp}$ in a beam of finite radius r_b due to the fringe field effect ($r_p \to 1$ when the beam is wide relative to the longitudinal modulation wavelength: $r_b \gg \lambda_{pm} = 2\pi/(k_{zq} + k_w)$).

The cubic equation has, of course, three solutions δk_i ($i = 1, 2, 3$), and the general solution for the radiation field amplitude and power is thus:

$$C_q(z) = \sum_{j=1}^{3} A_j\, e^{i\delta k_j z} \qquad [46]$$

$$P(z) = |C_q(z)|^2 \mathcal{P}_q \qquad [47]$$

The coefficients A_i can be determined from three initial conditions of the radiation and e-beam parameters $C_q(0)$, $\tilde{v}(0)$, $\tilde{i}(0)$, and can be given as a linear combination of them (here $\tilde{i} = A_e\tilde{J}_z$ is the longitudinal modulation current):

$$A_j = A_j^E(\omega)C_q^\omega(0) + A_j^v(\omega)\tilde{v}^\omega(0) + A_j^i(\omega)\tilde{i}^\omega(0) \qquad [48]$$

Alternatively stated, the exit amplitude of the electromagnetic mode can in general be expressed in terms of the initial conditions:

$$C_q(L) = H^E(\omega)C_q(\omega, 0) + H^v(\omega)\tilde{v}_q(\omega, 0) + H^i(\omega)\tilde{i}(\omega, 0) \qquad [49]$$

where

$$H^{(E,v,i)}(\omega) = \sum_{j=1}^{3} A_j^{(E,v,i)}(\omega)e^{i\delta k_j L} \qquad [50]$$

In the conventional FEL, electrons are injected in randomly, and there is no velocity prebunching ($\tilde{v}(\omega, 0) = 0$) or current prebunching ($\tilde{i}(\omega, 0) = 0$) (or equivalently $\tilde{n}(\omega, 0) = 0$). Consequently, $C_q(z)$ is proportional to $C_q(0)$ and one can define and calculate the FEL small-signal single-path gain parameter:

$$G(\omega) \equiv \frac{P(L)}{P(0)} = \frac{|C_q(\omega, L)|^2}{|C_q(\omega, 0)|^2} = |H^E(\omega)|^2 \qquad [51]$$

The FEL Gain Regimes

At different physically meaningful operating regimes, some parameters in eqn [42] can be neglected relative to others, and simple analytic expressions can be found for δk_i, A_i, and consequently $G(\omega)$. It is convenient to normalize the FEL parameters to the wiggler length: $\bar{\theta} = \theta L$, $\bar{\theta}_{pr} = \theta_{pr}L$, $\bar{Q} = QL^3$. An additional figure of merit parameter is the 'thermal' spread parameter:

$$\bar{\theta}_{th} = \frac{v_{zth}}{v_z}\frac{\omega}{v_z}L \qquad [52]$$

where v_{zth} is the axial velocity spread of the e-beam (in a Gaussian velocity distribution model: $f(v_z) = \exp[(v_z - v_{z0})/v_{zth}]/\sqrt{\pi}v_{zth}$). The axial velocity spread can result out of beam energy spread or angular spread (finite 'emittance'). It should be small enough, so that the general dispersion relation of eqn [37] reduces to eqn [42] (the practical 'cold beam' regime).

Assuming now a conventional FEL ($\tilde{v}_z(0) = 0$, $\tilde{i}(0) = 0$), the single path gain in eqn [51] can be calculated. We present next this gain expression in the different regimes. The maximum values of the gain expression in the different regimes are listed in Table 1.

Low gain

This is the regime where the differential gain in a single path satisfies $G - 1 = [P(L) - P(0)]/P(0) \ll 1$. It is not useful for FEL amplifiers but most FEL oscillators operate in this regime.

Table 1 The gain regimes maximum gain expressions

	Gain regime	*Parameters domain*	*Max. gain expression*
I	Tenuous beam low-gain	$\bar{Q},\ \bar{\theta}_{pr},\ \bar{\theta}_{th} < \pi$	$\frac{P(L)}{P(0)} = 1 + 0.27\bar{Q}$
II	Collective low-gain	$\bar{\theta}_{pr} > \frac{\bar{Q}}{2},\ \bar{\theta}_{th},\ \pi$	$\frac{P(L)}{P(0)} = 1 + \bar{Q}/2\bar{\theta}_{pr}$
III	Collective high-gain	$\bar{Q}/2 > \bar{\theta}_{pr} > \bar{Q}^{1/3},\ \bar{\theta}_{th},\ \bar{Q} > \pi$	$\frac{P(L)}{P(0)} = \frac{1}{4}\exp(\sqrt{2\bar{Q}/\bar{\theta}_{pr}})$
IV	Strong coupling high-gain	$\bar{Q}^{1/3} > \bar{\theta}_{pr},\ \bar{\theta}_{th},\ \bar{Q} > \pi$	$\frac{P(L)}{P(0)} = \frac{1}{9}\exp(\sqrt{3}\bar{Q}^{1/3})$
V	Warm beam	$\bar{\theta}_{th} > \bar{\theta}_{pr},\ \bar{Q}^{1/3},\ \pi$	$\frac{P(L)}{P(0)} = \exp(3\bar{Q}/\bar{\theta}_{th}^2)$

The three solutions of eqn [42] – namely the terms of eqn [46] – are reminiscent of the three eigenwaves of the uncoupled system ($\kappa = Q = 0$) : the radiation mode and the two plasma (space-charge) waves of the e-beam (the slow and fast waves, corresponding respectively to the forward and backward propagating plasma-waves in the beam rest reference-frame). In the low-gain regime, all three terms in eqn [46] are significant. Calculating them to first order in κ, results in analytical gain expressions in the collective ($\bar{\theta}_{pr} \gg \pi$) and tenuous-beam ($\bar{\theta}_{pr} \ll \pi$) regimes (note that $\bar{\theta}_{pr}/2\pi = f'_{pr}L/v_z$ is the number of plasma oscillations within the wiggler transit time L/v_z).

In most practical situations the beam current density is small enough, and its energy high enough, to limit operation to the tenuous-beam regime. The gain curve function is then:

$$G(\omega) - 1 = \bar{Q}F(\bar{\theta}(\omega)) = \bar{Q}\frac{d}{d\bar{\theta}}\mathrm{sinc}^2(\bar{\theta}/2) \qquad [53]$$

$$\bar{\theta}(\omega) \equiv \theta(\omega)L = 2\pi\frac{\omega - \omega_0}{\Delta\omega_L} \qquad [54]$$

where $\mathrm{sinc}(u) \equiv (\sin u)/u$, and in free space (no waveguide) propagation ($k_{zq} = \omega/c$), the FWHM frequency bandwidth of the $\mathrm{sinc}^2(\bar{\theta}/2)$ function is:

$$\frac{\Delta\omega_L}{\omega_0} = \frac{1}{N_w} \qquad [55]$$

The small signal gain curve is shown in Figure 8. There is no gain at synchronism – $\omega = \omega_0$. Maximum gain – $G - 1 = 0.27\bar{Q}$, is attained at a frequency slightly smaller than ω_0, corresponding to $\bar{\theta} = -2.6$. The small gain curve bandwidth is $\Delta\omega_{SG} \cong \Delta\omega_L/2$,

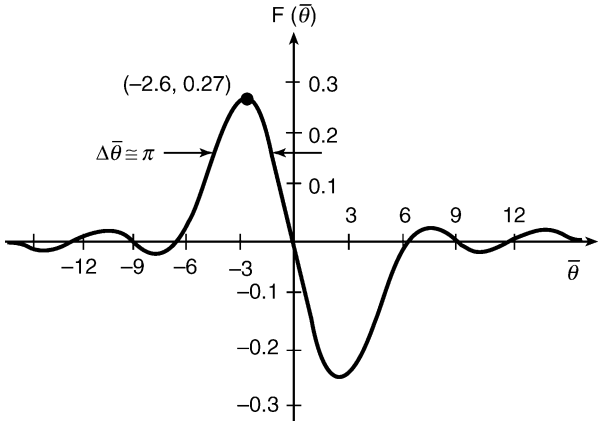


Figure 8 The low-gain cold-beam small-signal gain curve of FEL as a function of the detuning parameter $\bar{\theta}(\omega)$.

namely:

$$\frac{\Delta\omega_{SG}}{\omega_0} = \frac{1}{2N_w} \qquad [56]$$

High gain

This is the regime where the FEL gain in a single path satisfies $G = P(L)/P(0) \gg 1$. It is useful, of course, when the FEL is used as an amplifier.

Since the coefficients of the cubic eqn [42] are all real, the solutions δk_i ($i = 1, 2, 3$) must be either all real, or composed of one real solution δk_3 and two complex solutions, which are complex conjugate of each other: $\delta k_1 = \delta k_2^*$. In the first case, all terms in eqn [46] are purely oscillatory, there is no exponential growth, and the FEL operates in the low gain regime. In the second case, assuming $\mathrm{Im}(\delta k_1) < 0$, $\mathrm{Im}(\delta k_2) > 0$, the first term grows exponentially, and if L is long enough it will dominate over the other decaying ($j = 2$) and oscillatory ($j = 3$) terms, and

result in an exponential gain expression:

$$G(\omega) = \left(\frac{A_1}{A_1 + A_2 + A_3}\right)^2 \mathrm{e}^{2(\mathrm{Im}\ \delta k_1)L} \qquad [57]$$

If we focus on the tenuous-beam strong coupling (high–gain) regime $\theta_{\mathrm{pr}} \ll |\delta k|$, then the cubic eqn [42] gets the simple form:

$$\delta k\lfloor(\delta k)^2 - \theta^2\rfloor = \Gamma^3 \qquad [58]$$

where

$$\Gamma = Q^{1/3} = \left(\pi \frac{a_{\mathrm{w}}^2}{\gamma^3 \gamma_z^2 \beta_z^5} \frac{\omega/c}{A_{\mathrm{em}}} \frac{I_{\mathrm{b}}}{I_{\mathrm{A}}} A_{\mathrm{JJ}}^2\right)^{1/3} \qquad [59]$$

and $I_{\mathrm{A}} = 4\pi\varepsilon_0 m_{\mathrm{e}} c^3/e \cong 17\ \mathrm{kA}$ is the Alfven current. The solution of eqn [58] near synchronism ($\theta \cong 0$) is:

$$\delta k_1 = \frac{1 - \sqrt{3}\mathrm{i}}{2}\Gamma, \qquad \delta k_2 = \frac{1 + \sqrt{3}\mathrm{i}}{2}\Gamma, \qquad \delta k_3 = -\Gamma \qquad [60]$$

resulting in:

$$H^{\mathrm{E}}(\omega) = \frac{C_q(z)}{C_q(0)} = \frac{1}{3}\left[\mathrm{e}^{\frac{\sqrt{3}+\mathrm{i}}{2}\Gamma z} + \mathrm{e}^{\frac{-\sqrt{3}+\mathrm{i}}{2}\Gamma z} + \mathrm{e}^{-\mathrm{i}\Gamma z}\right] \qquad [61]$$

and for $\Gamma L \gg 1$:

$$G \cong \frac{1}{9}\mathrm{e}^{\sqrt{3}\Gamma L} \qquad [62]$$

The FEL gain is then exponential and can be very high. The gain exponential coefficient is characterized then by its third-order root scaling with the current, $\alpha I_{\mathrm{b}}^{1/3}$. The high-gain frequency detuning curve (found by solving eqn [58] to second order in θ) is:

$$G \cong \frac{1}{9}\mathrm{e}^{\sqrt{3}\Gamma L} \exp(-\theta^2/\Gamma L 3^{3/2}) \equiv \frac{1}{9}\mathrm{e}^{\sqrt{3}\Gamma L} \exp\left[-\frac{(\omega - \omega_0)^2}{\Delta\omega_{\mathrm{HG}}^2}\right] \qquad [63]$$

where $\Delta\omega_{\mathrm{HG}}$ is the 1/e half-width of the gain curve:

$$\frac{\Delta\omega_{\mathrm{HG}}}{\omega_0} = \frac{3^{3/4}}{2\pi} \frac{\lambda_{\mathrm{w}}}{(L/\Gamma)^{1/2}} \qquad [64]$$

Super-Radiance, Spontaneous-Emission and Self Amplified Spontaneous Emission (SASE)

Intense coherent radiation power can be generated in a wiggler or any other radiation scheme without any input radiation signal ($C_q(\omega, 0) = 0$) if the electron beam velocity or current (density) are prebunched. Namely, the injected e-beam has a frequency component $\tilde{v}(\omega)$ or $\tilde{i}(\omega)$ in the frequency range where the radiation device emits. In the case of pure density bunching ($\tilde{v}(\omega) = 0$), the coherent power emitted is found from eqns [46, 47, 49]:

$$(P_q)_{\mathrm{SR}} = P_q |H^i(\omega)|^2 |\tilde{i}(\omega, 0)|^2 \qquad [65]$$

A 'prebunched-beam FEL' emits coherent radiation based on the process of Super-radiant Emission (in the sense of Dike). Because all electrons emit in phase radiation wavepackets into the radiation mode, the resultant field amplitude is proportional in this case to the beam current I_{b} and the radiation. By contrast, spontaneous emission from a random electron beam (no bunching) is the result of incoherent superposition of the wavepackets emitted by the electrons and its power is expected to be proportional to the current I_{b}.

When the current to radiation field transfer function $H^i(\omega)$ is known, eqn [65] can be used to calculate the superradiant power, and in the high-gain regime also the amplified-superradiant power. The latter is the amplification of the superradiant radiation in the downstream sections of a long wiggler. Such unsaturated gain is possible only when the beam is partly bunched $\tilde{i}(\omega)(I_{\mathrm{b}})$ (because the FEL gain process requires enhanced bunching).

The expressions for the current to field transfer function, in the superradiant gain and the high-gain amplified superradiance limits respectively, are:

$$|H^i(\omega)| = \frac{(P_{\mathrm{pb}}/P_q)^{1/2}}{I_{\mathrm{b}}} \mathrm{sinc}(\theta L/2) \qquad [66]$$

$$\left|H^i(\omega)\right| = \frac{(P_{\mathrm{pb}}/P_q)^{1/2}}{3\Gamma L I_{\mathrm{b}}} \mathrm{e}^{(\sqrt{3}/2)\Gamma L}\ \mathrm{e}^{-(\omega-\omega_0)^2/2(\Delta\omega)_{\mathrm{HG}}^2} \qquad [67]$$

where

$$P_{\mathrm{pb}} = \frac{I_{\mathrm{b}}^2 Z_q}{32}\left(\frac{a_{\mathrm{w}}}{\gamma\beta_z}\right)^2 \frac{L^2}{A_{\mathrm{em}}} \qquad [68]$$

Z_q is the radiation mode impedance (in free-space $Z_q = \sqrt{\mu_0/\varepsilon_0}$). From these expressions one can calculate the power and spectral power of both coherent (superradiant) and partially coherent (spontaneous emission) radiation of FEL in the negligible gain and high gain regimes. The corresponding super-radiant power is in the negligible superradiance gain limit:

$$P_{\mathrm{SR}} = P_{\mathrm{pb}}\left|\frac{\tilde{i}(\omega)}{I_{\mathrm{b}}}\right|^2 \mathrm{sinc}^2(\theta L/2) \qquad [69]$$

(proportional, as expected, to the modulation current squared) and in the high-gain amplified superradiance limit (assuming initial partial bunching $|i(\omega)/I\mathrm{b}| \ll 1$):

$$P_{\mathrm{SR}} = P_{\mathrm{pb}}\left|\frac{\tilde{i}(\omega)}{I_{\mathrm{b}}}\right|^2 \frac{1}{9(\Gamma L)^2}\mathrm{e}^{\sqrt{3}\Gamma L}\,\mathrm{e}^{-(\omega-\omega_0)^2/(\Delta\omega)^2_{\mathrm{HG}}} \qquad [70]$$

The discussion is now extended to incoherent (or partially coherent) spontaneous emission. Due to its particulate nature, every electron beam has random frequency components in the entire spectrum (shot noise). Consequently, incoherent radiation power is always emitted from an electron-beam passing through a wiggler, and its spectral-power can be calculated through the relation:

$$\frac{\mathrm{d}P_q}{\mathrm{d}\omega} = \frac{2}{\pi}\mathcal{P}_q|H^i(\omega)|^2\frac{\langle|\tilde{i}(\omega)|^2\rangle}{T} \qquad [71]$$

Here $\tilde{i}(\omega)$ is the Fourier transform of the current of randomly injected electrons $i(t) = -e\sum_{j=1}^{N_T}\delta(t - t_{oj})$, where N_T is the average number of electrons in a time period T, namely, the average (DC) current is $I_{\mathrm{b}} = -eN_T/T$. For a randomly distributed beam, the shot noise current is simply $\langle|i(\omega)|^2\rangle/T = eI_{\mathrm{b}}$, and therefore the spontaneous emission power of the FEL, which is nothing but the 'synchrotron-undulator radiation', is given by (see eqn [66]):

$$\frac{\mathrm{d}P_q}{\mathrm{d}\omega} = \frac{1}{16\pi}eI_{\mathrm{b}}Z_q\frac{L^2}{A_{\mathrm{em}}}\left(\frac{a_{\mathrm{w}}}{\gamma\beta_z}\right)^2\mathrm{sinc}^2(\theta L/2) \qquad [72]$$

If the wiggler is long enough, the spontaneous emission emitted in the first part of the wiggler can be amplified by the rest of it (SASE). In the high-gain limit (see eqn [67]), the amplified spontaneous emission power within the gain bandwidth of eqn [64] is given by:

$$P_q = \frac{2}{\pi}\mathcal{P}_q eI_{\mathrm{b}}\int_0^\infty |H^i(\omega)|^2\mathrm{d}\omega = \frac{1}{9}P_{\mathrm{sh}}\,\mathrm{e}^{\sqrt{3}\Gamma L} \qquad [73]$$

where P_{sh} is an 'effective shot-noise input power':

$$P_{\mathrm{sh}} = \frac{2}{\sqrt{\pi}}\frac{eP_{\mathrm{pb}}}{I_0(\Gamma L)^2}(\Delta\omega)_{\mathrm{HG}} \qquad [74]$$

Saturation Regime

The FEL interaction of an electron with an harmonic electromagnetic (EM) wave is essentially described by the longitudinal component of the force in eqn [25], driven by the pondermotive force of eqns [27] and [28]:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\gamma_i m v_{zi}) = |\tilde{F}_{\mathrm{pm}}|\cos[\omega t - (k_z + k_{\mathrm{w}})z_i] \qquad [75]$$

$$\mathrm{d}z_i/\mathrm{d}t = v_{zi} \qquad [76]$$

As long as the interaction is weak enough (small signal regime), the change in the electron velocity is negligible – $v_{zi} \cong v_{z0}$, and the phase of the force-wave, experienced by the electron, is linear in time $\Psi_i(t) = [\omega - (k_z + k_{\mathrm{w}})v_{z0}](t - t_{0i}) + \omega t_{0i}$. Near synchronism condition $\theta \cong 0$ (eqn [24]), eqn [75] results in bunching of the beam, because different acceleration/deceleration forces are applied on each electron, depending on their initial phase $\Psi_i(0) = \omega t_{0i}(-\pi < \Psi_i(0) < \pi)$ within each optical period $2\pi/\omega$ (see Figure 7). Taylor expansion of v_{zi} around v_{z0} in eqns [75] and [76], and use of conservation of energy between the e-beam and the radiation field, would lead again to the small signal gain expression eqn [53] in the low gain regime.

When the interaction is strong enough (the nonlinear or saturation regime), the electron velocities change enough to invalidate the assumption of linear time dependence of Ψ_i and the nonlinear set of eqns [75] and [76] needs to be solved exactly.

It is convenient to invert the dependence on time $z_i(t) = \int_{t_{0i}}^{t} v_{zi}(t')\mathrm{d}t'$, and turn the coordinate z to the independent variable $t_i(z) = \int_0^z \mathrm{d}z'/v_{zi}(z') + t_{0i}$. This, and direct differentiation of $\gamma_i(v_{zi})$, reduces eqns [75] and [76] into the well-known pendulum equation:

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}z} = K_{\mathrm{s}}^2\sin\Psi_i \qquad [77]$$

$$\frac{\mathrm{d}\Psi_i}{\mathrm{d}z} = \theta_i \qquad [78]$$

where

$$\Psi_i = \int_0^z(\omega/v_{zi}(z') - k_z - k_{\mathrm{w}})\,\mathrm{d}z' \qquad [79]$$

$$\theta_i = \frac{\omega}{v_{zi}} - k_z - k_{\mathrm{w}} \qquad [80]$$

are respectively the pondermotive potential phase and the detuning value of electron i at position z.

$$K_{\mathrm{s}} = \frac{k\sqrt{a_{\mathrm{w}}a_{\mathrm{s}}A_{\mathrm{JJ}}}}{\gamma_0\gamma_{z0}\beta_{z0}^2} \qquad [81]$$

is the synchrotron oscillation wavenumber, where a_{w} is given in eqn [9], $a_{\mathrm{s}} = e|\tilde{E}|\omega/mc$, and $\gamma_0 = \gamma(0)$, $\gamma_{z0} = \gamma_z(0)$, and $\beta_{z0} = \beta_z(0)$ are the initial parameters of the assumed cold beam.

The pendulum eqns [77] and [78] can be integrated once, resulting in:

$$\tfrac{1}{2}\theta_i^2(z) - K_{\mathrm{s}}^2\cos\Psi_i(z) = C_i \qquad [82]$$

and the integration constant is determined for each electron by its detuning and phase relative to the

pondermotive wave at the entrance point ($z = 0$): $C_i = \frac{1}{2}\theta_i^2(0) - K_s^2 \cos \Psi_i(0)$.

The $\theta(z)$, $\Psi(z)$ phase-space trajectories of eqn [82] are shown in **Figure 9** for various values of C_i (corresponding to the initial conditions $\theta_i(0)$, $\Psi_i(0)$). The trajectories corresponding to $|C_i| > K_s^2$ are open; namely electrons on these trajectories, while oscillating, can slip-off out of the pondermotive–potential wave period to adjacent periods, ahead or backward, depending on the value of their detuning parameter $\bar{\theta}$. The trajectories corresponding to $|C_i| < K_s^2$ are closed, namely the electrons occupying these trajectories are 'trapped', and their phase displacement is bound to a range $|\Psi_i(z) - n\pi| < \Psi_{im} \equiv \arccos(|C_i|/K_s^2) < \pi$ within one pondermotive-wave period. The trajectory $C_i = K_s^2$ defines the 'separatrix':

$$\theta_i(z) = \pm 2K_s \cos(\Psi_i/2) \qquad [83]$$

which is sometimes referred to as the 'trap' or 'bucket'. Every electron within the separatrix stays trapped, and the ones out of it are free (untrapped). The height of the separatrix (maximum detuning swing) is $\Delta\theta_{trap} = 4K_s$. The oscillation frequency of the trapped electrons can be estimated for deeply trapped electrons ($\Psi_m \ll 2\pi$). In this case the physical pendulum eqns [77] and [78] reduce to the mathematical pendulum equation with an oscillation frequency K_s, in the z coordinate. This longitudinal oscillation, called 'synchrotron oscillation', takes place as a function of time at the 'synchrotron frequency' $\Omega_s = K_s v_z$.

Differentiation of $\theta_i(v_{zi})$ and $v_{zi}(\gamma_i)$ permits to describe the phase-space dynamics in terms of the more physical parameters $\delta v_{zi} = v_{zi} - v_{ph}$ and

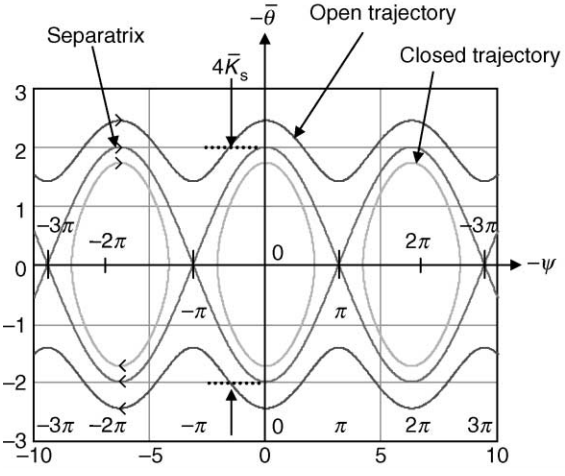


Figure 9 The (θ–Ψ) phase-space trajectories of the pendulum equation.

$\delta\gamma_i = \gamma_i - \gamma_{ph}$, where:

$$v_{ph} = \frac{\omega}{k_z + k_w} \qquad [84]$$

is the phase velocity of the pondermotive wave and $\gamma_{ph} \equiv (1 - \beta_{ph}^2)^{-1/2}$:

$$-\theta_i = \frac{\omega}{c^2\beta_{z0}^2}\delta v_{zi} = \frac{k}{\beta_{z0}^3\gamma_{z0}^2\gamma_0}\delta\gamma_i \qquad [85]$$

Figure 10 displays a typical dynamics of electron beam phase-space (γ, Ψ) evolution for the case of a cold beam of energy $\gamma(0)$ entering the interaction region at $z = 0$ with uniform phase distribution (random arrival times t_{0i}). The FEL is assumed to operate in the low-gain regime (typical situation in an FEL oscillator), and, therefore, the trap height (corresponding to $\Delta\theta_{trap} = 4k_s$):

$$\Delta\gamma_{trap} = 8\beta_z^2\gamma_z^2\gamma K_s/k \qquad [86]$$

remains constant along the interaction length. **Figure 10a** displays the e-beam phase-space evolution in the small signal regime. The uniform phase distribution evolves along the wiggler into a bunched distribution (compare to **Figure 7c**), and its average kinetic energy goes down $(\Delta E_k) = [\langle\gamma_i(L)\rangle - \gamma(0)]mc^2 < 0$, contributing this energy to the field of the interacting radiation mode, $\Delta P_q = (\Delta E_k)I_0/e$. In this case (corresponding in an FEL oscillator to the early stages of oscillation build-up), the electrons remain free (untrapped) along the entire length L.

Figure 10b displays the e-beam phase-space evolution in the large signal (saturation) regime (in the case of an oscillator – at the steady-state saturation stage). Part of the electrons are found inside the trap, immediately upon entering the interaction region ($z = 0$), and they lose energy of less than (but near) $mc^2\Delta\gamma_{trap}$ as they pass through the interaction region ($z = L$). A portion of the electrons remain outside the traps upon entrance. They follow open trajectories and lose less energy or may even become accelerated due to their interaction with the wave.

It can be appreciated from this discussion that a good design strategy in attempting to extract maximum power from the electron beam in the FEL interaction, is to set the parameters determining the synchrotron oscillation frequency K_s in eqn [81] so that only half a synchrotron oscillation period will be performed along the interaction length:

$$K_s L = \pi \qquad [87]$$

This is controlled in an amplifier by keeping the input radiation power $P_q(0)$ (and consequently a_s) small enough, so that K_s will not exceed the value set

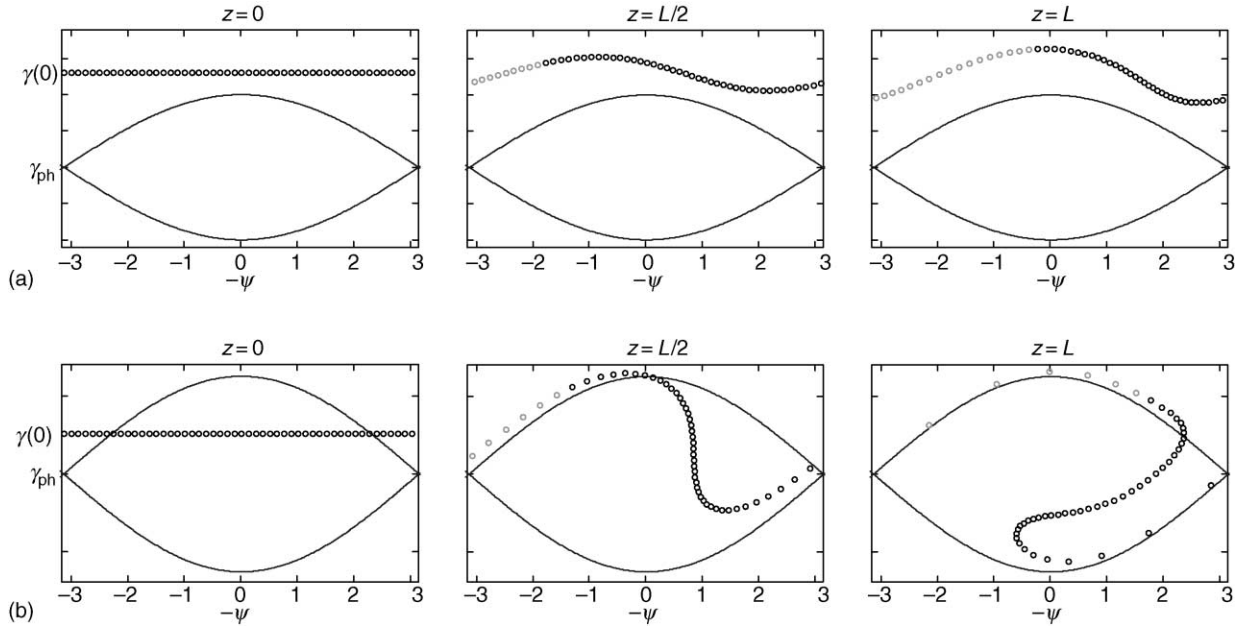


Figure 10 'Snapshots' of the ($\gamma-\Psi$) phase-space distribution of an initially uniformly distributed cold beam relative to the PM-wave trap at three points along the wiggler (a) Moderate bunching in the small-signal low gain regime. (b) Dynamics of electron beam trapping and synchrotron oscillation at steady state saturation stage of a FEL oscillator ($K_s L = \pi$).

by eqn [87]. In an oscillator, this is controlled by increasing the output mirror transmission sufficiently, so that the single path incremental small signal gain G-1 will not be much larger than the round trip loss, and the FEL will not get into deep saturation. When the FEL is over-saturated ($K_s L > \pi$), the trapped electrons begin to gain energy as they continue to rotate in their phase-space trajectories beyond the lowest energy point of the trap, and the radiative energy extraction efficiency drops down.

A practical estimate for the FEL saturation power emission and radiation extraction efficiency can be derived from the following consideration: the electron beam departs from most of its energy during the interaction with the wave, if a significant fraction of the electrons are within the trap and have positive velocity δv_{zi} relative to the wave velocity v_{ph} at $z = 0$, and if at the end of the interaction length ($z = L$), they complete half a pendulum swing and reverse their velocity relative to the wave $\delta v_{zi}(L) \cong -\delta v_{zi}(0)$. Correspondingly, in the energy phase-space diagram (Figure 10b) the electrons perform half a synchrotron oscillation swing and $\delta\gamma_i(L) = \gamma_i(L) - \gamma_{ph} = -\delta\gamma_i(0)$. In order to include in this discussion also the FEL amplifier (in the high gain regime), we note that in this case the phase velocity of the wave v_{ph} in eqn [84], and correspondingly γ_{ph}, are modified by the interaction contribution to the radiation wavenumber – $k_z = k_{z0} + \mathrm{Re}(\delta k)$, and also the electron detuning parameter (relative to the pondermotive wave) θ_i in eqn [80] differs from the beam detuning parameter θ in eqn [43]: $\theta_i = \theta - \mathrm{Re}(\delta k)$. Based on these considerations and eqn [85], the maximum energy extraction from the beam in the saturation process is:

$$\Delta\gamma = 2\delta\gamma_i(0) = 2\beta_{z0}^3\gamma_{z0}^2\gamma_0\frac{\mathrm{Re}\,\delta k - \theta}{k} \qquad [88]$$

where θ is the initial detuning parameter in eqn [43].

In an FEL oscillator, operating in general in the low-gain regime, $|\mathrm{Re}\,\delta k| \ll |\theta|$, oscillation will start usually at the resonator mode frequency, corresponding to the detuning parameter $\theta(\omega) = -2.6/L$, for which the small signal gain is maximal (see Figure 8). Then the maximum radiation extraction efficiency can be estimated directly from eqn [88]. It is, in the highly relativistic limit ($\beta_{z0} \cong 1$):

$$\eta_{ext} = \frac{\Delta\gamma}{\gamma_0} \cong \frac{1}{2N_w} \qquad [89]$$

In an FEL amplifier, in the high-gain regime $\mathrm{Re}\,\delta k = \Gamma/2 \gg |\theta|$, and consequently in the same limit:

$$\eta_{ext} \cong \frac{\Gamma\lambda_w}{4\pi} \qquad [90]$$

It may be interpreted that the effective wiggler length for saturation is $L_{eff} = 2\pi/\Gamma$.

Equation [90], derived here for a coherent wave, is considered valid also for estimating the saturation

efficiency also in SASE-FEL. In this context, it is also called 'the efficiency parameter' $-\rho$.

FEL Radiation Schemes and Technologies

Contrary to conventional atomic and molecular lasers, the FEL operating frequency is not determined by natural discrete quantum energy levels of the lasing matter, but by the synchronism condition of eqn [24] that can be predetermined by the choice of wiggler period, $\lambda_w = 2\pi/k_w$, the resonator dispersion characteristics $k_{zq}(\omega)$, and the beam axial velocity v_z.

Because the FEL design parameters can be chosen at will, its operating frequency can fit any requirement, and furthermore, it can be tuned over a wide range (primarily by varying v_z). This feature of FEL led to FEL development efforts in regimes where it is hard to attain high-power tunable conventional lasers or vacuum-tube radiation sources – namely in the sub-mm (far infrared or THz) regimes, and in the VUV down to soft X-ray wavelengths.

In practice, in an attempt to develop short wavelength FELs, the choice of wiggler period λ_w is limited by an inevitable transverse decay of the magnetic field away from the wiggler magnets surface (a decay range of $\approx k_w^{-1}$) dictated by the Maxwell equations. To avoid interception of electron beam current on the walls or on the wiggler surfaces, typical wiggler periods are made longer than $\lambda_w > 1$ cm. FELs (or FEMs – free electron masers) operating in the long wavelengths regime (mm and sub-mm wavelengths) must be based on waveguide resonators to avoid excessive diffraction of the radiation beam along the interaction length (the wiggler). This determines the dispersion relation $k_{zq}(\omega) = (\omega^2 - \omega_{coq}^2)^{1/2}/c$ where ω_{coq} is the waveguide cutoff frequency of the radiation mode q. The use of this dispersion relation in eqn [24] results in an equation for the FEL synchronism frequency ω_0. Usually the fundamental mode in an overmoded waveguide is used (the waveguide is overmoded because it has to be wide enough to avoid interception of electron beam current). In this case ($\omega_0 \gg \omega_{co}$) and certainly in the case of an open resonator (common in FELs operating in the optical regime) $k_{zq} = \omega/c$, and the synchronism condition in eqn [24] simplified to the well-known FEL radiation wavelength expression in eqn [6]:

$$\lambda = (1 + \beta_z)\beta_z\gamma_z^2\lambda_w \cong 2\gamma_z^2\lambda_w \qquad [91]$$

where γ_z, a_w are defined in eqns [7]–[9].

To attain strong interaction, it is desirable to keep the wiggler parameter a_w large (eqn [38]), however, if $a_w > 1$, this will cause reduction in the operating wavelength according to eqns [7] and [91]. For this reason, and also in order to avoid harmonic frequencies emission (in case of a linear wiggler), $a_w < 1$ in common FEL design. Consequently, considering the practical limitations on λ_w, the operating wavelength eqn [91] is determined primarily by the beam relativistic Lorentz factor γ (eqn [8]).

The conclusion is that for a short wavelength FEL, one should use an electron beam accelerated to high kinetic energy E_k. Also, tuning of the FEL operating-wavelength can be done by changing the beam energy. Small-range frequency tuning can be done also by changing the spacing between the magnet poles of a linear wiggler. This varies the magnetic field experienced by the e-beam, and effects the radiation wavelength through change of a_w (see eqns [7] and [91]).

Figure 11 displays the operating wavelengths of FEL projects all over the world versus their e-beam energy. FELs were operated or planned to operate over a wide range of frequencies, from the microwave to X-ray – eight orders of magnitude. The data points fall on the theoretical FEL radiation curve eqns [7], [8], and [91].

FEL Accelerator Technologies

The kind of accelerator used is the most important factor in determining the FEL characteristics. Evidently, the higher the acceleration energy, the shorter is the FEL radiation wavelength. However, not only the acceleration beam energy determines the shortest operating wavelength of the FEL, but also the e-beam quality. If the accelerated beam has large energy spread, energy instability, or large emittance (the product of the beamwidth with its angular spread), then it may have large axial velocity spread v_{zth}. At high frequencies, this may push the detuning spread parameter $\bar{\theta}_{th}$ (eqn [52]) to the warm beam regime (see Table 1), in which the FEL gain is diminished, and FELs are usually not operated.

Other parameters of the accelerator determine different characteristics of the FEL. High current in the electron beam enables higher gain and higher power operation. The e-beam pulse shape (or CW) characteristics, affect, of course, the emitted radiation waveform, and may also affect the FEL gain and saturation characteristics. The following are the main accelerator technologies used for FEL construction. Their wavelength operating-regimes (eqn [91]) (determined primarily by their beam acceleration energies), are displayed in Figure 12.

Modulators and pulse-line accelerators

These are usually single pulse accelerators, based on high voltage power supplies and fast discharge stored

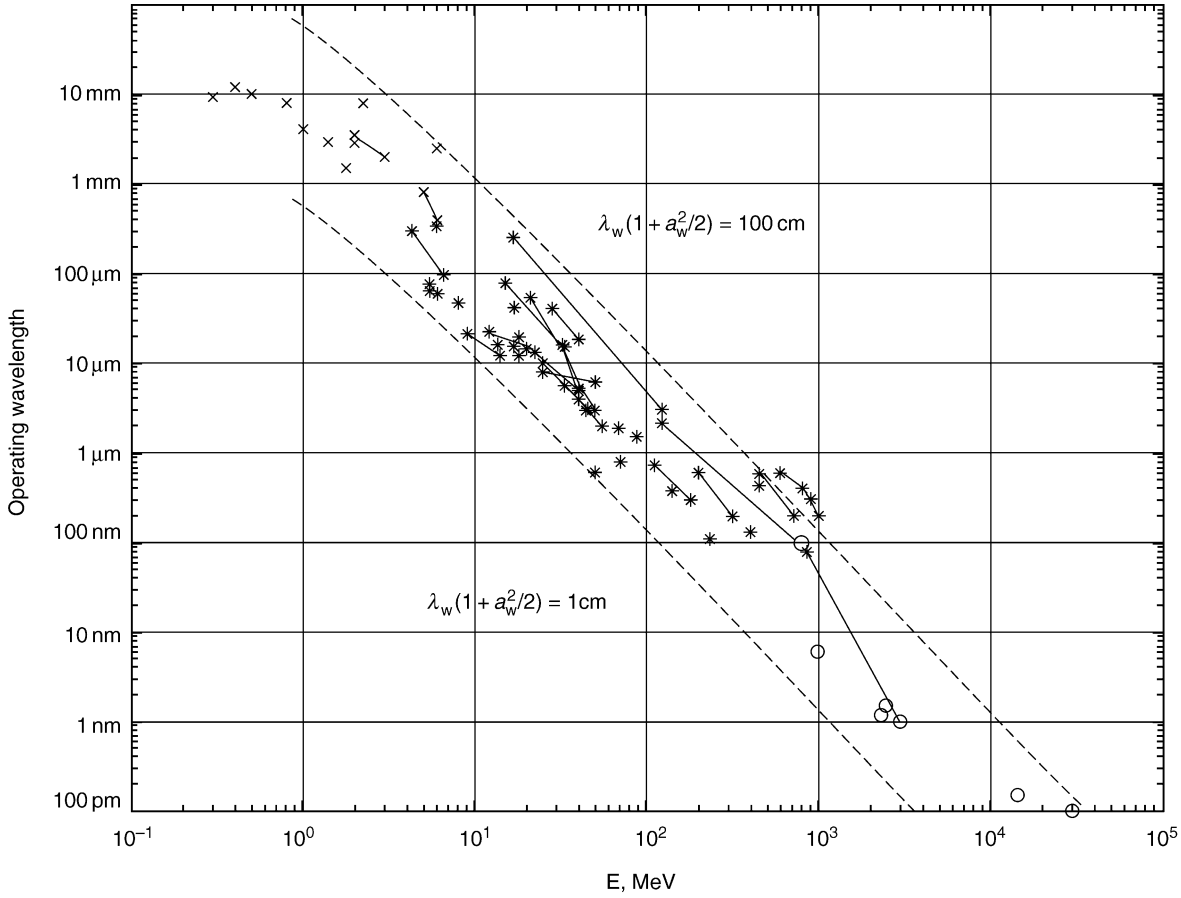


Figure 11 Operating wavelengths of FELs around the world vs. their accelerator beam energy. The data points correspond in ascending order of accelerator energy to the following experimental facilities: NRL (USA), IAP (Russia), KAERI (Korea), IAP (Russia), JINR/IAP (Russia), INP/IAP (Russia), TAU (Israel), FOM (Netherlands), KEK/JAERI (Japan/Korea), CESTA (France), ENEA (Italy), KAERI-FEL (Korea), LEENA (Japan), ENEA (Italy), FIR FEL (USA), mm Fel (USA), UCSB (USA), ILE/ILT (Japan), MIRFEL (USA), UCLA-Kurchatov (USA/Russia), FIREFLY (GB), JAERI-FEL (Japan), FELIX (Netherlands), RAFEL (USA), ISIR (Japan), UCLA-Kurchatov-LANL (USA/RU), ELSA (France), CLIO (France), SCAFEL (GB), FEL (Germany), BFEL (China), KHI-FEL (Japan), FELI4 (Japan), iFEL1 (Japan), HGHG (USA), FELI (USA), MARKIII (USA), ATF (USA), iFEL2 (Japan), VISA (USA), LEBRA (Japan), OK-4 (USA), UVFEL (USA), iFEL3 (Japan), TTF1 (Germany), NIJI-IV (Japan), APSFEL (USA), FELICITAI (Germany), FERMI (Italy), UVSOR (Japan), Super-ACO (France), TTF2 (Germany), ELETTRA (Italy), Soft X-ray (Germany), SPARX (Italy), LCLS (USA), TESLA (Germany). X, long wavelengths; $*$, short wavelengths; ○, planned short wavelengths SASE-FELs. Data based in part on H. P. Freund, V. L. Granatstein, Nucl. Inst. and Methods In Phys. Res. A249, 33 (1999), W. Colson, Proc. of the 24th Int. FEL conference, Argone, III. (ed. K. J. Kim, S. V. Milton, E. Gluskin). The data points fall close to the theoretical FEL radiation condition expression (91) drawn for two practical limits of wiggler parameters.

electric energy systems (e.g., Marx Generator), which produce short pulse (tens of nSec) Intense Relativistic Beam (IRB) of energy in the range of hundreds of keV to few MeV and high instantaneous current (order of kAmp), using explosive cathode (plasma field emission) electron guns. FELs (FEMs), based on such accelerators, operated mostly in the microwave and mm-wave regimes. Because of their poor beam quality and single pulse characteristic, these FELs were, in most cases, operated only as Self Amplified Spontaneous Emission (SASE) sources, producing intense radiation beams of low coherence at instantaneous power levels in the range of 1–100 MW. Because of the high e-beam current and low energy, these FEMs operated mostly in the collective high-gain regime (see Table 1).

Some of the early pioneering work on FEMs was done in the 1970s and 1980s in the US (NRL, Columbia Univ., MIT), Russia (IAP), and France (Echole Politechnique), based on this kind of accelerators.

Induction linacs

These too are single pulse (or low repetition rate) accelerators, based on induction of electromotive potential over an acceleration gap by means of an electric-transformer circuit. They can be cascaded to high energy, and produce short pulse (tens to hundreds

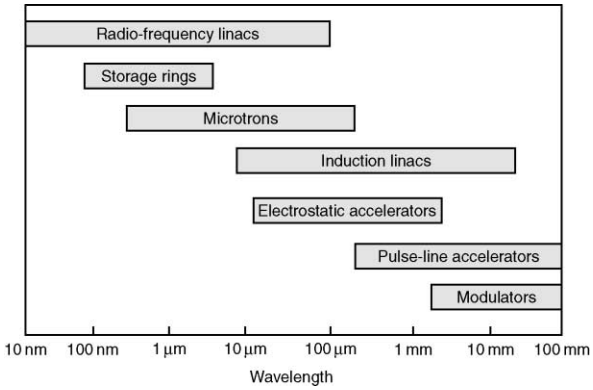


Figure 12 Approximate wavelength ranges accessible with FELs based on current accelerator and wiggler technologies.

of nSec), high current (up to 10 kA) electron beams, with relatively high energy (MeV to tens of MeV). The interest in FELs, based on this kind of accelerator technology, stemmed in the 1980s primarily from the SDI program, for the propose of development of a DEW FEL. The main development of this technology took place on a 50 MeV accelerator – ATA (for operating at 10 μm wavelength) and a 3.5 MeV accelerator – ETA (for operating at 8 mm wavelength). The latter experiment, operating in the high-gain regime, demonstrated record high power (1 GW) and energy extraction efficiency (35%).

Electrostatic accelerators

These accelerators are DC machines, in which an electron beam, generated by a thermionic electron-gun (typically 1–10 Amp) is accelerated electrostatically. The charging of the high voltage terminal can be done by mechanical charge transport (Van de Graaff) or electrodynamically (Crockford–Walton accelerator, Dynamitron). The first kind can be built at energies up to 25 MeV, and the charging current is less than mAmp. The second kind have terminal voltage less than 5 MeV, and the charging current can be hundreds of mAmps.

Because of their DC characteristics, FELs based on these kinds of accelerators can operate at arbitrary pulse shape structure and in principle – continuously (CW). However, because of the low charging current, the high electron beam current (1–10 Amp), required for FEL lasing must be transported without any interception along the entire way from the electron gun, through the acceleration tubes and the FEL wiggler, and then decelerated down to the voltage depressed beam-collector (multistage collector), closing the electric circuit back to the e-gun (current recirculation). The collector is situated at the e-gun potential, biased by moderate voltage high current power supplies, which deliver the current and power needed for circulating the e-beam and compensates for its kinetic energy loss in favor of the radiation field in the FEL cavity. This beam current recirculation is, therefore, also an 'Energy retrieval' scheme, and can make the overall energy transfer efficiency of the electrostatic-accelerator FEL very high.

In practice, high-beam transport efficiency in excess of 99.9% is needed for CW lasing, and has not been demonstrated yet. To avoid HV-terminal voltage drop during lasing, electrostatic-accelerator FELs are usually operated in a single pulse mode. Few FELs of this kind have been constructed. The first and main facility is the UCSB FEL shown in Figure 13. It operates in the wavelength range of 30 μm to 2.5 mm (with three switchable wigglers) in the framework of a dedicated radiation user facility. This FEL operates in the negatively charged terminal mode, in which the e-gun and collector are placed in the negatively charged HV-terminal inside the pressurized insulating gas tank, and the wigglers are situated externally at ground potential. An alternative operating mode of positively charged terminal internal cavity electrostatic-accelerator FEM was demonstrated in the Israeli Tandem–Accelerator FEM and the Dutch F.O.M. Fusion-FEM projects. This configuration enables operating with long pulse, high coherence, and very high average power. Linewidth of $\Delta\omega/\omega \cong 10^{-5}$ was demonstrated in the Israeli FEM and high power (730 kW over few microseconds) was demonstrated in the Dutch FEM, both at mm-wavelengths. The goal of the latter development project (which was not completed) was quasicontinuous operation at 1 MW average power for application in fusion plasma heating.

Radio-frequency (RF) accelerators

RF-accelerators are by far the most popular electron-beam sources for FELs. In RF accelerators, short electron beam bunches (bunch duration 1–10 pSec) are accelerated by the axial field of intense RF radiation (frequency about 1 GHz), which is applied in the acceleration cavities on the injected short e-beam bunches, entering with the accelerating-phase of the RF field. In microtrons, the electron bunches perform circular motion, and get incremental acceleration energy every time they re-enter the acceleration cavity. In RF-LINACs (linear accelerator), the electron bunches are accelerated in a sequence of RF cavities or a slow-wave structure, which keep an accelerating-phase synchronization of the traversing electron bunches with the RF field along a long linear acceleration length. The bunching of the electrons, prior to the acceleration step, is traditionally performed by bunching RF-cavities and a dispersive magnet (chicane) pulse compression system.

The FEL small signal gain, must be large enough to build-up the radiation field in the resonator from noise to saturation well within the macropulse duration.

RF-Linacs are essential facilities in synchrotron radiation centers, used to inject electron beam current into the synchrotron storage ring accelerator from time to time. Because of this reason, many FELs based on RF-LINACs were developed in synchrotron centers, and provide additional coherent radiation sources to the synchrotron radiation center users.

Figure 15 displays FELIX – a RF-LINAC FEL which is located in one of the most active FEL radiation user-centers in FOM – Holland.

Storage rings

Storage rings are circular accelerators in which a number of electron (or positron) beam bunches (typically of 50–500 pS pulse duration and hundreds of ampere peak current) are circulated continuously by means of a lattice of bending magnets and quadrupole lenses. Typical energies of storage ring accelerators are in the hundreds of MeV to GeVs range. As the electrons pass through the bending magnets, they lose a small amount of their energy due to emission of synchrotron radiation. This energy is replenished by a small RF acceleration cavity placed in one section of the ring. The electron beam bunch dimensions, energy spread, and emittance parameters are set in steady state by a balance between the electrons oscillations within the ring lattice and radiation damping due to the random synchrotron emission process. This produces high-quality (small emittance and energy spread) continuous train of electron beam bunches, that can be used to drive a FEL oscillator placed as an insertion device in one of the straight sections of the ring between two bending magnets.

Demonstrations of FEL oscillators, operating in a storage ring, were first reported by the French (LURE-Orsay) in 1987 (at visible wavelengths) and the Russians (VEPP-Novosibirsk) in 1988 (in the ultra-violet). The short wavelength operation of storage-ring FELs is facilitated by the high energy, low emittance and low energy spread parameters of the beam.

Since storage ring accelerators are at the heart of all synchrotron radiation centers, one could expect that FEL would be abundant in such facilities as insertion devices. There is, however, a problem of interference of the FEL operating as an insertion device in the normal operation of the ring itself. The energy spread increase, induced in the electron beam during the interaction in a saturated FEL oscillator, cannot be controlled by the synchrotron radiation damping process, if the FEL operating power is too high.

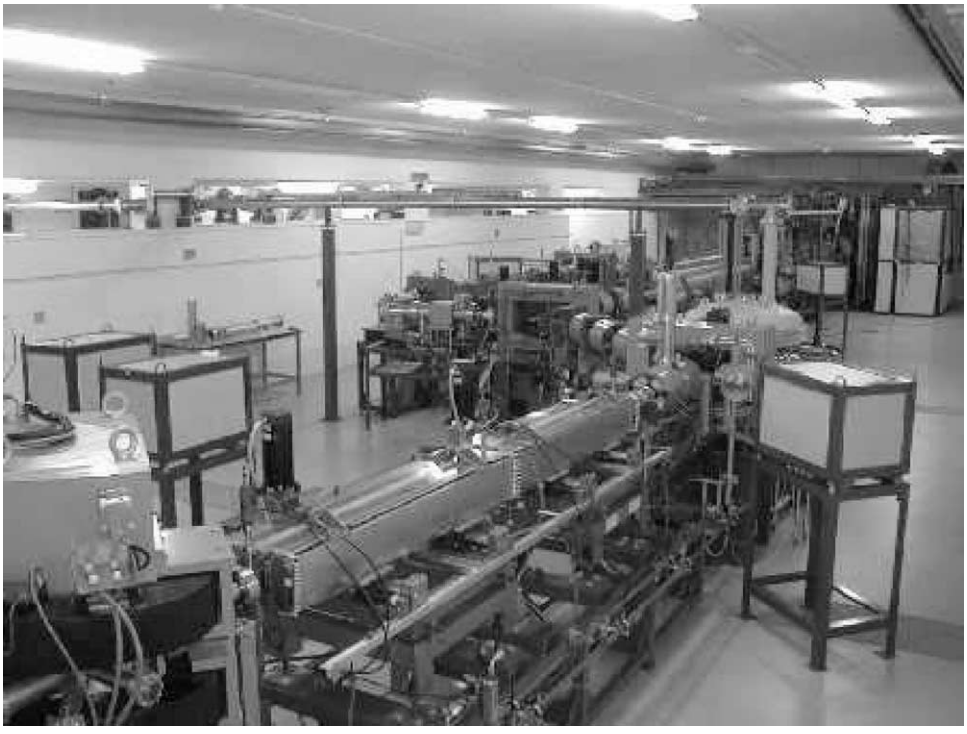


Figure 15 The FELIX RF-Linac FEL operating as a radiation users center in F.O.M. Netherlands. (Courtesy of L. van der Meer, F.O.M.)

This limits the FEL power to be kept as a fraction of the synchrotron radiation power dissipation all around the ring (the 'Renieri Limit'). The effect of the FEL on the e-beam quality, reduces the lifetime of the electrons in the storage ring, and so distrupts the normal operation of the ring in a synchrotron radiation user facility.

To avoid the interference problems, it is most desirable to operate FELs in a dedicated storage ring. This also provides the option to leave long enough straight sections in which long enough wigglers provide sufficient gain for FEL oscillation. Figure 16 displays the Duke storage ring FEL, which is used as a unique radiation user facility, providing intense coherent short wavelength radiation for applications in medicine, biology, material studies, etc.

Superconducting (SC) RF-LINACS

When the RF cavities of the accelerator are superconducting, there are very low RF power losses on the cavity walls, and it is possible to maintain continuous acceleration field in the RF accelerator with a moderate-power continuous RF source, which delivers all of its power to the electron beam kinetic energy. Combining the SC-RF-LINAC technology with an FEL oscillator, pioneered primarily by Stanford University and Thomas Jefferson Lab (TJL) in the US and JAERI Lab in Japan, gave rise to an important scheme of operating such a system in a current recirculating energy retrieval mode. This scheme revolutionized the development of FELs in the direction of high-power, high-efficiency operation, which is highly desirable, primarily for industrial applications (material processing, photochemical production, etc.).

In the recirculating SC-RF-LINAC FEL scheme the wasted beam emerging out of the wiggler after losing a fraction of only few percents (see eqn [89]) out of its kinetic energy, is not dumped into a beam-dump, as in normal cavity RF accelerators, but is re-injected, after circulation, into the SC-RF accelerator. The timing of the wasted electron bunches re-injection is such that they experience a deceleration phase along the entire length of the accelerating cavities. Usually, they are re-injected at the same cell with a fresh new electron bunch injected at an acceleration phase, and thus the accelerated fresh bunch receives its acceleration kinetic energy directly from the wasted beam bunch, that is at the same time decelerated. The decelerated wasted beam bunches are then dumped in the electron beam dump at much lower energy than without recirculation, at energies that are limited primarily just by the energy spread induced in the beam in the FEL laser-saturation process. This scheme, not only increases many folds the over-all energy transformation efficiency from wall-plug to radiation, but would solve significant heat dissipation and radioactivity activation problems in a high-power FEL design.

Figure 17 displays the TJL Infrared SC-RF-LINAC FEL oscillator, that demonstrated for the first time

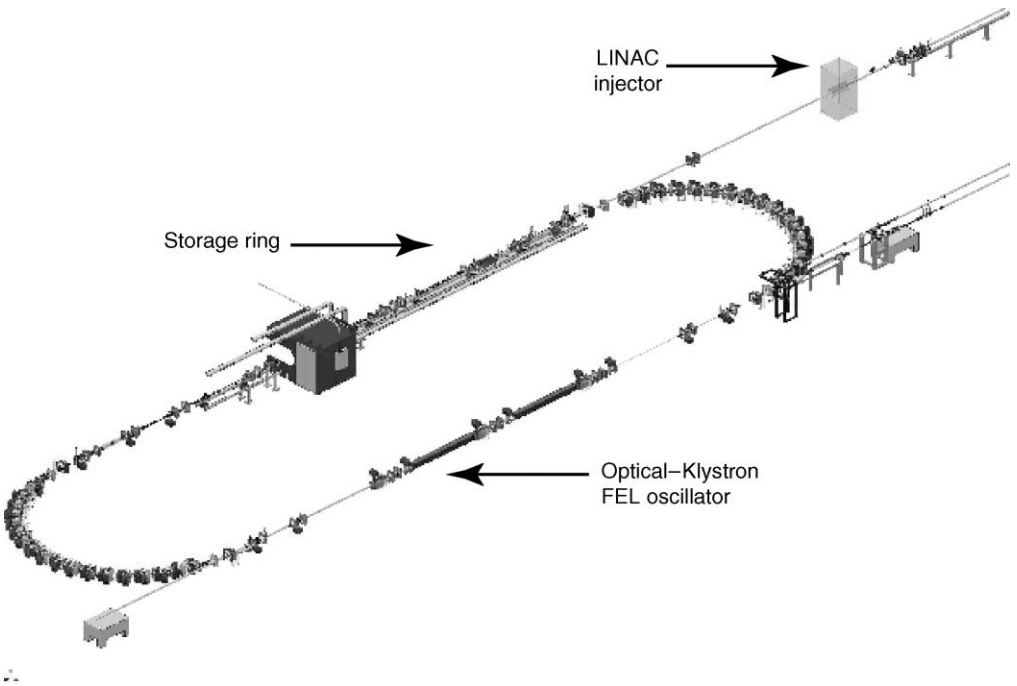


Figure 16 The Duke – University Storage Ring FEL operating as a radiation-users center in N. Carolina, USA. (Mendening: Matthew Busch, courtesy of Glenn Edwards, Duke FEL Lab.)

record high average power levels – nearly 10 kWatt at optical frequencies (1–14 μm). The facility is in upgrade development stages towards eventual operation at 100 kWatt in the IR and 1 kWatt in the UV. It operates in the framework of a laser material processing consortium and demonstrates important material processing applications, such as high-rate micromachining of hard materials (ceramics) with picoSecond laser pulses.

The e-beam current recirculation scheme of SC-RF-LINAC FEL has a significant advantage over the e-beam recirculation in a storage ring. As in electrostatic accelerators, the electrons entering the wiggler are 'fresh' cold-beam electrons from the injector, and not a wasted beam corrupted by the laser saturation process in a previous circulation through the FEL. This also makes it possible to sustain high average circulating current despite the disruptive effect of the FEL on the e-beam. This technological development has given rise to a new concept for a radiation-user facility light-source 4GLS (fourth-generation light source), which is presently in a pilot project development stage at Daresbury Lab in the UK (see Figure 18). In such a scheme, IR and UV FEL oscillators and XUV SASE-FEL can be operated together with synchrotron magnet dipole and wiggler insertion devices without disruptive interference. Such a scheme, if further developed, can give rise to new radiation-user, light-source facilities, that can provide a wider range of radiation parameters than synchrotron centers of previous generation.

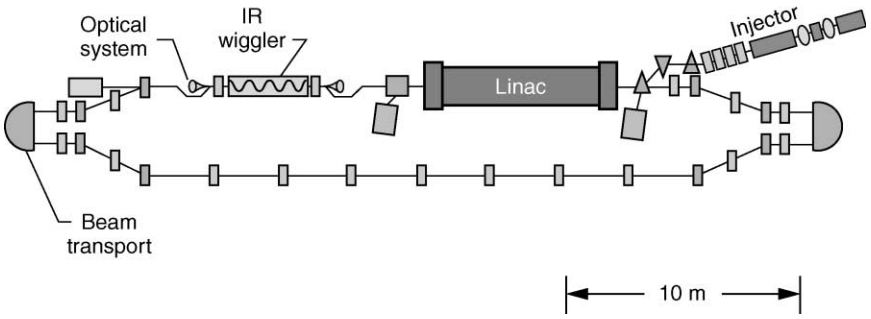


Figure 17 The Thomas Jefferson Lab. recirculating beam-current superconducting Linac FEL operating as a material processing FEL-user center in Virginia USA (Courtesy of S. Benson, Thomas Jefferson Laboratory).

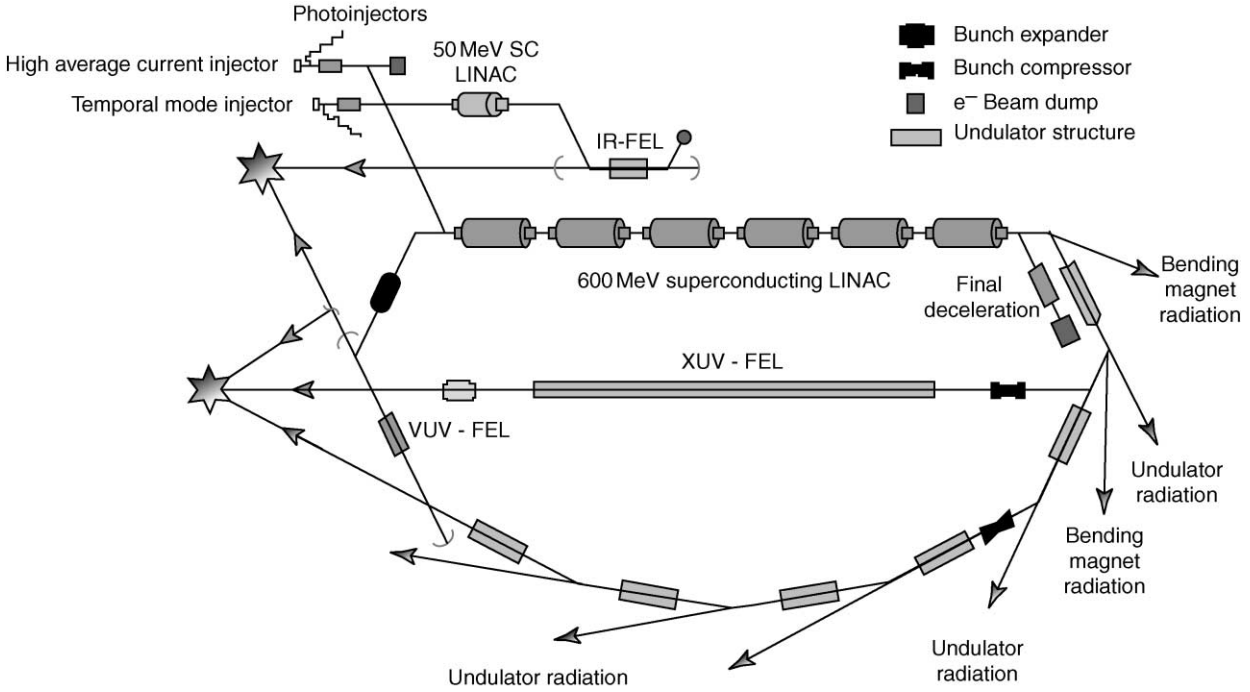


Figure 18 The Daresbury Fourth Generation Light-Source concept (4GLS). The circulating beam-current superconducting Linac includes SASE-FEL, bending magnets and wigglers as insertion devices. (Courtesy of M. Poole, Damesbury Laboratory)

Magnetic Wiggler Schemes

The optical klystron

The stimulated emission process in FEL (see Figure 7c) is based on velocity (energy) bunching of the e-beam in the first part of the wiggler, which turns into density bunching along the central part of the wiggler, and then the density-bunched electron beam performs 'negative work' on the radiation wave and emits radiative energy in the last part of the wiggler. In the OK, these steps are carried out in three separate parts of the wiggler: the energy bunching wiggler section, the dispersive magnet density buncher, and the radiating wiggler section (see Figure 7b).

A schematic of the OK is shown in Figure 19. The chicane magnetic structure in the dispersive section brings all electrons emerging from the bunching wiggler back onto the axis of the radiating wiggler, but provides variable delay $\Delta t_{di} = \int_{L_b}^{L_b+L_d} (v_{zi}^{-1} - v_{ph}^{-1})\, dz = [d(\Delta t_d)/d\gamma]\delta\gamma_i$ relative to the pondermotive wave phase to different electrons, which acquired different energy modulation increments $\delta\gamma_i = \gamma_i - \gamma_{ph}$ in the final section. The radiation condition is satisfied whenever the bunch-center phase satisfies $\Delta\varphi_d = \omega\Delta t_d = \pi/2 - 2m\pi$ (see Figure 7b). However, because the energy dispersion coefficient $d(\Delta t_d)/d\gamma$, is much larger in the chicane than in a wiggler of the same length, the density bunching amplitude, and consequently the OK gain, are much larger than in a uniform wiggler FEL of the same length.

The OK was invented by Vinokurov and Skrinsky in 1977 and first demonstrated in 1987 at visible wavelengths in the ACO storage ring of LURE in Orsay, France, and subsequently in 1988 at UV wavelengths, in the VEPP storage ring in Novosibirsk, Russia. The OK is an optimal FEL configuration, if used as an insertion device in a storage ring, because it can provide sufficient gain to exceed the high lasing threshold at the short operating wavelengths of a high-energy storage-ring FEL, and still conform with the rather short straight sections available for insertion devices in conventional synchrotron storage rings. It should be noted that the OK is equivalent to a long wiggler FEL of length L_{eff} of equal gain and therefore its axial velocity spread acceptance is small (this is determined from the cold beam limit $\bar{\theta}_{th} \ll \pi$ with L_{eff} used in eqn [52]). This too is consistent with storage ring accelerators, which are characterized by small energy spread and emittance of the electron beam.

Radiation emission at harmonic frequencies

In a linear wiggler (eqn [2]), the axial velocity:

$$\beta_z = [\beta^2 - (a_w/\gamma)^2 \cos^2 k_w z]^{1/2} \qquad [92]$$

is not constant. It varies with spatial periodicity $\lambda_w/2$, and in addition to its average value $\bar{\beta}_z = [\beta^2 - a_w^2/2\gamma^2]^{1/2}$, contains Fourier components of spatial frequencies $2mk_w$ $(m = 1, 2, \ldots)$. When $a_w \gg 1$, the axial oscillation deforms the sinusoidal trajectory of the electrons in the wiggler (eqns [22] and [23]), and in a frame of reference moving at the average velocity $\bar{\beta}_z$, the electron trajectories in the wiggling $(x-z)$ plane forms an figure 8 shape, rather than a pure transverse linear motion. In the laboratory frame this leads to synchrotron undulator emission in the forward direction at all odd harmonic frequencies of ω_0, corresponding to substitution of $k_w \to (2m+1)k_w$ $(m = 1, 2, 3, \ldots)$ in eqn [6]:

$$\omega_{2m+1} = (2m+1)\omega_0 \cong 2\gamma_z^2 c(2m+1)k_w \qquad [93]$$

All the stimulated emission gain expressions, presented earlier for the fundamental harmonic, are valid with appropriate substitution of

$$\theta_{2m+1} = \frac{\omega}{v_z} - k_z - (2m+1)k_w \qquad [94]$$

instead of eqn [43], and substitution of the harmonic-weight Bessel-function coefficient of

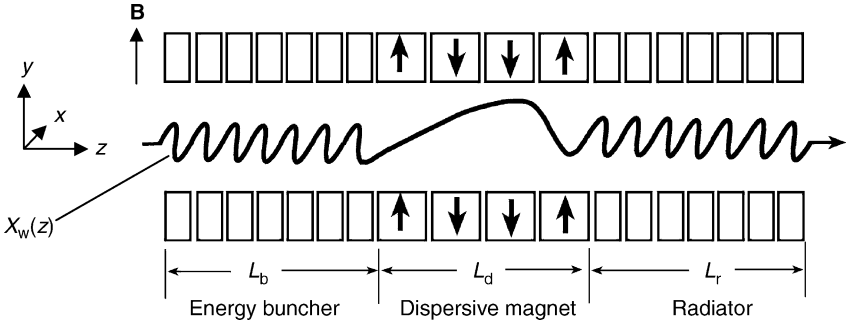


Figure 19 Schematics of the Optical–Klystron, including an energy bunching wiggler, a dispersive magnet bunching section and a radiating wiggler.

synchronism with the beam. Slowing down the PM wave can be done by the gradual increase of the wiggler wavenumber $k_w(z)$ (or decrease of its period $\lambda_w(z)$), so that eqns [29] or [91] keep being satisfied for a given frequency, even if v_z (or γ_z) goes down.

A more correct description of the nonlinear interaction dynamics of the electron beam in a saturated tapered-wiggler FEL is depicted in Figure 20: the electron trap synchronism energy $\gamma_{ph}(z)$ tapers down (by design) along the wiggler, while the trapped electrons are forced to slow down with it, releasing their excess energy by enhanced radiation. An upper limit estimate for the extraction efficiency of such a tapered wiggler FEL would be:

$$\eta_{ext} = \frac{\gamma_{ph}(0) - \gamma_{ph}(L)}{\gamma_{ph}(0)} \qquad [99]$$

and the corresponding radiative power generation would be: $\Delta P = \eta_{ext} I_b E_k / e$. In practice, the phase-space area of the tapered wiggler separatrix is reduced due to the tapering, and only a fraction of the electron beam can be trapped, which reduces correspondingly the practical enhancement in radiative extraction efficiency and power.

An alternative wiggler tapering scheme consists of tapering the wiggler field $B_w(z)$ (or wiggler parameter amplitude $a_w(z)$). If these are tapered down, the axial velocity and axial energy (eqn [7]) can still keep constant (and in synchronism with the PM wave) even if the beam energy γ goes down. Thus, in this scheme, the excess radiative energy extracted from the beam comes out of its transverse (wiggling) energy.

Efficiency and power enhancement of FEL by wiggler tapering have been demonstrated experimentally both in FEL amplifiers (first by Livermore, 1985) and oscillators (first by Los-Alamos, 1983). This elegant way to extract more power from the beam still has some limitations. It can operate efficiently only at a specified high radiation power level for which the tapering was designed. In an oscillator, a long enough untapered section must be left to permit sufficient small signal gain in the early stages of the laser oscillation build-up process.

FEL Oscillators

Most FEL devices are oscillators. As in any laser, in order to turn the FEL amplification process into an oscillation process, one provides a feedback mechanism by means of an optical resonator. In steady state saturation, $GR_{rt} = 1$, where R_{rt} is the round trip reflectivity factor of the resonator and $G = P(L)/P(0)$ is the saturated single-path gain coefficient of the FEL. To attain oscillation, the small signal (unsaturated) gain, usually given by the small gain expression in eqn [53], must satisfy the lasing threshold condition $G > 1/R_{rt}$, as in any laser.

When steady state oscillation is attained, the oscillator output power is:

$$P_{out} = \frac{T}{1 - R_{rt}} \Delta P_{ext} \qquad [100]$$

where $\Delta P_{ext} = \eta_{ext} I_0 (\gamma_0 - 1) mc^2/e$ and η_{ext} is the extraction efficiency, usually given by eqn [89] (low-gain limit).

Usually, FEL oscillators operate in the low-gain regime, in which case $1 - R_{rt} = L + T \ll 1$ (where L is the resonator internal loss factor). Consequently, then $P_{out} \cong \Delta P_{ext} T/(L + T)$, which would give a maximum value, depending on the saturation level of the oscillator. In the general case, one must solve the nonlinear force equations together with the resonator feedback relations of the oscillating radiation mode, in order to maximize the output power (eqn [100]) or efficiency by choice of optimal T for given L.

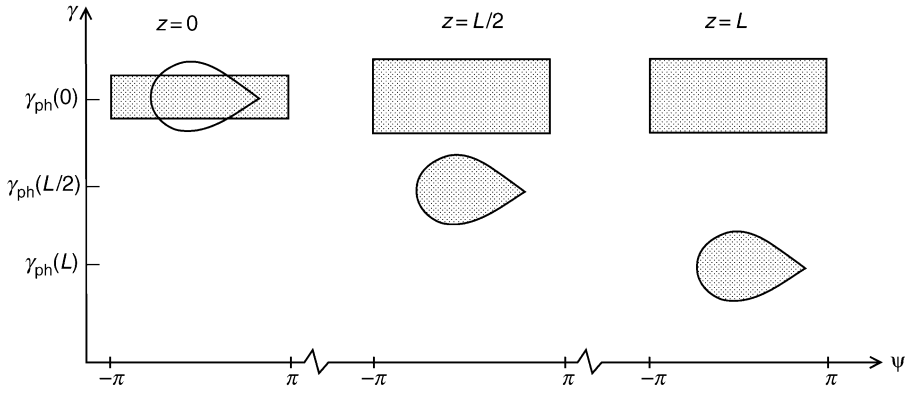


Figure 20 'Snapshots' of the trap at three locations along a tapered wiggler FEL.

In an FEL oscillator operating with periodic electron bunches (as in RF-acclerator based FEL), the solution for the FEL gain and saturation dynamics requires extension of the single frequency solution of the electron and electromagnetic field equations to the time domain. In principle, the situation is similar to that of a mode-locked laser, and the steady state laser pulse train waveform constitutes a superposition of the resonator longitudinal modes that produces a self-similar pulse shape with the highest gain (best overlap with the e-beam bunch along the interaction length). Because the e-beam velocity v_{z0} is always smaller (in an open resonator) than the group velocity of the circulating radiation wavepacket, the radiation wavepacket slips ahead of the electron bunch one optical period λ in each wiggling period (Slippage effect). This reduces the overlap between the radiation pulse and the e-beam bunch along the wiggler (see **Figure 14**) and consequently decreases the gain. Fine adjustment of the resonator mirrors (as shown in **Figure 14**) is needed to attain maximal power and optimal radiation pulse shape. The pulse-slippage gain reduction effect is negligible only if the bunch length is much longer than the slippage length $N_w\lambda$, which can be expressed as:

$$\tau_p \gg 2\pi/\Delta\omega_L \qquad [101]$$

where $\Delta\omega_L$ is the synchrotron undulator radiation frequency bandwidth (eqn [55]). This condition is usually not satisfied in RF-accelerator FELs operating in the IR or lower frequencies, and the

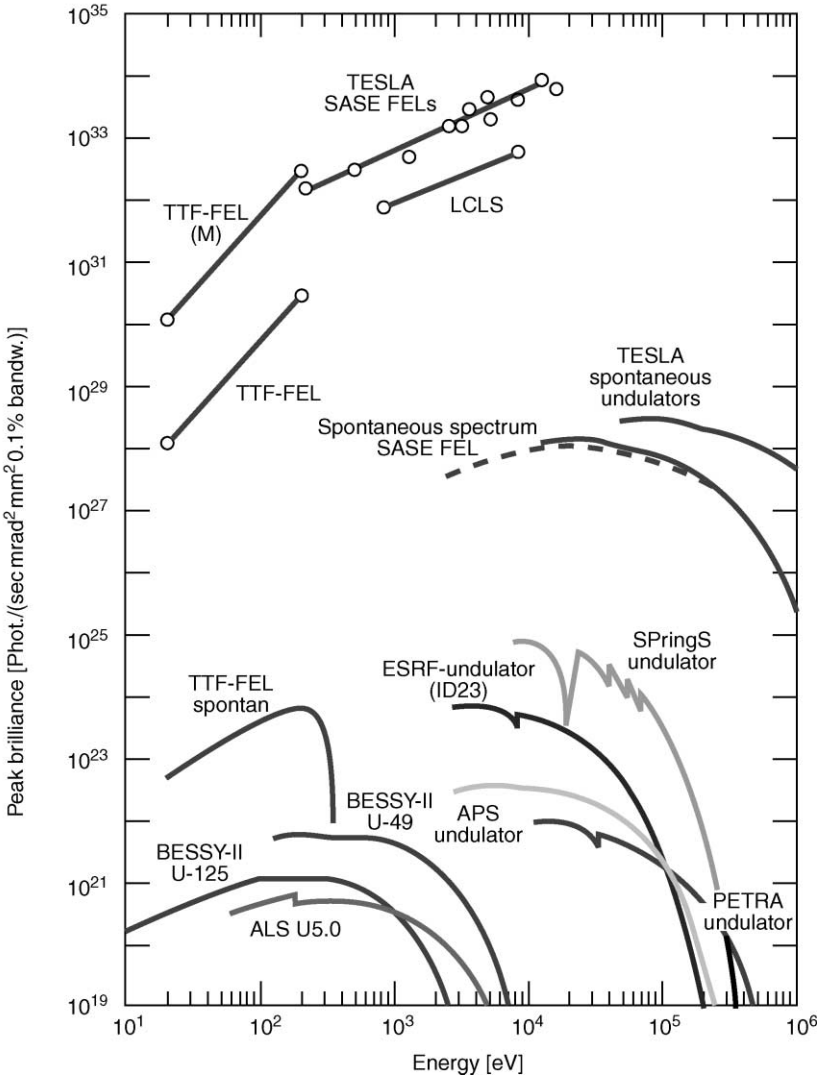


Figure 21 Anticipated peak brightness of SASE FELs (TTF-DESY, LCLS-SLAC) in comparison to the undulators in present third generation Synchrotron Radiation sources. Figure courtesy of DESY, Hamburg, Germany.

slippage effect gain reduction must be then taken into account.

An FEL operating in the cold-beam regime constitutes an 'homogeneous broadening' gain medium in the sense of conventional laser theory. Consequently, the longitudinal mode competition process that would develop in a CW FEL oscillator, leads to single-mode operation and high spectral purity (temporal coherence) of the laser radiation. The minimal (intrinsic) laser linewidth would be determined by an expression analogous to the Schawlow–Towns limit of atomic laser:

$$(\Delta f)_{\text{int}} = \frac{(\Delta f_{1/2})^2}{I_b/e} \qquad [102]$$

where $\Delta f_{1/2}$ is the spectral width of the cold resonator mode. Expression [102] predicts extremely narrow linewidth. In practice, CW operation of FEL was not yet attained, but Fourier transform limited linewidths in the range of $\Delta f/f_0 \cong 10^{-6}$ were measured in long-pulse electrostatic accelerator FELs. In an FEL oscillator, based on a train of e-beam bunches (e.g., an R.F. accelerator beam), the linewidth is very wide and is equal to the entire gain bandwidth (eqn [56]) in the slippage dominated limit, and to the Fourier transform limit $\Delta\omega \cong 2\pi/\tau_p$ in the opposite negligible-slippage limit (eqn [101]). Despite this slippage, it was observed in RF-LINAC FEL that the radiation pulses emitted by the FEL oscillator are phase corrected with each other, and therefore their total temporal coherence length may be as long as the

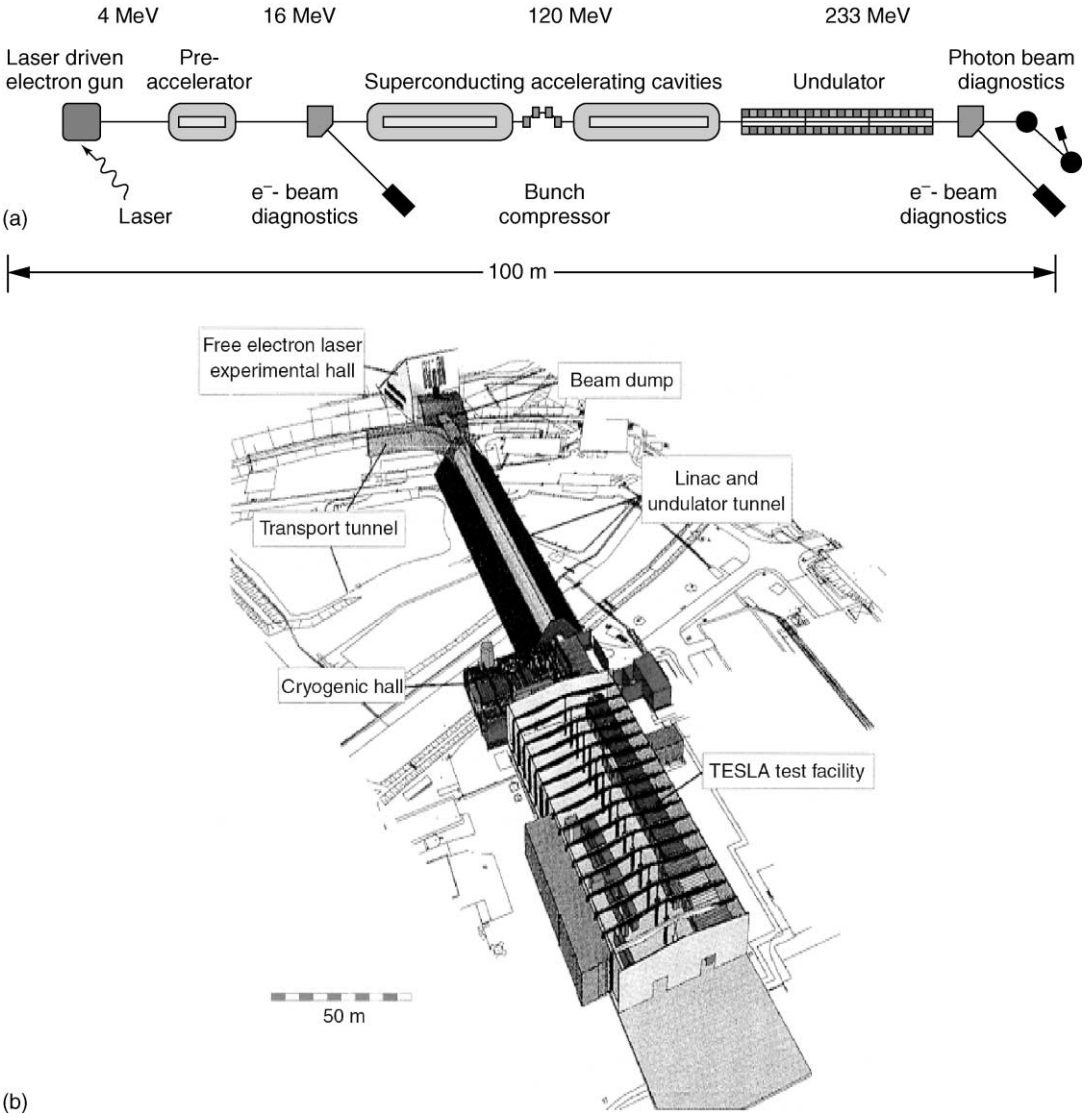


Figure 22 Phase 1 of the SASE FE L (TTF VUV-FEL1): (a) Accelerator layout scheme; (b) General view of the TESLA test facility. Figure courtesy of DESY, Hamburg, Germany.

Metal Vapor Lasers

D W Coutts, University of Oxford, Oxford, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Metal vapor lasers form a class of laser in which the active medium is a neutral or ionized metal vapor usually excited by an electric discharge. These lasers fall into two main subclasses, namely cyclic pulsed metal vapor lasers and continuous-wave metal ion lasers. Both types will be considered in this article, including basic design and construction, power supplies, operating characteristics (including principal wavelengths), and brief reference to their applications.

Self-Terminating Resonance-Metastable Pulsed Metal Vapor Lasers

The active medium in a self-terminating pulsed metal vapor laser consists of metal atoms or ions in the vapor phase usually as a minority species in an inert buffer gas such as neon or helium. Laser action occurs between a resonance upper laser level and a metastable lower laser level (Figure 1). During a fast pulsed electric discharge (typically with a pulse duration of order 100 ns) the upper laser level is preferentially excited by electron impact excitation because it is strongly optically connected to the ground state (resonance transition) and hence has a large excitation cross-section. For a sufficiently large metal atom (or ion) density, the resonance radiation becomes optically trapped, thus greatly extending the lifetime of the upper laser level such that decay from the upper laser level is channelled through the emission of laser radiation to the metastable lower laser level. Lasing terminates when the electron temperature falls to a point such that preferential pumping to the upper laser level is no longer sustained, and the build-up of population in the metastable lower laser level destroys the population inversion. Therefore after each excitation pulse, the resulting excited species in the plasma (in particular the metastable lower laser levels which are quenched by collisions with cold electrons) must be allowed sufficient time to relax and the plasma must be allowed to partially recombine before applying the next excitation pulse. The relaxation times for self-terminating metal vapor lasers correspond to operating pulse repetition frequencies from 2 kHz to 200 kHz.

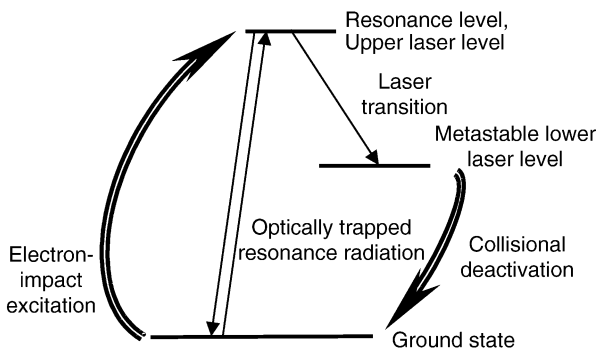


Figure 1 Resonance-metastable energy levels for self-terminating metal vapor lasers.

Many metal vapors can be made to lase in the resonance-metastable scheme and are listed together with their principal wavelengths and output powers in Table 1. The most important self-terminating pulsed metal vapor laser is the copper vapor laser and its variants which will be discussed in detail in the following sections. Of the other pulsed metal vapor lasers listed in Table 1, only the gold vapor laser (principal wavelengths 627.8 nm and 312.3 nm), and the barium vapor laser which operates in the infrared (principal wavelengths 1.5 μm and 2.55 μm) have had any commercial success. All the self-terminating pulsed metal vapor lasers have essentially the same basic design and operating characteristics as exemplified by the copper vapor laser.

Copper Vapor Lasers

Copper vapor lasers (CVLs) are by far the most widespread of all the pulsed metal vapor lasers. Figure 2 shows the energy level scheme for copper. Lasing occurs simultaneously from the $^2P_{3/2}$ level to the $^2D_{5/2}$ level (510.55 nm) and from the $^2P_{1/2}$ level to the $^2D_{3/2}$ level (578.2 nm). Commercial devices are available with combined outputs of over 100 W at 510.55 nm and 578.2 nm (typically with a green-to-yellow power ratio of 2:1).

A typical copper vapor laser tube is shown in Figure 3. High-purity copper pieces are placed at intervals along an alumina ceramic tube which typically has dimensions of 1–4 cm diameter and 1–2 m long. The alumina tube is surrounded by a solid fibrous alumina thermal insulator, and a glass or quartz vacuum envelope. Cylindrical electrodes made of copper or tantalum are located at each end of the plasma tube to provide a longitudinal discharge

Table 1 Principal self-terminating resonance-metastable metal vapor lasers

Metal	*Principal wavelengths (nm)*	*Powers*		*Total efficiencies*	*Pulse repetition frequency (kHz)*	*Technological development*
		Typical (W)	*Maximum (W)*			
Cu	510.55	2–70	2500 total	1%	4–40	Highly developed and commercially available
	578.2	1–50				
Au	627.8	1–8	20	0.13%	2–40	Commercially available
	312	0.1–0.2	1.2		1–8	
Ba	1500	2–10	12	0.5%	5–15	Have been produced commercially, but largely experimental
	2550	1	1.5			
	1130	0.5	1.0			
Pb	722.9		4.4	0.15%	10–30	Experimental
Mn	534.1 (>50%)		12 total	0.32%	~10	Experimental
	1290					

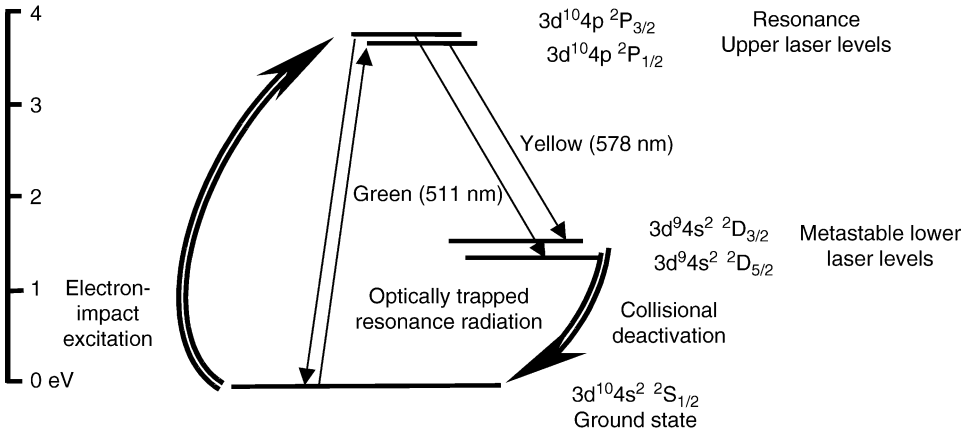


Figure 2 Partial energy level scheme for the copper vapor laser.

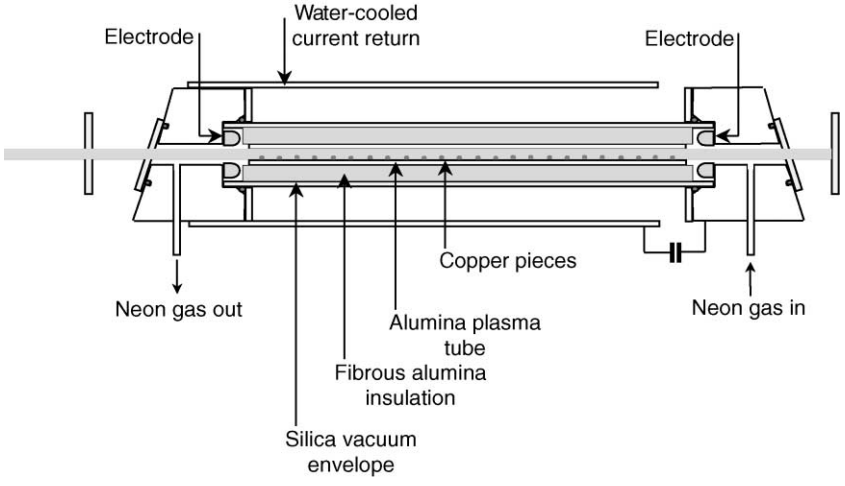


Figure 3 Copper vapor laser tube construction.

arrangement. The cylindrical electrodes and silica laser end windows are supported by water-cooled end pieces. The laser windows are usually tilted by a few degrees to prevent back reflections into the active medium. The laser head is contained within a water cooled metal tube to provide a coaxial current return for minimum laser head inductance. Typically a slow flow ($\sim$5 mbar l min^{-1}) of neon at a pressure of 20–80 mbar is used as the buffer gas with an approximately 1% H_2 additive to improve the afterglow plasma relaxation. The buffer gas provides a medium to operate the discharge when the laser is cold and slows diffusion of copper vapor out of the ends of the hot plasma tube. Typical copper fill times are of order 200–2000 hours (for 20–200 g copper load). Sealed-off units with lifetimes of order 1000 hours have been in production in Russia for many years.

During operation waste heat from the repetitively pulsed discharge heats the alumina tube up to approximately 1500 °C at which point the vapor pressure of copper is of approximately 0.5 mbar which corresponds to the approximate density required for maximum laser output power. Typical warm-up times are therefore relatively long at around one hour to full power.

One way to circumvent the requirement for high temperatures required to produce sufficient copper density by evaporation of elemental copper (and hence also reduce warm-up times) is to use a copper salt with a low boiling point located in one or more side-arms of the laser tube (Figure 4). Usually copper halides are used as the salt, with the copper bromide laser being the most successful. For the CuBr laser, a temperature of just 600 °C is sufficient to produce the required Cu density by dissociation of CuBr vapor in the discharge. With the inclusion of 1–2% H_2 in the neon buffer gas, HBr is also formed in the CuBr laser, which has the additional benefit of improving recombination in the afterglow via dissociative attachment of free electrons: $HBr + e^- \rightarrow H + Br^-$, followed by ion neutralization: $Br^- + Cu^+ \rightarrow Br + Cu^*$. As a result of the lower operating temperature and kinetic advantages of HBr, CuBr lasers are typically twice as efficient (2–3%) as their elemental counterparts. Sealed-off CuBr systems with powers of order 10–20 W are commercially produced.

An alternative technique for reducing the operating temperature of elemental CVLs is to flow a buffer gas mixture consisting of $\sim$5% HBr in neon at approximately 50 mbar l min^{-1} and allow this to react with solid copper metal placed within the plasma tube at about 600 °C to produce CuBr vapor *in situ*. The so-called Cu HyBrID (hydrogen bromide in discharge) laser has the same advantages as the CuBr laser (e.g., up to 3% efficiency) but at the cost of requiring flowing highly toxic HBr in the buffer gas, a requirement which has so far prevented commercialization of Cu HyBrID technology.

The kinetic advantages of the hydrogen halide in the CuBr laser discharge can also be applied to a conventional elemental CVL through the addition of small partial pressure of HCl to the buffer gas in addition to the 1–2% H_2 additive. HCl is preferred to HBr as it is less likely to dissociate (the dissociation energy of HCl at 0.043 eV is less than HBr at 0.722 eV). Such kinetic enhancement leads to a doubling in average output power, a dramatic increase in beam quality through improved gain characteristics, and shifts the optimum pulse repetition frequency for kinetically enhanced CVLs (KE-CVLs) from 4–10 kHz up to 30–40 kHz.

Preferential pumping of the upper laser levels requires an electron temperature in excess of the 2 eV range, hence high voltage (10–30 kV), high current (hundreds of A), short (75–150 ns) excitation pulses are required for efficient operation of copper vapor lasers. To generate such excitation pulses CVLs are typically operated with a power supply incorporating a high-voltage thyratron switch. In the most basic configuration, the charge-transfer circuit shown

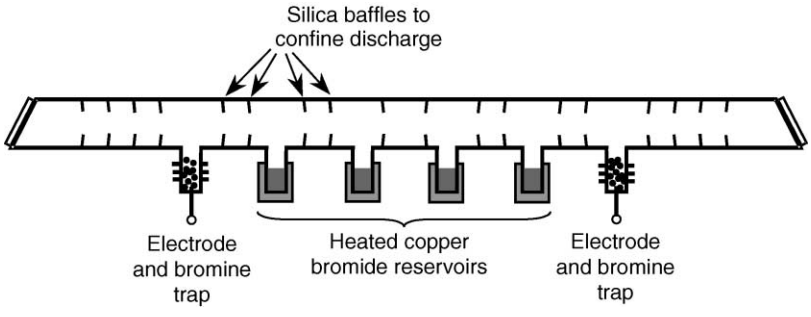


Figure 4 Copper bromide laser tube construction.

in Figure 5a, a dc high-voltage power supply resonantly charges a storage capacitor (C_S, typically a few nF) through a charging inductor L_C, a high-voltage diode and bypass inductor L_B, up to twice the supply voltage V_S in a time of order 100 μs. When the thyratron is triggered, the storage capacitor discharges through the thyratron and the laser head on a time-scale of 100 ns. Note that during the fast discharge phase, the bypass inductor in parallel with the laser head can be considered to be an open circuit. A peaking capacitor C_P ($\sim 0.5\ C_S$) is provided to increase the rate of rise of the voltage pulse across the laser head. Given the relatively high cost of thyratrons, more advanced circuits are now often used to extend the service lifetime of the thyratron to several thousand hours. In the more advanced circuit (Figure 5b) an LC inversion scheme is used in combination with magnetic pulse compression techniques and operates as follows. Storage capacitors C_S are resonantly charged in parallel to twice the dc supply voltage V_S as before. When the thyratron is switched, the charge on C_{S1} inverts through the thyratron and transfer inductor L_T. This LC inversion drags the voltage on the top of C_{S2} down to $-4V_S$. When the voltage across the first saturable inductor L_{S1} reaches a maximum ($-4V_S$) L_{S1} saturates, and allows current to flow from the storage capacitors (now charged in series) to the transfer capacitor ($C_T = 0.5C_{S2}$) thereby transferring the charge from C_{S1} and C_{S2} to C_T in a time much less than the initial LC inversion time. At the moment when the charge on transfer capacitor C_T reaches a maximum (also $-4V_S$), L_{S2} saturates, and the transfer capacitor is discharged through the laser tube, again with a peaking capacitor to increase the voltage rise time. By using magnetic pulse compression the thyratron switched voltage can be reduced by 4 and the peak current similarly reduced (at the expense of increased current pulse duration) thereby greatly extending the thyratron lifetime. Note that in both circuits a 'magnetic assist' L_A saturable inductor is provided in series with the thyratron to delay the current pulse through the thyratron until after the thyratron has reached high conductivity thereby reducing power deposition in the thyratron.

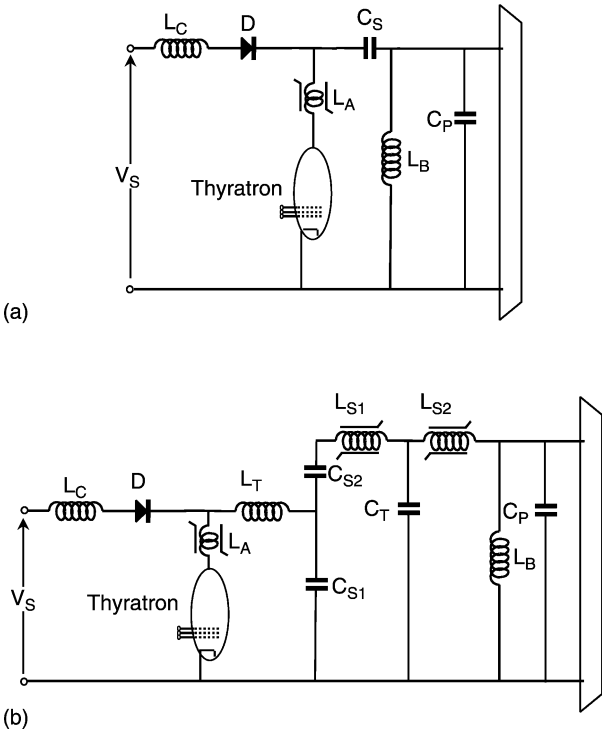


Figure 5 Copper vapor laser excitation circuits. (a) Charge transfer circuit; (b) LC inversion circuit with magnetic pulse compression.

Copper vapor lasers produce high average powers (2–100 W available commercially, with laboratory devices producing average powers of over 750 W) and have wall-plug efficiencies of approximately 1%. Copper vapor lasers also make excellent amplifiers due to their high gains, and the amplifiers can be chained together to produce average powers of several kW. Typical pulse repetition frequencies range from 4 to 20 kHz, with a maximum reported pulse repetition frequency of 250 kHz. An approximate scaling law states that for an elemental device with tube diameter D (mm) and length L (m) the average output power in watts will be of order $D \times L$. For example, a typical 25 W copper vapor laser will have a 25 mm diameter by 1 m long laser tube, and operate at 10 kHz corresponding to 2.5 mJ pulse energy and 50 kW peak power (50 ns pulse duration). Copper vapor lasers have very high single-pass gains (greater than 1000 for a 1 m long tube), large gain volumes and short gain durations (20–80 ns, sufficient for the intracavity laser light to make only a few round trips within the optical resonator). Maximum output power is therefore usually obtained in a highly 'multimode' (spatially incoherent) beam by using either a fully stable or a plane–plane resonator with a low reflectivity output coupler (usually Fresnel reflection from an uncoated optic is sufficient). To obtain higher beam quality a high magnification unstable resonator is required (Figure 6). Fortunately, copper vapor lasers have sufficient gain to operate efficiently with unstable resonators with magnifications ($M = R_1/R_2$) up to 100 and beyond. Resonators with such high magnifications impose very tight geometric constraints on propagation of radiation on repeated round-trips within the resonator, such that after two round-trips the divergence is typically diffraction-limited. Approximately half the stable-resonator

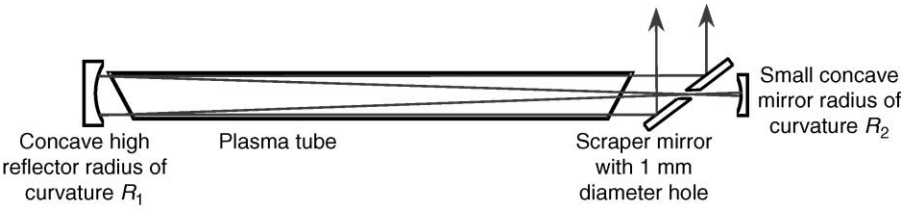


Figure 6 Unstable resonator configuration often used to obtain high beam quality from copper vapor lasers.

output power can therefore be obtained with near diffraction-limited beam quality by using an unstable resonator. Often, a small-scale oscillator is used in conjunction with a single power amplifier to produce high output power with diffraction-limited beam quality. Hyperfine splitting combined with Doppler broadening lead to an inhomogeneous linewidth for the laser transitions of order 8–10 GHz corresponding to a coherence length of order 3 cm.

The high beam quality and moderate peak power of CVLs allows efficient nonlinear frequency conversion to the UV by second harmonic generation (510.55 nm → 255.3 nm, 578.2 nm → 289.1 nm) and sum frequency generation (510.55 nm + 578.2 nm → 271.3 nm) using β-barium borate β-BaB_2O_4 (BBO) as the nonlinear medium. Typically average powers in excess of 1 W can be obtained at any of the three wavelengths from a nominally 20 W CVL. Powers up to 15 W have been obtained at 255 nm from high-power CVL master-oscillator power-amplifier systems using cesium lithium borate; CsLi B_6O_{10} (CLBO) as the nonlinear crystal.

Key applications of CVLs include pumping of dye lasers (principally for laser isotope separation) and pumping Ti:sapphire lasers. Medically, the CVL yellow output is particularly useful for treatment of skin lesions such as port wine stain birth marks. CVLs are also excellent sources of short-pulse stroboscopic illumination for high-speed imaging of fast objects and fluid flows. The high beam quality, visible wavelength and high pulse repetition rate make CVLs very suited to precision laser micromachining of metals, ceramics and other hard materials. More recently, the second harmonic at 255 nm has proved to be an excellent source for writing Bragg gratings in optical fibers.

Afterglow Recombination Metal Vapor Lasers

Recombination of an ionized plasma in the afterglow of a discharge pulse provides a mechanism for achieving a population inversion and hence laser output. The two main afterglow recombination metal vapor lasers are the strontium ion vapor laser (430.5 nm and 416.2 nm) and calcium ion vapor laser (373.7 nm and 370.6 nm) whose output in the violet and UV spectral regions extends the spectral coverage of pulsed metal vapor lasers to shorter wavelengths.

A population inversion is produced by recombination pumping where doubly ionized Sr (or Ca) recombines to form singly ionized Sr (or Ca) in an excited state: $Sr^{++} + e^- + e^- \rightarrow Sr^{+*} + e^-$. Note that recombination rates for a doubly ionized species are much faster than for singly ionized species hence recombination lasers are usually metal ion lasers. Recombination (pumping) rates are also greatest in a cool dense plasma, hence helium is usually used as the buffer gas as it is a light atom hence promotes rapid collisional cooling of the electrons. Helium also has a much higher ionization potential than the alkaline-earth metals (including Sr and Ca) which ensures preferential ionization of the metal species (up to 90% may be doubly ionized).

Recombination lasers can operate where the energy level structure of an ion species can be considered to consist of two (or more) groups of closely spaced levels. In the afterglow of a pulsed discharge electron collisional mixing within each group of levels will maintain each group in thermodynamic equilibrium yielding a Boltzmann distribution of population within each group. If the difference in energy between the two groups of levels (5 eV >> kT) is sufficiently large then thermal equilibrium between the groups cannot be maintained via collisional processes. Given that recombination yields excited singly ionized species (usually with a flux from higher to lower ion levels), it is possible to achieve a population inversion between the lowest level of the upper group and the higher levels within the lower group. This is the mechanism for inversion in the Sr^+ (and analogous Ca^+) ion laser as indicated in the partial energy level scheme for Sr^+ shown in Figure 7.

Strontium and calcium ion recombination lasers have similar construction to copper vapor lasers described above. The lower operating temperatures (500–800 °C) mean that minimal or no thermal insulation is required for self-heated devices.

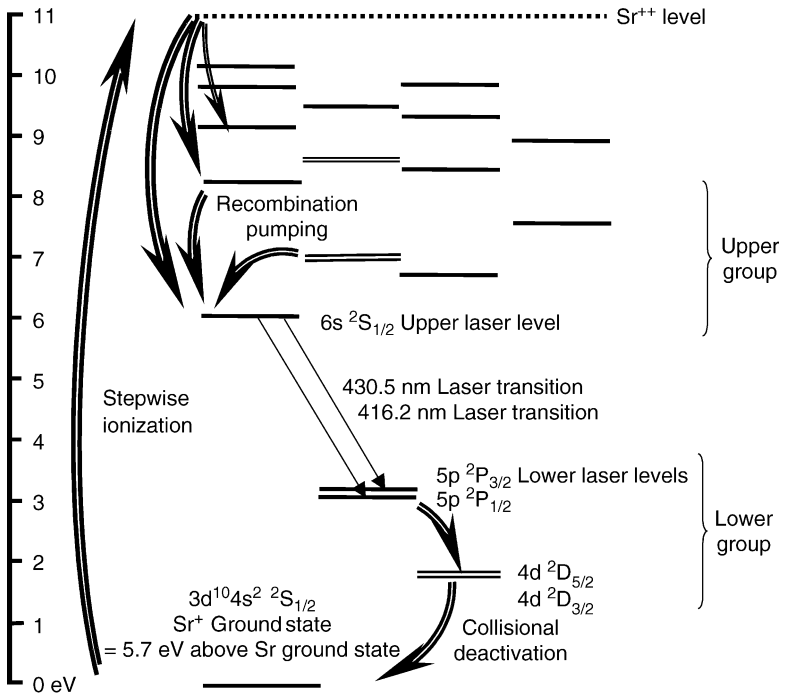


Figure 7 Partial energy level scheme for the strontium ion laser.

For good tube lifetime BeO plasma tubes are required due to the reactivity of the metal vapors. Usually helium at a pressure of up to one atmosphere is used as the buffer gas. Typical output powers at 5 kHz pulse repetition frequency are of order 1 W (~0.1% wall plug efficiency) at 430.5 nm from the He Sr^+ laser and 0.7 W at 373.7 nm from the He Ca^+ laser. For the high specific input power densities (10–15 W cm^{-3}) required for efficient lasing, overheating of the laser gas limits aperture scaling beyond 10–15 mm diameter. Slab laser geometries have been used successfully to aperture-scale strontium ion lasers. Scaling of the laser tube length beyond 0.5 m is not practical as achieving high enough excitation voltages for efficient lasing becomes problematic. Gain in strontium and calcium ion lasers is lower than in resonance-metastable metal vapor lasers such as the CVL, hence optimum output coupler reflectivity is approximately 70%. The pulse duration is also longer at around 200–500 ns resulting in moderate beam quality with plane–plane resonators.

For both strontium and calcium ion lasers the two principal transitions share upper laser levels and hence exhibit gain competition such that without wavelength-selective cavities usually only the longer wavelength of the pair is produced. With wavelength-selective cavities up to 60% (Sr^+) and 30% (Ca^+) of the normal power can be obtained at the shorter wavelength.

Many potential applications exist for strontium and calcium ion lasers, given their ultraviolet (UV)/violet wavelengths. Of particular importance is fluorescence spectroscopy in biology and forensics, treatment of neonatal jaundice, stereolithography, micromachining and exposing photoresists for integrated circuit manufacture. Despite these many potential applications, technical difficulties in power scaling mean that both strontium ion and calcium ion lasers have only limited commercial availability.

Continuous-Wave Metal Ion Lasers

In a discharge excited noble gas, there can be a large concentration of noble gas atoms in excited metastable states and noble gas ions in their ground states. These species can transfer their energy to a minority metal species (M) via two key processes: either charge transfer (Duffendack reactions) with the noble gas ions (N^+):

$$M + N^+ \rightarrow M^{+*} + N + \Delta E$$

(where M^{+*} is an excited metal ion state)

or Penning ionization with a noble gas atom in an excited metastable state (N^*):

$$M + N^* \rightarrow M^{+*} + N + e^-$$

In a continuous discharge in a mixture of a noble gas and a metal vapor, steady generation of excited metal ions via energy transfer processes can lead to a steady state population inversion on one or more pairs of levels in the metal ion and hence produce cw lasing.

Several hundred metal ion laser transitions have been observed to lase in a host of different metals. The most important such laser is the helium cadmium laser, which has by far the largest market volume by number of unit sales of all the metal vapor lasers. In a helium cadmium laser, Cd vapor is present at a concentration of about 1–2% in a helium buffer gas which is excited by a dc discharge. Excitation to the upper laser levels of Cd (Figure 8) is primarily via Penning ionization collisions with He^+ 2^3S_1 metastable ions produced in the discharge. Excitation via electron-impact excitation from the ion ground state may also play an important role in establishing a population inversion in Cd^+ lasers. Population inversion can be sustained continuously because the $^2P_{3/2}$ and $^2P_{1/2}$ lower laser levels decay via strong resonance transitions to the $^2S_{1/2}$ Cd^+ ground state, unlike in the self-terminating metal vapor lasers. Two principal wavelengths can be produced, namely 441.6 nm (blue) and 325.0 nm (UV) with cw powers up to 200 mW and 50 mW available respectively from commercial devices.

Typical laser tube construction (Figure 9) consists of a 1–3 mm diameter discharge channel typically 0.5 m long. A pin anode is used at one end of the laser tube together with a large-area cold cylindrical cathode located in a side-arm at the other end of the tube. Cadmium is transported to the main discharge tube from a heated Cd reservoir in a side-arm at 250–300 °C. With this source of atoms at the anode end of the laser a cataphoresis process (in which the positively charged Cd ions are propelled towards the cathode end of the tube by the longitudinal electric field) transports Cd into the discharge channel. Thermal insulation of the discharge tube ensures that it is kept hotter than the Cd reservoir to prevent condensation of Cd from blocking the tube bore. A large-diameter Cd condensation region is provided at the cathode end of the discharge channel. A large-volume side arm is also provided to act as a gas ballast for maintaining correct He pressure. As He is lost through sputtering and diffusion through the Pyrex glass tube walls, it is replenished from a high-pressure He reservoir by heating a permeable glass wall separating the reservoir from the ballast chamber which allows He to diffuse into the ballast chamber. Typical commercial sealed-off Cd lasers have operating lifetimes of several thousand hours. Overall laser construction is not that much more

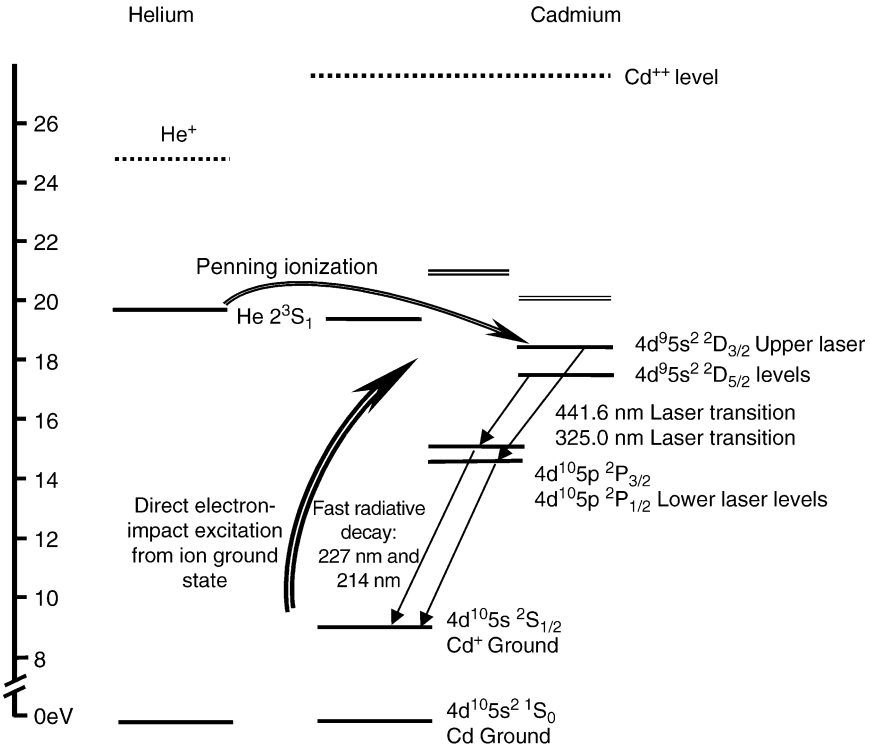


Figure 8 Partial energy level diagram for helium and cadmium giving HeCd laser transitions.

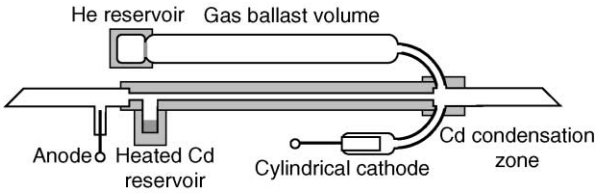


Figure 9 Helium cadmium laser tube construction.

complex than a HeNe laser, hence unit costs are considerably lower than low-power argon ion lasers which also provide output in the blue. Usually Brewster angle windows are provided together with a high Q stable resonator (97–99% reflectivity output coupler) to provide polarized output.

With their blue and UV wavelengths and relatively low cost (compared to low-power argon ion lasers), HeCd lasers have found wide application in science, medicine and industry. Of particular relevance is their application for exposing photoresists where the blue wavelength provides a good match to the peak photosensitivity of photoresist materials. A further key application is in stereolithography where the UV wavelength is used to cure an epoxy resin. By scanning the UV beam in a raster pattern across the surface of a liquid epoxy, a solid three-dimensional object may be built up in successive layers.

List of Units and Nomenclature

Boltzmann constant	[eV K^{-1}]	k
Excited metal ion		M^{+*}
Excited noble gas atom		N^{*}
Energy difference	[eV]	ΔE
Electron		e^{-}
Gas flow	[mbar l min^{-1}]	
Mirror curvatures	[m]	R_1, R_2
Metal atom		M
Noble gas atom		N
Noble gas ion		N^{+}
Pulse duration	[ns]	
Pulse repetition frequency	[kHz]	
Quality factor		Q
Temperature	[K]	T
Tube diameter	[mm]	D
Tube length	[m]	L
Wavelength	[nm], [μm]	
Unstable resonator magnification		M

See also

Nonlinear Sources: Harmonic Generation in Gases.

Further Reading

Little CE (1999) *Metal Vapor Lasers*. Chichester, UK: John Wiley.

Ivanov IG, Latush EL and Sem MF (1996) *Metal Vapor Ion Lasers, Kinetic Processes and Gas Discharges*. Chichester, UK: John Wiley.

Little CE and Sabotinov NV (eds) (1996) *Pulsed Metal Vapor Lasers*. Dordrecht, The Netherlands: Kluwer.

Lyabin NA, Chursin AD, Ugol'nikov SA, Koroleva ME and Kazaryan MA (2001) Development, production, and application of sealed-off copper and gold vapor lasers. *Quantum Electronics* 31: 191–202.

Petrash GG (ed.) (1989) *Metal Vapor and Metal Halide Vapor Lasers*. Commack, NY: Nova Science Publishers.

Silfvast WT (1996) *Laser Fundamentals*. New York: Cambridge University Press.

Noble Gas Ion Lasers

W B Bridges, California Institute of Technology, Pasadena, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

History

The argon ion laser was discovered in early 1964, and is still commercially available in 2004, with about $70 million in annual sales 40 years after this discovery. The discovery was made independently and nearly simultaneously by four different groups; for three of the four, it was an accidental result of studying the excitation mechanisms in the mercury ion laser (historically, the first ion laser), which had been announced only months before. For more on the early years of ion laser research and development, see the articles listed in the Further Reading section at the end of this article.

The discovery was made with pulsed gas discharges, producing several wavelengths in the blue and green portions of the spectrum. Within months, continuous operation was demonstrated, as well as oscillation on many visible wavelengths in ionized krypton and xenon. Within a year, over 100 wavelengths were observed to oscillate in the ions

of neon, argon, krypton, and xenon, spanning the spectrum from ultraviolet to infrared; oscillation was also obtained in the ions of other gases, for example, oxygen, nitrogen, and chlorine. A most complete listing of all wavelengths observed as gaseous ion lasers is given in the Laser Handbook cited in the Further Reading section. Despite the variety of materials and wavelengths demonstrated, however, it is the argon and krypton ion lasers that have received the most development and utilization.

Continuous ion lasers utilize high current density gas discharges, typically 50 A or more, and 2–5 mm in diameter. Gas pressures of 0.2 to 0.5 torr result in longitudinal electric fields of a few V/cm of discharge, so that the power dissipated in the discharge is typically 100 to 200 W/cm. Such high-power dissipation required major technology advances before long-lived practical lasers became available. Efficiencies have never been high, ranging from 0.01% to 0.2%. A typical modern ion laser may produce 10 W output at 20 kW input power from 440 V three-phase power lines and require 6–8 gallons/minute of cooling water. Smaller, air-cooled ion lasers, requiring 1 kW of input power from 110 V single-phase mains can produce 10–50 mW output power, albeit at even lower efficiency.

Theory of Operation

The strong blue and green lines of the argon ion laser originate from transitions between the 4p upper levels and 4s lower levels in singly ionized argon, as shown in Figure 1. The 4s levels decay radiatively to the ion ground state. The strongest of these laser lines are listed in Table 1. The notation used for the energy levels is that of the L–S coupling model. The ion ground state electron configuration is $3s^2 3p^5 (^2P^o_{3/2})$. The inner ten electrons have the configuration $1s^2 2s^2 2p^6$, but this is usually omitted for brevity. The excited states shown in Figure 1 result from coupling a 4p or 4s electron to a $3s^2 3p^4 (^3P)$ core, with the resulting quantum numbers S (the net spin of the electrons), L (the net angular momentum of the electrons), and J (the angular momentum resulting from coupling S to L) are represented by $^{2S+1}L_J$, where $L = 0, 1, 2, \ldots$ is denoted S, P, D, F, … The superscript 'o' denotes an odd level, while even levels omit a superscript. Note that some weaker transitions involving levels originating from the $3s^2 3p^4 (^1D)$ core configuration also oscillate. Note also that the quantum mechanical selection rules for the L–S coupling model are not rigorously obeyed, although the stronger laser lines satisfy most or all of these rules. The selection rule $|\Delta J| = 1$ or 0, but not $J = 0 \rightarrow J = 0$ is always obeyed. All transitions shown in Figure 1 and Table 1 belong to the second spectrum of argon, denoted Ar II. Lines originating from transitions in the neutral atom make up the first spectrum, Ar I; lines originating from transitions in doubly ionized argon are denoted Ar III, and so forth for even more highly ionized states.

Much work has been done to determine the mechanisms by which the inverted population is formed in the argon ion laser. Reviews of this extensive research

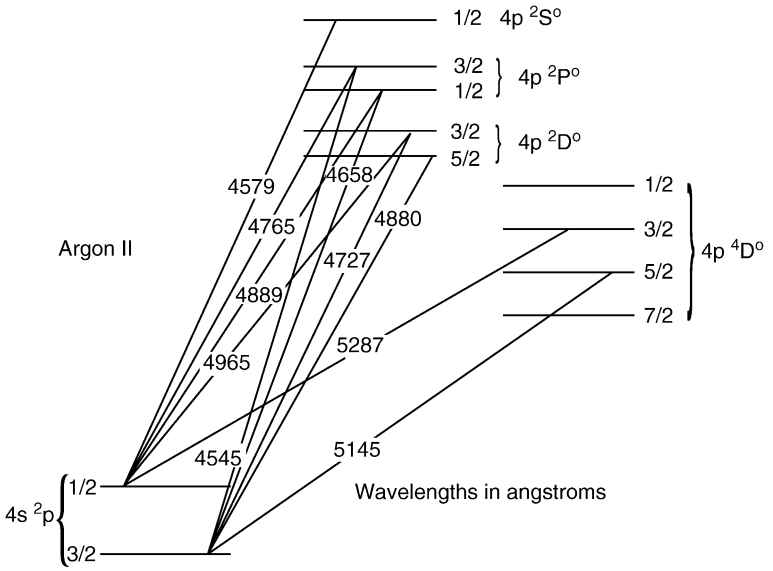


Figure 1 4p and 4s doublet levels in singly ionized argon, showing the strongest blue and green laser transitions.

Table 1 Ar II laser blue-green wavelengths

Wavelength (nanometers)	*Transition[a] (upper level) → (lower level)*	*Relative strength[b] (Watt)*
454.505	$4p\ ^2P^o_{3/2} \rightarrow 4s\ ^2P_{3/2}$	0.8
457.935	$4p\ ^2S^o_{1/2} \rightarrow 4s\ ^2P_{1/2}$	1.5
460.956	$(^1D)4p\ ^2F^o_{7/2} \rightarrow (^1D)4s\ ^2D_{5/2}$	–
465.789	$4p\ ^2P^o_{1/2} \rightarrow 4s\ ^2P_{3/2}$	0.8
472.686	$4p\ ^2D^o_{3/2} \rightarrow 4s\ ^2P_{3/2}$	1.3
476.486	$4p\ ^2P^o_{3/2} \rightarrow 4s\ ^2P_{1/2}$	3.0
487.986	$4p\ ^2D^o_{5/2} \rightarrow 4s\ ^2P_{3/2}$	8.0
488.903	$4p\ ^2P^o_{1/2} \rightarrow 4s\ ^2P_{1/2}$	[c]
496.507	$4p\ ^2D^o_{3/2} \rightarrow 4s\ ^2P_{1/2}$	3.0
501.716	$(^1D)4p\ ^2D^o_{5/2} \rightarrow 3d\ ^2D_{3/2}$	1.8
514.179	$(^1D)4p\ ^2F^o_{7/2} \rightarrow 3d\ ^2D_{5/2}$	[c]
514.532	$4p\ ^4D^o_{5/2} \rightarrow 4s\ ^2P_{3/2}$	10
528.690	$4p\ ^4D^o_{3/2} \rightarrow 4s\ ^2P_{1/2}$	1.8

[a]All levels are denoted in $L-S$ coupling with the (^{3}P) core unless otherwise indicated. Odd parity is denoted by the superscript 'o'.
[b]Relative strengths are given as power output from a commercial Spectra-Physics model 2080-25S argon ion laser.
[c]These lines may oscillate simultaneously with the nearby strong line, but are not resolved easily in the output beam.

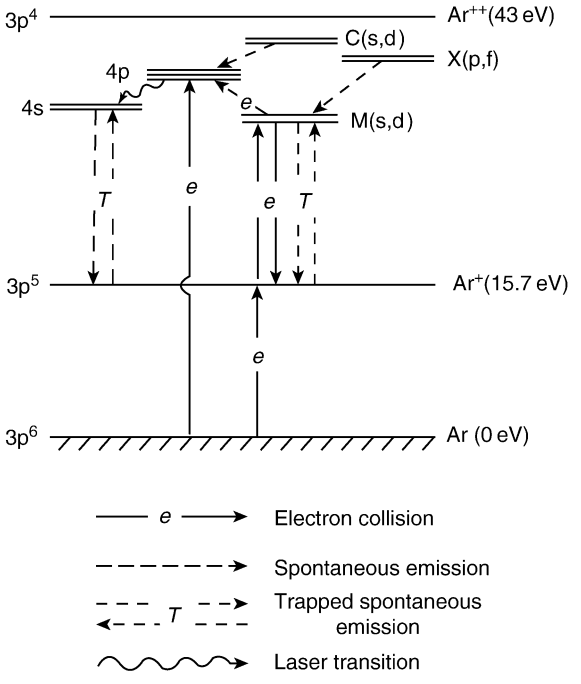


Figure 2 Schematic representation of energy levels in neutral and singly ionized argon, indicating alternative pathways for excitation and de-excitation of the argon ion laser levels.

are found in the Further Reading section. While a completely quantitative picture of argon ion laser operation is lacking to this day, the essential processes are known. Briefly, the 4p upper levels are populated by three pathways, as illustrated in Figure 2:

(i) by electron collision with the $3p^6$ neutral ground state atoms. This 'sudden perturbation process' requires at least 37 eV electrons, and also singles out the $4p\ ^2P^o_{3/2}$ upper level, which implies that only the 476 and 455 nm lines would oscillate. This is the behavior seen in pulsed discharges at very low pressure and very high axial electric field. This pathway probably contributes little to the 4p population under ordinary continuous wave (cw) operating conditions, however.

(ii) by electron collision from the lowest-lying s and d states in the ion, denoted M(s,d) in Figure 1. This only requires 3–4 eV electrons. These states have parity-allowed transitions to the ion ground state, but are made effectively metastable by radiation trapping (that is, there is a high probability that an emitted photon is re-absorbed by another ground state ion before it escapes the discharge region), or by requiring $|\Delta J|$ to be 2 to make the transition (forbidden by quantum selection rules). Thus, these levels are both created and destroyed primarily by electron collision, causing the population of the M(s,d) states to follow the population of the singly ionized ground state, which, in turn, is approximately proportional to the discharge current. Since a second electron collision is required to get from the M(s,d) states to the 4p states, a quadratic variation with current for the laser output power would be expected, and that is what is observed over some reasonable range of currents between threshold and saturation. Note that radiative decay from higher-lying opposite-parity p and f states, denoted X(p,f), can also contribute to the population of M(s,d), but the linear variation in population of the M(s,d) with discharge current is assured by electron collision creation and destruction.

(iii) by radiative decay from higher-lying opposite-parity s and d states, denoted C(s,d). These states are populated by electron collision with the $3p^5$ ion ground states, and thus have populations that also vary quadratically with discharge current. The contribution of this cascade process has been measured to be 20% to 50% of the 4p upper laser level population.

Note that it is not possible to distinguish between processes (ii) and (iii) by the variation of output power with discharge current; both give the observed quadratic dependence.

The radiative lifetimes of the 4s ^{2}P levels are sufficiently short to depopulate the lower laser levels by radiative decay. However, radiation trapping greatly lengthens this decay time, and a bottleneck

can occur. In pulsed ion lasers, this is exhibited by the laser pulse terminating before the excitation current pulse ends, which would seem to preclude continuous operation. However, the intense discharge used in continuous operation heats the ions to the order of 2300 K, thus greatly Doppler broadening the absorption linewidth and reducing the magnitude of the absorption. Additionally, the ions are attracted to the discharge tube walls, so that the absorption spectrum is further broadened by the Doppler shift due to their wall-directed velocities. The plasma wall sheath gives about 20 V drop in potential from the discharge axis to the tube wall, so most ions hit the wall with 20 eV of energy, or about ten times their thermal velocity. A typical cw argon ion laser operates at ten times the gas pressure that is optimum for a pulsed laser, and thus the radiation trapping of the $4s \rightarrow 3p^5$ transitions is so severe that it may take several milliseconds after discharge initiation for the laser oscillation to begin.

As the discharge current is increased, the intensities of the blue and green lines of Ar II eventually saturate, and then decrease with further current. At these high currents, there is a buildup in the population of doubly ionized atoms, and some lines of Ar III can be made to oscillate with the appropriate ultraviolet mirrors. Table 2 lists the strongest of these lines, those that are available in the largest commercial lasers. Again, there is no quantitative model for the performance in terms of the discharge parameters, but the upper levels are assumed to be populated by processes analogous to those of the Ar II laser. At still higher currents, lines in Ar IV can be made to oscillate as well.

Much less research has been done on neon, krypton, and xenon ion lasers, but it is a good assumption that the population and depopulation processes are the same in these lasers. Table 3 lists both the Kr II and Kr III lines that are available from the largest commercial ion lasers. Oscillation on lines in still-higher ionization states in both krypton and xenon have been observed.

Operating Characteristics

A typical variation of output power with discharge current for an argon ion laser is shown in Figure 3. This particular laser had a 4 mm diameter discharge in a water-cooled silica tube, 71 cm in length, with a 1 kG axial magnetic field. The parameter is the argon pressure in the tube before the discharge was struck. Note that no one curve is exactly quadratic, but that the envelope of the curves at different filling pressures is approximately quadratic. At such high discharge current densities (50 A in the 4 mm tube is approximately 400 A/cm^2) there is substantial pumping of gas out of the small-bore discharge region. Indeed, a return path for this pumped gas must be provided from anode to cathode ends of the discharge to keep the discharge from self-extinguishing. The axial electric field in this discharge was 3–5 V/cm, so the input power was of the order of 10 to 20 kW, yielding an efficiency of less than 0.1%, an unfortunate characteristic of all ion lasers.

Table 2 Ultraviolet argon ion laser wavelengths

Wavelength (nanometers)	*Spectrum*	*Transition*[a] *(upper level) → (lower level)*	*Relative strength*[b] *(Watt)*
275.392	III	$(^2D^o)4p\ ^1D_2 \rightarrow (^2D^o)4s\ ^1D^o_2$	0.3
275.6	?	?	0.02
300.264	III	$(^2P^o)4p\ ^1P_1 \rightarrow (^2P^o)3d\ ^1D^o_2$	0.5
302.405	III	$(^2P^o)4p\ ^3D_3 \rightarrow (^2P^o)4s\ ^3P^o_2$	0.5
305.484	III	$(^2P^o)4p\ ^3D_2 \rightarrow (^2P^o)4s\ ^3P^o_1$	0.2
333.613	III	$(^2D^o)4p\ ^3F_4 \rightarrow (^2D^o)4s\ ^3D^o_3$	0.4
334.472	III	$(^2D^o)4p\ ^3F_3 \rightarrow (^2D^o)4s\ ^3D^o_2$	0.8
335.849	III	$(^2D^o)4p\ ^3F_2 \rightarrow (^2D^o)4s\ ^3D^o_1$	0.8
350.358	III	$(^2D^o)4p\ ^3D_2 \rightarrow (^2D^o)4s\ ^3D^o_2$	0.05
350.933	III	$(^4S^o)4p\ ^3P_0 \rightarrow (^4S^o)4s\ ^3S^o_1$	0.05
351.112	III	$(^4S^o)4p\ ^3P_2 \rightarrow (^4S^o)4s\ ^3S^o_1$	2.0
351.418	III	$(^4S^o)4p\ ^3P_1 \rightarrow (^4S^o)4s\ ^3S^o_1$	0.7
363.789	III	$(^2D^o)4p\ ^1F_3 \rightarrow (^2D^o)4s\ ^1D^o_2$	2.5
379.532	III	$(^2P^o)4p\ ^3D_3 \rightarrow (^2P^o)3d\ ^3P^o_2$	0.4
385.829	III	$(^2P^o)4p\ ^3D_2 \rightarrow (^2P^o)3d\ ^3P^o_1$	0.15
390.784	III	$(^2P^o)4p\ ^3D_1 \rightarrow (^2P^o)3d\ ^3P^o_0$	0.02
408.904	IV?	?	0.04
414.671	III	$(^2D^o)4p\ ^3P_2 \rightarrow (^2P^o)4s\ ^3P^o_2$	0.02
418.298	III	$(^2D^o)4p\ ^1P_1 \rightarrow (^2D^o)4s\ ^1D^o_2$	0.08

[a]All levels are denoted in $L-S$ coupling with the core shown in ( ). Odd parity is denoted by the superscript 'o'.
[b]Relative strengths are given as power output from a commercial Spectra-Physics model 2085-25S argon ion laser.

Table 3 Krypton ion laser wavelengths

Wavelength (nanometers)	*Spectrum*	*Transition*[a] *(upper level) → (lower level)*	*Relative strength*[b] *(Watt)*
337.496	III	$(^2P^o)5p\ ^3D_3 \rightarrow (^2P^o)5s\ ^3P^o_2$	–
350.742	III	$(^4S^o)5p\ ^3P_2 \rightarrow (^4S^o)5s\ ^3S^o_1$	1.5
356.432	III	$(^4S^o)5p\ ^3P_1 \rightarrow (^4S^o)5s\ ^3S^o_1$	0.5
406.737	III	$(^2D^o)5p\ ^1F_3 \rightarrow (^2D^o)5s\ ^1D^o_2$	0.9
413.133	III	$(^4S^o)5p\ ^5P_2 \rightarrow (^4S^o)5s\ ^3S^o_1$	1.8
415.444	III	$(^2D^o)5p\ ^3F_3 \rightarrow (^2D^o)5s\ ^1D^o_2$	0.3
422.658	III	$(^2D^o)5p\ ^3F_2 \rightarrow (^2D^o)4d\ ^3D^o_1$	–
468.041	II	$(^3P)5p\ ^2S^o_{1/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.5
476.243	II	$(^3P)5p\ ^2D^o_{3/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.4
482.518	II	$(^3P)5p\ ^4S^o_{3/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.4
520.832	II	$(^3P)5p\ ^4P^o_{3/2} \rightarrow (^3P)5s\ ^4P_{3/2}$	–
530.865	II	$(^3P)5p\ ^4P^o_{5/2} \rightarrow (^3P)5s\ ^4P_{3/2}$	1.5
568.188	II	$(^3P)5p\ ^4D^o_{5/2} \rightarrow (^3P)5s\ ^2P_{3/2}$	0.6
631.024	III	$(^2D^o)5p\ ^3P_2 \rightarrow (^2P^o)4d\ ^3D_1$	0.2
647.088	II	$(^3P)5p\ ^4P^o_{5/2} \rightarrow (^3P)5s\ ^2P_{3/2}$	3.0
676.442	II	$(^3P)5p\ ^4P^o_{1/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.9
752.546	II	$(^3P)5p\ ^4P^o_{3/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	1.2
793.141	II	$(^1D)5p\ ^4F^o_{7/2} \rightarrow (^3P)4d\ ^2F_{5/2}$	0.3
799.322	II	$(^3P)5p\ ^4P^o_{3/2} \rightarrow (^3P)4d\ ^4D_{1/2}$	

[a]All levels are denoted in $L-S$ coupling with the core shown in ( ). Odd parity is denoted by the superscript 'o'.
[b]Relative strengths are given as power output from a commercial Spectra-Physics model 2080RS ion laser.

Technology

With such high input powers required in a small volume to produce several watts output, ion laser performance has improved from 1964 to the present only as new discharge technologies were introduced. The earliest laboratory argon ion lasers used thin-walled (≈1 mm wall thickness) fused silica discharge tubes, cooled by flowing water over the outside wall of the tube. The maximum input power per unit length of discharge was limited by thermal stresses in the silica walls caused by the temperature differential from inside to outside. Typically, ring-shaped cracks would cause the tube to fail catastrophically. Attempts to make metal–ceramic structures with alumina (Al_2O_3) discharge tubes to contain the plasma were made early on (1965) but were not successful. Such tubes invariably failed from fracture by thermal shock as the discharge was turned on. Later, successful metal–ceramic tubes were made with beryllia (BeO), which has a much higher thermal conductivity than silica or alumina and is much more resistant to thermal shock. Today, all of the lower power (less than 100 mW output) are made with BeO discharge tubes. Some ion lasers in the 0.5 to 1 W range are also made with water-cooled BeO discharge tubes.

A typical low-power air cooled argon ion laser is shown in Figure 4. The large metal can (2) on the right end of the tube contains an impregnated-oxide hot cathode, heated directly by current through ceramic feed-through insulators (3). The small (≈1 mm diameter) discharge bore runs down the center of the BeO ceramic rod (1), and several smaller diameter gas return path holes run off-axis parallel to the discharge bore to provide the needed gas equalization between cathode can and anode region. A copper honeycomb cooler is brazed to the cathode can, two more to the outer wall of the BeO cylinder, and one to the anode (4) at the left end of the tube. The laser mirrors (5) are glass-fritted to the ends of the tube, forming a good vacuum seal. Note that in this small laser, the active discharge bore length is less than half of the overall length.

The very early argon ion lasers were made with simple smooth dielectric tubes, allowing the continuous variation in voltage along the length of the tube required by the discharge longitudinal electric field. However, this variation can be step-wise at the discharge walls, and still be more or less smooth along the axis. Thus, the idea arose of using metal tube segments, insulated one from another, to form the discharge tube walls. The first version of this idea (1965) used short (≈1 cm) metal cylinders supported by metal disks and stacked inside a large diameter silica envelope, with each metal cylinder electrically isolated from the others. The tubes and disks were made of molybdenum, and were allowed to heat to incandescence, thus radiating several kilowatts of heat through the silica vacuum envelope to a water-cooled collector outside. While this eliminated the problems of thermal shock and poor thermal conductivity inherent in dielectric wall discharges, it made another

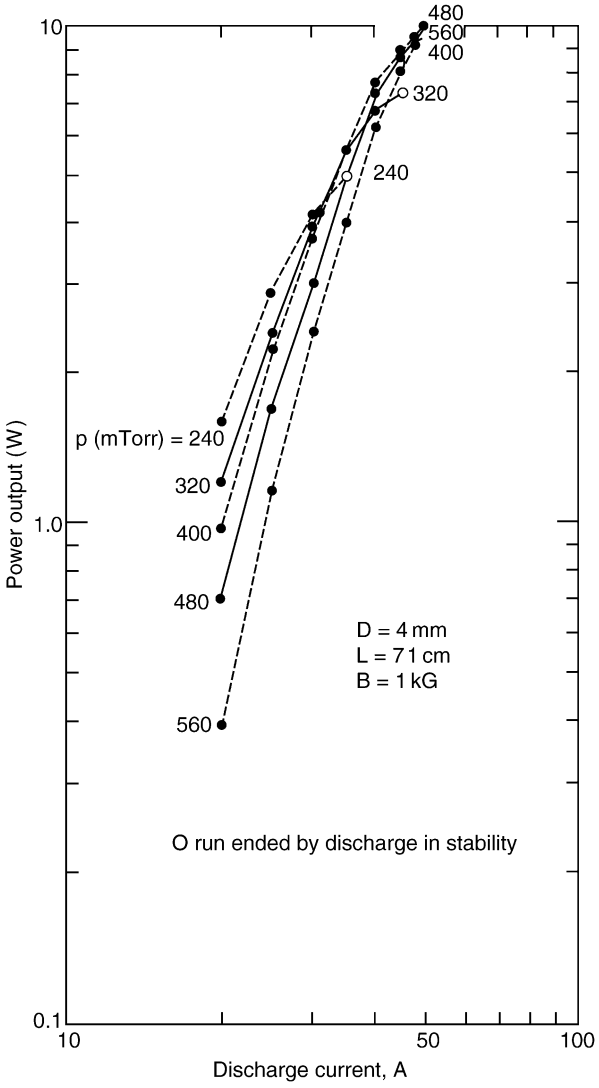


Figure 3 Laser output power (summed over all the blue and green laser lines) versus discharge current for a laser discharge 4 mm in diameter and 71 cm long, with a 1 kilogauss longitudinal magnetic field. The parameter shown is the argon fill pressure in mTorr prior to striking the discharge. The envelope of the curves exhibits the quadratic variation of output power with discharge current.

problem painfully evident. The intense ion bombardment of the metal tube walls sputtered the wall material, eventually eroding the shape of the metal cylinders and depositing metal films on the insulating wall material, thus shorting one segment to another.

Many different combinations of materials and configurations were investigated in the 1960s and 1970s to find a structure that would offer good laser performance and long life. It was found that simple thin metal disks with a central hole would effectively confine the discharge to a small diameter (on the order of the hole size) if a longitudinal d-c magnetic field of the order of 1 kiloGauss were used. The spacing between disks can be as large as 2–4 discharge diameters. Of the metals, tungsten has the lowest sputtering yield for argon ions in the 20 eV range (which is approximately the energy they gain in falling to the wall across the discharge sheath potential difference). An even lower sputtering yield is exhibited by carbon, and graphite cylinders contained within a larger diameter silica or alumina tube were popular for a while for ion laser discharges. Unfortunately, graphite has a tendency to flake or powder, so such laser discharge tubes became contaminated with 'dust' which could eventually find its way to the optical windows of the tube. Beryllia and silica also sputter under argon ion bombardment, but with still lower yields than metals or carbon; however, their smaller thermal conductivities limit them to lower-power applications.

The material/configuration combination that has evolved for higher-power argon ion lasers today is to use a stack of thin tungsten disks with 2–3 mm diameter holes for the discharge. These disks, typically 1 cm in diameter, are brazed coaxially to a larger copper annulus (actually, a drawn cup with a 1 cm hole on its axis). A stack of these copper/tungsten structures is, in turn, brazed to the inside wall of a large diameter alumina vacuum envelope. The tungsten disks are exposed to and confine the discharge, while the copper cups conduct heat radially outward to the alumina tube wall, which, in turn, is cooled by fluid flow over its exterior. Thus, the discharge is in contact only with a low-sputtering material (tungsten) while the heat is removed by a high thermal conductivity material (copper). Details differ among manufacturers, but this 'cool disk' technology seems to have won out in the end. A photo of a half-sectioned disk/cup stacked assembly from a Coherent Innova™ ion laser is shown in Figure 5. The coiled impregnated-tungsten cathode is also shown.

Sputtering of the discharge tube walls is not uniform along the length of the gas discharge. The small diameter region where the laser gain occurs is always joined to larger diameter regions containing cathode and anode electrodes (as shown in Figure 5, for example). A plasma double sheath (that is, a localized increase in potential) forms across the discharge in the transition region between large and small diameter regions (the discharge 'throats,' which may be abrupt or tapered). This sheath is required to satisfy the boundary conditions between the plasmas of different temperatures in the different regions. Such a double sheath imparts additional energy to the ions as they cross the sheath, perhaps

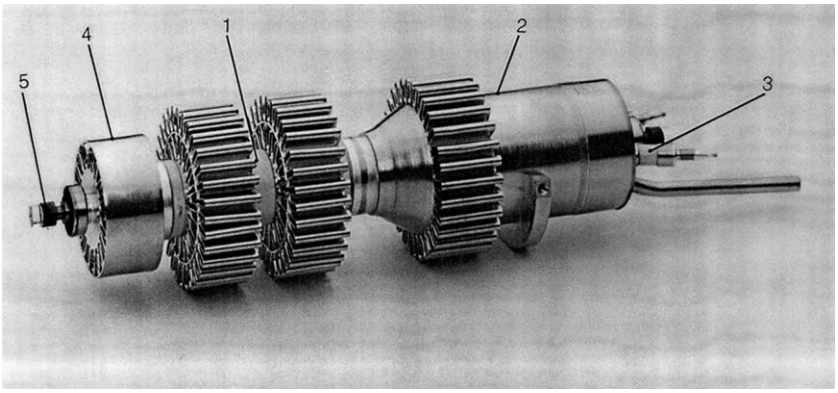


Figure 4 A typical commercial low-power, air-cooled argon ion laser. The cathode can (2) is at the right and the beryllia bore (1) and anode (4) is at the left. The cathode current is supplied through ceramic vacuum feed-throughs (3). Mirrors (5) are glass fritted onto the ends of vacuum envelope. (Photo courtesy of JDS Uniphase Corp.)

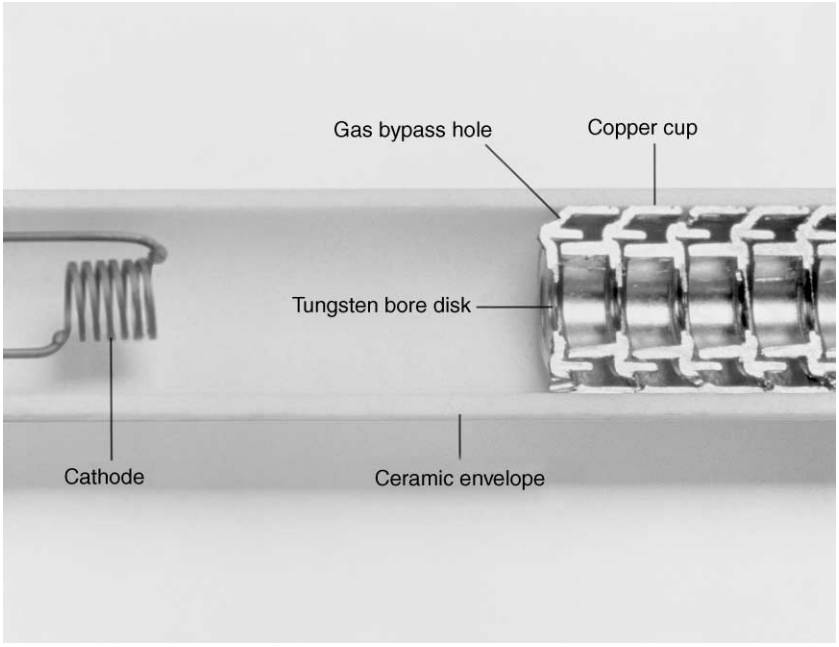


Figure 5 Discharge bore structure of a typical commercial high-power argon ion laser, the Coherent Innova™. Copper cups are brazed to the inner wall of a ceramic envelope, which is cooled by liquid flow over its outer surface. A thin tungsten disk with a small hole defining the discharge path is brazed over a larger hole in the bottom of the copper cup. Additional small holes in the copper cup near the ceramic envelope provide a gas return path from cathode to anode. Also shown is the hot oxide-impregnated tungsten cathode. (Photo courtesy of Coherent, Inc.)

an additional 20 eV. When these ions eventually hit the discharge walls near the location of the sheath, they have 40 eV of energy rather than the 20 eV from the normal wall sheath elsewhere in the small diameter discharge. Sputtering yield (number of sputtered wall atoms per incident ion) is exponentially dependent on ion energy in this low ion energy region, so the damage done to the wall in the vicinity of the 'throat' where the double sheath forms may be more than ten times that elsewhere in the small diameter bore region. This was the downfall of high-power operation of dielectric discharge bores, even BeO; while sputtering was acceptable elsewhere in the discharge, the amount of material removed in a small region near the discharge throat would cause catastrophic bore failure at that point. This localized increase in sputtering in the discharge throat is common to all

ion lasers, including modern cooled tungsten disk tubes. Eventually, disks near the throat are eroded to larger diameters, and more material is deposited on the walls nearby. Attempts to minimize localized sputtering by tapering the throat walls or the confining magnetic field have proven unsuccessful; a localized double sheath always forms somewhere in the throat. Because of the asymmetry caused by ion flow, the double sheath is larger in amplitude in the cathode throat than the anode throat, so that the localized wall damage is larger at the cathode end of the discharge than at the anode end.

Another undesirable feature of sputtering is that the sputtered material 'buries' some argon atoms when it is deposited on a wall. This is the basis of the well-known Vac-Ion® vacuum pump. Thus, the operating pressure in a sealed laser discharge tube will drop during the course of operation. In low-power ion lasers, this problem is usually solved by making the gas reservoir volume large enough to satisfy the desired operating life (for example, the large cathode can in Figure 4). In high-power ion lasers, the gas loss would result in unacceptable operating life even with a large reservoir at the fill pressure. Thus, most high-power ion lasers have a gas pressure measurement system and dual-valve arrangement connected to a small high-pressure reservoir to 'burp' gas into the active discharge periodically to keep the pressure within the operating range. Unfortunately, if a well-used ion laser is left inoperative for months, some of the 'buried' argon tends to leak back into the tube, with the gas pressure becoming higher than optimum. The pressure will gradually decrease to its optimum value with further operation of the discharge (in perhaps tens of hours). In the extreme case, the gas pressure can rise far enough so that the gas discharge will not strike, even at the maximum power supply voltage. Such a situation requires an external vacuum pump to remove the excess gas, usually a factory repair.

In addition to simple dc discharges, various other techniques have been used to excite ion lasers, primarily in a search for higher power, improved efficiency and longer operating life. Articles in the Further Reading section give references to these attempts. Radio-frequency excitation at 41 MHz was used in an inductively coupled discharge, with the laser bore and its gas return path forming a rectangular single turn of an air-core transformer. A commercial product using this technique was sold for a few years in the late 1960s. Since this was an 'electrode-less' discharge, ion laser lines in reactive gases such as chlorine could be made to oscillate, as well as the noble gases, without 'poisoning' the hot cathode used in conventional dc discharge lasers. A similar electrode-less discharge was demonstrated as a quasi-cw laser by using iron transformer cores and exciting the discharge with a 2.5 kHz square wave. Various microwave excitation configurations at 2.45 GHz and 9 GHz also resulted in ion laser oscillation. Techniques common to plasma fusion research were also studied. Argon ion laser oscillation was produced in Z-pinch and Θ-pinch discharges and also by high-energy (10–45 keV) electron beams. However, none of these latter techniques resulted in a commercial product, and all had efficiencies worse than the simple dc discharge lasers.

It is interesting that the highest-power output demonstrations were made in less than ten years after the discovery. Before 1970, 100 W output on the argon blue-green lines was demonstrated with dc discharges two meters in length. In 1970, a group in the Soviet Union reported 500 W blue-green output from a two-meter discharge with 250 kW of dc power input, an efficiency of 0.2%. Today, the highest output power argon ion laser offered for sale is 50 W output.

Manufacturers

More than 40 companies have manufactured ion lasers for sale over the past four decades. This field has now (2004), narrowed to the following:

Coherent, Inc.	http://www.coherentinc.com
INVERSion Ltd.	http://inversion.iae.nsk.su
JDS Uniphase	http://www.jdsu.com
Laser Physics, Inc.	http://www.laserphysics.com
Laser Technologies GmbH	http://www.lg-lasertechnologies.com
LASOS Lasertechnik GmbH	http://www.LASOS.com
Lexel Laser, Inc.	http://www.lexellaser.com
Melles Griot	http://lasers.mellesgriot.com
Spectra-Physics, Inc.	http://www.spectraphysics.com

Further Reading

Bridges WB (1979) Atomic and ionic gas lasers. In: Marton L and Tang CL (eds) *Methods of Experimental Physics, Vol 15: Quantum Electronics, Part A*, pp. 31–166. New York: Academic Press.

Bridges WB (1982) Ionized gas lasers. In: Weber MJ (ed.) *Handbook of Laser Science and Technology, Vol. II, Gas Lasers*, pp. 171–269. Boca Raton, FL: CRC Press.

Bridges WB (2000) Ion lasers – the early years. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 885–898.

Davis CC and King TA (1975) Gaseous ion lasers. In: Goodwin DW (ed.) *Advances in Quantum Electronics*, vol. 3, pp. 169–469. New York: Academic Press.

Dunn MH and Ross JN (1976) The argon ion laser. In: Sanders JH and Stenholm S (eds) *Progress in Quantum Electronics*, vol. 4, pp. 233–269. New York: Pergamon.

Weber MJ (2000) *Handbook of Laser Wavelengths*. Boca Raton, FL: CRC Press.

Optical Fiber Lasers

G E Town, Macquarie University, NSW, Australia

N N Akhmediev, Australian National University, ACT, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Optical fiber lasers were first demonstrated in the 1960s, and since then have developed to become versatile optical sources with many desirable properties. Aided by developments in associated technologies, such as fiber design and fabrication methods, semiconductor pump diode technology, and fiber-coupled and in-fiber components such as Bragg grating filters, optical fiber lasers now compete with other laser technologies in many applications, from telecommunications to materials processing.

An optical fiber laser is fundamentally an optical oscillator, which converts input pump power to coherent optical output power at one or more well-defined wavelengths. Optical oscillators require two basic elements: optical gain, and optical feedback. For sustained oscillation to occur the round-trip gain in the laser cavity must be unity, and the round-trip phase a multiple of 2π. In optical fiber lasers the optical gain is provided by an optical fiber amplifier using one or a combination of fundamental physical processes, e.g., stimulated emission, stimulated scattering, or nonlinear parametric processes. Early fiber amplifiers used stimulated Raman scattering to produce optical gain; however, rare-earth-doped fiber amplifiers, in which gain is provided by stimulated emission, are now more common. The reader is referred to **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems, for more details on doped fiber amplifiers, and to **Scattering:** Stimulated Scattering, **Nonlinear Optics, Applications:** Raman Lasers, **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed), and **Scattering:** Raman Scattering, for more details on stimulated scattering and parametric gain processes. Optical feedback may be provided in two fundamental ways, e.g., by using a closed ring of fiber, or from reflections from nonuniformities and discontinuities in the waveguide such as in Bragg grating filters or at fiber ends. The reader is referred to **Fiber Gratings** for more details on fiber Bragg gratings.

The main advantages of optical fiber lasers are derived from the confinement of the pump and signal in a small optical waveguide. In contrast with bulk lasers, the pump intensity in the fiber waveguide is largely independent of the laser length, resulting in large amplifier gain and low laser threshold, even for gain media with small absorption and emission cross-sections. The large gain enables lossy elements such as optical fiber-coupled devices and bulk optical elements to be incorporated into the laser cavity, providing additional control over the optical signal being generated. To obtain a large gain optical fiber amplifiers must usually be relatively long (i.e., from several centimeters to many meters), and so the linear and nonlinear properties of the fiber waveguide can have a significant influence on the optical signal being generated, resulting in some interesting and useful phenomena, e.g., as in soliton fiber lasers. The reader is referred to **Fiber and Guided Wave Optics:** Optical Fiber Cables, for fiber-based components, to **Fiber and Guided Wave Optics**: Dispersion, Light Propagation, and Nonlinear Effects (Basics), for a review of the linear and nonlinear properties of optical fibers, and to **Solitons:** Soliton Communication Systems, and Temporal Solitons, for the theory and applications of temporal solitons.

Whilst optical fibers may be fabricated in a range of materials, including polymer and crystalline materials, most fibers and fiber lasers are currently made of silica containing rare-earth ion dopants – see **Fiber and Guided Wave Optics:** Fabrication of Optical Fiber, for details. Lasers in silica optical fiber have much in common with other glass-host lasers, including a wide range of potential pump and lasing wavelengths, and broadband absorption and gain due to homogeneous and inhomogeneous broadening of the lasing energy levels in the amorphous glass host. For example, erbium-doped amplifiers in alumino-silicate glass fibers at room temperature have lasing transitions with homogeneous and

Table 1 Commonly used pump and signal transitions in rare-earth doped fiber lasers

Rare earth dopant	*Pump wavelengths [nm]*	*Signal wavelengths [nm]*	*Excited state lifetime [ms]*	*Energy levels*
Praseodymium	480, 585	885, 1080	0.1	4
Neodymium	590, 800	920 (900–950)	0.5	3
		1060 (1055–1140)		4
Samarium	488	651	1.5	4
Holmium	455, 650, 1150	2040	0.6	3
Erbium	800, 975, 1480	1550 (1530–1610)	10	3
Thulium	790, 1210 (1060–1320)	1480 (1460–1520)		
		1850 (1650–2050)	0.3	3
Ytterbium	920 (840–980)	975	0.8	3
		1040 (1010–1160)		4

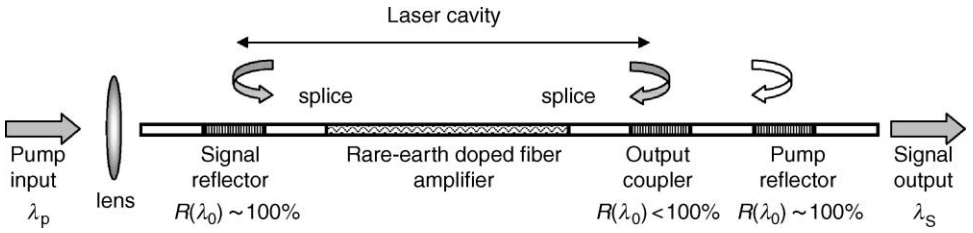


Figure 1 Example of a simple all-fiber optical oscillator. The rare-earth doped fiber provides optical gain, and the Bragg gratings provide narrowband optical feedback.

inhomogeneous linewidths of several nanometers, and can be optimized to provide around 30 dB gain over an 80 nm bandwidth centered around 1570 nm. The large gain-bandwidth is useful for achieving either wide tunability of the lasing wavelength and/or ultrashort pulse generation. Table 1 summarizes the most commonly used pump and signal transitions for a range of rare-earth dopants in silicate optical fibers.

Other desirable properties of silica fiber lasers include high efficiency, excellent thermal dissipation, substantial energy storage, high power capability, and compatibility with a wide range of optical fiber devices and systems. The characteristics of optical fiber lasers can be optimized for some very different applications, ranging from narrow linewidth and low-noise lasers to broadband pulsed lasers with high energy and/or high peak power. In the following sections the fundamentals of laser theory as applied to fiber lasers will be reviewed, highlighting the reasons underlying such versatility.

Fiber Laser Fundamentals

An example of a simple all-fiber laser is shown schematically in Figure 1. The laser contains the basic elements of optical gain (e.g., a rare-earth doped fiber amplifier) and optical feedback (e.g., a pair of Bragg grating filters). In the laser shown, the two Bragg gratings provide optical feedback only over a narrow range of wavelengths, which further restricts the range of possible lasing wavelengths. One grating must have a reflectivity less than unity to allow a proportion of the lasing signal to be coupled out of the laser cavity. At threshold, the gain in the amplifier exactly compensates for the loss in the output coupler.

The optical pump, typically from a semiconductor laser diode, is absorbed as is propagates along the fiber amplifier, e.g., by rare-earth ion dopants, which are raised to an excited state. The absorbed energy is stored and eventually emitted at a longer wavelength by either spontaneous or stimulated emission. Spontaneous emission produces unwanted noise and is associated with a finite excited state lifetime, whilst stimulated emission produces optical gain provided the gain medium is 'inverted', i.e., with more active ions in the upper lasing energy level than the lower lasing energy level. In practice some energy is also lost to nonradiative (thermal) emissions, e.g., in the transfer of energy between the pump and upper lasing energy levels. In an ideal gain medium the nonradiative transfer rate between the pump and upper lasing energy levels is much larger than the spontaneous or stimulated emission rates, so the inversion may usually be assumed to be independent of nonradiative transfer rates. Energy level diagrams for three- and four-level atoms are shown in Figure 2.

The local optical gain coefficient (i.e. power gain per unit length, in nepers[2]) in a fiber amplifier is proportional to the local inversion, i.e., $g(z) \sim \sigma_S \Delta N(z)$, where σ_S is the emission cross-section at the

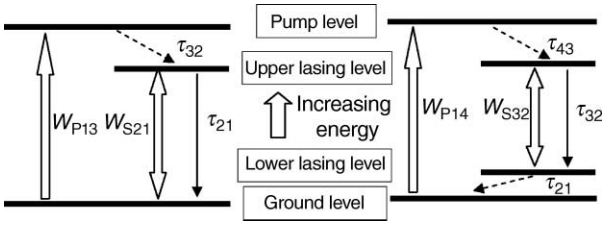


Figure 2 Energy level diagrams for three- and four-level laser transitions. *W* represents the probability of an optical transition per unit time associated with stimulated absorption or emission, and $1/\tau$ is the rate of spontaneous emission. Dotted lines represent nonradiative transitions (e.g., between the pump and upper lasing level).

signal wavelength, and $\Delta N(z)$ the difference in population density between the upper and lower lasing energy levels ($\Delta N > 0$ if the gain medium is inverted). The local inversion can be calculated by solving the rate equations including all stimulated and spontaneous transitions between the laser energy levels. The end-to-end amplifier power gain, $G = P_S(L_A)/P_S(0)$, in length L_A of amplifying fiber is then

$$G = \exp\left[\int_0^{L_A} g(z)\,\mathrm{d}z\right] \qquad [1]$$

which in decibels is

$$G = 4.34\int_0^{L_A} g(z)\,dz$$

A large gain coefficient is desirable for amplification with minimal added noise (i.e., low noise figure), and a large gain per unit pump power is desirable for low threshold lasing. For a given gain medium the only factor usually accessible to control the inversion and hence gain is the pump rate, i.e., the rate at which pump energy is absorbed by the gain medium, which for an unbleached amplifier is proportional to pump intensity. For example, in an ideal four-level unsaturated laser medium with uniform pump intensity, the small signal gain is approximately

$$\int_0^{L_A} g(z)\,\mathrm{d}z \approx \sigma_S \frac{P_{Pabs}\tau}{h\nu_P A} \qquad [2]$$

in which τ is the lifetime of the upper lasing level, P_{Pabs} is the pump power absorbed in the fiber amplifier with cross-sectional area A, h is Planck's constant, and ν_P is the pump frequency.

For a three-level laser under the same conditions, the small signal gain is

$$\int_0^{L_A} g(z)\,\mathrm{d}z \approx -\sigma_S N L_A + \sigma_S \frac{2P_{Pabs}\tau}{h\nu_P A} \qquad [3]$$

in which the first term corresponds to ground state absorption of the signal, in which N is the dopant ion density, and the factor of 2 in the second term is due to the fact that in an ideal three-level system every absorbed pump photon increases the inversion, ΔN, by two. Equation [3] can be approximated by multiplying [2] by the factor $(P_{Pabs} - P_{Psat})/(P_{Pabs} + P_{Psat})$, where $P_{Psat} = h\nu_P A/(\sigma_P\tau)$ is the pump power which must be absorbed to reduce the pump absorption coefficient to half its unpumped value (or the signal absorption coefficient to zero), and σ_P is the pump absorption cross-section. Unlike four-level lasers in which gain is available as soon as pump power is applied, three-level lasers require half the active ions to be excited before the inversion becomes positive and gain is produced.

The main advantages of optical fiber amplifiers are now clear; in an optical fiber waveguide the light is confined in a very small cross-sectional area, typically less than ten microns diameter. Consequently even a small coupled pump power can have a large intensity and produce a large gain coefficient, or low laser threshold. Even if the pump absorbed per unit length in the fiber amplifier is small (e.g., due to low dopant concentration or low absorption coefficient), this can often be compensated by using a long length of fiber with small intrinsic loss. Furthermore, rare-earth dopants have relatively long metastable state lifetimes, τ, which further assists in producing a large inversion throughout the fiber, and hence large overall gain or low lasing threshold, even at low pump powers. For example, a typical four-level neodymium doped fiber amplifier providing gain at $\lambda_S = 1.06\ \mu m$, pumped at $\lambda_P = 0.8\ \mu m$, with core diameter $7\ \mu m$, $\tau = 0.5$ ms, $\sigma_S = 1.4 \times 10^{-20}\ cm^2$, and $\sigma_P = 23 \times 10^{-21}\ cm^2$, could theoretically provide a small signal gain of 0.3 dB for every milliwatt of absorbed pump power.

Whilst the above analysis highlights the main advantages of fiber amplifiers, it is approximate in a number of respects. For example, eqns [2] and [3] assume both pump and signal have a uniform intensity distribution across the gain medium, and completely overlap. In single mode optical fibers the intensity distribution across the fiber core is not uniform but approximately Gaussian, with a width and effective area that depends on the numerical aperture of the waveguide and the wavelength. Consequently the overlap of the pump and signal beams with the doped fiber core and with each other is less than 100%, and the pump and signal intensities are higher in the center of the fiber than at the core-cladding boundary. The nonuniform distribution of pump and signal beams slightly modifies both the gain efficiency and gain saturation behavior

(i.e., the dependence of gain on signal power) in fiber amplifiers. More accurate analysis would also take into account factors such as the rare-earth dopant distribution (which may not be uniform), the variation of pump and signal intensities along the fiber, spectral variation in the absorption and emission cross-sections, spatial and spectral hole-burning in the gain medium, degeneracies in the energy levels, excited state absorption and other loss mechanisms, temperature, and spontaneous emission noise.

Continuous Wave Fiber Lasers

The ideal continuous wave (cw) laser converts input pump power with low coherence to a highly coherent optical output signal which is constant in amplitude and wavelength, with the spectral purity of the output (i.e., linewidth) limited only by cavity losses. Practical continuous wave lasers may be characterized by four parameters: the pump power required for the onset of oscillation (i.e., threshold); the conversion efficiency of pump to signal power above threshold; the peak wavelength of the optical output; and the spectral width of the optical output. Continuous wave fiber lasers often have a low threshold, high efficiency, and narrow linewidth relative to other types of laser. Output powers approaching 100 watts have been achieved using specialized techniques.

The threshold power is determined by the total cavity loss and gain efficiency (i.e., gain per unit of absorbed pump power) of the fiber amplifier. For example, for the laser shown in Figure 1 the internal losses are minimal, and the pump power required to reach threshold may be calculated by setting the product of the round-trip small-signal gain, and the output mirror reflectivity, G^2R, equal to unity. For the laser configuration shown in Figure 1, and using eqns [1] and [2] for a four-level laser, the pump power which must be absorbed in the amplifier for the laser to reach threshold is

$$P_{\text{th}} = \alpha_{\text{cav}} L_{\text{C}} \frac{h\nu_{\text{P}} A}{\sigma_{\text{S}} \tau} \qquad [4]$$

in which L_{C} is the cavity length, and $\alpha_{\text{cav}} = \alpha_{\text{int}} + \frac{1}{2} L_{\text{C}} \ln(1/R)$ is the total cavity loss per unit length, comprising both internal losses, α_{int}, and outcoupling losses through a mirror with reflectivity R. For example, if the internal losses were negligible and the transmittance, $T = 1 - R$, of the output reflector in the laser shown in Figure 1 was 50% (i.e. 3 dB), and the other fiber parameters were the same as above (i.e. $G = 0.3$ dB/mW), then the laser threshold would be 5 mW. In three-level fiber lasers ground state absorption of the signal is usually the dominant loss mechanism which must be overcome before a net gain is produced, and hence the threshold of three-level lasers is given approximately by eqn [4] with $\alpha_{\text{int}} = \sigma_{\text{S}} N L_{\text{A}} 2 L_{\text{C}}$, and increases with amplifier length. In either case the pump power required at the input of the laser to reach threshold, $P_{\text{P}}(0)$, may be approximated using the relation $P_{\text{Pabs}} = P_{\text{P}}(0) - P_{\text{P}}(L_{\text{A}}) \approx P_{\text{P}}(0)[1 - \exp(-\sigma_{\text{P}} N L_{\text{A}})]$.

For either three- or four-level lasers pumped above threshold, the inversion of the gain medium remains at its threshold level as the internal round-trip gain must remain unity; however a coherent optical output signal at the peak emission wavelength grows from noise until its amplitude is limited by saturation of the amplifier gain. Provided the output coupling loss is not too large (e.g., $T < 60\%$) such that the total signal intensity is constant along the whole length of the fiber amplifier, it can be shown that changes in pump power above threshold cause a proportional change in the signal power coupled out of the Fabry–Perot cavity in one direction, P^{+}_{Sout}, i.e.

$$P^{+}_{\text{Sout}} = \frac{(1-R)}{\delta_0} \frac{\nu_{\text{S}}}{\nu_{\text{P}}} (P_{\text{Pabs}} - P_{\text{th}}) \qquad [5]$$

in which $\delta_0 = 2\alpha_{\text{int}} L + \ln 1/R$ represents the total round-trip loss at threshold. The relationship between absorbed pump power and output signal power is shown schematically in Figure 3. The constant of proportionality is known as the slope or conversion efficiency, defined as $\eta_{\text{S}} = P_{\text{Sout}}/(P_{\text{Pabs}} - P_{\text{th}})$. In both three- and four-level fiber lasers with small intrinsic losses $\alpha_{\text{int}} \approx 0$, and small output coupling (i.e., $R \approx 1$), the slope efficiency can approach the intrinsic quantum efficiency, $\nu_{\text{S}}/\nu_{\text{P}}$.

Sustained oscillation can only occur if the optical path length around the cavity is an integral number of wavelengths. Wavelengths satisfying this condition are called 'modes' of the cavity, determined by

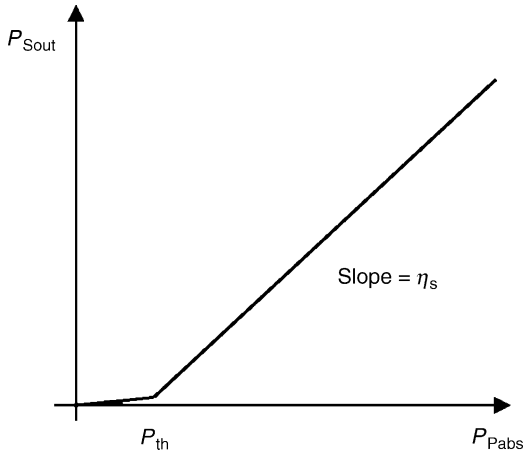


Figure 3 Representative plot of laser output power versus absorbed pump power.

$2nL_C = m\lambda_0$, where n is the effective refractive index seen by light propagating in the cavity, L_C is the cavity length, m is an integer, and λ_0 is the free-space wavelength. In Figure 1, the cavity length comprises the amplifying fiber, and two Bragg grating filters which may be regarded as having an effective length (i.e., less than the actual grating length) which depends on the grating characteristics. In fiber lasers the cavity is typically between several centimeters and many meters in length, hence there are usually very many potential lasing modes within the gain-bandwidth of the amplifying medium. For example, in the Fabry–Perot cavity of Figure 1, if the cavity length was 5 meters, the mode spacing would be $\nu_{m+1} - \nu_m = c/(2nL_C) = 20$ MHz, i.e. orders of magnitude less than the bandwidth of the optical amplifier.

Ideally the first mode to reach threshold would determine the laser output wavelength and the spectral width, or linewidth, of the laser output would be $\Delta\nu_L = h\nu/(2\pi\tau_C^2 P_{Sout})$, in which τ_C is the lifetime of photons in the cavity and determined by cavity losses, including outcoupling, and given by $\tau_C = t_R/\varepsilon$, where t_R is the cavity round trip time and ε the fractional energy loss in the cavity per round trip. In practice, the linewidth is usually significantly larger than the latter theoretical limit due to transient lasing of different modes within the bandwidth of the optical amplifier and environmental perturbations to the fiber cavity. Fiber amplifiers typically have a large homogeneously broadened gain-bandwidth, and being long and flexible are susceptible to acoustic and thermal perturbations, hence care must be taken to physically stabilize the laser cavity to minimize both mode-hopping and the output linewidth.

Whilst the CW laser shown in Figure 1 has the advantage of simple construction, it would not be ideal for narrow linewidth CW generation. The reason is that counterpropagating waves in the Fabry–Perot cavity form a standing wave, i.e., in which the local signal intensity varies due to interference between forward and backward propagating waves. The standing wave can cause spatial hole-burning in the gain medium (i.e., a spatial variation in gain, linked to the spatial variation in intensity) which in turn reduces the average gain for the lasing mode and promotes mode-hopping and hence spectral broadening of the output.

A preferable arrangement for many fiber lasers is a traveling wave laser, which can be realized using a ring configuration, as shown in Figure 4. In this configuration the optical isolator ensures unidirectional lasing and avoidance of spatial hole-burning. Additional components in the cavity can include a wavelength selective coupler (WSC), i.e., to couple the pump wavelength into the laser cavity whilst not coupling the signal wavelength out, and a narrowband filter (e.g. a small Fabry–Perot resonator) to further stabilize and/or narrow the lasing linewidth. Rare-earth doped fiber lasers with similar cavity configurations have been realized with linewidths of approximately 10 kHz, limited by environmental (e.g., acoustic) perturbations.

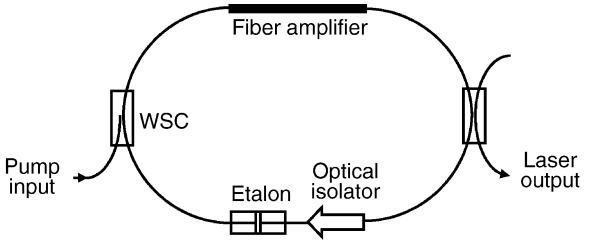


Figure 4 Schematic of a traveling wave laser, useful for low noise continuous wave oscillation. The wavelength selective coupler (WSC) is required to couple pump power into the laser cavity without coupling signal out.

Pulsed Fiber Lasers

There are two main types of fiber laser useful for generating short high-power pulses: Q-switched lasers, and mode-locked lasers. Q-switched fiber lasers are useful for generating large energy pulses (e.g., microjoules) with very high peak power (e.g., kilowatts) with relatively long pulse duration (e.g., tens of nanoseconds), whilst mode-locked fiber lasers are typically capable of generating ultrashort pulses (e.g., sub-picosecond) with moderate energy (e.g., nanojoules) and moderate-to-high peak power (e.g., tens of watts). Pulsed lasers typically produce less average output power than CW lasers (e.g., up to about 10 W). Both types of pulsed laser typically contain an element for loss modulation within the cavity, however the pulse generation mechanisms are very different.

Q-switched lasers operate by rapid switching of the cavity loss. Whilst the intracavity loss is high and the laser below threshold, energy is transferred from the pump to the lasing medium. The long excited state lifetime and small emission cross-sections of rare-earth dopants assists greatly in the latter process, so that a significant amount of energy can be stored in long fiber amplifiers before amplification of spontaneous emission noise begins to deplete the inversion. After the gain medium is fully inverted the cavity loss is suddenly reduced, and the laser is taken well above threshold, resulting in a rapid build-up of noise and the formation of a pulse which ideally extracts all the stored energy within a few round-trips in the laser cavity. The situation is shown schematically in Figure 5.

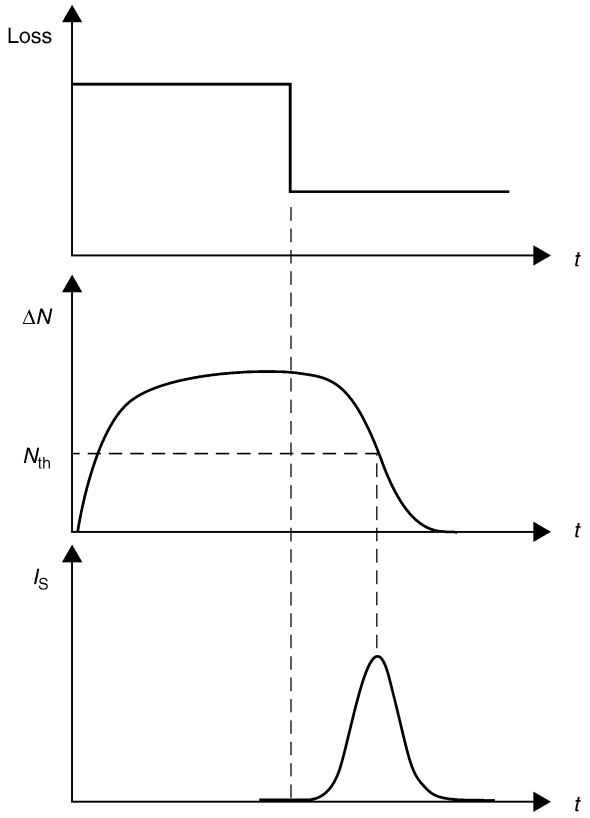


Figure 5 Pulse generation in a *Q*-switched laser. Initially the intracavity loss is high, the laser remains below threshold, and the inversion of the gain medium increases. When the intracavity loss is suddenly reduced, the laser goes well above threshold and spontaneous emission noise is rapidly amplified into a large pulse which extracts much of the energy previously stored in the gain medium.

An upper limit on the pulse energy obtainable from a *Q*-switched laser may be determined by assuming all active ions in the laser cavity are excited before the pulse is generated (i.e., at the time the cavity Q is switched high), that no ions remain excited immediately after the pulse is generated, and that the pulse generated is much shorter than the time required to pump all the ions into their excited state. The energy stored in the gain medium, and hence the maximum conceivable pulse energy, would then be $E = h\nu_S NV$, where N is the density of active ions and V is the cavity volume. For example, a typical erbium-doped fiber laser of length 10 m and dopant density $N = 10^{19}$ ions/cm in a single-mode fiber with core diameter 7 μm could ideally store up to 500 μJ of energy.

Two things limit the maximum inversion and energy storage achievable in practice: spontaneous emission noise (due to the finite excited-state lifetime) which is amplified and depletes the inversion, and unwanted feedback (e.g., due to the finite extinction ratio of the Q-switching element) which results in cw lasing which limits the inversion. Following from eqns [1] and [2], the unidirectional gain per unit of stored energy in a fiber laser may be expressed as $G = 4.34/(A_{core}F_{sat}) = 4.34/E_{sat}$ decibels per joule, where $F_{sat} = h\nu_S/\sigma_S$ is the saturation fluence, and $E_{sat} = A_{sat}F_{sat}$ is the saturation energy. For example, if the maximum gain achievable before the onset of cw lasing was 30 dB (i.e., even when the Q-switch introduces a large loss), then the maximum energy storable in the gain medium would be $E = 6.9E_{sat}$. For the same erbium-doped fiber parameters as used in the previous paragraph, the maximum storable energy would be $E \approx 70$ μJ.

Only a proportion of the energy stored in the gain medium may usually be converted into a Q-switched output pulse. The actual Q-switched pulse energy, duration, and peak power can be calculated by solving the rate equations for the photon flux and the population difference between the lasing energy levels in the gain medium. The equations must usually be solved numerically, however assuming the Q changes much more rapidly than the pulse build-up time, which is in turn much shorter than the time taken for a significant change in inversion to either pumping or spontaneous emission, the pulse parameters may be approximated analytically and are found to be determined by only two parameters, i.e., the population inversion (or stored energy) just before Q-switching, and the lifetime of photons within the cavity, τ_C, defined previously.

Under the conditions described above, standard analysis gives the Q-switched pulse duration as

$$T_P \approx \frac{\tau_C(r-1)}{(r-1-\ln r)} \qquad [6]$$

where r is the ratio of the inversion just before Q-switching (low Q) to the threshold inversion for cw lasing after Q-switching (Q high). In fiber lasers the gain can be very high, hence it is common for $r \gg 1$ and almost all available energy stored in the gain medium is extracted as a Q-switched laser pulse. The Q-switched pulse duration is then determined primarily by the cavity lifetime; shorter pulses can be obtained only by shortening the laser cavity and/or increasing the cavity loss. For example, for a 10 m long ring cavity in which the sum of cavity and outcoupling losses is 10 dB and $r \gg 1$, the pulse duration would be $T_P\tau_C = 22$ ns.

The exact proportion of energy stored in the gain medium that is converted to a Q-switched pulse depends on the number of energy levels in the lasing medium, the degeneracy of those levels, and the rate

of nonradiative energy transfer between those levels. For example, for an ideal three-level gain medium (e.g., erbium dopant) with no internal cavity losses the maximum available Q-switched pulse energy is half the stored energy, or $E_{\text{pulse}} = E/2$. It can be shown that the peak of the output pulse power occurs as the inversion decreases through threshold, and may be approximated by dividing the pulse energy by the pulse duration. Even for moderate pulse energies the peak power is potentially very high. For the erbium-doped fiber laser parameters used previously, the peak pulse power would be approximately 1600 W, i.e., sufficient to cause significant spectral broadening of the output wavelength due to stimulated Raman scattering and other nonlinear effects within the optical fiber.

Mode-locked lasers generate optical pulses by forcing many cavity modes to be locked in phase. As per Fourier theory, the pulse repetition rate is related to the spacing of the lasing modes, and the more modes that add coherently, the greater the pulse bandwidth and the shorter the transform-limited pulse duration. For example, the relationship between mode spacing and pulse repetition rate, and between the spectral and temporal envelopes of a train of Gaussian pulses is shown in Figure 6.

In mode-locked lasers a nonlinear or time-varying element must be present which synchronously modulates the optical signal circulating in the cavity in amplitude or phase at a frequency which is a harmonic of the cavity mode spacing. There are two main methods of mode locking:

(i) active mode locking (e.g., by an electro-optic modulator), in which amplitude or phase modulation is driven by an external source; and
(ii) passive mode locking (e.g., by a fast or slow saturable absorber), in which an element with nonlinear gain or loss responds to the optical signal in the laser cavity itself.

In both cases the modulation provided by the mode-locking element distributes energy between neighboring modes, effectively injection-locking them in phase. A variation of mode locking techniques often present in fiber lasers is soliton mode-locking, in which the active or passive modulation is weak, and a balance between nonlinearity and dispersion is the dominant process underlying pulse formation. Hybrid mode-locking techniques can also be used to combine the best aspects of different pulse-generation methods.

Whilst the latter frequency-domain view of mode locking is helpful, nonlinear effects are almost always significant in mode-locked fiber lasers, and hence the linear concept of a mode is not appropriate for detailed analysis. Hence it is usually preferable to describe pulse generation in mode-locked fiber lasers in the time domain; for stable operation the pulse at any point in the cavity must self-replicate after each round-trip, with the net effects of temporal and spectral loss, gain, dispersion and nonlinearity all canceling.

An optical pulse propagating in a fiber laser is usually described by a slowly varying complex envelope, $\Psi(z,t)$, where $|\Psi(z,t)|^2$ is the instantaneous power in a frame of reference moving at the pulse group velocity, and $\arg[\Psi(z,t)]$ is the instantaneous phase relative to the optical carrier. The various elements in the laser cavity, such as the passive fiber, optical fiber amplifier, filters, and mode-locking devices, each modify the complex pulse envelope as it propagates around the laser cavity.

For example, in the absence of other effects, group velocity dispersion (GVD) causes different wavelengths to travel at different velocities along the fiber, resulting in 'chirping' and temporal broadening of optical pulses by an amount proportional to the distance propagated. The effect can be described mathematically as the accumulation of a phase shift which is quadratic in frequency, i.e., $\Psi(z,\omega) = \Psi(0,\omega)\exp[i\beta_2\omega^2 z/2]$. Taking the derivative with respect to propagation distance, and using $\partial/\partial t = -i\omega$, the incremental change in the complex envelope

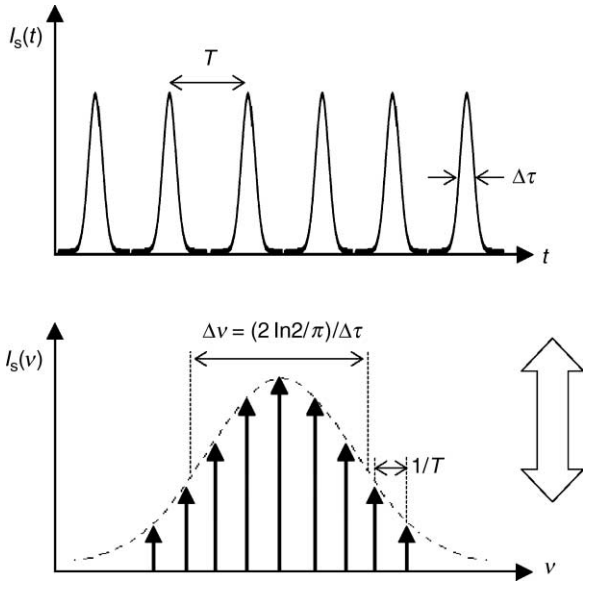


Figure 6 Schematic showing the Fourier-transform relationship between an infinite train of Gaussian pulses and the lasing modes in a mode-locked laser. The temporal and spectral widths are measured full width at half the maximum intensity.

with propagation distance is

$$\frac{\partial \Psi}{\partial z} = -\mathrm{i}\frac{\beta_2}{2}\frac{\partial^2 \Psi}{\partial t^2} \qquad [7]$$

If the GVD parameter, β_2, is positive then long wavelengths propagate faster than shorter wavelengths, and the dispersion is called 'normal'. If the reverse is true, the dispersion is 'anomalous'. In standard single-mode step index optical fibers the dispersion passes through zero at wavelengths around 1300 nm, and is anomalous for longer wavelengths. It is often necessary to control the net dispersion in a fiber laser, and this can be done by incorporating specially designed fibers with wavelength-shifted or otherwise modified dispersion characteristics, or by using chirped Bragg grating reflectors.

Whilst silica is not commonly regarded as a nonlinear medium (i.e., in which the optical properties are a function of the intensity of light), in ultrashort pulsed fiber lasers the peak powers are usually such that nonlinear effects cannot be ignored. In silica the main nonlinearity is associated with the Kerr effect, and is readily observed as an intensity-dependent refractive index, i.e. $n = n_0 + n_2 I$, where the nonlinear index, n_2, is typically about 3×10^{-20} m^2/W in silica fibers. Other nonlinear effects related to the Kerr nonlinearity which can sometimes be important are cross-phase-modulation and four-wave mixing. In the absence of other effects, the nonlinear index results in self-phase modulation (SPM), or a phase shift in the optical carrier which is proportional to the instantaneous intensity of the signal. Mathematically, $\Psi(z,t) = \Psi(0,t)\exp[\mathrm{i}\gamma|\Psi(0,t)|^2 z]$, in which $\gamma = n_2\omega_0/(cA)$ is the nonlinear coefficient of the fiber, ω_0 is the radian frequency of the optical carrier, and A is the effective core area for the guided mode. The incremental change per unit length in the complex pulse envelope induced by self-phase modulation is then

$$\frac{\partial \Psi}{\partial z} = \mathrm{i}\gamma|\Psi|^2\Psi \qquad [8]$$

For example, self-phase-modulation is used to realize a nonlinear switching element in the passively mode-locked laser cavity shown schematically in Figure 7. For obvious reasons the laser is called a 'figure-8' laser, though it is actually a ring laser into which a device known as a nonlinear amplifying loop-mirror (NALM) has been incorporated. The NALM acts as a fast saturable absorber, with transmittance dependent on the instantaneous optical intensity. The NALM operates as follows: light propagating through the optical isolator is divided equally into co- and counter-propagating waves in the loop-mirror on the right-hand side. The light propagating in a clockwise direction is amplified immediately, and propagates around the loop with high intensity, whilst the counter-clockwise propagating signal has a low intensity for most of its time in the loop. The difference in intensity means that the co- and counter-propagating waves experience different nonlinear phase shifts. When they are recombined at the coupler, high-intensity light is transmitted back into the ring, whilst low-intensity light is reflected back to the isolator, and lost. In this way the NALM provides both gain and a nonlinear transmittance which depends on the instantaneous power of the input, described by

$$T \approx G\left\{1 - \frac{1}{2}\left[1 + \cos\left(\frac{1}{2}\gamma L P_{\mathrm{in}} G\right)\right]\right\} \qquad [9]$$

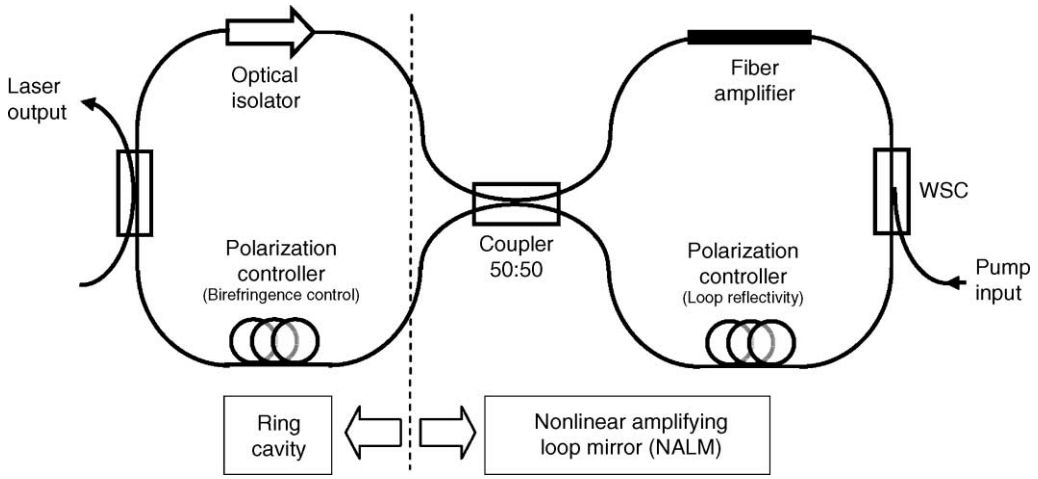


Figure 7 Schematic of a 'figure-8' fiber laser. The laser is passively mode-locked by the nonlinear amplifying loop mirror on the right, which acts as a fast saturable absorber, transmitting and amplifying high-power optical signals whilst reflecting low optical power signals.

in which G is the linear gain of the amplifier, and L the loop length, P_{in} is the input power, and T the ratio of transmitted power to input power. The nonlinear transmittance of the NALM promotes pulse formation and shortening in the laser cavity, and hence passively mode-locks the laser. The polarization controllers are to compensate for unwanted birefringence in the optical fiber, and to control the transmittance of the loop-mirror at low input powers. The fiber dispersion and nonlinearity, together with other factors, further influence the pulses generated in the laser, as described in the following paragraphs.

The combined effects of dispersion and self-phase-modulation can give rise to some interesting and useful physical phenomena. In lossless optical fibers, or fiber lasers in which loss is balanced by gain, pulse propagation is described by the nonlinear Schrödinger equation (NLSE), i.e., a nonlinear wave equation including the effects of both group velocity dispersion and intensity-dependent refractive index. The NLSE may be derived by combining eqns [8] and [9], i.e.

$$\mathrm{i}\frac{\partial \Psi}{\partial z} - \frac{\beta_2}{2}\frac{\partial^2 \Psi}{\partial t^2} + \gamma|\Psi|^2\Psi = 0 \qquad [10]$$

It can be shown that if the dispersion is anomalous ($\beta_2 < 0$) then for a particular shape of pulse envelope the distributed effects of dispersion and nonlinearity can exactly cancel, so $\partial\Psi/\partial z = 0$, and the pulse propagates without changing shape. Such pulses are called fundamental solitons, the solution for which is $\Psi(z,t) = \Psi_0\,\mathrm{sech}(t/T_0)\exp[\mathrm{i}z/(2L_D)]$, in which the peak amplitude is $\Psi_0 = 1/\sqrt{\gamma L_D}$, T_0 is the pulse width parameter, and $L_D = T_0^2/\beta_2$ is known as the dispersion length. For example, in a standard silica fiber at $\lambda = 1.55\ \mu\mathrm{m}$, $\beta_2 \approx -20\ \mathrm{ps}^2/\mathrm{km}$ and $\gamma \approx 10\ \mathrm{W}^{-1}\mathrm{km}^{-1}$, hence a soliton with $T_0 = 1$ ps would have peak power 2 W and energy $E_0 = 2|\beta_2|/(\gamma T_0) = 4$ pJ. The energy of a soliton is quantized, in that pulses with initial energy between 0.5 and 1.5 times E_0 can evolve into a fundamental soliton during propagation. Consequently solitons behave somewhat like particles, and are robust even in the presence of significant perturbations. The pulses which form spontaneously in the figure-8 fiber laser have many of the properties of solitons.

Mode-locked fiber lasers, such as that shown in Figure 7, can often be described by a generalized nonlinear Schrödinger equation, in which extra terms are included in eqn [10] to account for the average effects of an excess linear gain or loss, nonlinear gain and loss (e.g., due to a saturable absorber), spectral filtering, and temporal modulation, etc. Assuming the effects of the added elements can be regarded as distributed over the length of the cavity rather than lumped, such that the envelope of any pulse generated in the laser does not change shape significantly as it propagates around the cavity, then the master differential equation describing propagation of an optical signal within a fiber laser which is passively mode-locked by a fast saturable absorber can be written

$$\mathrm{i}\frac{\partial U}{\partial \zeta} + \left(\frac{D}{2} - \mathrm{i}\beta\right)\frac{\partial^2 U}{\partial \tau^2} + (1 - \mathrm{i}\varepsilon)|U|^2 U = \mathrm{i}\delta U - (\nu - \mathrm{i}\mu)|U|^4 U \qquad [11]$$

Equation [11] is known as the Ginzburg–Landau equation, which describes propagation of the normalized pulse envelope, $U(\zeta, \tau)$, as a function of delayed time, τ, and normalized distance, ζ, and under the combined influence of group velocity dispersion ($D = \pm 1$ for anomalous or normal group velocity dispersion, respectively), dispersion due to band-limiting by amplification in a homogeneously broadened gain medium ($\beta > 0$), nonlinear refractive index, nonlinear gain or loss (ε), linear gain (δ), and quintic terms, μ and ν, which if less than zero relate to saturation of the nonlinear gain and refractive index, respectively. Note that if the right-hand side of the equation is set to zero, and $\varepsilon = \beta = 0$, then the equation reduces to the normalized form of the nonlinear Schrödinger equation. Stable pulses generated within the laser are solutions of the above equation with $\partial U/\partial\zeta = 0$, and hence the effect of all terms must balance. Consequently it is possible for stable pulses to be generated even with normal dispersion in the cavity, though the pulses are then generally strongly chirped, and have an envelope which is approximately Gaussian shaped rather than the hyperbolic secant shape expected with anomalous dispersion. Stable soliton-like pulses can be found by solving the Ginzburg–Landau equation for specific ranges of parameters. There are many numerical solutions, and some analytical solutions, but in general the dynamics of pulse formation and propagation in passively mode-locked fiber lasers can be quite complex, and for some sets of parameters remarkable and unexpected numerical solutions can be found. One such example which has been observed experimentally is the 'exploding' soliton, which unlike the usual soliton is periodically unstable, as shown in the plot of pulse evolution in Figure 8.

Other Fiber Lasers

The lasers discussed in the previous sections are indicative of the main types of fiber laser in terms of their output characteristics; however it should be

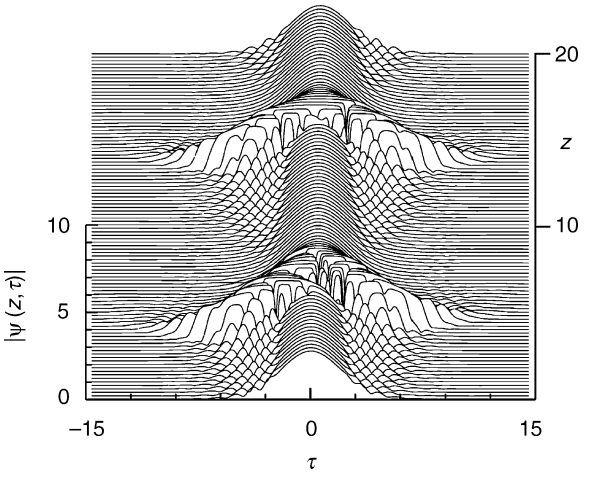


Figure 8 Two periods of the evolution of an exploding soliton found by numerical solution of the quintic Ginzburg–Landau equation with normalized parameters $\varepsilon = 1.0$, $\delta = 0.1$, $\beta = 0.125$, $\mu = 0.1$, and $\nu = 0.6$. The process never repeats itself exactly in successive 'periods'; however the pulse always returns to the same shape. Such solutions cannot be found in analytic form; nevertheless they are as common as stationary pulse solutions and exist for a wide range of parameters.

noted that there is a wide range of other fiber laser configurations with similar behaviors incorporating a variety of other fiber and/or bulk elements. For example, specially designed fibers in which the pump and/or signal has a large cross-sectional area can be used to realize fiber lasers with a high output power, comparable to that in many bulk laser systems. Linear and/or nonlinear processing of the output of fiber lasers can be used to increase their peak output power and/or pulse energy, or to change their output wavelength (i.e., by harmonic generation) from that achievable in the laser cavity itself. In mode-locked fiber lasers it is now common to include passive semiconductor devices which provide a slow saturable absorber action to initiate and stabilize soliton mode-locking.

It should also be noted that there are a number of fiber lasers which generate optical outputs using different or modified physical processes compared to those described in previous sections. For example, it is possible to realize up-conversion fiber lasers in which the output wavelength is shorter than the pump wavelength, multiwavelength fiber lasers (pulsed and cw) which produce outputs at several wavelengths simultaneously, and lasers with limited output coherence such as chaotic fiber lasers and broadband superfluorescent sources, etc. Lastly, whilst the rare-earth doped fiber lasers dealt with in previous sections have some very desirable properties, it should be recognized that optical fiber lasers based on other optical gain mechanisms, such as stimulated scattering and parametric effects, can also be realized (e.g., as in fiber Raman lasers, and fiber Brillouin lasers). In the latter cases the lasing wavelength is limited only by available pump wavelengths. The reader is referred to the references in the Further Reading for information on the wide range of fiber lasers that are possible in practice.

List of Units and Nomenclature

Absorption cross-section	σ_a	$[m^{-2}]$
Amplifier length	L_A	[m]
Area	A	$[m^2]$
Cavity length	L_C	[m]
Complex amplitude	Ψ	$[W^{0.5}]$
Dispersion length	L_D	[m]
Distance	z	[m]
Emission cross-section	σ_e	$[m^{-2}]$
Frequency	ν	[Hz]
Gain coefficient	g	$[m^{-1}]$
Group velocity dispersion	β_2	$[s^2m^{-1}]$
Inversion ratio	r	
Lifetime	τ	[s]
Loss coefficient	α	$[m^{-1}]$
Nonlinear coefficient	γ	$[m^{-1}W^{-1}]$
Nonlinear refractive index	n_2	$[m^2W^{-1}]$
Number density	N	$[m^{-3}]$
Power gain	G	
Pump power	P_P	[W]
Pump wavelength	λ_P	[m]
Reflectivity	R	
Saturation energy	E_{sat}	[J]
Saturation fluence	F_{sat}	$[J.m^{-2}]$
Signal power	P_S	[W]
Signal wavelength	λ_S	[m]
Threshold power	P_{th}	[W]

See also

Fiber and Guided Wave Optics: Dispersion; Fabrication of Optical Fiber; Light Propagation; Nonlinear Effects (Basics); Optical Fiber Cables. **Fiber Gratings**. **Lasers:** Up Conversion Lasers. **Nonlinear Optics, Applications:** Raman Lasers. **Optical Amplifiers:** Basic Concepts; Erbrium Doped Fiber Amplifiers for Lightwave Systems. **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed). **Solitons:** Temporal Solitons. **Scattering:** Raman Scattering; Stimulated Scattering. **Ultra-Fast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Agrawal GP (2001) *Nonlinear Fiber Optics*, 3rd edn. San Diego, CA: Academic Press.

Agrawal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego, CA: Academic Press.
Akhmediev NN and Ankiewicz A (1997) *Solitons*. London: Chapman and Hall.
Becker PC, Olsson NA and Simpson JR (1999) *Erbium-Doped Fiber Amplifiers: Fundamentals and Technology*. San Diego, CA: Academic Press.
Desurvire E (1994) *Erbium-Doped Fiber Amplifiers: Principles and Applications*. New York: Wiley.
Digonnet MJF (ed.) (2001) *Rare-Earth-Doped Fiber Lasers and Amplifiers*, 2nd edn. New York: Marcel Dekker.
Duling IN (ed.) (1995) *Compact Sources of Ultrashort Pulses*. Cambridge, UK: Cambridge University Press.
France PW (ed.) (1991) *Optical Fibre Lasers and Amplifiers*. Glasgow, UK: Blackie.
Jeunhomme LB (1990) *Single-Mode Fiber Optics: Principles and Applications*, 2nd edn. New York: Dekker.
Kashyap R (1999) *Fiber Bragg Gratings*. San Diego, CA: Academic Press.
Siegman AE (1986) *Lasers*. Oxford, UK: Oxford University Press.
Weber MJ (ed.) (1995) *CRC Handbook of Laser Science and Technology*. Boca Raton, FL: CRC Press.

Organic Semiconductors and Polymers

G A Turnbull, University of St Andrews, Fife, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Organic semiconductors, and conjugated polymers in particular, are emerging as an interesting new class of visible laser media. The polymers are electroluminescent, broadband visible emitters and exhibit large optical gain coefficients. As plastics, they are amenable to simple processing, and can be shaped into complex structures. This combination of photonic, electronic, and processing properties make these materials extremely well suited to forming novel microlasers, with future prospects as inexpensive, plastic diode lasers operating throughout the visible spectrum.

Conjugated Polymers

Plastics are widely used in low-cost optical components such as lenses, diffraction gratings, optical fibers, and adhesives. Their common use derives from a combination of qualities: good optical transmission, a wide range of mechanical properties and simple fabrication via molding: extrusion or solution processing. Functional optical polymers, including liquid crystals and photorefractives, have also been widely applied to polarization control and for nonlinear- and electro-optics. Conjugated polymers are a remarkable class of functional plastics that exhibit some unusual properties. Firstly, they are plastics that can conduct electricity; and in fact are electrically semiconducting. Secondly, they are plastics that can be stimulated either optically or electrically to emit visible light.

These two unusual properties derive from a particular chemical structure specific to this family of plastics. In all conjugated polymers, the backbone of the molecule is a chain of carbon atoms joined by alternating single and double covalent bonds. The simplest example is illustrated in **Figure 1a**, which shows the chemical structure of the polymer poly(acetylene). Structural integrity of the polymer is provided by a string of single σ-bonds between adjacent carbon atoms. The additional bonding electrons in the double bonds are located in π-orbitals oriented perpendicular to the polymer chain. Adjacent π-orbitals overlap, and their electrons can be widely delocalized along the polymer backbone. The delocalized electrons can move relatively freely along the chain, giving the electronic properties of the material. The π-orbitals can be transformed between bonding and anti-bonding states by the absorption or emission of light, leading to strong optical transitions. While poly(acetylene) is a very inefficient light emitter, other conjugated polymers, notably those that include phenyl rings in the backbone (**Figure 1b–d**), can be highly emissive. Photoluminescence quantum

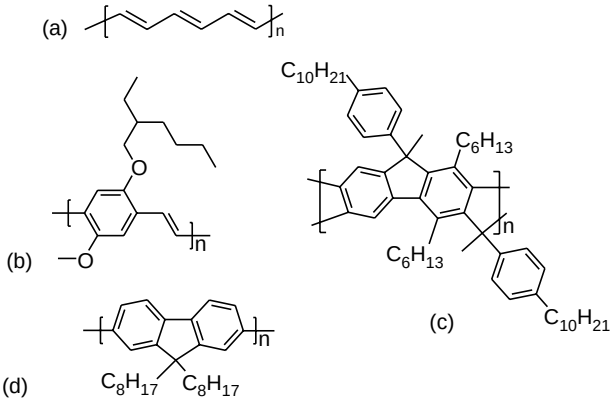


Figure 1 Chemical structures of several conjugated polymers. (a) poly[acetylene]; (b) poly[2-methoxy-5-(2′-ethylhexyloxy)-1,4-phenylene vinylene], (MEH-PPV); (c) ladder-type poly[1,4-phenylene], (MeLPPP); (d) poly[9,9-dioctylfluorene], (PFO).

efficiencies, approaching unity in solution and as high as 60% in the semiconducting solid state, have been realized.

Organic Semiconductor Lasers

While organic dye lasers have been widely used since the mid-1960s, the history of organic semiconductor lasers is much more recent. Other than a few early observations of lasing in molecular crystals in the 1970s by Karl and others, the field of organic semiconductor lasers began in earnest following the related discovery in 1990 by Burroughes *et al*. of electroluminescence in a conjugated polymer. In 1992, Moses demonstrated lasing for the first time in a solution of the conjugated polymer MEH-PPV (Figure 1b). However, it was another four years – once material synthesis had been sufficiently refined – before stimulated emission and then true laser action were observed in polymers in the semiconducting solid state. In the following year the first demonstrations were made of dye-doped small molecular semiconductor lasers. Since that time, there has been a great growth in interest in the field with an average of ~30 research papers per year. Further details of the development of the field of organic semiconductor lasers can be found in the Further Reading.

So what are the particular properties of conjugated polymers that have stimulated interest in developing organic semiconductor lasers? Firstly these materials are broadband visible emitters. Through appropriate chemical design, their emission can be tuned throughout the visible spectrum, from 400 nm to 700 nm (Figure 2). The polymers exhibit large optical gain coefficients, because the optical transitions are dipole-allowed. Compared with small molecular dyes, polymers exhibit a relatively low degree of self-absorption of the emitted light, and much reduced quenching of the luminescence with increased concentration. This means that it is possible to work with much higher chromophore densities than conventional laser dyes, which is advantageous for achieving very low threshold lasing in extremely compact resonators. Furthermore, because these materials are semiconducting, they have the potential for direct electrical excitation. This could lead to semiconductor diode lasers operating throughout the visible spectrum and, in particular, in the blue and green spectral bands where currently there are few commercially viable inorganic semiconductor sources. Finally, these materials are plastics and so are amenable to simple processing, and can be shaped and structured into complex feedback resonators.

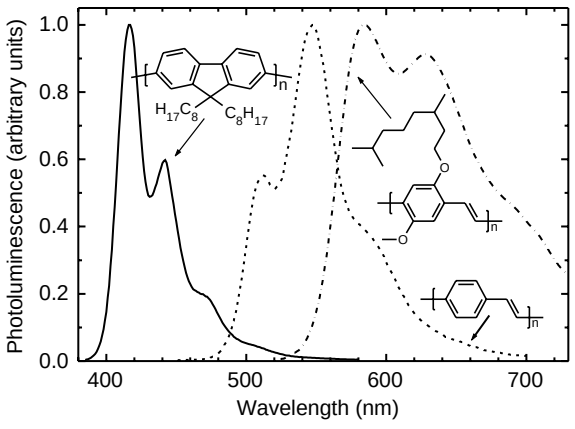


Figure 2 Photoluminescence spectra of three commonly used conjugated polymers; chemical structures inset.

Organic Semiconductor Gain Materials

A great advantage of organic laser media is that a very wide range of materials can be readily synthesized, and through minor changes in the molecular structure one can relatively simply tune the emission properties of the material. Many different organic semiconductors have been applied as lasers, though they fall into three major categories. These are conjugated polymers, dye-doped small molecular semiconductors, and molecular single crystals. Conjugated polymers make up the most diverse materials set, most notably including derivatives of poly(paraphenylene-vinylene), poly(paraphenylene), and poly(fluorene) (Figure 1b–d). The second category comprises small molecular semiconductors that are co-evaporated under vacuum with a dopant laser dye. In this blend of materials, the organic semiconductor functions simply as an energy transfer host for the emitting dye molecules. In the third category are organic single crystals, including materials such as polyacenes and thiophene oligomers. These materials have the advantage of having the highest electrical mobilities of any organic semiconductors, which is important for electrical excitation. However, they can be demanding to grow into high-quality single crystals, and so lack the key processing advantages of the other families of organic semiconductors.

Figure 3a shows a typical exciton energy level structure of a semiconducting polymer. Similar to a conventional laser dye, the energy levels comprise a manifold of singlet states and a manifold of triplet states, each with vibrational sublevels. The laser transition occurs from the bottom of the first excited singlet state S_1 to the vibronic overtones of the ground

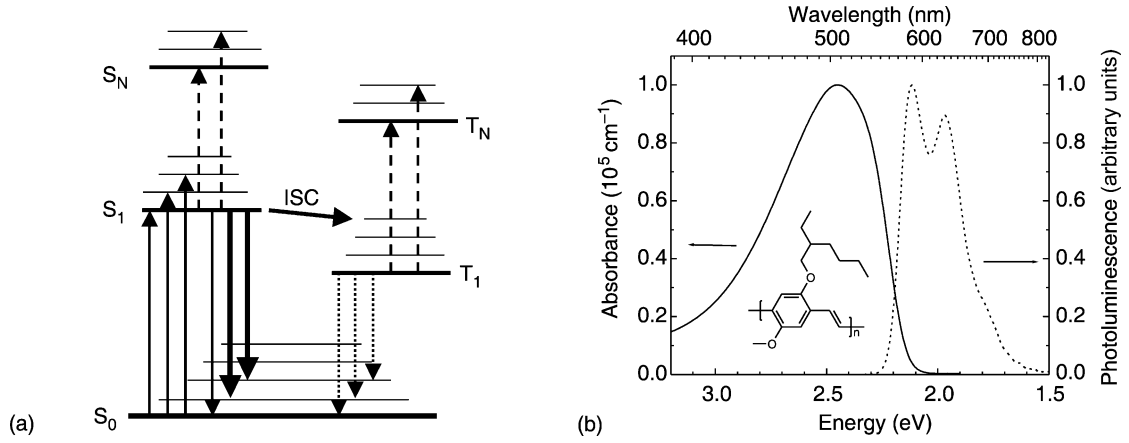


Figure 3 (a) Generic energy level structure and optical transitions of a conjugated polymer. S_N refers to the Nth singlet state; T_N refers to the Nth triplet state; ISC: intersystem crossing. (b) Absorption and photoluminescence spectra of the polymer MEH-PPV.

state S_0, so forming a classic 4-level laser system. The optical transition is dipole-allowed and so has a high oscillator strength. As a result, conjugated polymers typically exhibit simulated emission cross-sections in the range 10^{-16} cm^2 to 10^{-15} cm^2, while the population inversion lifetime is as short as a few hundred picoseconds.

There are, however, several other transitions that can compete with the stimulated emission process. There can be excited-state absorption from the S_1 singlet state into higher-lying singlet states. Additionally intersystem crossing from the S_1 state to the lowest excited triplet state T_1 can subsequently lead to excited triplet absorption. Other processes in the polymer can also impede the stimulated emission. Firstly, charged excitations can exhibit absorption bands that overlap with the emission. Secondly, at high excitation densities, exciton–exciton annihilation can significantly deplete the excited states available for stimulated emission. Ultimately, singlet–singlet annihilation is the limiting factor for the maximum excitation density possible, and while polymer films can have a repeat-unit density of $\sim 10^{21}$ cm^{-3}, the excited state density is typically limited to around 10^{18} cm^{-3}. In many materials this is not restrictive on achieving optical gain because stimulated emission can occur at densities of 10^{17} cm^{-3} and below.

An important factor that distinguishes the photophysics of semiconducting polymers from conventional laser dyes is an internal process of exciton migration between the absorption and emission of a photon. Excitations are confined to short straight sections of the polymer between kinks or twists in the chain. These straight sections, or 'sites', are typically a few repeat units in length, though there is a statistical spread of site lengths in any given material. When formed, excitons tend to hop to the longest, lowest-energy, sites before emitting a photon. The exciton migration process has two effects. First, it means the emission takes place from a subset of the chain segments, resulting in a lower degree of inhomogeneous broadening of the emission spectrum (vibronic peaks are better resolved than in absorption (see Figure 3b)). Second, the peaks of the absorption and emission spectra tend to be widely separated in wavelength, reducing the level of self-absorption at the lasing wavelength.

As mentioned above, blending materials is a particularly attractive approach towards further reducing self-absorption in organic semiconductor lasers. In such systems, a narrow bandgap dopant chromophore is blended with a concentration of 1–2% in a wider bandgap semiconducting host. Emission from the host is strongly quenched via nonradiative Förster energy transfer to the guest molecule. The blended material exhibits an absorption profile that is dominated by the host absorption, while the emission profile is almost completely that of the guest dopant. By strongly separating the emission peak from the absorption edge of the host, it is possible to achieve a substantially reduced (by perhaps an order of magnitude or more) residual absorption coefficient at the lasing wavelength. In doing so, the combined material acts much more like an ideal 4-level laser medium and can thereby have substantially lower pump thresholds for stimulated emission.

Measuring Gain

The dynamics of the excited state population in semiconducting polymers have been widely studied using femtosecond transient absorption experiments.

This is a pump–probe technique in which the transmission of the broadband probe pulse is compared in the presence and absence of a pump pulse that excites molecules into the first excited state. Such measurements have the time resolution of the pump pulses used and so can typically be ~100 fs. By systematically delaying the arrival of the broadband probe pulse, it is possible to map out the evolution of the gain dynamics across the emission spectrum of the material with very high time resolution.

Figure 4 shows a series of snapshots of the absorption induced by a pumping pulse in a poly(thiophene) derivative. This figure shows the changing absorption of the sample between 200 fs and 400 ps after excitation. Within the absorption band of the material, one finds that the level of absorption is slightly reduced because the pump pulse has partly depleted the ground state. In the emission band there is also a reduction in the absorption that mimics in shape the photoluminescence spectrum of the polymer. This negative absorption – or optical gain – shows direct evidence for stimulated emission through much of the emission band. In the low energy tail of the emission, however, there is an increase in absorption that arises due to the excited-state absorption process mentioned earlier. Clearly, for a good laser material, the spectra of excited-state and charge-induced absorptions should overlap as little as possible with the photoluminescence.

Many other investigations of gain in semiconducting polymer films have used a phenomenon known as spectral line narrowing to explore indirectly the amplifying properties of the medium. The experimental procedure for such studies is as follows. A thin film of the semiconducting material, of typically 100 nm thickness, is excited by a pulsed pump laser focused to a stripe of dimensions ~100 μm wide by a few millimeters in length. A significant fraction of the light emitted by the material is trapped in the film by total internal reflection at the polymer–air and polymer–substrate interfaces, and is thereby waveguided along the length of the excitation stripe. This waveguided spontaneous emission can be amplified by stimulated emission before being emitted out the edge of the film.

One finds that, above a particular threshold of pump energy, there is a superlinear increase in the output energy with pump intensity. Associated with this is a dramatic narrowing of the broad emission spectrum to a linewidth of typically less than 10 nm, as illustrated in Figure 5. These two features are signatures of gain narrowing via the process of amplified spontaneous emission. The line-narrowed feature usually appears at the first vibronic overtone of the optical transition, where the gain exceeds the losses in the material by the greatest amount. This particular wavelength experiences the greatest amplification and, above a particular pumping density, undergoes a runaway effect in which most of the light is stimulated to emit at that wavelength at the expense of the rest of the spectrum.

In the case of amplified spontaneous emission or mirrorless lasing as it is also known, the spontaneous emission guided along the length of the stripe acts as

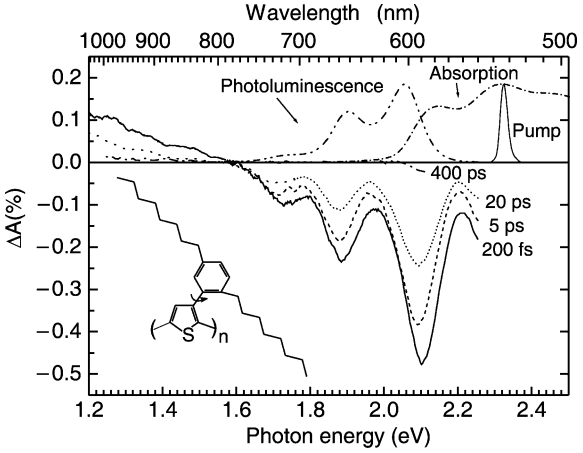


Figure 4 Transient absorption spectra of poly[3-(2,5-dioctylphenyl)-thiophene]; chemical structure inset. (Reproduced from Ruseckas A, Theander M, Valkunas L, Andersson MR, Ingänas O and Sundström V (1998) Energy transfer in a conjugated polymer with reduced inter-chain coupling. *Journal of Luminescence* 76 & 77: 474–477. Copyright (1998), with permission from Elsevier.)

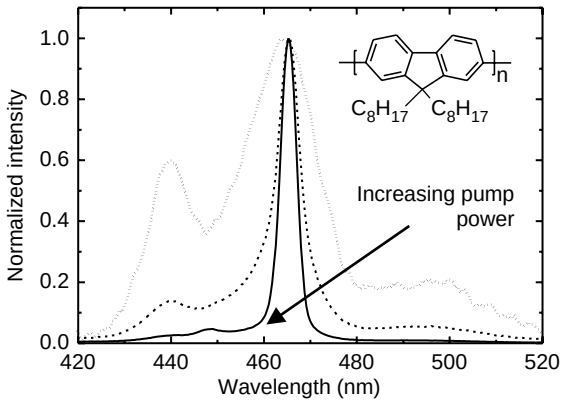


Figure 5 The change in spectral shape with increasing excitation density (from 10 to 133 μJ cm^{-2}) of the edge emission from a PFO film – the normalized spectra show a significant spectral line narrowing at high excitation densities. (Adapted from Heliotis G, Bradley DDC, Turnbull GA and Samuel IDW (2002) Light amplification and gain in polyfluorene waveguides. *Applied Physics Letters* 81: 415–417. Copyright (2002), with permission from American Institute of Physics.)

the 'probe pulse'. In these experiments it is therefore necessary to extract information indirectly about the net gain in the material. By integrating the spontaneous and stimulated emission along the length of the stripe, one finds that the wavelength-dependent output intensity $I(\lambda)$ of the spectral line narrowed light is given by the relationship

$$I(\lambda) \propto \frac{1}{g(\lambda)}[\exp[g(\lambda)l] - 1] \qquad [1]$$

where I_p is pumping intensity, l is the length of the stripe, and $g(\lambda)$ is the net gain coefficient. Therefore, by monitoring the intensity of the line-narrowed emission as a function of the length of the pump stripe, one can extract the value of the net gain for the material. Similarly, by progressively moving the stripe further and further away from the edge of the film, it is possible to evaluate the waveguide losses of light propagating through an unpumped region of the waveguide. Waveguide losses in conjugated polymers typically lie in the range of 3–50 cm^{-1}. Net gains can be as high as 60 cm^{-1} or 260 dB cm^{-1} at modest pumping densities of 4 kW cm^{-2}. These substantial gains, in very small propagation lengths, therefore highlight the attraction of waveguide-based resonators for polymer lasers. In the next section we will discuss the most commonly used feedback structures that have been used for polymer lasers.

Polymer Laser Resonators

The first demonstration of cavity lasing in a conjugated polymer was made by Moses in 1992. The resonator consisted simply of two flat dielectric mirrors on either side of a cuvette containing the polymer in solution. Lasing occurred at the peak wavelength of the gain profile. Subsequent developments in solid-state polymer lasers have involved more sophisticated resonators that exploit the capacity of these materials for very simple processing. For example, solution-processing methods such as spin-casting or dip-coating, make it possible to fabricate very high-quality optical waveguides. Wavelength-scale microstructures can also be readily formed, allowing the straightforward fabrication of laser microresonators.

Many different resonators have been explored for solid-state polymer lasers. Some of the key types are shown in Figure 6, including different geometries of both microcavities and distributed feedback structures. The first resonator used, by Tessler *et al.*, was a planar microcavity (Figure 6a), in which the polymer film was sandwiched between two mirrors. The structure supported standing-wave resonances perpendicular to the plane of the film, at three

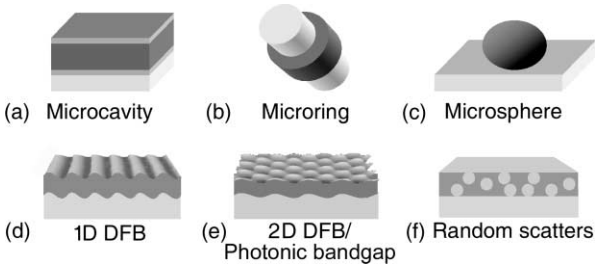


Figure 6 Resonators used for conjugated polymer lasers. (a) Planar microcavity; (b) annular microcavity; (c) spherical microcavity; (d) distributed feedback resonator; (e) 2D DFB/ photonic crystal resonator; (f) random laser.

wavelengths in the emission band of the polymer. Above laser threshold, light was preferentially stimulated into only one of these modes. Microcavities of other geometries have also been studied, including annular and spherical microresonators (Figure 6b,c), in which the optical field is confined in waveguide modes, whispering gallery modes, or a superposition of the two. Such structures provide a longer interaction length in the gain medium for stimulated emission; though tend to lase on many longitudinal modes and have no directional output beam.

The final key group of resonators are distributed feedback (DFB) structures, which researchers have progressively favored to attain the low pump threshold densities that will be required to realize electrically pumped lasing. Figure 6d shows a typical DFB laser structure. The device is based on a thin polymer film that is typically 100 nm in thickness deposited on top of a silica substrate. The silica–polymer–air structure forms an asymmetric slab waveguide that supports only the lowest order transverse electric mode TE_0. To provide the feedback, the substrate is modulated with a wavelength-scale periodic grating structure. These corrugations give rise to Bragg scattering of the waveguide mode. Through careful selection of the grating period, one may arrange that optical wavelengths close to the peak of the polymer gain will be Bragg scattered from the propagating TE_0 mode to the counter-propagating TE_0 mode. This is achieved when the Bragg equation

$$2n_{\text{eff}}\Lambda = m\lambda_B \qquad [2]$$

is satisfied, where Λ is the grating period, n_{eff} is the effective refractive index of the waveguide, m is an integer, and λ_B is the Bragg wavelength. In so doing, one may arrange that two counter-propagating waveguide modes can be coupled together, thereby providing the distributed resonant feedback for the laser process. Commonly in semiconducting polymer

lasers, a second-order (i.e., $m = 2$) Bragg grating has been favored to provide the feedback for the laser, so that the first-order scattering provides a surface-emitted output coupling, perpendicular to the plane of the guide. This surface-emitting geometry is particularly attractive in polymer lasers because it is relatively difficult to achieve a high optical quality edge to the waveguide. DFB lasers have the advantage over simple microcavity resonators in that they can combine long interaction lengths, limited by the scattering length of the Bragg grating, with excellent wavelength selection, which is controlled by the grating period.

In addition to conventional 1D DFB gratings, a number of more exotic 2D and 3D DFB, and photonic crystal resonators have been explored in semiconducting polymers. Even randomly located scattering sites in a polymer film have been used to provide a weak closed-loop cavity feedback. Examples of 2D gratings are given in Figure 7, which shows atomic force micrographs of an eggbox structure, formed from two orthogonal gratings, and a circular DFB grating. Such 2D DFB structures can significantly improve the operation of surface-emitting polymer lasers compared with 1D gratings. Improved feedback yields lower oscillation thresholds, combined with higher slope efficiencies and significantly improved beam quality. Indeed, such surface emitting DFB lasers can produce nearly diffraction limited laser beams.

Typical operating characteristics of an egg-box DFB polymer laser are shown in Figure 8. This laser, which is based on the material MEH-PPV, is transversely pumped with a pulsed pump laser focused onto a small region of the waveguide; up to 90% of the incident pump light is absorbed in the 100 nm thick film. The laser has a threshold pump energy of 4 nJ and an external slope efficiency of $>7\%$, and operates on a single frequency close to the Bragg wavelength of the periodic waveguide. Lasing does not occur exactly at the Bragg wavelength, because at this wavelength the propagation of light is suppressed by the periodic structure. This leads to a photonic stopband within which the emission is reduced. At the edges of this stopband the microstructure can support standing wave electric field patterns of the same spatial period. At these wavelengths there can be a strong interaction between the electric field and excited states in the gain medium, which leads to low-threshold band edge lasing as observed in the figure.

Although DFB resonators have advantages over microcavity devices, the complicated lithographic and etching steps involved in their fabrication can detract from the key material advantage of simple processing. To address this issue, several different techniques for simply replicating the complicated microstructures have been explored. These include UV embossing of flexible plastic substrates, and hot embossing and micromolding of the active

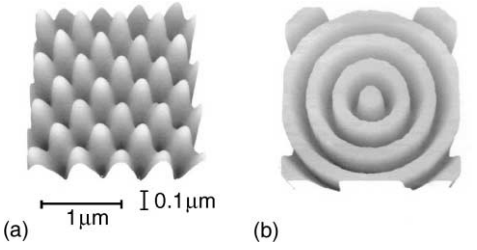


Figure 7 Atomic force microscope images of (a) 'egg-box' DFB grating, and (b) circular DFB grating; grating period is 400 nm in each structure.

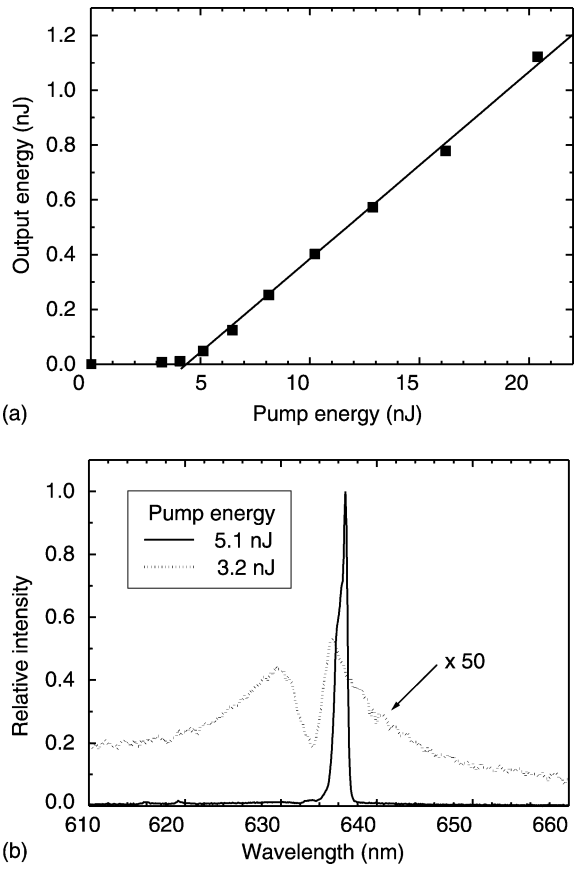


Figure 8 (a) Energy characteristic of an egg-box DFB laser; (b) normalized emission spectra above and below lasing threshold, showing lasing in the edge of the photonic stopband. (Reprinted from Turnbull GA, Andrew P, Barnes WL and Samuel IDW (2002) Operating characteristics of a semiconducting polymer laser pumped by a microchip laser. *Applied Physics Letters* 82: 313–315. Copyright (2003). With permission from American Institute of Physics.)

polymer itself. Such techniques can very readily reproduce structures with feature sizes as small as 100 nm and could allow future mass production of photonic microstructures in these materials.

Towards Plastic Diode Lasers

A typical experimental configuration for characterising optically pumped polymer lasers is shown in Figure 9. The laser is mounted in a vacuum chamber in order to isolate the polymer from oxygen and water. A pulsed visible or ultraviolet pump laser is attenuated and then focused onto the waveguide to transversely pump the polymer laser. The emission is collected using a CCD spectrograph for spectral measurements and energy meter and beam profilers for other characteristics. Typical pulsed pump lasers, which have been used, include nitrogen laser pumped dye lasers, and spectral harmonics of regeneratively amplified Ti^{3+}-sapphire lasers or flashlamp pumped Nd^{3+}:YAG lasers. Recent advances have shown the thresholds to be low enough to be successfully pumped with a modest Nd^{3+} pulsed microchip laser.

So while optically pumped polymer lasers are now proving to be compact, functional light sources, a major motivation for research in this field remains the prospect of electrically driven plastic diode lasers. Such lasers could be very low-cost, flexible light sources that would access any wavelength throughout the visible spectrum. In particular, there are exciting prospects of producing blue and green diode lasers in spectral regions currently difficult to access with inorganic semiconductors. The lowest threshold densities reported for optically pumped organic semiconductor lasers are of the order of 100 W cm^{-2}. To achieve similar excitation densities electrically, one requires pump current densities of $\sim$200 A cm^{-2}. This is substantially more than the few mA cm^{-2} commonly used in polymer LEDs for display applications. Presently, such high current densities are infeasible with continuous excitation, but can be exceeded with short current pulses of a few tens of nanoseconds duration.

Electroluminescence in conjugated polymers was first achieved in 1990 by Burroughes and co-workers. Since that time there has been remarkable progress made in the science and technology of organic LEDs. Commercial products are now on the market, and the feasibility of both flexible and full-color light-emitting displays, in which the individual pixels have been ink-jet printed, have been demonstrated in the laboratory. Despite such rapid progress in the field of organic LEDs, there have been no successful demonstrations to date of electrically pumped lasing in any organic semiconductor. In this section we highlight some of the challenges that must be faced if electrically pumped lasing is to be achieved. To understand the problems associated with electrical driving of organic semiconductor lasers, it is useful first to review the generic structure of organic light-emitting diodes.

A generic organic LED structure is shown in Figure 10. The device consists of one or more organic layers sandwiched between two electrical contacts. Usually a transparent anode is used, commonly made from ITO, which allows light to be emitted from the front face of the LED. A low work-function metal, such as calcium or aluminum, is used as the cathode. The organic part of the device may consist of several layers. These include one or more charge transport layers, plus a light-emitting layer in which the electrons and holes combine to form excitons before emitting light. Typically, the combined thickness of these organic layers is as little as 100 nm. Such thin devices are necessary to keep operating voltages low, to overcome potential barriers for injection, and because charge mobilities in organic semiconductors are very low (10^{-5} to 10^{-2} cm^2 V^{-1} s^{-1}).

The inclusion of electrical contacts into organic semiconductor laser structures can be problematic since they need to be situated very close to the emitting region. Metallic contacts can quench the luminescence and introduce losses to the waveguide mode, thereby increasing the oscillation thresholds. Furthermore, the ITO anode has a high refractive index and so can reduce the spatial overlap of the

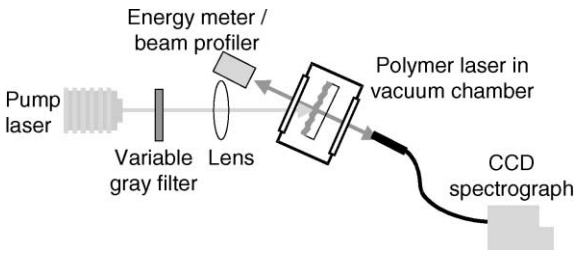


Figure 9 Schematic of a typical experimental configuration for characterizing polymer lasers.

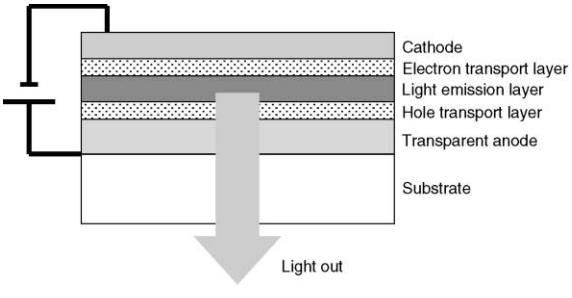


Figure 10 Schematic structure of a generic organic light-emitting diode.

guided mode with the active medium. Such issues indicate that the design of organic LEDs needs to be somewhat modified for achieving electrically pumped lasing. In particular, there needs to be a trade-off between the optimum electronic and photonic designs. To address these issues, various device structures have been proposed that distance the optical mode from metallic contacts, while reducing the thickness of the ITO layer to optimize the mode confinement in the active material.

A more demanding obstacle to electrically pumped lasing is the additional excited state absorption associated with the electrical pumping scheme. As many as three-quarters of the generated excitons in an organic LED are formed in the triplet state which is nonemissive and can lead to excited-state absorption, albeit principally at lower energies than the lasing wavelength. Additionally, injected charges that have not yet formed excitons, exhibit strong absorption, which overlaps significantly with the peak of the material gain. Indeed, this charge-induced absorption has so far prevented the demonstration of injection lasing in conjugated polymers, even in optimized device structures mentioned above. A possible alternative route for efficient pumping of polymer lasers could be through indirect electrical pumping. For example, an inorganic diode laser or LED, or even an organic LED, could be used as an electrically efficient optical pump for the semiconducting polymer laser. Recent advances in GaN-based LEDs and lasers provide interesting prospects for such alternative systems.

Research on semiconducting polymer lasers continues, both to reduce the excitation densities needed for threshold and to investigate new material systems that could overcome the obstacle of charge-induced absorption. By building on recent developments in resonator geometries, material blends, and compact pumping schemes, semiconducting polymer laser systems, whether optically or electrically pumped, should soon become practical, visible light sources for a range of applications.

List of Units and Nomenclature

Charge mobility	$[\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}]$
Current density	$[\mathrm{A\,cm}^{-2}]$
Energy density	$[\mu\mathrm{J\,cm}^{-2}]$
Excited state densities	$[\mathrm{cm}^{-3}]$
Gain coefficient α ($\exp(\alpha z)$)	$[\mathrm{cm}^{-1}]$
Intensity/power density	$[\mathrm{kW\,cm}^{-2}]$
Optical gain	$[\mathrm{dB\,cm}^{-1}]$
Stimulated emission cross-section	$[\mathrm{cm}^2]$
Waveguide propagation loss coefficient γ ($\exp(-\gamma z)$)	$[\mathrm{cm}^{-1}]$
Wavelength	[nm]

See also

Diffractive Systems: Diffractive Laser Resonators. **Lasers:** Dye Lasers; Planar Waveguide Lasers. **Optical Amplifiers:** Basic Concepts. **Optical Materials:** Plastics. **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguide.

Further Reading

Burroughes JH, Bradley DDC, Brown AR, *et al.* (1990) Light-emitting-diodes based on conjugated polymers. *Nature* 347: 539–541.

Friend RH, Gymer RW, Holmes AB, *et al.* (1999) Electroluminescence in conjugated polymers. *Nature* 397: 121–128.

Karl N (1972) Laser emission from an organic molecular crystal. *Physica Status Solidi (a)* 13: 651–655.

Kranzelbinder G and Leising G (2000) Organic solid-state lasers. *Reports on Progress in Physics* 63: 729–762.

McGehee MD and Heeger AJ (2000) Semiconducting (conjugated) polymers as materials for solid-state lasers. *Advanced Materials* 12: 1655–1668.

Miyata S and Nalwa HS (eds) (1997) *Organic Electroluminescent Materials and Devices*. Amsterdam: OPA.

Moses D (1992) High quantum efficiency luminescence from a conducting polymer in solution – a novel polymer laser-dye. *Applied Physics Letters* 60: 3215–3216.

Pope M and Swenberg CE (1999) *Electronic Processes in Organic Crystals and Polymers*. New York: Oxford University Press.

Samuel IDW (2000) Polymer electronics. *Philosophical Transactions of the Royal Society of London A* 358: 193–210.

Scherf U, Riechel S, Lemmer U and Mahrt RF (2001) Conjugated polymers: lasing and stimulated emission. *Current Opinion Solid State and Materials Science* 5: 143–154.

Tessler N (1999) Lasers based on semiconducting organic materials. *Advanced Materials* 11: 363–370.

Tessler N, Denton GJ and Friend RH (1996) Lasing from conjugated-polymer microcavities. *Nature* 382: 695–697.

Planar Waveguide Lasers

S Bhandarkar, Alfred University, Alfred, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Lasers have evolved to become true marvels of technology. From their invention in the early 1960s, laser light has changed the way society operates, its application stretching to virtually every aspect of human activity. The versatility of lasers is derived from the fact that there are several different types, each with a unique set of advantages. Solid-state lasers, made from specific semiconducting materials, are commonly used in the optical networks that underlie the greatest machine on Earth – the Internet. High power gas lasers, such as CO_2 lasers, are used in industrial machining, whereas rare-earth doped and dye lasers have found applications ranging from eye surgery to tattoo removal.

Planar waveguide lasers are a relatively new class of lasers made possible by advances in the field of optical waveguides. Over the past decade, with the aim to miniaturize optical components, planar waveguide technology has been developed to a high degree of sophistication. Likewise, Er and other rare-earth doping was studied extensively owing to its importance in all-optical amplification. These elements turned out to be the building blocks for the development of fiber lasers and subsequently planar waveguide lasers.

This chapter will discuss the basic concepts underlying this category of lasers, the variations therein, and the main characteristics. It is important to understand the basic features of waveguides and lasers, which is included as a prelude. We will then discuss planar waveguide lasers with emphasis on the essential elements: the laser materials, the various designs, their performance, and the resultant advantages as well as the applications.

Waveguides

A waveguide is a general conduit for efficient propagation of an electromagnetic field. It is usually defined by making a region of higher refractive index within a dielectric medium. This behaves as an attractive potential well that confines light in the form of bound modes, much like a potential well that confines electrons to a bound state. The propagating field is referred to as the lightwave and is governed by Maxwell's equation. For a weakly guiding approximation, as is the case when the refractive index contrast (Δn) is small, Maxwell's equation can be reduced to a scalar wave equation for the transverse electric field. For a given wavelength (λ) and polarization, the scalar wave equation in a nonmagnetic medium is

$$\nabla^2 E(x) + \frac{4\pi^2}{\lambda^2} n(x)^2 E(x) = 0 \qquad [1]$$

where E is the electric field. For a waveguide aligned in the z direction, the field is the sum of the two propagating transverse modes, each with the form $E(x) = \Phi(x, y)e^{ik_z z}$ where $\Phi(x, y)$ is the field distribution and k_z is the propagation constant. The wave equation for the transverse mode reduces to

$$\frac{\lambda^2}{4\pi^2}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\Phi(x, y) + n(x, y)^2\Phi(x, y) = (\bar{n})^2\Phi(x, y) \qquad [2]$$

Here, $\bar{n}$ is referred to as the effective refractive index and is given by $\bar{n} = k_z\lambda/2\pi$. The effective index dictates the occurrence of bound modes – these states occur when $\bar{n}$ is between n_{core} and n_{cladding}. For smaller values of $\bar{n}$, there is a continuum of unbound and radiation modes. These are leaky modes that are dispersed quickly over a short distance. Only bound modes are guided in the core of the waveguide structure.

Planar Waveguides

Waveguides formed on a flat substrate are called planar waveguides. These are typically made by stepwise deposition of films of dielectric materials (typically glass). The waveguide core is defined by one of several methods, the most common of which uses lithography and etching. In this case, a film of an appropriately higher index material is deposited on an 'under-cladding' film and then selectively etched to define a core. This is generally covered with a suitable 'upper-cladding' layer so that the index contrast is controlled in all directions. Other methods of forming the core include patterned ion exchange or ion implantation or increasing the index of the medium using a femtosecond laser, etc. In all these cases, the end result is the formation of an embedded higher-index region whose locus defines the optical path for the lightwave. A circuit made from these waveguides is known as planar lightwave circuits (PLC).

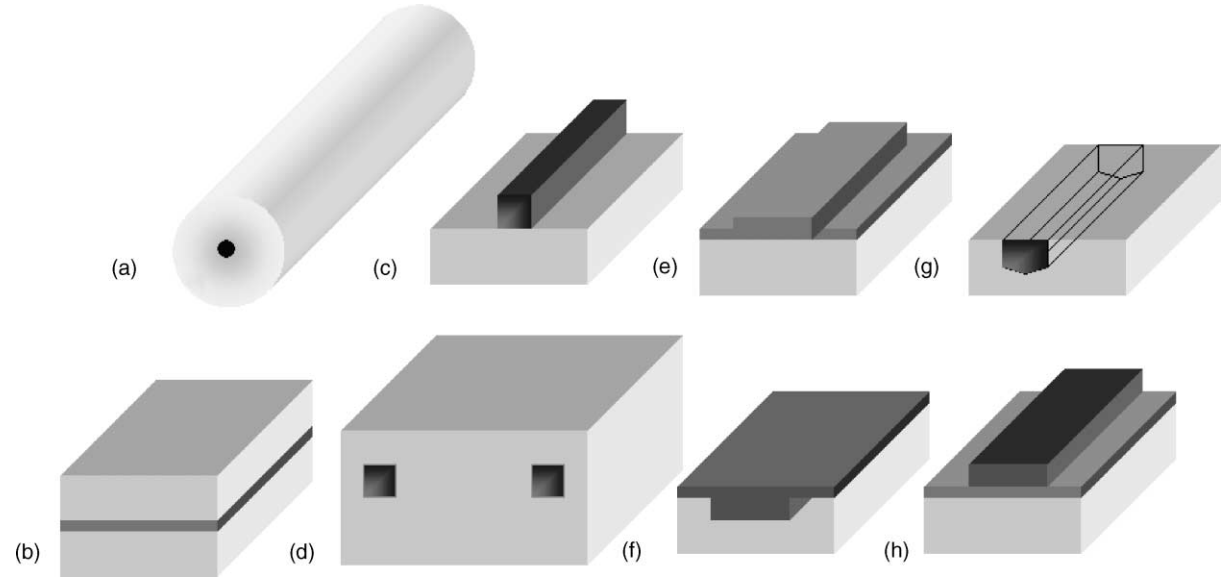


Figure 1 Range of different waveguide designs, with the core represented by the darker regions: (a) Optical fiber; (b) slab waveguide; (c) channel waveguide; (d) embedded channel waveguides with a layer of the overcladding; (e) rib waveguide; (f) inverted rib; (g) ion diffused; and (h)strip loaded, comprising of a strip of a different higher index material (that serves to confine light) on a core layer.

Waveguide cores may be designed in several different configurations, as shown in Figure 1. These designs allow for different fabrication processes as well as the control over the modal properties of the lightwave. For a given index contrast, the number of modes traveling in the waveguide is primarily determined by the dimensions of the waveguide. The aspect ratio of the waveguide affects the propagation of the transverse electric (TE) and the transverse magnetic (TM) modes, due to different propagation constant k_z for each mode.

Lasers

The generation of laser light requires four basic conditions to be satisfied:

- a source capable of light emission;
- an energy supply source – typically electrical, optical, or chemical;
- population inversion to the excited state in the lasing medium; and
- absence of parasitic effects that would absorb the emitted light.

Laser light is often a result of either electronic or vibrational excitation. This is because these excitations involve transfer between energy states that coincide with the radiation of UV, visible, or infrared light. Population inversion is an important requirement in lasers since the emission is stimulated. This involves having a greater fraction of the atoms or molecules in the excited state so that a stimulating photon can cause a radiative decay to the lower energy level. The stimulated light from most lasing media is weak, which requires unacceptably long lengths to generate sufficient power. Instead, as shown in Figure 2, lasers are made with two parallel mirrors with an axial light-generating source in between. This configuration is known as a resonant cavity, a resonator, or an oscillator. The back mirror is usually fully reflective but a partial front mirror allows a portion of the light in the cavity to pass through. The two mirrors cause the light to bounce to and fro, forcing it into making multiple passes through the cavity. Light emanating from this structure is thus amplified and is also coherent (that is the lightwaves are in phase

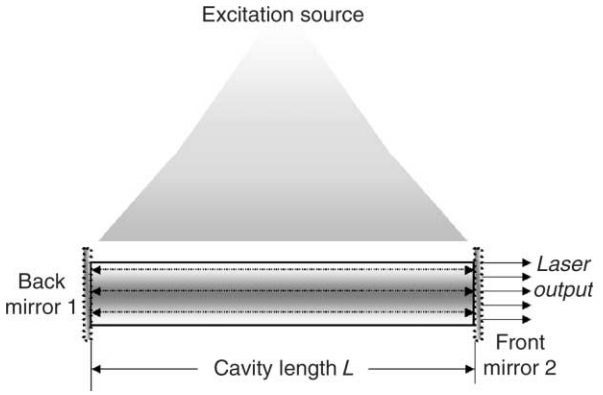


Figure 2 Conceptual representation of a simple Fabry–Perot laser.

with each other spatially and temporally). This has come to represent the signature output of a laser.

Planar Waveguide Lasers (PWL)

By way of an example, let us consider a solid state, rare-earth doped laser, such as an Nd-doped YAG laser. All rare-earth ions are able to fluoresce, some (e.g., Nd) more efficiently than others, and hence can be light emitters. These ions have to be dissolved (doped) into a solid host, say, a glass or a single crystal such as YAG (**Y**ttrium **A**luminum **G**arnet). The excitation of these ions is done optically with light of a suitably higher energy (or lower wavelength). When this 'pump' light is made incident on a rod of the rare earth doped material, it is absorbed by the rare earth ions and used to reach an excited electronic state. Amongst all the possible higher energy states, only certain ones are sustainable because the electron resides in that state for an acceptably long period of time (called the lifetime). Hence, population inversion can be more readily achieved in these longer lifetime states and stimulated processes such as amplification and lasing are readily possible. The laser light generated in the doped rod is emitted across its entire cross-section.

Planar waveguide lasers are designed to miniaturize the kinds of stand-alone solid state lasers describe above. In principle, a PWL is a laser where the light emitting resonant cavity is a planar waveguide. This topic of our interest is a relatively recent development that renders some unique advantages. These can be broadly classified into two categories: one relating to making efficient high-power lasers and the other, to the integration of lasers with other optical components.

An example of the first type is a Nd-YAG laser configured into a planar geometry so that the laser light is now contained within the waveguide. One design that has evolved to achieve this well is shown in Figure 3. The light is guided by the waveguiding structure setup by the sapphire planes. Since both the excitation (or pump) and the lasing modes are well confined, high modal overlap is achieved. These PWLs can be edge-pumped or face-pumped and high powers (measured in the order of Watts) can be achieved.

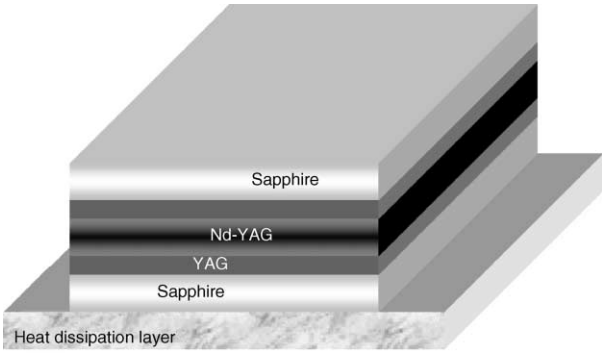


Figure 3 End view of a planar waveguide form of a Nd-YAG laser made using sapphire (single crystal Al_2O_3) for cladding.

In the above design, the cavity is defined by two parallel reflectors bracketing the light-generating medium and is known as a Fabry–Perot etalon. The distance between the two mirrors sets up a wavelength filtering mechanism whereby only resonant wavelengths are sustained. These are related to the intra-mirror distance by the simple equation shown below:

$$L_c = \frac{a\lambda}{2n} \qquad [3]$$

where L_c is the length of the cavity, a is an integer, n is the refractive index of the medium, and λ is the wavelength in vacuum. As the cavity length increases, the number of possible resonant wavelengths increases and the spacing between wavelengths is reduced. These are called the longitudinal modes (Figure 4a). Likewise, lateral modes (Figure 4b) can be generated in the cavity as well, tending to make the operation of the laser multimoded, as shown in Figure 5. This is characteristic of a Fabry–Perot laser, which can be overcome by embellishing the basic etalon design.

As stated above, another big advantage here is that the platform for fabricating PWLs can be chosen to be the same as to make PLCs. Waveguide comprizing the laser can be continued further and configured into other desirable components, such as splitters, couplers, etc. Thus the laser does not have to be separately coupled to the waveguide. This is a major advantage as this connecting process is time-consuming and incurs loss of light. In the field of photonics, the ability to integrate active components with passive ones onto the same chip is considered to be the first step toward photonic chips, analogous to the integrated circuit chip that has revolutionalized our world. This is the driver behind continued research and development of active elements such as PWLs that can be coupled with other actives, such as thin film electro-optic modulators or acousto-optic filters, etc.

Waveguide Laser Materials

The waveguiding structure is comprised of materials that are either dielectrics or semiconductors, both inorganic and organic. For their homogeneity and broad transmittance, high silica glasses are ideal

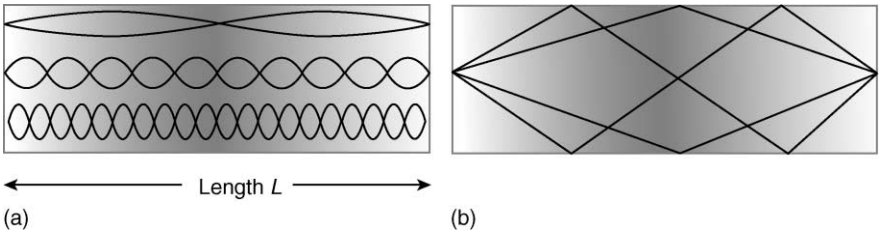


Figure 4 Generation of longitudinal (a) and lateral modes (b) in a Fabry–Perot laser.

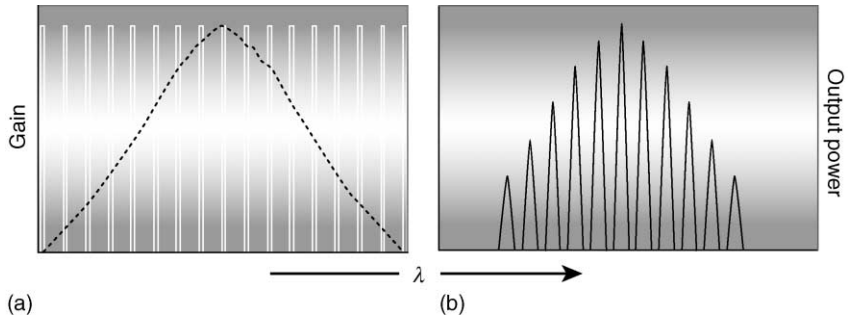


Figure 5 The output of a Fabry–Perot laser: (b) is governed by the number of modes within the gain spectrum of the lasing medium shown by the dotted lines (a). The spacing between the modes can be altered changing the length of the cavity. The smallest mode in output spectrum (b) needs to have enough gain to overcome the lasing threshold.

materials for waveguiding. The chief advantages of high silica waveguides are easy index and mode matching to standard optical fibers, low propagation loss, and good durability. When crystalline materials are used, they have to be single-crystalline to avoid grain boundary scattering (for instance, Ti-diffused waveguide in single-crystal $LiNbO_3$). If semiconductors are used, the bandgap of the material needs to be sufficiently high so as to avoid absorption of the photons. Silicon waveguide (bandgap = 1.1 eV) on a silica cladding (SOI) is commonly used system in PLCs used for telecommunication applications.

The PWLs described above are fabricated in rare-earth doped glass waveguides. Waveguides have to be much shorter than fiber, so the challenge of incorporating the rare-earth ions at high concentration in the host material has to be overcome. Planar waveguide lasers are a compressed version fiber lasers that have the unique advantage that they can be directly integrated with other components on the same chip using very large-scale integration (VLSI)-style processing. The concentration of Er (or the lasing ion in general) in a PWL needs to be many times that in fiber amplifiers (~100 ppm). The need for high gain in a short length requires high pump powers and high lasing ion concentration, both of which lead to nonideal behavior. Since rare earth ions tend to have multiple excited electronic states, there can be absorption at the excited state level as well. When a higher energy state is separated from the upper emission level by the energy in one laser photon, the ions can undergo excited state absorption or ESA (Figure 6). Rare-earth ions in close proximity tend to undergo cooperative upconversion processes, where two excited ions transfer energy between each other so that neither ion may fluoresce at the desired wavelength. These processes can negatively affect the

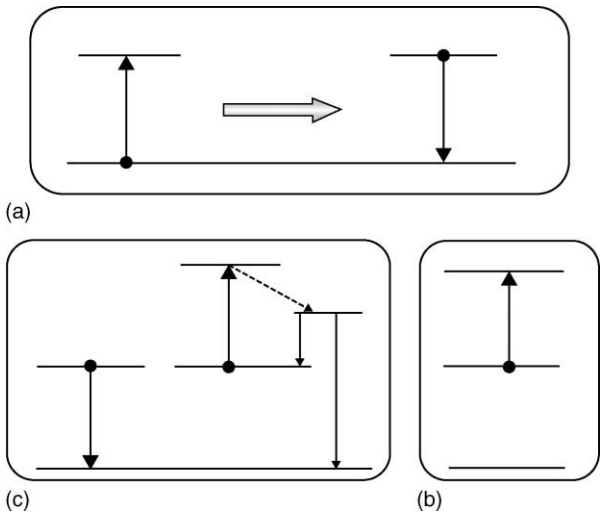


Figure 6 The energy level diagram for a representative two level system. The inset (a) shows the excitation and emission process. Excited state absorption is shown in (b) and cooperative upconversion is shown in (c).

useful gain of the system. Ensuring high spatial dispersion of the Er atoms is critical to minimize this effect. Special glass compositions, such as phosphate glasses, are used here to ensure that the Er is not phase separated or clustered. These glasses also provide a high induced cross-section for gain. Glasses with low phonon energy are also desirable to minimize phonon-mediated nonradiative decays in the rare earths. Light emitting semiconductors and polymers containing fluorescent dyes have also been used to make PWLs.

Now that we have gained an overall perspective of PWLs, we can discuss different manifestations. The following discussion uses a Er doped system as a model since there is a substantial body of work on this system.

Distributed Bragg Reflector (DBR) PWL

In planar waveguide systems, it is difficult to construct efficient mirrors or reflective facets perpendicular to the waveguide. However, mirrors can be replaced by Bragg gratings, which be fabricated relatively easily. In-line Bragg gratings result in wavelength specific reflection as per eqn [4]:

$$\Lambda = \frac{m\lambda}{2\bar{n}} \qquad [4]$$

where Λ is the grating pitch, m is the order of the grating, and λ is the central Bragg wavelength. Gratings have another important benefit in that they ensure that unwanted frequencies outside the limited reflection spectrum are dispersed before they reach the lasing threshold.

A DBR laser is similar in principle to a Fabry–Perot design but uses a Bragg grating on either end of planar Er-doped waveguide. This is shown schematically in Figure 7. Gratings can be made in a number of different ways. One popular way is to exploit the photosensitivity of glasses using 193 nm or 244 nm UV light to periodically alter the refractive index of a glass waveguide. Photosensitivity occurs by a combination of molecular changes (such as the formation of oxygen deficiency sites or color centers) as well as changes in the material strain in the region exposed to the UV light (Figure 7c). Alternatively, gratings can also be formed using etching to make periodic grooves in the waveguide. These are called corrugated or surface relief gratings (Figure 7a and 6b).

The back Grating 1 has a high reflectance of almost 100%, whereas Grating 2 is a weaker grating. The reflectance strength of Grating 2 is determined by taking into account the gain coefficient of the material and the output power required from the laser. The reflection intensity can be altered according to the following equation:

$$R = \tanh^2\left[\frac{\pi L_G \Delta n_{\text{eff}} \eta(V)}{\lambda_B}\right] \qquad [5]$$

where L_G is the grating length, Δn_{eff} is the index modulation of the grating, λ_B is the Bragg wavelength, $\eta(V)$ is the confinement factor, a function of the waveguide parameter V that represents the fraction of the integrated mode intensity contained in the core. Increasing the reflectance strength also widens the linewidth of reflected light. The overlap of the spectral width between the back (Grating 1) and the front grating (Grating 2) becomes the effective output window within the gain spectrum of the lasing material (Figure 8).

Not unexpectedly, a DBR structure generates longitudinal and lateral modes. These modes are, however, limited to the overlapping section of the reflection spectrum of the two gratings, as opposed to the entire gain spectrum as in Fabry–Perot lasers (Figure 9). Yet, this can lead to problems such as spectral hole burning and mode hopping, whereby the dominant mode hops from one wavelength to another. The separation between the modes needs to

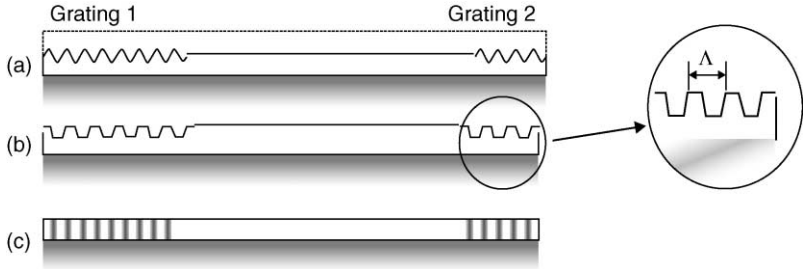


Figure 7 DBR planar waveguide laser. The PWLs shown in (a) and (b) have sinusoidal and trapezoidal surface-relief gratings respectively etched in their cores. An in-line grating of the kind seen in (c) could be fabricated using the photosensitivity of the glass. The top cladding is shown only in (a).

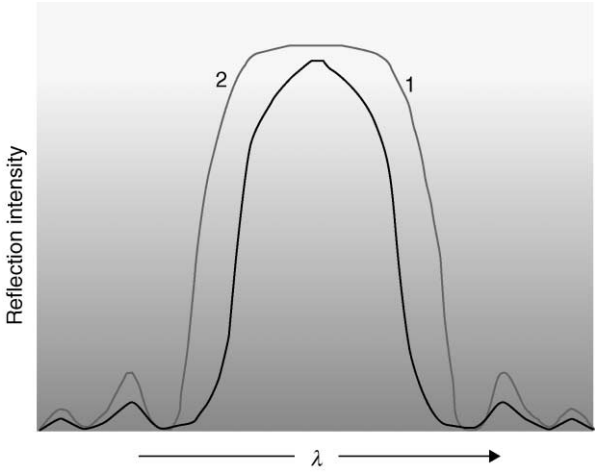


Figure 8 The reflection spectrum of the back (1) and the front (2) gratings. As the intensity of the reflected light is made to increase, so does the linewidth. Hence, the narrower spectrum of partially reflecting grating 2 tends to dictate the operating wavelength range of a DBR.

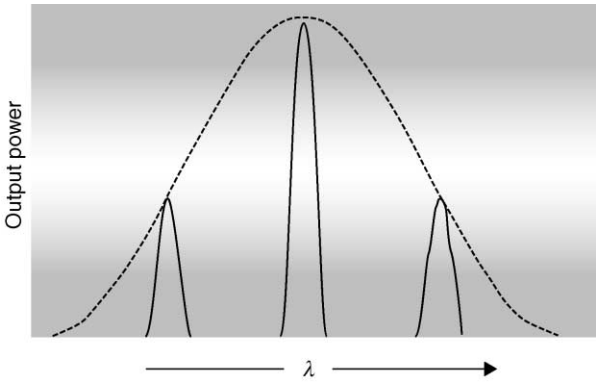


Figure 9 A sample output from a DBR laser, shown in this case with 3 modes. The dotted line shows the effective overlap between the two gratings. The existence of the modes depends on several factors, in particular the pump intensity.

be made sufficiently high compared to the spectral overlap so that single-mode operation can be achieved. But in practice the length of the cavity (L_c in eqn [3]) is determined by factors such as power and space and may not be independently altered to achieve this. However, the reflection spectrum of the back and the front gratings could be slightly shifted in opposite directions to realize a smaller overlap within which only one or a few modes may exist (Figure 10). The occurrence of a single mode also depends on the gain profile of the lasing material and the optical pump power. In Er-doped PWLs, modes other than the primary one can be suppressed by using a pump power below a threshold limit and stable operation can be thus achieved.

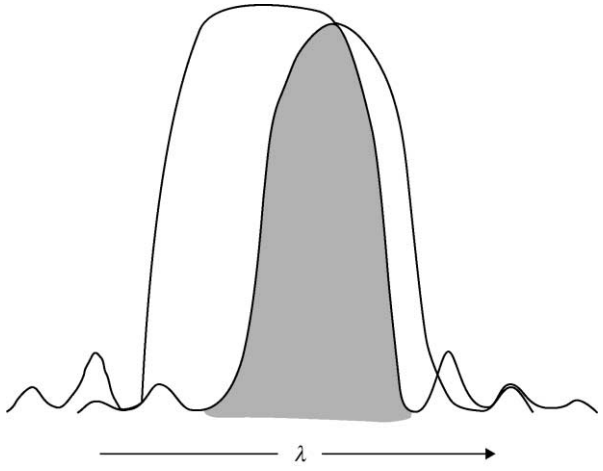


Figure 10 Shifting the front and back gratings is one means of controlling the width of the overlap region, shown in gray. This way a single mode operation can be achieved even for a longer cavity with many modes.

Distributed Feedback (DFB) PWL

The modal problems listed above can be addressed by having a single, distributed grating that is placed along the entire length of the lasing waveguide itself (Figure 11). Alternatively, the gratings may be also placed in the upper or the lower claddings to minimize loss of optical power due to scattering because of the longer grating. In this case, it is the evanescent portion of the light that interacts with the grating. In all cases, the periodicity of the grating leads to a condition for a single lasing wavelength based on constructive interference. The strongest mode occurs when the period Λ of the Bragg grating satisfies the primary Bragg condition (for first order grating or $m = 1$) in eqn [4]. The device will also function if the grating pitch is equal to small integer multiples of $(\lambda/2\bar{n})$.

The grating needs to be made strong enough to generate sufficient feedback (reflections), which also widens the linewidth of the spectrum. To ensure a single narrow linewidth output, a quarter wave phase shift is introduced in the grating, as seen in Figure 11. This phase shift creates a transmission fringe in the spectrum (Figure 12) that significantly narrows the linewidth of the output signal. This design is considered to be standard for the DFB lasers of today. Figure 13 shows a typical output of a DFB laser.

Any of the geometries seen in Figure 1 can be used to make the waveguide for the laser. Due to processing constraints, rib (c) and ion diffused (d) configurations are commonly used. Waveguide losses, characteristically due to scattering from sidewall roughness, have to be low. Otherwise, this adds a

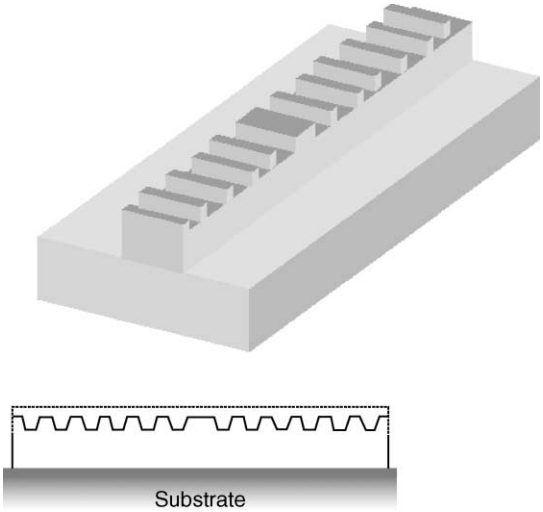


Figure 11 A DFB planar waveguide laser.

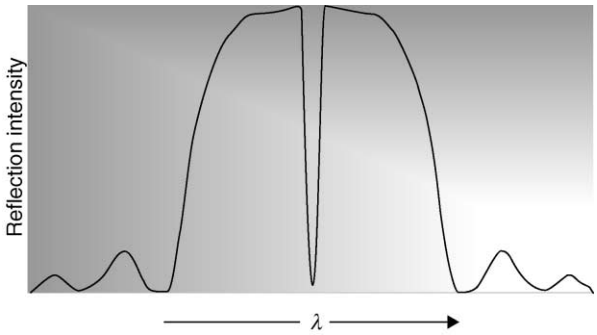


Figure 12 The reflection spectrum of a grating with a central $\lambda/4$ shift in the pitch, showing the transmission fringe.

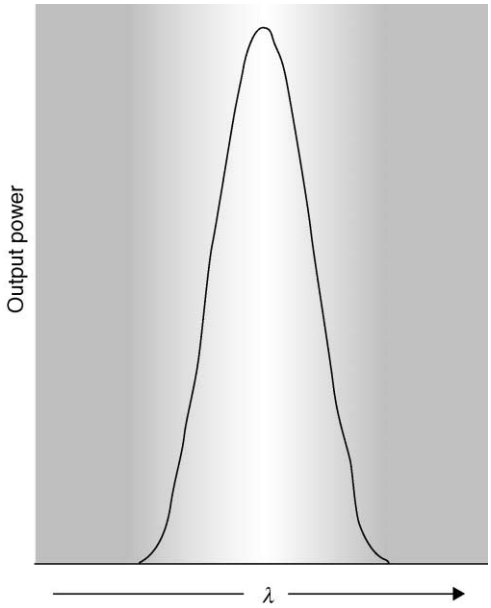


Figure 13 The typical output of a DFB PWL. The linewidth can be made to be as narrow as a few hundred kHz.

significant overhead onto the pump power and lowers the output of the PWL. Birefringence caused by thermal expansion coefficient differences and excessive sidewall roughness also need to be minimized as they can substantially alter the loss of the TE mode compared to TM.

Though their spectral properties are ideal as sources for long haul transmission, DFB lasers tend to be less powerful than the corresponding DBR designs. But the DBR PWLs are very sensitive to reflections that can cause a broadening of the laser linewidth or can act as a source of noise. In both cases, single mode lasers with very narrow linewidths in the kHz range with low radiance and noise, have been demonstrated. But neither is known to have been commercialized yet.

PWL Arrays

One of the unique advantages of the PWLs is that they are often processed using VLSI techniques. One important implication of this is the ability to make a PWL array where multiple lasers are formed simultaneously on the same chip, each with a unique output (Figure 14). Such arrays are likely to be very useful in future dense wave division multiplexing (DWDM) optical network systems where as many as 128 channels (and more) are planned. Since packaging of individual lasers can be up to 75% of the total cost, making chips with multiple lasers is very attractive. The lasers may be DBRs or DFBs whose output wavelength is controlled by changing the grating parameters. To vary the individual wavelengths either the effective index and/or the pitch of the grating can be changed. Thus, any output spectrum can be designed and executed merely by making the appropriate lithography mask.

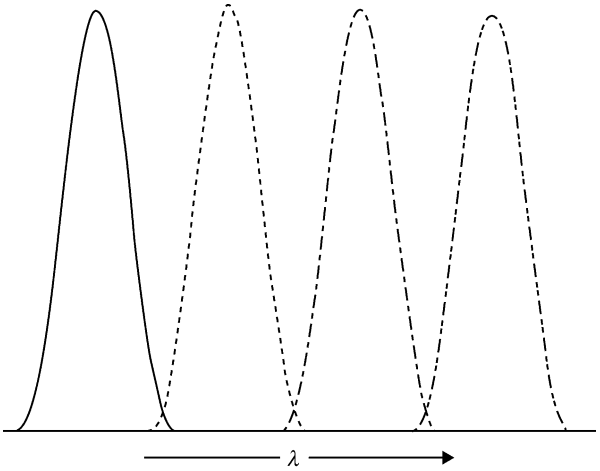


Figure 14 Idealized output from a 4-PWL array.

The gratings are often written holographically, where the wafer is exposed to an interference pattern caused by two coherent light beams. The wafer underneath can be sensitized with a photosensitive glass layer or a film a photoresist. It is not efficient to do multiple holographic exposures to individually change each of the gratings in the array. One way of doing this with a single holographic exposure is by constructing the waveguides at different relative angles. Then the component of the index modulation, resolved along the axis of the waveguide, would vary for each waveguide depending on the relative waveguide angle (Figure 15); this would lead to differences in $\bar{n}$. Alternatively, the waveguide array can be made of the same height but of different widths. This is easily done since waveguide height is usually deposited uniformly but the width is a function of the mask layout. Here the waveguide and the grating are orthogonal to each other and the variation in wavelengths is achieved by the fact that the effective propagation constant for each of the waveguides is different.

Arrays of PWLs have been made on a silicon substrate using deposition of an Er containing glass that is subsequently patterned into waveguide, and ultimately laser, arrays. Another approach has been to start with a phosphate glass as the host and the substrate. A phosphate glass is a superior host for rare-earth ions such as Er since the sensitization efficiency is nearly unity and much high doping levels are possible before concentration quenching effects set in. Waveguides are created by ion-exchange by immersion in a hot bath containing suitable ions (Na for Li, for instance). When starting with a uniformly doped glass, one issue in the assimilation of the laser array with other components is the fact that Er-doped region does not start and end with the waveguide laser section. Recently, the Er has been deposited selectively in the laser waveguide in phosphate glass using methods such as ion implantation, bringing the concept of integration closer to reality.

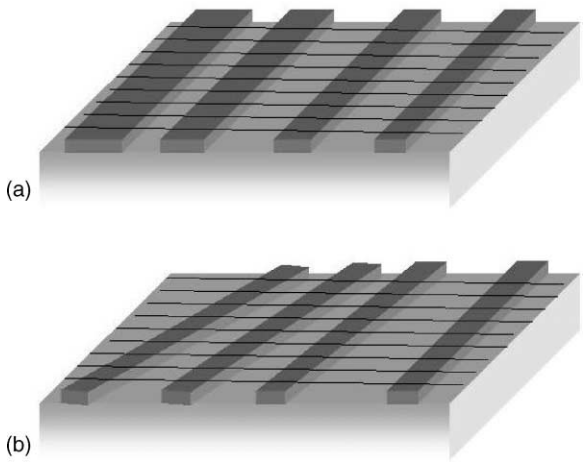


Figure 15 Two different means of making planar waveguide laser array with a single holographic exposure to write the gratings. The darker rectangular blocks represent the core channels or ribs. In (a) the width and hence the effective index of the waveguide is varied whereas the relative angle between the waveguides is altered in (b). The lines represent the direction of the gratings, which is orthogonal only to the left most waveguide in (b).

Pumping

One of the key issues here is the optical pumping of the rare-earth PWLs to generate acceptably high output power. Individual diode lasers or an array of diode lasers all on a single chip (called a diode bar) have been utilized as pumps. Both single and multimode diode lasers are available and have been used. There are pros and cons for either choice: whereas single mode lasers are less powerful (in the range of few 100 mW), it is difficult to convert the multimoded output of their counterparts to useful power. The choice is determined by the waveguide design, which needs to be optimized for high overlap of the lasing mode with the pump. Carefully tapered waveguides can be used distribute the pump power from a diode bar to the various lasers in an array without significant loss.

PWL Stability

Stable operation of a PWL is one of the big challenges. In DBR PWLs, it is seen that there is a threshold for pump power only below which single moded operation can be achieved. Additionally, any PWL die needs to be made immune to ambient temperature changes. These changes can affect many of the critical output parameters of the PWL: wavelength, linewidth, power and in the case of an array, the wavelength separation. This occurs primarily due to the thermal shift in the grating pitch as well as the thermo-optic effect. It is overcome by supplanting the laser die onto a heat spreader chip made from conducting material such as Cu or BeO, followed by a thermoelectric cooler. Ironically, the laser's output wavelength and linewidth are fine-tuned and 'fixed' using temperature to overcome any drifts due to processing errors or other reasons.

Reliability is another important issue that needs to be addressed, especially for telecommunication applications where these types of components have to pass high quality standards such as the Telecordia qualification in the USA.

Other PWL Systems

Many of the PWLs to date have been developed using Er doping, which emits light in the range of interest for telecommunications (the C-band and even the L-band). Here, Yb can be added as a co-dopant to improve power conversion efficiency. Other fluorescing ions have also been used: Nd, Tm, Ho, and Cr. Several research teams have successfully demonstrated Nd-YAG PWLs, in particular. Ridge waveguide lasers have been made using semiconductor materials InGaAs/InP heterojunctions in both Fabry–Perot and DFB configurations. These systems are electrically pumped. DFB PWLs have also been made using dye doped polymers.

It must be mentioned that there are other ways to use waveguides to make lasers. The primary example is a configuration where the waveguide resonator is external to the active region. Designs, such as ring resonator lasers and an external cavity lasers, fall in this category.

Future research and development of PWLs would undoubtedly utilize photonic crystal structures. These structures have a periodic fluctuation in the index of refraction in two or three dimensions. Owing to the resultant Bragg reflections, specific allowed and forbidden states are created which can be used to control the confinement of light. Hitherto, these structures have been used to form mirrors or hexagonal ring resonators. These more sophisticated laser designs are likely to enable new functions such as tunability of the output spectrum.

Summary

Planar waveguide lasers are a special class of laser where light is confined to a waveguide. They have distinctive advantages that include high optical gains, low laser thresholds, narrow linewidths in the kHz range, and optimal thermal management. Significantly, they afford a platform that can be readily expanded to include an array of lasers, each with an unique output, and can ultimately be combined with multiple optical components on the same chip. Such compact integrated devices are bound to take photonics to a new high level of capability.

List of Units and Nomenclature

a	integer denoting number of laser modes
E	electric field
k_z	propagation constant for travel in z direction
L_c	length of laser cavity
L_G	length of grating
m	order of the grating, an integer
n	refractive index of the medium
$\bar{n}$	modal effective refractive index
R	reflectance
V	normalized frequency or the V parameter of the waveguide
Δn_{eff}	modulation of the effective index of the grating
η	confinement factor or fraction of the light confined in the region of interest
λ	wavelength of the lightwave
λ_B	resonant wavelength of a Bragg grating or Bragg wavelength
Λ	grating pitch
Φ	electric field distribution

See also

Lasers: Carbon Dioxide Lasers; Dye Lasers. **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems.

Further Reading

Carroll JE, Whiteaway JEA, Plumb RGS and Plumb D (1998) *Distributed Feedback Semiconductor Lasers*. London: IEE Publishing.

Ghafouri-Shiraz H (2003) *Principles of Semiconductor Laser Diodes and Amplifiers: Analysis and Transmission Line Laser*. London: Imperial College.

Hecht J (2001) *Understanding Lasers: An Entry Level Guide*. New York: Wiley Interscience.

Lee JR, Baker HJ, Friel JG, Hilton GJ and Hall DR (2002) High-average-power Nd:YAG planar waveguide laser, face-pumped by 10 laser diode bars. *Optical Letters* 27: 524–526.

Milonni PW and Eberly JH (1988) *Lasers*. New York: John Wiley and Sons.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: John Wiley & Sons.

Shepherd DP, Hettrick SJ, Li C, *et al.* (2001) High-power planar dielectric waveguide lasers. *Journal of Physics D: Applied Physics* 34: 2420–2432.

Siegman AE (1986) *Lasers*. Harendon, Virginia: University Science Books.

Svelto O (ed.) (1998) *The Principles of Lasers*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Veasey DL, Funk DS and Sanford A (1999) Arrays of distributed-Bragg-reflector waveguide lasers at 1536 nm in Yb/Er codoped phosphate glass. *Applied Physics Letters* 74(6): 789–791.

Young M (2001) *Optics and Lasers: Including Fibers and Optical Waveguides*. NewYork: Springer-Verlag New York, Inc.

Semiconductor Lasers

S W Koch, Philipps-University, Marburg, Germany

M R Hofmann, Ruhr-Universität Bochum, Bochum, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article discusses the principle of operation and the basic material engineering aspects for the design of semiconductor lasers. The physics of the gain medium is analyzed in detail and the role of many-body effects due to the Coulomb interaction of the carriers for the optical properties of the active semiconductor material is demonstrated. The two main types of laser resonators for diode lasers, Fabry–Perot resonators and vertical cavity structures, are described. Finally, the dynamics of semiconductor lasers, including the principles of short pulse generation and nonequilibrium gain dynamics, is addressed.

Basic Principles

The optical emission from semiconductor lasers arises from the radiative recombination of charge carrier pairs, i.e., electrons and holes in the active area of the device. The conduction-band electron fills the valence-band hole by simultaneously transferring the energy difference to the light field. This is called radiative recombination. Hence, the frequency of the laser light is mainly determined by the bandgap of the semiconductor laser material. In order to achieve lasing, one needs carrier inversion, i.e., a sufficient number of electrons in the conduction band so that the probability of light absorption is lower than the probability of emission.

It has been shown that a large variety of semiconductor materials is suited for semiconductor lasers. The most stringent requirement for the material is that it has a direct bandgap. In other words, the maximum of the valence band and the minimum of the conduction band have to be at the same position in k-space (in the center of the Brillouin zone). The elemental semiconductors silicon and germanium do not fulfill this condition and therefore cannot be used for semiconductor lasers.

In order to achieve carrier inversion, it is necessary to pump the semiconductor laser, i.e., to excite a sufficient number of electrons from the valence band into the conduction band. Even though this pumping in semiconductors, as in other solid state laser materials, can be done optically, the large technological impact of semiconductor lasers is at least in part due to the possibility of electrical pumping with a few tens of milli-amperes at low voltages.

The original version of such a semiconductor laser device is shown in Figure 1, The laser structure consists of a GaAs-based p-n junction which is biased in the forward direction. Electrons are injected from the n region and holes from the p region, respectively. Due to the diode characteristics of the p-n junction, semiconductor lasers are often also called laser diodes. In the device shown in Figure 1, the laser mirrors are formed by the cleaved facets of the semiconductor crystal. Because of the high refractive index of GaAs, the semiconductor–air interface provides a reflectivity of about 32% which is sufficient for laser emission.

However, the early type of laser, the so-called homojunction device, was not very efficient because of the low carrier density in the active area and because of the weak overlap between the inverted region and the optical mode. Considerable improvement towards room-temperature continuous wave (cw) operation was achieved by the introduction of the so-called double heterostructure. Its principle is shown in Figure 2.

The basic idea behind the double heterostructure laser is that the active material (e.g., GaAs) is embedded in another semiconductor with a slightly larger bandgap (e.g., (AlGa)As). This causes a potential well in which electrons and holes are confined in order to achieve high carrier densities in the active area. Moreover, the smaller refractive index of the surrounding high-bandgap material helps to guide the optical field within the region of highest inversion. Additional ridge-shaped lateral structuring is frequently used to form a complete optical waveguide.

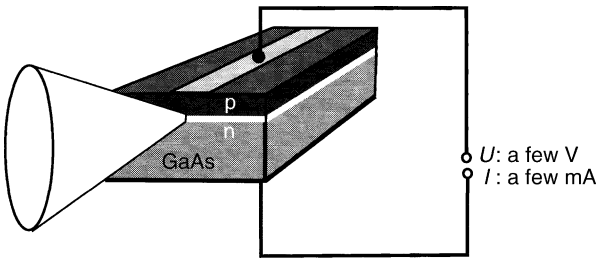


Figure 1 Principle of operation of a semiconductor laser.

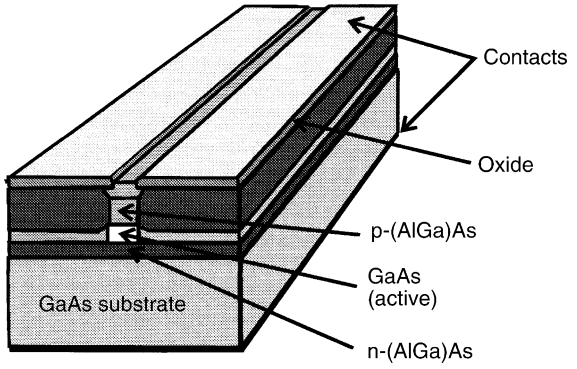


Figure 2 Double heterostructure laser diode.

Nowadays, even more advanced laser structures can be fabricated, as modern crystal growth techniques such as molecular beam epitaxy (MBE) and metal organic chemical vapor deposition (MOCVD) have become available. These techniques allow the grower not only to determine the composition of the semiconductors with remarkable precision, but also to define the shape virtually on an atomic scale. In particular, it is now possible to grow microstructures so small that their electronic and optical properties deviate substantially from those of the corresponding bulk materials. So-called quantum-well (QW) structures, in particular, turned out to be superior to bulk materials in many respects and are currently the active medium in most commercial semiconductor lasers. The gain medium in QW laser structures consists of one or more films with a thickness of a few atomic monolayers. These are embedded in a barrier material with a larger bandgap that serves to confine the carriers to the wells. In QWs the quantum confinement restricts the free carrier motion to the two dimensions within the plane of the films, since the well thickness is comparable to or less than the thermal wavelength of the carriers. Hence, one speaks of quasi-two-dimensional structures where no free motion perpendicular to the films is possible. The corresponding energy states of the carriers are quantized, i.e., only discrete values are allowed for that part of the carrier kinetic energy which is related to the confined direction.

A major consequence of this restriction of the carrier motion to two dimensions is a change in the (single particle) density of states. This density of states is altered in QWs in comparison to bulk material. The density of states for bulk material is zero at the bandgap energy and rises with the square root of the energy for higher energies. In contrast, the density of states for a two-dimensional carrier system in a QW is a step-like function with a density of states different from zero at the QW-well bandgap (the bandgap energy of the active material plus the quantization energies of electrons and holes in the well). This increase of the density of states at the bandgap promises important advantages of QW structures over bulk material. Most importantly, the number of states that have to be occupied before inversion is obtained is considerably reduced in QW structures. This reduces the threshold current and increases the modulation speed of the laser structures. Accordingly, most of the modern laser structures are QW structures.

These arguments have encouraged many researchers to develop structures with further increased carrier confinement in two dimensions (quantum wires) and three dimensions (quantum dots). The single-particle density of states is even more favorable in quantum wires and quantum dots than in QWs: in both cases the single-particle density of states peaks at the effective bandgap energy. This argument seems to promise even better performance of quantum wire and quantum dot lasers as compared to QW devices. However, the single-particle density of states description completely neglects the Coulomb interaction of the carriers which strongly influences the optical and electronic properties of semiconductor structures. The absorption is modified by excitonic effects that lead to additional resonances below the bandgap and to changes of the continuum absorption. These changes are often quantified by the so-called Sommerfeld factor, which depends on the dimensionality of the system, i.e., it differs for bulk, QW, quantum wire and quantum dot structures. For example, the Sommerfeld factor completely suppresses the singularity of the density of states at the bandgap of quantum wire structures. Accordingly, with inclusion of Coulomb effects, the benefit of the quantum confinement is much less than expected from the single-particle density of states.

The use of quantum wire and quantum dot structures as a semiconductor laser gain medium is still in its infancy even though they can be grown with various degrees of precision. However, quantum dot structures, especially, hold great promise for the future, since they can be grown in large numbers using techniques similar to those already used for QWs. The possibility of layer growth under strained conditions, such as QWs between barrier material with a slightly different atomic lattice constant, also results in a greater flexibility to design the laser gain medium with the desired emission wavelength.

In principle, the emission wavelength of a semiconductor laser can be roughly designed by the choice of the material. Different semiconductors have different bandgaps between the valence band and conduction band and the bandgap energy determines the spectral range of the emission. One can tune the bandgap and thus the spectral range of

the emission by fabricating semiconductor alloys as, e.g., (GaIn)As or (AlGa)As. Presently, a large variety of different materials is commercially available covering, for example, the red ((AlGaIn)P), the near infrared ((AlGa)As, (GaIn)As), the telecom ((GaIn)(AsP)) and even the blue–ultraviolet range ((GaIn)N).

Great care has to be taken that the crystal used for a semiconductor laser is of almost perfect quality. Dislocations in the crystal, for example, have to be avoided as well as any kinds of defects since they act as nonradiative recombination centers. Nonradiative recombination competes with the radiative recombination that is essential for the laser process. Nonradiative recombination effects include Auger recombination and recombination via defects. Auger recombination is the process where an electron and a hole recombine and completely transfer the energy difference to another carrier (electron or hole) which is excited to a higher state. This is an intrinsic effect that depends mainly on the band structure and the relevant values of the transition probabilities (transition matrix elements). The probability for Auger recombination is temperature- and carrier-density-dependent and can be minimized, e.g., by properly cooling the device and by keeping the carrier densities as low as possible.

Nonradiative recombination via defects is an extrinsic effect that depends on the material quality and can thus be minimized by optimized growth. However, trying to avoid dislocations limits the choice of materials for laser structures severely. The growth processes require 'host' crystals, so-called substrates, on which the real structure can be grown. GaAs, InP, SiC and sapphire wafers are available in sufficient quality as standard substrates for optoelectronic applications. The epitaxially grown material adapts to the crystal structure of the substrate which implies that the lattice constant of the substrate and structure material should be very similar. If the lattice constants do not agree, i.e., if there is a structural mismatch, a certain amount of strain develops which increases with increasing lattice mismatch. Too much strain leads to the formation of dislocations which are detrimental to laser applications. Control of the dislocation density is currently one of the critical issues for the wide-gap III–V materials, such as GaN, because no substrate is available that matches the GaN lattice constant. However, a small amount of strain is often even desirable as long as it does not cause dislocations. Strain influences the band structure of the material in a predictable way and can therefore be used to tailor the laser emission properties.

Thus laser materials for a certain desired wavelength range have to fulfill two important requirements: first, the bandgap of the material has to fit the desired photon energy; and second, and more crucial, an appropriate substrate with the right lattice constant has to be available. This second condition excludes many material compositions from laser applications. In many cases, both conditions can only be simultaneously met with quaternary semiconductor alloys (for example (GaIn)(AsP) on InP substrates).

Physics of the Gain Medium

From a material physics point of view, the important structural aspects of a certain laser material are summarized in the band structure of the gain medium, i.e., in the quantum mechanically allowed energy states of the material electrons. The calculation of the band structure of a particular gain medium is therefore a crucial aspect of semiconductor laser modelling. Additionally, however, one has to understand the relevant properties of the excited electron–hole plasma in the active region of the laser.

In its ground state, i.e., without excitation and at very low temperatures, a perfect semiconductor is an insulator where no electrons are in the conduction-band states. The presence of a population inversion, i.e., occupied conduction band and empty valence band states in a certain range of frequencies, is, however, a requirement for obtaining optical gain and light amplification from a semiconductor. Sufficiently strong pumping leads to the generation of an electron–hole plasma, where the term 'electron' specifically refers to conduction-band electrons and 'holes' to the missing electrons in the originally full valence bands. Holes are also quasiparticles, just as the electrons. Electrons in a solid are quasiparticles with an effective mass that is considerably different from the free-electron mass in vacuum and that is determined by the band structure, i.e., the interaction with all the ions of the solid. The properties of the missing electrons, such as a positive charge for the missing negative electron charge, a spin opposite to that of the missing electron, etc., are attributed to the holes. As a consequence of their electrical charges, the excited electrons and holes interact via the Coulomb potential which is repulsive for equal charges (electron–electron, hole–hole) and attractive for opposite charges (electron–hole). Furthermore, electrons and holes obey Fermi–Dirac statistics and the Pauli exclusion principle that forbids two Fermions from occupying the same quantum state. The consequences of the Pauli principle are often referred to as 'band-filling effects' in the physics of semiconductor lasers.

The theory therefore has to account not only for the band-structure but also for this band filling, i.e., the detailed population of the electron and hole states. Taking into account only the band structure and band filling leads to a so-called free-carrier theory. But the electron–hole plasma in a semiconductor gain medium is an interacting many-body system because of the carrier interactions and the exclusion principle. In a many-body system the properties of any one carrier are influenced by all the other carriers in their respective quantum states. Generally, advanced many-body techniques are needed to systematically analyze such an interacting electron–hole plasma and to compute the resulting optical properties, such as gain, absorption or refractive index, all of which are changing in a characteristic way with the changing electron–hole density.

It is often helpful to investigate the characteristic time-scales and the most direct consequences of the various interaction effects even though the consequences of the different many-body effects are in general not additive. The fastest interaction, under typical laser conditions, is the carrier–carrier Coulomb scattering that acts on scales of femto- to picoseconds to establish Fermi–Dirac distributions of the carriers within their bands. These distributions are often referred to as 'quasi-equilibrium distributions' since they describe electrons and holes in the conduction and valence bands which may be characterized by an 'electronic temperature' or 'plasma temperature.' The plasma temperature is a measure for the mean kinetic energy of the respective carrier ensemble and relaxes towards the lattice temperature as a consequence of electron–phonon (= quantized lattice vibration) coupling, i.e., the carriers can emit or absorb phonons. The coupling to longitudinal optical phonons in polar materials is especially strong with scattering times in the picosecond range. The corresponding times for the electron–acoustic phonon scattering are in the nanosecond range. Generally, the carrier–phonon interaction provides an exchange of kinetic energy of the carriers with the crystal lattice in addition to the relaxation of the initial nonequilibrium distribution. Deviations between plasma and lattice temperature may occur, especially when lasers are subject to strong pumping or rapid dynamic changes.

The Coulomb interaction not only causes rapid carrier scattering but also directly influences the laser gain spectrum. Gain, i.e., negative absorption, occurs in the spectral region of population inversion, in other words, where the net probability for a test photon to be absorbed is lower than the probability to be amplified by stimulated electron–hole recombination. Energetically, the gain region is bounded from below by the effective (or renormalized) semiconductor bandgap and from above by the electron–hole chemical potential. Here, the chemical potential is the energy at which neither absorption nor amplification occurs, i.e., where the semiconductor medium is effectively transparent.

The effective bandgap energy for elevated carrier densities is significantly below the fundamental absorption edge of an unexcited semiconductor. This 'bandgap renormalization' is a consequence of the fact that the Pauli exclusion principle reduces the repulsive Coulomb interaction between carriers of equal charge. Furthermore, the Coulomb screening, i.e., the density-dependent weakening of the effective Coulomb interaction potential, contributes to the reduction of the bandgap.

Another many-body effect originates in the Coulomb attraction between an electron and a hole, and leads to an excitonic or Coulomb enhancement of the interband transition probability. All these effects contribute to the detailed characteristics of the optical spectra of a semicondcutor gain medium. Figure 3 shows an example of calculated spectra in comparison to experimental data.

The example demonstrates that the microscopic theory correctly accounts for the changes in the gain spectrum with changing carrier density by means of a consistent treatment of band structure, band-filling, and many-body Coulomb effects.

The Resonator

In addition to a gain medium, a laser needs a positive feedback process for the amplified stimulated emission. Hence, the inverted medium has to be

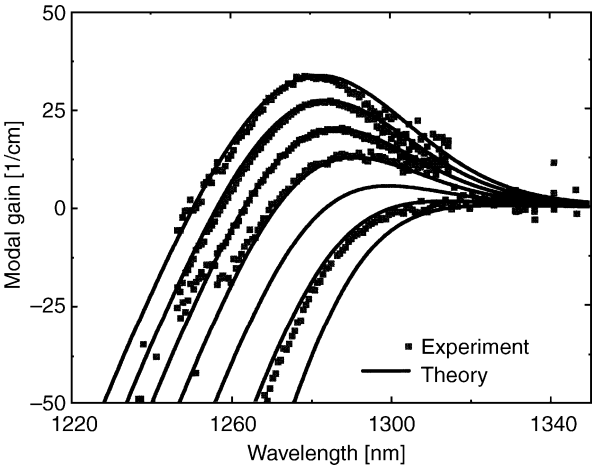


Figure 3 Comparison of experimental and theoretical gain spectra of a (GaIn)(Nas)/GaAs laser.

embedded into a resonator. One of the conceptually simplest versions of such a resonator is a so-called Fabry–Perot resonator which consists of two parallel plane mirrors. A double heterostructure Fabry–Perot laser is shown in Figure 2. Such Fabry–Perot resonators are very easy to fabricate for semiconductor lasers: the semiconductor crystal is cleaved perpendicular to the optical waveguide. The cleaved edges provide a reflectivity of about 32% due to the high refractive index of the semiconductor. This is sufficient to provide enough feedback for laser operation.

However, despite its conceptual simplicity, the cleaved-edge Fabry–Perot concept also introduces a number of problems. First, the light output occurs on two sides of the laser, whereas for most applications all the output power should be emitted out of one facet only. For practical applications, this problem can be solved by depositing coatings onto the cleaved facets of the structure. For example, deposition of a 10% reflectivity coating on one facet and of a 90% reflectivity coating on the other facet concentrates the laser emission almost completely to the facet of lower reflectivity. Antireflection coatings with reflectivities below 10^{-4} are used when laser diodes are coupled to external cavities, e.g., in tunable lasers.

A second problem with Fabry–Perot lasers is that the emission wavelength is often not well defined. A Fabry–Perot resonator of length L (optical length nL) has transmission maxima – so-called modes – at all multiples of $c/2nL$ where c is the speed of light and n the refractive index. That means that for typical devices with a length of 500 μm, hundreds of modes are present even within the limited gain bandwidth of a semiconductor. These modes often compete with each other resulting in multimode emission which may become particularly complex and uncontrollable when the laser is modulated for high-speed operation.

This problem can be overcome in so-called distributed feedback (DFB) lasers. In these devices, the refractive index is periodically modulated along the waveguide. In a simplified picture, this periodic pattern causes multiple reflections at different locations within the waveguide. For certain wavelengths, these multiple reflections interfere constructively to provide enough 'distributed' feedback for laser operation. With the DFB concept, single-mode emission can be realized even for high-speed modulation of the lasers. Alternatively, the region of distributed feedback can be separated from the region of amplification in so-called distributed Bragg reflector (DBR) lasers. These are devices with multiple sections where one section is responsible for the optical gain and another section (the DBR section) is responsible for the optical feedback.

So far, we have discussed so-called edge emitting lasers where the light emission is in the plane of the active area of the device. The lengths of these lasers are typically hundreds of microns and thus correspond to a few hundreds of optical wavelengths. This causes one of the major disadvantages of edge emitters – the multimode emission. Another disadvantage is the poor elliptic beam profile due to diffraction at the small rectangular output aperture. Finally, the fabrication procedure of edge emitters is complex since it requires multiple steps of cleaving before it is even possible to test the laser.

A novel alternative semiconductor laser concept was realized in the early 1990s: the vertical-cavity surface-emitting laser (VCSEL). A typical VCSEL structure is shown in Figure 4. In VCSELs, the light emission is parallel to the growth direction, i.e. perpendicular to the epitaxial layers of the structure. The cavity length of a VCSEL is in the order of a few (1–5) multiples of half the wavelength of the emitted light in the material. These small dimensions imply that the interaction length between the active medium and the light field in the resonator is very short. Therefore, the mirrors of a VCSEL have to be of very good quality; reflectivities of more than 99% are required in most configurations.

Very high reflectivities can be achieved with dielectric multilayer structures, so-called Bragg reflectors. Bragg reflectors consist of a series of pairs of layers of different refractive indices, where each layer has the thickness of a quarter of the emission wavelength. Most preferably, Bragg reflectors are epitaxially grown together with the active area of the device in only one growth process.

Small VCSEL resonators with Bragg reflectors at both ends of the cavity are called microcavities. They are designed in such a way that only one longitudinal mode is present in the region of the semiconductor gain spectrum. Accordingly, VCSELs are well suited to provide single-mode characteristics. Moreover, a round aperture for light emission can be realized with a diameter of a few micrometers so that

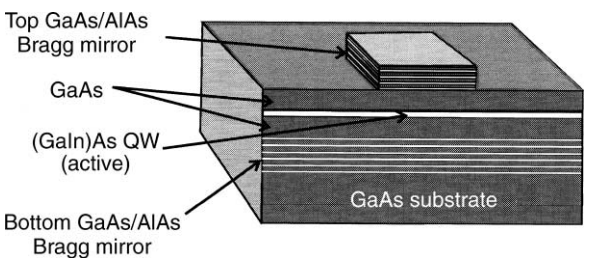


Figure 4 Vertical-cavity surface-emitting laser (VCSEL).

diffraction is weak and the beam profile is almost perfect.

VCSELs can be tested on wafer without further processing and can be arranged in two-dimensional arrays. A major disadvantage of VCSELs is their temperature dependence. The bandgap of the active material and the cavity mode depend on temperature in such a way that the cavity mode shifts away from the region of maximum gain when the temperature is varied. This severely limits the temperature range of operation for VCSELs.

Semiconductor Laser Dynamics

The dynamical behavior of semiconductor lasers is rather complex because of the coupled electron–hole and light dynamics in the active material. Like other laser materials, semiconductor lasers exhibit a typical threshold behavior when the pump intensity is increased. For weak pumping, the emission is low and spectrally broad; the device operates below threshold and shows (amplified) spontaneous emission. At threshold, the optical gain compensates the losses in the resonator (i.e., internal losses and mirror losses) and lasing action starts. Ideally the output power increases linearly with the injection current above threshold. The derivative of this curve is proportional to the differential quantum efficiency. The differential quantum efficiency can reach values close to unity and even the overall efficiency of the diode laser, i.e., the conversion rate from electrical power into optical power, can be as high as 60% which is far higher than for any other laser material.

Ideally, the laser emission is single mode above threshold. However, as discussed above, edge-emitting diode lasers tend to emit on multiple longitudinal modes and the competition of these modes can lead to complex dynamical behavior which is significantly enhanced as soon as the laser receives small amounts of optical feedback. Semiconductor lasers are extremely sensitive to feedback and tend to show dynamically unstable, often even chaotic, behavior under certain feedback conditions.

Like other lasers, semiconductor lasers respond with relaxation oscillations when they are suddenly switched on or disturbed in stationary operation. Relaxation oscillations are the damped periodic response of the total carrier density and the optical output intensity as a dynamically coupled system. Both are important for many aspects of semiconductor laser dynamics. The relaxation oscillation frequency determines the maximum modulation frequency of diode lasers. This frequency is strongly governed by the differential gain in the active material, i.e., by the variation of the optical gain with carrier density.

An important application of relaxation oscillations is short pulse generation by gain switching of laser diodes. The idea of gain switching is to inject a short and intense current pulse or an optical pump pulse into the laser structure. This pulse initiates relaxation oscillations but it is so short that already the second oscillation maximum is not amplified. As a consequence, only one intense pulse is emitted. The shortest pulse widths achieved with this scheme so far are in the few picoseconds range. For short pulse generation, a further complexity of the diode laser comes into play. The emitted pulses are observed to be strongly chirped, i.e., the center frequency changes during the pulse. This complex self-phase-modulation is a result of the strong coupling of the real and the imaginary parts of the susceptibility in the semiconductor. If the gain (which is basically given by the imaginary part of the susceptibility) changes during a pulse emission, the refractive index (basically the real part of the susceptibilty) also changes considerably. This leads to a self-phase-modulation if a strong optical pulse is emitted. This strong phase–amplitude coupling is also responsible for the high feedback sensitivity of diode lasers, for complex spatio-temporal dynamics (e.g., self-focusing, filamentation), and for a considerable broadening of the emission linewidth. The phase–amplitude coupling is often quantified with the so-called linewidth enhancement or α factor. This parameter, however, is a function of carrier density, photon wavelength, and temperature.

Diode lasers are also suited for the generation or amplification of extremely short pulses with durations below 1 picosecond. The concept used for such extremely short pulse generation is called mode-locking. The idea of mode-locking is to synchronize the longitudinal modes of the laser so that their superposition is a sequence of short pulses. The most successful concept to achieve mode-locking of diode lasers is to introduce a saturable absorber into the laser cavity and to additionally modulate the injection current synchronously to the cavity roundtrip frequency. A saturable absorber exhibits nonlinear absorption characteristics: the absorption is high for weak intensities and weak for high intensities so that high peak power pulses are supported. Such absorbers can be integrated as a separate section into multiple-section diode lasers which have already been successfully used for subpicosecond pulse generation. Furthermore, electro-absorption modulators, passive waveguide sections, and tunable DBR segments for wavelength tuning can also be integrated into such multiple-section devices.

However, when subpicosecond pulses are generated or amplified in semiconductor laser amplifiers the dynamics of the pulse generation or pulse amplification is strongly governed by complex nonequilibrium carrier dynamics in the semiconductor. The most prominent effects are spectral hole burning and carrier heating. Spectral hole burning occurs if an intense laser field removes carriers from a certain area in k-space faster than carrier–carrier scattering can compensate for. This leads to nonthermal carrier distributions and a gain suppression in certain energy regimes. As mentioned above, a nonthermal carrier distribution relaxes quickly (100 femtoseconds) into a thermal (quasi-equilibrium) distribution if the intense laser field is switched off, but this carrier distribution may have a plasma temperature much higher than the lattice temperature. This effect, which is referred to as carrier heating, also leads to a transient reduction of the gain on a few picoseconds timescale until the carrier distributions have cooled down to the lattice temperature by interactions with phonons. These ultrafast effects are of course particularly important for amplification and generation of short pulses but they generally influence the whole dynamical behavior of diode lasers including, e.g., the modulation behavior.

See also

Coherent Transients: Ultrafast Studies of Semiconductors. **Quantum Optics:** Entanglement and Quantum Information.

Further Reading

Chow WW and Koch SW (1999) *Semiconductor-Laser Fundamentals: Physics of the Gain Materials*. Berlin: Springer-Verlag.

Chuang SL (1995) *Physics of Optoelectronic Devices*. New York: Wiley.

Delfyett PJ, Dienes A, Heritage JP, Hong MY and Chang YH (1994) Femtosecond hybrid mode-locked semiconductor laser and amplifier dynamics. *Applied Physics* B 58: 183–195.

Haug H and Koch SW (2004) *Quantum Theory of the Optical and Electronic Properties of Semiconductors*, 4th edn. Singapore: World Scientific Publisher.

Hofmann M (1999) Gain and emission dynamics of semiconductor lasers. *Recent Research Developments, in Applied Physics* 2: 269–290.

Koch SW, Jahnke F and Chow WW (1995) Physics of semiconductor microcavity lasers, review article. *Semiconductor Science and Technology* 10: 739–751.

Osinski M and Buus J (1987) Linewidth broadening factor in semiconductor lasers – an overview. *IEEE Journal of Quantum Electronics* QE-23: 9–29.

Up-Conversion Lasers

A Brenier, University of Lyon, Villeurbanne, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Under photoluminescence excitation, a luminescent center located inside a transparent material usually emits at a longer wavelength than the excitation. This is because of the existence of de-excitation losses inside the material. Thus, efficient down-conversion lasers can be built, for example, the commercial Nd^{3+}:$Y_3Al_5O_{12}$ laser which works at 1064 nm, under usual pumping near 800 nm. More surprisingly, it has been demonstrated that fluorescence can also be emitted at shorter wavelength than the excitation. This is possible if the losses are reduced and overcome by so-called 'up conversion mechanisms'. Generally speaking, a laser emitting at a shorter wavelength than the excitation is classified as an up-conversion laser.

These lasers are able to generate coherent light in the red, green, blue, and uv spectral ranges. A point of practical importance is that some of them can be pumped in the near-infrared range (around 800 nm or 980 nm) by commercially available laser diodes. They can be based on all-solid-state technology and share its advantages: compactness of devices, low maintenance, and efficiency. They are needed for a variety of applications: color displays, high density optical data storage, laser printing, biotechnology, submarine communications, gas-laser replacement, etc.

Because the emitted photons have higher energy than the excitation ones, more than one pump photon is needed to generate a single laser photon, so the physical processes involved in the up-conversion laser are essentially nonlinear. Two kinds of laser devices can be considered. In the first one, the excitation of the luminescent center up-migrates through the electronic levels and reaches the initial laser level. This step is due to some luminescence mechanisms. Then, in a second step, lasing at a short wavelength occurs through a transition towards the final laser level. In the second kind of device, the pump excitation activates a down-conversion lasing, but

the laser wave is maintained inside the cavity, cannot escape, and is up-converted in short wavelengths by the second-order nonlinear susceptibility of the crystal host. This requires bifunctional nonlinear optical and laser crystals. The device is thus a self-frequency doubling laser or a self-sum frequency mixing laser. If the nonlinearity is not an intrinsic property of the laser crystal but is due to a second nonlinear optical crystal located inside or outside the cavity, the device operates intracavity or extracavity up conversion.

Up-Conversion from Luminescence Mechanisms

The numerous $^{2S+1}L_J$ energy levels ($4f^n$ configuration) of the trivalent rare earth ions, inserted in crystals or glasses, offer many possibilities for up conversion, particularly from long-lived intermediate levels that can be populated with infrared pumping. Crystal field induced transitions between them are comprised of sharp lines, because the $4f^n$ electrons are protected from external electric fields by the $5s^2p^6$ electrons. Their fluorescence spectra can exhibit cross-sections high enough (10^{-19} cm^2) to lead to laser operation in the visible range. For up-conversion purposes, the most popular ions are Er^{3+}, Pr^{3+}, Tm^{3+}, Ho^{3+}, and Nd^{3+}. Up conversion in Er^{3+} was used for the first time in 1959, to detect infrared radiation.

Energy Transfers

Two ions in close proximity interact electrostatically (Coulomb interaction) and can exchange energy. The ion which gives energy is called the sensitizer S or donor, and the energy receiving ion is the activator A or acceptor. Two examples discussed below are shown in Figure 1. The most typical case of such nonradiative energy transfer in trivalent rare earth ions in insulating materials is due to electrical dipole–dipole interaction. For the latter, the transition probability $W(s^{-1})$ takes the form:

$$W = \frac{3\hbar^4 c^4}{4\pi n^4}\frac{1}{R^6}\frac{Q_A}{\tau_{rad}}\int \frac{f_A(E)f_S(E)}{E^4}dE \qquad [1]$$

where R is the S–A distance, E is the photon energy of the transition operated by each ion, n the refractive index, and Q_A is the integrated absorption cross-section of the A-transition:

$$Q_A = \int \sigma_A(E)dE, \qquad f_A = \frac{\sigma_A}{Q_A} \qquad [2]$$

In eqn [1], τ_{rad} is the radiative emission probability of S for the involved transition calculated from the emission probability A_S:

$$\frac{1}{\tau_{rad}} = \int A_S(E)dE, \qquad f_S = \tau_{rad}A_S \qquad [3]$$

The integral in eqn [1] is called the overlap integral and shows that the transfer is efficient if the luminescence spectrum of S is resonant with the absorption spectrum of A (corresponding to the S and A transitions involved in the nonradiative transfer). The transitions are often not in perfect resonance but remain efficient enough to be used in practice. Then the energy mismatch is compensated by phonons, and the energy transfer is said to be phonon-assisted.

The energy transfer visualized in Figure 1a is an up-conversion cross-relaxation. It is often met in the rare earth ions cited above when they are sufficiently concentrated in the material. The S and A ions can be of the same or of different chemical species. Successive energy transfers can promote the activator from its ground state towards higher energy levels, up to the one from which lasing can occur. This is the so-called 'Addition de Photons par Transferts d'Energie', sometimes appointed 'APTE'. Yb^{3+} is the most used ion as sensitizer and leads to the following transfers in examples of Er^{3+} or Ho^{3+} co-doping:

$$\begin{aligned} {}^2F_{5/2}(Yb) + {}^4I_{15/2}(Er) &\rightarrow {}^2F_{7/2}(Yb) + {}^4I_{11/2}(Er) \\ {}^2F_{5/2}(Yb) + {}^4I_{11/2}(Er) &\rightarrow {}^2F_{7/2}(Yb) + {}^4F_{7/2}(Er) \end{aligned} \qquad [4]$$

$$\begin{aligned} {}^2F_{5/2}(Yb) + {}^5I_8(Ho) &\rightarrow {}^2F_{7/2}(Yb) + {}^5I_6(Ho) \\ {}^2F_{5/2}(Yb) + {}^5I_6(Ho) &\rightarrow {}^2F_{7/2}(Yb) + {}^5S_2(Ho) \end{aligned} \qquad [5]$$

If several sensitizers give their energy simultaneously to a single activator, promoting it directly from its ground state towards a high energy level without intermediate levels, the sensitization is said to be 'cooperative'.

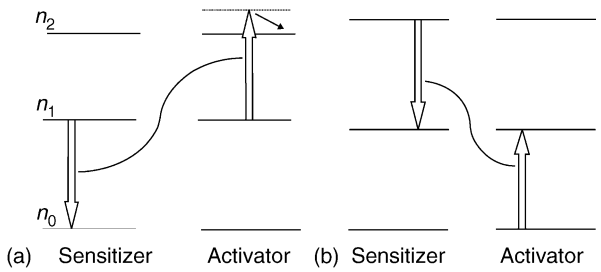


Figure 1 (a) Up-conversion cross-relaxation energy transfer. (b) Down-conversion cross-relaxation energy transfer. The example in (a) is phonon assisted, the example in (b) is resonant.

The energy transfer shown in Figure 1b is a down-conversion cross-relaxation. It is the inverse of the previous one. At first sight it is, of course, undesirable in an up-conversion laser. However, we notice that one ion in the upper level will lead to two ions in the intermediate level. This fact can be exploited in special conditions (see below) and be beneficial for up conversion.

The dynamical study of the energy transfers, starting from the microscopic point of view contained in eqn [1], is very complicated because of the random locations of the ions or because of other processes, such as energy migration among sensitizers and back transfers. A popular method leading to useful predictions in practice is the rate equation analysis dealing with average parameters. The time evolution of the population densities n_i of the various levels i are described by first-order coupled equations, solved with adequate initial conditions and including the pumping rate. As an example, the rate equations for levels 1 and 2 in Figure 1a take the form:

$$\frac{dn_1}{dt} = -\frac{n_1}{\tau_1} - 2kn_1^2 + \frac{\beta_{21}}{\tau_2}n_2 + W \qquad [6]$$

$$\frac{dn_2}{dt} = -\frac{n_2}{\tau_2} + kn_1^2 \qquad [7]$$

$$n_0 + n_1 + n_2 = 1 \qquad [8]$$

where $\tau_{1,2}$ are the lifetimes of levels 1 and 2, k is the up-conversion transfer rate, β_{21} is the branching ratio for the $2 \rightarrow 1$ transition, and W is the pump rate (not shown in Figure 1).

Sequential Two-Photon Absorption

Pumping the ground state of a luminescent ion (Figure 2a) with a wave at angular frequency ω, leads to population with density n_1 of an excited state in a first step. Because the rare earth ions have numerous energy levels, it is possible for the pump to also be resonant with the transition from level 1 towards a higher energy level 2, and is further absorbed. Such a process leading to n_2 up-conversion population is a sequential two-photon absorption. If there is no $1 \rightarrow 2$ resonant transition for the ω frequency, it is possible to use a second pump wave at a different angular frequency ω' resonant with the $1 \rightarrow 2$ transition (Figure 2b). Of course, in practice, it is advantageous when a single pump beam is required for up conversion.

Let us show the relevant terms of the rate equation analysis describing the process in Figure 2a:

$$\frac{dn_1}{dt} = -\frac{n_1}{\tau_1} + I\sigma_0 n_0 - I\sigma_1 n_1 + \frac{\beta_{21}}{\tau_2}n_2 \qquad [9]$$

$$\frac{dn_2}{dt} = -\frac{n_2}{\tau_2} + I\sigma_1 n_1 \qquad [10]$$

$$n_0 + n_1 + n_2 = 1 \qquad [11]$$

where I is the photon density of the pump (photons $s^{-1}\,cm^{-2}$) and σ_0 and σ_1 are the absorption cross-sections (cm^2) of the $0 \rightarrow 1$ and $1 \rightarrow 2$ transitions, respectively. The n_2 population density shows a nonlinear I dependence, most often close to I^2.

Photon Avalanche

In rare earth doped materials, an up-conversion emission, intense enough for visible lasing, can result from a more complex mechanism. In this case, absorption of the pump is not resonant with a transition from the ground state (weak ground-state absorption), but on the contrary, the photon pump energy matches the gap between two excited states (high excited state absorption). These levels are labeled 1 and 2 in Figure 3. Level 1 generally has a rather long lifetime in order to accumulate

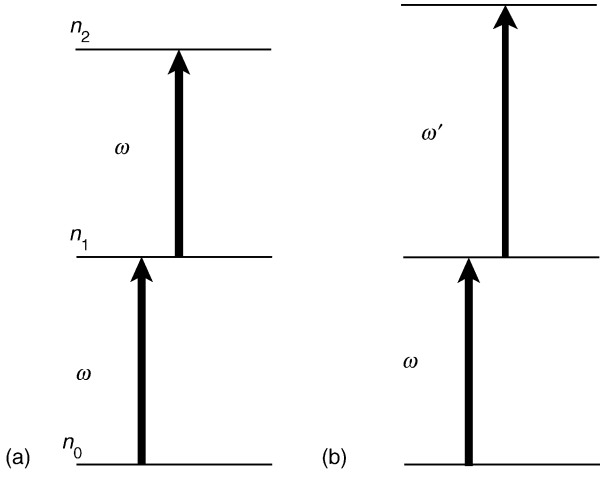


Figure 2 Two-photon absorption up-conversion.

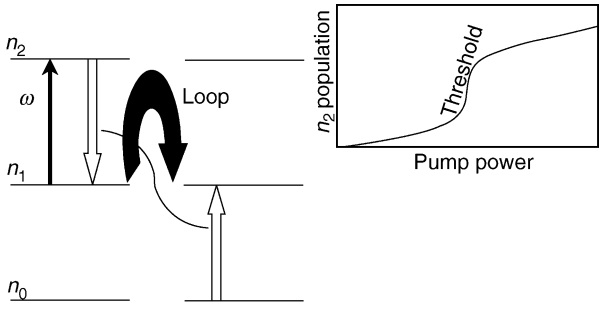


Figure 3 Photon avalanche up-conversion.

population and level 2 is the initial up-conversion laser level. When the pump light is turned on, the populations of levels 1 and 2 first grow slowly, and then can accelerate exponentially. This is generally referred to as 'photon avalanche'. The term 'looping mechanism' also applies, because we can see in Figure 3 that the excited state absorption $1 \rightarrow 2$ is followed by a cross-relaxation energy transfer (quantum efficiency up to two) that returns the system back to level 1. So after one loop, the n_1 population density has been amplified, and is further increased after successive loops. This pumping scheme was discovered in 1979, with $LaCl_3:Pr^{3+}$ crystal used as an infrared photon counter.

The rate equation analysis applied to the three-level system in Figure 3 leads to the following time evolutions:

$$\frac{dn_1}{dt} = -\frac{n_1}{\tau_1} + I\sigma_0 n_0 - I\sigma_1 n_1 + \frac{\beta_{21}}{\tau_2} n_2 + 2kn_0 n_2 \qquad [12]$$

$$\frac{dn_2}{dt} = -\frac{n_2}{\tau_2} + I\sigma_1 n_1 - kn_0 n_2 \qquad [13]$$

$$n_0 + n_1 + n_2 = 1 \qquad [14]$$

where the parameters have the same meaning as in eqns [6]–[11].

The steady state limit of the populations at infinite time, upon continuous excitation starting at $t = 0$, can be obtained from canceling the time derivative in the left-hand side of the equation. Let us point out the interesting result. Two different dynamical regimes and analytical expressions for $n_2(\infty)$ are obtained. In the usual case of low pumping rate in the ground state and if the parameters of the systems satisfy: $k < \frac{1-\beta_{21}}{\tau_2}$, then the first regime is observed, whatever the pumping rate $I\sigma_1$ in the excited state.

If the parameters satisfy:

$$k > \frac{1 - \beta_{21}}{\tau_2} \qquad [15]$$

the first regime is still observed at low pumping rate $I\sigma_1$ in the excited state, more precisely at pumping rate low enough such that

$$I\sigma_1 < \frac{\left(k + \frac{1}{\tau_2}\right)\frac{1}{\tau_1}}{k - \frac{(1 - b_{21})}{\tau_2}}$$

But interestingly, the second regime, the so-called photon avalanche, is obtained at high pumping rate $I\sigma_1$ in the excited state, that is to say if:

$$I\sigma_1 > \frac{\left(k + \frac{1}{\tau_2}\right)\frac{1}{\tau_1}}{k - \frac{(1 - b_{21})}{\tau_2}} \qquad [16]$$

So the right-hand side of eqn [16] appears to be a pump threshold separating two dynamical regimes when the condition in eqn [15] is satisfied, as the insert in Figure 3 shows.

Another way to distinguish between the two regimes is given by linearization of the system of eqns [12]–[14] ($n_0 \cong 1$). Then the description will only be valid close to $t = 0$. We know from standard mathematical methods that the solutions of eqns [12]–[14] are then a combination of $\exp(\alpha t)$ terms. If the threshold condition in eqn [16] is satisfied, a parameter α in the combination becomes positive, so the population of level 2 increases exponentially (photon avalanche) at early times and the dynamics of the system is unstable.

The physical meaning of eqns [15] and [16] is clarified by introducing the yield η_{CR} of the cross-relaxation process:

$$\eta_{CR} = \frac{k}{k + 1/\tau_2} \qquad [17]$$

and the yield η of the $2 \rightarrow 1$ de-excitation:

$$\eta = \frac{b_{21}/\tau_2}{k + 1/\tau_2} \qquad [18]$$

Then, the threshold condition in eqn [16] can be written as:

$$\left(\frac{I\sigma_1}{I\sigma_1 + 1/\tau_1}\right)(2\eta_{CR} + \eta) > 1 \qquad [19]$$

Starting with one ion in level 1, the first parenthesis in eqn [19] is the yield with which it will be promoted to level 2, and the second parenthesis is the yield of the backward path $2 \rightarrow 1$. So, the left-hand side of eqn [19] is the new number of ions after a looping path $1 \rightarrow 2 \rightarrow 1$, which has to be higher than 1 for the avalanche to occur. If we consider a lot of successive loops, the origin of the exponential behavior is clear. The role of the first ion in level 1 (and the weak pumping rate in the ground state $I\sigma_0$) is limited to initialize the process. Note that the value of the first parenthesis in eqn [19] is less than 1, so the second parenthesis has to be higher than 1. This condition is equivalent to eqn [15].

Materials: The Energy Gap Law

As we can see from the above mechanisms, up conversion needs intermediate energy levels and is not efficient if these cannot accumulate populations due to a short lifetime. The spontaneous rate of de-excitation of an electronic level is the sum of a radiative part W^{rad} and of a nonradiative part W^{nrad}. The radiative part can be obtained through the absorption cross-section corresponding to the transition or from the Judd–Ofelt theory which also exploits the absorption spectrum. The nonradiative rate is then obtained by subtracting W^{rad} from the measured fluorescence decay rate of the level. Based on the measurements in many hosts for the different trivalent rare earth ions, at different temperatures, the nonradiative decay rate of a J-level towards the next lower J′-level was found to have the form:

$$W_{\mathrm{JJ'}}^{\mathrm{nrad}} = B\exp(-\beta\Delta E_{\mathrm{JJ'}})\left(1-\exp\left(-\frac{\omega_{\max}}{kT}\right)\right)^{-p} \quad [20]$$

where $\Delta E_{\mathrm{JJ'}}$ is the energy gap between the two levels, $p = \Delta E_{\mathrm{JJ'}}/\omega_{\max}$, $\omega_{\max}$ being the energy of an effective optical phonon of the host. The values of B, β and $\omega_{\max}$ are found by fitting with experimental data. An example of such a fit is represented in Figure 4.

Equation [20] is the well-known and familiar 'energy gap law'. We have represented it for several oxide and nonoxide hosts at $T = 300$ K in Figure 5. It is clear that chloride, bromide, fluoride will be favorable hosts for up conversion due to their low phonon energy ($LaBr_3$: $\omega_{\max} = 175\ \mathrm{cm}^{-1}$, $LiYF_4$: $\omega_{\max} = 400\ \mathrm{cm}^{-1}$, $Y_3Al_5O_{12}$: $\omega_{\max} = 700\ \mathrm{cm}^{-1}$, silicate glass: $\omega_{\max} = 1100\ \mathrm{cm}^{-1}$). Let us mention also a fluorozirconate glass, ZBLAN, widely used in up-conversion fiber lasers.

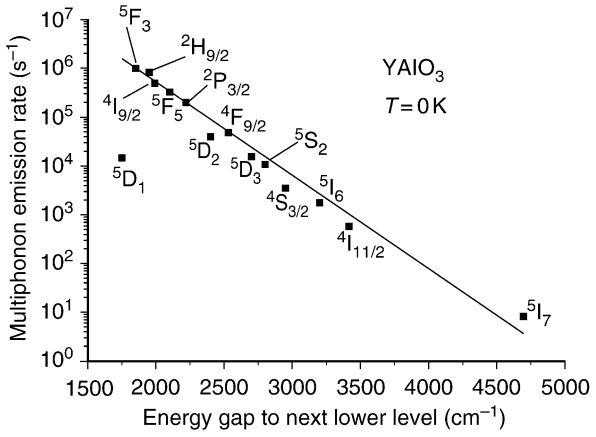


Figure 4 Energy-gap dependence of nonradiative decay rates in $YAlO_3$.

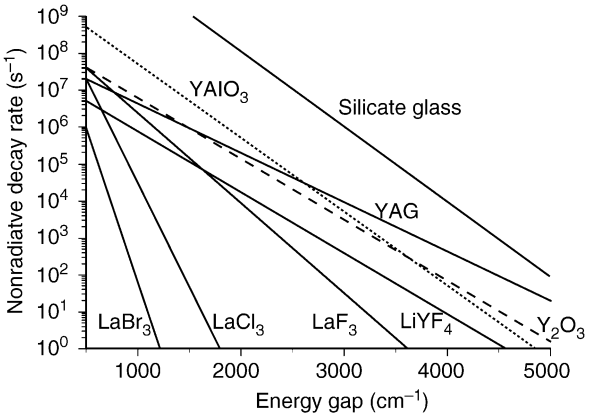


Figure 5 Energy-gap law for different hosts at 300 K.

Up-Conversion from Second-Order Optical Nonlinearity

The inter-atomic electric field acting on the electrons inside a medium has a magnitude of $\sim 10^8$ V/cm and derives from a potential that is anharmonic. The electric field **E** of an electromagnetic wave propagating through the medium drives the electrons beyond the quadratic region of the potential in the case of high **E**. So, the electron response and the associated polarization **P** take the form of a function of **E**:

$$\mathbf{P}(\mathbf{E}) = \varepsilon_0(\chi^{(1)}:\mathbf{E} + \chi^{(2)}:\mathbf{E}^2 + \chi^{(3)}:\mathbf{E}^3 + \cdots) \quad [21]$$

The different terms in eqn [21] are responsible for many optical phenomena because in Maxwell electromagnetic theory, the polarization is the source of the waves. We have selected in Table 1 the terms relevant for linear effects and for the purpose of frequency up conversions: second-harmonic generation (SHG) and sum-frequency mixing, due to second-order terms. It should be noticed that the latter have a nonzero value only in noncentrosymmetric materials.

Figure 6 visualizes the up conversion from SHG: two photons of the fundamental field at angular frequency ω_1 are annihilated while one photon at twice the angular frequency is created. A similar picture could be drawn for sum-frequency up conversion: SHG is the degenerate case where $\omega_1 = \omega_2$.

The second-order nonlinear processes have an efficiency given by an effective coefficient d_{eff} given hereafter:

$$d_{\mathrm{eff}} = \sum_{ijk} e_i(\omega_3)d_{ijk}e_j(\omega_1)e_k(\omega_2) \quad [22]$$

where $e_i(\omega_l)$ is the ith component of the unit vector of the electric field of the wave l at angular frequency ω_l and $d_{ijk} = (1/2)\chi_{ijk}$.

Table 1 Relevant terms for frequency up-conversion

Optical process	*Term*	*New wave*
Linear		
Refractive index, absorption, stimulated emission	$\chi^{(1)}(-\omega;\omega)$	
Second order nonlinear		
Second harmonic generation	$\chi^{(2)}(-\omega_3;\omega_1,\omega_1)$	$\omega_3 = 2\omega_1$
Sum-frequency mixing	$\chi^{(2)}(-\omega_3;\omega_1, \pm\omega_2)$	$\omega_3 = \omega_1 + \omega_2$

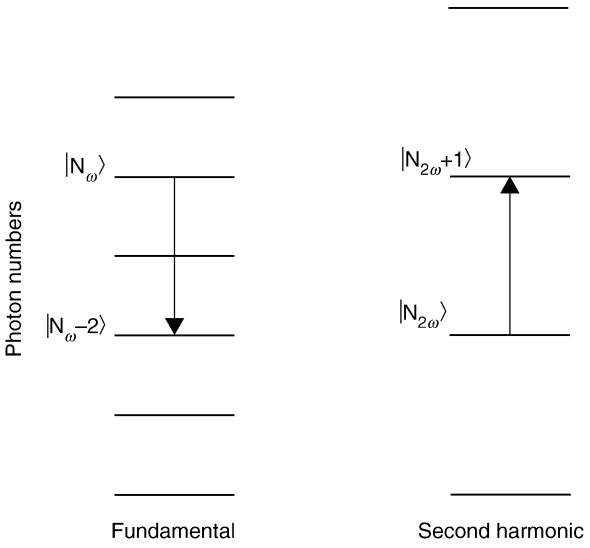


Figure 6 Up-conversion from second harmonic generation.

Moreover, the wave propagation equations in the slow varying envelope approximation, given hereafter for the E_3 electric field of the sum-frequency mixing wave:

$$\frac{dE_3}{dz} = \frac{i\omega_3 d_{eff}}{n_3 c} E_1 E_2 \quad \exp(i(k_1 + k_2 - k_3)z) \qquad [23]$$

$$(\text{with } k_j = \omega_j n_j / c)$$

impose that the frequency conversion is efficient in practice, only if the following phase matching condition is satisfied:

$$\omega_1 n_1 \uparrow (\theta, \varphi) + \omega_2 n_2 \uparrow (\theta, \varphi) = \omega_3 n_3 \downarrow (\theta, \varphi) \qquad [24]$$

where $n_i \uparrow$ and $n_i \downarrow$ are respectively the upper and lower refractive index in the direction of propagation (θ, φ) (n_i in eqn [23] are the same with a simplified notation). Equation [24] is restricted here for simplicity to collinear type I (waves 1 and 2 and the same polarization) and is the expression of photon momentum conservation. It is usually achieved from the birefringence of the crystals. When this is not possible, another technique can be used, namely quasi-phase matching, which involves reversing periodically the sign of the nonlinear optical coefficient.

In bifunctional crystals, the laser effect and the $\chi^{(2)}$ interaction occur simultaneously inside the same crystal. In the case of the self-frequency doubling laser, the angular frequency $\omega_1 = \omega_2$ in eqn [24] is that of the fundamental laser wave. In the case of the self-sum frequency mixing laser, ω_1 corresponds to the fundamental laser wave and ω_2 to the pump wave. As we can see from eqns [22]–[24], the polarization of the laser emission is crucial in order for the device to work. So, the laser stimulated emission cross-section σ for propagation in the phase matching direction (θ, ϕ) in the adequate polarization has to be evaluated. This can be performed with the formula:

$$\sigma(\theta, \varphi) = \frac{\sigma_X \sigma_Y \sigma_Z}{\sqrt{\sigma_Y^2 \sigma_Z^2 e_X^2 + \sigma_X^2 \sigma_Z^2 e_Y^2 + \sigma_X^2 \sigma_Y^2 e_Z^2}} \qquad [25]$$

where $\sigma_{X,Y,Z}$ are the emission cross-sections for X, Y, Z-polarizations and $e_{X,Y,Z}$ are the components of the unit vector or the electric field of the laser wave.

To summarize, the conditions necessary for a crystal to be a bi-functional material are: (i) it must be noncentrosymmetric; (ii) it must accept fluorescent doping (usually with Nd^{3+} or Yb^{3+}); and (iii) it must be phase matchable for its laser emission. In practice, only a few crystals satisfy these requirements and have been tried with some success. The main ones are $LiNbO_3$, $LaBGeO_5$, $Ba_2NaNb_5O_{15}$, β'-$Gd_2(MoO_4)_3$, $YAl_3(BO_3)_4$, $GdAl_3(BO_3)_4$, $CaY_4(BO_3)_3O$, and $CaGd_4(BO_3)_3O$.

Up-Conversion Lasers Based on Luminescence Mechanisms in Materials

Energy Transfers

Up-conversion lasers from energy transfers have the advantage of using a single pump laser. This latter populates firstly a long-lived intermediate level. A typical example, based on $LiYF_4$:Er^{3+} crystal, is provided by Figure 7.

The pump at 802 nm from a GaAlAs diode laser, or at 969 nm, populates efficiently at the 8 ms lifetime $^4I_{11/2}$ level. The green laser, at 551 nm, corresponds to the $^4S_{3/2} \rightarrow {}^4I_{15/2}$ transition and it can be shown that energy transfer is an essential part of the mechanism by stopping abruptly the pump laser: green lasing continues to occur for several hundred microseconds because the $^4S_{3/2}$ level is still fed (an excited-state

absorption would stop immediately). The involved energy transfer is:

$$^4I_{11/2} + {}^4I_{11/2} \rightarrow {}^4F_{7/2} + {}^4I_{15/2} \qquad [26]$$

and leads to a 20 mW laser threshold and 2.3 mW green output power at 50 K temperature upon 95 mW of incident power.

By using mirrors with high transmission in the green range but having high reflectivity at blue wavelengths, blue laser emission at 470 nm can be sustained, corresponding to the $^2P_{3/2} \rightarrow {}^4I_{11/2}$ transition. It originates from a mechanism involving three up-conversion energy transfers:

$$^4I_{11/2} + {}^4I_{11/2} \rightarrow {}^4F_{7/2} + {}^4I_{15/2}$$
$$^4I_{11/2} + {}^4I_{13/2} \rightarrow {}^4F_{9/2} + {}^4I_{15/2} \qquad [27]$$
$$^4S_{3/2} + {}^4F_{9/2} \rightarrow {}^4I_{15/2} + {}^2K_{13/2}$$

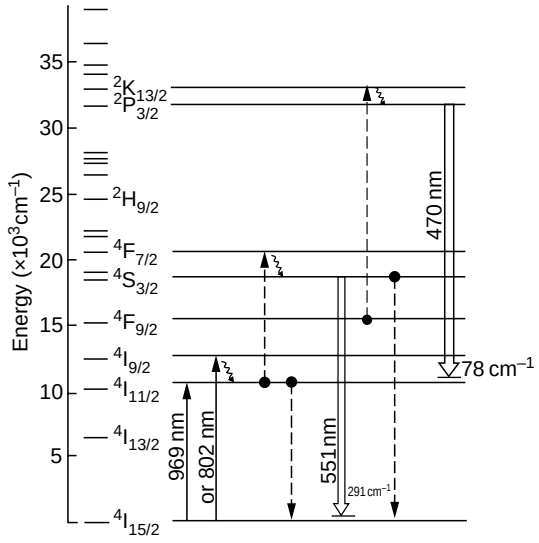


Figure 7 Energy level diagram of $LiYF_4:Er^{3+}$. The dashed lines indicate the energy transfers responsible for the green and blue up-conversion lasing. Reproduced with permission of American Institute of Physics from Macfarlane RM, Tong F, Silversmith AJ and Lenth W (1988) Violet CW neodymium upconversion lasers. *Applied Physics Letters* 52(16): 1300.

The illustration in Figure 7 shows up-conversion energy transfers between ions of the same chemical species, but co-doping the laser host with two ions of different chemical species provides the opportunity of separating their roles: one species is devoted to pump absorption (this is the sensitizer) and the other one is devoted to lasing (this is the activator). Due to the development of efficient semiconductor InGaAs diodes emitting near 980 nm, the most widely used sensitizer is Yb^{3+}. Table 2 provides a list of several up-conversion lasers based on energy transfer mechanisms, operating at room temperature. We can see that among the most efficient lasers, are those based on rare-earth ions doped glass fibers. The reason is that in fibers, the pump and laser waves remain confined inside a core (typically 5 μm diameter) and have high energy densities over a long length (tenths of cms), which is favorable for all up-conversion mechanisms.

Sequential Two-Photon Absorption

An example of an up-conversion laser pumped by a two-photon absorption mechanism is represented in Figure 8. The material is $LaF_3:Nd^{3+}$ and violet lasing at 380 nm corresponds to the $^4D_{3/2} \rightarrow {}^4I_{11/2}$ transition.

Transition from the ground state up to the $^4F_{5/2}$ level is firstly obtained from an infrared beam at 790 nm in order to feed the $^4F_{3/2}$ intermediate level after a fast nonradiative de-excitation. The $^4F_{3/2}$ level has a rather long lifetime: 700 μs, so it accumulates

Table 2 Main room temperature up-conversion lasers operating from energy transfers

Laser material	*Laser wavelength (nm)*	*Pump wavelength (nm)*	*Output power*
$BaY_{1.4}Yb_{0.59}Ho_{0.01}F_8$	670	1540 + 1054	
$BaYYb_{0.998}Tm_{0.002}F_8$	649	1054	1%
$BaYYb_{0.99}Tm_{0.01}F_8$	799–649–510–455	960	
$BaYYb_{0.99}Tm_{0.01}F_8$	348	960	
$LiYF_4$:Er(1%):Yb(3%)	551	966	37 mW
$LiY_{0.89}Yb_{0.1}Tm_{0.01}F_4$	810–792	969	80 mW
$LiY_{0.89}Yb_{0.1}Tm_{0.01}F_4$	650	969	5 mW
$LiKYF_5$:Er (1%)	550	808	150 mW
Fiber:Yb:Pr	635	849	20 mW
Fiber:Yb:Pr	635	1016	6.2 mW
Fiber:Yb:Pr	521	833	0.7 mW
Fiber:Yb:Pr	635	860	4 mW

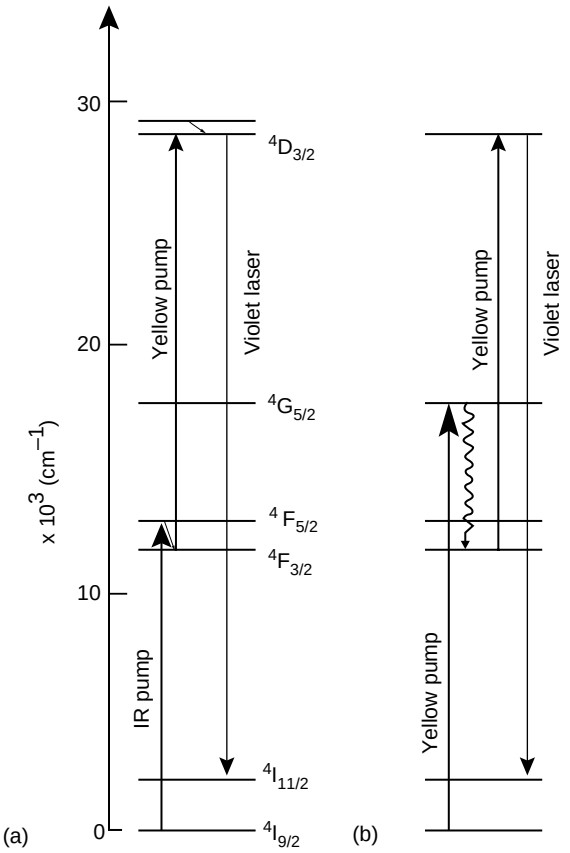


Figure 8 Simplified energy level diagram of LaF_3:Nd^{3+} and up-conversion lasing from two-photon absorption. (a) Two different pump wavelengths, (b) doubly resonant single pump wavelength. Reproduced with permission of American Institute of Physics from Lenth W, Silversmith AJ and Macfarlane RM (1988) Green infrared pumped erbium up conversion lasers. In: TAM AC, Gole JL and Stwalley WC (eds) *Advances in Laser Science – III*. AIP Conference Proceedings n°172: 8–12.

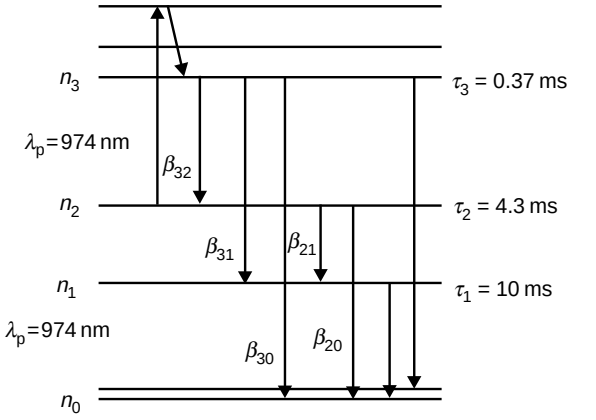


Figure 9 Simplified energy level diagram of $LiYF_4$:Er^{3+} and up-conversion lasing from two-photon absorption (doubly resonant). Reproduced with permission of Institute of Physics Publishing from Huber G (1999) Visible cw solid-state lasers. *Advances in Lasers and Applications*, Bristol and Philadelphia: Scottish Universities Summer School in Physics & Institute of Physics Publishing, 19.

enough population that a second yellow pumping at 591 nm (Figure 8a) can efficiently populate the $^4D_{3/2}$ initial laser level. With 110 mW of infrared pump and 300 mW of yellow pump, 12 mW of violet output were obtained at 20 K temperature.

A simpler device is obtained if a single-pump beam is used. This is the case in Figure 8b, where a yellow pump at 578 nm provides both ground state and excited state absorption. This doubly resonant scheme is less efficient in LaF_3:Nd^{3+} than the previous one but in Figure 9, we show another example: $LiYF_4$:Er^{3+}, working at room temperature. The lasing $^4S_{3/2} \rightarrow {}^4I_{15/2}$ transition at 551 nm delivers 20 mW upon 1600 mW pumping at 974 nm.

The simplified Er^{3+} level scheme of Figure 9 contains four main levels with population densities: $n_0 = [{}^4I_{15/2}]$, $n_1 = [{}^4I_{13/2}]$, $n_2 = [{}^4I_{11/2}]$, $n_3 = [{}^4S_{3/2}]$. The rate equations of populations of this system can

Table 3 Main room temperature up-conversion lasers operating from two-photon absorption

Laser material	*Laser wavelength (nm)*	*Pump wavelength (nm)*	*Output power*
$Y_3Al_5O_{12}$:Er 1%	561	647 + 810	
$LiYF_4$:Er 1%	551	647 + 810	0.95 mJ
$LiYF_4$:Er 1%	551	810	40 mW
$LiYF_4$:Er 1%	551	974	45 mW
$LiYF_4$:Er	551	974	20 mW
KYF_4:Er 1%	562	647 + 810	0.95 mJ
$LiLuF_4$:Er	552	974	70 mW
$LiLuF_4$:Er	552	970	213 mW
$LiYF_4$:Tm 1%	453–450	781 + 649	0.2 mJ
Fiber:Tm	480	1120	57 mW
Fiber:Tm	480	1100–1180	45 mW
Fiber:Tm	455	645 + 1064	3 mW
Fiber:Tm	803–816	1064	1.2 W
Fiber:Tm	482	1123	120 mW
Fiber:Yb:Tm	650	1120	
Fiber:Ho	540–553	643	38 mW
Fiber:Ho	544–549	645	20 mW
Fiber:Er	548	800	15 mW
Fiber:Er	546	801	23 mW
Fiber:Er	544	971	12 mW
Fiber:Er	543	800	
Fiber:Pr	635	1010 + 835	180 mW
Fiber:Pr	605	1010 + 835	30 mW
Fiber:Pr	635	1020 + 840	54 mW
Fiber:Pr	520	1020 + 840	20 mW
Fiber:Pr	491	1020 + 840	7 mW
Fiber:Nd	381	590	74 μW
Fiber:Nd	412	590	500 μW

be written and solved because all the implied parameters (lifetimes, branching ratios, ground, and excited states absorption cross-sections) have been measured in this material. Predictive models of the up-conversion green laser are successful, matching experimental measurements, and confirm that the doubly resonant mechanism is realistic at low doping concentration.

Table 3 summarizes the performances of the main up-conversion lasers based on sequential photon absorption and working at room temperature.

Photon Avalanche

Photon avalanche up-conversion was observed mainly in Nd^{3+}, Tm^{3+}, Er^{3+}, and Pr^{3+} doped materials. As an illustration of this mechanism, let us consider the case of the continuous wave Pr^{3+}–Yb^{3+}-doped fluorozirconate fiber laser (3000 ppm Pr, 20 000 ppm Yb). With two Ti:sapphire pump lasers operating at 852 and 826 nm (Figure 10), an output power at 635 nm as high as 1.02 W was obtained with an incident pump power of 5.51 W (19% slope efficiency).

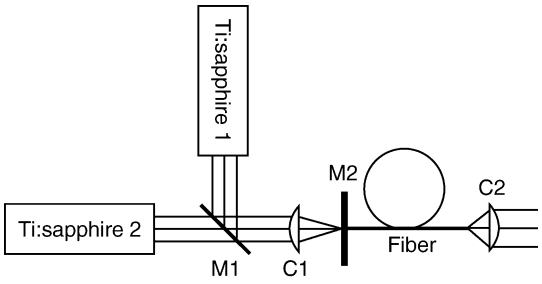


Figure 10 Experimental set-up for Yb–Pr-doped fiber up-conversion laser. M1: dichroic mirror, C1, C2: lenses, M2 dichroic input mirror. Reproduced with permission of Optical Society of America from Sandrock T, Scheife H, Heumann E and Huber G (1997) High power continuous wave up-conversion fiber laser at room temperature. *Optics Letters* 22(11): 809.

This remarkable result can be explained quantitatively by modeling the photon avalanche with a five-level diagram and restricted in Figure 11 to a single laser pump at 850 nm. The laser emission corresponds to the $^3P_0 \rightarrow {}^3F_2$ transition. The five relevant levels have the population densities: $n_0 = [{}^3H_4(Pr)]$, $n_1 = [{}^1G_4(Pr)]$, $n_2 = [{}^3P_0(Pr)]$, $n_3 = [{}^2F_{7/2}(Yb)]$, $n_4 = [{}^2F_{5/2}(Yb)]$.

The two components of the mechanism are:

(i) the excited state absorption of the pump by the spin allowed transition: $^1G_4 \rightarrow {}^1I_6$
(ii) the process leading to the doubling of the $^1G_4$ population, in this instance through two energy transfers:

$$\begin{aligned} {}^1I_6(Pr) + {}^2F_{7/2}(Yb) &\rightarrow {}^1G_4(Pr) + {}^2F_{5/2}(Yb) \\ {}^2F_{5/2}(Yb) + {}^3H_4(Pr) &\rightarrow {}^2F_{7/2}(Yb) + {}^1G_4(Pr) \end{aligned} \quad [28]$$

The rate equation analysis predicts that the avalanche threshold occurs at about 1 W pump and the long fluorescence rise time is somewhat shortened at higher pump power, as observed experimentally.

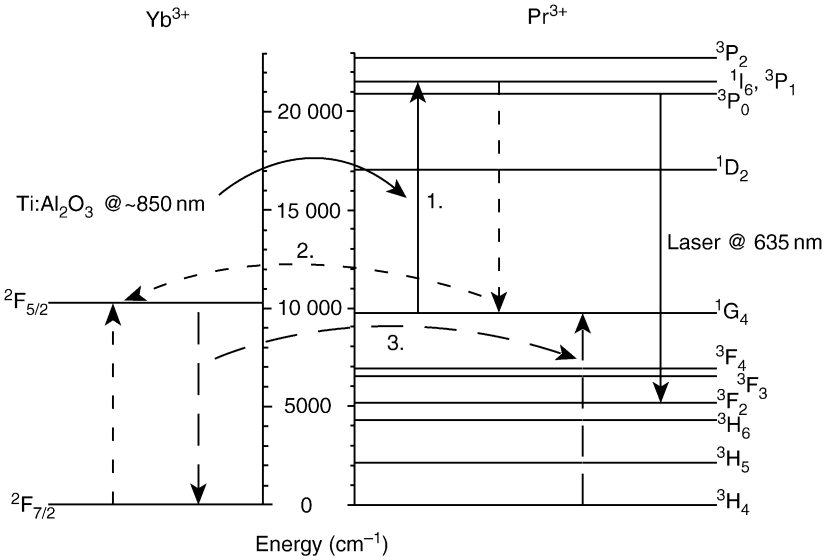


Figure 11 Energy level diagram of fluorozirconate: Yb^{3+}:Pr^{3+} fiber. The dashed lines indicate the energy transfers. Reproduced with permission of the Optical Society of America from Sandrock T, Scheife H, Heumann E and Huber G (1997) High power continuous wave up-conversion fiber laser at room temperature. *Optics Letters* 22(11): 809.

Table 4 Main room temperature up-conversion lasers operating from photon avalanche

Laser material	*Laser wavelength (nm)*	*Pump wavelength (nm)*	*Output power*
BaY_2F_8:Yb:Pr	607.5	822	55 mW
BaY_2F_8:Yb:Pr	638.7	841	26 mW
$LiY_{0.89}Yb_{0.1}Pr_{0.01}F_4$	720	830	1%
$LiY_{0.89}Yb_{0.1}Pr_{0.01}F_4$	639.5	830	
Fiber:Yb:Pr	635–637	780–880	300 mW
Fiber:Yb:Pr	605–622	780–880	44 mW
Fiber:Yb:Pr	517–540	780–820	20 mW
Fiber:Yb:Pr	491–493	780–880	4 mW
Fiber:Yb:Pr	635	850	1.02 W
Fiber:Yb:Pr	635	850	2.06 W
Fiber:Yb:Pr	520	850	0.3 W

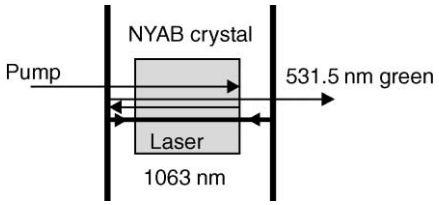


Figure 12 Scheme of a self-frequency doubling laser. Note that in reality the three beams are superimposed.

Table 4 summarizes the performances of the main up-conversion lasers based on photon avalanche mechanisms.

Up-Conversion Lasers Based on Bi-Functional Crystals

The laser stimulated emission in a bi-functional crystal was first obtained in 1969 with Tm^{3+} in $LiNbO_3$ but the most used ion is, of course, Nd^{3+} working in the two channels:

$$^4F_{3/2} \rightarrow {}^4I_{11/2} \qquad [29]$$

$$^4F_{3/2} \rightarrow {}^4I_{13/2} \qquad [30]$$

and more recently Yb^{3+} working in the channel:

$$^2F_{5/2} \rightarrow {}^2F_{7/2} \qquad [31]$$

Channels in eqns [29] and [31] operate around 1060 nm wavelength and the channel in eqn [30] near 1338 nm.

The Self-Frequency Doubling Laser

A typical laser scheme is shown in Figure 12.

Table 5 Main results of Nd^{3+} and Yb^{3+} green self-frequency doubling lasers near 530 nm (eqns [29] and [31])

Crystal	*Input power (mW)*	*Output power (mW)*	*Pumping*
$LiNbO_3$:MgO:Nd	215	1	Dye laser
$YAl_3(BO_3)_4$:Nd	870	10	Laser diode
$YAl_3(BO_3)_4$:Nd	280	3	Laser diode
$YAl_3(BO_3)_4$:Nd	400	69	Laser diode
$YAl_3(BO_3)_4$:Nd	1380	51	Laser diode
$YAl_3(BO_3)_4$:Nd	369	35	Laser diode
$LiNbO_3$:MgO:Nd	850	18	Dye laser
$LiNbO_3$:Sc_2O_3:Nd	65	0.14	Ti:sapphire laser
$LiNbO_3$:MgO:Nd	100	0.2	Laser diode
$YAl_3(BO_3)_4$:Nd	1600	225	Laser diode
$YAl_3(BO_3)_4$:Nd	2200	450	Ti:sapphire laser
$(Y,Lu)Al_3(BO_3)_4$:Nd	880	24	Laser diode
$LiNbO_3$:ZnO:Nd	430	0.65	Ti:sapphire laser
$Ba_2NaNb_5O_{15}$:Nd	270	46	Ti:sapphire laser
$CaY_4(BO_3)_3O$:Nd	900	62	Laser diode
$CaGd_4(BO_3)_3O$:Nd	1600	192	Ti:sapphire laser
$CaGd_4(BO_3)_3O$:Nd	1250	115	Laser diode
$CaGd_4(BO_3)_3O$:Nd	1560	225	Ti:sapphire laser
$YAl_3(BO_3)_4$:Yb	11 000	1100	Laser diode
$GdAl_3(BO_3)_4$:Nd	2.8 mJ/pulse	0.12 mJ/pulse	Pulsed dye laser

The laser beam cannot escape from the cavity and dichroic mirrors are used. The input mirror has high transmission at the pump wavelength and is highly reflective at laser- and second-harmonic wavelengths. The output mirror is highly reflective at laser wavelength and has high transmission for second harmonic. Channels in eqns [29] and [31] generate green light near 530 nm and the channel in eqn [30] generates red light near 669 nm.

The most exploited bi-functional crystal is $YAl_3(BO_3)_4$:Nd^{3+}, well-known as NYAB. It is a negative uniaxial trigonal crystal (extraordinary index lower than the ordinary one) and its laser emission is easily observed in ordinary polarization with a high cross-section: 2×10^{-19} cm^2 at 1063 nm. Type I phase matching occurs at a polar angle $\theta = 30.7°$ and 26.8° for channels in eqns [29] and [30] respectively. The effective nonlinear optical coefficient d_{eff} is close to 1.4 pm V^{-1} at azimuthal angle $\phi = 0°$.

The Yb^{3+} ion has some advantages over the Nd^{3+} ion due to its very simple energy level scheme: there is no excited state absorption at the laser wavelength, no up-conversion losses, no concentration quenching, and no absorption in the green. The small Stokes shift between pump and laser emission reduces the thermal loading of the material during laser operation. A self-doubling laser based on $YAl_3(BO_3)_4$:Yb^{3+} of crystal has produced 1.1 W green power upon 11 W diode pumping.

The drawback of $YAl_3(BO_3)_4$ is that it is rather difficult to grow because it is not congruent. This difficulty was overcome by the discovery at Ecole Nationale Supérieure de Chimie de Paris of the rare-earth calcium oxoborate $CaGd_4(BO_3)_3O$ which can be grown to a large size by the Czokhralski method. It is a monoclinic biaxial crystal and doped with Nd^{3+}, it was firstly exploited (channel in eqn [29]) in the XY principal plane at $\theta = 90°$, $\phi = 46°$ with ($d_{eff} = 0.5$ pm V^{-1}). It was soon recognized that the optimum phase matching direction ($d_{eff} = 1.68$ pm × V^{-1}) occurred out of the principal planes, in the direction $\theta = 66.8°$, $\phi = 132.6°$, and high green power can be obtained: 225 mW under 1.56 W pump.

Table 5 summarizes the main self-frequency doubling results obtained with the channel in eqn [29].

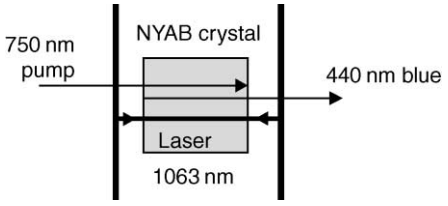


Figure 13 Scheme of a self-sum frequency mixing laser. Note that in reality the three beams are superimposed.

The Self-Sum Frequency Mixing Laser

The example in Figure 13 gives the main features of such a laser. The input mirror has high transmission at the pump wavelength and both mirrors are highly reflective at laser wavelength. The output mirror has high transmission at sum frequency mixing wavelength.

The laser wave at angular frequency ω_1 in Table 1 corresponds to the channel in eqns [29] or [30] in the case of Nd^{3+}. The wave at angular frequency ω_2 in Table 1 has two different roles: first it excites the Nd^{3+} laser center and secondly its nonabsorbed part is up converted by the second-order nonlinear process. ω_2 (corresponding to the wavelength λ_2) is then chosen to match the main Nd^{3+} absorption lines:

$$\begin{aligned} &{}^4I_{9/2} \rightarrow {}^4G_{5/2} - {}^2G_{7/2}\ (\lambda_2 = 590\ \text{nm}) \\ &{}^4I_{9/2} \rightarrow {}^4F_{7/2} - {}^4S_{3/2}\ (\lambda_2 = 750\ \text{nm}) \\ &{}^4I_{9/2} \rightarrow {}^4F_{5/2} - {}^2H_{9/2}\ (\lambda = 800\ \text{nm}) \\ &{}^4I_{9/2} \rightarrow {}^4F_{3/2}\ (\lambda_2 = 880\ \text{nm}) \end{aligned} \qquad [32]$$

The most efficient self-frequency mixing lasers based on channels in eqns [29]–[31] are gathered in Table 6.

See also

Materials Characterization Techniques: $\chi^{(2)}$. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation.

Table 6 Main results of Nd^{3+} self-sum frequency mixing lasers based on eqn [29]

Crystal	*Pump wavelength (nm)*	*Generated wavelength (nm)*	*Output power*
$YAl_3(BO_3)_4$:Nd	740–760	436–443	0.16 mJ/pulse
$GdAl_3(BO_3)_4$:Nd	740–760	436–443	0.403 mJ/pulse
$CaGd_4(BO_3)_3O$:Nd	811	465	1 mW
$YAl_3(BO_3)_4$:Nd	585–600	377–383	0.25 mJ/pulse
$GdAl_3(BO_3)_4$:Nd	587–597	378–382	0.105 mJ/pulse
$YAl_3(BO_3)_4$:Nd	488, 515	330, 380	0.2 mW

Further Reading

Brenier A (2000) The self-doubling and summing lasers: overview and modeling. *Journal of Luminescence* 91: 121–132.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Huber G (1999) Visible cw solid-state lasers. *Advances in Lasers and Applications*. Bristol and Philadelphia: Scottish Universities Summer School in Physics & Institute of Physics Publishing.

Joubert MF (1999) Photon avalanche up-conversion in rare earth laser materials. *Optical Materials* 11: 181–203.

Kaminskii AA (1996) *Crystalline Lasers: Physical Processes and Operating Schemes*. CRC Press, Inc.

Lenth W and Macfarlane RM (1992) Upconversion lasers. *Optics & Photonics News* 3: 8–15.

Powell RC (1998) *Physics of Solid-State Laser Materials*. New York: Springer-Verlag.

Risk WP, Gosnell T and Nurmikko AV (2003) *Compact Blue-Green Lasers*. Cambridge, UK: Cambridge University Press.

Weber MJ (2001) *Handbook of Lasers*. Boca Raton, FL: CRC Press.

Yariv A (1995) *Quantum Electronics,* Third Edition. John Wiley & Sons.

LASER-INDUCED DAMAGE OF OPTICAL MATERIALS

A J Glass and A H Guenther, The University of New Mexico, Albuquerque, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Shortly after the invention of the laser, in the early 1960s, it became apparent that laser performance was limited by the ability of optical materials to withstand high energy and power densities. Broadly speaking, laser damage arises from two causes: (i) thermal effects, which are associated with high average power; and (ii) dielectric breakdown, which is associated with high values of the local electric field. Early researchers quickly recognized that most of the observed damage effects arose from extrinsic causes, such as impurities and inclusions in the bulk material, or particles or defects on exposed optical surfaces. At that time, optical materials had not been developed for high-power use, and lacked the quality needed for laser applications. In fact, in many situations, the laser provided the inspection light source used to assess the properties of the materials, such as absorption, defects, scattering centers, surface topography, and dynamic behavior. Thus, the field of laser damage studies was broadened to include many aspects of material preparation and characterization. At high-power densities, nonlinear effects become important, often limiting performance without creating any irreversible change in the material's properties. The most general definition of laser-induced damage includes all effects arising in optical materials and components that limit or degrade the performance of laser systems.

The pursuit of improved performance at high values of power and energy densities was motivated by several factors and applications. In general, improved tolerance to high flux and fluence leads to smaller, lighter, and more efficient systems. These are particularly important for military and energy applications. Since the 1970s, the laser fusion program alone has produced an extensive body of research on laser-induced damage in high-power nanosecond-scale systems, which has enabled orders of magnitude improvement in the damage resistance and extraction efficiency of glass lasers. More recently, with the development of fiber laser systems producing and transporting a kilowatt of average power in a fiber of less than 100 μm diameter, laser damage phenomena on an entirely different spatial scale have become important.

Since the early 1970s, the literature of laser damage research has been collected in a series of annual publications of the Boulder Damage Symposium, initiated by the authors of this article, which was first held in 1969. Recently the papers presented at the Damage Symposium in the years 1969 to 1998 were collected and published in CD form by the SPIE. A CD containing the proceedings of the next five years (1999–2003) will be released by SPIE this year. Also in the late 1960s, in the Soviet Union, Academician Aleksei Bonch-Bruevich launched a comprehensive research project on the nonresonant interaction of laser radiation with matter, opening up a new branch of photonics research, known in the Soviet Union as power optics. In 1969, he founded the All-Union Conference on Interaction of Optical Radiation with Matter (AURICOM), which was held every two to three years until 1990. After 1990, the conference was renamed Nonresonant Laser-Matter Interaction (NLMI), and chaired by Professor Mikhail Libenson. NLMI was held in St Petersburg in 1996, 2000, and

2003. The proceedings of the 1996 AURICOM and subsequent NLMI conferences have been published by the SPIE. We are saddened to note that Professor Libenson died in February 2004.

The Boulder Damage Symposium grew out of a one-day Symposium convened in Boulder in 1969 under the auspices of the ASTM Subcommittee on Lasers, to establish standards for laser materials. In their summary of the first Symposium, the conference organizers remarked:

> The entire question of the nature of the standards remains to be addressed. Should they take the form of energy density at which catastrophic damage is likely to occur, or should they be specified in terms of a mean number of shots at a given level of energy density before certain degradation of performance is measured? The latter would seem to express the kind of information that the buyer of laser materials would find most useful. Once a standard is agreed upon, a meaningful test configuration must be established. It is likely that this will be a well-controlled oscillator amplifier chain with the test specimen an active element of the system, but this is not certain. Another point to be determined is that of the accuracy to which its specifications should be written. This depends in part on the kind of quality control the producers of laser glass believe is feasible.

The questions raised in this first Symposium remain at the core of laser damage studies today. How do we define damage, in terms of the physical effect on the material or in terms of the degradation of performance of the system? How do we measure damage thresholds or values? How do we specify the properties and preparation of the sample, including the relevant experimental and material variables? How can we ensure that damage values are reproducible, unless we make the measurements in a very well controlled and characterized system? With what precision should damage levels be stated, in order to be useful? Extensive research was required in order to answer these questions, and to elucidate the fundamental nature of damage mechanisms. Many of the results of this research are documented in the 35 year record of the Boulder Damage Symposium and the corresponding Russian-based conferences.

From the beginning of this area of research, investigators devoted a significant effort to establishing the 'first causes' of damage phenomena. However, they quickly realized that in most cases, damage occurred at surfaces and material interfaces, or was mediated by impurities or inclusions in bulk materials. As material fabrication processes improved, and higher-quality materials became available, observed damage thresholds approached the intrinsic limits of the pure material. This was particularly true of silicate glass, metal reflectors, and in some cases, polymers. In fact, silicate glass has become one of the purest noncrystalline materials available in bulk quantities. However, for practical purposes, most laser systems are limited by damage at surfaces, coatings, or impurities. Impurities aggregate at grain boundaries in crystalline materials, and coalesce into inclusions in glasses, resulting in absorption, scattering, and field enhancement.

The coherent nature of laser light makes any scattering phenomena potentially damaging, because the scattered light interferes constructively with the laser beam, leading to the formation of intensity hot spots. Every interface in an optical system is potentially a source of constructive interference between the reflected and incident waves. Multilayer dielectric coatings are particularly vulnerable to damage, due to the presence of residual stress as well as field enhancement from Fresnel reflection. This has led to attempts to mitigate the problem by use of specific designs, such as avoiding AR coatings by using Brewster-angle surfaces, or use of nonquarter wave layers to move the peak field away from the vicinity of boundaries where impurities tend to collect. Other damage-sensitive design considerations can be found in the literature. Multiple reflections are the basis of photonic crystal optics, and applications of these devices to high-power density systems will undoubtedly be limited by field-enhanced damage phenomena.

In CW operation, of course, thermal effects dominate. They may be irreversible, creating permanent damage sites in the medium, or reversible, such as color centers that can be bleached out. As laser pulse durations vary from nanosecond to picosecond and femtosecond durations, the nature of the dominant damage mechanisms changes as well. As a rule of thumb, the scaling of damage thresholds as $\tau^{1/2}$, where τ is the pulse duration, was established empirically, and was found to be valid over a wide range of pulse durations. This scaling implies the presence of diffusion effects, either of heat or of electrons, in the vicinity of a damage center. As might be expected, this scaling does not apply for very short pulses (femtoseconds or less), where the extremely high field values can lead to instantaneous breakdown of the material.

Whether the damage is due to field enhancement or is impurity-mediated, it is difficult to establish a reproducible value for a damage threshold, unless the sample is macroscopic in extent, and the sampling laser beam is reproducibly of high spatial and temporal quality, preferably truly single-mode. Damage measurements on microscopic areas are unlikely to be representative of the entire sample. Damage will occur first at a point where the field is strongest or the material is weakest, and a sufficient

area must be tested to ensure that it includes such sites. Also, unless the beam quality of the laser used for testing is well controlled and well characterized, it will not be possible to reproduce the test results. It is equally important to characterize the samples, including their method of preparation, cleaning, and test environment. For damage research, as opposed to simple threshold testing, postmortem examination is also required, to elucidate the mechanism of the observed damage effect. However, we now have several techniques, such as photothermal deflection, that can provide useful pre-catastrophic indicators of the onset of laser damage, and thus are appropriate for quality control.

Professor Roger Wood had published a series of reviews and textbooks on laser damage in optical materials, which provide an overview of the field, with appropriate emphasis on measurement techniques, characterization of the test setup and sample, and implications for system design. He has also edited a collection of seminal publications in the field, published by the SPIE in 1990.

It is gratifying to those working in this field to observe how much of the research carried out to elucidate and, if possible, avoid the deleterious effects of laser damage has been applied to the constructive purpose of laser-assisted material processing and manufacturing. The introduction of low-cost, reliable femtosecond lasers has stimulated activity in this field, since it enables the experimenter to control to exquisite precision the power and energy deposited on a material. With femtosecond pulses, one can ablate material from the surface of an object while minimizing any collateral damage to the interior material. Surface layers of dirt can be removed from priceless art works such as oil paintings, without concern for damage to the object itself. Objects with arbitrary shape can be built up from metal particles, using laser additive manufacturing techniques. Thin films and photonic crystal structures can be created using laser-assisted deposition. Laser irradiation provides a power tool for surface cleaning, conditioning, and annealing. Laser interaction with materials and laser-assisted manufacturing techniques comprise a vital area of research in the materials science community.

In both the US and Russia, study of the interaction of intense coherent light with optical materials began with a disarmingly simple objective; to understand the fundamental mechanisms of interaction. In the review and summary of the 1969 ASTM Symposium, the editors wrote:

> In view of the number of problems remaining to be resolved, it is suggested that another Symposium on laser damage be held in 1970. Hopefully, at that time a better understanding of the nature of damage in laser glass will have been obtained. Higher threshold values will have been reproducibly achieved, and some agreement can be arrived at regarding useful and realistic standards.

Thirty-five years later, damage levels in optical materials have been significantly improved, and realistic and useful standards have been developed under the auspices of the International Organization for Standardization (ISO), and are now accepted globally. However, as the operating regimes of lasers have moved to femtosecond temporal scale and micrometer dimensions, new phenomena have arisen that require further research. An entirely new technology of laser-enabled manufacturing and materials conditioning has developed, based on the same physics as laser damage. Combining the disciplines of materials science and applied optics, the interaction of intense light with materials remains a scientifically challenging and economically important field of research.

See also

Optical Coatings: Laser Damage in Thin Film Coatings.

Further Reading

Bonch-Bruevich AM, Konov VI and Libenson MN (eds) (1990) *Optical Radiation Interaction with Matter*. SPIE Proceedings, vol. 1440.

Glass AJ and Guenther AH (eds) (1969) *Damage in Laser Glass Symposium: Review and Summary*. ASTM Special Technical Publication 469.

ISO Online Catalogue, Section 31.260, http://www.iso.ch/iso/en/CatalogueListPage.CatalogueList

Laser-Induced Damage in Optical Materials, Collected Papers 1969–1998, (CD-ROM), SPIE, Bellingham, Washington (1999).

Laser-Induced Damage in Optical Materials, Collected Papers 1999–2003, (CD-ROM), SPIE, Bellingham, Washington (2004) (To be published).

Libenson MN (ed.) (1997) *Non-Resonant Laser-Matter Interaction (NLMI-9)*. SPIE Proceedings, vol. 3093.

Libenson MN (ed.) (2001) *Non-Resonant Laser-Matter Interaction (NLMI-10)*. SPIE Proceedings, vol. 4423.

Libenson MN (ed.) (2004) *Non-Resonant Laser-Matter Interaction (NLMI-11)*. SPIE Proceedings, vol. 5506, (To be published).

Wood RM (ed.) (1986) *Laser Damage in Optical Materials*. Bristol, UK: Adam Hilger.

Wood RM (ed.) (1990) *Selected Papers on Laser Damage in Optical Material*. SPIE Milestone Series, vol. MS24. SPIE Optical Engineering Press.

Wood RM (ed.) (2003) *Laser-Induced Damage in Optical Materials*. Bristol, UK: Institute of Physics.

Wood RM (ed.) (2003) *The Power- and Energy-Handling Capability of Optical Materials, Components, and Systems*. SPIE Press Tutorial Texts in Optical Engineering, vol. TT60.

LIGHT EMITTING DIODES

J Schanda, University of Veszprém, Veszprém, Hungary

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Short Historic Overview

During the second half of the nineteenth century, electric lighting became possible. In 1879, Edison demonstrated his incandescent lamp, and although some electric arc lamps competed with the incandescent lamp bulb for a few years, the first 50 years of electrical lighting was dominated by the incandescent lamp. Gas discharge lamps (both low- and high-pressure lamps) became more widely distributed only by the end of the third decade of the twentieth century, and widespread distribution of the fluorescent lamp, that still dominates our offices and many other lighting applications today, was generally accepted only after the Second World War.

The early incandescent lamps had an efficacy of a few lm/W, but by 1920 this could be increased by up to 10–15 lm/W. Gas discharge lamps started with approximately 30 lm/W, but by 1960 their efficacy increased to about 70 lm/W. Fluorescent lamps still provide only 70 to 100 lm/W; high-pressure lamp efficacy could be increased above this value. The different lamp types can be used, however, in different applications, depending on the size and the unit power of the lamp type. Nonincandescent light generation in solid-state materials dates back to the 1920s: a faint glow of a SiC crystal was observed when current passed through it at a point contact. In the 1930s, electroluminescence of ZnS powder layers was detected (Destriaux effect). By the mid-1960s, it became clear that by using the then available technology, good luminous efficiency could be expected only in III–V compounds (GaAs, GaP, and their ternary and quaternary compounds using Al, In, and N in the compounds). Pure GaAs has a bandgap in the infrared (IR), thus one can build good IR-emitters using this material, but no visible light emitter. However, GaAs proved to be an excellent coherent light-emitting material and enabled the construction of IR-emitting semiconductor lasers.

First visible light-emitting diodes were fabricated using a GaAsP alloy. Figure 1 shows how the efficacy of III–V compound LEDs can be increased by introducing new compositions, and later, new technologies and better light extraction techniques. In the figure we also show the efficacy values of average incandescent, high-pressure Hg- and fluorescent lamps. The first LEDs emitted in the red part of the spectrum, and only by 1975 was the green LED invented. By 1985, practically any color between red and green became possible, and the efficacy of these lamps reached a level where for monochromatic light applications, their efficacy could compete favorably with filtered incandescent light. LEDs emit in a relatively narrow wavelength band (20 nm–30 nm bandwidth), and if one has to filter from the continuous spectrum of an incandescent lamp such a narrowband, the loss becomes considerable. LEDs needed only a few volts to function and it became general practice to build LEDs with approximately a 20 mA current load. These LEDs became very popular for interior signaling (on instruments, household appliances, and in similar applications). The goal was, however, to be able to produce white light, but the blue-emitting LED was still missing. Only after 1990 could the problem of producing p–n junctions from GaN, that had a bandgap large enough to emit blue light, be solved. Different ternary and quaternary compounds, as shown in Figure 1, enabled efficacies comparable to incandescent lamp efficacy, and the race for producing white light began.

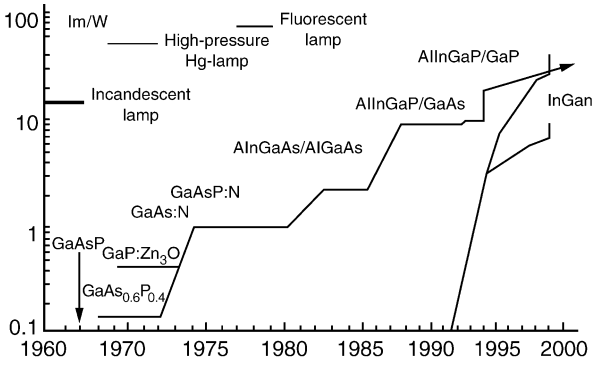


Figure 1 Increase of luminous efficacy of LEDs during the past 40 years. Courtesy of Lumileds Lighting.

Physical Fundamentals

Figure 2 shows a forward-biased p–n junction. As injected free electrons and holes are driven towards the junction, they can radiatively recombine, emitting a photon. The probability of photon emission versus nonradiative recombination will depend on the band

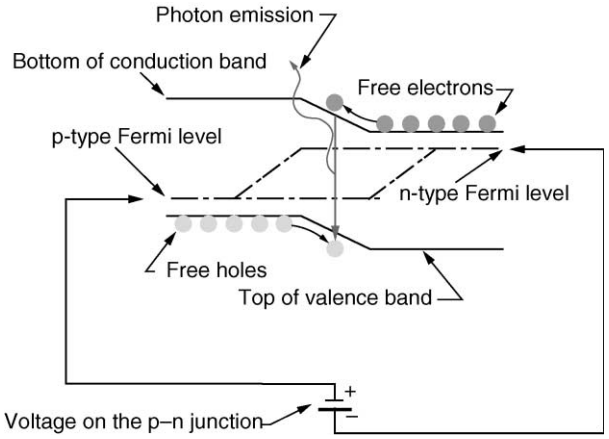


Figure 2 Forward biased p–n junction, light generation via recombination of free charge carriers.

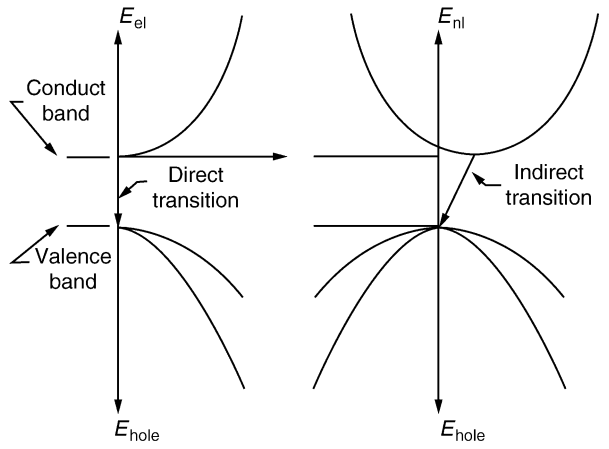


Figure 3 Band to band recombination in direct and indirect band semiconductor.

structure of the semiconductor. Figure 3 shows the two possibilities of a direct and an indirect bandgap semiconductor. GaAs is, for example, a direct bandgap material. In this material free electrons at the bottom of the conduction band can recombine with free holes at the top of the valence band, as the momentum of the two 'particles' is the same and a photon emission takes care of the principle of energy conservation. GaP is an indirect bandgap material where the momentum of the free electrons at the bottom of the conduction band differs from the momentum of the free holes at the top of the valence band, and thus the recombination needs not only the fulfilment of the conservation of energy, but also of momentum. Thus, besides photon emission, phonon emission or absorption has to take place. Thus, the probability of the transition will decrease and competing nonradiative processes will become important. Using N doping produces very shallow

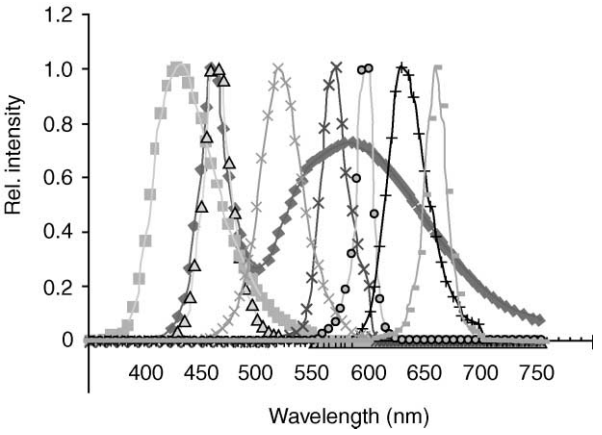


Figure 4 Relative spectral power distribution of some LED structures.

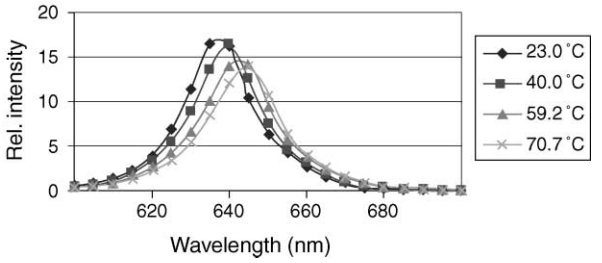


Figure 5 Spectral power distribution of a red LED at different temperatures.

donor states and from these an efficient recombination can take place. This enabled the production of early relatively efficient green-emitting LEDs.

GaN is ideal for producing LEDs, as it is a direct gap material and has a high bandgap. The technological difficulties are, however, enormous. Modern high-efficiency LEDs use (Al)InGaN-based materials for shorter wavelength LEDs, and AlInGaP-based materials for longer wavelength LEDs. Figure 4 shows a number of LED spectra of different wavelengths including a white LED. Its spectrum consists of a blue maximum produced by the LED emission and a yellow broader band produced by the luminescent material excited by the blue emission of the LED.

LEDs are semiconductor materials, and as with every semiconductor the LED characteristics are also temperature dependent. Not only the efficiency of the current carrier–photon emission process depends on the temperature, but the bandgap of the base material also depends on the temperature, and thus both the efficacy and the relative spectral distribution of the emission change with temperature. Figure 5 shows the spectral power distribution of a red LED at different temperatures.

Modern High Brightness LEDs

Besides the high quantum efficiency of the semiconductor junction, two further requirements to achieve high luminous efficacy are the efficient light extraction from the junction and keeping the temperature of the semiconductor chip low.

All semiconductors for producing LEDs have a high refractive index, thus light reflection at the semiconductor–air interface is relatively high, and for obliquely impinging rays the critical angle for total internal reflection is soon reached. Different manufacturers shaped the semiconductor chip in different forms to secure maximum light extraction. Figure 6 shows the cross-section of an LED, where the side planes have been cut obliquely to secure good light extraction.

The other problem that has to be solved is good heat conduction from the chip to a heat sink. At the time of writing, highest single chip power loads were in the order of 5 W. The structure of such a construction is seen in Figure 7. The GaN-based LED is grown on a visible light transparent sapphire (Al_2O_3) substrate. In modern LEDs the structure itself is quite complicated; in the vicinity of the p–n junction some nanometer thick layers are deposited forming quantum wells (QW(s)). The p-type original upper layer is metalized and soldered to the submount serving as heat-sink. The n-type layer gets a second soldered heat conducting current leading to the submount. As seen, the original structure grown on the sapphire substrate is inverted, and now the light can be extracted through the sapphire layer. The metalization on the p-type layer serves at the same time as a mirror. High brightness LEDs are now available in package sizes of over 100 lumen.

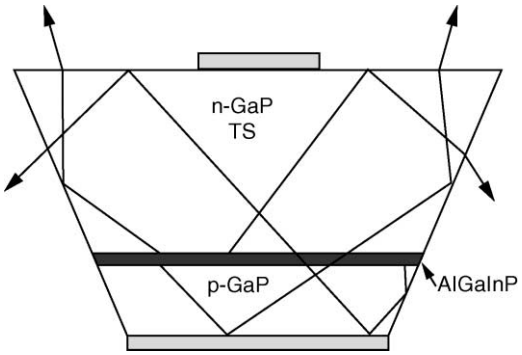


Figure 6 Truncated pyramid structure to secure good light extraction from the LED chip. Courtesy of Lumileds Lighting.

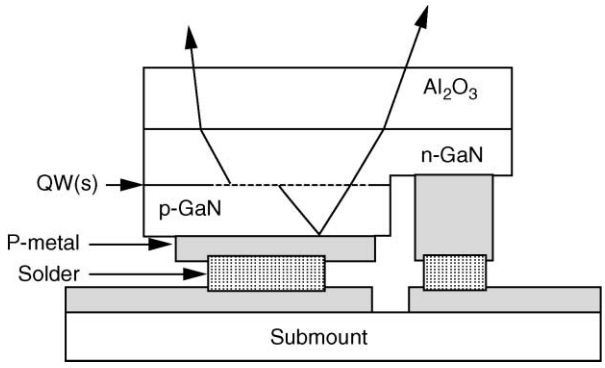


Figure 7 Inverted, flip-chip structure for good heat-sinking of the LED. Courtesy of Lumileds Lighting.

White LED Structures

The advantages of LEDs (low drive voltage, extreme long lifetime of 50000 to 100000 hours, robust shock-resistant construction) can open up a new era of lighting engineering, if white light emission can be obtained. The LED structure in itself is a narrowband emitter, thus – with inorganic LEDs – it is not possible to produce white light within one chip. Organic LEDs (OLEDs) promise the production of broadband and white light emission, but these are still in an experimental phase. At present they are used in smaller, portable displays, such as those used on portable phones, cam-coders, etc. To get further information, go to the Further Reading section at the end of this article.

At present three techniques to generate white light get thorough investigation: first, mixing the light of a red, a green, and a blue LED. This should theoretically give the highest efficiency; the problem is, however, that the three LEDs have different temperature dependence and aging characteristics, and thus it is difficult to keep the chromaticity of the mixed color constant (the three colored LEDs can be used in LED-based displays to produce full-color reproduction). At present the following technique is widely used. A blue LED excites a yellow phosphor, so that the mixture of the blue and yellow lights produce a white appearance. The color rendering of such white LEDs is not too good, and therefore alternative techniques using a green and a red phosphor are receiving increasing attention. The third method uses the same principle, but with an LED emitting in the near ultraviolet. This has the advantage that it provides a greater selection of phosphors, but as the energy difference between the UV photon and the photon emitted by the phosphor is larger than in the case of a blue-emitting LED, the theoretical efficiency is lower.

The phosphor layer is deposited on the chip surface, and for blue-emitting diodes special care has to be taken that the LED light and the phosphor

light are properly mixed so that the emitted light shows no directional color shift. Early white LEDs suffered much from the fact that the spectral distribution of the light changed considerably with viewing angle.

The problems with phosphor-coated white LEDs are how a suitable correlated color temperature and color rendering can be obtained. Early white LEDs had very bluish chromaticity, their correlated color temperature was very high, above 7000 K. Preferred white light has different chromaticity in different cultural regions, in the Far East 6000 K to 6500 K is preferred, in Europe and North America for general lighting a correlated color temperature of 4000 K is well accepted and for social and home lighting even lower correlated color temperatures, 2800 K to 3200 K are preferred. Recently, white LEDs based on blue chips and special phosphor blends have been produced that meet these requirements.

A further problem is the color rendering of the white LEDs. The Commission Internationale d'Eclairage (CIE), the International Commission on Illumination, provides standards, recommendations, and technical reports in the field of light and lighting, including vision, color, and image processing. Photometry and colorimetry are based on the standards and recommendations of the CIE.

The CIE established a method of calculating a color-rendering index in 1965 and updated it in 1974. Despite several attempts, the lighting community has been unable to agree on a revision of this outdated document since the 1970s. Thus, color-rendering indices are still calculated based on this method, although several investigations have shown that the calculated indices do not correspond with the visual observation. CIE, therefore, established a technical committee (TC) dealing only with the question of LED color rendering. It is hoped that this TC will be able to come up with some recommendations soon and, therefore, that LED manufacturers can tailor the emission spectra of their LEDs for optimal efficacy and color rendering. Until this happens, however, the published color rendering indices, especially the General Color Rendering Index (R_a) have to be handled carefully. The color appearance of objects under LED illumination might deviate considerably from what one expects based on the R_a index.

Application of LEDs

LEDs are the solution for practically every signaling application. Figure 8 shows an LED traffic light. Here it is also shown that, despite the fact that LEDs have extremely long life expectancy, the failure of the electronic circuitry can lead to unpleasant dark spots and rows in the display.

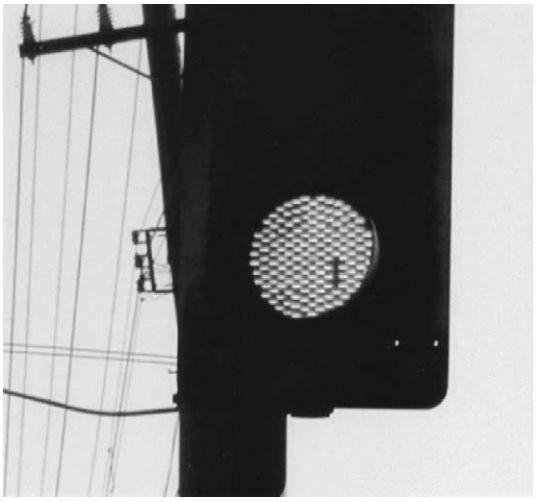


Figure 8 Green LED traffic light. Due to an electronic failure some LEDs stay dark.

LEDs are increasingly used in large active posters and displays. In such applications the highly saturated color of LED light is an advantage: LED displays can be very bright and emit saturated colors.

LEDs have an advantage in every situation where colored light has to be produced. Thus, they become the preferred automotive stoplights, dashboard lights, and are widely used in other signaling applications on the roads, at airfields, and in other traffic situations; much decorative lighting is now using LEDs.

Experiments are underway to find the best applications for white LEDs. Their drawback at the moment is still that their single unit power is below ten watts, as we often need ten to a hundred times larger units. For such applications, some LED clusters have been built, but these are still not competitive with other light sources. At places where the ruggedness of the LEDs is of an advantage, white LED lights are used – for example, in aircraft reading lights. One area where we will probably soon find higher-intensity white LEDs is car headlamps. Both the efficacy and the luminous flux output of the single LED units is increasing continuously. Applications, unbelievable only a few years ago, can now be realized efficiently with LEDs. Thus it is very difficult to foresee in which applications – beyond signaling – LEDs will first supersede other light sources, but it is certain that within the next decade we will see many new applications of LED sources.

List of Units and Nomenclature

Chromatic adaptation	Adaptation by stimuli in which the dominant effect is that of different relative spectral distributions.
Chromaticity	Partial description of the color stimulus, describing only chromatic aspects and not the 'intensity' (the luminance) of the stimulus. In light sources, chromaticity is an important descriptor, as the level of illumination can be changed by changing the distance between the source and the illuminated object. The luminance of the object will depend on the illumination and the reflection (transmission) characteristics of the object.
Color	In color science one distinguishes between color (perception), characteristic of visual perception that can be described by attributes of hue, brightness (or lightness) and colorfulness (or saturation or chroma), and color stimulus, a specification of the optical stimulus in terms of operationally defined values, such as three tristimulus values, or luminance, and chromaticity (or dominant wavelength and saturation).
Color rendering	Measure of the degree to which the psychophysical color of an object illuminated by the test illuminant conforms to that of the same object illuminated by the reference illuminant, suitable allowance having been made for the state of chromatic adaptation.
Correlated color temperature	The temperature of the Planckian radiator whose perceived color most closely resembles that of a given stimulus at the same brightness and under specified viewing conditions.
Efficacy (luminous efficacy of a source)	Quotient of the luminous flux emitted by the power consumed by the source.
Luminous efficiency	Ratio of radiant flux weighted according to $V(\lambda)$, the spectral luminous efficiency of the human eye, to the corresponding radiant flux.
Spectral luminous efficiency (of the human observer)	Ratio of the radiant flux at wavelength λ_m to that at wavelength λ, such that both radiations produce equally intense luminous sensations under specified photometric conditions and λ_m is chosen so that the maximum value of this ratio is equal to 1.

See also

Incoherent Sources: Lamps. **Lasers:** Semiconductor Lasers.

Further Reading

Fukuda M (1991) *Reliability and Degradation of Semiconductor Lasers and LEDs*. Boston, MA: Artech House.

Rudden MN and Wilson J (1993) *Elements of Solid State Physics*. Chichester, UK: Wiley.

Sze SM (1985) *Semiconductor Devices: Physics and Technology*. New York: Wiley.

Wenckebach WT (1999) *Essentials of Semiconductor Physics*. Chichester, UK: Wiley.

Žukauskas A, Shur MS and Gaska R (2002) *Introduction to Solid-State Lighting*. New York: Wiley.

M

MAGNETO-OPTICS

Contents

Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping
Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering

Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping

H Pascher, Universität Bayreuth, Bayreuth, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A complete understanding of the band structure of a semiconductor close to the Fermi energy requires information on the fundamental energy gap, the effective masses and the effective g-factors of the valence and conduction band. Magneto-optical methods in principle allow one to determine all these quantities; they are particularly well suited to measure the spin splittings of the energy bands.

Faraday rotation is the rotation of the plane of polarization of linearly polarized light propagating through a sample parallel to an external magnetic field. The magneto-optical Kerr effect is the analog for reflection. Though in general it yields much smaller angles of rotation than the Faraday effect, it has to be applied in spectral ranges where the investigated sample is not transparent.

If a semiconductor is excited by circularly polarized light with photon energies not too high above the energy gap a spin polarization in the conduction band can be obtained (optical pumping). A small transverse magnetic field reduces the degree of this spin polarization (the Hanle effect). Owing to the spin polarization of the photo-excited conduction electrons the photoluminescence radiation is partly circularly polarized. Thus mounting the sample in a microwave resonator and applying a longitudinal magnetic field, electron spin resonance transitions can be detected optically by monitoring the degree of polarization of the photoluminescence (ODCESR: optically detected conduction electron spin resonance). Via a 'nuclear Hanle effect' mediated by the hyperfine interaction, nuclear magnetic resonance transitions can also be detected optically (ODNMR). ODCESR and ODNMR are summarized under ODMR (optically detected magnetic resonance).

In modern optics nonlinear optical effects have gained particular importance. Since electron spin resonances in semiconductors are Raman-active transitions, coherent Raman methods such as CARS (coherent anti-Stokes Raman scattering) or CSRS (coherent Stokes Raman scattering) are of great importance for the determination of spin splittings of energy bands and the investigation of magnetic structures in semiconductors.

In the present contribution the physical basis for these methods will be discussed and typical examples for their application will be presented.

For Faraday rotation and CARS the systems from which these examples will be taken are magnetic or dilute magnetic semiconductors. In these materials paramagnetic ions like Mn^{++} or Eu^{++} are incorporated. There are two types of exchange interactions: (i) between the magnetic ions which in most cases causes an antiferromagnetic coupling; and (ii) between the ions and free carriers which causes a

strong modification of the spin splittings of the energy bands with respect to nonmagnetic materials. Thus the Faraday effect or CARS are particularly well-suited methods for the investigation of such systems.

As examples of optical pumping and ODMR, results obtained with III–V semiconductors will be presented, which show the most important applications of these experimental methods.

Faraday Rotation and Magneto-optical Kerr Effect

Basic Principles

As already pointed out in the Introduction, the Faraday effect is the rotation of the plane of polarization of linearly polarized light traversing a sample parallel to a magnetic field. The angle of rotation Θ_{Far} is proportional to the length l of the sample and to the strength B of the magnetic field:

$$\Theta_{\mathrm{Far}} = VlB \qquad [1]$$

where V is called Verdet's constant.

For a phenomenological understanding of the effect one can consider plane polarized light as a superposition of right- and left-hand circularly polarized components. If the indices of refraction are different for both components, then at the end of the sample, when the two components are recombined, a phase shift occurs which manifests itself as a rotation of the polarization plane. The angle resulting from this consideration is

$$\Theta_{\mathrm{Far}} = \frac{l\omega}{2c}(n_- - n_+) \qquad [2]$$

where ω is the frequency, c the velocity of light, and n_- and n_+ are the refractive indices for left- and right-hand polarized light, respectively. If one neglects ellipticity of the light behind the sample only the real parts of the complex refractive indices $\tilde{n}_\pm$ are important,

$$\tilde{n}_\pm = n_\pm - \mathrm{i}\kappa_\pm \qquad [3]$$

The imaginary parts, however, are needed to understand the magneto-optical Kerr effect which occurs in the light reflected from the sample surface. According to Fresnel's formula for vertical incidence the field amplitude of the reflected beam is:

$$\tilde{E}^{\mathrm{r}}_\pm = \frac{1-\tilde{n}_\pm}{1+\tilde{n}_\pm}\tilde{E}^{\mathrm{i}}_\pm = \tilde{r}_\pm \tilde{E}^{\mathrm{i}}_\pm \qquad [4]$$

where the superscript 'r' refers to the reflected beam and 'i' to the incident beam. The complex reflection coefficient can be split into an amplitude and a phase factor:

$$\tilde{r}_\pm = r_\pm \exp(\mathrm{i}\psi_\pm) \qquad [5]$$

By comparing eqns [4] and [5] the phase shifts during reflection can be expressed in terms of the refractive index:

$$\psi_\pm = \arctan\frac{2\kappa_\pm}{1-n_\pm^2-\kappa_\pm^2} \qquad [6]$$

Finally the Kerr angle of rotation is obtained from the phase shifts in an analogous way to the Faraday rotation:

$$\Theta_{\mathrm{Kerr}} = \tfrac{1}{2}(\psi_- - \psi_+) \qquad [7]$$

It is seen that both effects can be understood in terms of the refractive index. In a simple oscillator model the refractive index has the form given in eqn [8]:

$$\tilde{n}^2_\pm = \tilde{\varepsilon}_\pm(\omega,B) = \varepsilon_\infty + \sum_{i,j}\frac{F^\pm_{i,j}}{\hbar\omega^\pm_{i,j} - \hbar\omega - \frac{1}{2}\mathrm{i}\Gamma^\pm_{i,j}} \qquad [8]$$

with $\hbar\omega^\pm_{i,j}$ representing the transition energies between the energy levels i and j in σ^+ and σ^- polarization, respectively, and $F^\pm_{i,j}$ and $\Gamma^\pm_{i,j}$ are the respective oscillator strength and linewidth parameter.

From eqn [8] it is seen that when tuning the frequency, ω, resonant enhancements of the rotation angles are expected whenever the frequency corresponds to an allowed electronic transition, e.g., an exciton in a wide-gap semiconductor. This is the reason why the Faraday and magneto-optical Kerr effect can yield information on interband transition energies. In the case of such a resonance, the refractive index will be dominated by the particular term in the sum, corresponding to that transition. By a Taylor expansion one can then show that the difference of the refractive indices for σ^+ and σ^- polarization and thus the angles of rotation (see eqns [2] and [7]) are proportional to the difference of the transition energies for right- and left-hand circularly polarized light, and this difference depends on the spin splittings of the valence and conduction band. So we understand why the most important information obtained from Faraday or magneto-optical Kerr effect studies is on the spin splittings of the energy bands.

A quantum mechanical description of the relation between the material properties and the magneto-optical effects beyond the simple oscillator model given here is published by Boswarva *et al.* and the reader is referred to this paper for

further studies. A matrix formalism to describe the effects for a multilayer system has been developed by Nies and Kessler.

Examples of Experimental Results

The straightforward method to measure Faraday rotation is to put the sample between polarizers and sweep the magnetic field. Then the transmitted intensity is proportional to a $\cos^2$ function. Subsequent maxima demonstrate a rotation by 180°. There have been numerous successful investigations done in this way, but the possibilities of detecting small angles of rotation are very limited in this method. Therefore differential methods have been developed either using a polarizing beam splitter, or a photo-elastic modulator which periodically switches the state of polarization. Rotation angles of the order of 10^{-4} rad can be detected in this way.

Figure 1 shows the Faraday rotation of an epitaxial film of $Pb_{0.933}Eu_{0.067}Te$ (thickness 11μm) for two temperatures. The plot demonstrates behavior typical for a diluted magnetic semiconductor in its paramagnetic phase. The spin splittings of the valence and conduction band and accordingly the Faraday rotation are proportional to the magnetization of the sample which behaves like a Brillouin function. (The lines without noise are fitted Brillouin functions.)

Examples of the application of the magneto-optical Kerr effect to the investigation of semiconductors, particularly multilayer systems, are published, e.g., by Pohlt *et al.* and Krenn *et al.*

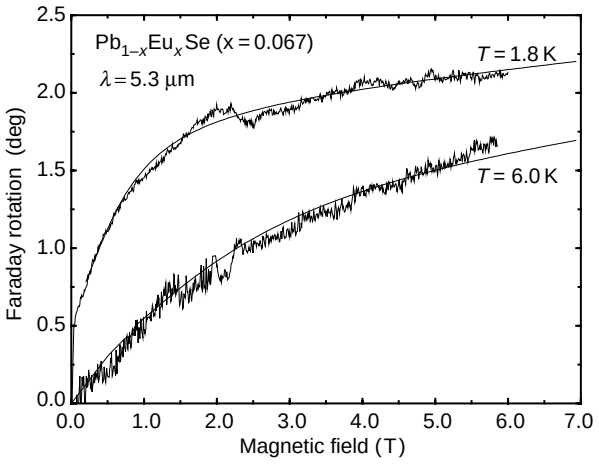


Figure 1 Faraday rotation as a function of magnetic field of a PbEuSe epitaxial film for two temperatures. Lines with noise: experimental recordings; lines without noise: fitted Brillouin functions.

Coherent Raman Scattering

Basic Principles

A more direct method for the determination of spin splittings of energy bands is provided by Raman scattering. There is a huge number of investigations of spin transitions in semiconductors using spontaneous Raman scattering (for a review of spontaneous Raman scattering in dilute magnetic semiconductors see the article by Ramdas and Rodriguez). However, there are several advantages in the spectroscopic application of the nonlinear Raman techniques compared to spontaneous Raman scattering:

(i) The scattered radiation is emitted as a collimated beam which is more easily detectable than the light spontaneously scattered into the entire solid angle. This advantage is particularly important for the investigation of narrow-gap semiconductors which have to be investigated by infrared light, a spectral range where the detectors are less sensitive than in the visible.
(ii) Fluorescence can be strongly suppressed by spatial filtering.
(iii) The spectral resolution is determined by lasers instead of a classical spectrometer.
(iv) Raman scattering cross-sections can be calculated from the intensity of a collimated beam without taking care of the aperture of the detecting system.

The coherent Raman techniques are special cases of optical four-wave mixing where two laser beams with frequencies ω_L and ω_S ($\omega_L > \omega_S$) are superimposed in the sample. Due to nonlinear interactions radiation with new frequencies is generated, among them:

$$\omega_{CARS} = 2\omega_L - \omega_S \qquad [9]$$

If one identifies ω_L with the frequency of the pump laser and ω_S with that of the scattered Stokes radiation of a spontaneous Raman experiment, then ω_{CARS} is the frequency of the anti-Stokes radiation. For the intensity of the radiation generated at ω_{CARS} one finds:

$$I_{CARS} = \frac{9}{16c_0^4\varepsilon_0^2}\frac{\omega_{CARS}^2}{n_{CARS}n_S n_L^2}\left|\chi^{(3)}\right|^2 l^2 I_S I_L^2 \times \left[\frac{\sin(\overrightarrow{\Delta k}\,\vec{l}/2)}{\overrightarrow{\Delta k}\,\vec{l}/2}\right]^2 \qquad [10]$$

where c_0 is the velocity of light, l is the length of the sample, I denotes the intensity, and n the refractive index at the respective frequency. $\chi^{(3)}_{kjkj}$ is the third-order nonlinear susceptibility of the scattering medium and

$$\overrightarrow{\Delta k} = 2\vec{k}_{\mathrm{L}} - \vec{k}_{\mathrm{S}} - \vec{k}_{\mathrm{CARS}} \qquad [11]$$

is the phase mismatch. $\overrightarrow{\Delta k} = 0$ can be achieved by noncollinear alignment of the laser beams. However, for very thin epitaxial layers with $l \ll \pi/\Delta k$ the phase factor in eqn [10] can be approximated by 1.

The information on the investigated material is contained in the nonlinear susceptibility. It can be calculated by time-dependent perturbation theory taking into account a ground state $|a>$ and intermediate states $|b>$, $|c>$ and $|d>$. It turns out that $\chi^{(3)}$ is proportional to a sum over all states of the system as intermediate states; in the sum 24 terms like the one given in eqn [12] have to be added:

$$\frac{\mu_{ijkl}}{(\omega_{ba} - \omega_{\mathrm{L}} - \mathrm{i}\Gamma_{ba})(\omega_{ca} - \omega_{\mathrm{L}} + \omega_{\mathrm{S}} - \mathrm{i}\Gamma_{ca})(\omega_{da} - \omega_{\mathrm{CARS}} - \mathrm{i}\Gamma_{da})} \qquad [12]$$

μ_{ijkl} is a product of four dipole matrix elements between states $|a>:|b>$; $|b>:|c>$; $|c>:|d>$ and $|d>:|a>$ with field polarizations as given by the indices. $\omega_{\alpha\beta} = (E_\alpha - E_\beta)/\hbar$ and $\Gamma_{\alpha\beta}$ is the linewidth of the respective transition. Resonances of the susceptibility are found by inspection of the denominators. $\chi^{(3)}$ exhibits Raman-like resonances whenever $\hbar(\omega_{\mathrm{L}} - \omega_{\mathrm{S}})$ is equal to the excitation energy of a Raman allowed transition (see the second factor in the denominator of eqn [12]). If in this case the laser beams are polarized according to the Raman selection rules then μ_{ijkl} is not zero. For spin transitions in a semiconductor in a magnetic field (free electron spin flip as well as transitions within the magnetic ions) this is the case if in the Voigt configuration ($\vec{k}_{\mathrm{L,S}} \perp \vec{B}$) the beams with frequencies ω_{L} and ω_{S} are polarized perpendicular to one another. In the case of a spin transition $\omega_{ca} = g^* \mu_{\mathrm{B}} B$ (g^* being the effective Landé-factor and μ_{B} the Bohr magneton). Thus holding the frequencies ω_{L} and ω_{S} fixed and tuning the magnetic field a resonance in I_{CARS} will be observed from the position of which the effective g-factor can be obtained. Using tunable dye lasers, of course, the laser frequencies can be tuned as well.

Close to the resonance the real part of the susceptibility becomes small and the imaginary part is not negligible. Furthermore, one-photon resonances may occur if the laser frequency ω_{L} corresponds to the energy of a real electronic transition, e.g., an allowed interband transition. If one-photon and Raman-like resonances occur simultaneously one talks about resonance CARS.

As discussed in the literature $\chi^{(3)}$ can be split into a resonant part χ_{r}, summing up all the resonant terms, and a nonresonant part χ_{nr}. In a semiconductor there are contributions to χ_{nr} due to free carriers and a background due to bound valence electrons. The background contribution in general is real. As can be seen from eqn [10],

$$I_{\mathrm{CARS}} \propto \left|\chi^{(3)}_{\mathrm{CARS}}\right|^2 = |\chi_{\mathrm{nr}}|^2 + |\chi'|^2 + |\chi''|^2 + 2 \cdot \chi_{\mathrm{nr}} \chi' \qquad [13]$$

χ' and χ'' being the real and imaginary parts of the resonant contribution. Depending on the relative strength of resonant and nonresonant contributions to $\chi^{(3)}$ the last term in eqn [13] may cause complicated lineshapes which make it difficult to find the correct transition energy in spectroscopic applications of CARS. If more than one resonance is not completely resolved the lineshape is even more complicated and a careful analysis has to be performed to get the correct resonance positions and linewidths of the transitions.

Examples of Experimental Results

Figure 2 shows a scheme of a CARS setup suitable to measure spin transitions in a semiconductor. The scheme is drawn for backscattering geometry; forward detection is accomplished by directing the incoming beams onto the sample from the opposite side to that plotted.

Two dye lasers (frequencies ω_{L} and ω_{S}) are pumped by a frequency-doubled Nd:YAG laser. A Glan–Thomson prism combines the two beams. The monochromator is set to $2\omega_{\mathrm{L}} - \omega_{\mathrm{S}}$. The iris and the polarizer PZ3 reduce stray light. The experiment is performed either at fixed magnetic field and tuning one of the dye lasers (with the monochromator swept accordingly) or by keeping the laser frequencies fixed and sweeping the magnetic field. The latter method is also applicable in the mid-infrared using step-tunable CO or CO_2 lasers.

As an example, in Figure 3 a spectrum obtained from a CdMnTe alloy crystal is reproduced. There are two peaks observed due to the paramagnetic resonance of the Mn 3d electrons and due to the spin resonance of photo-excited free electrons in the conduction band. The inset shows the paramagnetic resonance on an extended scale. The linewidth of 0.5 cm^{-1} is not limited by the experimental resolution, which in the present case is about a factor of 10 better.

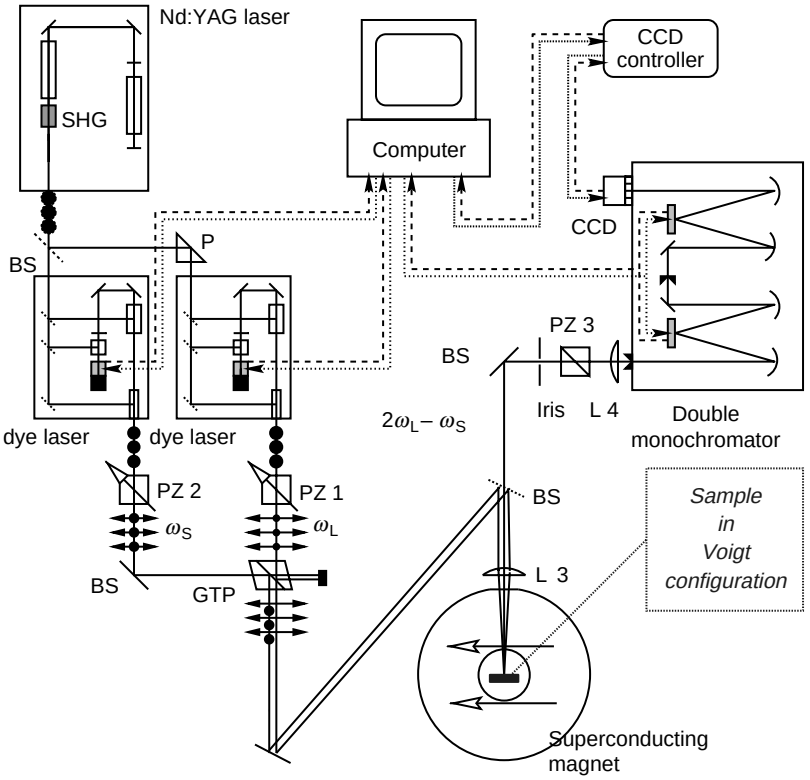


Figure 2 Scheme of the experimental arrangement for CARS. Bold full lines: light beams; dashed and dotted lines: data connections; BS: beam splitter; PZ1,2: Pockels cells for rotation of linear polarization; PZ3: polarizer; GTP: Glan–Thomson prism for superposition of beams polarized orthogonally to each other; L: lenses.

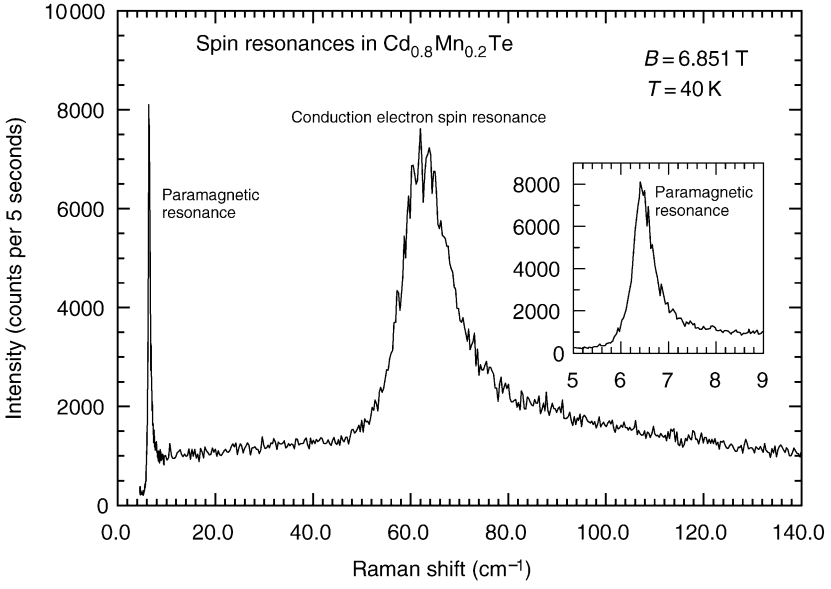


Figure 3 Coherent Raman intensity of a CdMnTe bulk crystal as a function of frequency difference $\omega_L - \omega_S$. Inset: Region of paramagnetic resonance in an extended scale. Adapted from Rupprecht R, Müller B, Pascher H, *et al.* (1998) Coherent Raman scattering from magnetic excitations in diluted magnetic semiconductors: Bulk crystals of CdMnTe. *Physical Review B* 58: 16123. Copyright (1998) by the American Physical Society.

Optical Pumping

The basic principles of optical pumping have been developed by Kastler. A material (gas or solid) is excited by absorption of circularly polarized light. Lampel was the first to demonstrate the possibility of optical spin orientation of free electrons in a semiconductor. The method has the most important applications in zincblende materials. Thus it is explained here with the example of GaAs.

Principles of Optical Spin Orientation

Most of the III–V and II–VI semiconductors have direct gaps at the center of the Brillouin zone. The conduction band (cb) is doubly degenerate with spin. The valence band consists of the heavy hole band (hh; $m_j = \pm 3/2$), the light hole band (lh; $m_j = \pm 1/2$) and the split-off band (so; $m_j = \pm 1/2$), each doubly degenerate. In bulk material hh and lh are degenerate at $k = 0$. The transition probabilities for interband transitions between hh and cb are three times larger than for transitions between lh and cb. Due to the selection rules, e.g., for the absorption of σ^+ circularly polarized light, transitions $-3/2 \rightarrow -1/2$ and $-1/2 \rightarrow +1/2$ are allowed, the first having three times higher probability. The electronic spin polarization P is defined as:

$$P = \frac{n_+ - n_-}{n_+ + n_-} \qquad [14]$$

where n_+ and n_- are the densities of electrons with the two spin orientations with respect to the direction of propagation of light. These considerations yield $P_\mathrm{i} = -0.5$ as the initial spin polarization under excitation with σ^+ circularly polarized light with a photon energy between the fundamental gap and the energy difference from the spin–orbit split valence band to the conduction band. Due to relaxation processes the actual electronic steady state polarization P is smaller than P_i:

$$P = P_\mathrm{i} \frac{T_1}{T_1 + \tau} \qquad [15]$$

with T_1 being the spin relaxation time, which in zero magnetic field is equal to the transverse relaxation time T_2, and τ is is the lifetime of conduction electrons.

The spin polarization of the electrons can be detected by analyzing the degree of circular polarization of the luminescence light which is defined as:

$$\Pi = \frac{I_+ - I_-}{I_+ + I_-} \qquad [16]$$

where I_+ and I_- are the intensities of σ^+ and σ^- polarized light, respectively. Since for emission due to free-electron/free-hole or free-exciton recombination the same transition probabilities as for absorption are obtained, one finds:

$$\Pi = 0.5\, P = 0.25 \frac{T_1}{T_1 + \tau} \qquad [17]$$

In the case of quantum wells, the lh-hh degeneracy is lifted due to different confinement energies. Then the maximum polarization of cb $\rightarrow$ hh as well as for cb $\rightarrow$ lh luminescence light is 0.5.

Of course these considerations are only true for intrinsic or p-type materials where the number of photo-excited electrons is high compared to the equilibrium number of electrons.

Hyperfine Interaction of Spin-Polarized Electrons and Atomic Nuclei

In GaAs three isotopes exist with spin 3/2, namely ^{69}Ga, ^{71}Ga and ^{75}As. Owing to the small energy separation of the nuclear spin levels in moderate magnetic fields, in thermal equilibrium no significant nuclear spin polarization exists even at a temperature of 2 K. However, the hyperfine interaction between electrons and nuclei allows so-called flip-flop processes, that is a simultaneous reversal of electronic and nuclear spins. This causes a considerable dynamic spin polarization of the nuclei if the electronic spin system is polarized due to optical pumping. Under most experimental conditions and when nuclear spin relaxation mechanisms other than via the hyperfine interaction are negligible, the mean nuclear spin is approximately:

$$\langle I \rangle = \frac{I(I+1)}{S(S+1)} \frac{P}{2} \qquad [18]$$

where I and S are the operators for nuclear and electron spin. These dynamically polarized nuclei act on the electrons and shift their spin resonance magnetic field by an effective nuclear field (the Overhauser shift). This shift is proportional to $\langle I \rangle$ and to the probability of finding the electrons at the nuclear sites, and is inversely proportional to g^*. In GaAs due to the small value of g^* this field can reach several tesla.

Effect of a Constant Transverse Magnetic Field – The Hanle Effect

In a transverse (with respect to the direction of propagation of light) magnetic field the spin-oriented electrons precess around the field with an

angular frequency $\Omega = g^*\mu_B B/\hbar$. If the spin lifetime T_S ($1/T_S = 1/T_2 + 1/\tau$) is long enough, this precession diminishes the average spin polarization and consequently the degree of circular polarization of the photoluminescence according to:

$$\Pi(B) = \Pi(0)\frac{1}{1+\Omega^2 T_S^2} \qquad [19]$$

This is a Lorentzian with a width at half height of $\Delta B = \hbar/(g^*\mu_B T_S)$.

It is seen that by measuring $\Pi(B)$ the spin lifetime can be determined if g^* is known. Together with eqn [17] the spin and interband lifetime can be determined in a steady-state experiment. If, on the other hand, T_S is known from a time-resolved experiment, g^* can be found from a Hanle effect measurement.

Optically Detected Conduction Electron Spin Resonance (ODCESR)

Optical spin orientation of conduction electrons is the presupposition for ODCESR. Suppose now, a longitudinal magnetic field $\mathbf{B}_0$ is applied on the sample ($\mathbf{B}_0$ parallel to $\mathbf{k}$, the direction of propagation of light, which is parallel to the direction of the optically oriented spins). If the field is not too high the considerations of optical spin orientation remain unchanged. In higher fields the spin polarization may be modified due to the energy splitting of the spin levels. In this case the spin polarization is enhanced or diminished, depending on the direction of the field parallel or antiparallel to $\mathbf{k}$. If now an additional transverse microwave field B_1 of frequency ω_0, which fulfills the resonance condition $\hbar\omega_0 = g^*\mu_B B_0$, is applied, transitions between the two spin levels will occur which resonantly decrease the spin polarization. This decrease manifests itself in a decrease of the degree of circular polarization of the photoluminescence.

The method has several advantages compared to conventional ESR. Microwave photons with frequency of about 10^{10} Hz change the state of polarization of optical photons with frequency of about 10^{14} Hz. Thus the sensitivity is that of the optical detection which is much higher than in conventional ESR. Even electrons in a single quantum well or heterostructure can be detected by the optical method. There is also the possibility to distinguish between electrons in different quantized levels in heterostructures due to the different spectral position of their photoluminescence. Another advantage is that the electrons are photo-excited; the sample does not need to be n-doped. High-purity samples can be investigated which exhibit narrower linewidths.

Optically Detected Nuclear Magnetic Resonance (ODNMR)

With the presuppositions provided in the preceeding sections the basic principle of ODNMR is easily understood. As in ODCESR we assume that an external magnetic field $\mathbf{B}_0$ is applied parallel to $\mathbf{k}$. Perpendicular to it a radio-frequency field $\mathbf{B}_2$ with frequency ω is applied. The NMR resonance frequency is $\omega_0 = \gamma_n \cdot B_0$ with γ_n being the gyromagnetic ratio of the nucleus under consideration. In the frame rotating with ω around the direction of $\mathbf{B}_0$ an effective field $\mathbf{B}_{\text{eff}} = \mathbf{B}_0 - \boldsymbol{\omega}/\gamma_n + \mathbf{B}_2$ acts on the nuclei. The nuclear magnetization is oriented parallel to $\mathbf{B}_{\text{eff}}$. It is seen that in resonance $\mathbf{B}_{\text{eff}}$ is parallel to $\mathbf{B}_2$, which means perpendicular to $\mathbf{k}$. Via the hyperfine interaction this nuclear field acts on the electrons, diminishing their spin polarization via the Hanle effect. So the nuclear magnetic resonance can be detected as a reduction of the degree of circular polarization of the photoluminescence.

The simplified model given above does not take into account several effects which influence the lineshape, position and intensity of the observed resonance lines.

Examples of Experimental Results for Optical Pumping and Optically Detected Magnetic Resonances

These experiments can be performed at liquid helium temperature in a standard photoluminescence setup. In addition one has to provide the possibility of applying transverse (Hanle effect) or longitudinal (ODCESR, ODNMR) magnetic fields. For ODCESR the sample is mounted in a microwave resonator with optical access; for ODNMR a radio-frequency coil is wound around the sample. The state of circular polarization of the exciting laser light has to be controlled. The measured quantity is the sign and degree of circular polarization of the luminescence light at a certain wavelength. The sensitivity can be enhanced by periodically σ^+/σ^--modulating either the exciting light or the detection. Which one of these possibilities is used depends on whether or not a high nuclear spin polarization is wanted.

As pointed out above the dynamically polarized nuclei, particularly in GaAs, strongly affect the

ODCESR resonant field. To get reliable values for the effective g-factor, the nuclear spin polarization must be completely suppressed. This can be achieved by σ^+/σ^--modulation of the exciting light with a modulation period much shorter than the nuclear spin relaxation time. But due to imperfections, e.g., if the intensities of σ^+ and σ^- light or their time intervals are not equal, residual optical fields may remain. To get precise values for g^* in addition the NMR has to be saturated by a suitable radio-frequency field. On the other hand, the strong Overhauser shift of the CESR can be useful for the determination of nuclear spin relaxation times. To perform such experiments the nuclei are first polarized by optical pumping with fixed circular polarization and afterwards, for the detection, σ^+/σ^--modulation is switched on.

For the optical detection of NMR a high nuclear spin polarization is favorable. Thus for those experiments one uses fixed polarization of the exciting beam and modulates the detected light.

Figure 4 shows examples of results obtained with a p-channel AlGaAs/GaAs single heterostructure. In this sample due to the doping profile a triangular potential well at the interface in the valence band of the GaAs layer is built, in which quantized hole levels exist. The luminescence band shown in Figure 4a between 821 nm and 825 nm is due to recombination of electrons with holes in these quantized states. For the experiments shown in the other panels of the figure the spectrometer is set at a position within this band. The other lines are due to excitons and impurities and are not interesting for what follows.

In Figure 4b the results of a Hanle experiment are reproduced. The line without noise is the fitted Lorentzian. The fit yielded $\tau = 8.1$ ns and $T_1 = 4.7$ ns.

In Figure 4c a typical ODNMR spectrum of the same sample is plotted. One observes resonances due to ^{75}As at 1.36 MHz, ^{69}Ga at 1.9 MHz and ^{71}Ga at 2.4 MHz. In the inset the ^{75}As resonance is plotted in an extended scale. A quadrupole splitting into three components is clearly resolved. This result demonstrates that the tetrahedral symmetry at the nuclear site is broken either due to local defects or due to the (rather small) lattice mismatch close to the AlGaAs/GaAs interface.

Finally Figure 4d shows the ODCESR spectrum. For the experiment the exciting beam was σ^+/σ^--modulated and in addition during the sweep the NMR of all three nuclei was permanently saturated to make sure that no Overhauser shift

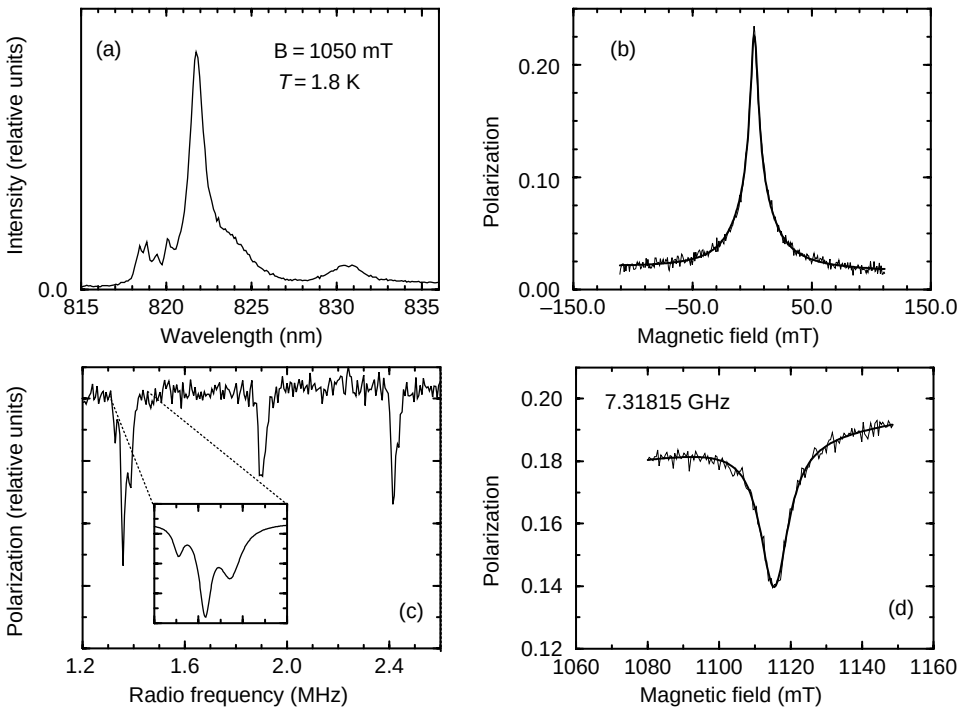


Figure 4 (a) Photoluminescence spectrum of a p-channel AlGaAs/GaAs heterostructure; (b) Hanle curve (luminescence polarization as a function of transverse magnetic field); (c) ODNMR recording; (d) ODCESR recording. Adapted from Schreiner M, Pascher H, Denninger G, *et al.* (1997) Nuclear spin relaxation in AlGaAs/GaAs heterostructures observed via optically detected magnetic resonance (ODMR) experiments. *Solid State Communications* 102: 715. Copyright (1997), with permission from Elsevier.

affects the result. The g-factor calculated from this plot is $|g^*| = 0.468$, which is larger than that found for bulk GaAs. This value was also used for fitting the Hanle curve.

List of Units and Nomenclature

Angle	[rad]
Angular frequency ω	[rad s^{-1}]
Bohr magneton μ_B	9.2742×10^{-24} J T^{-1}
Degree of circular polarization of light Π	
Electric field E	[V m^{-1}]
Gyromagnetic ratio γ	[T^{-1} s^{-1}]
Imaginary unit i	
Intensity I	[W m^{-2}]
Magnetic displacement B	[T]
Permittivity of free space ε_0	8.8542×10^{12} V s A^{-1} m^{-1}
Planck's constant $\hbar$	1.0546×10^{-34} J s
Refractive index n	
Speed of light in vacuum c_o	2.9979×10^{8} m s^{-1}
Spin polarization P	
Susceptibility χ	

CARS	coherent anti-Stokes Raman scattering
CSRS	coherent Stokes Raman scattering
ODCESR	optically detected conduction electron spin resonance
ODMR	optically detected magnetic resonance
ODNMR	optically detected nuclear magnetic resonance

Acknowledgments

The author thanks W. Herbst, B. Müller, R. Rupprecht and M. Schreiner for making unpublished results available.

See also

Magneto-Optics: Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering. **Semiconductor Physics:** Light Scattering; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Boswarva IM, Howard RE and Lidiard AB (1964) Faraday effect in semiconductors. *Proceedings of the Royal Society* A269: 125–141.

Dyakonov MI and Perel VI (1984) Theory of optical spin orientation of electrons and nuclei in semiconductors. In: Meier F and Zakharchenya BP (eds) *Optical Orientation*, pp. 11–71. Amsterdam: North-Holland.

Eesley GL (1981) *Coherent Raman Spectroscopy*. Oxford: Pergamon Press.

Furdyna JK (1988) Diluted magnetic semiconductors. *Journal of Applied Physics* 64: R29–R64.

Geist F and Pascher H (1996) Resonance CARS in narrow gap semiconductors. *Journal of Raman Spectroscopy* 27: 289–295.

Geist F, Herbst W, Mejia-Garcia C, *et al.* (1997) Magneto-optical investigations of Eu-based diluted magnetic lead chalcogenides. *Physical Review B* 56: 13042.

Hermann C and Weisbuch C (1984) Optical detection of conduction electron spin resonance in semiconductors and its application to $\mathbf{k}\cdot\mathbf{p}$ perturbation theory. In: Meier F and Zakharchenya BP (eds) *Optical Orientation*, pp. 463–508. Amsterdam: North-Holland.

Krenn H, Herbst W, Pascher H, *et al.* (1999) Interband Faraday and Kerr rotation and magnetization of $Pb_{1-x}Eu_xTe$ in the concentration range $0 \times \sim 1$. *Physical Review B* 60: 8117.

Levenson MD and Kano SS (1987) *Introduction to Nonlinear Laser Spectroscopy*. San Diego: Academic Press.

Nies R and Kessler FR (1989) Coherent and incoherent magnetooptical reflection and transmission effects of multilayer structures. *Phys. Stat. Sol. (a)* 111: 639–654.

Oh E, Bartholomew DU, Ramdas AK, Furdyna JK and Debeska U (1988) Interband Faraday effect in CdMnSe. *Physical Review B* 38(18): 13183–13190.

Paget D and Berkovits VL (1984) Optical investigation of hyperfine coupling between electronic and nuclear spins. In: Meier F and Zakharchenya BP (eds) *Optical Orientation*, pp. 404ff. Amsterdam: North-Holland.

Pascher H, Geist F, Springholz G, *et al.* (1997) Coherent anti-Stokes Raman scattering in diluted magnetic IV–VI epilayers and superlattices. In: Shen SC, Tang DY, Zheng GZ and Bauer G (eds) *Narrow Gap Semiconductors*, pp. 269–278. Singapore: World Scientific.

Pohlt M, Herbst W, Pascher H, Faschinger W and Bauer G (1998) Magneto-optical Kerr effect in short-period CdTe/MnTe superlattices. *Physical Review B* 57(16): 9988–9994.

Ramdas AK and Rodriguez S (1988) Raman scattering in diluted magnetic semiconductors. In: Furdyna JK and Kossut J (eds) *Diluted Magnetic Semiconductors, Semiconductors and Semimetals*, vol. 25, pp. 345–412. London: Academic Press.

Schreiner M, Pascher H, Denninger G, *et al.* (1997) Nuclear spin relaxation in AlGaAs/GaAs heterostructures observed via optically detected resonance (ODMR) experiments. *Solid State Communications* 102: 715–720.

Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering

C R Pidgeon, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The addition of a magnetic field to an optical study of a semiconductor generally improves the specificity of the experiment, but in particular affects primarily the electrons rather than the lattice so that the magneto-optical properties are determined to a large degree by the electrons alone. The major interest has been in the determination of the electronic energy band structure of semiconductors, both in their bulk and low-dimensional (two-, one-, and zero-dimensional) quantum forms. The magneto-optical experiments fall mainly into two categories according to whether the effects are associated with intraband (within a single conduction or valence band) or interband (valence to conduction band) transitions; there are also hybrid (i.e., higher-order) effects such as spin-flip Raman scattering or two-photon absorption where both inter- and intraband transitions are involved in a two-stage process. Interband transitions occur typically in the visible to mid-infrared (MIR) region, and intraband transitions in the MIR to far-infrared (FIR) region.

Theoretical Background

The most common band structure involved in semiconductor magneto-optics is the so-called quasi-Ge model, which gives an accurate description of group IV, and zincblende (III–V and II–VI) semiconductors, so it is convenient to describe the theory in terms of this model. In this model the small terms associated with inversion asymmetry are neglected to a first approximation. At the Brillouin zone center (i.e., at $k = 0$) the conduction band has s-like symmetry and the triply degenerate valence band p-like symmetry. In the presence of strong spin–orbit coupling, part of this degeneracy is raised resulting in the doubly degenerate (at $k = 0$) heavy and light mass valence bands and a so-called spin–orbit split-off band. The reader should go to the general reviews quoted under 'Further Reading' for the equivalent coverage of other common systems such as the wurzite II–VI and the lead salt systems.

The dynamics of a free electron subject to an external magnetic field have been given by a variety of authors, but the most universally used treatment is that of Landau. For the motion of an electron in a semiconductor we treat first the case of a simple parabolic energy band. For the case of the magnetic field, B, along the z-direction, k_z remains a good quantum number, but k_x and k_y are subsumed into the new magnetic orbital (Landau) quantum number, n. The original three-dimensional energy band, $E = \hbar k^2/2m^*$, is coelesced into a series of one-dimensional Landau sub-bands in the k_z-direction (where the motion along k_z remains unchanged):

$$E_n^{\pm}(k_z) = (n + 1/2)\,\hbar\omega_c + \frac{\hbar k_z^2}{2m^*} \qquad [1]$$

The cyclotron frequency is $\omega_c = eB/m^*$ and μ_B is the Bohr magneton. The resulting electron energy eigenvalues at $k_z = 0$ in an external magnetic field (i.e., Landau sub-band extrema) are referred to as Landau levels, LLs. In practice each state is doubly degenerate with spin which results in an additional term $\pm g^*\mu_B B$ on the right-hand side, where g^* is the effective g-factor. (Note that in the presence of strong spin–orbit coupling and small m^*, there is a corresponding g^* which becomes negative and substantially different from the free-electron g-factor; the largest value for a binary elemental compound is that for InSb, where $g^* \approx -50$.) The situation for a simple parabolic conduction band near $k = 0$ is shown in Figure 1a,b. Intraband optical absorption associated with direct transitions between successive Landau sub-bands is referred to as cyclotron resonance (CR) absorption. There are corresponding Landau sub-bands in the valence band, and direct transitions between valence and conduction sub-bands give rise to the so-called interband magneto-optical (IMO) effects.

Expressions for the density of states (DOS) in three, two, and one dimensions in the absence of a magnetic field are given elsewhere ('semiconductor basics' – Pidgeon). In the presence of a magnetic field the three-dimensional energy band is reduced to one-dimensional sub-bands so that the energy dependence of the DOS has the one-dimensional form with the energy given by eqn [1]. Performing this substitution and multiplying by a term for the degeneracy of the Landau levels (LLs) gives the following expression for the total DOS per unit volume:

$$g(E) = \frac{(2m^*)^{1/2}}{4\pi^2\hbar}\,\frac{eB}{\hbar}\sum_n \left[E - (n + 1/2)\,\hbar\omega_c\right]^{-1/2} \qquad [2]$$

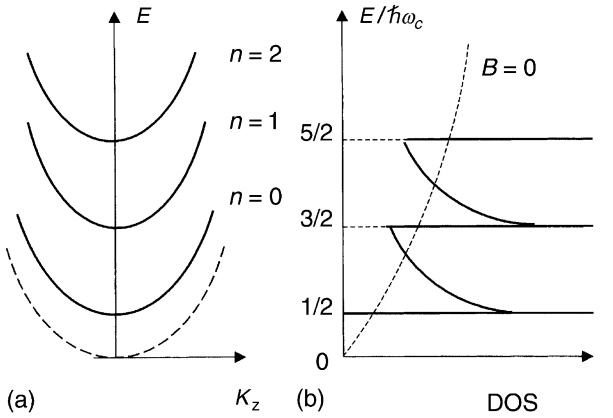


Figure 1 (a) Landau sub-bands according to eqn [1] with the magnetic field applied in the *z*-direction. The dashed line shows the parabolic conduction band for k along k_z and $B = 0$. (b) Schematic outline of the density of states (DOS) for free electrons in a magnetic field. It comprises the sum of identical contributions from each Landau sub-band. The density of states with $B = 0$ is shown dashed.

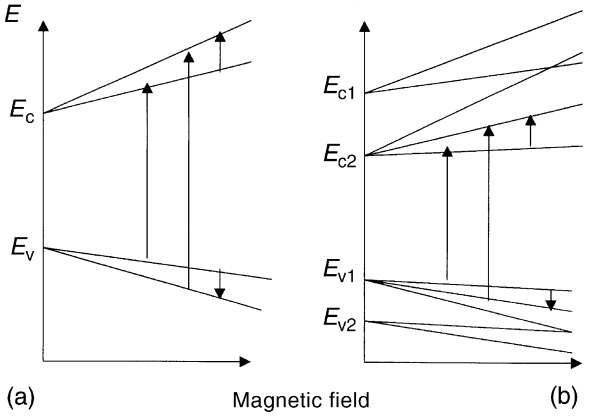


Figure 2 Fan plots of the conduction and valence band Landau levels versus magnetic field for simple parabolic bands: (a) a three-dimensional semiconductor; (b) quasi-two-dimensional systems. Some CR and interband magneto-optical transitions are shown by vertical arrows.

Thus, the DOS for Landau sub-bands is quite different from the zero-field DOS, and is a series of square-root singularities with divergences at the energies $(n + 1/2)\hbar\omega_c$. The two cases are illustrated schematically in Figure 1b. Since the optical absorption strength for both interband and intra-band transitions is proportional to the joint density of states for the initial and final states, this result shows the important advantage of the use of a magnetic field in semiconductor spectroscopy because it means that the absorption lines for CR or IMO transitions are strongly peaked at the sub-band extrema (i.e., at $k_z = 0$). This allows accurate measurement of LL energy differences for transitions observed in a single experiment, and hence, by comparison to theoretical models, leads to the best determination of fundamental energy-band parameters. The basic LL selection rule can be shown straightforwardly from time-dependent perturbation theory to be $\Delta n = 0$, $\Delta m_s = 0$ for IMO transitions and $\Delta n = 1$, $\Delta m_s = 0$ for CR absorption, where m_s is the spin quantum number (i.e., $\pm\frac{1}{2}$, for spin up and spin down). The cyclotron resonance absorption occurs only for one sense of circularly polarized light, but the IMO effect can occur for all polarizations. The choice then of magnetic field direction, crystal orientation, and polarization of the optical radiation allows extremely detailed determination of the energy band parameters for particular crystal symmetries.

The data are often summarized in a so-called fan diagram of the energy levels, or transition energies, versus magnetic field. Such a plot is shown for hypothetical conduction and valence band LLs in Figure 2a, together with interband and CR transitions. The states are labeled by the Landau quantum number, with selection rules $\Delta n = 0$ and $\Delta n = 1$ for interband and CR absorption transitions, respectively (or $\Delta n = -1$ for CR emission). If the energies of the transitions themselves are plotted versus the magnetic field, then the intercept of the fan plots at $B = 0$ gives the zero-field energy gap. Note that in the presence of excitonic effects (see below) the intercept will give the energy gap plus the exciton binding energy.

For a strictly two-dimensional system the DOS per unit area is independent of energy: $g(E) = m^*/\pi\hbar^2$. In the case of quasi two-dimensional confinement in the *z*-direction the motion in the *z*-direction is quantized, and the total energy becomes (*see* **Semiconductor Physics:** Quantum Wells and GaAs-Based Structures):

$$E = E_i + \frac{\hbar^2}{2m^*}(k_x^2 + k_y^2) \qquad [3]$$

Equation [3] describes a set of confinement sub-bands (parabolas induced by geometric, as opposed to magnetic, confinement), the bottoms of which are at the energies E_i, where i is the sub-band index. The energies, E_i, depend on the shape of the confining potential. The DOS per unit area is in this case a series of step functions of height $m^*/\pi\hbar^2$. From the point of view of magneto-optics, the main effect of the quantum confinement is to suppress the k_z motion, so that in the presence of an external magnetic field along the *z*-direction, the motion is completely quantized. The energy eigenvalues are given by:

$$E_{n,i} = E_i + (n + 1/2)\,\hbar\omega_c \qquad [4]$$

The total DOS becomes:

$$g(E) = \frac{m^* \omega_c}{\pi \hbar} \delta[E - (n + 1/2)\,\hbar\omega_c] \qquad [5]$$

a series of delta-functions located at the energy positions of the LLs. The DOS is obviously more singular in two dimensions than in three dimensions, and leads to substantially sharper magneto-optical and magneto-transport effects. For confined structures the selection rules for interband and intersub-band transitions are identical to the bulk case for the electromagnetic (em) radiation polarized in the xy plane provided that the external magnetic field is along the confinement (z-) direction. The only difference is that the conservation of momentum in the z-direction ($\Delta k_z = 0$) is replaced by a selection rule on the sub-band index, $\Delta i = 0$. A fan plot for this configuration is shown schematically in Figure 2b. For each confinement sub-band transition there is a set of LL transitions. When the energy positions are plotted against B, this leads to a set of Landau 'fans', each of which extrapolates to the sum of a valence plus conduction confinement sub-band energy at $B = 0$. Allowed transitions are shown for the lowest two confinement sub-bands in Figure 2b. Again this picture may be complicated in the presence of excitonic effects at low temperature.

Turning again to the bulk semiconductor picture, eqns [1]–[5] apply for a single parabolic energy band and are valid for the conduction band in the quasi-Ge model. In actual semiconductors further complications arise. Most importantly the triply degenerate p-like valence band (heavy hole, light hole, and split-off band) must be treated by so-called degenerate perturbation theory. In the quasi-Ge model the conduction and three valence bands are treated together as one coupled band system using the so-called $\mathbf{k} \cdot \mathbf{p}$ method (described elsewhere by Zawdzki). With spin this results in an 8×8 matrix equation which has to be diagonalized for the eigenvalues and eigenvectors of the system. The spin–orbit coupling results in so-called complex, mixed-spin valence bands and anomalously spaced low quantum number LLs (the so-called quantum effects). The overall energy band scheme for InSb, near the fundamental direct gap in the center of the Brillouin zone for conduction and (heavy and light mass) valence bands is shown schematically in Figure 3, for no magnetic field and for an applied magnetic field. Each LL is doubly degenerate with spin, the a-set representing the so-called spin-up and the b-set the spin-down states.

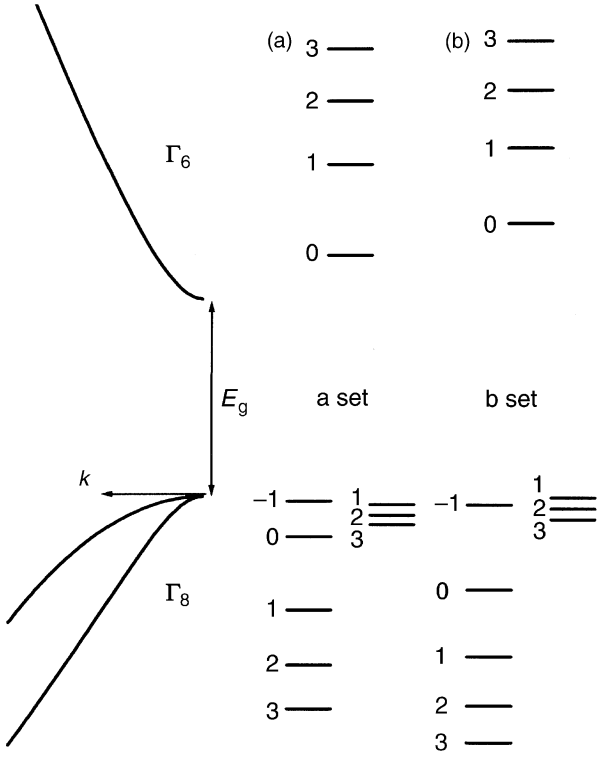


Figure 3 (a) Schematic diagram of the conduction and valence bands for InSb (symmetry labels Γ_6 and Γ_8 respectively) at $B = 0$. (b) Computed Landau levels for the conduction and valence bands of InSb at 4 K ($E_g = 0.23$ eV) with $B = 5$ T along the [100] direction, grouped according to the so-called a- and b-spin sets (see text).

Experimental

The first serious use of magneto-optics to study effective mass and energy band structure of semiconductors was in the microwave cyclotron resonance (i.e., intraband) investigations of Ge and Si in the early 1950s. Classically it can be shown that the ratio of the radiation power absorbed, P, to power absorbed in zero magnetic field, P_0, is given by:

$$\frac{P}{P_0} = \frac{1 + (\omega^2 + \omega_c^2)\tau^2}{[1 + (\omega_c^2 - \omega^2)\tau^2]^2 + 4\omega^2\tau^2} \qquad [6]$$

where τ is the collision time of the carriers. This expression for P/P_0 describes cyclotron resonance absorption and is plotted as a function of frequency in Figure 4 for different values of the parameter $\omega_c\tau$. Clearly the peak absorption occurs at $\omega = \omega_c$, namely at cyclotron resonance. Also evident is the very important requirement for observation of a clear resonance that $\omega_c\tau > 1$; this condition corresponds physically to a travel time between collisions of the particle of at least $1/2\pi$ of a revolution so that the particle can interact for a sufficiently long time

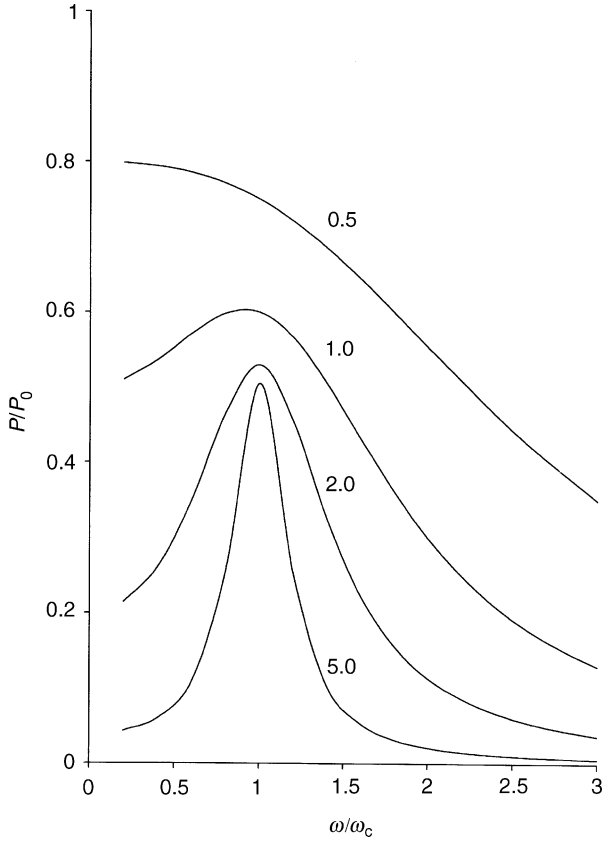


Figure 4 Ratio of the radiation power absorbed to the power absorbed at zero magnetic field, P/P_0, plotted as a function of ω/ω_c according to eqn [6], for different values of $\omega_c\tau$ from 0.5 to 5.0.

with the electromagnetic wave to absorb energy. For pure materials and low temperatures (where the condition $\omega_c\tau > 1$ is achievable) this is the best technique for obtaining the details of the electron and hole band parameters of a semiconductor near $k = 0$. The most thorough and successful studies of this type have been achieved in pure Si, Ge, and InSb where (with the aid of optical excitation to populate the levels shown in Figure 3) all possible electron, light and heavy hole cyclotron resonances have been resolved. In this way the valence band quantum effects, described above, have been observed in detail. The technique is equally applicable in low-dimensional systems where CR of both electrons and holes has been resolved in a large number of n- and p-type semiconductor heterostructures and quantum wells.

In interband – either absorption or photoluminescence (PL) – spectroscopy at low temperatures, the photo-excited electrons and holes interact and form excitons (*see* **Semiconductor Physics:** Excitons). In many cases the excitonic state may be described in the so-called envelope function approximation as a hydrogenic system, in which the electron and hole are bound by their mutual Coulomb attraction. The energy spectrum of such a hydrogenic exciton is given by

$$E_n = \frac{\mu e^4}{2\hbar^2\varepsilon^2 n^2} \qquad [7]$$

where n is the principal quantum number ($n = 1, 2, 3, \ldots$, but note that the symbol n is also used for the Landau quantum number above), μ is the electron–hole reduced effective mass, and ε is the background dielectric constant. In low-dimensional (i.e., confined) semiconductor systems of the type described above the degeneracy of the valence bands at $k = 0$ is removed, and two types of excitons are observed; these are referred to as 'heavy' and 'light' hole excitons in the literature, where heavy and light refer to the effective masses along the confinement direction. These excitons can move freely (in the remaining two unconfined directions) or be bound to an impurity or defect. In these systems exciton absorption may be observed involving not only the ground-state excitons, but also excitons associated with higher confinement sub-bands. This is shown for the so-called asymmetric quantum well system $Al_yGa_{1-y}As/In_xGa_{1-x}As/GaAs$ in Figure 5a. The data are taken both by the direct PL method (where the exciting beam has an energy much larger than the bandgap and the electrons and holes de-excite instantaneously to the bottom of the bands before recombining as exiton recombination radiation), and so-called PL excitation spectroscopy (where the exciting beam is a tunable laser, tuned into resonance with the successive exciton resonances).

The interband Landau transitions can be observed either in PL or as excitations in a magneto-absorption or magnetoreflection experiment. Excitons are affected differently by the application of an external magnetic field depending whether the ratio $\gamma = \hbar\omega_c/2E_B$ (where E_B is the exciton binding energy with $n = 1$ above) is less than or greater than one. In the so-called magnetic limit where $\gamma \gg 1$, the excitonic ground state moves approximately as $\frac{1}{2}\hbar\omega_c$, i.e., linearly with B. In this regime the magnetic field energy is the dominant term in the exciton Hamiltonian. The various states of the exciton (ground and excited states) become associated with the interband LL transition, and exhibit an energy dependence that is nearly identical to the free carrier LL transitions. The energies of the various excitonic transitions are shifted down with respect to the corresponding LL transitions by an amount equal to the binding energy of the excitonic state to that LL. Typical PL and PLE (photoluminescence excitation)

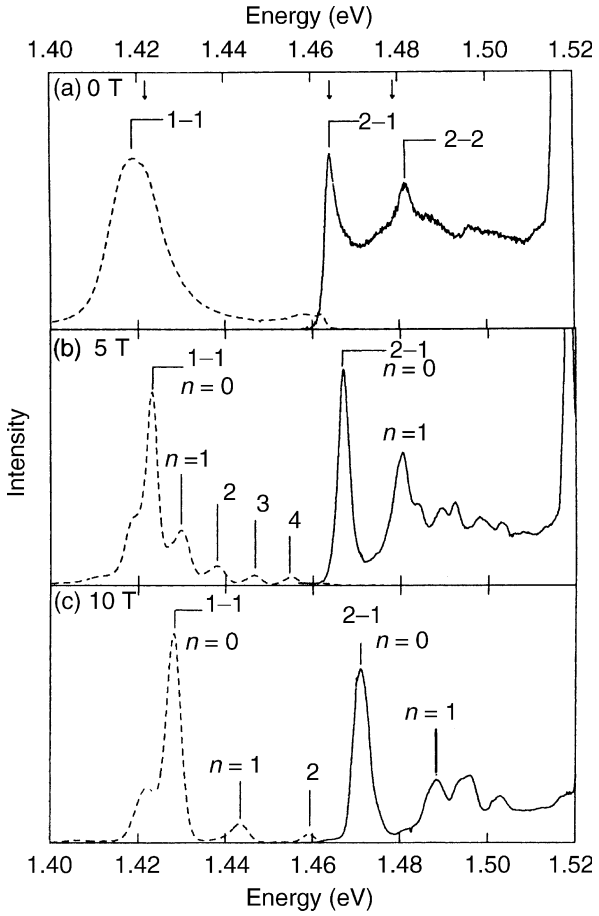


Figure 5 Interband transition data obtained in zero magnetic field (a) and fields of 5 T (b) and 10 T (c) from asymmetric quantum wells. The dashed lines indicate data obtained from PL and the solid lines data from PLE. The arrows in (a) indicate the calculated interband transition energies $1 \rightarrow 1$, $2 \rightarrow 1$, and $2 \rightarrow 2$, the latter being clearly resolved in PLE (the asymmetric wells mean that transitions with $\Delta i \neq 0$ are allowed). In (b) and (c) the development of the spectra into LL transitions is clearly seen, although beyond 1.48 eV the analysis is obviously extremely complicated by the overlap of LL transitions associated with all three of the above sub-bands.

spectra for interband magneto-optical transitions are shown in Figure 5b,c for a fixed applied magnetic field in the asymmetric quantum well system $Al_yGa_{1-y}As/In_xGa_{1-x}As/GaAs$.

Spin Flip Raman Scattering

In addition to the interband transitions discussed above, Raman scattering of intersub-band transitions between conduction or valence band LLs can also be studied. It provides a versatile means of probing band edge parameters of electrons and holes in semiconductors since the excitation (normally a laser pump) is not confined to the resonance frequencies under investigation, i.e., any pumping energy below the band edge will do. Electronic Raman scattering is then allowed between LLs via virtual interband transitions. In the presence of spin–orbit coupling strong spin-flip processes via electric dipole transitions become allowed giving rise to the so-called spin-flip Raman (SFR) cross-section. This is a single-particle process in which electrons or holes interacting with the radiation change their spin state. This process can be made to go stimulated particularly easily when the pump radiation is resonant with the bandgap of the semiconductor, forming the spin-flip Raman laser. The most important practical case has probably been that of InSb, resonantly pumped near 5 μm with the CO laser. In this case the small effective mass of a narrow-gap semiconductor in the presence of substantial spin–orbit coupling leads to both an extremely large anomalous effective g-factor ($g^* \sim -50$) – giving a large tuning range, $\Delta E = \hbar\omega_{SF} = g^*\beta B$ – and a strong cross-section which is larger than that of the free electron by a factor $(m_0/m^*)^2$. Efficient cw operation is possible with threshold powers less than 50 mW.

SFR scattering has enabled the observation of a great variety of excitations in nonmagnetic and dilute magnetic semiconductors, including: free electrons and holes, electrons bound to donors and holes bound to acceptors, scattering from bound magnetic polarons, and scattering due to SF transitions within the Zeeman multiplets of paramagnetic ions. In addition to pure SF transitions, combined resonances with combinations such as $\omega_{CR} \pm \omega_{SF}$ have also been detected. The process has also been used in a similar manner to electron spin resonance to study coherent processes, including Raman photon echo spectroscopy in semiconductors.

See also

Magneto-Optics: Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Häfele HG (1991) Spin-flip raman scattering. In: Landwehr G and Rashba WI (eds) *Landau Level Spectroscopy in Modern Problems in Condensed Matter Sciences*, 27.1, 207–275.

Petrou A and McCombe BD (1991) Magnetospectroscopy of confined semiconductor systems. In: Landwehra G and Rashba WI (eds) *Landau Level Spectroscopy in Modern Problems in Condensed Matter Sciences*, 27.1, 679–775.

Pidgeon CR (1980) Free carrier optical properties of semiconductors. In: Balkanski M (ed.) *Optical Properties of Solids. Handbook of Semiconductors*, pp. 223–328. Amsterdam, The Netherlands: North-Holland.

Zawadzki W (1991) Intraband and interband magneto-optical transitions in semiconductors. In: Landwehr G and Rashba WI (eds) *Landau Level Spectroscopy in Modern Problems in Condensed Matter Sciences*, 27.1, 483–512.

MATERIALS CHARACTERIZATION TECHNIQUES

Contents

$\chi^{(2)}$
$\chi^{(3)}$

$\chi^{(2)}$

R C Eckardt, Cleveland Crystals, Inc., Highland Heights, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Techniques used to measure second-order susceptibility $\chi^{(2)}$ or nonlinear optical coefficients are described in this article, and some measured values are quoted. These are the parameters that describe the property of a material that is responsible for second-harmonic generation and other three-wave nonlinear optical interactions. These other interactions include sum- and difference-frequency generation in which two waves combine to generate a third wave of a frequency that is the sum or difference of the first two waves. Also included are optical parametric generation and optical parametric oscillation, in which a single pump wave generates two resulting waves with frequencies that sum to the frequency of the pump wave. The measurement techniques involve measuring the efficiency of nonlinear frequency conversion by these processes.

The measurements require a great deal of care and precision, and it is advisable to perform a comparison with previous measurements and obtain correlation with accepted standards whenever possible. A great deal of effort continues to be expended by researchers to establish correlations between measurements and provide additional values that are accepted as standards. A number of techniques have been used for these measurements sometimes yielding results that disagree. There are, however, values derived from early measurements that have proven accurate. In these measurements it is necessary to accurately control and characterize the polarization, power, spectral distribution, and spatial and temporal intensity distribution of the incident radiation and accurately measure the frequency-converted light, which can be many orders of magnitude smaller.

The development and availability of single-longitudinal-mode and nearly diffraction-limited lasers has greatly reduced the ambiguity and simplified increased accuracy of measurements of second-order nonlinear optical coefficients. The improved optical quality of nonlinear-optical crystals has also been an important component of the increased accuracy of measurement. Measurement techniques have also improved, providing absolute measurements of optical intensity over wide dynamic ranges at different wavelengths. Until the early 1990s, there was a range of measured values for nonlinear optical coefficients of common nonlinear materials, and numerous definitions were in use. Caution is required when consulting early publications on this topic, and some of this ambiguity is carried over to present time. Different definitions vary by factors of 2 and the permittivity of free space, $\varepsilon_0 \cong 8.854 \times 10^{-12}$ C/N $\times$ m^2. It is important that the definition of a nonlinear optical coefficient is clear before a value is used. This can be done through a statement of the nonlinear electric polarization or the coupled equations for nonlinear frequency conversion.

The standard techniques that have been used to measure nonlinear optical coefficients are covered in detail in publications listed in the Further Reading at the end of this article. Brief descriptions of some of these techniques, such as Maker-fringe and wedge,

optical parametric fluorescence, and phase-matched second-harmonic generation, will be presented in this article. Each of these techniques has a unique set of advantages and disadvantages. There was some lively controversy over the technique that provided the most accurate measurements and those measurements that should be dismissed. Measurements based on phase-matched second-harmonic generation had the advantage of being most similar to practical application of nonlinear optical frequency conversion. However, the large divergence and multiple-longitudinal-mode operation of early lasers made this method unreliable. Even relative comparisons of nonlinear optical coefficients between different materials were inaccurate because highly structured laser pulses had different conversion characteristics due to differences in bandwidth acceptance, group velocity walkoff, and birefringent walkoff in different materials.

Nonlinear optical coefficients of improved accuracy were required for analysis of optical harmonic conversion in the large, well-characterized laser systems used in inertial-confinement fusion research. The development of these systems provided a standardization on the value of $d_{36} = 0.39\ \text{pm V}^{-1}$ ($1\ \text{pm} = 10^{-12}\ \text{m}$) for second-harmonic generation of neodymium laser radiation in the material potassium dihydrogen phosphate (KH_2PO_4 or KDP). The development of injection-seeded Q-switched lasers provided single-longitudinal-mode laser output that allowed the phase-matched second-harmonic techniques to be extended to provide more accurate absolute measurements and relative comparisons on KDP and a number of other materials. Correlation was achieved on a larger number of relative and absolute measurements of nonlinear optical coefficients to provide some systematic agreement. It is not well understood why the earlier measurements referenced to optical parametric fluorescence produced values that were higher by about 60%. Measurement techniques have been further refined, and more exacting measurements are leading agreement, using different methods of measurement on an increasing number of nonlinear optical materials.

Definitions

The definition of the nonlinear optical coefficient can be specified in a number of ways. Examples given here are how the coefficient relates to the nonlinear electric polarization in a material, the coupled equations for nonlinear frequency conversion, and expressions describing second-harmonic generation with a monochromatic plane-wave fundamental.

The second-order optical nonlinearity is reported as both the second-order susceptibility $\chi^{(2)}$ and the second-order nonlinear optical coefficient d, which usually differ in definition by a factor of two. These are third-rank tensors that relate products of components of the electric field vector E_j and E_k to the vector components of the nonlinear polarization P_i^{NL}. The second-order susceptibility typically is used to express instantaneous values:

$$P_i^{NL}(t) = \varepsilon_0 \sum_{j,k} \chi_{ijk}^{(2)} E_j(t) E_k(t) \qquad [1]$$

whereas the nonlinear optical coefficient more often is used to relate the vector amplitudes of frequency components of these quantities. The electric field at the fundamental angular frequency ω can be expressed as

$$E_j(\omega, \mathbf{r}, t) = \frac{1}{2}\{E_j(\omega) \exp i(\mathbf{k}_\omega \cdot \mathbf{r} - \omega t) + c.c.\} \qquad [2]$$

and the nonlinear polarization at the harmonic frequency 2ω can be expressed as

$$P_i^{NL}(2\omega, \mathbf{r}, t) = \frac{1}{2}\left\{P_i^{NL}(2\omega) \exp i(2\mathbf{k}_\omega \cdot \mathbf{r} - 2\omega t) + c.c.\right\} \qquad [3]$$

The intensity is given by $I = (nc\varepsilon_0/2)|E|^2$. The relationship between the complex amplitudes expressed in terms on the nonlinear optical coefficient tensor is

$$P_i^{NL}(2\omega) = \varepsilon_0 \sum_{j,k} d_{ijk}(-2\omega; \omega, \omega) E_j(\omega) E_k(\omega) \qquad [4]$$

When waves of three different frequencies are involved, the expression becomes

$$P_i^{NL}(\omega_3) = 2\varepsilon_0 \sum_{j,k} d_{ijk}(-\omega_3; \omega_2, \omega_1) E_j(\omega_2) E_k(\omega_1) \qquad [5]$$

Expressing the coefficient as $d_{ijk}(-\omega_3; \omega_1, \omega_2)$ allows for dispersion and indicates that fields at frequencies ω_1 and ω_2 interact to produce a nonlinear polarization at frequency ω_3, the sum of the first two frequencies. The relationship between the second-order electrical susceptibility and the second-order nonlinear optical coefficient is

$$\chi_{ijk}^{(2)} = 2 d_{ijk} \qquad [6]$$

The dispersion of nonlinear optical coefficients is approximated by the use of Miller's delta:

$$\delta_{ijk} = \frac{d_{ijk}(-\omega_3; \omega_2, \omega_1)}{(n_i^2(\omega_3) - 1)(n_j^2(\omega_2) - 1)(n_k^2(\omega_1) - 1)} \qquad [7]$$

The indices of refraction in the direction of the electric field components $E_m(\omega_n)$ are $n_m(\omega_n)$ for $m=i,j,k$ and $n=1,2,3$. There is some theoretical and experimental justification for treating the components of Miller's delta as constants. Earlier definitions of Miller's delta included multiplication by ε_0, the permittivity of free space.

The nonlinear optical coefficient is usually reduced to a scalar quantity, the effective nonlinear coefficient or d_{eff}, that relates only the specific vector components of the fields involved in the interaction. The effective nonlinear optical coefficient will be discussed in more detail later. The coupled equations for second-harmonic generation using d_{eff} are

$$\frac{\mathrm{d}E_{2\omega}(z)}{\mathrm{d}z} = \mathrm{i}\frac{\omega d_{\text{eff}}}{n_{2\omega}c}E_{\omega}(z)E_{\omega}(z)\exp(-\mathrm{i}\,\Delta k z) \quad [8]$$

$$\frac{\mathrm{d}E_{\omega}(z)}{\mathrm{d}z} = \mathrm{i}\frac{\omega d_{\text{eff}}}{n_{\omega}c}E_{\omega}(z)E^{*}_{\omega}(z)\exp(\mathrm{i}\,\Delta k z) \quad [9]$$

Here ω is the angular frequency of the fundamental wave; z is distance in the direction of propagation, $n_{2\omega}$ and n_{ω} are the refractive indices of the harmonic and fundamental respectively; and c is the speed of light in free space. The complex electric field amplitudes of the harmonic and fundamental, $E_{2\omega}(z)$ and $E_{\omega}(z)$, respectively, are assumed to be slowly varying functions of position compared with the oscillation of the carrier waves $\exp \mathrm{i}(\mathbf{k}_{2\omega}\cdot\mathbf{r} - 2\omega t)$ and $\exp \mathrm{i}(\mathbf{k}_{\omega}\cdot\mathbf{r} - \omega t)$, with $\mathbf{k}_{2\omega}$ and $\mathbf{k}_{\omega}$ the wavevectors of harmonic and fundamental; the asterisk indicates complex conjugate. The quantity Δk is the wavevector mismatch and is given by $\Delta k = k_{2\omega} - 2k_{\omega}$. These are simplified expressions that apply to nearly monochromatic plane waves. Two solutions to the coupled equations are

$$I_{2\omega}(z) = \frac{2\omega^2 d_{\text{eff}}^2 l^2 I_{\omega}^2(0)}{n_{2\omega}n_{\omega}^2c^3\varepsilon_0}\frac{\sin^2(\Delta k z/2)}{(\Delta k z/2)^2} \quad [10]$$

for insignificant depletion of the fundamental amplitude and

$$I_{2\omega}(z) = I_{\omega}(0)\tanh^2\sqrt{\frac{2\omega^2 d_{\text{eff}}^2 z^2 I_{\omega}(0)}{n_{2\omega}n_{\omega}^2c^3\varepsilon_0}} \quad [11]$$

for the case of perfect phase matching $\Delta k = 0$ and allowing depletion of the fundamental. The quantity $I_{2\omega}(z)$ is the harmonic intensity after propagating a distance z in the nonlinear material, and $I_{\omega}(0)$ is the fundamental intensity at the input surface of the nonlinear material where $I_{2\omega}(0) = 0$.

The above relationships for nonlinear polarization, the coupled equations for second-harmonic generation, or harmonic conversion of a monochromatic plane wave, each determine the definition of the second-order nonlinear optical coefficient.

Symmetry

Symmetry has an important role in the measurement of nonlinear optical coefficients. Indeed the second-order nonlinear optical coefficient must be identically zero if the material has inversion symmetry. This restricts materials to noncentrosymmetric crystals or isotropic materials in which inversion symmetry is removed by such techniques as application of an electric field or polymeric materials in which alignment of molecules is achieved by some poling process. The discussion here is limited to noncentrosymmetric crystals.

The nonlinear optical coefficient is a tensor of rank 3. Representation of the tensor in a rotated coordinate system is described by

$$d'_{ijk} = \sum_{l,m,n} R_{il}R_{jm}R_{kn}d_{lmn} \quad [12]$$

where R_{il} is the matrix that describes rotation of the initial coordinate system xyz to the new coordinate system $x'y'z'$. The representation of a vector r_l in the original coordinate system becomes $r'_{\mathrm{i}} = \sum_l R_{il}r_l$ in the new coordinate system. A required symmetry property of the nonlinear optical tensor, that exchange of the second two indices of a component does not change the value of the component $d_{ijk} = d_{ikj}$, allows the third-rank tensor to be written as a 3 by 6 matrix:

$$\begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} \end{pmatrix} \quad [13]$$

with the components of the matrix related to the tensor components by

$$\begin{aligned} &d_{i1} = d_{i11} \quad d_{i4} = d_{i23} = d_{i32} \\ &d_{i2} = d_{i22} \quad d_{i5} = d_{i13} = d_{i31}, \; i = 1,2,3 \\ &d_{i3} = d_{i33} \quad d_{i6} = d_{i12} = d_{i21} \end{aligned} \quad [14]$$

A further property, tensor components with any permutation of the indices are equal $d_{ijk} = d_{kij}$, called Kleinman symmetry or Kleinman's conjecture, is often applied if there is no absorption at any of the wavelengths involved. Kleinman symmetry assures that the amount of energy depleted from the

fundamental is equal to the amount of energy gained by the second harmonic. This reduces the number of independent components of the nonlinear optical matrix to 10: d_{11}; $d_{12} = d_{26}$; $d_{13} = d_{35}$; $d_{14} = d_{25} = d_{36}$; $d_{15} = d_{31}$; $d_{16} = d_{21}$; d_{22}; $d_{23} = d_{34}$; $d_{24} = d_{32}$; and d_{33}. Sometimes components equal, according to Kleinman symmetry, are measured individually, either as a check of the measurement or to check the validity of the conjecture.

Crystal symmetry places additional constraints on the components of the nonlinear optical matrix. These constraints determine the components that must be zero and determine additional relationships between the nonzero components. Matrices for selected crystal point groups are given in Table 1. Effective nonlinear optical coefficients can be found by a coordinate frame transformation in which the rotated axes align with the direction of propagation and the eigen polarizations. Effective nonlinear coefficients for a few crystal point groups are given in Table 2. More complete listings of both the reduced matrices and effective nonlinear optical coefficients are given elsewhere.

Table 1 Nonlinear coefficient matrices for selected crystal point groups

Orthorhombic *mm*2

$$\begin{pmatrix} 0 & 0 & 0 & 0 & d_{15} & 0 \\ 0 & 0 & 0 & d_{24} & 0 & 0 \\ d_{31} & d_{32} & d_{33} & 0 & 0 & 0 \end{pmatrix}$$

Hexagonal 6

$$\begin{pmatrix} 0 & 0 & 0 & d_{14} & d_{15} & 0 \\ 0 & 0 & 0 & d_{15} & -d_{14} & 0 \\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{pmatrix}$$

Tetragonal $\bar{4}2m$

$$\begin{pmatrix} 0 & 0 & 0 & d_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{25} & 0 \\ 0 & 0 & 0 & 0 & 0 & d_{36} \end{pmatrix}$$

Hexagonal 6*mm*

$$\begin{pmatrix} 0 & 0 & 0 & 0 & d_{15} & 0 \\ 0 & 0 & 0 & d_{15} & 0 & 0 \\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{pmatrix}$$

Trigonal 3*m* $m \perp$ X

$$\begin{pmatrix} 0 & 0 & 0 & 0 & d_{15} & -d_{22} \\ -d_{22} & d_{22} & 0 & d_{15} & 0 & 0 \\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{pmatrix}$$

Cubic $\bar{4}3m$

$$\begin{pmatrix} 0 & 0 & 0 & d_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & d_{14} \end{pmatrix}$$

Measurements

Usually the symmetry of a nonlinear crystal is determined by techniques such as X-ray diffraction or surface morphology. Crystal symmetry determines the nonzero components of the second-order nonlinear optical tensor. Nonphase-matched second-harmonic-generation techniques, such as Maker-fringe and wedge measurements, allow selection of polarizations and propagation directions relative to the crystal lattice that isolates individual components of the tensor. Phase-matched measurement techniques usually require propagation in a direction for which an effective nonlinear optical coefficient applies and can involve more than one tensor component. Parametric fluorescence and sum- and difference-frequency-generation measurements are phase-matched processes that involve three wavelengths. All of the measurement techniques have been used with pulsed and continuous-wave laser output. In all of these methods, except parametric fluorescence, it is necessary to accurately characterize the incident radiation. Measurements relative to a standard material avoid some of the concerns of characterization of the fundamental radiation. Relative measurements are important in that they allow correlation between different sets of measurements. Measurement sets, that include both relative and absolute measurements on a number of materials, are useful in developing correlations that achieve more accurate compilations of values.

It is desirable to have an incident beam with a Gaussian-like transverse intensity distribution and close to diffraction limited propagation. Gaussian beams have advantages that they are most easily characterized and their propagation is described with relatively simple analytical methods. Accurate temporal characterization of the incident light is also essential. Here it is desirable to have stable single-mode laser operation to produce narrow bandwidth light that is well within phase-matching acceptance bandwidths and that has temporal structure slow enough to be easily characterized and slow enough

Table 2 Effective nonlinear optical coefficient for a few symmetries

*mm*2, a point group with biaxial birefringence		
$\phi = 0$, two waves $E \| Y$, one wave $E \| XZ$	$d_{\text{eff}} = d_{32} \sin\theta$	
$\phi = 90°$, two waves $E \| XY$, one wave $E \| Z$	$d_{\text{eff}} = d_{31} \cos^2\phi + d_{32} \sin^2\theta$	
$\phi = 90°$, two waves $E \| X$, one wave $E \| YZ$	$d_{\text{eff}} = d_{31} \sin\theta$	
Some crystal symmetry groups with uniaxial birefringence		
point group	two e rays and one o ray	two o rays and one e ray
$\bar{4}2m$	$d_{\text{eff}} = d_{14} \sin^2 2\theta \cos 2\phi$	$d_{\text{eff}} = -d_{14} \sin\theta \sin 2\phi$
3*m*	$d_{\text{eff}} = d_{22} \cos^2\theta \cos 3\phi$	$d_{\text{eff}} = d_{15} \sin\theta - d_{22} \cos\theta \sin 3\phi$
6 and 6*mm*	$d_{\text{eff}} = 0$	$d_{\text{eff}} = d_{15} \sin\theta$

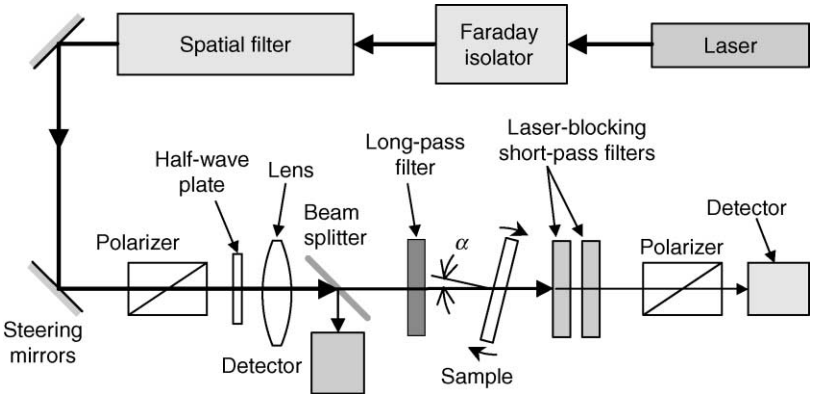


Figure 1 Schematic illustration of an experimental setup to record Maker-fringe measurements.

that group-velocity mismatch will be of no concern. Generating such laser output usually requires special techniques, such as injection-seeded pulse operation or stabilized single-frequency cw operation. Often spatial filters are required to improve the beam quality on the laser output. Even though these topics are only mentioned here, the techniques are important and demand much effort.

The analysis of measurements can also be complicated. Consider the Maker-fringe measurement technique (Figure 1). In this method a focused fundamental beam is incident on a thin slab of oriented nonlinear optical material for nonphase-matched second-harmonic generation. The thickness of the slab is large enough that the function $\sin(\Delta kz/2)$ has gone through many cycles. As the crystal is rotated both the path length in the material l and the wave vector mismatch Δk change, and $\sin^2(\Delta k \cdot l/2)$ oscillates between maxima and minima (Figure 2). The coherence length $L_c = \pi/\Delta k$ is normally small enough that conversion remains in the region of small depletion of the fundamental. Knowledge of the material dispersion and birefringence, required to calculate Δk, or L_c, can be obtained from the fringe spacing. Ideally, the incident fundamental beam will be focused to a spot size w_0, small enough to provide measurable harmonic conversion but large enough that the near-field approximation, in which diffraction is not significant, can be used for analysis of harmonic generation. In this case integration over the incident fundamental intensity distribution $I_\omega(r, z = 0) = I_{\omega,0} \exp(-2r^2/w_0^2)$ yields an expression for the harmonic power:

$$W_{2\omega}(\alpha) = \frac{8\omega^2 d_{\text{eff}}^2(\alpha) W_\omega^2 L_c^2(\alpha)}{\pi^3 w_0^2 n^3 c^3 \varepsilon_0} \sin^2\left(\frac{\pi l(\alpha)}{2L_c(\alpha)}\right) \times \text{correction factors} \quad [15]$$

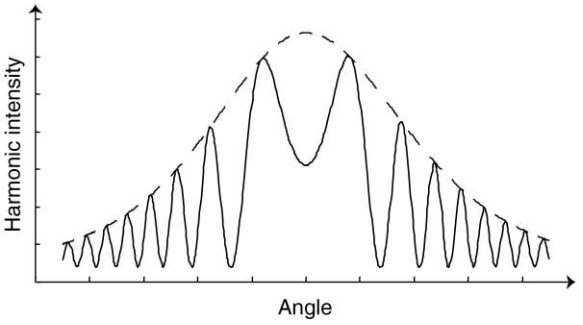


Figure 2 The envelope and spacing of the fringes allow determination of the second-order nonlinear optical coefficient and the coherence length.

where α is the angle of incidence of the fundamental radiation, $W_{2\omega}(\alpha)$ is the second harmonic power, and W_ω is the incident fundamental power. This equation applies to cw radiation. It is necessary to integrate over the temporal distributions for pulsed fundamental light. The effective nonlinear coefficient $d_{\text{eff}}(\alpha)$, the path length in the crystal $l(\alpha)$, and the coherence length $L_c(\alpha)$ change with the angle of incidence α. The correction factors include Fresnel losses at the crystal surfaces, harmonic generation at the surfaces, multiple reflections between the surfaces, absorption at either fundamental or harmonic wavelengths, and walkoff due to non-normal incidence and birefringence.

The wedge technique (Figures 3 and 4) also uses nonphase-matched second-harmonic generation. A wedged sample is translated through a focused fundamental beam, and the above equation applies with the angle α constant. The crystal length l varies with transverse position. It is important that the wedge angle and the fundamental beam spot size are chosen so that the entire beam fits well within a maximum of conversion and the beam is large enough to avoid divergence due to diffraction. Both the Maker-fringe

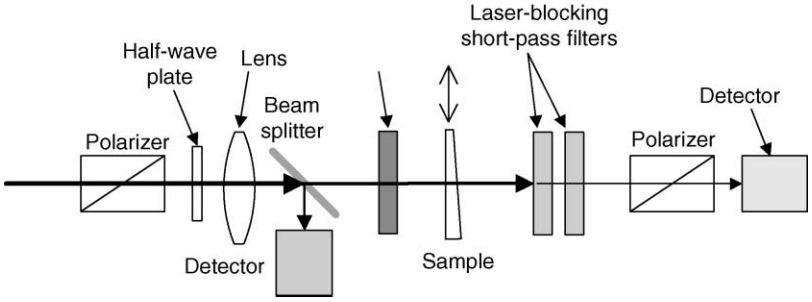


Figure 3 The setup for wedge measurements is similar to that used for Maker-fringe measurements but with a wedged sample translated in the fundamental beam to change path length.

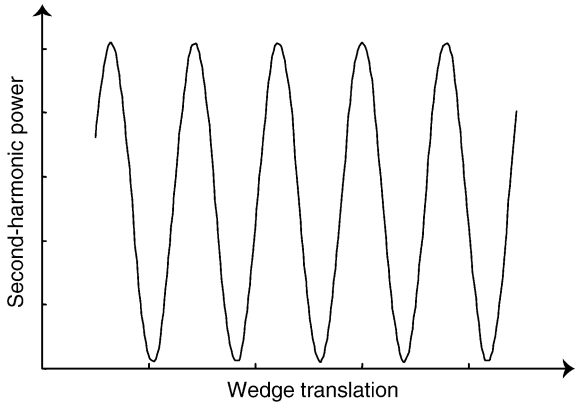


Figure 4 The harmonic intensity cycles between a minimum and maximum when the path is changed by one coherence length.

and wedge measurements have the advantage of being able to isolate individual components of the nonlinear optical tensor. Harmonic power can be low with cw beams requiring chopped, lock-in detection methods or even photon-counting techniques. Even with pulsed fundamental radiation, peak harmonic power may only be at the level of milliwatts. It is necessary to isolate the harmonic radiation from the fundamental radiation that could have peak power in the range of 10^5 W for a Q-switch laser.

Phase-matched second harmonic generation (Figure 5) offers a measurement technique that is similar to practical applications. Harmonic signals are large, thus simplifying detection. However, it is helpful to keep conversion levels below a few percent to allow the use of small depletion approximations in data analysis. Good optical quality of the nonlinear optical material and good quality and accurate characterization of the fundamental radiation are important in this measurement technique. The analysis is simplified for low levels of conversion, spectral distribution well within the phase-matching bandwidth, no absorption or other losses, and no significant diffraction or walkoff. If the fundamental intensity is a pulse with distribution of $I_\omega(r,t) = I_{\omega,0} \exp(-t^2/\tau^2) \exp(-2r^2/w_0^2)$, integration yields second-harmonic energy:

$$U_{2\omega} = \int_{-\infty}^{\infty} \mathrm{d}t \int_0^{\infty} 2\pi r \, \mathrm{d}r I_{2\omega}(r, z = l)$$

$$\approx \frac{2\sqrt{2}\omega^2 d_{\mathrm{eff}}^2 l^2 U_\omega^2}{n^3 c^3 \varepsilon_0 \pi^{3/2} \tau w_0^2} \frac{\sin^2(\Delta k l/2)}{(\Delta k l/2)^2} \qquad [16]$$

where $U_{2\omega}$ is the second harmonic energy and U_ω is the fundamental energy, r is the radial coordinate, t is time, and l is the length of the material. Harmonic conversion can be measured as Δk and varied, either by crystal rotation or temperature change, to confirm that optical quality of the nonlinear material and incident fundamental radiation are a level of tuning that is accurately predicted by theory (Figure 6). Even this simplified equation shows the importance of accurate characterization of the fundamental radiation and second-harmonic conversion in obtaining values of the nonlinear optical coefficient. Conversion efficiency is calculated more accurately with detailed analytical techniques or numerical methods, provided accurate and complete characterization of the fundamental radiation is available.

Parametric fluorescence, in principle, offers an experimental simplification over second-harmonic generation techniques. Parametric fluorescence (Figure 7) is a phase-matched technique in which a pump beam of angular frequency ω_p produces a signal beam and an idler beam of angular frequencies $\omega_\mathrm{s} + \omega_\mathrm{i} = \omega_\mathrm{p}$ by the process of spontaneous parametric generation. It is only necessary to measure the ratio of the generated signal power W_s and the incident pump power W_p. The signal power is given by

$$W_\mathrm{s} = \frac{2(h/2\pi)\omega_\mathrm{s}^4 \omega_\mathrm{i} d_{\mathrm{eff}}^2 n_\mathrm{s} l W_\mathrm{p} \Omega}{(2\pi)^2 \varepsilon_0 c^5 n_\mathrm{i} n_\mathrm{p} |\partial \Delta k / \partial \omega_\mathrm{s}|_{\omega p}} \qquad [17]$$

where h is Planck's constant, Ω is the solid angle of detection (inside the crystal), and n_s, n_i, and n_p are

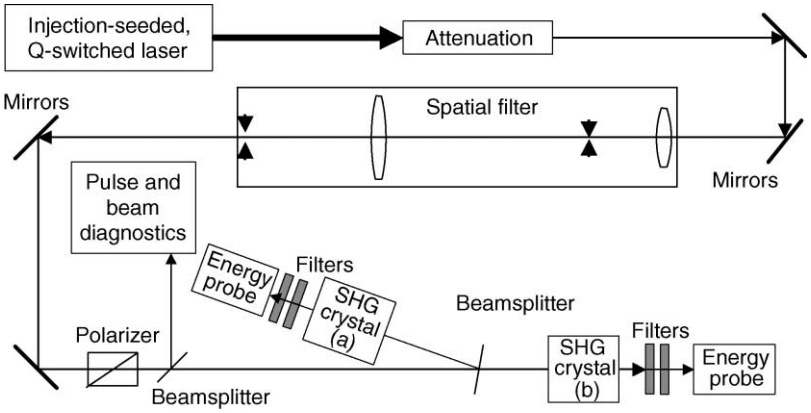


Figure 5 A setup for a relative measurement of second-order nonlinear optical coefficient based on phase-matched second-harmonic generation is shown.

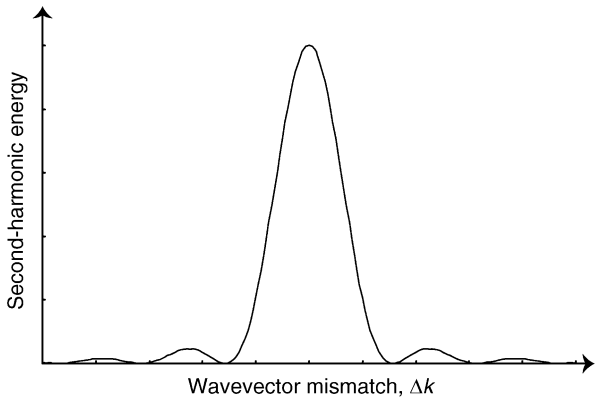


Figure 6 Observation of a theoretically predicted phase-matching tuning curve confirms required optical quality.

the indices of refraction of the signal, idler, and pump respectively. The absolute value of the partial derivative of $\Delta k = k_p - k_s - k_i$, with respect to signal frequency with the pump frequency held constant, is

$$|\partial \Delta k/\partial \omega_s|_{\omega p} = \frac{1}{c}\left| n_s - n_i + \omega_s \frac{dn_s}{d\omega_s} - \omega_i \frac{dn_i}{d\omega_i} \right| \quad [18]$$

It is still not well understood why the carefully performed earlier measurements by this method produced nonlinear optical coefficient values larger than those obtained by other techniques.

Tabulations of measurements of nonlinear optical coefficients are included in handbooks. The Internet posting 'SNLO' listed in the Further Reading is a good source of up-to-date values. Table 3 lists some nonlinear optical coefficients obtained from several sources. Signs of these coefficients are not always well determined, and they are reproduced where given and may only indicate relative polarity. Point group symmetry is given with some comments about axis assignments for the *mm*2 point group. The fundamental wavelength is given for second-harmonic measurements and twice the shortest wavelength is given for three-wave processes. Many of the values obtained by Shoji *et al.* are based on carefully performed wedge measurements with single frequency cw lasers and taking detailed account of multiple reflections in the analysis. Their measurements for congruent $LiNbO_3$ reported agreement between Maker-fringe and wedge measurements, and phase-matched difference-frequency generation measurements, and finally agreement with parametric fluorescence. The measurements of Alford and Smith combined optical-parametric amplification and second-harmonic generation at a number of different wavelengths and correlation with values previously reported in the literature to examine the validity of using a constant Miller's delta, and reported best-fit values for the second-order nonlinear coefficient. Constant Miller's delta provided reasonably accurate wavelength scaling, with good agreement for most materials in recent measurements.

The measurements of the nonlinear optical coefficients of $KNbO_3$, listed in Table 3 are of particular interest. These measurements were performed by the separated-beam wedge technique (Figure 8). With this technique, a large wedge angle is used in a relatively long crystal and with a larger fundamental beam than in the conventional wedge technique. The beam is large enough that many fringes are present in the second-harmonic at the exit surface. The fringe structure causes focused harmonic radiation to diffract into two clearly resolved spots. Integration of an earlier equation, with constant fundamental field amplitude $E_\omega(0)$, provides a simplified expression of the second-harmonic field $E_{2\omega}(z,t)$ in terms of a free wave with propagation constant

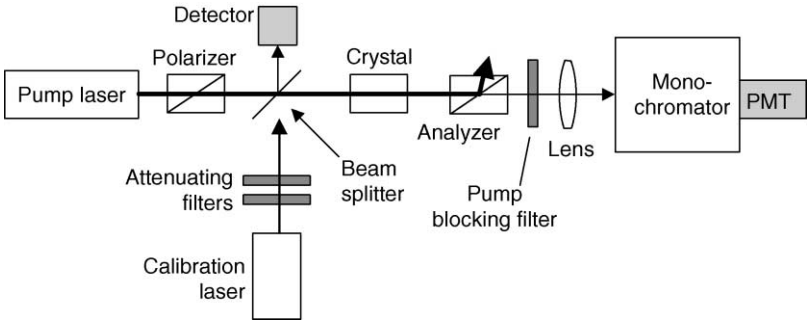


Figure 7 Spontaneous optical parametric fluorescence requires measurement of the ratio of the power of the signal fluorescent to the power of the pump radiation. The calibration laser is used to measure the signal transmission of the filtering system used to block the pump radiation from reaching the photomultiplier tube (PMT).

Table 3 Selected nonlinear optical coefficients

Material	*Point group*	*Nonlinear optical coefficient (pm V^{-1})*		*Wavelength*	*Source*
$AgGaS_2$	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	11.2	10.6 μm	*a*
		$d_{14} = d_{24} = d_{36}$	17.5	1064 nm	*a*
$AgGaSe_2$	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	33	10.6 μm	*a*
BaB_2O_4 (BBO)	3*m*	d_{22}	2.3	1064 nm	*a–f*
		$d_{16} = d_{21} = -d_{22}$	−2.3	1064 nm	*a*
		$d_{15} = d_{31} = d_{24} = d_{32}$	0.04	1064 nm	*d*
		d_{33}	0.04	1064 nm	*d*
CdS	6*mm*	d_{33}	32	10.6 μm	*a*
		$d_{15} = d_{31} = d_{24} = d_{32}$	−16	10.6 μm	*a*
		d_{33}	19.1	1064 nm	*e*
		$d_{15} = d_{31}$	10.1	1064 nm	*e*
		$d_{24} = d_{32}$	10.7	1064 nm	*e*
CdSe	6*mm*	d_{33}	36	10.6 μm	*a*
		$d_{15} = d_{31}\, d_{24} = d_{32}$	−18	10.6 μm	*a*
CdTe	$\bar{4}3m$	$d_{14} = d_{24} = d_{36}$	109	1064 nm	*e*
		$d_{14} = d_{24} = d_{36}$	168	10.6 μm	*c*
$CsLiB_6O_{10}$	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	0.74	1064 nm	*f*
GaAs	$\bar{4}3m$	$d_{14} = d_{24} = d_{36}$	83	10.6 μm	*a*
			170	1064 nm	*e*
GaP	$\bar{4}3m$	$d_{14} = d_{24} = d_{36}$	37	10.6 μm	*a*
			70.6	1064 nm	*e*
KD_2PO_4 (KD^*P)	$\bar{4}2m$	d_{36}	0.37	1064 nm	*a,b*
KH_2PO_4 (KDP)	$\bar{4}2m$	d_{36}	0.39	1064 nm	*a,b,e,g,h*
$KNbO_3$	*mm*2	d_{33}	21.9	1064 nm	*h*
$Z = Z_L = x = l$		d_{15}	13	1064 nm	*h*
$X = Y_L = z = s$		d_{31}	12.4	1064 nm	*h*
$Y = X_L = -y = m$		d_{24}	9.2	1064 nm	*h*
		d_{32}	8.9	1064 nm	*h*
$KTiOAO_4$ (KTA)	*mm*2	$d_{15} = d_{31}$	4.2	1064 nm	*e*
$Z = Z_L = z = m$		d_{15}	2.91	2128 nm	*g*
$X = Y_L = -y = s$		$d_{24} = d_{32}$	2.8	1064 nm	*e*
$Y = X_L = x = l$		d_{33}	16.2	1064 nm	*e*
$KTiOPO_4$ (KTP)	*mm*2	d_{33}	14.6	1064 nm	*e*
$Z = Z_L = z = m$		d_{15}	3.7	1064 nm	*e*
$X = Y_L = -y = s$		d_{31}	3.7	1064 nm	*e*
$Y = X_L = x = l$		d_{24}	1.9	1064 nm	*e*
		d_{32}	2.2	1064 nm	*e*
LiB_3O_5 (LBO)	*mm*2	d_{33}	±0.04	1064 nm	*a,b*
$Z = Z_L = y = s$		$d_{15} = d_{31}$	±0.85	1064 nm	*a,b*
$X = Y_L = z = m$		d_{15}	1.04	1064 nm	*g*
$Y = X_L = x = l$		$d_{24} = d_{32}$	∓0.67	1064 nm	*a,b*
$LiIO_3$	6	d_{33}	4.5	1064 nm	*a,b*
		$d_{15} = d_{31}$	4.4	1064 nm	*a,b*
congruent	3*m*	d_{33}	−27.2	1064 nm	*a,b*
$LiNbO_3$		$d_{31} = d_{15} = d_{32} = d_{24}$	−4.35	1064 nm	*a,b*
		$d_{22} = -d_{21} = -d_{16}$	2.1	1064 nm	*a,b*

Table 3 (*Continued*)

Material	*Point group*	*Nonlinear optical coefficient (pm V^{-1})*		*Wavelength*	*Source*
Agreement of measurements of d_{31} by Maker fringe and wedge SHG (4.6 pm V^{-1}) and parametric fluorescence and parametric amplification (4.3 pm V^{-1})					
				1064 nm	[e]
5%MgO:$LiNbO_3$	3*m*	d_{33}	25	1064 nm	[e]
		d_{31}	4.4	1064 nm	[e]
$LiTaO_3$	3*m*	d_{33}	13.8	1064 nm	[e]
		d_{31}	0.85	1064 nm	[e]
$NH_4H_2PO_4$ (ADP)	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	0.47	1064 nm	[a,e]
αSiO_2	32	d_{11}	0.3	1064 nm	
$ZnGeP_2$	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	69	10.6 μm	[a,b]

[a]Roberts DA (1992) Simplified characterization of uniaxial and biaxial nonlinear optical crystals: a plea for standardization of nomenclature and conventions. *IEEE Journal of Quantum Electronics* 28: 2057–2074.
[b]Dmitriev VG (1997) *Handbook of Nonlinear Optical Crystals*, 2nd edn. Berlin: Springer-Verlag.
[c]Sutherland RL (2003) *Handbook of Nonlinear Optics*, 2nd edn, chs 4 and 5. New York: Marcel Dekker.
[d]Shoji I, Nakamura H, Ohdaira K, *et al.* (1999) Absolute measurement of second-order nonlinear-optical coefficients of β-BaB_2O_4 for visible to ultraviolet second-harmonic wavelengths. *Journal of the Optical Society of America B* 16: 620–624.
[e]Shoji I (1997) Absolute scale of second-order nonlinear-optical coefficients. *Journal of the Optical Society of America B* 14: 2268–2294.
[f]Shoji I, Nakamura H, Ito R, *et al.* (2001) Absolute measurement of second-order nonlinear-optical coefficients of $CsLiB_6O_{10}$ for visible-to-ultraviolet second-harmonic wavelengths. *Journal of the Optical Society of America B* 18: 302–307.
[g]Alford WJ and Smith AV (2001) Wavelength variation of the second-order nonlinear coefficients of $KNbO_3$, $KTiOPO_4$, $KTiOAsO_4$, $LiNbO_3$, $LiIO_3$, β-BaB_2O_4, KH_2PO_4, and LiB_3O_5 crystals, a test of Miller wavelength scaling. *Journal of the Optical Society of America B* 18: 524–533.
[h]Pack MV, Armstrong DJ and Smith AV (2003) Measurement of the $\chi^{(2)}$ tensor of the potassium niobate crystal. *Journal of the Optical Society of America B* 20: 2109–2116.

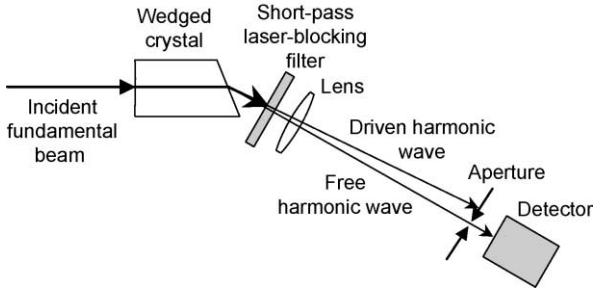


Figure 8 In the separated-beam wedge measurement technique, a fundamental beam of large transverse size generates harmonic radiation that is separated into a free harmonic wave and a driven harmonic wave after exiting the crystal.

$k_{2\omega}$ and a driven wave with propagation constant $2k_\omega$:

$$E_{2\omega}(z,t) = \frac{\omega d_{\text{eff}} E_\omega^2(0)}{n_{2\omega} c} \frac{1}{\Delta k} \{\exp i(k_{2\omega} z - 2\omega t) + c.c.\} - \frac{\omega d_{\text{eff}} E_\omega^2(0)}{n_{2\omega} c} \frac{1}{\Delta k} \{\exp i(2k_\omega z - 2\omega t) + c.c.\} \quad [19]$$

The two beams with different propagation constants are refracted at different angles at the wedged exit surface. The detailed analysis must include losses due to absorption and reflections at the surfaces and also nonlinear effects at the surfaces. Precision again is required in characterizing the fundamental beam and incident and generated intensities or energies as well as the geometry of the measurement. The large wedge eliminates problems of multiple reflections, and it is no longer necessary to carefully choose the beam spot size to fit within one fringe maximum, yet be large enough to allow use of the near field approximation. The accuracy of this measurement is demonstrated when the coefficients are expressed as Miller-delta values: $\delta_{24} = 0.153\,\text{pm V}^{-1}$, $\delta_{32} = 0.150\,\text{pm V}^{-1}$, $\delta_{15} = 0.195\,\text{pm V}^{-1}$, and $\delta_{31} = 0.192\,\text{pm V}^{-1}$. These values satisfy Kleinman symmetry, $\delta_{24} = \delta_{32}$ and $\delta_{15} = \delta_{31}$, well within the reproducibility of the measurement, which was about ± 3%.

The nonlinear optical coefficients of BBO $d_{22}(1064\text{ nm}) = 2.3\text{ pm V}^{-1}$ and KDP $d_{36}(1064\text{ nm}) = 0.39\text{ pm V}^{-1}$ are gaining agreement from a number of independent measurements. The apparent disparity in the values for GaAs $d_{36}(10.6\,\mu\text{m}) = 83\text{ pm V}^{-1}$ and $d_{36}(1064\text{ nm}) = 170\text{ pm V}^{-1}$, is resolved when the coefficients are converted to Miller's delta values: $\delta_{36} = 0.089\text{ pm V}^{-1}$ and $\delta_{36} = 0.086\text{ pm V}^{-1}$, respectively. The same thing can be done for GaP, for which $d_{36}(10.6\,\mu\text{m}) = 37\text{ pm V}^{-1}$ and $d_{36}(1064\text{ nm}) = 70.6\text{ pm V}^{-1}$ convert to $\delta_{36} = 0.085\text{ pm V}^{-1}$ and $\delta_{36} = 0.077\text{ pm V}^{-1}$, respectively. However, the older and newer values for other materials such as CdTe and CdS do not achieve this agreement. Other materials such as CLBO and LBO need additional investigation to resolve differences. The measurement of nonlinear optical coefficients is a continuing process.

Other techniques for measurement involving different nonlinear-optical processes are possible. Parametric amplification was mentioned above. Generally more complicated processes are less well suited for measurement of fundamental properties. Integrated-optical devices such as waveguides are even more complex. Often nonlinearity is measured in fractional conversion per unit of pump power to characterize overall device performance. There are survey techniques for evaluating materials such as grinding the material to a powder, illuminating with intense laser radiation, and looking for harmonic conversion. A surprising amount of information, such as a rough estimate of nonlinearity and the possibility of birefringent phase matching, can be gleaned by sorting the power and measuring the scattered harmonic radiation. Second-harmonic reflections from surfaces and thin film can be used in some instances. Electric fields can be use to orient molecules in polymers and solute molecules in solution to measure the hyperpolarizability of individual molecules.

Reporting Frames

Standardized axis specifications avoid confusion, but nonstandard axes can prove convenient in some applications provided care is used. The crystal $KNbO_3$ has *mm*2 point group; it could also be *m*2*m* or 2*mm* depending on the choice of axes to describe the crystal properties. The axis assignment for the orthorhombic class *mm*2, as it is used in Table 3, follows the IEEE/ANSI Std. 176-1987: Z is the polar axis, X is parallel to the shorter of the remaining unit cell dimensions, and Y is chosen to complete a right-handed coordinate frame. Polarity of the axes can be assigned by pyroelectric properties and piezoelectric properties of the crystal. The common assignments of the axes in the literature are given by X_L, Y_L, Z_L in the table. Reference to the biaxial indices of refraction are given by the lower case x, y, z with $n_x < n_y < n_z$. Finally the axes are identified by the unit cell dimensions $s < m < l$. Either a plus or minus rotation about the Y axis will bring the crystal into an *xyz* frame. However, a $+90$ degree rotation will give coefficients that are opposite in sign from a -90 degree rotation. The signs of the nonlinear optical coefficients are often not well determined because sign is not obtained from conversion efficiency in a single crystal. However, coefficient signs can be compared in the same crystal or between two crystals to determine if they are the same or opposite.

The crystals KTP or KTA are used here as an example. A 90° rotation about the Z-axis gives a new coordinate system $X'Y'Z'$ aligned with the *xyz* frame determined by the indices of refraction. In the new frame, the component of the nonlinear optical tensor, in terms of the standard frame, are $d_{zxx} = d_{32}$ and $d_{zyy} = d_{31}$. All that is needed here is to exchange d_{32} and d_{31} to change the standard frame to the one commonly used in the literature, X_L, Y_L, Z_L. When the largest index of refraction is not aligned with the polar axis, as is the case for LBO and $KNbO_3$, the specification of tensor components can become confusing if a standard frame is not used.

Often crystallographers and crystal manufacturers can identify axes by surface morphology or by poling processes of ferroelectric crystals. With other crystal symmetries a change in the reporting frame that does not change linear optical properties may change the sign of some of the nonlinear optical tensor components while leaving others unchanged. For a crystal of *3m* point group, a 180° rotation about Z will change the sign of d_{22}, d_{21}, and d_{16} and leave the other coefficients unchanged. The *xyz* frame with $n_x < n_y < n_z$ is convenient for calculations of optical propagation, and the standard XYZ frame is unambiguous for reporting of tensor properties.

Conclusion

There is still a range of values for nonlinear optical coefficients of many materials. Standardization is evolving on more commonly used materials. Greater precision in measurement and extension of measurement techniques are providing better accuracy. However, caution is still required when accessing values. It is necessary to scrutinize the sources of values to obtain the highest accuracy. Accurate measurements need to be extended to difficult spectral ranges of the ultraviolet and infrared. The ambiguity of the sign of tensor components needs to be resolved. In recent measurements, carefully executed wedge measurements are producing more reliable results. The separated beam wedge technique appears to be particularly promising. Phase-matched second-harmonic generation and other practical applications of nonlinear optical materials will continue to serve as techniques to verify accuracy. Current analytical and numerical techniques provide a method for precise modeling of nonlinear optical processes and the need for more accurate material characterization.

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Nonlinear Optics, Basics:** $\chi^{(2)}$-Harmonic Generation.

Further Reading

Alford WJ and Smith AV (2001) Wavelength variation of the second-order nonlinear coefficients of $KNbO_3$, $KTiOPO_4$, $KTiOAsO_4$, $LiNbO_3$, $LiIO_3$, β-BaB_2O_4, KH_2PO_4, and LiB_3O_5 crystals, a test of Miller wavelength scaling. *Journal of the Optical Society of America B* 18: 524–533.

Dmitriev VG (1997) *Handbook of Nonlinear Optical Crystals*, 2nd edn. Berlin: Springer-Verlag.

Gehr RJ (1998) Separated-beam non-phase-matched second-harmonic method of characterizing nonlinear optical crystals. *Journal of the Optical Society of America B* 15: 2109–2116.

Kurtz SK (1975) Measurement of nonlinear optical susceptibilities. In: Rabin H and Tang CL (eds) *Quantum Electronics: A Treatise, vol. 1, part A*, pp. 209–281. New York: Academic Press.

Pack MV (2003) Measurement of the $\chi^{(2)}$ tensor of the potassium niobate crystal. *Journal of the Optical Society of America B* 20: 2109–2116.

Roberts DA (1992) Simplified characterization of uniaxial and biaxial nonlinear optical crystals: a plea for standardization of nomenclature and conventions. *IEEE Journal of Quantum Electronics* 28: 2057–2074.

Shoji I (1997) Absolute scale of second-order nonlinear-optical coefficients. *Journal of the Optical Society of America B* 14: 2268–2294.

SNLO nonlinear optics code available from Smith AV Sandia National Laboratories, Albuquerque, NM 87185-1423. (an Internet publication) http://www.sandia.gov/imrl/XWEB1128/xxtal.htm.

Sutherland RL (2003) *Handbook of Nonlinear Optics*, 2nd edn, chs 4 and 5. New York: Marcel Dekker.

$\chi^{(3)}$

P P Banerjee, University of Dayton, Dayton, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Over the past 30 years, several methods have been used to measure and characterize optical nonlinearities. With increasing interest in nonlinear optics beyond the traditional harmonic generation, third-order nonlinear processes have been shown to be responsible for self-focusing and defocusing of optical beams, phase conjugation, optical bistability, etc. Third-order nonlinear processes in Kerr materials arise from the cubic dependence of the polarization $P = P^{L} + P^{NL}$, $P^{NL} = \chi^{(3)}{:}E{\cdot}E{\cdot}E$ in a material on the optical field E, through the nonlinear susceptibility tensor $\chi^{(3)}$. The cubic nonlinearity in turn results in a nonlinear refractive index coefficient n_2 defined through:

$$n = n_0 + n_2|E_p^{\omega_0}|^2 \qquad [1]$$

where $E_p^{\omega_0}$ is the phasor part of E : $E = Re[E_p^{\omega_0}\exp \times j\omega_0 t]$. The relation between $\chi^{(3)}$ and n_2 can be expressed as:

$$n_2 = (3/8n_0)\chi^{(3)}_{1111}(-\omega_0; \omega_0, \omega_0, -\omega_0) \qquad [2]$$

In some materials, an effective $\chi^{(3)}$ or an effective n_2 can be defined as well. For instance, in a second-order nonlinear material, an effective $\chi^{(3)}$ is established through cascading of $\chi^{(2)}$s, and in the presence of phase mismatch between the fundamental and the second harmonic. In other materials, such as liquid crystals, an effective n_2 can be defined due to the nonlinear dependence of the reorientation of the director axis on the optical field. In photorefractive materials, the induced nonlinearity stems from the nonlinear dependence of the induced space charge distribution in the material on the optical field. Some of the techniques used to measure the third-order nonlinearity in materials include degenerate four-wave mixing (DFWM) or phase conjugation, interferometry, ellipse rotation, and beam self-refraction and/or distortion and bending.

For instance, in DFWM, two contradirectional propagating pumps and the probe beam having the same frequency ω_0 interact to give rise to the conjugate, whose amplitude depends on the magnitude of the cubic nonlinearity coefficient. The conservation of energy and momentum describes a parametric interaction of the form $\omega_0 + \omega_0 - \omega_0 \rightarrow \omega_0$ (in frequency space) and, $\mathbf{k}_{probe} + \mathbf{k}_{pump1} - \mathbf{k}_{pump2} \rightarrow \mathbf{k}_{conj}$ (in $\mathbf{k}$-space). The conjugate beam is clearly distinguishable by its spatial separation from the other interacting waves. However, alignment of the three waves is critical to obtain a good phase conjugate; furthermore, the sign of the nonlinearity cannot be determined.

In the interferometric method, a laser beam is split into two parts, one serving as the reference, and the other being incident on the nonlinear sample. As in the case of interferometers, if the two beams are then recombined and introduced through a lens, an

interference pattern forms at the focal plane of the lens. While the information about the magnitude of the nonlinearity can be found from the fringes and their change when the incident power is varied, information of the sign of the nonlinearity is still not readily inferable.

In this article, we will concentrate on self-refraction, particularly in light of the fact that it results in a sensitive and at the same time simple method called the *z*-scan©, requiring the propagation of a single beam through the nonlinear sample. We comment that self-diffraction does also cause self-bending of an asymmetric beam due to the asymmetric phase change induced through the cubic nonlinearity; however, extracting the information of the sign and magnitude of the nonlinearity can be effectively done only if the sample is thin, implying that there is negligible linear diffraction of the beam as it passes through the sample. In the rest of the article, we will therefore concentrate on the *z*-scan method, since it is an elegant and simple single (symmetrical) beam technique that yields information on the magnitude and sign of the cubic nonlinearity, as well as its real and imaginary parts. We will describe in detail how the *z*-scan can measure the real part of the cubic nonlinearity, for both thin and thick samples, as well as showing its application in determining the reorientational nonlinearity of liquid crystals and the nonlinearity of photorefractive materials, where photovoltaic effect and diffusion of charge carriers contribute to the effective nonlinearity.

Since the mid-1990s *z*-scan has evolved as a standard method for determining effective nonlinear refractive index coefficients of nonlinear materials. The name *z*-scan is derived from the fact that the nonlinear test sample is scanned along the direction of propagation (z) of a beam about the back focal plane of an external lens illuminated by an incident Gaussian beam from a laser. The far-field diffraction pattern yields information about the nonlinearity of the test material. The *z*-scan technique has been used to measure and characterize the nonlinear refractive index coefficient and nonlinear absorption coefficient in materials such as CS_2, ZnSe, Cs atoms, multiphoton absorption in GaN, organic dyes, thermal nonlinearities in liquids, liquid crystals, photorefractive materials, etc., to name a few. *z*-scan measurements have been made using non-Gaussian incident beams, and *z*-scans have been performed using both CW and pulsed laser sources.

The first extensive series of work on the *z*-scan technique, including analysis and application to the measurement of the cubic nonlinearity, was performed by the group of Van Stryland at the Center of Electro-Optics and Lasers (CREOL) at the University of Central Florida. Other investigators later performed similar measurements and characterization during the study of nonlinear optical properties of a solution of chlorophyll in acetone, and Chinese tea.

Theory of the Thin Sample *z*-Scan

We will first analyze propagation of a (circularly symmetric) Gaussian beam which is initially focused by an external lens and is then incident on a 'thin' slice of a cubically nonlinear Kerr medium. This case is important because, as explained below, it provides a basis for the measurement of the nonlinear refractive index coefficient n_2.

Figure 1 shows a simple ray diagram to demonstrate what happens if the cubically nonlinear material is thin. The propagating Gaussian beam induces its own lens, which is positive for a material having a positive n_2, for reasons that will become clear to the reader later on in this section. Note that if the position of the induced lens is at the back focal plane of the external lens, the former does not affect the propagation of the beam behind the nonlinear material. The situation is thus the same as the propagation of the externally focused Gaussian beam in a linear diffraction-limited environment. However, if the separation between the external lens and the sample is less than the focal length f of the external lens, the beam in the far field is wider than the linear diffraction-limited case. Conversely, if the lens-sample separation is larger than f, the beam width is smaller. This is the basis of the *z*-scan: for a material with a positive n_2, the scanning of a thin nonlinear sample through the back focal plane of the external lens, starting from a position to the left of the focus, yields far-field profiles which are initially larger and eventually smaller than the width in the linear diffraction-limited case. The opposite is true for materials with a negative n_2. For very large displacements of the sample from the back focal plane of the

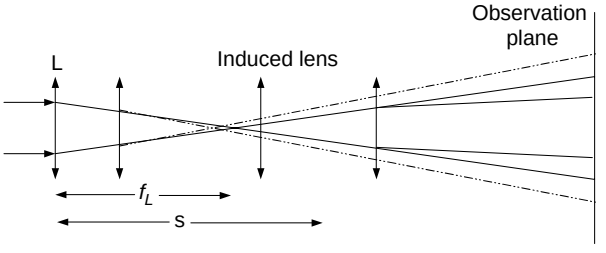


Figure 1 Geometrical optics picture of *z*-scan setup for nonlinear refractive index n_2. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

external lens, the beam power is small enough that nonlinear lensing is negligible, and hence the beam widths return to the linear diffraction-limited case. The variation of the on-axis intensity or transmittance as a function of scan distance is complementary to the variation of the width, due to conservation of power, to be explained further below. A sketch of the on-axis transmittance during a z-scan for a positive n_2 material, is shown in Figure 2.

The simpler ray picture described above can be augmented by using a q parameter approach to Gaussian beam propagation. For the sake of simplicity, we prefer to introduce the q parameter approach here, instead of the more rigorous Gaussian decomposition of the induced phase as computed by Van Stryland's group. The q of a Gaussian beam in air is defined through the relation $1/q = 1/R - \mathrm{i}(2/k_0 w^2)$, where R and w denote the radius of curvature and the beamwidth, respectively, and k_0 is the propagation constant in air. Assume a Gaussian beam with a waist w_0 (implying initially planewave fronts) incident on the external lens. The Gaussian can be described by $q = q_0 = \mathrm{i}z_{R0}$, where $z_{R0} = k_0 w_0^2/2$ is called the Rayleigh length of the incident Gaussian. The laws of q transformation of a Gaussian beam are:

(i) for translation through a distance d, the q changes to $q + d$; and
(ii) for a Gaussian with q-parameter q incident on a lens, the q immediately behind the lens, is given by $(1/q - 1/f)^{-1}$.

Using these laws, the $q = q_{\mathrm{obs}}$ of the Gaussian at a distance d from the nonlinearly induced lens of the focal length f_{ind}, which is located at a distance $s = f + \Delta s$ behind the external lens, can be found as:

$$q_{\mathrm{obs}} = \frac{f_{\mathrm{ind}}[q_0 f + (f - q_0)(f + \Delta)]}{f_{\mathrm{ind}}(f - q_0) - [q_0 f + (f - q_0)s]} + d \qquad [3]$$

In deriving eqn [3], we have used the laws of q transformation on the initial Gaussian beam of q-parameter q_0 passing through the external lens, being translated a distance $f + \Delta s$, passing through the induced lens, and finally being translated a distance d to the observation plane.

To determine $f_{\mathrm{ind}}(\Delta s)$, assume a Gaussian of the form:

$$|E_{\mathrm{p}}^{\omega_0}(x, y, z)| = E(z)\exp(-(x^2 + y^2)/w^2(z)) \qquad [4]$$

Since

$$n = n_0 + n_2|E_{\mathrm{p}}^{\omega_0}|^2 \approx n_0 - 2n_2E^2(z)(x^2 + y^2)/w^2(z) \qquad [5]$$

the nonlinearly induced phase change is

$$\Delta\phi = k_0 n\Delta z \approx k_0[n_0 - 2n_2E^2(z)(x^2 + y^2)/w^2(z)]\Delta z \qquad [6]$$

so that

$$f_{\mathrm{ind}}(z) = n_0 w^2(z)/4n_2a^2(z)\Delta z \qquad [7]$$

Note that $f_{\mathrm{ind}}(z)$ is inversely proportional to $n_2E^2(z)$ and Δz. The longer the sample length, the more the nonlinear induced phase, implying a stronger lensing (smaller f_{ind}). Also, the stronger the nonlinearity parameter n_2 and/or on-axis intensity ($E^2(z)$), the stronger the focusing since the change in n is larger.

We must point out that the focal length f_{ind} of the induced lens is also a function of Δs, since it depends on the beam intensity at the location of the nonlinear sample. In fact, the focal length f_{ind} of the nonlinearly induced lens can be expressed in terms of the beam power P, the beam width w at the position Δs, f (the focal length of the external lens), the linear refractive index n_0, the nonlinear refractive index coefficient n_2 defined through eqn [1] where $|E_{\mathrm{p}}^{\omega_0}| = E$ is the optical field in the nonlinear material of length L, assumed 'thin' for now. The term 'thin' means that the sample length is smaller than the depth of focus of the

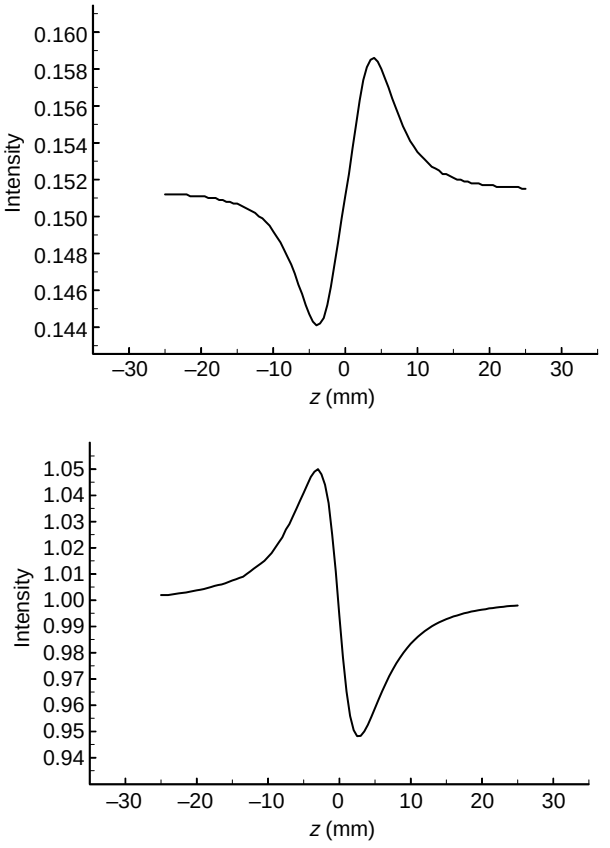


Figure 2 Typical z-scan plots for positive (top) and negative (bottom) cubic nonlinearities. The vertical axis is in arbitrary units.

externally focused Gaussian beam. In this case, we can assume that the beam width is a constant through the sample, and the induced focal length is then only directly dependent on the total thickness L of the sample. The focal length of a nonlinearly induced 'thin' Kerr lens can be written as

$$f_{\text{ind}}(\Delta s) = \frac{n_0 w^2(\Delta s)}{4n_2|E|^2 L} \approx \frac{\pi f^4 n_0}{k_0^4 w_0^4 \eta_0 P L n_2} \qquad [8]$$

for small Δs, where η_0 is the (linear) characteristic impedance of the nonlinear medium. Note that in eqn [8], the beam width w is given by

$$w^2(\Delta s) = w_{\text{f}}^2\left[1 + \frac{4(\Delta s)^2}{k_0 w_{\text{f}}^4}\right] \text{ and } w_{\text{f}} \cong \frac{2f}{k_0 w_0} \qquad [9]$$

In eqn [9], w_{f} denotes the waist size of the Gaussian beam at the back focal plane of the external lens. In deriving eqn [8], one has to use the power conservation relation $|E|^2 w^2 = (4/\pi)\eta P$ to re-express the dependence of the induced focal length on the field amplitude in terms of the corresponding width at the same point.

Our final interest is to find the imaginary part of $1/q_{\text{obs}}(= -2/k_0 w_{\text{obs}}^2)$, in order to ascertain the width w_{obs} of the Gaussian beam on the observation plane. Simple but tedious algebra leads us to the relation:

$$\frac{\frac{k_0 w_{\text{obs}}^2}{2}}{= \frac{[(ff_{\text{ind}}d - f^2 d + f^2 f_{\text{ind}}) + f(f_{\text{ind}} - d)\Delta s]^2 + z_{R0}^2[(f_{\text{ind}} - d)\Delta s + f_{\text{ind}} d]^2}{z_{R0} f^2 f_{\text{ind}}^2}} \qquad [10]$$

It can be shown simply that when $\Delta s = 0$, and for $|f_{\text{ind}}| \gg f$, $w_{\text{obs}} = w_0 d/f$ for large d. In other words, in the linear case, i.e., when $f_{\text{ind}} = \infty$, as well as in the nonlinear case when $\Delta s = 0$, the far-field width of the Gaussian beam is what is predicted by geometrical optics. Furthermore, from eqn [10], the rate of change of the width with Δs around $\Delta s = 0$, (assuming $z_{R0} \gg f$, and for large d) can be found as

$$\frac{\text{d}w_{\text{obs}}}{\text{d}\Delta s} = \frac{-2d^2 z_{R0}}{k_0 w_{\text{obs}} f^2 f_{\text{ind}}} = -\frac{w_0 d}{f f_{\text{ind}}} \qquad [11]$$

provided we invoke the relation for w_{obs} derived above. In eqn [11], f_{ind} refers to the value of the induced focal length (eqn [8]) with $\Delta s = 0$. Equation [11] may be written as

$$\frac{\text{d}w_{\text{obs}}}{\text{d}\Delta s} = -\frac{k_0^4 w_0^5 P L n_2}{2\pi n_0 f^5} \qquad [12]$$

This means that in a material with a positive (negative) n_2, implying a positive (negative) induced lens, the rate of change of the observed width with the position of the nonlinear sample about the focal plane of the external lens will be negative (positive), in agreement with the results from geometrical optics. Furthermore, the rate of change is proportional to the angle of convergence w_0/f of the externally focused Gaussian, and to the position d of the observation plane.

In z-scan, it is often important to relate the on-axis transmittance in the far field to the scan distance s for measurement purposes. Since for power conservation, the intensity is inversely proportional to the width of the beam in the far field, a positive n_2 would imply a positive slope as a function of scan distance with increasing s, and a negative slope for negative n_2. The value of the slope of the on-axis transmittance could give the information of the n_2 of the material. Note that the z-scan graph in this case is an odd function of the scan distance, measured about the back focal plane of the external lens.

In practice, a detector with a finite aperture area is placed on-axis, and the detected power monitored as a function of scan distance. The transmitted power through the aperture is obtained by

$$\begin{aligned} P_{\text{T}} &= \frac{1}{2}\int_0^{2\pi}\int_0^{r_{\text{a}}} \frac{E_{\text{a}} E_{\text{a}}^*}{\eta_0}\text{d}A \\ &= \frac{\pi w_0}{2}\left(\frac{E_0^2}{2\eta_0}\right)\left(1 - \exp\left(-\frac{2r_{\text{a}}^2}{w_{\text{obs}}^2}\right)\right) \\ &= P\left(1 - \exp\left(-\frac{2r_{\text{a}}^2}{w_{\text{obs}}^2}\right)\right) \end{aligned} \qquad [13]$$

where r_{a} is the radius of aperture. The rate of change of P_{T} with Δs can be found as

$$\frac{\text{d}P_{\text{T}}}{\text{d}\Delta s} = \frac{-4r_{\text{a}}^2 P}{w_{\text{obs}}^3}\exp\left(-\frac{2r_{\text{a}}^2}{w_{\text{obs}}^2}\right)\frac{\text{d}w_{\text{obs}}}{\text{d}\Delta s} \qquad [14]$$

Substituting eqn [12] into eqn [14] and using the approximation $w_{\text{obs}} = w_0 d/f$, eqn [14] can be rewritten as

$$\frac{\text{d}P_{\text{T}}}{\text{d}\Delta s} \cong \frac{4r_{\text{a}}^2 f^2 P}{f_{\text{ind}} d^2 w_0^2}\exp\left(-\frac{2r_{\text{a}}^2 f^2}{d^2 w_0^2}\right) \qquad [15]$$

from which

$$\frac{\text{d}P_{\text{T}}}{\text{d}\Delta s}(\Delta s \approx 0) \cong \frac{4k_0^4 w_0^2 \eta_0 r_{\text{a}}^2 P^2 L}{\pi f^2 d^2 n_0}\exp\left(-\frac{2r_{\text{a}}^2 f^2}{d^2 w_0^2}\right) n_2 \qquad [16]$$

When r_a approaches 0;

$$\frac{dP_T}{d\Delta s}(\Delta s \approx 0) = 4\left(\frac{k_0^2 w_0 r_a P}{fd}\right)^2 \frac{\eta_0 L}{\pi n_0} n_2 \qquad [17]$$

Therefore, n_2 can be estimated as

$$n_2 = \left(\frac{\pi n_0}{\eta_0 L}\right)\left(\frac{d}{2k_0^2 w_0 P}\right)^2 \left(\frac{f}{r_a}\right)^2 \frac{dP_T}{d\Delta s} \qquad [18]$$

Thus, if one knows the slope of the z-scan graph around $\Delta s \approx 0$, the value of n_2 can be found. Another way to state this result is to say that the n_2 is related to the peak-to-valley change in the on-axis transmittance ΔT_{p-v}, which is proportional to $dP_T/d\Delta s$.

In passing, we comment that the z-scan method can be also adapted to determine the nonlinear absorption coefficient α_2 of a material, defined through the relation $\alpha = \alpha_0 + \alpha_2 I$, where α is the total absorption coefficient and I denotes the intensity. In order to achieve this, the aperture is assumed to be completely open, in order to collect all the light behind the nonlinear absorptive sample. For $\alpha_2 > 0$, maximum absorption is experienced when the nonlinear sample, assumed thin, is at the back focal plane of the lens since the optical intensity is a maximum due to the external focusing. The z-scan graph is an even function of scan distance, measured about the back focal plane of the external lens.

Thick Sample *z*-Scan

Contrary to the thin sample case, a thick sample is defined as one whose thickness may be comparable to the depth of focus of the externally focused Gaussian beam. In this case, the beamwidth changes during propagation through the sample, so that the induced focal length also changes with propagation. This makes the problem substantially more involved. Indeed, to track the far-field profile in this case, one needs to describe the propagation of the Gaussian beam inside the nonlinear sample using the q parameter formalism. Since both the width and the radius of curvature of the beam at the exit face of the sample are required to predict the far-field behavior, one can write a simple differential equation describing the variation of the q through the nonlinear sample. This equation can be composed from the laws of q-transformation described earlier, and the effects of propagation and nonlinear lensing can be added assuming each is small over an infinitesimal distance of propagation, much like the numerical method called the split step beam propagation method (BPM), often used to analyze such problems. We will not present details of the computation, rather show calculation results for a typical thick sample case, and point out differences with the thin sample predictions for z-scan.

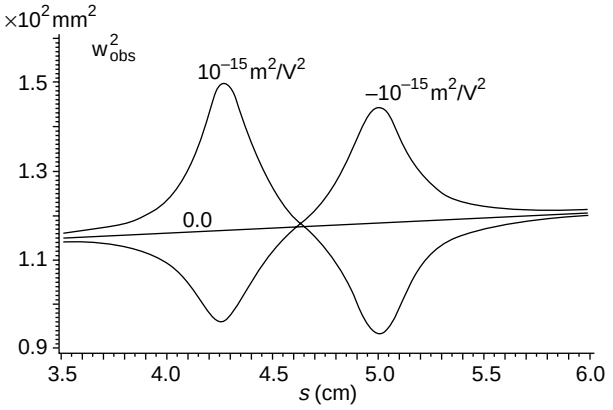


Figure 3 Plot of width of the beam in the far field as a function of the scan distance s. The light wavelength is 640 nm. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications*. New York: Marcel Dekker.

Figure 3 shows the beamwidth as a function of scan distance for the case where the observation plane is at a distance $d = 1$ m from the exit face of a 1 cm thick. Assume that a hypothetical sample with $n_0 = 1$, $n_2 = \pm 10^{-15}$ m^2/V^2 is scanned about the back focal plane of an external lens with focal length $f = 5$ cm. For comparison, the case of a linear sample is also superposed on the same graph. Note that with the variation of s, the far-field width initially increases, and then decreases before starting to increase again, for a material with a positive n_2. Note that around the point where the variations intersect the graph for the linear case, the slope of the curves are negative and positive for materials with positive and negative n_2, respectively. The graphs for the variation of on-axis transmittance are opposite to the variation of the widths, as mentioned earlier. This is also in agreement with the results predicted from geometrical optics and the q formalism for a thin sample, described earlier.

How do thick and thin sample z-scans differ? Important differences are: (i) the position of the points where the graphs for positive and negative, $n_2 s$, intersect the graph for the linear case; and (ii) the slope of the graphs, predicted from thin and thick sample calculations. For instance, for the thin sample approximation, the value of the slope around $s = f$ is about -4 for a material with a $n_2 = 10^{-15}$ m^2/V^2. Recall from the previous section, the z-scan plot for a thin sample always intersects the graph for the linear case at the back focus of the external lens,

and therefore, it is easy to determine the slope of the z-scan at this point. From the graph in the thick sample case, it is perhaps difficult to ascertain at exactly which point one must determine the slope. A figure of merit would be the minimum value, which occurs for a value of s slightly smaller than $f = 5$ cm. A conservative estimate yields a slope of approximately -1 for this particular example. Note, however, that in the thick lens analysis, we have taken a paraxial approximation to the Gaussian at every stage of its propagation through the nonlinear sample. Strictly speaking, this is not true, and to alleviate this problem, Van Stryland's group included a correction factor with the width of the Gaussian beam.

We would like to mention that values of $n_2 \approx -10^{-15}$ m^2/V^2 can indeed be measured for thermal nonlinearities in liquids. In fact, this is a typical value that we measured for our solution of chlorophyll in ethanol during our first experiments on z-scan. Values of effective $n_2 \approx 10^{-14}$ m^2/V^2 can be measured for liquid crystals with orientational nonlinearities, and will be described below.

Example of z-Scan: Effective Cubic Nonlinearity of Liquid Crystals

The induced refractive index in a nematic liquid crystal, due to reorientational nonlinearity, is a function of position in the crystal and depends on the intensity of the incident optical plane wave, its angle with respect to the direction of propagation, and on the applied voltage. Because of the fact that the nonlinearity depends on the direction of propagation of the planewave and the director axis of the crystal, modeling Gaussian beam propagation is rather complicated. Moreover, the director axis has a certain distribution in the material due to the applied voltage, and hence, the electrostatic field in the material, as well as the optical field. This poses a level of complexity a degree higher than that discussed above for the simple thick sample case. Since the induced refractive index is not a constant, and more importantly, a function of the spatial frequency of the optical beam (which determines the angle between a planewave component of the Gaussian beam spectrum and the director axis distribution under external field), one is compelled to analyze optical propagation using split-step BPM. A typical set of simulated z-scan plots is shown in Figure 4, assuming one transverse dimension for simplicity, and the same parameters as in the experiment described below. Actual experimental results from a z-scan of a biased liquid crystal sample is shown in Figure 5 for transmitted power versus

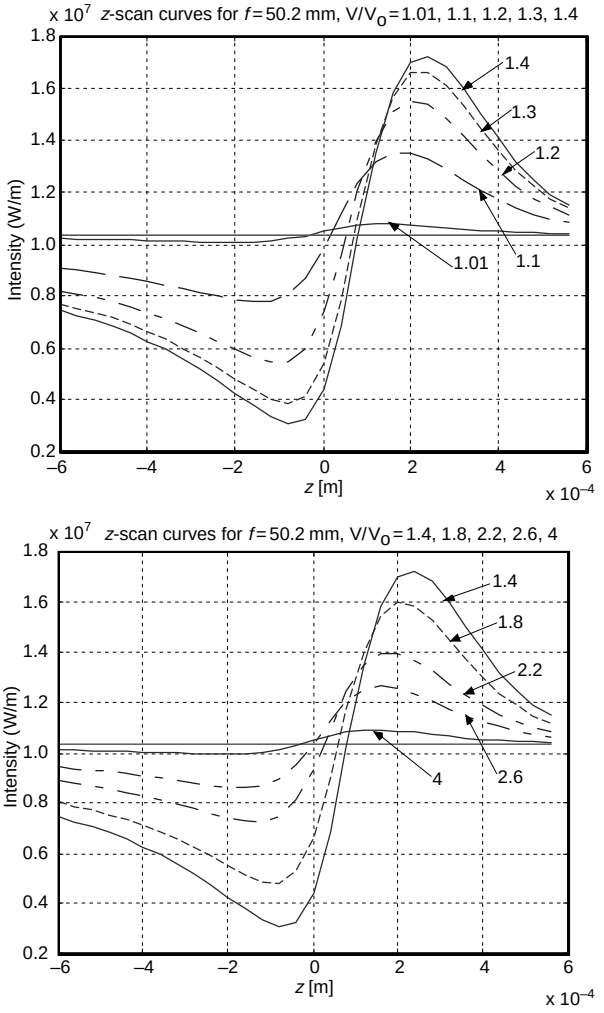


Figure 4 z-scan simulation results for $f = 50.2$ mm and $V/V_o = 1.01$, 1.1, 1.2, 1.3, 1.4 with $E_{op} = 2 \times 10^5$ V/m. z-scan simulation results for $f = 50.2$ mm and $V/V_o = 1.4$, 1.8, 2.2, 2.6, 4 with $E_{op} = 2 \times 10^5$ V/m. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

sample position in a z-scan experiment for a liquid crystal sample for applied voltages 0.8, 1.0, 1.2, 1.8, and 2.2 V. In this case, the focal length of the external lens is approximately 50 mm and aperture diameter on the observation plane is 1 mm. The scan range is 0.15 inches or approximately 4 mm. The scan step, ΔZ, is varied from about 0.002 inches to about 0.01 inch, depending on the position of the sample. This is done so as to get the maximum information around the back focal plane of the external lens where the measurement values are to be used to find the slope of the z-scan graph, and hence the effective nonlinear refractive index coefficient.

The plots of the effective $n_2 = n_{2\text{eff}}$ (simulated and measured) are shown in Figure 6a and b, respectively. As seen from both plots, the $n_{2\text{eff}}$ increases and

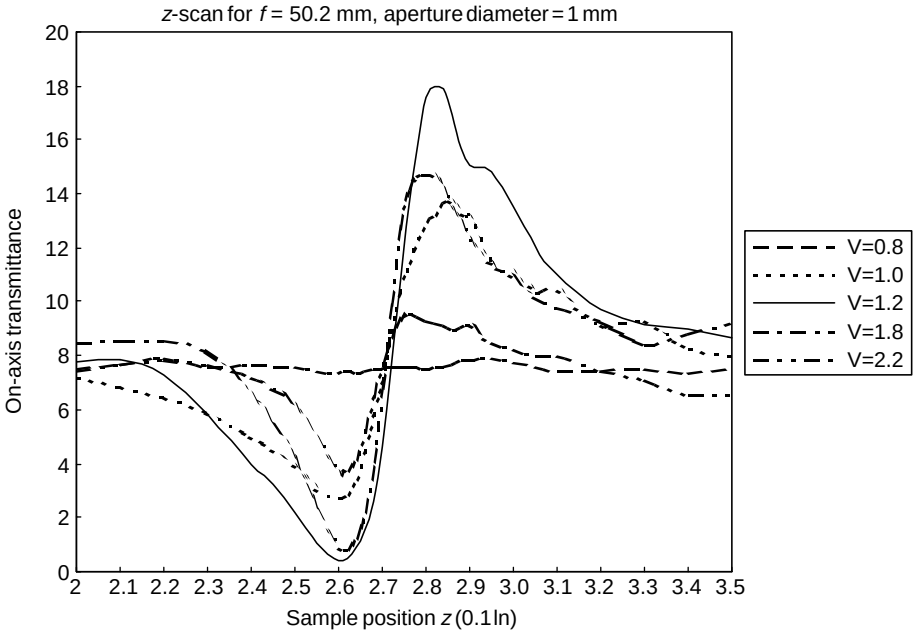


Figure 5 Measured *z*-scan graph for $f = 50.2$ mm and aperture diameter $= 1$ mm. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

reaches a maximum at $V/V_o = 1.4$, before starting to decrease. Simulated and experimental $n_{2\text{eff}}$ have similar trends. Furthermore, there is reasonable agreement between the calculated and measured values of $n_{2\text{eff}}$.

Discussion: Effective n_2 of Other Materials

The *z*-scan is a powerful tool in the measurement and characterization of effective cubic nonlinearities of different materials. The method is simple and easy to set up, and can be readily automated. We provide two other examples of *z*-scans below, as examples. For instance, photorefractive materials are commonly used for dynamic holography, in order to record and reconstruct volume holograms. Photorefractive materials are doped light-sensitive electro-optic materials in which optical radiation creates mobile charge carriers which redistribute themselves to create an induced space-charge field. This electrostatic field in turn modulates the refractive index which phase modulates the optical beam. A figure of merit for the effectiveness of such a material, e.g., lithium niobate, for holographic storage, is the donor to acceptor ratio in such materials. It turns out that this ratio, along with the photovoltaic coefficient, can be related to an effective n_2 in lithium niobate. This is also responsible in producing an elliptic beam in the far field by starting out from a circularly symmetric Gaussian beam focused onto the sample. The ellipticity stems from the tensor nature of the electro-optic coefficients. Typical simulated *z*-scan plots for lithium niobate are shown in Figure 7, and carries the signature variation of the scan in the variation of the beam ellipticity with scan distance. From this plot, we are able to determine the effective n_2 and hence, the donor to acceptor concentration ratio.

In other photorefractive materials, such as barium titanate, in which diffusion of charge carriers plays a dominant effect, the induced change in the refractive index is proportional to the gradient of the intensity, rather than the intensity itself, as was the case in all previous examples. In this case, as seen in Figure 8, it can be shown that the *z*-scan plot for on-axis transmittance with a small finite aperture on the observation plane is an even function of scan distance. This, however, should not be confused with the open aperture *z*-scan that is also even, but which yields information of the nonlinear absorption coefficient, as described earlier. Experiments with a new material, manganese aluminum oxide, which is thought to have both photovoltaic and diffusive contribution to space charge field distribution, yields a *z*-scan that has both odd and even components. Through careful analysis of the *z*-scan results, one is able to determine the relative contributions of each of these effects. Care should be taken in the interpretation of the results in this case, since in an actual experiment, the aperture size cannot be infinitely small, and thus, an optimum aperture size must be determined that allows for sufficient detected power,

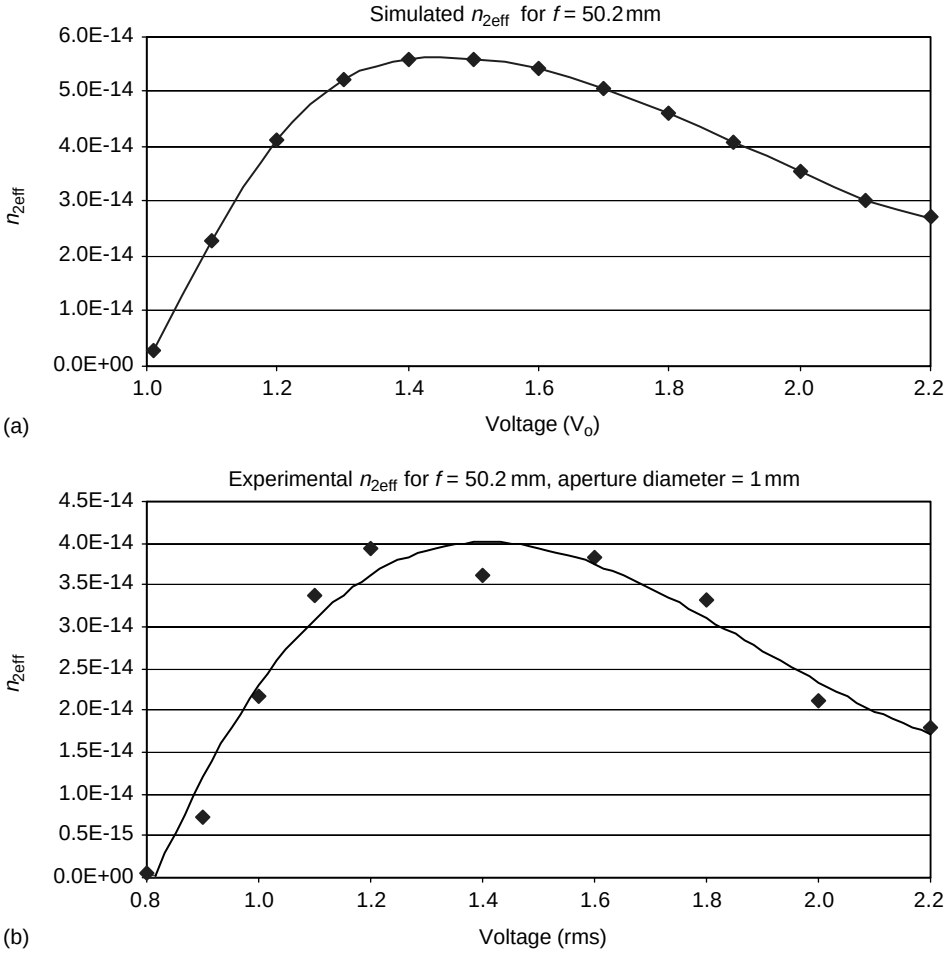


Figure 6 (a) Calculated n_{2eff} for $f = 50.2$ mm with $E_{op} = 2 \times 10^5$ V/m; (b) Experimental n_{2eff} for $f = 50.2$ mm with aperture diameter = 1 mm. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

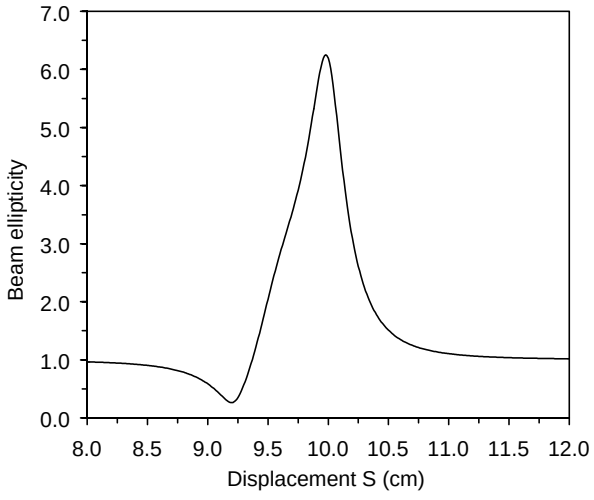


Figure 7 Typical simulated *z*-scan plot showing variation of beam ellipticity as a function of scan length. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

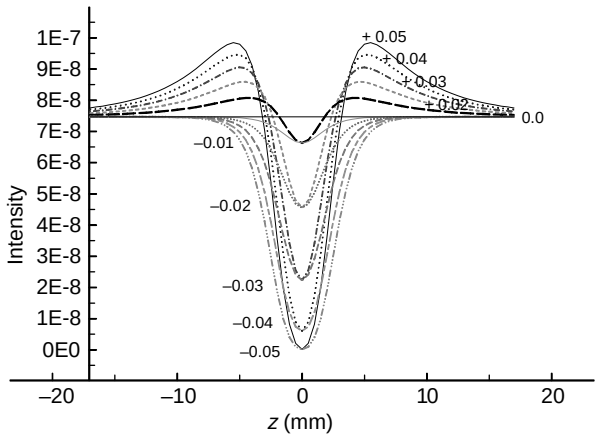


Figure 8 Typical *z*-scan plots of a diffusion dominated photorefractive material, with $\Delta n = C\nabla I$. Different values of C are written on the graph.

but at the same time allows one to neglect any effects of nonlinear absorption.

In summary, it is hoped that this article illuminates readers on the history of the *z*-scan, its application to thin and thick samples, its effectiveness in measurement and characterization of effective third-order optical nonlinearities for a wide class of nonlinear materials, such as $\chi^{(3)}$ materials, materials with cascaded second-order nonlinearities, liquid crystals, photorefractives, etc., and its potential for automation, while at the same time pointing out areas where care should be taken in the interpretation of the measured data. With appropriate modification in the theory, it can also be applied to characterization of linear lenses, possibly voltage controllable lenses and lenslet arrays, suitable for dynamic imaging in biomedical applications.

See also

Materials for Nonlinear Optics: Liquid Crystals for NLO. **Nonlinear Optics, Basics:** Four-Wave Mixing; $\chi^{(3)}$–Third-Harmonic Generation.

Further Reading

Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications*. New York: Marcel Dekker.

Banerjee PP, Danilieko A, Hudson T and McMillen D (1998) P-scan analysis of induced inhomogeneous optical linearities. *Journal of the Optical Society of America* 15: 2446.

Banerjee PP, Misra RM and Maghraoui M (1991) Theoretical and experimental studies of propagation of beams through a finite sample of a cubically nonlinear material. *Journal of the Optical Society of America* 8: 1072.

Boyd R (2002) *Nonlinear Optics*, 2nd edn. New York: Academic Press.

Butcher P and Cotter D (1990) *The Elements of Nonlinear Optics*. New York: Cambridge University Press.

Chapple PB, Staromlynska J and McDuff RG (1994) *Z*-scan studies in the thin- and the thick-sample limits. *Journal of the Optical Society of America B* 11: 975.

Noginov M, Helzer SW, Loutts GB, Banerjee PP, Morrisey M and Kim Y (2003) Study of photorefraction response and diffraction efficiency in Mn:$YalO_3$ crystals. *Journal of the Optical Society of America B* 20: 1233.

Sheik-Bahae M, Said AA and Van Stryland EW (1989) High-sensitivity, single beam n_2 measurements. *Optics Letters* 14: 955.

Sheik-Bahae M, Said AA, Wei TH, Hagan DJ and Van Stryland EW (1990) Sensitive measurement of optical nonlinearities using a single beam. *IEEE Journal of Quantum Electronics* 26: 760.

Sutherland RL (1996) *Handbook of Nonlinear Optics*. New York: Marcel Dekker.

Yeh P (1993) *Introduction to Photorefractive Nonlinear Optics*. New York: Wiley.

MATERIALS FOR NONLINEAR OPTICS

Contents

Liquid Crystals for NLO
Organic Nonlinear Materials

Liquid Crystals for NLO

I C Khoo, Pennsylvania State University, University Park, PA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Liquid crystals are complex organic molecular systems that exhibit various mesophases characterized by the degree of order. In this review, we shall focus mainly on thermotropic liquid crystals which exhibit various mesophases in the temperature range between the crystalline and the isotropic liquid states. In particular, we shall discuss the extraordinarily large optical nonlinearities of nematic liquid crystals, as well as the electronic nonlinearities of isotropic-phase liquid crystals. Recent studies have shown that these materials are promising candidates for advanced photonic applications.

Nematic liquid crystals (NLC) possess directional ordering, i.e., when placed in thin cells with surface

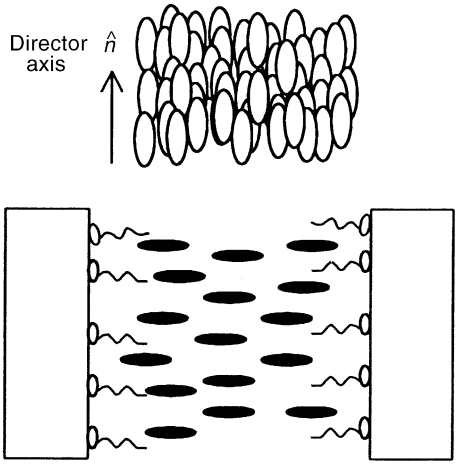


Figure 1 Nematic liquid crystals (NLC) which exhibit directional ordering characterized by the director axis. The lower figure shows how the NLC molecules are aligned in a typical cell with surfactant coated windows.

treated boundaries, the NLC molecules will self-assemble into a birefringent crystal characterized by a so-called director axis $\hat{n}$, cf. Figure 1. For light polarized parallel to $\hat{n}$, the refractive index is n_e, while for light polarized perpendicular to $\hat{n}$, the refractive index is n_0. Typically n_e is ~1.7 and n_0 is ~1.5. This large birefringence $\Delta n = n_e - n_0$ of NLC holds throughout the near UV to the far infrared spectral regime, cf. Figure 2. Perhaps the most important characteristics of NLCs is the easy susceptibility of the birefringent director (crystalline) axis orientation to externally applied fields, which has led to their ever increasing widespread use in various opto-electronics information and image display and processing devices. In analogy to dc and ac field effects, the director can easily be reoriented by an optical field, and thus cause a large optically induced index of refraction change. By doping the NLC with photosensitive or photo-charge producing agents, this process of refractive index change can be enhanced considerably. The refractive indices of NLC are also very strongly temperature dependent.

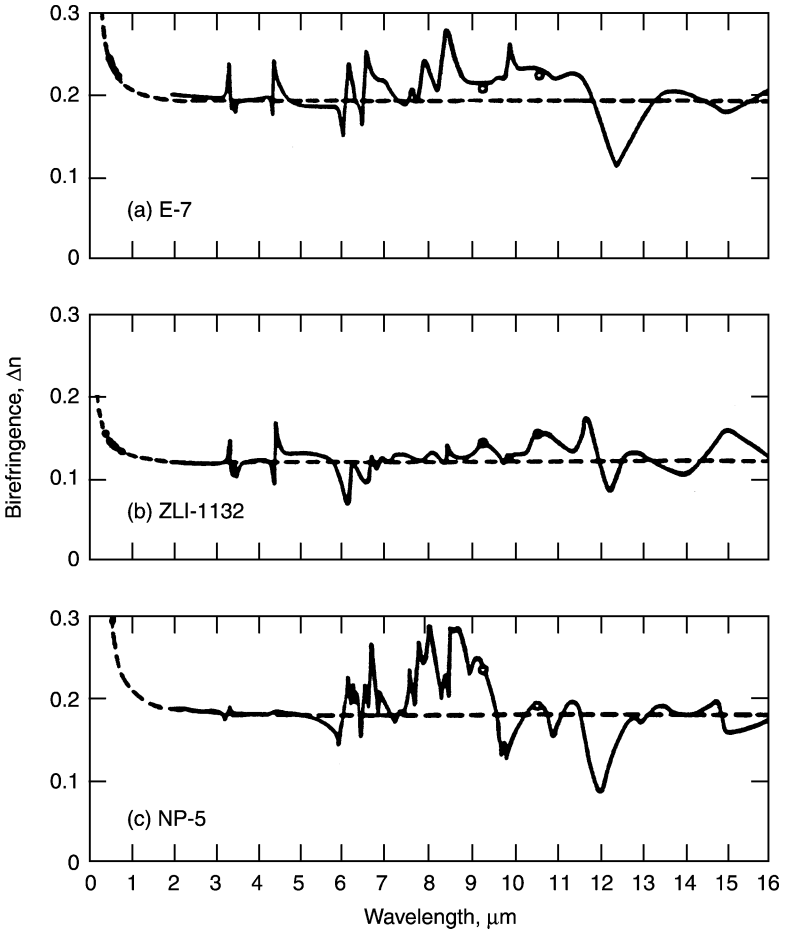


Figure 2 Broadband birefringence of nematic liquid crystals from the near ultraviolet to far infrared.

Accordingly, a large refractive index change can also be generated with laser absorption induced density and order parameter changes. In the following sections, we will broadly classify these nonlinear optical processes originating from the collective response of NLC to external fields as *nonelectronic nonlinearities*.

On the other hand, individual NLC molecules also respond to an optical field through the electronic dipolar interaction. A typical electronic energy level structure of liquid crystal molecules is shown in Figure 3. The dipolar interaction gives rise to the usual nonlinear polarization characterized by the second, third and higher-order nonlinear susceptibilities $\chi^{(2)}, \chi^{(3)}, \ldots$ etc. The complex imaginary parts of these linear and nonlinear susceptibilities account for the photonic absorption processes. Most liquid crystalline molecules possess small linear absorption in the visible region (400–700 nm), and begin to absorb near the UV region (<350 nm), cf. Figure 4. This absorption spectrum translates into a (usually) large nonlinear two-photon absorption spectrum in the visible. In some organic molecules, the intermediate absorption cross-section σ_i or the excited state absorption cross-section σ_e is much larger than the ground state linear absorption σ_g, giving rise to reverse saturable absorption (RSA) and excited state

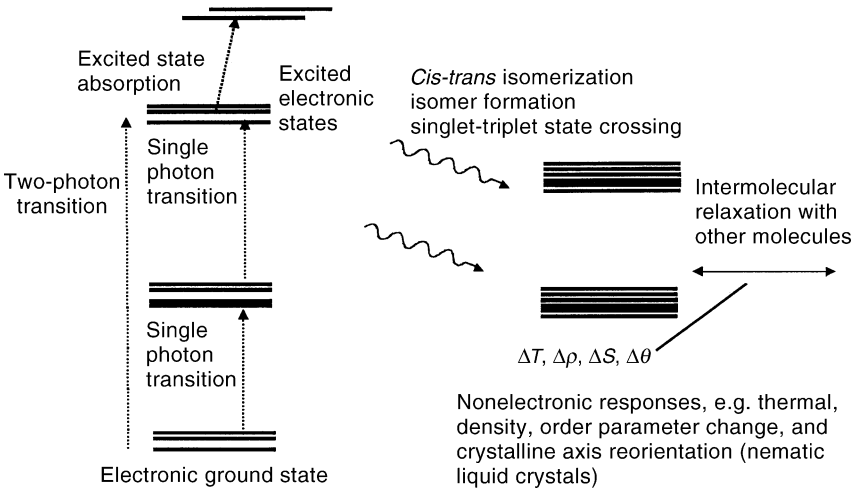


Figure 3 Schematic depiction of the typical energy level structure of liquid crystal molecules and various linear and nonlinear optical processes.

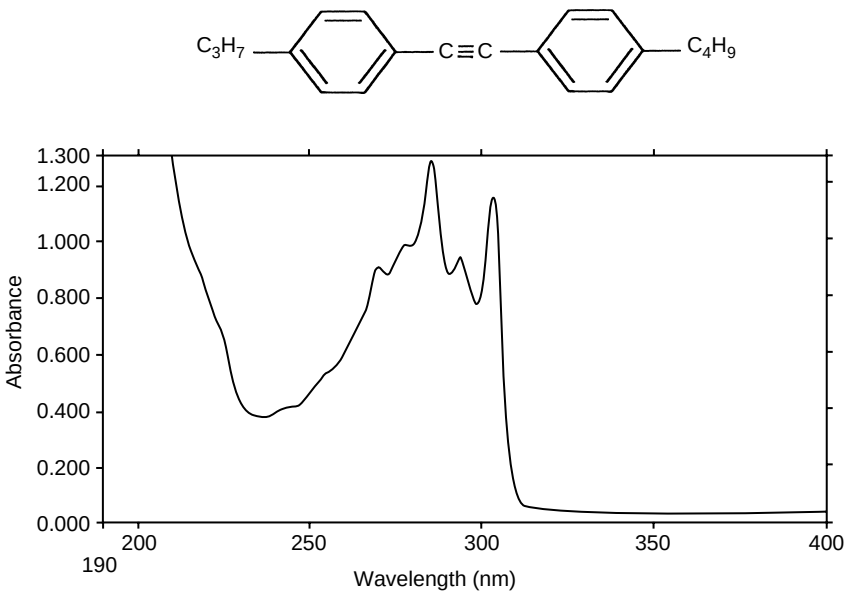


Figure 4 Molecular structure and absorption spectrum of typical isotropic liquid crystal molecules.

absorption (ESA), respectively. As we will see presently, these so-called *electronic nonlinearities* are quite useful for optical limiting applications.

Following photo-absorption, the liquid crystal molecules could undergo intersystem crossing, *cis-trans* or other conformational changes, and form isomers or charge transfer complexes. These processes will contribute to the nonelectronic nonlinearities such as reorientational, order parameter changes and thermal effects.

Orientational Nonlinearities

Purely Optically Induced Orientation

Figure 5 depicts schematically various processes that could occur in an aligned nematic film upon irradiation by an optical field. The reorientation of the director axis $\hat{n}$ is governed by the system's tendency to minimize the overall free energy, which consists of several bulk and surface interaction terms, and the interaction energy with the applied field. In purely dielectric nonabsorbing nematic liquid crystals, director axis reorientation is caused by the torque Γ_{opt} between the anisotropic nematic and the optical electric field,

$$\Gamma_{opt} = \left(\frac{\Delta\varepsilon}{8\pi}\right)(\hat{n}\cdot\mathbf{E}^*)(\hat{n}\times\mathbf{E}) \qquad [1]$$

which competes with the elastic restoring torque Γ_{el} of the surrounding NLC molecules:

$$\Gamma_{el} = \frac{(K_1\sin^2\theta + K_3\cos^2\theta)\mathrm{d}^2\theta}{\mathrm{d}z^2} + \left[(K_1 - K_3)\sin\theta\cos\theta\right]\left(\frac{\mathrm{d}\theta}{\mathrm{d}z}\right)^2 \qquad [2]$$

where the K's are the elastic constants, and $\Delta\varepsilon$ is the (optical) dielectric anisotropy.

As a result of the director axis reorientation, an optical wave propagating through the medium will 'see' a refractive index change. A frequently employed method for measuring the refractive index change in nematic liquid crystal films is the grating diffraction setup, as shown in Figure 6a. Two polarized coherent writing beams are overlapped on a nematic film, which is probed by a polarized reading beam. If the writing beams are copolarized, they impart an optical intensity grating onto the nematic film $I = I_0(1 + \cos qx)$, where q is the grating wavevector, cf. Figure 6b. On the other hand, if the writing beams are orthogonally polarized, the resulting intensity is uniform over the illuminated area, and instead a polarization grating is created, cf. Figure 6c. The polarization state of the light in the overlapped region varies from circular, through elliptical to linear in accordance to the phase difference between the two coherent writing beams. Using a linearly polarized probe beam, this setup allows one to distinguish various intensity and reorientation dependent effects, and assess their contribution to the induced refractive index change Δn.

The magnitude of the induced reorientation angle, and therefore the extraordinary index change, depends on whether the interaction is transient or steady state. In the steady state case, to the first order of approximation, Δn is proportional to the optical intensity I, i.e., $\Delta n = n_2^{SS} I$. Under some experimental conditions, one can also show that the transient nonlinear index coefficient $n_2\ (\tau_p)$ is given by $n_2\ (\tau_p) \sim n_2^{SS}\tau_p/\tau_r$, where $\tau_r = \gamma d^2/K\pi^2$ is the orientational relaxation time constant. For $d = 100\ \mu$m, $\gamma = 0.1$ poise and $K = 10^{-6}$ dynes, we have $\tau_r \sim 1$ s. Typical observed values of $n_2^{SS}(\theta)$ are

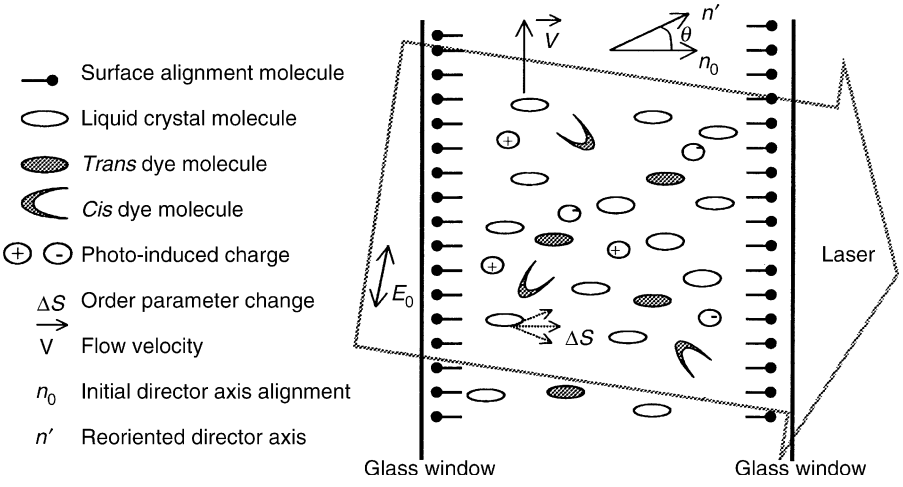


Figure 5 Schematic depiction of various laser-induced optical processes in a nematic liquid crystal film.

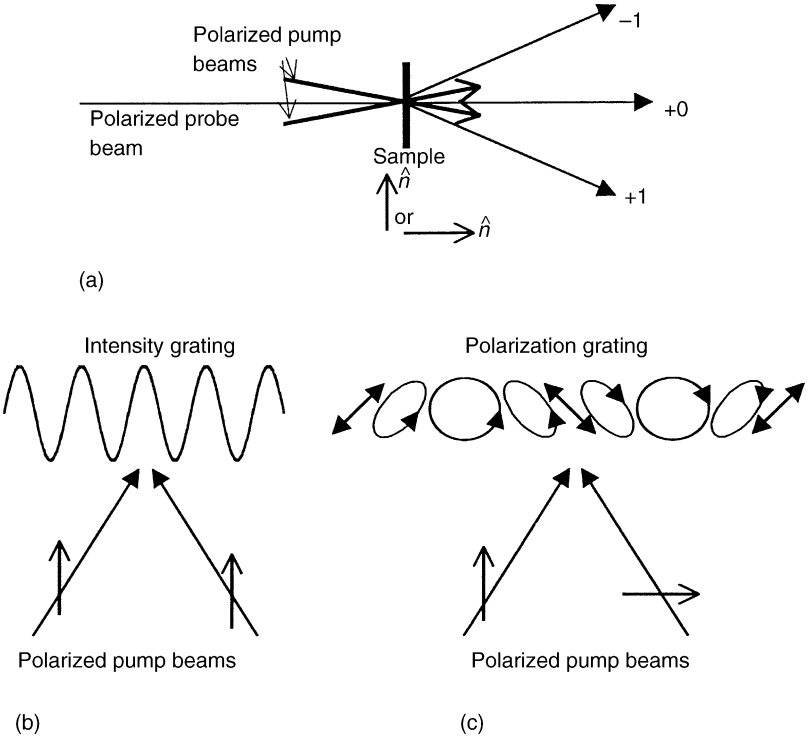


Figure 6 (a) Schematic depiction of grating formation and probe diffraction from the index grating induced in a nematic film. (b) Optical intensity grating formed by interfering two coherent copolarized writing beams. (c) Optical polarization grating formed by interfering two coherent crossed-polarized writing beams.

in the range of 10^{-5}–10^{-4} cm^2 W^{-1}. In the transient case, if $\tau_p = 1$ ns, then n_2 $(\tau_p) \sim 10^{-12}$ cm W^{-1} [$\chi^{(3)} \sim 2 \times 10^{-11}$ esu].

Excited Dye-Dopant Induced Intermolecular Torque

Studies of nematic liquid crystals doped with anthraquinone or dichroic dyes have shown that the excited dye molecules could exert an intermolecular torque Γ_{mole} that could be much larger in magnitude than Γ_{op}. By considering the interaction energies between ground and excited dyes with the surrounding LC molecules, one can show that Γ_{mole} may be written as:

$$\Gamma_{mole} = \zeta \Gamma_{op} \qquad [3]$$

where the proportionality constant ζ can be much larger than unity. The nonlinear index coefficients n_2 obtained in these films are on the order of 10^{-3} cm^2 W^{-1}. The response times for the director axis reorientation and relaxation are similar to the optically induced reorientation process discussed in the preceding section.

Photorefractive Effect

Another mechanism, namely, orientational photorefractivity, could also result in a very large optical nonlinearity. This effect occurs in nematic liquid crystals doped with photo-charge producing agents such as fullerene C_{60} or rhodamine 6G dye, in conjunction with a small applied dc field. For a 25 μm thick film with a dc bias voltage of ~1 V, the observed nonlinear coefficient n_2 is in the range of 10^{-3}–10^{-2} cm^2 W^{-1}. Figure 7 illustrates the orientational photorefractive effect. An incident optical field excites the doped liquid crystals to form charge transfer complexes (CTCs). The CTCs subsequently dissociate and create dc space charge fields through ionic diffusion, migration and other electrodynamical processes. The space charge fields, in combination with the applied dc field, create a dc field induced torque of the form:

$$\Gamma_{dc} = \left(\frac{\Delta\varepsilon_{dc}}{8\pi}\right)(\hat{n}\cdot\mathbf{E}^*_{int})(\hat{n} \times \mathbf{E}_{int}) \qquad [4]$$

Here the total internal dc field $\mathbf{E}_{int}$ contains the applied dc field $\mathbf{E}_{dc}$ and the generated dc space charge fields $\mathbf{E}_{sc}$. This causes director axis reorientation and refractive index change.

The space charge fields consist of one photorefractive and two anisotropy components. For an incident optical intensity grating function of the form $I_{op} = I_0(1 + m \cos q\xi)$, the photorefractive space charge

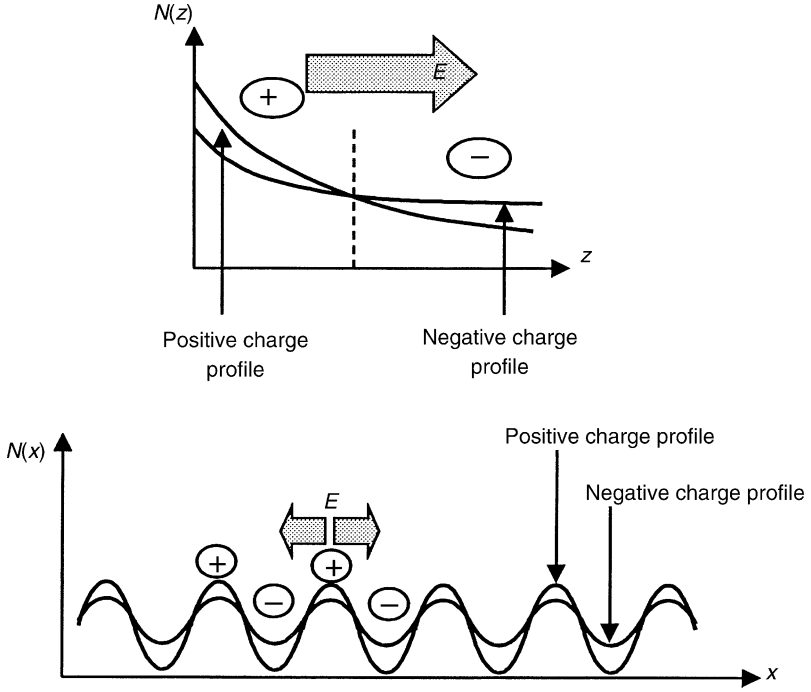


Figure 7 Schematic depiction of photo-induced space charge field formation in a nematic liquid crystal illuminated by a monotonic (upper curve) or a sinusoidal (lower curve) laser intensity function.

field is given by:

$$E_{\mathrm{ph}} = q\nu\, mk_{\mathrm{B}}T\left(\frac{\sigma-\sigma_{\mathrm{d}}}{2e\sigma}\right)\cos\left(\frac{q\xi-\pi}{2}\right) \qquad [5a]$$

where k_{B} is the Boltzmann constant, σ is the illuminated conductivity, σ_{d} is the dark state conductivity, $\nu = (D_+ - D_-)/(D_+ + D_-)$, where D_+ and D_- are the diffusion constants for the positively and negatively charged ions, respectively; m is the optical intensity modulation factor, $q = 2\pi/\Lambda$ is the magnitude of the grating wavevector, and ξ is the coordinate along q.

Two other sources of space charge fields are also at work, and they arise from the conductivity and dielectric anisotropies of the nematic under the action of the dc bias field E_{dc}:

$$E_{\Delta\sigma} = E_{\mathrm{dc}}\left(\frac{(\sigma_{//}-\sigma_{\perp})\sin\theta\cos\theta}{\sigma_{//}\sin 2\theta + \sigma_{\perp}\cos 2\theta}\right) \qquad [5b]$$

$$E_{\Delta\varepsilon} = E_{\mathrm{dc}}\left(\frac{(\varepsilon_{//}-\varepsilon_{\perp})\sin\theta\cos\theta}{\varepsilon_{//}\sin 2\theta + \varepsilon_{\perp}\cos 2\theta}\right) \qquad [5c]$$

where $(\sigma_{//} - \sigma_{\perp})$ is the low-frequency **conductivity anisotropy** and $(\varepsilon_{//} - \varepsilon_{\perp})$ the **dielectric anisotropy** of the liquid crystal, and θ is the reorientation angle of the director axis.

In addition to these space charge field induced effects, the **flows** of the charged ions and complexes under the applied dc field and the generated space charge fields also contribute significantly to reorienting the director axis. For an interaction geometry similar to Figure 5, the flow creates a shear torque Γ_{shear} of magnitude proportional to the velocity gradient

$$\Gamma_{\mathrm{shear}} = \frac{\mathrm{d}\mathbf{v}(\zeta)}{\mathrm{d}\zeta}(\alpha_2\sin^2\theta - \alpha_3\cos^2\theta) \qquad [6]$$

where the α's are the viscosity coefficients, and $\mathbf{v}(\zeta)$ is the flow velocity. The combined action of the above space charge fields and the flow effects on director axis reorientation have been quantitatively formulated and shown to describe experimental results very well, including, in particular, the observation that maximal diffraction is obtained for a grating constant $\lambda_{\mathrm{g}} \sim 2d$. Furthermore, the flow-reorientation and adsorption of the excited dopant molecules on the surfaces are believed to be responsible for quasi-permanent and permanent director axis reorientation in rhodamine 6G- and C_{60}-doped nematic films, respectively.

Supra-Nonlinear Methyl-Red Doped NLC – Observed Phenomena

Recent studies in methyl-red doped nematic liquid crystals (MRNLCs) have ushered in the era of supra-nonlinearity. Unlike photorefractive nematic liquid

crystals, no external bias field is needed. The observed n_2 values could be as large as $10\,\mathrm{cm^2\,W^{-1}}$. In MRNLC, a large photovoltage drop across the sample is also observed upon irradiation. The polarity of the photovoltage is reversed if the light impinges from the opposite cell window. The extremely nonlinear optical response of MRNLC is attributed to nematic director axis reorientation by the optically induced space charge fields, similar to the effect discussed in the previous section. The unusually large response is possibly due to several factors, including the higher photo-charge producing abilities of methyl red, and the large difference in the positive and negative photo-ion diffusion constants. Another cause of director axis reorientation is the intermolecular torque exerted by the photo-excited dye molecules on the NLC. This is particularly evident in experiments using an input polarization grating on planar aligned samples. Because the laser excited dye molecules are adsorbed and strongly 'anchored' on the cell walls, they cause permanent realignment of the nematic liquid crystal. This process usually occurs upon prolonged illumination of the liquid crystal cells with a low-power laser, and can be sped up with higher laser power.

Trans–Cis Isomerization and Order Parameter Change

Some nematic liquid crystals are azo-compounds, i.e., their molecular structures are analogous to the typical one depicted in Figure 8. An azo-compound when irradiated at its absorption wavelength will change from the *trans* to the *cis* form, cf. Figure 8. *Trans–cis* isomerization of these azo-nematic liquid crystals or azo-dopants in NLC's could also give rise to a refractive index change through modification of the NLC's order parameter and/or director axis reorientation. The *trans* species are oblong in shape, and align parallel to the director axis. On the other hand, the 'bent' structure of the *cis* species is incongruous to the nematic order, and thus causes an order parameter change, and therefore a refractive index modification. Such refractive index changes owing to order parameter modification are similar to thermal effects discussed in the next section. The response [on-] time for these *trans–cis* isomerization induced order-parameter changes can be quite fast (as fast as nanoseconds in some cases). The relaxation time (typically in the millisecond regime) is decided by the director axis relaxation rate as discussed earlier.

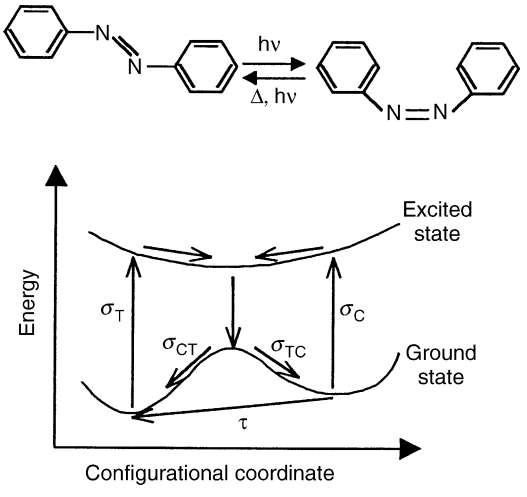


Figure 8 *Trans–cis* isomerization of azo-molecules upon optical illumination. The upper diagram depicts molecular structural changes; the lower diagram shows the excitation energy versus the molecular coordinate associated with these photo-excited processes.

The other index-changing mechanism is the tendency of the *trans* molecule to align with its transition moment perpendicular to the optical polarization in order to minimize the interaction (absorption) energies. This in turn causes the NLC director axis to realign through the intermolecular torque effect. Both mechanisms are characterized by a nonlinear index coefficient n_2 on the order of $10^{-3}\,\mathrm{cm^2\,W^{-1}}$.

Thermal and Density Effect

Laser induced temperature and density changes are described by time-dependent coupled hydrodynamics equations. In grating diffraction experiments, cf. Figure 6, the problem can be simplified to a one-dimensional one, i.e., along the grating wavevector direction, if the grating spacing is smaller than the nematic film thickness and the beam size. In that case, the thermal and Brillouin (sound wave) decay time constants are given by, respectively:

$$\tau_T = \frac{\rho_0 C_v}{\lambda_T q^2} \qquad [7a]$$

$$\tau_B = \frac{2\rho_0}{\eta q^2} \qquad [7b]$$

where $q = 2\pi/\lambda_q$ is the grating wavevector, ρ_0 the unperturbed density, C_v the specific heat, λ_T the thermal conductivity, and η the viscosity. The values of these parameters depend on the director axis alignment relative to the optical fields, thermal diffusion and sound propagation directions. For typical values of $\eta = 7 \times 10^{-2}\,\mathrm{kg\,m^{-1}\,s^{-1}}$, $\rho_0 = 10^3\,\mathrm{kg\,m^{-3}}$, $D = \lambda_T/\rho_0$, $C_v \sim 1.6 \times 10^{-7}\,\mathrm{m^2\,s^{-1}}$ and a grating spacing $\lambda_q = 20\,\mu\mathrm{m}$, we have $\tau_T \sim 50\,\mu\mathrm{s}$, whereas $\tau_B \sim 200\,\mathrm{ns}$. Such a vast difference between the relaxation times allows one to distinguish the contribution from thermal and density origins.

Table 1 Nonelectronic nonlinearities of nematic liquid crystals

Nonlinear processes	*Typical magnitude*	*Time-scale*	*Liquid crystal*
Orientational nonlinearities			
Purely optically induced n_2	10^{-4} cm^2 W^{-1}	ms	pure 5CB
Excited dopant assisted n_2	10^{-3} cm^2 W^{-1}	ms	dyed 5CB
Photorefractive	10^{-3} cm^2 W^{-1}	ms	5CB
Photorefractive + Methyl-red doped 5CB	10 cm^2 W^{-1}	ms	
Trans–Cis and order parameter modification	10^{-4} cm^2 W^{-1}	ms	dyed 5CB
Thermal nonlinearities			
n_2	10^{-6} cm^2 W^{-1}	ms	E7
n_2	10^{-11} cm^2 W^{-1}	ns	E7
Density effect			
electrostrictive n_2	10^{-11} cm^2 W^{-1}	ns	E7

In the steady state, i.e., when the laser–nematic interaction time is much longer than τ_T and τ_B, the nonlinear index coefficient is given by:

$$n_2^{SS}(\Delta T) = \left(\frac{dn}{dT}\right)\frac{\alpha\tau_r}{\rho_0 C_V} \qquad [8]$$

where dn/dT is the appropriate thermal index coefficient. The absorption constant α of liquid crystals varies widely in the visible–infrared spectrum. Near the phase transition temperature, $dn_{//,\perp}/dT$ also changes dramatically. The magnitude of $n_2^{SS}(\Delta T)$ could therefore vary by several orders of magnitude. For $\alpha = 100\ \text{cm}^{-1}$, $dn_{//,\perp}/dT \sim 10^{-3}\ \text{K}^{-1}$, and $\lambda_q = 20\ \mu\text{m}$, we have $n_2^{SS}(\Delta T) \sim 10^{-6}\ \text{cm}^2\ \text{W}^{-1}$. This is the typical value observed in wave mixing experiments with infrared [CO_2] or visible lasers.

In the transient regime ($\tau_p \ll \tau_r, \tau_B$), the nonlinear index coefficient is proportional to the laser pulse duration:

$$n_2(\tau_p) = n_2^{SS}(\Delta T)\tau_p/\tau_r \qquad [9]$$

With $\tau_r \sim 0.5 \times 10^{-4}$ s and $\tau_p = 1$ ns, $n_2(\tau_p) \sim 5 \times 10^{-11}\ \text{cm}^2\ \text{W}^{-1}$. Such nonlinearities have been observed in transient dynamic grating diffraction studies involving nanosecond laser pulses. In these studies, it is found that the contribution from the electrostrictive effect (movement of the nematic due to the optical intensity gradient) to the refractive index change can be as large as the absorptive thermal component.

In Table 1, we give a concise summary of these nonelectronic optical nonlinearities observed in nematic liquid crystals.

Electronic Nonlinearities

Second and Third Order Nonlinear Susceptibilities

Electronic interaction involves the perturbation of the electronic wavefunctions by the impinging optical fields. The induced dipole moment of a molecule is given by:

$$\mathbf{d} = \boldsymbol{\alpha}:\mathbf{E} + \boldsymbol{\beta}:\mathbf{EE} + \boldsymbol{\gamma}:\mathbf{EEE} + \cdots \qquad [10]$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are the linear, second-order and third-order nonlinear polarizability tensors, $\mathbf{E}$ is the optical electric field and : denotes tensorial operation. The macroscopic polarization $\mathbf{P}$ (dipole moment per unit volume) is

$$\mathbf{P} = \varepsilon_0\boldsymbol{\chi}^{(1)}:\mathbf{E} + \boldsymbol{\chi}^{(2)}:\mathbf{EE} + \boldsymbol{\chi}^{(3)}:\mathbf{EEE} + \cdots \qquad [11]$$

where $\boldsymbol{\chi}^{(1)}$, $\boldsymbol{\chi}^{(2)}$, and $\boldsymbol{\chi}^{(3)}$ are the linear, second and third-order nonlinear susceptibilities, respectively. The macroscopic parameters $\boldsymbol{\chi}$ are related to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ by the local field correction factors.

A well-known process associated with the second-order nonlinear polarization is second harmonic generation, in which two incident photons of frequency ω are converted into a photon of frequency 2ω. Symmetry rules require that a material should be non-centrosymmetric in order to exhibit second-order nonlinear polarization. In centrosymmetric nematic liquid crystals, an applied dc field or discontinuity at a surface or interface, or director orientational curvature, could provide the required symmetry-breaking mechanism.

In ferroelectric or smectic C^* liquid crystals, the molecules possess permanent polarization, i.e., the centrosymmetry is broken, and it is possible to observe second harmonic generation associated with the second-order nonlinear polarization. Typically, the observed nonlinear coefficients β_{2HG} are on the order of $\sim 10^{-14}\ \text{mV}^{-1}$. In contrast to inorganic crystals of much larger dimensions, these thin liquid crystal films are not useful for practical devices. Nonetheless, because of the sensitivity of second harmonic generation to surface and interface conditions, second harmonic generation has been proven to be an effective spectroscopic tool.

In dc field induced second harmonic generation, the nonlinear polarization involved is actually the

Table 2 Electronic optical nonlinearities of nematic liquid crystals

Nonlinear optical phenomena and nonlinearities	*Magnitude*	*Time-scale (pulse duration of laser used)*
Second harmonic generation (dc field induced),		
$\chi_{//}^{(3)}$	2.7×10^{-12} esu	ns
$\chi_{\perp}^{(3)}$	0.55×10^{-12} esu	ns
Second harmonic generation β_{2HG}	10^{-12} mV^{-1}	ps (ferroelectric)
Two-photon absorption, β	5 cm GW^{-1}	80 ps
Effective nonlinear absorption, β_{eff}	200 cm GW^{-1}	10 ns
Degenerate four-wave mixing		
electronic $\chi^{(3)}(\omega; \omega, -\omega, \omega)$	10^{-12} esu	ns

third-order nonlinear polarization given by $P(2\omega) = \boldsymbol{\chi}^{(3)}E(0)E(\omega)E(\omega)$. Typically measured values of $\boldsymbol{\chi}^{(3)}$ are $\sim 10^{-12}$ esu, expressed in the commonly used cgs units. The conversion factor between cgs and SI unit is given by

$$\boldsymbol{\chi}^{(3)}[\text{esu}] = 3^4 \times 10^{17} \boldsymbol{\chi}^{(3)}[\text{SI unit}] \qquad [12]$$

There is relatively little work done on measuring or studying the third-order nonlinear electronic polarizabilities of nematic liquid crystals. Most of the studies are centered on the molecular correlation effects associated with the nematic–isotropic phase transition. The typical magnitude of electronic $\boldsymbol{\chi}^{(3)}$ is 10^{-12} esu.

Nonlinear Absorption and Limiting

Two-photon absorption occurs where the molecule is excited to a real electronic excited state by simultaneous absorption of two photons from the incident light. Since the absorption is proportional to the square of the optical intensity, the incident light will experience an increasing absorption effect, and its transmission at high intensity will be 'quenched' to an almost constant level. Two-photon absorption is usually characterized by the two-photon absorption coefficient β, which is related to the imaginary part of $\boldsymbol{\chi}^{(3)}$. For liquid crystals, they are on the order of 5×10^{-12}–10^{-11} mW^{-1} (equivalent to 0.5–5 cm GW^{-1} in the more commonly used units).

Liquid crystalline materials in their (less scattering) isotropic phase have been extensively studied in the last few years for application in optical limiting devices as a result of their large nonlinear absorption constant. In the nanosecond time regime, the RSA and ESA effects, cf. Figure 3, act to give rise to a very large effective nonlinear absorption coefficient β on the order of several hundreds of cm GW^{-1}. In combination with special device structure such as a nonlinear fiber array, it has become possible to clamp the transmitted nanosecond laser pulse energy to below the retina damage level (below 0.5 μJ).

In Table 2, we give a concise summary of these electronic optical nonlinearities of nematic liquid crystals.

Concluding Remarks

We have presented a concise review of the nonlinear optics of liquid crystals in their nematic and isotropic phases. Nematic liquid crystals in their various pure and doped forms possess many interesting nonlinear optical responses, with nonlinearities ranging over 16 orders of magnitude. Some of the extraordinarily large nonlinearities enable the performance of several all-optical switching and limiting, image modulation and sensing processes at unprecedented low threshold power. In particular, self-defocusing and optical limiting effects with nanowatt-power lasers, and incoherent–coherent image conversion at μW cm^{-2} optical intensity have been demonstrated. Liquid crystal films have been used in selective attenuation of glares and blinding laser light. Stimulated scattering that used to be the domain of high-power lasers can now be performed in thin nematic films with mW-power cw laser. Readers can consult the Further Reading section for some of these advanced photonic devices and applications based on nonlinear optics of liquid crystals. Current active studies by several research groups worldwide are likely to result in even more interesting nonlinear optical effects and useful applications in the near future.

List of Units and Nomenclature

Absorption coefficient	α [cm^{-1}]
Density	ρ_0 [kg m^{-3}]
Eleastic constant	K [dynes]
Intensity (optical)	I [W cm^{-2}]
Nonlinear susceptibility	$\chi^{(3)}$ [esu]
Second harmonic generation coefficient	β_{2HG} [mV^{-1}]
Specific heat	C_V [m^2 s^{-1}]
The nonlinear index coefficients	n_2 [cm^2 W^{-1}]
Time constants	τ [s]

Two-photon absorption constant	β [cm GW^{-1}]
Viscosity	η [kg m^{-1} s^{-1}]
Viscosity coefficient	γ [poise]

See also

Nonlinear Optics, Basics: Cascading; Kramers–Krönig Relations in Nonlinear Optics; Nomenclature and Units; Photorefraction.

Further Reading

Khoo IC (1995) *Liquid Crystals: Physical Properties and Nonlinear Optical Phenomena*. New York: Wiley Interscience.

Khoo IC (1996) Orientational photorefractive effects in nematic liquid crystal films. *IEEE Journal of Quantum Electronics* 32: 525–534.

Khoo IC and Liang Y (2000) Stimulated orientational and thermal scatterings and self-starting optical phase conjugation with nematic liquid crystals. *Physical Review* E62: 6722–6733.

Khoo IC and Wu ST (1993) *Optics and Nonlinear Optics of Liquid Crystals*. Singapore: World Scientific.

Khoo IC, Chen PH, Wood MV and Shih MY (1999) Molecular photonics of a highly nonlinear organic fiber core liquid for picosecond–nanosecond optical limiting effect. *Chemical Physics* 245: 517–531.

Khoo IC, Wood MV, Shih MY and Chen PH (1999) Extremely nonlinear photosensitive liquid crystals for image sensing and sensor protection. *Optics Express* 4(11): 431–442.

Khoo IC, Shih MY, Wood MV, *et al.* (1999) Dye-doped photorefractive liquid crystals for dynamic and storage holographic grating formation and spatial light modulation. *IEEE Proceedings* 87: 1897–1911.

Marrucci L and Papara D (1997) Photoinduced molecular reorientation of absorbing liquid crystals. *Physical Review* E56: 1765–1772.

Rau H (1990) Photoisomerization of Azobenzenes. In: Rabeck FJ (ed.) *Photochemistry and Photophysics*, vol. 2, chap. 4, pp. 119–141. Boca Raton, FL: CRC Press.

Shen YR (1990) *Principles of Nonlinear Optics*. New York: Wiley Interscience.

Organic Nonlinear Materials

F Kajzar and I Rau*, Commissariat à l'Energie Atomique, Gif sur Yvette, France

© 2005, Elsevier Ltd. All Rights Reserved.

Nonlinear Medium Polarization

Macroscopic

Photons are neutral, noninteracting bosons. Interaction is possible only through a material medium. Indeed, photons can be absorbed by electrons leading to excited states. A reverse transition may go through radiative channels, with emission of photons, usually with a different energy than the excitation energy. Interaction of photons with matter may also lead to other types of excitation, such as vibronic, rotational, libronic, etc., or just to a change in the medium polarization. This change may be expanded in the power series of the electric forcing field strength with *i*th component given by:

$$P_I(\omega_\sigma) = P_{0I}(\omega_\sigma) + \varepsilon_0[K_1\chi^{(1)}_{IJ}(-\omega_\sigma;\omega_\sigma)E^{\omega_\sigma}_J + K_2\chi^{(2)}_{IJK}(-\omega_\sigma;\omega_1,\omega_2)E^{\omega_1}_J E^{\omega_2}_K + K_3\chi^{(3)}_{IJKL}(-\omega_\sigma;\omega_1,\omega_2,\omega_3)E^{\omega_1}_J E^{\omega_2}_K E^{\omega_3}_L + \cdots] \quad [1]$$

where the expansion coefficients $\chi^{(n)}$ are $(n+1)$ rank three-dimensional tensors, describing linear $(n=1)$ and nonlinear $(n>1)$ responses of the medium, $P_{0i}(\omega_\sigma)$ is its permanent polarization (at frequency ω_σ) and the coefficients K_i $(i=1,2,3,\ldots)$ take account of the conventions for the Fourier transform of electric field and polarization vectors used, as well as the degeneracy of nonlinear optical processes. For historical reasons two conventions in quantitative description of NLO processes are used:

$$\mathrm{E}(\mathrm{r},t) = \tfrac{1}{2}[\mathrm{E}(\mathrm{r})\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{c.c.}] \qquad \text{(Convention I)} \qquad [2]$$

and

$$\mathrm{E}(\mathrm{r},t) = \mathrm{E}(\mathrm{r})\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{c.c.} \qquad \text{(Convention II)} \qquad [3]$$

and similarly for the polarization field. As researchers often use different conventions, it is important to bear this in mind when comparing NLO properties of materials coming from different sources. The tabulated data and the formulas used here are within convention I, with the permutation factors included explicitly in coefficients K_i (eqn [1]). In centrosymmetric media and in dipolar approximation, the static polarization P_0, as well as all even order NLO susceptibilities, are equal to zero, as it follows from time reversal symmetry.

*On leave from POLITECHNICA University of Bucharest, Romania.

Relations Between Macroscopic and Microscopic Polarizations

Microscopic Polarization

Similar expansions to eqn [1] also obey the conventions on a molecular level. Similarly, as for bulk polarization, the ith component of the molecule dipole moment can also be expanded into the series of the local field strength giving:

$$\begin{aligned}\mu_i(\omega_\sigma) &= \mu_{0i}(\omega_\sigma) + K_1\alpha_{ij}(-\omega_\sigma;\omega_\sigma)E_j^{\omega_\sigma} \\ &\quad + K_2\beta_{ijk}(-\omega_\sigma;\omega_1,\omega_2)E_j^{\omega_1}E_k^{\omega_2} \\ &\quad + K_3\gamma_{ijkl}(-\omega_\sigma;\omega_1,\omega_2,\omega_3)E_j^{\omega_1}E_k^{\omega_2}E_l^{\omega_3} + \cdots \\ &= \mu_{0i}(\omega_\sigma) + \mu_{\text{ind}}(\omega_\sigma,\omega_1,\omega_2,\ldots,\omega_n,\alpha,\beta\ldots,\zeta, \\ &\quad E^{\omega_1},E^{\omega_2},\ldots,E^{\omega_n}) \qquad [4]\end{aligned}$$

where α is the linear molecular polarizability, β is the first order and γ is the second molecular hyperpolarizability. μ_{in} in eqn [4] denotes the induced dipole moment. The expansions [1] and [4] may be also considered as the Tamm–Dancoff perturbation development as functions of an increasing number of interacting photons. In this case, the second term on the right-hand sides of eqns [1] and [4] describes the interaction of two photons, the third term that of three photons, etc. The electric field in eqn [4] is the effective field experienced by the molecule and is given by

$$E = fE \qquad [5]$$

where f is the local field factor, depending on the form of molecule. In eqns [1] and [4] Einstein's notation is used.

For a single crystal and for noninteracting molecules there exists a straightforward relationship between the macroscopic and microscopic quantities given by:

$$\begin{aligned}\chi^{(2)}_{IJK}(-\omega_\sigma;\omega_1,\omega_2) &= \sum_n f_n^{\omega_\sigma}f_n^{\omega_1}f_n^{\omega_2}N_n \\ &\quad\times\sum_{ijk} a_{Ii}a_{Jj}a_{Kk}\beta^n_{ijk}(-\omega_\sigma;\omega_1,\omega_2) \qquad [6]\end{aligned}$$

for the second-order NLO susceptibility and

$$\begin{aligned}\chi^{(3)}_{IJKL}(-\omega_\sigma;\omega_1,\omega_2,\omega_3) &= \sum_n f_n^{\omega_\sigma}f_n^{\omega_1}f_n^{\omega_2}f_n^{\omega_3}N_n \\ &\quad\times\sum_{ijkl} a_{Ii}a_{Jj}a_{Kk}a_{Ll}\gamma^n_{ijkl}(-\omega_\sigma;\omega_1,\omega_2,\omega_3) \qquad [7]\end{aligned}$$

for the third-order NLO susceptibility, respectively. The coefficients a_{Ij} in eqns [6] and [7] are Wigner rotation matrix elements, N_n denoting the number density of species (n). The subscripts ($ijkl$) refer to the molecule, and ($IJKL$) to the laboratory reference frames, respectively.

For disordered (or partly ordered) materials, the macroscopic NLO susceptibility depends on the orientation average of nonzero molecular hyperpolarizability components:

$$\begin{aligned}\chi^{(2)}_{IJK}(-\omega_\sigma;\omega_1,\omega_2) &= \sum_n f_n^{\omega_\sigma}f_n^{\omega_1}f_n^{\omega_2}N_n \\ &\quad\times\sum_{ijk} <\beta^n_{ijk}(-\omega_\sigma;\omega_1,\omega_2)>_{IJK} \qquad [8]\end{aligned}$$

and similarly for $\chi^{(3)}$ susceptibility. The average in eqn [8] is taken over all NLO molecule orientations and all β tensor components. Fortunately, the number of the nonzero β tensor components (21 for β tensor) is limited by the molecular symmetry and in the case of charge transfer (CT) molecules, only the component in the charge transfer direction β_{xxx} may be retained. The others are usually significantly smaller and can be neglected with respect to this one.

Second-Order Nonlinear Optical Materials

The research on nonlinear optical materials is boosted by the potentiality of their applications in photonics, and particularly in optical signal processing, image recognition, frequency conversion, and light amplification. New fields of applications open in biology, medicine (photodynamic therapy), material processing (optical engineering), etc. Depending on the targeted application, the materials showing enhanced second-, third-, and/or higher-order nonlinear optical properties are developed and optimized. The optimization is performed on both molecular and macroscopic levels. Indeed, the bulk nonlinear optical response depends not only on the individual hyperpolarizabilities but also on the way the molecules are assembled into the bulk material. As already mentioned, only the noncentrosymmetric materials exhibit the second-order nonlinear optical response.

Molecules

The most promising molecules for the application in second-order NLO devices are intramolecular CT compounds, composed from an electron donating (D), an electron accepting (A) part, linked by a π conjugated backbone (**Figure 1**). The CT molecules exhibit a large first-order susceptibility β tensor component β_{xxx}, enhanced in the charge transfer direction x. The value of β_{xxx} depends on the strength of electron donating and electron accepting groups (amount of the charge transfer)

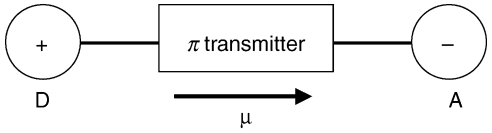


Figure 1 Schematic representation of a charge transfer molecule.

Table 1 Chemical structures of commonly used electron donating and electron accepting groups

Donors	*Acceptors*
$-N(CH_3)_2$	$-NO$
$-NH_2$	$-NO_2$
$-OCH_3$	$-CHO$
$-OH$	$-CN$
$-CH_3$	$-SO_2CH_3$
$-SCH_3$	$-SO_2CF_3$
$-N(CH_3)_2$	$=C(CN)_2$

and on the conjugation length of the transmitter. Table 1 shows the chemical structures of typical electron donating and electron accepting groups. Chemical structures of selected CT molecules are given in Table 2, together with the values of the β_{xxx} tensor component, as determined by the electric field induced second-harmonic generation technique (EFISH).

The dependence of β_{xxx} on the conjugation length L of the conjugated transmitter is given by the scaling law:

$$\beta_{xxx} \propto L^{\alpha_s} \qquad [9]$$

with $\alpha_s > 1$. Usually the conjugated backbone is composed of a number of alternate single/double (or single/triple) bonds and in this case, the scaling law is tested as a function of the number of double (or triple) bonds N:

$$\beta_{xxx} \propto N^{\alpha_\beta} \qquad [10]$$

For a number of CT oligomers with the same electron donating and electron accepting groups, differing only in the length of the conjugated transmitter, the exponent $\alpha_\beta \sim 2.5$. The β_{xxx} value saturates for a larger number of double bonds ($N > 4$, in the case of phenyl oligomers used as transmitter).

A number of studies are in progress on octupolar molecules which exhibit large β hyperpolarizabilities. However, their processing into efficient, noncentrosymmetric bulk materials is still at its initial stage and up to now no practical application has been demonstrated.

Bulk Materials

In order to obtain noncentrosymmetric materials, different approaches are used such as:

(i) single crystals growth;
(ii) epitaxy or heteroepitaxy;
(iii) Langmuir–Blodgett films;
(iv) self-assemblies;
(v) intermolecular charge transfer structures;
(vi) poled polymers; and
(vii) chiral structures.

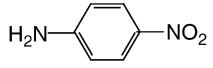
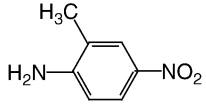
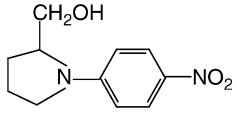
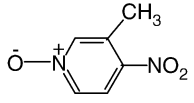
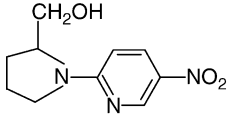
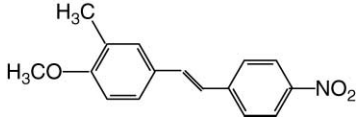
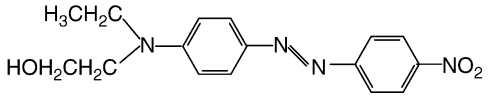
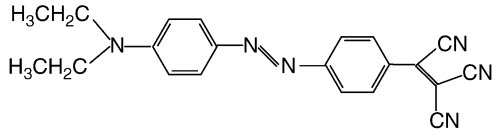
The growth of good-quality optical single crystals is costly and difficult, especially as the charge transfer molecules usually exhibit a large ground state dipole moment (Table 2). The strong dipole–dipole interaction usually leads to their antiparallel alignment, and as a consequence, a centrosymmetric bulk structure. Only a few molecules crystallize in a noncentrosymmetric structure. One of the approaches used consists of design and synthesis of molecules with a negligible ground state dipole moment such as POM (Table 3). Therefore, a lot of effort has been put into forcing molecules to build into noncentrosymmetric bulk materials (ii)–(vi). Owing to the fact that most of applications are targeted in waveguiding configuration, a significant effort has been made to grow noncentrosymmetric thin films (ii)–(vi). While the chiral molecules form spontaneously noncentrosymmetric films, in other techniques, this is obtained more easily by application of an external electric field (poled polymers), by use of an adequate substrate (epitaxy, pseudoepitaxy, self assembly), intramolecular charge transfer in ordered layered structures (v), and interaction between molecules and molecules–substrate (Langmuir–Blodgett films). In the last case, three types of well-ordered structures, with controllable thickness are obtained: X, Y, Z (Figure 2). Only X and Z type structures give noncentrosymmetric thin films exhibiting second-order NLO properties.

Poled polymers

Up to now, the best results in getting efficient materials for second-order NLO application are obtained with poled polymers. The materials consist of a polymer matrix, exhibiting excellent optical properties, functionalized with active NLO chromophores (CT molecules). Four kinds of structures, shown schematically in Figure 3, are realized:

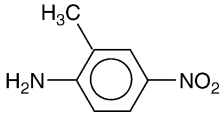
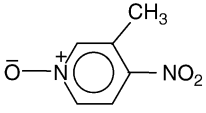
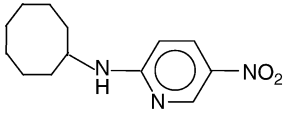
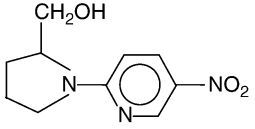
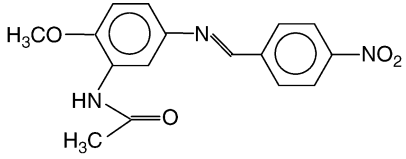
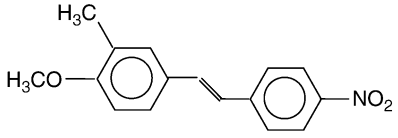
(i) guest–host systems;
(ii) side chain polymers;
(iii) main chain polymers; and
(iv) thermally or photo-crosslinked polymers.

Table 2 Molecular first hyperpolarizabilities and dipole moments of selected CT molecules

Molecule	*Chemical structure*	λ_{max} *[nm]*	*Solvent*	*Ground state dipole moment,* μ *[D]*	*Product* $\mu\beta$ $[10^{-77}$ Cm^5 $V^{-1}]$	β $[10^{-40}$ m^4 $V^{-1}]$	λ_{meas} *[nm]*	β_0 $[10^{-40}$ m^4 $V^{-1}]$
p-NA (*para*-nitroaniline)		358	1,4 dioxane	5.5	1.30 0.74	71.0 40.2	1064 1907	35
MNA (2-methyl-4-nitroaniline)		361	1,4 dioxane	7.4	1.72 0.98	69.9 39.8	1064 1907	33
NPP (*N*-(4-nitrophenyl)-(L)-prolinol)		397	acetone	7.3	4.28	176	1064	67
POM (3-methyl-4-nitropyridine-1-oxide)		320	from crystal data	0	0	36	1064	21
PNP (1-(*N*-prolinol)-5-nitropyridine)		376	1,4 dioxane	7.2	2.28 1.41	95 59	1064 1907	38
MMONS (3-Methyl-4-methoxy-4′ nitrostilbene)		366	1,4 dioxane	5.2	2.14	123.9	1907	90
DR1 (4-[*N*-ethyl-*N*-(2-hydroxyethyl)]amino-4′-nitroazobenzene (dispersed red 1))		509 455	DMSO 1,4 dioxane			525 234.1	1356 1907	197 150
DADAB (4-(*N*,*N*-diethylamino)-4′-tricyanovinylazo benzene)		582	DMSO			1640	1580	648

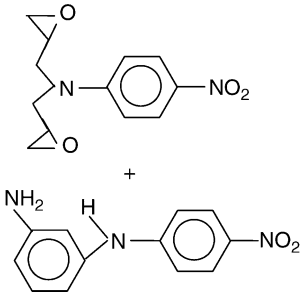
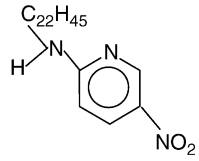
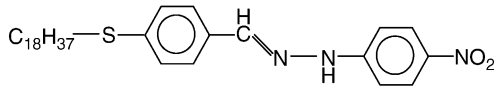
$1D = (1/3) \times 10^{-39}$ Cm.

Table 3 Bulk second-order nonlinear optical properties of selected organic materials

Material	*Chemical structure*	*Form*	*Point group*	λ_{meas} *[nm]*	*Refractive index*	*SHG susceptibility* $\chi^{(2)}_{ijk}(-2\omega;\omega,\omega)$ *[pm* V^{-1}*]*	*Electrooptical coefficient, r, [pm* V^{-1}*]*
MNA (2-methyl-4-nitroaniline)		Single crystal	*m*	514.5 1064	2.0	$\chi^{(2)}_{111} = 500$ $\chi^{(2)}_{122} = 76$	70 ± 14
POM (3-methyl-4-nitropyridine-1-oxide)		Single crystal	222	514.5 600 10 600 1064	1.712 1.919 1.638	$\chi^{(2)}_{123} = 19.2$ $\chi^{(2)}_{123} = \chi^{(2)}_{251} = \chi^{(2)}_{312}$	4.0 5.1 3.7
COANP (2 cyclooctylamino-5 nitropyridine)		Single crystal	*mm*2	514.5 600 10 600 1064	1.672 1.781 1.647	$\chi^{(2)}_{322} = 108$ $\chi^{(2)}_{333} = 48$	9.3 ± 1.5 15.4 ± 3.2
PNP (2-(N-prolinol)-5-nitropyridine)		Single crystal	2	514.5 600 10 600 1064	1.990 1.788 1.467	$\chi^{(2)}_{211} = 170$ $\chi^{(2)}_{223} = 55$	8.0 ± 1.4 13.0 ± 2.2
MNBA (4′-nitrobenzylidene-3-acetamino-4-methoxyaniline)		Single crystal	*m*	514.5 600 10 600 1064	2.024 1.585 1.652	$\chi^{(2)}_{111} = 437$	–
MMONS (3-Methyl-4-methoxy-4′-nitrostilbene)		Single crystal	*mm*2	514.5 600 10 600 1064	1.569 1.693 2.129	$\chi^{(2)}_{333} = 368^a$ $\chi^{(2)}_{223} = 142^a$	19.3 ± 4 39.9 ± 7

PMMA + DR 1 (Poly(methyl methacrilate) + 4-[N-ethyl-N-(2-hydroxyethyl)] amino-4′-nitroazobenzene (Dispersed red 1))		PP/GH	1580	$\chi^{(2)}_{333} = 5$	
			633		$r_{33} = 2.5$
PMMA + DCV (Poly(methyl metacrilate) + 4-dicyanovinyl-4′-[N-ethyl-(2 hydroxyethyl)] azobenzene)		PP/SC	1580	$\chi^{(2)}_{333} = 42$	
			799		$r_{33} = 18$
PMMA + 3RDCVXY (Poly(methyl metacrilate) + 4-dicyanovinyl-4′ (diethylamino) diazobenzene)		PP/SC	1064	$\chi^{(2)}_{333} = 840$	
			633		$r_{33} = 40$

Table 3 (*Continued*)

Material	*Chemical structure*	*Form*	*Point group*	λ_{meas} *[nm]*	*Refractive index*	*SHG susceptibility* $\chi^{(2)}_{ijk}(-2\omega;\omega,\omega)$ *[pm V^{-1}]*	*Electrooptical coefficient, r, [pm V^{-1}]*
NNDN + NAN (N,N-diglycidyl-4-nitroaniline + N-(2-aminophenyl)-4-nitroaniline)		PP/CL		1064		$\chi^{(2)}_{311} = 28$	
				530.9			$r_{13} = 6.5$
DCANP		LB film		632.8	1.598 (TE_0-mode)		
				1064		$\chi^{(2)}_{333} = 19.5 \pm 2.5$	
FA06		LB film		1064		$\chi^{(2)}_{322} = 20$	

PP, poled polymer; GH, guest host system; SC, side chain polymer; CL, cross linked polymer; LB, Langmuir–Blodgett film.
[a] Calibrated with KTP single crystal ($\chi^{(2)}_{333} = 27.4$ pm/V).

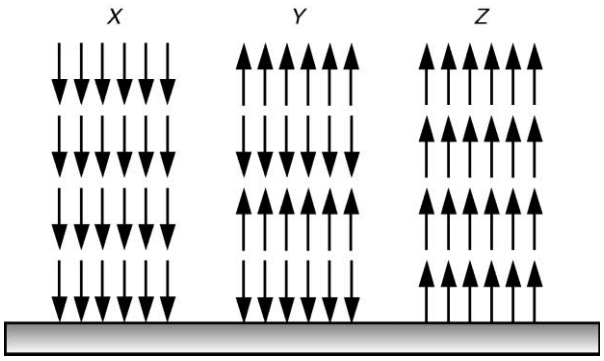


Figure 2 Schematic representation of *X, Y, Z* type Langmuir–Blodgett films. The arrows symbolize the direction of dipole moments, which usually are tilted with respect to the normal to film surface.

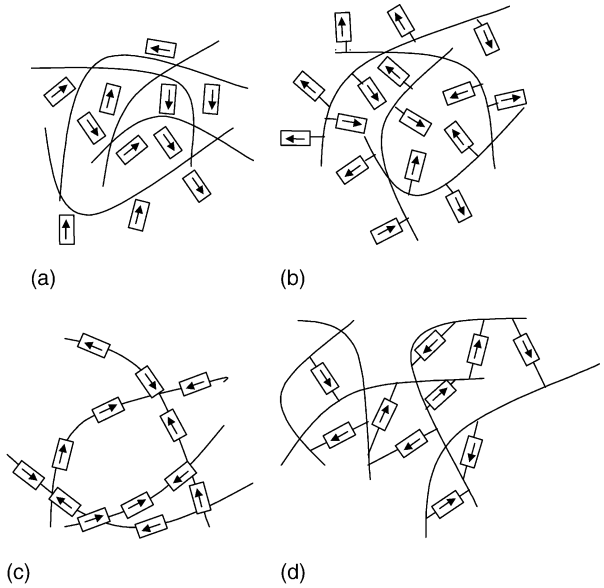


Figure 3 Schematic representation of functionalized polymers for second-order NLO effects: guest–host systems (a), side chain polymers (b), main chain polymers (c), photo and thermally cross-linking polymers (d). Arrows represent the NLO chromophores.

The guest–host systems are just solid solutions of NLO chromophores in an inert matrix. The concentration of NLO chromophores is limited because of already-mentioned strong dipole–dipole interaction leading to antiparallel alignment, resulting in an increase in propagation losses and decrease of $\chi^{(2)}$ susceptibility. The polar order is obtained by applying an external field. In order to increase the rotational mobility of chromophores, the material is heated to the glass transition temperature of polymer. The interaction of dipole moments with the external field forces them to orient in its direction. The orientation is frozen by cooling the material (usually a thin film) to room temperature. However, the stability induced in this way results in a polar order that is poor. Better stability and higher concentration of chromophores is obtained by attaching chemically the NLO active molecules to the polymer backbone (side-chain polymers). However, it requires a significant effort in chemical synthesis. The stability of induced order depends strongly on the difference between the glass transition and the operation temperatures. The larger the difference, the less relaxation is observed. Therefore, a lot of effort is going into developing functionalized high glass transition temperature polymers such as polyimides, polyurethanes, polyvinyl carbazoles, etc.

The main-chain polymers (iii) attract a relatively weak interest. This is due, not only to the difficulties encountered with chemical synthesis, but also with the efficient chromophore orientation. More promising is the last class of functionalized polymers (iv). Here the active chromophores are used to link covalently the polymer chains through a thermal or photo-crosslinking process. However, the poling process is very delicate as the glass transition temperature is increasing with the poling time, so the time process has to be carefully monitored. Also the cross-linking may lead to shrinkage and crackling of thin films and as a consequence, to the increase of propagation losses.

A special class of functionalized materials, similar to poled polymers, represent sol–gels. Here the active chromophores either fill the pores of sol–gel or are chemically bonded to the matrix. The poling process is very similar to that applied to the thermally and photo cross-linking polymers and also has to be carefully monitored. The main advantage of these materials consists of excellent optical properties, and the disadvantage on a relatively low chromophore content, limiting the bulk $\chi^{(2)}$ susceptibility value. The representative values of second-order nonlinear optical susceptibility $\chi^{(2)}_{ijk}$ for different materials for second-order NLO optics, are listed in Table 3.

Poling efficiency

The DC poling of functionalized polymers leads to birefringent noncentrosymmetric materials with two nonzero $\chi^{(2)}$ tensor components: the diagonal:

$$\chi^{(2)}_{ZZZ}(-\omega_0;\omega_1,\omega_2) = NF\beta_{xxx}(-\omega_\sigma;\omega_1,\omega_2) \times \langle\cos^3\Theta\rangle \quad [11]$$

and the off-diagonal:

$$\chi^{(2)}_{XXZ}(-\omega_\sigma;\omega_1,\omega_2) = \frac{1}{2}NF\beta_{xxx}(-\omega_\sigma;\omega_1,\omega_2) \times \langle\sin^2\Theta\cos\Theta\rangle \quad [12]$$

where we assumed for simplicity, that only one nonlinear species, as is usually the case. Z, in eqns [11] and [12], is the direction of the applied external field and Θ is the angle between the electric field and the molecule dipole moment. The efficiency of poling is measured by the ratio:

$$a = \frac{\chi^{(2)}_{ZZZ}(-\omega_\sigma; \omega_1, \omega_2)}{\chi^{(2)}_{XXZ}(-\omega_\sigma; \omega_1, \omega_2)} \qquad [13]$$

It varies between 1 and ∞ (perfect alignment). For isotropic poled polymers $a = 3 \div 6$ and for nematic liquid crystal side-chain polymers, values of $a \propto 18$ were obtained.

Usually, the second-order NLO susceptibility is measured by the second-harmonic generation technique ($\omega_\sigma = 2\omega$; $\omega_1 = \omega_2 = \omega$), giving the electronic part of $\chi^{(2)}$ susceptibility, with ultrafast response time as the generated harmonic wave has to follow the fundamental one. This susceptibility is important, not only for frequency conversion through the second-harmonic generation, but also for sum and difference frequency generation as well as for optical parametric oscillation (OPO) and amplification (OPA). Another important effect described by $\chi^{(2)}(-\omega; \omega, 0)$ susceptibility is the linear electro-optic effect (Pockels effect). It finds application in electro-optic modulation for signal transmission. Traditionally, this effect is described by the electro-optic coefficient r_{ijk}, which is a third-rank three-dimensional tensor. The variation of the medium refractive index under the applied electric field is given by the following formula:

$$\Delta n_i = -\frac{2}{n_i^3} r_{ij} E_j \qquad [14]$$

The electro-optic coefficient is directly linked to the $\chi^{(2)}(-\omega; \omega, 0)$ susceptibility:

$$r_{ijk} = -\frac{2\chi^{(2)}_{ijk}(-\omega; \omega, 0)}{(n_i n_j)^2} \qquad [15]$$

The measured values of electro-optic coefficient for some selected organic single crystals and thin films are also given in Table 3.

Third-Order Nonlinear Optical Materials

There is no symmetry restriction on molecules and materials for third- (or higher-odd-) order nonlinear optical effects. These materials are used in significantly larger classes of applications than the materials for second-order NLO effects, particularly in logic systems, all optical switching, frequency conversion (e.g., frequency tripling in third-harmonic generation), light amplification, optical bistability, all optical switching, etc. In the case of organic materials, it was shown that the conjugated quasi 1D π electron systems, such as semiconducting polymers (Figure 4) exhibit an enhanced third-order NLO response.

Molecules

Similarly, as in the case of second-order NLO materials, the NLO response depends strongly on the conjugation length. In the case of conjugated quasi 1D molecules, such as polymers with a chemical structure as shown in Figure 4, the γ tensor component in the conjugation direction (largest hyperpolarizability) is strongly enhanced and described by a scaling law:

$$\gamma_{xxxx} \propto L^{\alpha_t} \qquad [16]$$

where L is the conjugation length and the exponent $\alpha_t > 1$. The scaling law, eqn [16], can also be presented as dependent on γ on a number of π conjugated double or triple bonds N:

$$\gamma_{xxxx} \propto N^{\alpha_\gamma} \qquad [17]$$

It was extensively studied for a series of conjugated oligomers polyenes, oligothiophenes, etc., and the exponential coefficients are listed in Table 4, together with the calculated values. For a larger value of double (or triple) bonds, a saturation in $\chi^{(3)}$ susceptibility is observed.

The molecular second hyperpolarizability measurements are derived from solution measurements. Some typical values, obtained by the third-harmonic generation technique (THG) are listed in Table 5. Although other NLO property determination techniques were used, the THG method gives the fast, electronic part of molecular hyperpolarizability.

Bulk Materials

Importance of Order

Similarly, as in the case of materials for second-order NLO effects for practical applications the value of $\chi^{(3)}$ susceptibility is important, which on the other hand depends on the molecular second hyperpolarizability γ. Although the noncentrosymmetry is not required in the case of third-order NLO materials, the values of the bulk $\chi^{(3)}$ susceptibility depend on the order, as it follows from eqn [7]. This is particularly clear in the case of conjugated molecules, with enhanced γ_{xxxx} component. In the case of parallel polarizations

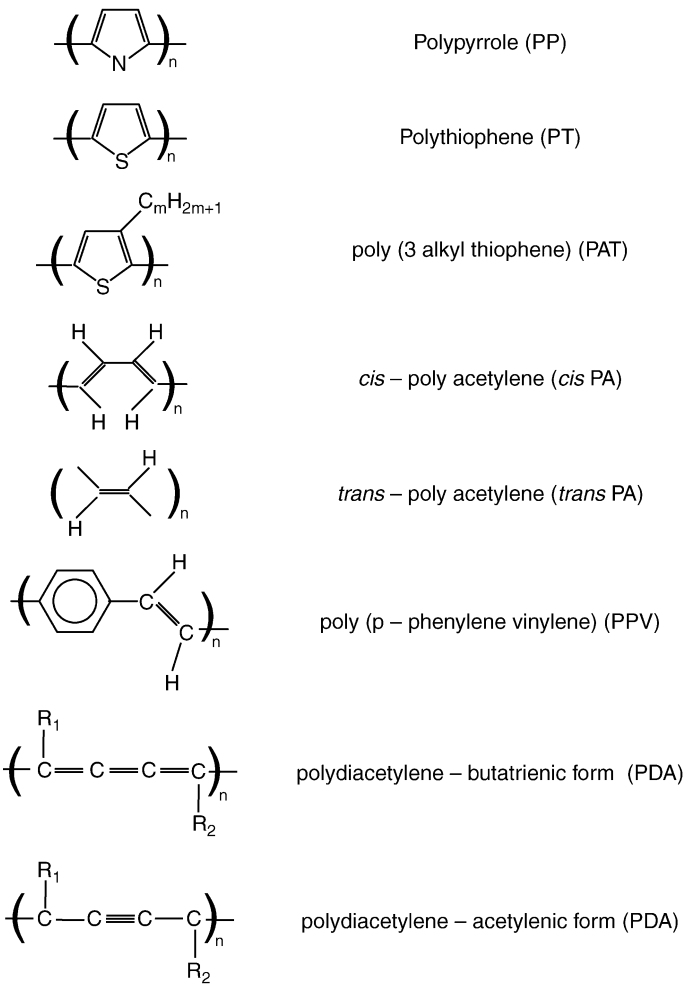


Figure 4 Chemical structures of selected conjugated quasi 1D organic polymers.

Table 4 Calculated (calc) and measured (meas) values of the scaling low exponent $\gamma \propto N^{\alpha_\gamma}$ for different conjugated oligomers

Oligomers	$\alpha_{\gamma\,calc}$	*Remark*	$\alpha_{\gamma\,meas}$	*Remark*
Polyene oligomers	5	free electron model	2.82	CT polyenes ($\omega = 0$)
		INDO ($\omega = 0$)	3.5	$\lambda_\omega = 1.907$ μm
	3	SCF/STO-3G	2.3	Symmetric A-A
	3.27	SCF/3-21G		THG 1.907 μm
	3.14	$N = 2$	3.3	Asymmetric D-A
	3.67	$N = 6$, decrease		THG 1.907 μm
	3.98	For $N > 6$ DVB-PPP	4.4	Symmetric D-D
				EFISH 1.34 μm
				no saturation up to $N = 14$
	3.8	neutral solitons		
	3.85	charged polarons		
	6.32	*trans* form		
	5.4	*cis* form		
	4.7			
Thiophene oligomers	4.5	CNDO $N < 6$	4.05	DFWM in solutions
		Saturation $N > 6$	4.54	EFISH in solid solutions $N = 3, 5, 6$
			3.4	THG 1.907 μm

of $\omega_\sigma, \omega_1, \omega_2, \omega_3$ fields, the bulk susceptibility is given by:

$$\chi^{(3)}(-\omega_\sigma; \omega_1, \omega_2, \omega_3) = NF\gamma_{xxxx}(-\omega_\sigma; \omega_1, \omega_2, \omega_3) \times \langle \cos^4 \Phi \rangle \quad [18]$$

where F is the total field factor and Φ is the angle between the polarization direction of $\omega_\sigma, \omega_1, \omega_2, \omega_3$ fields and a chosen direction in the laboratory reference frame. The value of $\langle \cos^4 \Phi \rangle$ varies between 0.2 for an isotropic distribution of polymer chains and 1 for perfectly aligned, respectively. It gives a factor of 5 difference in the macroscopic $\chi^{(3)}$ susceptibility. Some examples of third-order NLO susceptibilities for representative conjugated polymers are given in Table 6. These were obtained by third-harmonic generation technique which measures the fast, electronic contribution to $\chi^{(3)}$.

Table 5 Molecular off resonance second hyperpolarizability γ for selected organic molecules

Molecule	$\gamma(-3\omega; \omega, \omega, \omega)$ 10^{-50} m^5/V^2	λ_{meas} nm
Carbon tetrachloride	5.1	1064
	4.7	1907
Chloroform	3.8	1064
	3.2	1907
Acetone	2.5	1064
	2.6	1907
Methanol	1.16	1064
	1.21	1907
Ethanol	1.8	1064
	2.1	1907
Propanol	2.4	1064
	2.5	1907
Water	0.46	1064
Nitrobenzene	9.2	1907
Chlorobenzene	5.8	1907

Calibrated with silica value: $\chi^{(3)}(-3\omega; \omega, \omega, \omega) = 3.9 \times 10^{-22}$ m^2/V^2 at 1907 nm.

Two-Photon Absorbing Materials

Two-photon absorbing (TPA) materials are a special class of third-order NLO materials exhibiting a complex Kerr susceptibility $\chi^{(3)}(-\omega; \omega, -\omega, \omega)$. In this case, the nonlinear refractive index, derived from eqn [1] is complex too.

$$n = n_0 + n_2 I = n_0 + (n_2^r + \kappa_2) I \quad [19]$$

where I is the light intensity, n_2^r is the real part and κ_2 is the imaginary part (nonlinear extinction coefficient) of the nonlinear index of refraction, respectively. The nonlinear index of refraction is directly related to the Kerr susceptibility $\chi^{(3)}(-\omega; \omega, -\omega, \omega)$:

$$n_2 = \frac{3\chi^{(3)}(-\omega; \omega, -\omega, \omega)}{4\varepsilon_0 c n_0^2} \quad [20]$$

where c is the light velocity. If $\chi^{(3)}(-\omega; \omega, -\omega, \omega)$ is complex (close to a resonance) then n_2 is complex too and the light beam propagating in such a medium in direction z is attenuated through the two-photon absorption process:

$$I(z) = I_0 \, e^{-(\alpha_L + \alpha_{TPA} I)z} \quad [21]$$

Table 6 Resonant and nonresonant values of cubic susceptibility $\chi^{(3)}$ for selected organic materials

Material	*Form*	λ_{meas} nm	$\chi^{(3)}(-3\omega; \omega, \omega, \omega)$ 10^{-22} m^2/V^2	*Remark*
PDA-PTS	Single crystal	1890	119 (70) × 10^3	∥ chain, 3ph
			<1400	⊥ chain
		2620	22.4 (14) × 10^3	∥ chain
			<280	⊥ chain
PDA-TCDU	Single crystal	1890	9.8 (7) × 10^3	∥ chain
			<70	⊥ chain
		2620	51.8 (20) × 10^2	∥ chain
			<40	⊥ chain
DEANS	Single crystal	1900	1.4 × 10^4	∥ (mol. axis)
			127	⊥ (mol. axis)
PDA-AFA	LB film (blue)	1350	7.8 (1.4) × 10^3	2ph
PDA-AFA	LB film (blue)	1907	11.6 (1.2) × 10^3	3ph
PA	Isotropic oriented	1907	78.4 (22) × 10^4	3ph
			2.4 × 10^6	3ph
PDA-C_4UC_4	Oriented film	1907	2.9 × 10^4	∥ chain, 3ph
			3 × 10^2	⊥ chain
C_{60}	Isotropic film	1064	11.5 (1.1) × 10^3	3ph
C_{70}	Isotropic film	1420	13 (1.4) × 10^3	3ph
PPV	Isotropic film	1450	1.96 × 10^4	3ph

Numbers in brackets are standard deviations. Data calibrated with silica: $\chi^{(3)}(-3\omega;\omega,\omega,\omega) = 3.9 \times 10^{-22}$ m^2/V^2; 2ph, two photon resonance enhancement; 3ph, three photon resonance enhancement; for chemical structures see **Figure 4**.

Table 7 Two photon absorption cross-sections σ_{TPA}, maximum TPA wavelength (λ_{max}), measurement wavelength (λ_{meas}), pulse duration and measurements technique for a series of organic molecules

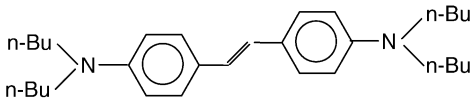
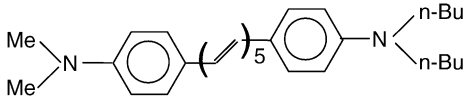
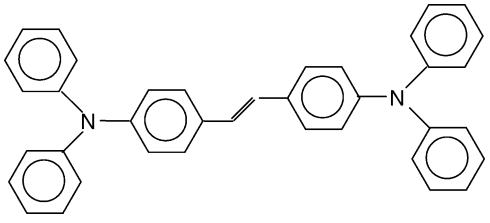
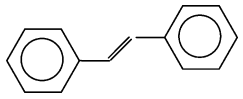
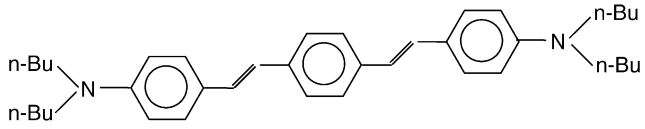
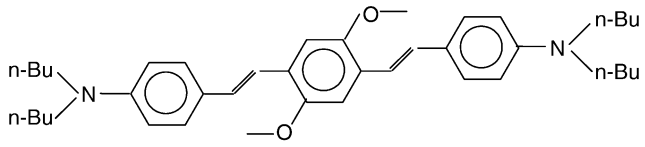
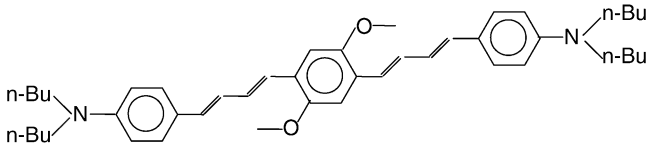
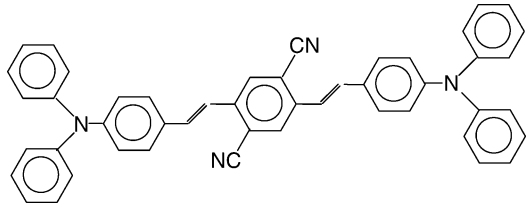
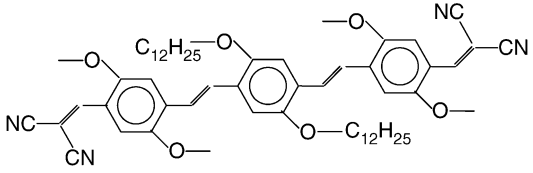
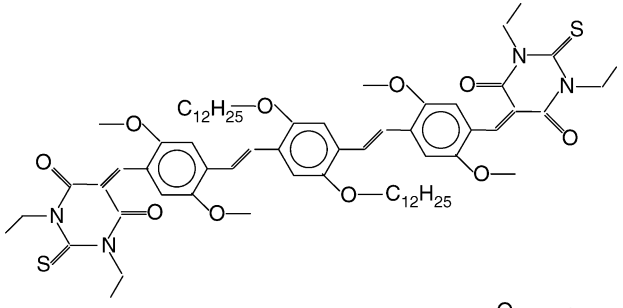
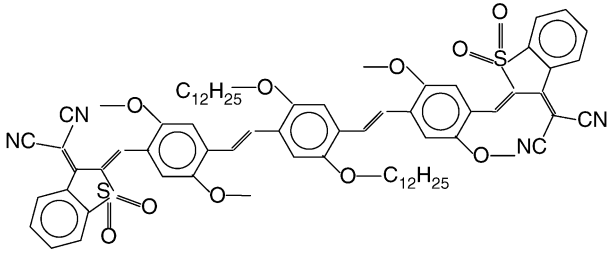
Molecule/material	*Chemical structure*	*Solvent*	λ_{max}^{TPA} *[nm]*	σ_{TPA} *[GM]*	*Pulse duration*	λ_{meas} *[nm]*	*Measurement technique*
BDBAS (Bis(di-*n*-butylamino)stilbene)		Toluene	600	310	4 ps	600	NLT
				9300	5 ns	600	
				1000	5 ns	700	
			600	210	5 ns	600	TPF
				6	5 ns	698	
				110	100 fs	620	
		Acetone		17 700	5 ns	600	NLT
		Toluene	755	10 000	5 ns	700	NLT
				17 000	5 ns	755	
				290	4 ps	700	
			730	300	5 ns	698	TPF
				790	5 ns	730	
BDPAS (Bis(diphenylamino)stilbene)		Toluene		13 000	5 ns	700	NLT
				110	4 ps	700	
Trans-stilbene		Toluene	514	12	5 ns	514	TPF
		Toluene	730	995	5 ns	730	TPF
			~725	635	100 fs	725	
		Toluene	730	900	5 ns	730	TPF
			~725	680	100 fs	725	

Table 7 (*Continued*)

Molecule/material	*Chemical structure*	*Solvent*	λ_{max}^{TPA} *[nm]*	σ_{TPA} *[GM]*	*Pulse duration*	λ_{meas} *[nm]*	*Measurement technique*
		Toluene	775 ~750	1250 1270	5 ns 100 fs	775 750	TPF
		Toluene	835 810	1940 3670	5 ns 100 fs	835 810	TPF
		Toluene	825 940 815 910	480 620 650 470	5 ns 100 fs 	825 940 815 910	TPF
		Toluene	970	1750	5 ns	970	TPF
		Toluene	975 945	4400 3700	5 ns 100 fs	975 945	TPF

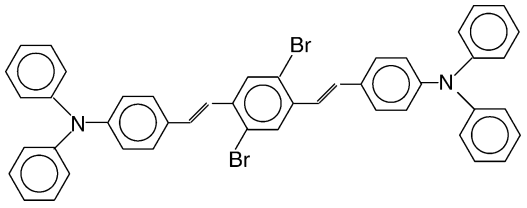
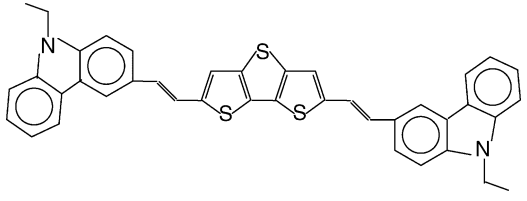
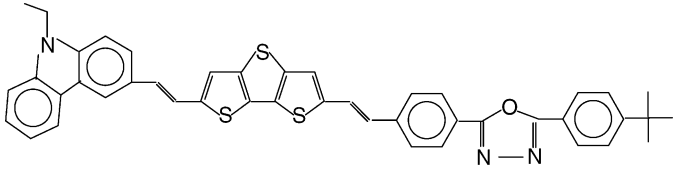
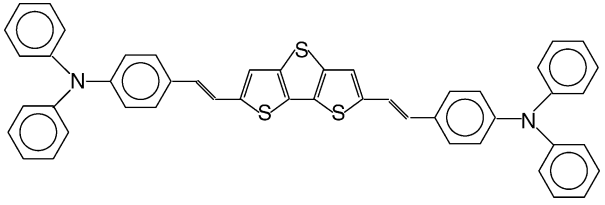
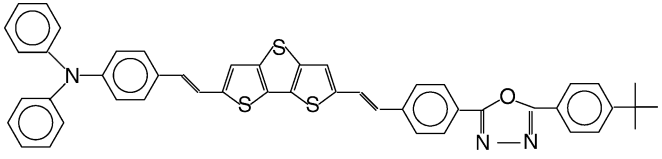
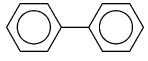
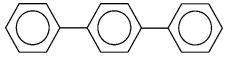
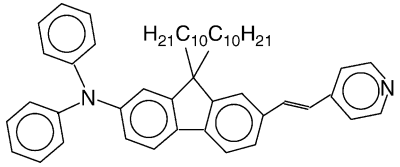
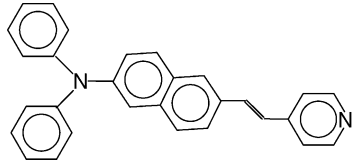
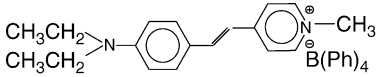
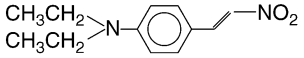
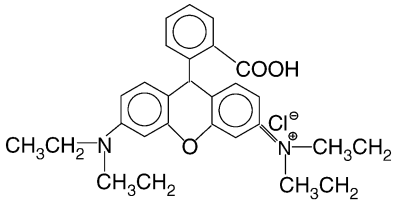
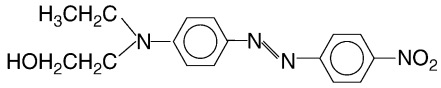
		Toluene	~800	450	5 ns	800	TPF
CZ-DTT-CZ (2,6 -Bis [2-(9-ethylcarbazol-3-yl) vinyl] thiopheno [3′,2′-2,3] thiopheno [4,5-b] thiophene)		1,1,2,2-Tetra chloro-ethane		105 000	8 ns	810	NLT
CZ-DTT-OX (2-[2-(4-(5-(4-*tert*-Buthyl phenyl-1,3,4-oxadiazol-2-yl) phenyl) vinyl)-6-[2-(9-ethyl carbazol-3-yl) vinyl] thiopheno [3′,2′-2,3] thiopheno [4,5-b] thiophene)		1,1,2,2-Tetra chloro-ethane		33 500	8 ns	810	NLT
TP-DTT-TP ([4-(2-(6-(2-(4-Diphenylamino) phenyl)vinyl) thiopheno [2′,3′-4,5] thiopheno [3,2-b] thiophen 2-yl-vinyl) phenyl] diphenylamine)		1,1,2,2-Tetra chloro-ethane		199 000	8 ns	810	NLT
TP-DTT-OX ([4-(2-(6-(2-(4-(5-(4-*tert*-Buthylphenyl-1,3,4-oxadiazol-2-yl)) phenyl) vinyl) thiopheno [2′,3′-4,5] thiopheno [3,2-b] thiophen 2-yl)-vinyl) phenyl] diphenylamine)		1,1,2,2-Tetra chloro-ethane		119 000	8 ns	810	NLT
Biphenyl				80			Theoretical calculus
Terphenyl				330			Theoretical calculus

Table 7 (*Continued*)

Molecule/material	*Chemical structure*	*Solvent*	λ_{max}^{TPA} *[nm]*	σ_{TPA} *[GM]*	*Pulse duration*	λ_{meas} *[nm]*	*Measurement technique*
Sexiphenyl				2090			Theoretical calculus
AF 50		Tetrahydro-furan Benzene		11 560 19 300	8 ns	800	NLT
AF 80		Tetrahydro-furan		6840	8 ns	800	NLT
ASPT		Poly hydroxyethyl metacrylate		12 000	10 ns	1064	NLT
DEANST		Tetrahydro-furan		200	46 ps	1064	KE
Rhodamine B		Methanol		50	100 fs	730	TPES
DR1 (4-[*N*-ethyl-*N*-(2-hydroxyethyl)] amino-4′-nitroazobenzene (dispersed red 1))		Tetrahydro-furan		102	42 ps	1064	KE

KE, optical Kerr effect; NLT, nonlinear transmission; TPF, two photon fluorescence; TPS, two photon excitation spectroscopy.
1 GM = 10^{-50} cm^4 s/photon-molecule.

Table 8 Two photon absorption coefficients for different molecules and extrapolated bulk values

Molecule	α_{TPA} *cm/GW*	λ_{max} *nm*	λ_{TPA} *nm*	*Concentration g/l*	*Solvent*	*Bulk value cm/GW*
DR 1	0.8	490	835	70	THF	11.4
DBANS	1.1	450	808	70	THF	15.7
DCNP	0.06	460	920	30	THF	2.3
DEANST	0.64	440	920	105	THF	6.1
DCM	0.1	480	950	11	DMSO	9.1
PDA-PTS	80	620	1064 not max of λ_{TPA}	Bulk	Single crystal	80
PDA-PTS	800	620	925	Bulk	Single crystal	800
Poly(3-octyl thiophene)	0.27	435 (142)	720 (94)	25	THF	9.6
Poly(3-octyl thiophene) regioregular	0.85	550 (170)	730 (70)	11.2	THF	76

λ_{max} corresponds to the maximum linear and λ_{TPA} to the maximum nonlinear absorption wavelengths, respectively, DCNP: 3-(1,1-dicyanoethenyl)-1-phenyl-4,5-dihydro-1H-pyrazole, for other molecules see **Figures 4** and **5** and preceding tables.

where α_L is the linear and α_{TPA} is the two-photon absorption coefficient, respectively. The last coefficient can be explicitly expressed by the imaginary part of the nonlinear index of refraction:

$$\alpha_{TPA} = \frac{2\pi\kappa_2}{\lambda_\omega} \qquad [22]$$

where λ_ω is the wavelength of the propagating beam. On the molecular level one also defines the two-photon absorption cross-section given by

$$\sigma_{TPA} = \frac{\hbar\omega}{N}\alpha_{TPA} \qquad [23]$$

where N is the number density of molecules. The values of TPA absorption cross-sections and two-photon absorption coefficients for selected organic molecules and solids are listed in Tables 7 and 8, respectively.

Photorefractive Materials

The photorefractive effect consists of the generation of a refractive index grating in an NLO material by its inhomogeneous illumination. This is a necessary condition but not a sufficient one. The sufficient condition consists of the possibility of creating the space charge field, which combined with the applied external field, leads to the refractive index modulation through the Pockels effect. In that case, the refractive index grating is phase mismatch with the intensity grating, leading to the transfer of photons from one beam to another, as occurs in two-beam coupling experiments, and thus applies to light amplification. Although the photorefractive effect was first observed in ferroelectric single crystals, the researchers at IBM demonstrated, at the beginning of 1990s, the photorefractive effect in organic composite materials. However, the main contribution in these materials comes from the rotational mobility of nonlinear optical (NLO) chromophores under the applied external electric field. Thus, to exhibit a photorefractive effect, the organic material has to possess the following properties:

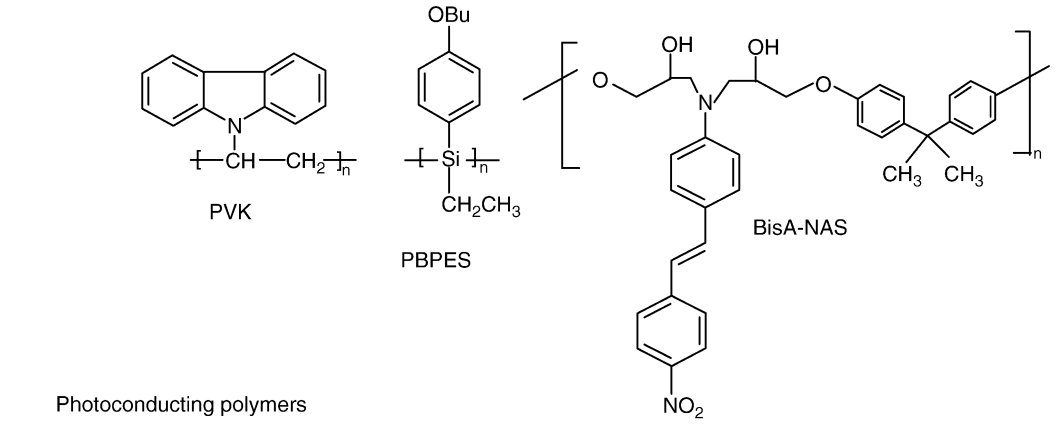
(i) charge generation and separation capacity;
(ii) charge mobility;
(iii) linear electro-optic (Pockels) effect; and
(iv) rotational mobility.

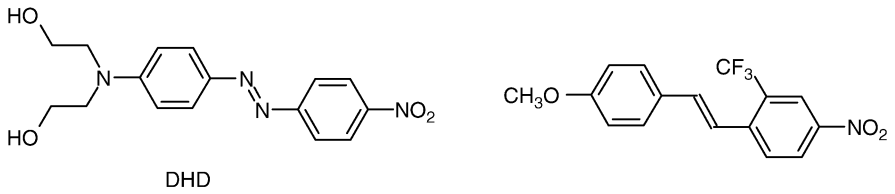
Composite Materials

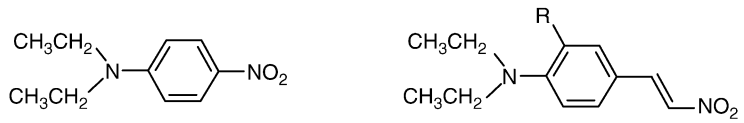
The different properties satisfying the above requirement to exhibit a photorefractive effect, are obtained by mixing together four types of molecules:

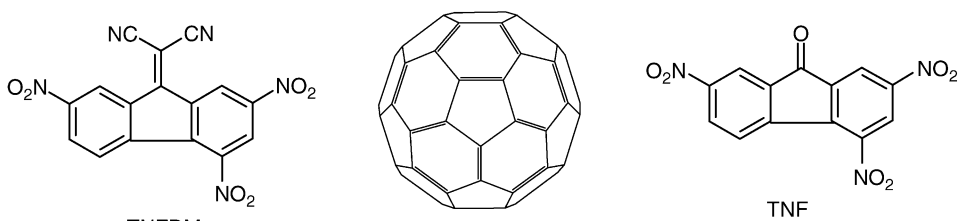
(i) photoconducting polymer;
(ii) highly second order nonlinear CT molecules;
(iii) photoensitizer for charge generation under illumination; and
(iv) plasticizer, lowering the glass transition temperature, thus increasing the rotational mobility.

Some examples of different types of molecules used are shown in **Figure 5**. Although these structures exhibit large exponential gain coefficients Γ and high diffraction efficiencies η, their main disadvantage consists of slow, seconds-duration response times









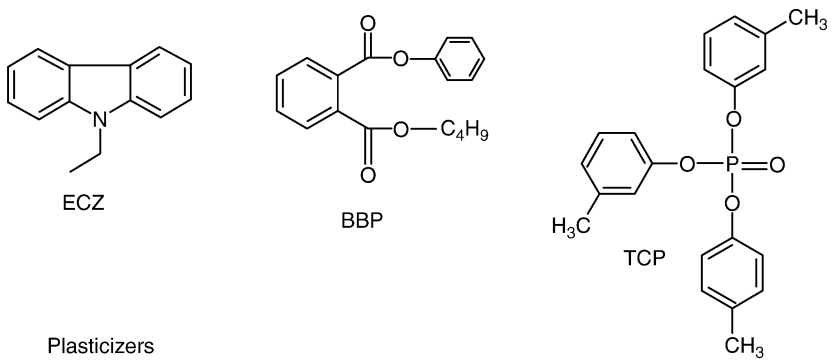


Figure 5 Chemical structures of photoconducting polymers, CT chromophores, photosensitizers and plasticizers commonly used for the fabrication of photorefractive polymers.

Table 9 Photorefractive properties of selected single crystals at zero external field, composite photorefractive polymers and liquid crystals

Structure	*Form*	*Thickness* μm	*Composition w%*	*Laser intensity* W/cm^2	*Exponential gain* $\Gamma\ cm^{-1}$	*Electric field* $V/\mu m$	*Response time s*	*Diffraction efficiency %*	λ_{meas} *nm*
COANP: TCNQ	Single crystal			0.08			100–1000 depending on grating period Λ		514.5
MNBA	Single crystal			0.1	2		10–100		514.5
DMNPAA/PVK/TNF/ECZ	Composite material	105	50: 33: 16: 1		220 (p-polarized) −40 (s-polarized)	90	1	86	514.5
Doped: 5 CB/C_{60}	Composite material	25	95: 5		2890	0.07	0.3		514.5
Doped: 8 OCB/5 CB /PER/NI	Composite material	37			640	0.04	0.04		514.5
SCLCP/E7/C_{60}	Composite material	50	49.975: 49.975: 0.05		600	0.7	3.5		632.8
PVA/E44/C_{60}	PDLC	25				40	On: 8 s Off: $>$3 days	52	
PMMA/E49/ETZ/TNF	PDLC	53	45: 33: 21: 1		41	22		8	675
POT/LC	HPCPLC	10			965	0.6	0.011	26	514.5
POT/DR1/LC	HPCPLC	10			2600	0.6	0.010	31	514.5
PVK/TNF/LC	HPCPLC	10			3700	0.9	0.010	45	514.5

LC, liquid crystals; PDLC, polymer dispersed liquid crystals; HPCPLC, hybrid photoconducting polymer–liquid crystal; COANP-2-cyclooctalamino-5 nitropyridine, TCNQ, 7,7,8,8-tetracyanoquinodimethane; MNBA4′, nitrobenzylidene-3 acetamino-4-methoxyaniline; 5-CB-4′, (*n*-octyloxy)-4-cyanobiphenyl; 8OCB-4′, (*n*-octhyloxy)-4-cyanobiphenyl; PER, perylene; NI, *N,N*-di(*n*-octyl)1,4,5,8-naphtalenediimide; SCLCP, side chain liquid crystalline polymer (PMMA-4-cyanophenyl benzoate); E7, E44, E49, commercial LC mixtures (Merck, Germany); other molecules in **Figure 5** and **Tables 2** and **3**.

and require large external operation electric fields (~100 V/μm).

Photorefractive Liquid Crystals

In order to improve the performances of organic photorefractive materials the use of nematic liquid crystals was proposed as active media. These materials exhibit a large rotational mobility, requiring a relatively small electric field with response time of milliseconds. The reorientation of domains requires $2 \div 3$ orders of magnitude electric fields as compared with the composite materials (reorientation of individual molecules). Several types of structures of liquid crystal based photorefractive structures are realized, such as:

- composite materials in which the liquid crystals mesogens are NLO chromophores;
- polymer dispersed liquid crystals (PDLCs);
- hybrid photoconducting polymers – liquid crystals (HPCPLCs), in which the rotational mobility is separated from the charge mobility. Moreover the use of photoconducting polymer allows the modulation of the electric field on the liquid crystal layer, enhancing the effect.

The performances of different, representative, organic photorefractive structures are listed in Table 9.

See also

Nonlinear Optics, Basics: Photorefraction; $\chi^{(2)}$–Harmonic Generation. **Polarization:** Introduction.

Further Reading

Agranovich VM and La Rocca GC (eds) (2002) Organic Nanostructures: Science and Applications. *Proceed. International School of Physics "Enrico Fermi", Course CXLIX*. Amsterdam: IOS Press.

Bosshard Ch, Sutter K, Pretre Ph, Hulliger J, *et al.* (1995) In: Kajzar F and Garito AF (eds) *Organic Nonlinear Optical Materials*, *Advances in Nonlinear Optics*, vol. 1. Amsterdam: Gordon and Breach Publishers.

Bredas JL and Silbey R (eds) (1991) *Conjugated Polymers*. Dordrecht: Kluwer Academic Publishers.

Charra F, Agranovich V and Kajzar F (2003) *Organic Nanophotonics*. NATO Science Series 2. Dordrecht: Kluwer Academic Publishers.

Chemla DS and Zyss J (eds) (1987) *Nonlinear Optical Properties of Organic Molecules and Polymers*. Orlando, FL: Academic Press.

Graja A, Bulka B, Kajzar F (eds) (2002) *Molecular Low Dimensional-and Nano-Structured Materials for Advanced Applications*. NATO Science Series 2. Dordrecht: Kluwer Academic Publishers.

Hornac LA (ed.) (1992) *Polymers for Lightwave and Integrated Optics. Technology and Applications*. New York: Marcel Dekker, Publishers.

Kajzar F and Agranovich VM (eds) (2000) *Multiphoton and Light Driven Multielectron Processes in Organics: New Phenomena, Materials and Applications*. Dordrecht: Kluwer Academic Publishers.

Kajzar F and Reinisch R (eds) (1998) *Beam Shaping and Control with Nonlinear Optics*. NATO ASI Series, Series B: Physics, vol. 369. New York: Plenum Press.

Kajzar F and Swalen J (eds) (1996) *Thin Films for Waveguiding Nonlinear Optics*. Amsterdam: Gordon and Breach Scientific Publishers.

Kajzar F, Agranovich VM and Lee CY-C (1996) *Photoactive Organic Materials: Science and Application*. Dordrecht: Kluwer Academic Publishers.

Kuzyk M and Dirk C (eds) (1998) *Characterization Techniques and Tabulations for Organic Nonlinear Optical Materials*. New York: Marcel Dekker.

Lee KS and Wegner G (eds) (2003) *Polymers for Photonics Applications II: Nonlinear Optical, Photorefractive and Two-Photon Absorption Polymers*. Advances in Polymer Sciences 161. Berlin: Springer-Verlag.

Lessard R and Franke H (eds) (1997) *Materials Characterization and Optical Probe Techniques*. Critical Reviews of Optical Science and Technology, vol. CR69. Bellingham: SPIE.

Messier J, Kajzar F, Prasad P and Ulrich D (1989) *Nonlinear Optical Effects in Organic Polymers*. Dordrecht: Kluwer Academic Publishers.

Messier J, Kajzar F and Prasad P (1991) *Organic Molecules for Nonlinear Optics and Photonics*. Dordrecht: Kluwer Academic Publishers.

Miyata S and Sasabe H (eds) (1997) *Poled polymers and their Application to SHG and EO Devices*. Amsterdam: Gordon and Breach Scientific Publishers.

Prasad PN and Williams DJ (1991) *Introduction to Nonlinear Optical Effects in Molecules and Polymers*. New York: Wiley.

Richardson T (ed.) (1999) *Functional Polymers and Polymeric Materials*. Chichester, UK: John Wiley & Sons.

MICROSCOPY

Contents

Overview
Confocal Microscopy
Imaging Multiple Photon Fluorescence Microscopy
Interference Microscopy
Nonlinear Microscopy
Phase Contrast Microscopy

Overview

C J R Sheppard, National University of Singapore, Singapore

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The basic construction and imaging properties of the optical microscope, one of the most important optical instruments, are described. But the restrictive meaning of a microscope as a device that the user looks down to see a magnified image is far from the present terminology that includes scanning systems (where the image is stored in a computer), nonvisible radiation, such as electrons or X-rays, and a variety of different contrast mechanisms.

According to the *Oxford English Dictionary* a microscope is 'an instrument magnifying objects by means of lenses so as to reveal details invisible to the naked eye'. Thus, basically a microscope forms a magnified image of an object. The first microscopes used visible light and formed an image showing variations in the intensity of light scattered by the object. This intensity is in general related to the optical properties of the object in a complicated way.

Nowadays there are many different types of microscope. Our understanding of a microscope must therefore be generalized in the following ways.

- The microscope may not be a conventional imaging system using lenses or mirrors. For example, it may be a scanning imaging system. A confocal system is a combination of a conventional and a scanning system. Another important category of microscopes is that of probe microscopes, including near-field optical microscopes. Some probe microscopes, for example the atomic force microscope (AFM) do not even seem at first glance to rely on the direct use of radiation.
- The microscope may not use visible light, but other electromagnetic radiation, or even other forms of radiation.
- The image can be formed using a variety of different contrast mechanisms. Some of these, such as phase contrast, are designed to image particular optical properties of the sample. Others image the generation of a form of radiation when the object is stimulated by another form of radiation. In particular, in principle, virtually any form of spectroscopy can be the basis of building up an image by measuring the spatial variations in the signal.

Different Forms of Microscopes

The Conventional Microscope

In a simple microscope, an objective lens forms a real, magnified, and inverted image of an object. In a compound microscope, an eyepiece is added, forming a virtual image that is viewed by eye to give real image on the retina. The eye can be replaced by CCD camera to record moving or still images. In early microscopes, the objective lens and eyepiece are mounted in the two ends of a brass tube. The objective screws into the tube using an RMS (Royal Microscopical Society) thread (0.8″ Whitworth 36 threads per inch), until its mounting face is flush with the end of the tube. The length of the tube is the mechanical tube length. The distance from the mounting face of the objective to the plane of the real image is the optical tube length, which varied according to different manufacturers in the range 160–210 mm. Now most objectives have infinity tube length, so that they collimate the light from

the object. The collimated light is brought to a focus by an additional tube lens. An objective of correct tube length should always be used as otherwise spherical aberration is introduced. As the objective, tube lens, and eyepiece are designed together as a system, great care should be exercised when mixing components from different manufacturers.

Abbe showed that in order to form a perfect lateral image the objective must satisfy a different condition called the Herschel condition. Otherwise the image of an off-axis point will suffer from coma. An aberration-free system satisfying the sine condition is called an aplanatic system. A microscope produces a three-dimensional (3D) image of a 3D object. A point of the object closer to the lens appears further from the lens in the image. If the lateral magnification is M, the longitudinal magnification is approximately equal to M^2. However, it can be shown that it is impossible to devise an instrument that can produce a perfect 3D image of a 3D object, because to have a perfect transverse image the system must satisfy Abbe's sine condition, but to have a perfect axial image the system must satisfy the Herschel condition. These two conditions are not mutually compatible. Actually, we find that the longitudinal magnification is not exactly equal to M^2, and is not even constant in space. Different depths of the object are imaged differently: the microscope objective only produces a good image for one plane of focus, and other planes will exhibit spherical aberration. However, all of these problems can be overcome by bringing the part of the object under observation into focus, without refocusing the eye or other detector. A 3D image can thus be generated by scanning the object stage in the axial direction, and recording a stack of images from different depths in a computer. For modern objectives of infinite tube length, focusing can be alternatively achieved by piezoelectric scanning of the objective lens.

Resolution

For an object consisting of a single bright point object in a dark background, the image produced by a perfect microscope, according to paraxial diffraction theory, is the so-called Airy disk, consisting of a bright central spot, surrounded by a series of rings. The intensity in a meridional cross-section through the Airy disk is the same for any aperture of objective, or wavelength of light, but its width varies. The intensity of the first bright ring is 1.75% of that at the peak. The radius of the first dark ring of the Airy disk is

$$r_0 = \frac{0.61\lambda}{n \sin \alpha} \qquad [1]$$

where λ is the wavelength, n is the refractive index of the immersion medium, and α is the angle subtended at the edge of the objective aperture. The radius depends on $n \sin \alpha$, which is called the numerical aperture and should be as large as possible for good resolution. Thus, high-magnification lenses often use an immersion fluid, usually oil with a refractive index of 1.518 (ISO 8036/1).

The optical coordinate v is defined as

$$v = \frac{2\pi}{\lambda} rn \sin \alpha \qquad [2]$$

so that the dark ring of the Airy disk occurs at a value of v of $2\pi \times 0.61 = 3.83$. Resolution is sometimes expressed in terms of the full width at half maximum (FWHM) of the image of a point object, which is 3.232 in optical coordinates. A full nonparaxial theory does not give a very different figure.

Resolution of a microscope is often specified by two-point resolution, which describes whether the image of two points can be distinguished from that of a single point. According to the Rayleigh criterion of two-point resolution, two points are just resolved if the second point is placed on the first dark ring of the first. The separation is then $0.61\lambda/(n \sin \alpha)$. A cross-section through the image in an incoherent optical system of two bright point objects of equal strength and different separations is shown in **Figure 1**. The intensity in the image between the points decreases as the separation is increased. The ratio of the intensity midway between the points to that at the points for the Rayleigh criterion to be satisfied is then found to be 0.735.

The concept of resolution should be distinguished from sensitivity or precision. An object much smaller than the resolution limit can still be detected, perhaps weakly, in a microscope: this detection of weak contrast depends on the sensitivity of the microscope rather than its resolution. The size of an object can also be measured with a precision much greater than the resolution of the microscope.

Illumination

If the object is not self-luminous, in order to be imaged in a microscope it must be illuminated. A semitransparent object, such as a biological slice, is illuminated in transmission. For observation of bulk objects, or surfaces, we use illumination in the reflection geometry, called epi-illumination. Either a tungsten halogen lamp or an arc source is usually used. In critical illumination, the source is focused on to the object by a condenser lens. The disadvantage of this approach is that variations in emission of the source are imaged directly into the image.

Cheaper microscopes avoid this problem by using a diffuser. Better microscopes use instead the Köhler illumination system, where the source is placed in the front focal plane of the condenser lens. Figure 2 shows a Köhler illumination system in which the source is imaged into the front focal plane of the condenser lens by another, projector, lens. The system incorporates a separate field stop and aperture stop. In practice, the size of the field stop should be reduced to illuminate as small a region of the object as necessary in order to minimize stray light. An important part of setting up a microscope to operate properly consists in centering the aperture stop of the condenser. This is often done using a so-called Bertrand lens that allows imaging of the aperture stop.

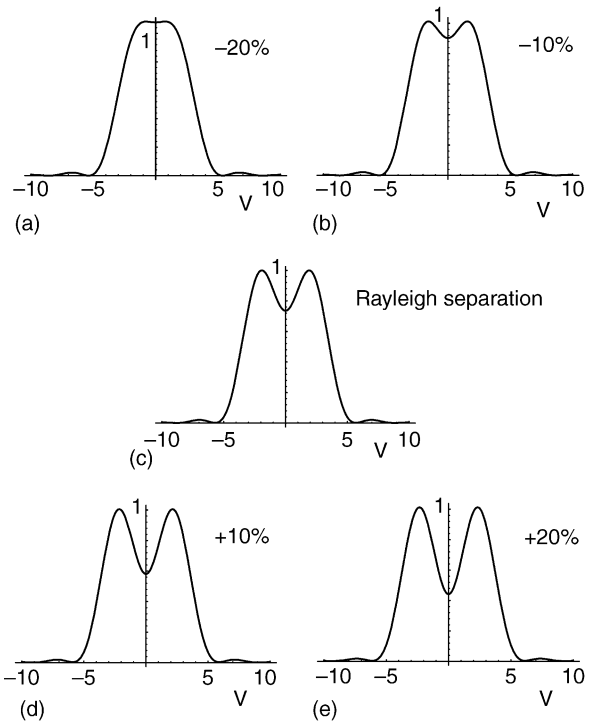


Figure 1 A cross-section through the image of two equal point objects in an incoherent microscope. The separation of the points is that for the Rayleigh criterion to be satisfied, and for changes of ±10, 20% of the Rayleigh separation.

Image formation

The Rayleigh criterion was originally specified for an incoherent optical system. It is thus applicable in fluorescence microscopy, as fluorescence is an incoherent process as two fluorescent point objects fluoresce independently. But for the image of a trans-illuminated object, the resolution depends on the coherence of the illumination. This is controlled by the numerical aperture of the condenser lens as compared with that of the objective. The ratio of these numerical apertures is called the coherence parameter, S, so that

$$S = \frac{n_c \sin \alpha_c}{n \sin \alpha} \qquad [3]$$

For $S = 0$ the illumination is purely coherent, while for $S \rightarrow \infty$ imaging becomes purely incoherent, thus corresponding to the Rayleigh criterion, so S should really be called an incoherence parameter. An important in-between case is when $S = 1$, when the apertures of objective and condenser are equal. This is termed matched, full, or complete illumination. According to the generalized Rayleigh criterion, the points are just resolved when the ratio of the intensity midway between the points to that at the points is 0.735, which is called the generalized Rayleigh criterion. We require that the distance between the points is as small as possible when this occurs. The generalized Rayleigh separation is the same for incoherent illumination and for $S = 1$, and corresponds to a distance of $2v_0 = 3.83$ in optical coordinates. Note, however, that for other separations of the points, the ratio is different for these two cases: it is not true, but often erroneously stated, that $S = 1$ corresponds to incoherent imaging. Resolution improves as the aperture of the condenser is increased ($2v_0 = 5.15$ in optical coordinates for coherent illumination) reaching a maximum when $S = 1.46$, when resolution is 9% better than for incoherent imaging ($2v_0 = 3.58$). This maximum in resolution is achieved when the numerical aperture of the condenser is larger than that of the objective, which is not

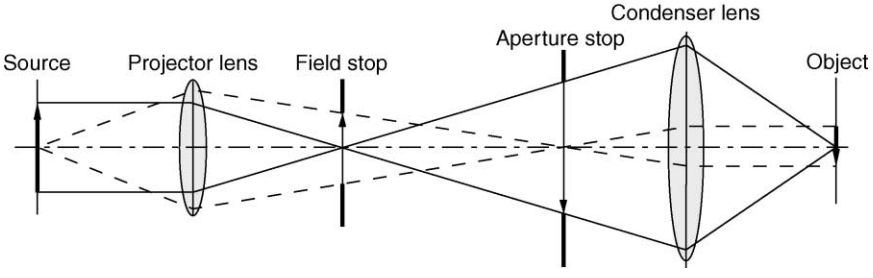


Figure 2 Geometry of the Köhler illumination system. In practice an arrangement with a smaller number of lenses is usually used.

physically achievable if the objective has the highest possible numerical aperture. Similarly, it is impossible to achieve incoherent illumination with very large objective numerical apertures.

The effect of coherence on the resolution of an image depends on the form of the object, and on the particular resolution criterion employed. In practice, contrast also depends on the aperture of the condenser. Usually, opening the condenser improves resolution, but reduces contrast, so there will be an optimum condenser aperture size.

Abbe theory of microscope imaging

Abbe argued that for coherent illumination of a grating, the grating spectra can be observed in the back focal plane of the objective. The tube lens then forms an image from the grating components. Thus, this is an early description of Fourier optics. The strength of the grating components can be altered in the back focal plane to give various optical effects, such as phase contrast. Abbe's theory was not properly appreciated at the time because it was known that in practice, resolution could be improved by using a larger condenser aperture, which does not give coherent illumination. There was much controversy about the merits of Köhler versus critical illumination, and different sizes of condenser aperture.

We now know that Köhler and critical illumination are in principle equivalent. The aberrations of the condenser lens are not important, so that the source and condenser together behave simply as a partially coherent effective source. Imaging can still be described by Fourier optics within the framework of partially coherent imaging theory.

Although Köhler and critical illumination are equivalent in principle, Köhler illumination is in practice better because of its improved illumination uniformity.

Depth of focus

The longitudinal image of a point object is found to be approximately invariant when expressed in terms of an optical coordinate u, defined as

$$u = \frac{8\pi}{\lambda} zn \sin^2 \frac{\alpha}{2} \qquad [4]$$

The axial resolution, and thus also the depth of focus, varies strongly with aperture. The FWHM for longitudinal imaging is 11.13 in optical coordinates.

Microscope objectives

Because microscope objectives often have large apertures, aberrations will be very strong, unless the objective is designed correctly. The basic design principle for a high aperture lens is the aplanatic front element (Figure 3). It is found that all rays from a point A in a sphere of refractive index n, where A is a distance r/n from its center, appear to come from a point B, distance rn, without any aberration. This principle is used by employing a front element that has one surface of the same radius of curvature as the sphere. The front surface is a sphere centered on A (Figure 3b), so that the rays are not deviated on crossing it.

A low-power achromat uses just two components with different optical dispersion to cancel chromatic aberration and correct for other aberrations (an achromatic doublet). For higher powers, two separated doublets are necessary. For the highest apertures, an aplanatic front is used. However, this can never converge the light, but only makes it less divergent. It is thus used together with an achromatic doublet. For oil immersion, the interface between the oil and the front element is not very important, so it is made planar to simplify manufacture. In practice, it is

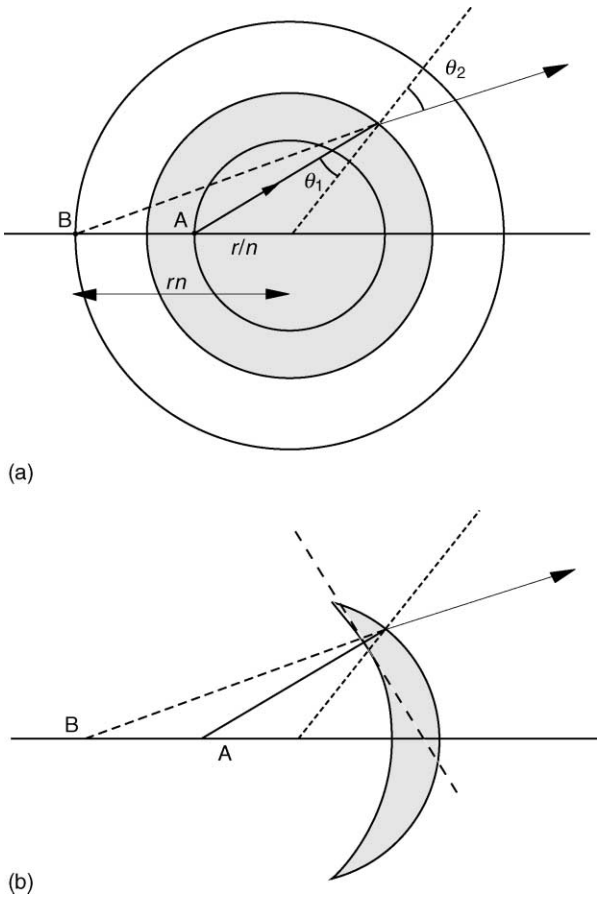


Figure 3 The principle of the aplanatic front: (a) an aplanatic surface; (b) a meniscus lens that acts as an aplanatic system.

usually made planar even for a dry lens, and the resulting aberrations cancelled out elsewhere in the objective. Apochromats, corrected for three colors, use more elements, which may include two or more stages of aplanatic front. The final objective thus consists of a number of elements that must be accurately aligned relative to one another. The objective is adjusted by the manufacturer to give a good star image (image of a point object). First a spacer is selected to optimize spherical aberration. Then a sleeve is adjusted to make the objective parfocal with others. Finally a screw is used to center the assembly to remove coma.

Conventional and Scanning Microscopes

Conventional microscope

Various different types of transmission microscope are illustrated in Figure 4. We consider image formation in a conventional microscope as illustrated schematically in Figure 4a, showing a microscope with critical illumination. A large area incoherent source is focused by the condenser lens on to the specimen, illuminating a comparatively large area of the specimen, corresponding to the whole field of view of the objective. Information from each illuminated point in the specimen is simultaneously transmitted by the objective lens to form the primary image. The objective is responsible for forming the image, with the condenser playing only a secondary role in determining the resolution of the system, through control of the coherence of the illumination. Figure 4b shows a conventional microscope in which the image is measured point by point by a detector. In practice this could be achieved by using a CCD detector.

Scanning microscope

An image can be generated by a scanning system, as illustrated in Figure 4c. A probe of light is formed by demagnification of a source, and is scanned over the object in a raster. The transmitted (or reflected) light is detected by a photodetector and thus builds up an image. The size of the probe of light limits the resolution of the system, that is, the smallest detail in the object that can be seen in the image. It should be noted that the magnification of the image is given simply by the ratio of the distance scanned in the image to the distance scanned by the probe. It is thus unrelated to the demagnification of the source. The image magnification can be altered without changing the lens and is not in any way related to the resolution.

It has been shown that the imaging properties of scanning and conventional microscopes are identical under analogous conditions, this being known as the principle of equivalence. This property is based on the principle of reciprocity, which is a very general physical law that holds even with diffraction,

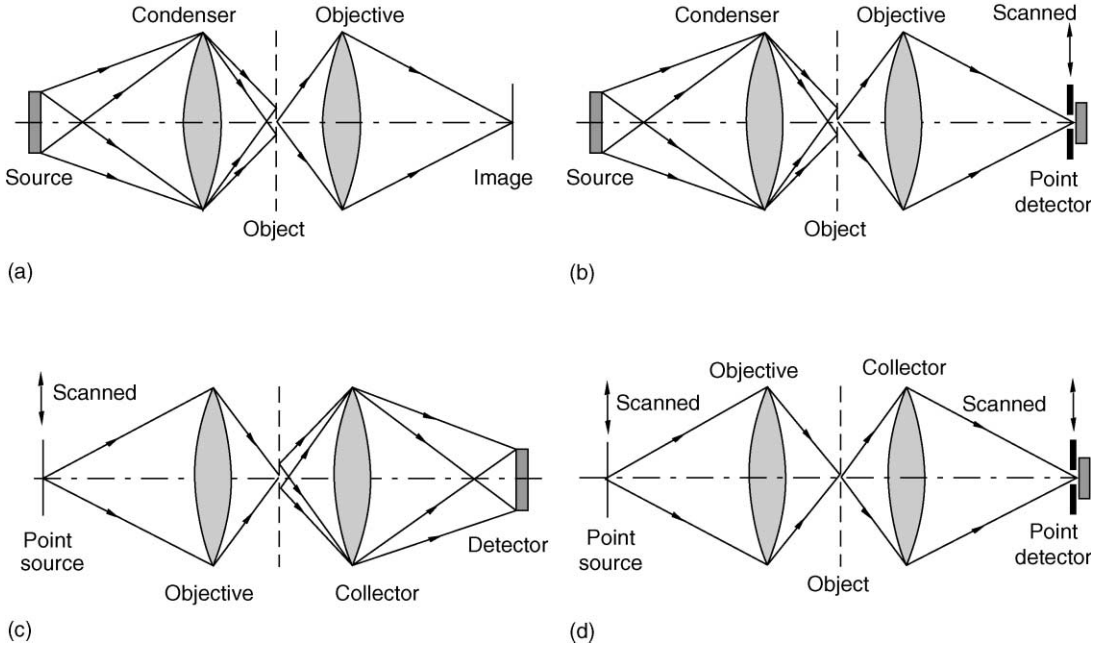


Figure 4 Different geometries of transmission microscope: (a) a conventional microscope; (b) a conventional microscope with a point detector; (c) a scanning microscope; (d) a confocal microscope.

absorption, multiple scattering, and stray light. The principle of equivalence is generally valid except if there are nonreciprocal magnetic or polarization effects, or energy losses involved (such as in fluorescence microscopy, and with inelastic scattering in electron microscopy).

Scanning systems exhibit a number of important advantages over conventional systems. Broadly, these are based on two classes of property. First, in a scanning system the image is in the form of an electronic signal. As a result, this is advantageous for quantitative measurements, as well as for image processing, including image enhancement, image restoration, and image analysis. Second, in a scanning system the object is illuminated by a focused spot, which extends the range of imaging modes available.

Confocal microscope

Finally in Figure 4d, we combine the arrangements of Figures 4b and c to give a confocal scanning optical microscope, in which a point source illuminates just a small region of the object, and a confocal point detector detects light from this illuminated region. If the point source and detector are scanned in unison, a two-dimensional image is generated. However, this system now behaves very differently from the previous ones. The confocal microscope is thus not a special case of the general partially coherent conventional imaging system. From the symmetry of Figure 4d, it is clear that the two lenses play an equal part in the imaging process. This results in an improvement in resolution. In fact, the confocal system behaves as a coherent imaging system, but with a sharper effective point spread function than in a conventional coherent microscope.

Although Figure 4 is drawn for the transmission geometry, in practice most confocal systems operate in the reflection or epi-illumination mode, in which the same objective lens is used both for illumination and detection. For a specimen placed in the focal plane, the properties of confocal transmission and reflection systems are identical. However, once the object is moved from the focal plane, certain differences arise: in particular in confocal reflection a strong optical sectioning effect occurs that allows a single section through a thick object to be imaged. This is the major advantage of the confocal microscope arrangement. Figure 4 also applies equally well to fluorescence imaging. However, in this case, because of the incoherent nature of fluorescence emissions after excitation by coherent light, imaging is then incoherent.

Confocal microscopes can be achieved by using either a point detector, in practice performed by placing a pinhole in front of a photodetector, or by using a coherent detector. Such a coherent detector, sensitive to the amplitude of the incident radiation, is not directly available for light, but is available for acoustic radiation, as in a scanning acoustic microscope. For light, a confocal effect can also be achieved by using an interferometric method to synthesize a coherent detector. Thus, in interference microscopes using the high aperture condenser and objective lenses, an optical sectioning effect arises similar to that in the confocal microscope. Another way of producing a coherent detector is to use a single-mode optical fiber.

Probe Microscopes

Scanning tunneling microscope

The scanning tunneling microscope was the first of the family of probe microscopes, relying on a fundamentally different principle from the usual forms of microscope. A physical tip in the nanometer range is brought close, also in the nanometer range, to a conducting sample. If an electric potential is applied, electrons can tunnel across between the sample and the tip. By scanning the tip mechanically across the sample, an image can be generated with a resolution in the subnanometer range. The surface topography can also be measured with subnanometer sensitivity. While interference microscopy can measure surface topography with subnanometer sensitivity, in this case the profile is averaged over the lateral resolution of the microscope, of the order of the wavelength in dimensions. The scanning tunneling microscope can image individual atoms. The signal is related to the work function of the material. Altering the bias voltage allows the band structure of the material to be investigated.

Atomic force microscope

In the atomic force microscope, the force between the surface and the tip is measured using a cantilever beam and used to build up an image. Atomic force microscopy can be performed in either a contact or noncontact mode. It can be used with insulating specimens. Optical methods, often using a position-sensitive detector, are used to measure displacements of the cantilever.

Near-field scanning optical microscope (NSOM)

In a near-field scanning optical microscope, a very small tip or aperture is scanned relative to the specimen to attain resolution greater than the classical limit set by the wavelength of the radiation. As Figure 5 shows, various different designs of near-field microscope have been proposed. It is evident that

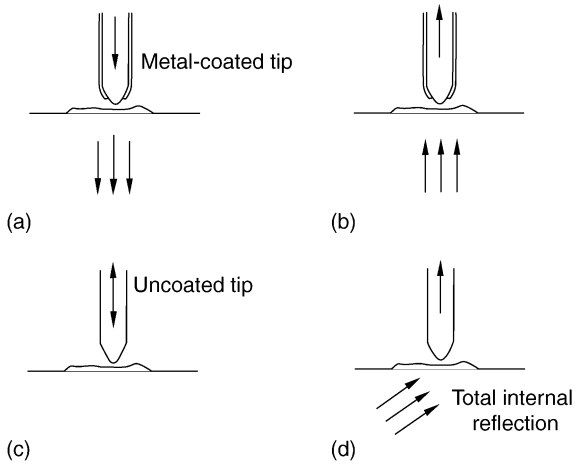


Figure 5 Different forms of near-field microscope: (a) illumination-mode NSOM; (b) collection-mode NSOM; (c) a near-field probe used for both illumination and collection in a confocal arrangement; (d) photon tunneling microscope (PTM). Reproduced with permission from Sheppard CJR (1978) The scanning optical microscope. *The Physics Teacher* 16: 648–651. © 1978 American Association of Physics Teachers.

these again form into the conventional/scanning categories described earlier. The sample can be illuminated using the near-field probe, as in Figure 5a. This is called the illumination-mode NSOM. Or the sample can be uniformly illuminated, and a signal detected using a near-field probe as in Figure 5b, giving a collection-mode NSOM. Or a confocal arrangement can be used, with a near-field probe used for both illumination and collection (c). In this case, an uncoated tip can be used; as otherwise the detected signal will be very small. Finally, Figure 5d shows the photon tunneling microscope (PTM), in which the sample is illuminated with evanescent waves produced by total internal reflection and an uncoated tip used to probe the evanescent field in the presence of the sample.

Different Types of Radiation

Electromagnetic Radiation

UV radiation and X-rays

The first microscopes used visible light, but, in fact, we do not need to use visible light, but can use any electromagnetic radiation, over a broad range of different available wavelengths. In Figure 6 we show the wavelength of the electromagnetic spectrum, illustrating how the wavelength in air varies with frequency. Instead of using visible light, shorter wavelengths allow greater resolution to be achieved. Ultra-violet light is commonly used to excite fluorescence. X-rays have also successfully been used to image biological samples.

IR and microwave radiation

In the longer wavelength region, infra-red radiation is used for observation of semiconductors or for molecular spectroscopic imaging. Imaging of emitted midinfra-red radiation can show up variations in temperature, such as hot spots in semiconductors. Microwaves have also been used, but resolution is limited by the long wavelength. For this reason microwaves (and mid-IR) have also been used for near-field microscopy.

Other Radiation

Electron microscope

The other important class of radiation that can be used for illumination is matter waves. The most common of these, of course, is electrons, the wavelength of which, as a function of acceleration voltage, is shown in Figure 7. The bend in the curve at very high voltages is caused by relativistic effects. The wavelength of the electrons is very small: even for 100 V electrons, the wavelength is only 0.1 nm. But aberrations caused by the electron lenses limit

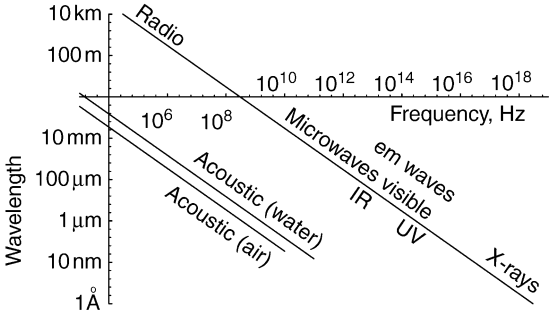


Figure 6 The wavelength of different forms of electromagnetic radiation and acoustic radiation. Reproduced with permission from Sheppard CJR (2002) The generalized microscope. In: Diaspro A (ed.) *Confocal and Two-Photon Microscopy: Foundations, Applications and Advances*, pp. 1–18. New York: Wiley-Liss.

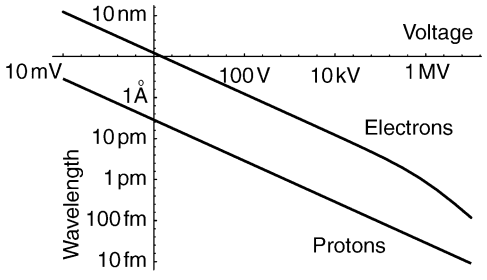


Figure 7 The wavelength of matter waves (electrons and protons) as a function of accelerating voltage. Reproduced with permission from Sheppard CJR (2002) The generalized microscope. In: Diaspro A (ed.) *Confocal and Two-Photon Microscopy: Foundations, Applications and Advances*, pp. 1–18. New York: Wiley-Liss.

resolution so that at present atomic resolution is only just achievable, even with much higher voltages.

In a transmission electron microscope (TEM), a beam of electrons is produced by an electron gun. The specimen is illuminated via a condenser lens (or lenses). Usually these are magnetic lenses in which a current in a coil produces a magnetic field that focuses the electrons. The electrons that are transmitted through the specimen are then focused by a series of lenses to form a final image that can be viewed directly on a phosphor screen. An X-ray detector can be used to characterize the elements in the specimen. An electron-energy-loss spectrometer, together with a detector, can measure the energy of the electrons in a particular region of the image.

Other matter radiation

Instead of electrons, other particle beams can be used for illumination. Figure 7 shows the wavelength of a beam of protons, as function of the accelerating voltage. Because of the larger mass of protons, relativistic effects are not apparent for protons in the curve in Figure 7. Microscope images have also been formed using ions, and even neutral particles such as neutrons. Neutral atom microscopes are also under development.

Acoustic radiation

Instead of electromagnetic waves, acoustic radiation can be used to form a scanning acoustic microscope (SAM). Ultrasonic waves are focused into a water immersion medium. Mechanical scanning is again used to build up an image. Figure 6 shows the wavelength of the acoustic wave as a function of frequency. By using microwave frequencies, wavelengths, and therefore resolution, in the submicron range, can be generated. The SAM acts as a confocal microscope because the detector is sensitive to the amplitude of the acoustic wave. The acoustic microscope provides information on variations in the elastic properties (viscosity and elasticity) of the sample.

Different Contrast Mechanisms

The bright field microscope detects the intensity of the light transmitted through the object. A microscope with a condenser aperture of appreciable size behaves as a partially coherent system, and the image will, in general, depend on the phase of the light coming from the object as well as its intensity. Many objects are weakly scattering, so the contrast of this phase information is weak. In this case, the microscope effectively images variations in the transmittance or reflectance of the sample. Different designs of microscope can be used to image various properties of the object, thus providing new and complementary information. Phase contrast microscopy provides but one example of a different contrast mechanism. In fact, in general, any physical or chemical interaction of the illuminating radiation with the sample can be used as the basis for a contrast mechanism. These different contrast mechanisms can be performed in either a conventional or scanning arrangement, or in a confocal system.

Phase Contrast Microscopy

Phase contrast is a widely used technique in both biological microscopy (in transmission) and materials microscopy (in reflection). Its importance in biological microscopy stems, firstly, from the fact that contrast may otherwise be too weak to be visible, but secondly because the phase variations are caused by changes in optical thickness, related to physical density. In reflection microscopy, phase is related to surface height variations, but also depends on the optical material properties of the sample.

Polarization Microscopy

The polarization state of light transmitted through or reflected from a sample can also be changed. This change is detected in polarization microscopy, for example by placing the sample between crossed polarizers. The birefringence detected can be either material birefringence, or form birefringence (caused by the shape of the microstructure). Polarization microscopy is widely used in biological microscopy, and in mineralogy.

Fluorescence Microscopy

In fluorescence microscopy, electrons excited by illumination decay to the ground state and emit photons of light. Autofluorescence is natural fluorescence of the sample. Fluorescent dyes can be used as labels for particular biological or chemical constituents. Fluorescence microscopy is usually performed in the epi-illumination geometry to assist rejection of light from the source. Often it is performed in a confocal microscope, allowing 3D localization of the fluorescent labels.

Raman Microscopy

Raman spectroscopy can be performed on a microscopic scale to produce images of presence of particular molecular bonds. It is usually performed in a scanning system, often of the confocal type.

This is one example of the wider class of spectroscopic microscopies: in principle, any form of spectroscopy can be used as the basis for a microscope contrast mode.

Nonlinear Microscopy

Another class of contrast mechanisms is based on nonlinear optical interactions. Again, any nonlinear mechanism can be used as the basis for a contrast mechanism. These include two- (or multiphoton) fluorescence, second- (or higher-order) harmonic generation, sum or difference frequency generation, and coherent anti-Stokes Raman scattering (CARS). Probably other nonlinear optical mechanisms will be exploited in microscopy in the future.

Other Contrast Mechanisms

The contrast mechanisms described above, all rely on illumination and detection of light. But in many other modes, radiation of one form can be converted into another. These include cathodoluminescence microscopy (electrons to light), X-ray microanalysis (electrons to X-rays), photo-emission microscopy (light, usually UV, to electrons), photo-acoustic microscopy (light to ultrasonics), photothermal microscopy (light to thermal waves), optical or electron beam induced current (OBIC/EBIC). Again, in many of these, the illuminating or emitted radiation can be focused, or a combination can be used in a confocal arrangement.

See also

Microscopy: Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy; Interference Microscopy; Nonlinear Microscopy; Phase Contrast Microscopy.

Further Reading

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge: Cambridge University Press.

Bradbury S and Bracegirdle B (1997) *Introduction to the Optical Microscope*. Oxford: Bios Scientific Publishers.

Inoue S and Spring KR (1997) *Video Microscopy: The Fundamentals*, 2nd edn. New York: Plenum.

Martin LC (1966) *Theory of the Microscope*. London: Blackie.

Paesler MA and Moyer P (1996) *Near-Field Optics: Theory, Instrumentation, and Applications*. New York: Wiley InterScience.

Pluta M (1988) *Advanced Light Microscopy, Volume 1: Principles and Basic Properties*. Amsterdam: Elsevier.

Sheppard CJR (1987) Scanning optical microscopy. In: Barer R and Cosslett VE (eds) *Advances in Optical and Electron Microscopy*, vol. 10, pp. 1–98. London: Academic Press.

Sheppard CJR (2002) The generalized microscope. In: Diaspro A (ed.) *Confocal and Two-Photon Microscopy: Foundations, Applications, and Advances*, pp. 1–18. New York: Wiley-Liss.

Slayter EM and Slayter HS (1992) *Light and Electron Microscopy*. Cambridge: Cambridge University Press.

Wiesendanger R (1994) *Scanning Probe Microscopy and Spectroscopy*. Cambridge: Cambridge University Press.

Wilson T and Sheppard CJR (1984) *Theory and Practice of Scanning Optical Microscopy*. London: Academic Press.

Confocal Microscopy

T Wilson, University of Oxford, Oxford, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

It is probably fair to say that the development and wide commercial availability of the confocal microscope has been one of the most significant advances in light microscopy in the recent past. The main reason for the popularity of these instruments derives from their ability to permit the structure of thick specimens of biological tissue to be investigated in three dimensions. It achieves this important goal by resorting to a scanning approach together with a novel (confocal) optical system.

The traditional wide-field conventional microscope is a parallel processing system which images the entire object field simultaneously, which is a severe requirement for the optical components. We can relax this requirement if we no longer try to image the whole object at once. The limit of this relaxation is to require a good image of only one object point at a time. The price that we have to pay is that we must scan in order to build up an image of the entire field.

A typical arrangement of a scanning confocal optical microscope is shown in Figure 1 where the system is built around a host conventional microscope. The essential components are some form of mechanism for scanning the light beam (usually from a laser) relative to the specimen and appropriate photodetectors to collect the reflected or transmitted light.

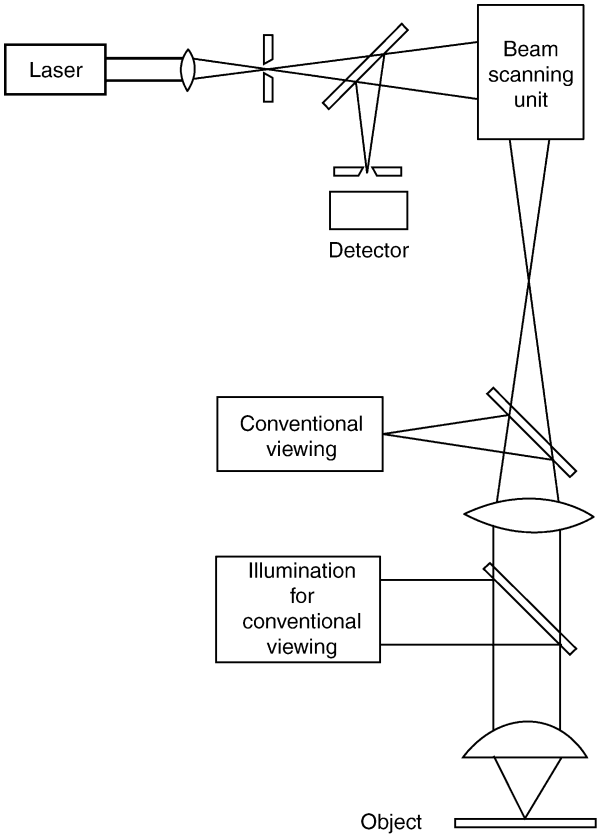


Figure 1 Schematic diagram of a confocal microscope.

Most of the early systems were analog in nature, but it is now universal, due to the serial nature of the image formation, to use a computer both to drive the microscope and to collect, process, and display the image.

In the beam scanning confocal configuration of Figure 1, the scanning is typically achieved by using vibrating galvanometer-type mirrors or acousto-optic beam deflectors. The use of the latter gives the possibility of TV-rate scanning, whereas vibrating mirrors are often relatively slow when imaging an extended region of the specimen, although significantly higher scanning speeds are achievable over smaller scan regions. We should note that other approaches to scanning may be implemented, such as specimen-scanning and lens-scanning. These methods, although not generally available commercially, do have advantages in certain specialized applications. Since this article is necessarily limited in length much additional material may be found in other sources listed in the Further Reading at the end of this article.

Image Formation in Scanning Microscopes

We will not discuss the fine detail of the optical properties of confocal systems. However, the essence is shown in Figure 1 where we see that the confocal optical system consists simply of a point source of light which is then used to probe a single point on the specimen. The strength of the reflected or fluorescence radiation from the single object point is then measured via a point, pinhole detector. The confocal – point source and point detector – optical system therefore merely produces an 'image' of a single object point and hence some form of scanning is necessary to produce an image of an extended region of the specimen. However, the use of single-point illumination and single-point detection results in novel imaging capabilities which offer significant advantages over those possessed by conventional wide-field optical microscopes. In essence, these are enhanced lateral resolution and, perhaps more importantly, result in a unique depth discrimination or optical sectioning property. It is this latter property which leads to the ability to obtain three-dimensional images of volume specimens.

The improvement in lateral resolution may at first seem implausible. However, it can be explained simply by the principle which states that resolution can be increased at the expense of field of view, which can then be increased by scanning. One way of taking advantage of this principle is to place a very small aperture extremely close to the object. The resolution is now determined by the size of the aperture rather than the radiation. In the confocal microscope we do not use a physical aperture in the focal plane but rather use the back-projected image of a point detector in conjunction with the focused point source. Figure 2 indicates the improvement in lateral resolution that may be achieved.

The confocal principle was introduced in an attempt to obtain an image of a slice within a thick specimen which was free from the distracting presence of out-of-focus information from surrounding planes. The confocal optical systems fulfills this requirement and its inherent optical sectioning or depth discrimination property has become the major

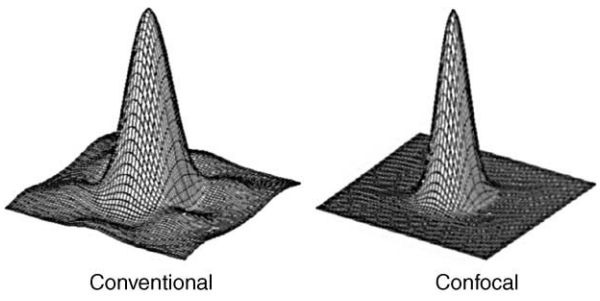


Figure 2 The point spread functions of the conventional and confocal microscopes showing the improvement in lateral resolution which may be obtained in the confocal case.

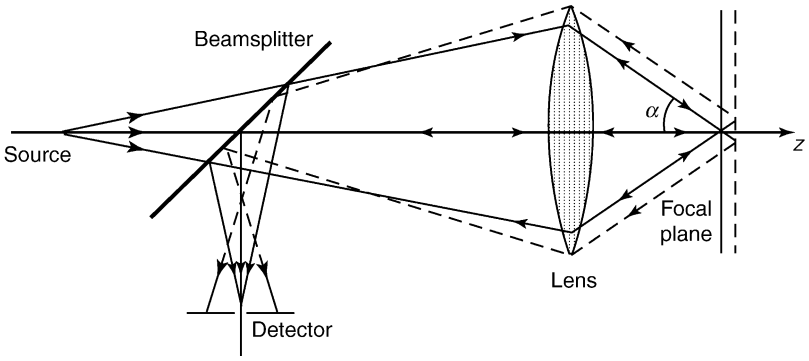


Figure 3 The origin of the optical sectioning or depth discrimination property of the confocal optical system.

motivation for using confocal microscopes, and is the basis of many of the novel imaging modes of these instruments. The origin of the depth discrimination property may be understood easily from Figure 3, where we show a reflection-mode confocal microscope and consider the imaging of a specimen with an undulating surface. The full lines show the optical path when an object feature lies in the focal plane of the lens. At a later scan position, the object surface is supposed to be located in the plane of the vertical dashed line. In this case, simple ray tracing shows that the light reflected back to the detector pinhole arrives as a defocused blur, only the central portion of which is detected and contributes to the image. In this way the system discriminates against features which do not lie within the focal region of the lens. A very simple method of both demonstrating the effect and giving a measure of its strength is to scan a perfect reflector axially through focus and measure the detected signal strength. Figure 4 shows a typical response. These responses are frequently termed the $V(z)$, by analogy with a similar technique in scanning acoustic microscopy. A simple paraxial theory models this response as

$$I(u) = \left[\frac{\sin(u/2)}{u/2}\right]^2 \qquad [1]$$

where u is a normalized axial coordinate which is related to real axial distance, z, via

$$u = \frac{8\pi}{\lambda} nz \sin^2(\alpha/2) \qquad [2]$$

where λ is the wavelength and $n \sin \alpha$ the numerical aperture. As a measure of the strength of the sectioning, we can choose the full width at half intensity of the $I(u)$ curves. Figure 5 shows this value as a function of numerical aperture for the specific case of imaging with red light from a helium neon laser. These curves were obtained using a high aperture theory which is more reliable than eqn [1]

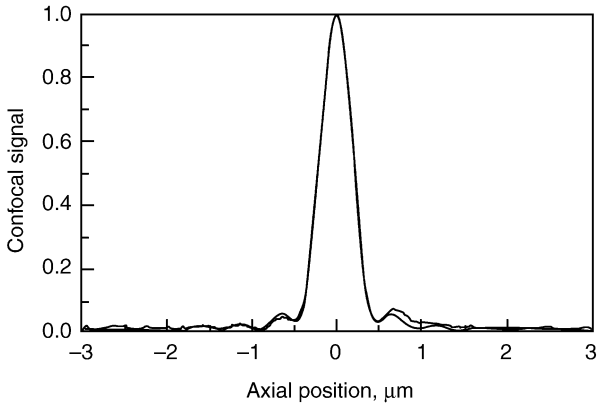


Figure 4 The variation in detected signal as a plane reflector is scanned axially through focus. The measurement was taken with a 1.3 numerical aperture objective and 633 nm radiation.

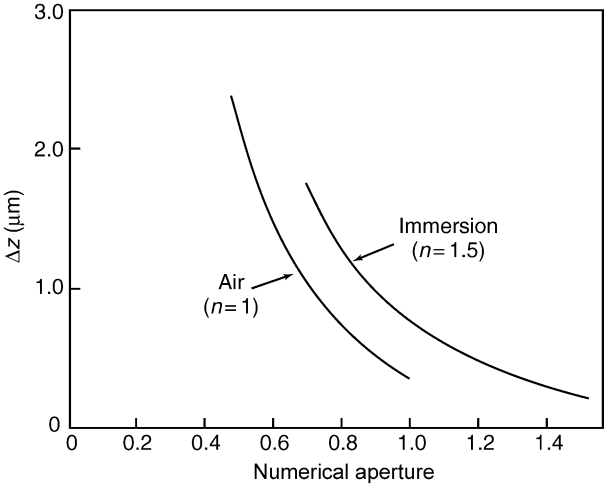


Figure 5 The optical sectioning width as a function of numerical aperture. The curves are for red light (0.6328 μm wavelength). Δz is the full width at the half-intensity points of the curves of $I(u)$ against u.

at the highest values of numerical aperture. We note, of course, that these numerical values refer to nonfluorescence imaging. The qualitative explanation of optical sectioning, of course, carries over to the fluorescence case but the actual value of the optical sectioning strength is different in the fluorescence case.

Applications of Depth Discrimination

Since this property is one of the major reasons for the popularity of confocal microscopes, it is worthwhile, at this point, to review briefly some of the novel imaging techniques which have become available with confocal microscopy.

Figure 6 illustrates the essential effect: Figure 6a shows a conventional image of a planar microcircuit which has deliberately been mounted with its normal at an angle to the optic axis. We see that only one portion of the circuit, running diagonally, is in focus. Figure 6b shows the corresponding confocal image: here the discrimination against detail outside the focal plane is clear. The areas which were out of focus in Figure 6a have been rejected. Furthermore, the confocal image appears to be in focus throughout the visible band, which illustrates that the sectioning property is stronger than the depth of focus.

This suggests that if we try to image a thick translucent specimen, we can arrange, by the choice of our focal position, to image detail exclusively from one specific region. In essence, we can section the specimen optically without having to resort to mechanical means. Figure 7 shows an idealized schematic of the process. The portion of the beehive-shaped object that we see is determined by the focus position. In this way it is possible to take a through-focus series and obtain data about the three-dimensional structure of the specimen. If we represent the volume image by $I(x, y, z)$, then by focusing at a position $z = z_1$, we obtain, ideally, the image $I(x, y, z_1)$. This, of course, is not strictly true in practice because the optical section is not infinitely thin.

It is clear that the confocal microscope allows us to form high-resolution images with a depth of focus sufficiently small that all the detail which is imaged appears in focus. This suggests immediately that we can extend the depth of focus of the microscope by adding together (integrating) the images taken at different focal settings without sacrificing the lateral resolution. Mathematically, this extended-focus image is given by

$$I_{\mathrm{EF}}(x, y) = \int I(x, y, z)\mathrm{d}z \qquad [3]$$

As an alternative to the extended-focus method, we can form an auto-focus image by scanning the object axially and, instead of integrating, selecting the focus at each picture point by recording the maximum in the detected signal. Mathematically, this might be written as

$$I_{\mathrm{AF}}(x, y) = I(x, y, z_{\max}) \qquad [4]$$

where $z_{\max}$ corresponds to the focus setting giving the maximum signal. The images obtained are somewhat similar to the extended focus and, again, substantial increases in depth of focus may be obtained. We can go one step further and turn the microscope into a noncontacting surface profilometer. Here we simply display $z_{\max}$.

It is clear by now that the confocal method gives us a convenient tool for studying three-dimensional structures in general. We essentially record the image as a series of slices and play it back in any desired fashion. Naturally, in practice it is not as simple as this, but we can, for example, display the data as an $x - z$ image rather than an $x - y$ image. This is somewhat similar to viewing the specimen from the side. As another

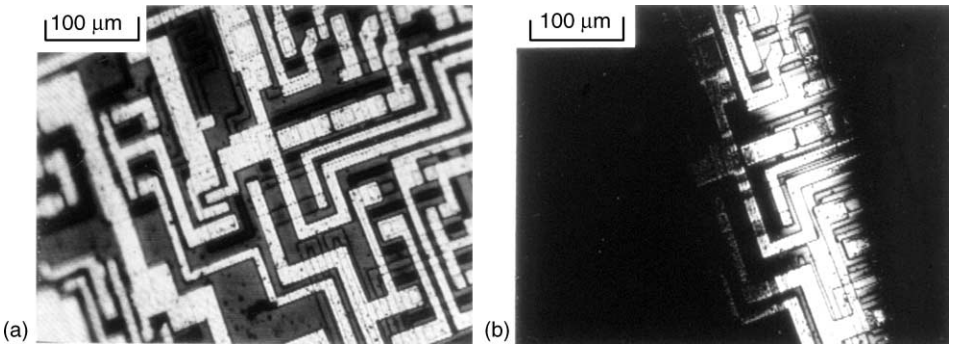


Figure 6 (a) Conventional scanning microscope image of a tilted microcircuit: the parts of the object outside the focal plane appear blurred. (b) Confocal image of the same microcircuit: only the part of the specimen within the focal region is imaged strongly.

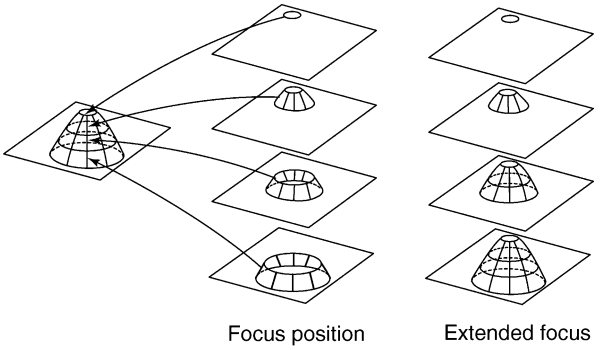


Figure 7 An idealization of the optical sectioning property showing the ability to obtain a through-focus series of images, which may then be used to reconstruct the original volume object at high resolution.

example, we might choose to recombine the data as stereo pairs by introducing a slight lateral offset to each image slice as we add them up. If we do this twice, with an offset to the left in one case and the right in another, we obtain, very simply, stereo pairs. Mathematically, we form images of the form:

$$\int I(x \pm \gamma z, y, z)\mathrm{d}z \qquad [5]$$

where γ is a constant. In practice it may not be necessary to introduce offsets in both directions to obtain an adequate stereo view.

Our discussion so far on these techniques has been by way of a simplified introduction. In particular, we have not presented any fluorescence images. The key point is that, in both brightfield and fluorescence modes, the confocal principle permits the imaging of specimens in three dimensions. The situation is, of course, more involved than we have implied. A thorough knowledge of the image formation process, together with the effects of lens aberrations and absorption, is necessary before accurate data manipulation can take place.

In conclusion to this section it is important to emphasize that the confocal microscope does not produce three-dimensional images. It essentially produces very high-quality two-dimensional images of a (thin) slice within a thick specimen. A three-dimensional rendering of the entire volume specimen may then be generated by suitably combining a number of these two-dimensional image slices from a through focus series of images.

Fluorescence Microscopy

We now turn our attention to confocal fluorescence microscopy, because this is the imaging mode which is usually employed in biological applications. Although we have introduced the confocal microscope in terms of bright-field imaging the comments we have made concerning the origin of the optical sectioning, etc., carry over directly to the fluorescence case although the numerical values describing the strength of the optical sectioning are, of course, different and we will return to this point later.

If we assume that the fluorescence in the object destroys the coherence of the illuminating radiation and produces an incoherent fluorescent field proportional to the intensity of the incident radiation, $I(v, u)$, then we may write the effective intensity point spread function, which describes image formation in the incoherent confocal fluorescence microscope, as

$$I(v, u)I\left(\frac{v}{\beta}, \frac{u}{\beta}\right) \qquad [6]$$

where the optical coordinates u and v are defined relative to the primary radiation and $\beta = \lambda_2/\lambda_1$ is the ratio of the fluorescence to the primary wavelength. We note that $v = (2\pi/\lambda)rn \sin \alpha$, where r denotes the actual radial distance.

This suggests that the imaging performance depends on the value of β. In order to illustrate this we show in Figure 8 the variation in detected signal strength as perfect fluorescent sheet scans through focus. This serves to characterize the strength of the optical sectioning in fluorescence microscopy in the same way that the mirror was used in the brightfield case. We note that the half-width of these curves is

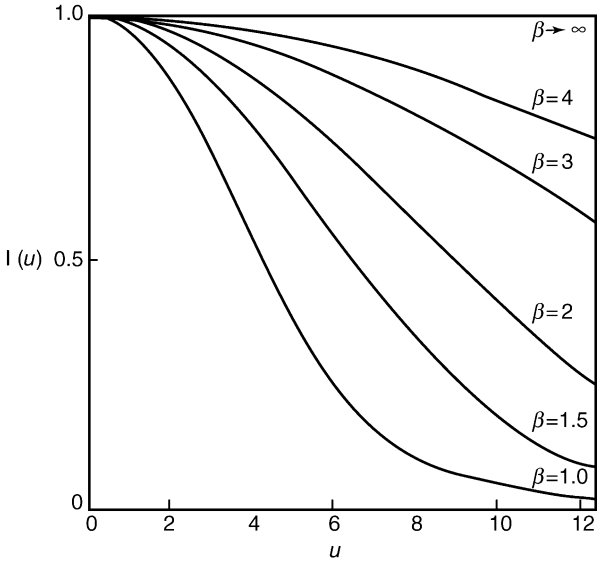


Figure 8 The detected signal as thin fluorescent sheet is scanned axially through focus for a variety of fluorescent wavelengths. We note that if we measure the sectioning by the half width of these curves, the strength of the sectioning is essentially proportional to β.

essentially proportional to β and so for optimum sectioning the wavelength ratio should be as close to unity as possible.

We have just discussed what we might call one-photon fluorescence microscopy in the sense that a fluorphore is excited by a single photon of a particular wavelength. It then returns to the ground state and emits a photon at the (slightly longer) fluorescence wavelength. It is this radiation which is detected via the confocal pinhole. Recently, however, much interest has centered on two-photon excitation fluorescence microscopy. This process relies on the simultaneous absorption of two, longer wavelength photons, following which a single fluorescence photon is emitted. The excitation wavelength is typically twice that used in the one-photon case. The beauty of the two-photon approach lies in the quadratic dependence of the fluorescence intensity on the intensity of the illumination. This leads to fluorescence emission which is always confined to the region of focus. In other words the system possesses an inherent optical sectioning property. Other benefits of two-photon fluorescence over the single photon case include the use of red or infra-red lasers to excite ultra-violet dyes, confinement of photo-bleaching to the focal region (the region of excitation), and the reduced effects of scattering and greater penetration. However, it should also be remembered that, compared with single photon excitation, the fluorescence yield of many fluorescent dyes under two-photon excitation is relatively low.

In order to make a theoretical comparison between the one- and two-photon modalities, we shall assume that the emission wavelength λ_{em} is the same, irrespective of the mode of excitation. Since the wavelength required for single-photon excitation is generally shorter than the emission wavelength, we may write it as $\gamma\lambda_{em}$ where $\gamma < 1$. Since two-photon excitation requires the simultaneous absorption of two photons of half the energy we will assume that the excitation wavelength may be written as $2\gamma\lambda_{em}$ which has been shown to be a reasonable approximation for many dyes.

If we now introduce optical coordinates u and v, normalized in terms of λ_{em}, we may now write the effective point spread functions in the one-photon confocal and two-photon case as

$$I_{1p\text{-}conf} = I\left(\frac{v}{\gamma}, \frac{u}{\gamma}\right)I(v,u) \qquad [7]$$

and

$$I_{2p} = I^2\left(\frac{v}{2\gamma}, \frac{u}{2\gamma}\right) \qquad [8]$$

respectively. We note that although a pinhole is not usually employed in two-photon microscopy it is possible to include one if necessary. In this confocal two-photon geometry the effective point spread function becomes

$$I_{2p\text{-}conf} = I^2\left(\frac{v}{2\gamma}, \frac{u}{2\gamma}\right)I(v,u) \qquad [9]$$

If we now look at eqns [7] and [8] in the $\gamma = 1$ limit, we find that

$$I_{1p\text{-}conf} = I^2(v,u) \qquad [10]$$

and

$$I_{2p} = I^2\left(\frac{v}{2}, \frac{u}{2}\right) \qquad [11]$$

We now see that because of the longer excitation wavelength used in two-photon microscopy the effective point spread function is twice as large as that of the one-photon confocal in both the lateral and axial directions. The situation is somewhat improved in the confocal two-photon case but it is worth remembering that the advantages of two-photon excitation microscopy are accompanied by a reduction in optical performance as compared to the single-photon case.

From the practical point of view it is worth noting that the two-photon approach has certain very important advantages over one-photon excitation, in terms of image contrast, when imaging through scattering media apart from the greater depth of penetration afforded by the longer wavelength excitation. In a single-photon confocal case it is quite possible that the desired fluorescence radiation from the focal plane may be scattered after generation in such a way that it is not detected through the confocal pinhole. Since the fluorescence is generated throughout the entire focal volume it is also possible that undesired fluorescence radiation, which was not generated within the focal region, may be scattered so as to be detected through the confocal pinhole. In either case this leads to a reduction in image contrast. The situation is, however, completely different in the two-photon case. Here the fluorescence is generated only in the focal region and not throughout the focal volume. Furthermore since all the fluorescence is detected via a large area detector – no pinhole is involved – it is not so important if further scattering events take place. This leads to high-contrast images which are less sensitive to scattering. This is particularly important for specimens which are much more scattering at the fluorescence ($\lambda/2$) wavelength than the excitation (λ) wavelength.

The Use of Structured Illumination to Achieve Optical Sectioning

An alternative approach to obtain optical sectioning is to use structured illumination in a conventional microscope. This approach is attractive since the conventional light microscope already possesses many desirable properties: real-time image capture, standard illumination, ease of alignment, etc. However, it does not produce optically sectioned images in the sense usually understood in confocal microscopy. In order to see how this deficiency may be corrected via a simple modification of the illumination system let us begin by looking at the theory of image formation in a conventional fluorescence microscope and start by asking in what way the image changes as the microscope is defocused. We know that in a confocal microscope the image signal from all object features attenuates with defocus and that this does not happen in a conventional microscope. However, when we look closely at the image formation process we find that it is only the zero spatial frequency (constant) component which does not change with defocus (Figure 9); all other spatial frequencies actually do attenuate with defocus to a greater or lesser extent. Figure 10 illustrates this by showing the image of a single spatial frequency one-dimensional bar pattern object for increasing degrees of defocus. When the specimen is imaged in focus, a good image of the bar pattern is obtained. However with increasing defocus the image becomes progressively poorer and weaker until it eventually disappears leaving a uniform gray level. This observation is the basis of a simple way to perform optical sectioning in a conventional microscope.

If we simply modify the illumination path of the microscope so as to project onto the object the image of a one-dimensional, single spatial frequency fringe pattern, then the image we see through the microscope will consist of a sharp image of those parts of the object where the fringe pattern is in focus together with an out-of-focus blurred image of the rest of the object. In order to obtain an optically sectioned image it is necessary to remove the blurred out-of-focus portion as well as the fringe pattern from the in-focus optical section. There are many ways to do this – one of the simplest involves simple processing of three images taken at three different spatial positions of the fringe pattern. The out-of-focus regions remain fairly constant between these images and the relative spatial shift of the fringe pattern allows the three images to be combined in such a way as to remove the fringes. This permits us to retrieve both an optically sectioned image as well as a conventional image in real time.

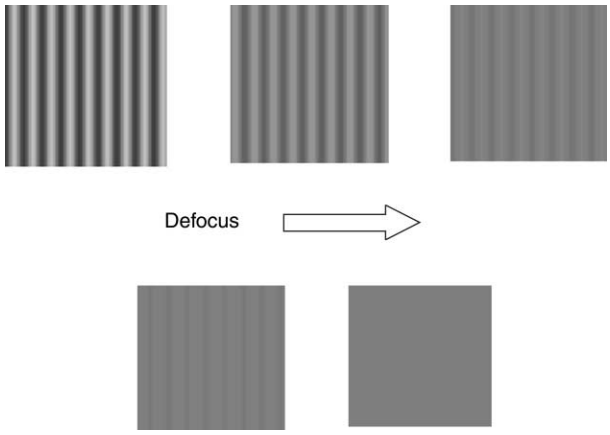


Figure 10 The image of a one-dimensional single spatial frequency bar pattern for varying degrees of defocus. We note that for sufficiently large values of defocus the bar pattern is not imaged at all.

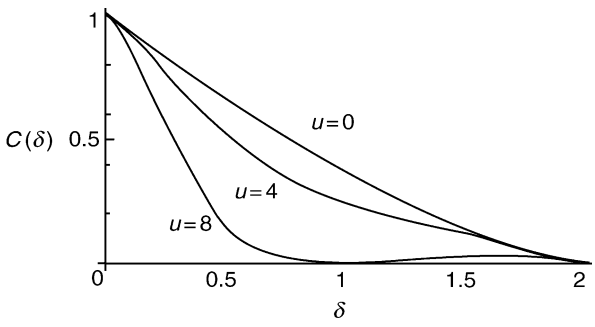


Figure 9 The transfer function for a conventional fluorescence microscope for a number of values of defocus (measured in optical units). We notice that the response for all nonzero spatial frequencies decays with defocus.

Since the approach involves processing three conventional microscope images, the image formation is fundamentally different from that of the confocal microscope. However, the depth discrimination or optical sectioning strength is very similar and this approach, which requires very minimal modifications to the instrument, has been used to produce high-quality three-dimensional images of volume objects which are directly comparable to those obtained with confocal microscopes. A schematic of the optical arrangement is shown in Figure 11a, together with experimentally obtained axial responses in Figure 11b. It will be noted that these responses are substantially similar to those obtained in the true confocal case. As an example of the kind of images which can be obtained with this kind of microscope, we show in Figure 12 two images

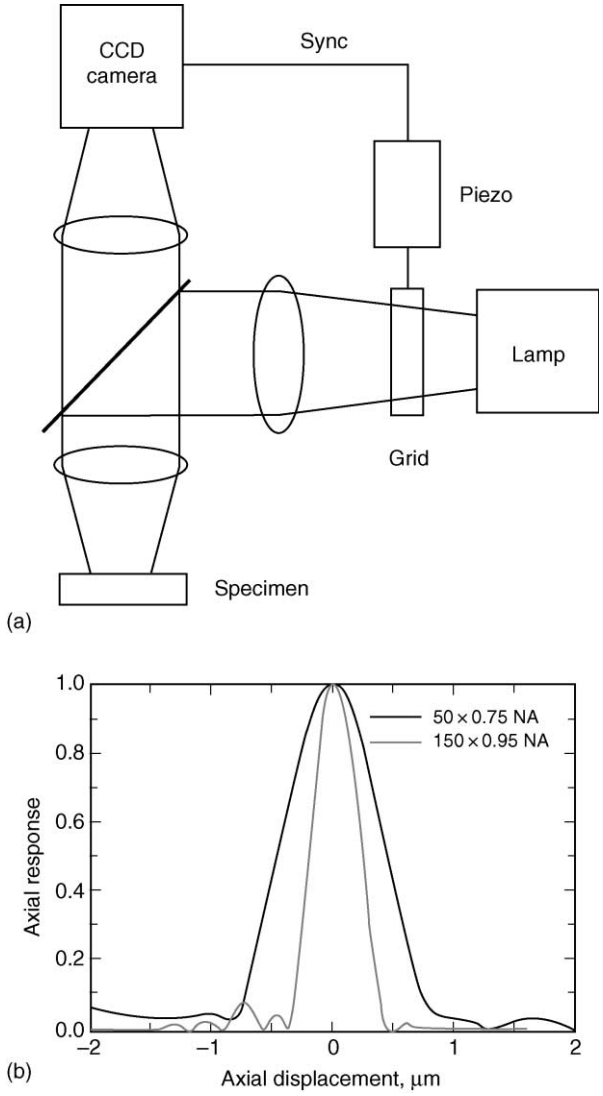


Figure 11 (a) Shows the optical system of the structured illumination microscope together with experimentally obtained axial responses in (b) which confirm the optical sectioning ability of the instrument.

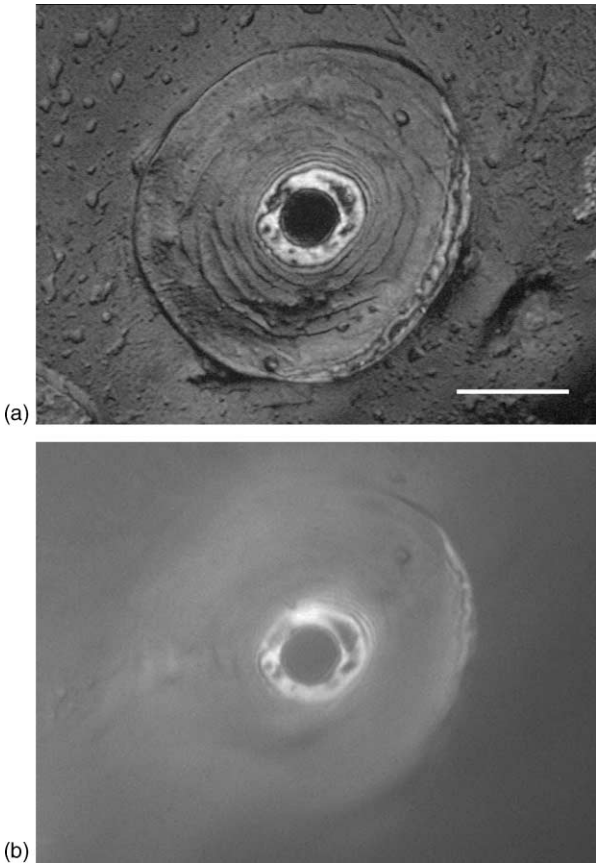


Figure 12 Two images of the region around the spiracle of a head louse. As in the previous figure, (a) is an autofocus image and (b) shows a mid-focus conventional image. The scale bar depicts a length of 10 μm.

of a spiracle of a head louse. The first is an auto-focus image of greatly extended depth of field constructed from a through-focus series of images. The second is a conventional image taken at mid-focus. The dramatic increase in depth of field is clear when compared with a mid-focus conventional image. We note that these images were taken using a standard microscope illuminator as light source. Indeed, the system is so light efficient that good-quality optical sections have been obtained or transistor specimens using simply a candle as light source!

Imaging using fluorescence light is also possible using this technique. However, an alternative approach that does not require a physical grid is possible if a laser is used as the light source. In this system the laser illumination is split into two beams which are allowed to interfere at an angle in the fluorescent specimen. This has the effect of directly 'writing' a one-dimensional fringe pattern in the specimen. Spatial shift of the fringe pattern is achieved by varying the phase of one of the interfering beams. As before, three images are taken from which both the optically sectioned image and conventional image may be obtained. The beauty of this approach is that no imaging optics are required at the illuminating wavelength and system.

Summary

We have discussed the origin of the optical sectioning property in the confocal microscope in order to introduce the range of imaging modes which this unique form of microscope leads to. A range of optical architectures have also been described. By far the most universal is that shown in Figure 1 where a confocal module is integrated around a conventional optical microscope. Other, more recent, real-time implementations have also been described. A number

of practical aspects of confocal microscopy have not been discussed. This is because they are readily available elsewhere, such as advice on the correct choice of detector pinhole size or because they are still the focus of active research, such as the development of new contrast mechanisms for achieving enhanced three-dimensional resolution such as the stimulated emission depletion method, STED or 4-Pi which are still under development but offer great promise.

See also

Microscopy: Imaging Multiple Photon Fluorescence Microscopy; Overview.

Further Reading

Corle TR and Kino GK (1996) *Confocal Scanning Optical Microscopy and Related Imaging Systems*. New York: Academic Press.

Denk W, Strikler JH and Webb WW (1990) Two photon laser scanning fluorescence microscopy. *Science* 248: 73.

Diaspro A (2002) *Confocal and Two-Photon Microscopy*. New York: Wiley-Liss.

Gu M (1996) *Principles of Three-Dimensional Imaging in Confocal Microscopes*. Singapore: World Scientific.

Hell S and Stelzer EHK (1992) Fundamental improvements of resolution with a 4Pi-confocal fluorescence microscope using two-photon excitation. *Optics Communication* 93: 277.

Hell S and Wichmann J (1994) Breaking the diffraction limit by stimulated emission. *Optics Letters* 19: 870.

Lukosz W (1966) Optical systems with resolving power exceeding the classical limit. *Journal of the Optical Society of America* 56: 1463.

Masters BR (1996) *Confocal Microscopy*. Bellingham: SPIE.

Minsky M (1961). Microscopy Apparatus. United States Patent 3,013,467, Dec 19, 1961 (Filed Nov 7, 1957).

Neil MAA, Juskaitis R and Wilson T (1997) Method of obtaining optical sectioning by using structured illumination in a conventional microscope. *Optics Letters* 22: 1905.

Pawley JB (ed.) (1995) *Handbook of Biological Confocal Microscopy*. New York: Plenum Press.

Wilson T (ed.) (1990) *Confocal Microscopy*. London: Academic Press.

Wilson T and Sheppard CJR (1984) *Theory and Practice of Scanning Optical Microscopy*. London: Academic Press.

Imaging Multiple Photon Fluorescence Microscopy

M Previte, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The field of light microscopy started with Ernst Abbe's introduction of the diffraction-limited objective lens in the late nineteenth century. Over the last 60 years, new imaging modalities, such as video, fluorescence, confocal, near field, and more recently multi-photon microscopy, have enabled the experimentalist to observe a plethora of new phenomena in the biological and physical sciences. Of these techniques, multi-photon microscopy has emerged not only as a valuable imaging technique but it also forms the technology base for emerging fields of optical memory and microfabrication.

The history of multi-photon microscopy (MPM) began with Göppert-Mayer's theoretical prediction of two-photon excitation in 1931. Although the theoretical groundwork was established, it took another 30 years for this process to be experimentally realized when Kaiser and Garrett observed two-photon absorption fluorescence emission from CaF_2:Eu^{+2}. Soon after this observation was made, Singh and Bradley witnessed higher-order multi-photon processes in 1964.

Historically, three-dimensional imaging was first achieved using confocal microscopy. In contrast to MPM, photo-interaction and fluorescence emission occurs throughout the excitation light path in confocal microscopy. Confocal microscopy selects the signal from the focal plane by the use of a conjugated pinhole aperture. In order to construct 3D images, successive planes must be imaged. As a result, the entire sample volume is exposed to the excitation light multiple times, and inevitably, photobleaching and photodamage occurs throughout the imaging volume. Furthermore, the pinhole aperture in the emission path causes additional signal reduction in optically thick specimens where the emission photon is scattered. Consequently, the utility of confocal microscopy is limited for imaging living biological specimens. The introduction of multi-photon microscopy, by Denk and Webb in 1990, addressed many of these limitations. By combining the methodology of multi-photon excitation with scanning microscopy, MPM offered advantages such as 3-D localized, femtoliter excitation volumes,

reduced scattering and absorption of the infrared excitation light by the specimen, and a large separation between the excitation and emission wavelengths. MPM can produce excellent three-dimensional images with increased signal to noise ratios.

Theoretical Basis of Multi-Photon Microscopy

Multi-photon Excitation Selection Rules

Unlike direct ground state to excited state one-photon transitions, multi-photon transitions depend on the presence of virtual electronic intermediate states to reach an excited state (see **Figure 1**).

Using perturbative expansion, the probability of excitation for one-photon transitions is described by the first-order solution, and multi-photon transitions can be determined by including the higher order solutions. The interaction of an electric field vector E_λ with a molecule's dipole, $\vec{r}$, is described by the respective one-photon and two-photon transition probabilities:

$$P_{1h\nu} \approx \langle \Psi_f | \vec{E}_\lambda \cdot \vec{r} | \Psi_i \rangle^2$$

$$P_{2h\nu} \approx \left| \sum_m \frac{\langle \Psi_f | \vec{E}_\lambda \cdot \vec{r} | \Psi_m \rangle \langle \Psi_m | \vec{E}_\lambda \cdot \vec{r} | \Psi_i \rangle}{\varepsilon_\lambda - \varepsilon_{m,i}} \right|^2 \quad [1]$$

where Ψ_i, Ψ_m, and Ψ_f are the initial, intermediate, and final transition states, $\vec{E}_\lambda$ is the electric field associated with the excitation light, ε_λ is the energy difference between the ground state and excited states, $\varepsilon_{m,i}$ is the energy difference between the intermediate state m and the ground state, i, and the summation is over all possible intermediate states m. Although electronic transitions, that result from the interaction of a strong electric field with a molecular dipole, are affected by vibrational mode overlap (Franck Condon factors) and anharmonic perturbative effects, the presence of intermediate transition states creates a set of general selection rules for multi-photon absorption processes.

Molecular states have eigenfunctions with a defined parity: 'gerade' describing even or symmetric states (no change in the sign of the wavefunction when the sign changes in the spatial coordinates) and 'ungerade' describing odd or antisymmetric states (the sign of the wavefunction changes when the sign changes in the spatial coordinates). Selection rules arise from the physics inherent in these even and odd eigenfunctions and the odd dipole operator that couples transitions between ground and excited states. In one-photon excitation, transitions are allowed for ground state and excited state transitions of opposite parity. In contrast, two-photon excitation occurs between ground and excited states with the same parity, by coupling through intermediate states with opposite parity. In general, ground and excited states with the same parity have a high excitation probability for an even number of photons; ground and excited states with different parity have a high transition probability for excitations involving an odd number of photons.

In the absence of major vibrational perturbations, inductive effects, or solvent relaxation, the probability of the transition depends on the magnitude of the integral overlap between the ground, intermediate, and excited states. As a result, molecule absorption cross-sections for one-photon, two-photon, three-photon, or multi-photon excitation processes can be estimated. Choosing fluorophores with higher multi-photon cross-sections allows more efficient excitation with lower excitation intensity. Finally, it should be noted that while the absorption cross-sections for one-, two-, and multi-photon excitation of a given fluorophores are different, the molecules rapidly relax to the same vibrational level in the excited electronic state. The excited state residence time (fluorescence lifetime), and the fluorescence decay processes depend only on the molecular structure and its microenvironment. Therefore, the fluorescence

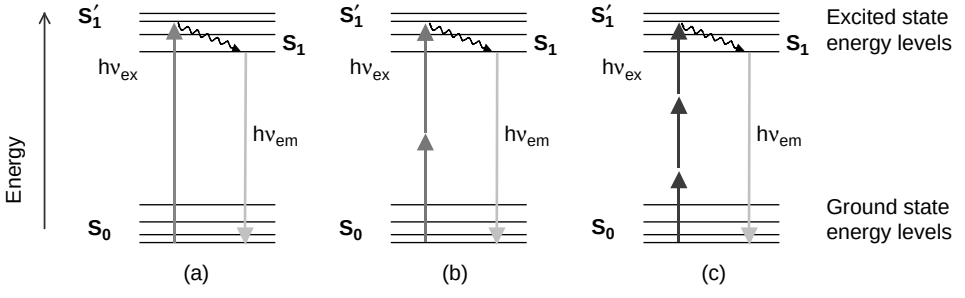


Figure 1 Jablonski diagram of a fluorescent dye molecule in one-, two- and *n*-photon excitation mode.

quantum yield and emission spectra is independent of the initial excitation process.

Enhanced Microscope Resolution Based on Multi-photon Excitation

The number of photons that are absorbed in a multi-photon process is proportional to the fluorophore concentration, $C(\vec{r},t)$, the molecule's absorption cross-section, σ_n, which depends on n, the number of photons involved in the absorption process, and the intensity of the excitation source, such that

$$N_{\text{abs}}(t) = \int \mathrm{d}\,\vec{r} I_0^n(\vec{r},t) C(\vec{r},t)\sigma_n \qquad [2]$$

When estimating $N_{\text{abs}}(t)$, the spatial and temporal intensity distribution of the excitation intensity are considered independently:

$$I_0(\vec{r},t) = S(\vec{r})I_0(t) \qquad [3]$$

such that for an n-photon process, $I_0(t)$ is the temporal dependence of the excitation intensity at the maxima of the spatial distribution and $S(\vec{r})$ is the unitless spatial distribution of the excitation light. Combining eqns [2] and [3] gives

$$N_{\text{abs}}(t) = \int \mathrm{d}\vec{r} S^n(\vec{r}) I_0^n(t) C(\vec{r},t)\sigma_n \qquad [4]$$

If it is assumed that fluorophore distribution is uniform and the system does not undergo photobleaching or ground-state depletion, $C(\vec{r},t)$ can be considered a constant and the previous equation can be rewritten as

$$N_{\text{abs}}(t) = C\sigma_n I_0^n(t) \int_V S^n(\vec{r})\mathrm{d}\vec{r} \qquad [5]$$

By expanding the incident laser beam and overfilling the back aperture of the objective lens, the paraxial form of the normalized intensity point spread function (PSF, $h^2[u,v]$) determines the properties for a diffraction-limited lens. The dimensionless radial distance, v, from the optic axis and the distance from the in-focus plane, u, are given by

$$v = \frac{2\pi(\text{NA})r}{\lambda} \qquad [6]$$

$$u = \frac{2\pi(\text{NA})^2 z}{n_0\lambda} \qquad [7]$$

where $\text{NA} = n_0 \sin\theta$ (θ being the half-angle of collection for the lens and n_0 is the index of refraction the medium) is the numerical aperture. Expressing the spatial distribution function with the optical units, v, u, and θ gives the following equation:

$$S(\vec{r}) = h^2[u,v] = 2\left|\int J_0(v\theta)\mathrm{e}^{-(1/2)\mathrm{i}u\theta^2}\theta\,\mathrm{d}\theta\right|^2 \qquad [8]$$

where J_0 is the zero-order Bessel function. If an oil immersion objective with a $\text{NA} < 1.0$ is used, the paraxial approximation accurately estimates the spatial distribution, and the integral of $S(\vec{r})$ evaluates as

$$\int_{V=\infty} S^n(\vec{r})\mathrm{d}\vec{r} = a_n\left[\frac{n_0\lambda^3}{8\pi^3(\text{NA})^4}\right] \qquad [9]$$

where $a_n = \int_0^\infty 2\pi v\,\mathrm{d}v \int_{-\infty}^{\infty} \mathrm{d}u[h(u,v)]^{2n}$. Since this integral cannot be solved analytically, the values of a_n are determined using numerical integrations of $h^2(u,v)$ over all space. As a result, the volume integral of the spatial distribution is always proportional to NA^{-4} and therefore independent of the number of photons involved in the excitation process. For higher NA objectives, the paraxial form of the normalized intensity PSF may no longer be an adequate approximation. The expression for temporal part of eqn [5] can be written as

$$I_0(t) = \left(\frac{\pi\text{NA}^2}{\lambda^2}\right)P(t) \qquad [10]$$

where $P(t)$ is the excitation power and λ is the center wavelength of the pulsed excitation source. The temporal dependence can be described with the time average expression:

$$\langle I_0^n(t)\rangle = \left(\frac{\pi\text{NA}^2}{\lambda^2}\right)^n \frac{1}{T}\int_0^T P^n(t)\mathrm{d}t \qquad [11]$$

where T is an arbitrary time interval for continuous wave lasers and $T = 1/f_{\text{p}}$ for a pulsed laser with a repetition rate, f_{p}. Since multi-photon microscopy uses pulsed laser systems, the integral in eqn [11] must be evaluated accordingly. For a pulsed laser with a pulse width, τ, repetition rate, f_{p}, and averaged power, P_0, the pulse profile can be approximated as

$$P(t) = \frac{P_0}{f_{\text{p}}\tau} \qquad \textit{for } 0 < t < \tau \qquad [12]$$

$$P(t) = 0 \qquad \textit{for} \qquad \tau < t < \frac{1}{f_{\text{p}}} \qquad [13]$$

such that eqn [11] can be expressed as

$$\langle I_0^n(t)\rangle = \left(\frac{\pi\text{NA}^2}{\lambda^2}\right)^n \left(\frac{P_0}{f_{\text{p}}\tau}\right)^n \frac{1}{T}\int_0^T \mathrm{d}t \qquad [14]$$

and evaluated as

$$\langle I_0^n(t)\rangle = \left(\frac{\pi \mathrm{NA}^2}{\lambda^2}\right)^n (P_0)^n \left(\frac{1}{f_\mathrm{p}\tau}\right)^{n-1} \qquad [15]$$

Assuming no stimulated emission and no self-quenching, the quantity of emitted photons collected per unit time, $F(t)$, is

$$F(t) = \frac{1}{n}\phi\eta N_\mathrm{abs}(t) \qquad [16]$$

where η and ϕ are the fluorescence quantum efficiency of the dye and collection efficiency of the system, respectively. Since n photons are needed for each individual excitation event, $1/n$ is a normalization factor for N_abs which describes the total number of photons absorbed. Rewriting this expression in terms of the average number of absorbed photons, $\langle N_\mathrm{abs}(t)\rangle$, gives the average fluorescence intensity:

$$\langle F(t)\rangle = \frac{1}{n}\phi\eta\langle N_\mathrm{abs}(t)\rangle \qquad [17]$$

where $\langle N_\mathrm{abs}(t)\rangle$ is

$$\langle N_\mathrm{abs}(t)\rangle = C\sigma_n\langle I_0^n(t)\rangle \int_V S^n(\vec{r})\mathrm{d}\vec{r} \qquad [18]$$

such that the measured fluorescence intensity or time averaged photon flux is expressed as

$$\langle F(t)\rangle = \frac{1}{n}\phi\eta C\sigma_n\langle I_0^n(t)\rangle \int_V S^n(\vec{r})\mathrm{d}\vec{r} \qquad [19]$$

Since the time averaged fluorescence is proportional to $\langle I_0^n(t)\rangle$, eqns [9] and [15] can be substituted into eqn [19] to give a practical expression that relates average fluorescence to measurable laboratory quantities, such as P_0, τ_p, f_p, C, σ_n, n, η, and ϕ, NA, and associated numerical constants such that

$$\langle F(t)\rangle = \frac{1}{n}\phi\eta C\sigma_n\left(\frac{\pi \mathrm{NA}^2}{\lambda^2}\right)^n (P_0)^n \left(\frac{1}{f_\mathrm{p}\tau}\right)^{n-1} \times a_n\left[\frac{n_0\lambda^3}{8\pi^3(\mathrm{NA})^4}\right] \qquad [20]$$

As eqn [19] shows, the fluorescence intensity has an nth order dependence on the spatial and temporal distribution of the excitation power. Therefore, appreciable multi-photon probability occurs only at a region of high temporal and spatial concentration of photons. Confocal pinholes can be eliminated making this technique better suited for 3D biological imaging.

The higher order (>2) dependence of the fluorescence signal upon the excitation intensity distribution results in axial confinement. Experimentally, it can be shown that over 80% of fluorescence signal emanates from a 1 μm thick region about the focal point for objectives with a numerical aperture of 1.25 under two-photon excitation. This confinement is significantly different from one-photon excitation where a spatially uniform fluorescent sample contributes equal fluorescence intensities from each z-section above and below the focal plane (assuming negligible excitation attenuation). For an excitation wavelength of 960 nm, the typical point spread function of two-photon absorption has FWHM of 0.3 μm in the radial direction and 0.9 μm in the axial direction. In comparison with confocal microscopy, the localization of the volume of excitation instead of the volume of observation, as in confocal microscope, allows 3D imaging with minimal photobleaching and damaging.

Basic Imaging Applications of Multi-Photon Microscopy

Multi-photon microscopy finds an increasingly broad range of applications, especially in biological and medical areas. For example, multi-photon microscopy has been utilized to study the physiological and biological states in various highly scattering tissues such as cornea, skin, and pancreas islet. Also, multi-photon microscopy is used in embryology to study the development of sea urchin, *Caenorhabditis elegans*, hamster, and zebra fish. Neurobiological applications include qualifying calcium dynamics dendritic spines, memory and structural plasticity, hemodynamics, and plaque evolution in Alzheimer's disease. MPM has been applied to study pharmacological distribution in tissues and in the quantification of transdermal drug delivery. Due to minimally invasive capability of multi-photon microscopy, intra-vital microscopy imaging into thick tissues of living animals is possible leading to applications in the study of cancer biology.

Although the advantages of MPM are less compelling in studying cell cultures where the specimen is optically thin, a number of uniquely exciting uses have been identified. For instance, the significantly broader two-photon absorption spectra of many fluorophores allow convenient imaging of specimens labeled with multiple color fluorophores. The localized excitation volume reduces photo-interaction with fluorophores outside the excitation volume allowing better 3D images to be obtained from cellular systems labeled with easily photobleachable fluorophores. The use of three and higher photon excitation has also allowed the excitation and imaging of important biomolecules with deep ultraviolet

absorption. Another interesting application of multi-photon excitation in cellular systems is based on the ability of this method to produce localized chemical reactions. Through the photoactivation or photolysis of 'caged' molecules, their chemical structures are altered to produce biologically active molecules, fluorescent species, or other observable chemical changes. Multi-photon excitation has been used to locally uncage neurotransmitters for mapping ion channels in neurons and to uncage calcium in studying signal transduction events. In cellular systems, MPM is being used to image normally nonfluorescent biomolecules, such as serotonin in neurons, based on multi-photon induced chemical conversion and multi-photon excitation. Finally, multi-photon microscopy has also found applications in studying the conformation, aggregation, diffusion, and transport of single molecules and proteins. Two-photon fluorescence correlation spectroscopy and photon counting histograms are powerful techniques that have been applied to study protein aggregation, conformation changes, and diffusion *in vitro* with applications ranging from cell biology to high throughput drug discovery. The large separation of excitation and emission spectra of two-photon excitation also allows high sensitivity detection of fluorescence from single chromophore; however, the higher photobleaching rate at the excitation volume for some chromophores may lessen or even reverse this advantage.

Advanced Multi-Photon Microscopy Techniques

As described in the previous sections, excitation processes involving multiple numbers of photons offers considerable advantages for the imaging process in fluorescent microscopy. 4Pi Confocal Microscopy and high-speed Multifocal Multi-photon Microscopy MMM are advanced multi-photon microscopy techniques that capitalize on these advantages to improve image resolution and acquisition times in fluorescence imaging.

4Pi Confocal Microscopy

In contrast to near-field microscopy, far-field light microscopes, i.e. confocal and multi-photon microscopes, can create three-dimensional images of the interior of cells. The lateral and axial resolution of these systems is limited between 180 nm and 300 nm and 500–1000 nm, respectively. This limitation is dictated by the Abbe theory where resolution is inversely proportional to the numerical aperture of the focusing lens. When focusing with a single lens, geometric constraints limit the numerical aperture to about 1. 4Pi microscopy was developed to overcome this limitation. By utilizing two opposing objective lenses whose spherical wavefronts are coherently combined, a greater total effective numerical aperture is achieved. This results in a higher axial resolution, while the lateral resolution of the 4Pi microscope remains unchanged with respect to its confocal counterpart.

In the setup of a 4Pi microscope, the illuminating wavefronts emanate from a common point light source. The path is split by a beamsplitter, guided through two optically identical paths, and through two opposing high NA lenses that share a common focus. The fluorescence light is then collected through one lens and a confocal pinhole. In the axial direction, this arrangement yields a PSF with one main maximum lobe that is 3–4 times sharper along the optic axis than the PSF of a conventional confocal microscope, and two undesired side-lobes.

These side lobes can produce artifacts in the image and their effect needs to be suppressed for unambiguous 3D imaging. Lobe suppression can be achieved in a 4Pi system using two-photon excitation resulting in a quadratic decrease in the ratio of the side-lobe and the main lobe amplitudes. Using an excitation wavelength of 800 nm and 1.4 NA oil-immersion objectives, a FWHM along the axial direction of the main lobe is about 145 nm and side lobes are approximately 25–35% of the maximum lobe intensity. These side lobes can be further mathematically reduced by means of a linear filter after (or during) image acquisition. The filter can be applied at lobe heights below 50%, and an effective 3D PSF consisting of a single spot with a 250 nm lateral and 145 nm axial FWHM is achieved. When combined with nonlinear image restoration, 4Pi confocal microscopy achieves a 3D resolution in the 100 nm range. This resolution is similar to that obtained with near-field optical microscopy in biological applications with an added advantage of 3D imaging of the cell interior.

Since resolution improvement in 4Pi microscopy can be negated by the presence of aberrations, the usage of oil immersion lenses requires the specimen to be fixed and mounted in glycerol to minimize aberrations resulted from index mismatch. The sample thickness is further limited in the range of 20–30 μm. Due to significant index mismatches between oil and water, the imaging of live cells in aqueous medium is extremely difficult, if not impossible. In order to circumvent this difficulty, water immersion objective lenses are introduced into the 4Pi system. These lenses have a lower aperture angle $\alpha = 62°$ (NA = 1.2) than oil immersion lenses

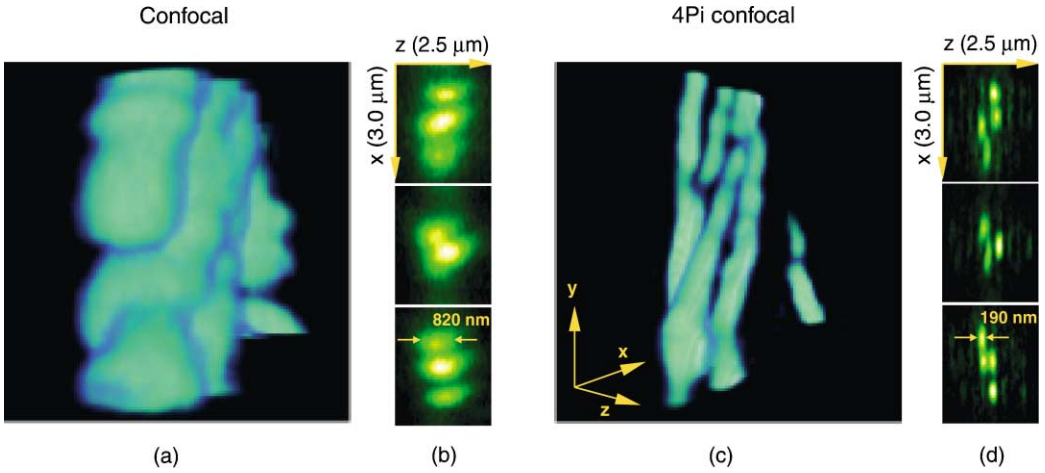


Figure 2 (a) Two-photon confocal and (c) 4Pi confocal point deconvolved images of an immunolabelled microtubule network in fixed NIH3T3 cells. The 3D rendered images in (a) and (c) were generated out of a large XZ-image-stack, partly shown in (b) and (d). As indicated in the XZ-images, the axial resolution is 820 nm and 190 nm for the two-photon confocal and the two-photon 4Pi confocal images, respectively. The images were taken with water immersion lenses (NA = 1.2). Reproduced from Bahlmann K, Jakobs S, *et al.* (2001) 4Pi-confocal microscopy of live cells. *Ultramicroscopy* 87(3): 155–164, with permission from Elsevier.

$\alpha = 68°$ (NA = 1.4) and therefore are expected to produce higher side lobes, which are reduced by 35% with careful aberration studies such that linear filtering is possible. Nonetheless, the use of water-immersion lenses solves two major problems. First, they extend the possible sample thickness to the working distance of the objective lens (~200 μm); second they allow the imaging of live cells in aqueous medium. As a result, microtubule networks (see Figure 2) and live bacteria can be imaged with a resolution of 280 nm in lateral and 190 nm in axial direction. The water immersion 4Pi microscope, utilizing two-photon excitation, is capable of the highest 3D resolution so far reported in live cell imaging to date.

Multifocal Multi-Photon Microscopy (MMM)

Improving imaging speed is critical for applying 3D microscopy in biological specimens. An MMM scans an object with an array of high-aperture foci whereas a conventional multi-photon microscope scans with a single beam. An MMM combines nonlinear fluorescence excitation with parallelized 3D imaging. The technical design consists of a densely packed microlens array; each microlens transmitting a collimated beam at a slightly different angle, overfilling the back aperture of the objective. Consequently, from the arrangement of the microlenses, an array of foci is reproduced in the image plane. The foci are scanned rapidly across the image plane to ensure full coverage of the focal plane and real-time imaging (a stack of 50 images of 30–70 μm diameter can be acquired within a few seconds). The foci's movements are either generated by rotating a microlens array that is arranged in a Lasserie pattern or by rectilinear scanning of a square microlens matrix.

MMM development was initiated because nonlinear optical damaging processes limit the usable focal peak intensity to about 200 GW/cm^2, which can be reached by a tightly focused 7–15 mW femtosecond laser beam. This power consumption makes use of only a fraction of the laser output of a typical mode-locked titanium-sapphire laser light source. The MMM better exploits the 1–2 W of the beam power by dividing the power among 25–40 foci and scanning each in parallel. With this method, typical acquisition times are cut from ~1000 ms to 10–50 ms. MMM provides an excellent method for real time 3D multi-photon imaging.

A fundamental problem in parallelized 3D microscopy is that increasing density of the foci over a certain level decreases the axial resolution. The reduction of the distance between the foci increases not only the speed or image brightness, but it also increases the interference between neighboring focal fields; especially in planes shortly before and after the focal plane. For neighboring foci in high NA objective lenses, interference effects can be significant. The relatively short laser pulses that are commonly used in multi-photon excitation provide a means to reduce this interference by using time multiplexing. Experimentally, the femtoseconds pulses of the Ti:sapphire laser, going through each lenslet, can be scheduled to arrive at the focal plane at different times by a mask of glass in front of each microlens, varying the optical path lengths of these beams. The different arrival

time eliminates interference effects. The resolution of the parallelized system remains unchanged as compared with the conventional one foci system. In conclusion, the MMM enables increased image acquisition speed by a factor of the number of foci in the optical plane without compromising image resolution.

Advanced Multi-Photon Microscopy Applications

As described by the previous section, MPM technology is undergoing rapidly development with improved resolution and increased instrument throughput. Recently, MPM has also found applications in exciting new areas of optical memory and microfabrication.

Multi-Photon 3D Optical Memory

Researchers in the field of optical memory are challenged to develop instrumentation that records three-dimensional, high density, robust bitwise information in low-cost and readily available storage materials. Initially, early optical memory techniques relied on higher energy pulsed recording beams and red shifted reading beams. Aided by the development of the Ti:sapphire laser and high numerical aperture objectives that can confine laser light to a volume of the order, λ^3, the power of the focused femtosecond pulses at the focal point can reach over 10^{18} W/cm^2 with electrical field strengths reaching 2.7×10^{10} V/cm. These fields are substantial enough to locally change the optical properties of selected materials. Thus, multi-photon processes were intertwined with the field of optical memory. Recent experiments showed that with properly chosen materials and instrumentation, high density, three-dimensional data storage was possible.

Optical memory is based on photochromic material whose optical properties are changed upon irradiation such that initial and final chemical states have different absorption spectra or indices of refraction. In 1989, Parthenopoulos and Rentzepis first showed that 3D information could be stored in a polymer film doped with photochromic molecules. Although the recorded information is lost when the photo-excited compound spontaneously reverts back to the original nonfluorescent compound, Dvornikov and Rentzepis later improved the material such that the stored information was stably recorded at room temperature.

For photochromic materials, the recording and reading mechanisms often required multiple laser sources and complex optical setups. Kawata *et al.* addressed these issues with the development of an alternative optical method that used photorefractive crystals as the storage media, a Ti:sapphire as the recording device, and a phase-contrast microscope configuration for data readout of bits separated by 20 μm. Using this method, Kawata avoided the complication of different recording and read-out laser sources, thereby simplifying the instrumentation. However, in photorefractive material, poor refractive index matching, between the storage medium and the objective, results in aberration and degrades storage densities. Recent advances in objective lens design have attempted to address these limitations. Kawata and Kawata describe many of these optical memory devices in a recent review. A recent publication by Olson *et al.* describes a method for multi-photon data storage in molecular glasses and highly cross-linked polymers. These materials prove to be highly efficient and robust for long-term data storage.

Although the field of optical memory has made great progress since 1989, further advances are required before this field becomes a widely adopted technology. For example, simpler recording and reading techniques are required to insure a cost-effective and reliable data storage technology. Consequently, further progress in the production of compact ultrafast lasers and high-speed scanning instrumentation will be needed before broader acceptance of optical memory storage devices.

Multi-Photon 3D Microfabrication

Another promising emerging application of multi-photon excitation is 3D microfabrication. 3D microfabrication processes can be performed with two-photon absorption. Maruo *et al.* and Sun *et al.* provide examples of microfabricated structures generated with femtosecond pulses to induce photopolymerization. Much like the optical memory techniques, this methodology relies on photo-initiators with large two-photon cross-sections to initiate polymerization of surrounding monomer or polymer molecules. Kawata *et al.* and Sun *et al.* showed that two-photon photopolymerization could be used successfully to create intricate three-dimensional microstructures. These systems offer lower polymerization thresholds, higher polymerization rates, and wider dynamic write ranges such that rapid fabrication and low-power writing are possible. Recently, Cumpston *et al.* described a class of π-conjugated compounds with large two-photon cross-sections, which can be effectively used for two-photon polymerization initiators in three-dimensional optical data storage and microfabrication processes.

As with most of these processes, the techniques have proven to be successful, but the mechanisms behind the multi-photon initiated chemical processes are not completely understood. With further work to optimize the reaction parameters, it is possible to design experiments that will allow new materials to be synthesized or patterned based on processes such as dielectric breakdown, microexplosions, optical damage, and pyrolytic effects.

Conclusion

Relying on femtoliter excitation volumes and large separations between excitation and emission wavelengths, multi-photon microscopy has emerged as a powerful technology that allows for increased resolution and enhanced signal-to-noise ratios. It has also enabled scientists to avoid complications in imaging biological systems due to decreased sample photodamage and in some instances, photobleaching of fluorophores. Advancements in multi-photon imaging applications, such as 4Pi Confocal microscopy and multifoci multi-photon microscopy, have further enhanced the resolution of imaging technology and improved the speed of data acquistion respectively. Furthermore, the application of multi-photon technology in the fields of microfabrication and optical memory promises to address some of the difficulties involved in designing three-dimensional high-density storage devices and microfabricated components. Overall, multi-photon microscopy provides the scientific community with an invaluable tool in the study and design of biological and physical systems.

See also

Microscopy: Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy; Overview.

Further Reading

Bahlmann K, Jakobs S, *et al.* (2001) 4Pi-confocal microscopy of live cells. *Ultramicroscopy* 87(3): 155–164.

Bewersdorf J, Pick R, *et al.* (1998) Multifocal multiphoton microscopy. *Optical Letters* 23: 655.

Denk W, Strickler JH, *et al.* (1990) Two-photon laser scanning fluorescence microscopy. *Science* 248(4951): 73–76.

Kawata S and Kawata Y (2000) Three-dimensional optical data storage using photochromic materials. *Chem. Rev.* 100(5): 1777–1788.

Olson CE, Previte MJR, *et al.* (2002) Efficient and robust multiphoton data storage in molecular glasses and highly crosslinked polymers. *Nature Materials* 1(4): 225–228.

Pawley JB (ed.) (1995) *Handbook of Biological Confocal Microscopy.* New York: Plenum Press.

So PT, Dong CY, *et al.* (2000) Two-photon excitation fluorescence microscopy. *Annu. Rev. Biomed. Eng.* 2: 399–429.

Xu C and Webb WW (1997) Multiphoton excitation of molecular fluorophores and nonlinear laser microscopy. In: Lakowicz JR (ed.) *Topics in Fluorescence Spectroscopy.* New York: Plenum Press.

Xu C, Zipfel W, *et al.* (1996) Multiphoton fluorescence excitation: new spectral windows for biological nonlinear microscopy. *Proceedings of the National Academy of Science of the USA* 93(20): 10763–10768.

Interference Microscopy

E Novak, Veeco Instruments, Inc., Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Interference microscopy is a powerful noncontact metrology technique for reflection and transmission characterization of samples. The field can be divided roughly into two main categories, polarization-splitting interferometers and amplitude-splitting interferometers. In both cases, the light incident on a test sample is first divided into two beams and later recombined to form an interference pattern. This interference pattern is analyzed by one of a variety of methods to determine the difference in the optical paths the two beams traversed. This optical path difference may be used to quantify surface shape, refractive index variations, or thickness variations of the sample to a very high accuracy. This article describes the background of the two categories of interference microscopy, specifics of their implementations for different types of measurements, current state of the technologies, and examples of different measurements performed using these techniques.

Background

Interference fringes were first observed in a microscope by Robert Hooke in 1665. In the late 1800s, several practical versions of interference microscopes

entered use to examine transparent specimens, though none gained widespread commercial success. Then in the 1950s, differential interference contrast (DIC) microscopy was invented to increase image contrast when examining semi-transparent samples under a microscope, particularly biological specimens. As this technique does not require staining of samples, and avoids the halo effect common in phase contrast microscopy (*see* **Microscopy:** Phase Contrast Microscopy), such systems became widely used in a variety of biological applications. In the early 1960s, Krug, Rienitz and Shulz published a comprehensive book to summarize all of the various interference microscope designs that had been produced thus far, showing the large depth and breath of research to that point.

With the advent of affordable computers and CCD cameras in the 1970s and 1980s, various commercial systems were developed using a variety of interferometric configurations to quantify the measured interference patterns from small samples. Due to their noncontact nature and ability to provide area measurements, as opposed to the two-dimensional traces produced by contact stylus techniques, such systems gained rapid acceptance in a variety of industries. Performance has steadily improved such that vertical resolution is sub-nm and accuracy on the nm-level is readily achievable. Lateral resolution is limited only by the optics of the microscope, with high numerical aperture objectives achieving resolution below 0.5 μm. Today, commercial systems are used in such measurements as index variations of biological samples, roughness of paper, microfluid channel volume, MEMS properties, hard drive head dimensions, drug coating thickness, and machine part characterization.

Polarization-Division Interference Microscopes

While the concepts behind polarization division interference microscopes have been known for well over a century, they did not gain widespread use until after Georges Nomarski invented the DIC microscope, also referred to as a polarization interference contrast microscope, in 1955. A schematic of a transmission-based and reflection-based Nomarski setup is shown in Figure 1, where the displacement of the beams has been greatly exaggerated for clarity.

The illumination in Nomarski systems is generally from an unpolarized broadband source, such as a

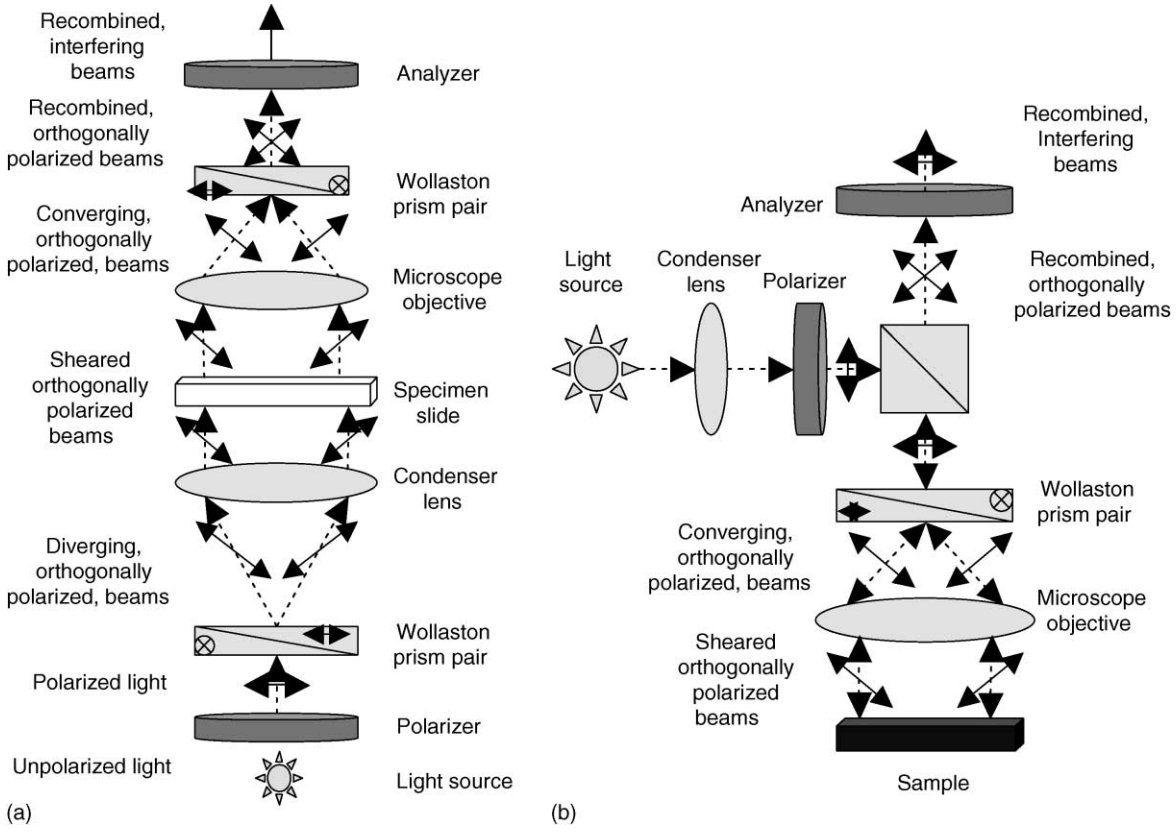


Figure 1 (a) Schematic of transmission-based Nomarski interferometer and (b) reflection-based Nomarski interferometer.

tungsten–halogen bulb. A polarizer in the illumination path is then used to set the polarization state of the light incident on a modified Wollaston prism pair. The modified Wollaston prism consists of two birefringent wedge prisms which divide the beam in two parts, each with orthogonal polarization states. The two beams travel in slightly different directions from one another. The condenser lens brings the light to illuminate the sample, with each of the two orthogonally polarized beams slightly sheared from one another. This shear is typically much less than the resolution limit of the microscope objective, typically from tens of nanometers to several hundred nanometers, such that only one image is seen by the observer. After passing through or reflecting from the sample and the imaging microscope objective, a Wollaston prism pair is used to recombine the two polarized beams. Finally, an analyzer is used to bring both beams to a common polarization state, allowing interference between the two beams.

As the system is very sensitive to the position and optical properties of the prisms, different Wollaston prism pairs are used for each objective, usually rotating into the path automatically on a turret as the objectives are changed. In a reflection system, this prism also recombines the displaced beams after they reflect from the test piece; a transmission system will usually have only one Wollaston prism pair after the sample that can handle all objectives, instead of requiring a different recombining prism for each objective.

The observed interference pattern is an image of the gradient in the sample in the direction of shear created by the Wollaston prism pairs. In order to fully characterize the surface, the sample or optics must be rotated to capture shear images in both orthogonal directions. Essentially the technique acts as a high-pass filter, emphasizing edges and lines on the test piece.

With a white-light source, dispersion effects will cause one side of a feature to appear brighter or in a specific color range whereas the other side appears darker or in different colors. This gradient effect confers a pseudo three-dimensional appearance to the specimen. The images are quite colorful, with different colors shearing differently; by adjusting the Nomarski prisms along the optical axis, one can vary the retardation between the two beams, producing more or less contrast or color variation in the observed images.

The Nomarski system has several appealing qualities for surface and index characterization over other interferometric and polarization-based techniques. First, it is a common-path interferometer; with the exception of the slight shear, both beams pass through exactly the same optics. This means that these interferometers are mostly insensitive to vibration, as both interfering beams will be affected in the same manner. In addition, the common path nature of the system reduces the quality requirements on the optics, improving measurement quality and reducing overall system cost. In addition, Nomarski techniques do not suffer the halos associated with phase contrast imaging techniques (*see* **Microscopy:** Phase Contrast Microscopy) and provide better out of focus rejection and higher light levels since they can use the entire aperture of the microscope objective.

For reflection-based systems, the observed fringes relate to the surface slope along the direction of shear; for transmission systems, index of refraction gradients or thickness variations are observed, producing much higher contrast images of biological samples than conventional bright-field techniques. The good out-of-focus rejection mentioned earlier makes possible more accurate optical sectioning of samples, where each vertical layer of the specimen is in focus while all others are out of focus. These features, combined with the fact that high contrast images can be obtained without chemical staining, have made these systems widely used in biology. In addition, the high sensitivity to slope has made DIC systems commonly used in semiconductor inspection applications and other precision surface characterization and defect inspection applications, where automatic image processing is combined with the highly sensitive measurements to quantify both surface and subsurface defects. For instance, Figure 2 shows a bright-field image of a laser diode facet as

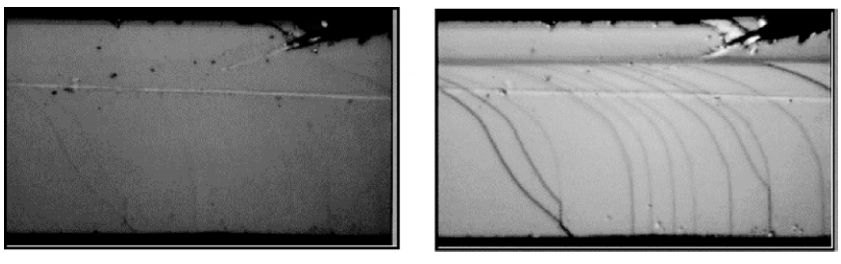


Figure 2 Bright field (left) and Nomarski (right) images of laser diode facet defects. Courtesy of Royce Instruments, Inc., Napa, CA.

well as a Nomarski image, with the facet defects greatly emphasized by the Nomarski optics.

Initial systems using differential interference contrast were strictly qualitative, with the user manually rotating the incoming polarization to produce the best contrast for a given sample and visually evaluating the resulting image. As mentioned above, visual inspection of biological specimens and defect review remain primary applications of these devices, though quantitative. By translating the Wollaston prisms laterally or rotating the analyzer in the system, one can phase shift the interference pattern. Storing multiple phase shifted images on a computer, one can obtain quantitative data about the shear using standard phase shifting interferometry equations. In addition to phase shifting through translation of the prism, quantitative data may be performed through analysis of the colors of the observed fringes or through geometric phase shifting techniques. More details on these methods can be found in the articles on **Interferometry:** Phase Measurement Interferometry; and **Interferometry:** White Light Interferometry.

Once quantitative information about the shear between the beams is obtained, the resulting data may be integrated to get back the original surface or index profile. Performing this operation on data with shear in two orthogonal directions allows the complete surface profile to be obtained. Unfortunately, since Nomarski setups are typically employed as add-on modules to standard microscope platforms rather than custom-designed metrology instrumentation, such quantitative phase shifting capability has not been widely commercialized; only a few custom or laboratory systems are in use when compared with white-light optical profilers described below. However, many research groups report nm-level surface characterization, and have developed a variety of techniques to accurately quantify thickness and index changes. The primary limitation of the quantitative techniques comes from the depth of focus of the microscope objective, which limits the vertical heights, thickness range, and index gradients that can be measured in a single measurement. Phase shifting at multiple vertical locations can be performed, but unless greatly increases the complexity of the measurement and reconstruction algorithms.

Amplitude-Division Interference Microscopes

White Light Optical Profilers

Probably the most common amplitude division interferometric microscopes are white-light interferometers, also often called optical profilers. These systems consist of three main elements: the base microscope, with illuminator and imaging optics; an interferometric element to divide and recombine the wavefront, and a scanning mechanism by which the optical path between the test and reference beams may be varied in a known manner. As the name implies, these systems typically employ spectrally broad sources, generally from a tungsten–halogen bulb. Unlike Nomarski interferometers, white light interferometers are used only in reflection-mode, as their noncommon-path nature and low-coherence source preclude interference from thick samples measured in transmission.

Figure 3 shows a schematic of a typical white light interferometer. Light from the illuminator arm, typically in a Kohler configuration, is coupled into the system with a first beamsplitter and passes through a second beamsplitter which is typically inside the microscope objective. This beamsplitter divides the beam into two parts, one of which travels to the sample and reflects from it, while the other reflects from a very high-quality reference surface. The reflected light is then recombined by the same beamsplitter and the resulting interfering light is imaged onto a camera. The static interferogram thus produced is usually by changing the optical path difference between the test and reference beams through translation of the objective with respect to the sample. The resulting signal is analyzed by one of a variety of methods as described below.

White light interferometers generally have two distinct modes of operation. In the first mode, generally referred to as phase shifting mode, a filter

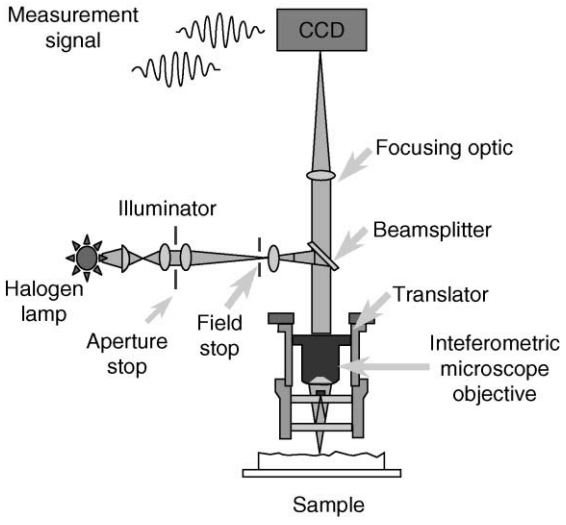


Figure 3 Schematic of white light optical profiler showing the illuminator, and translating interferometric objective.

is used to produce quasi-monochromatic light for the measurement process. This produces a large number of high-contrast interference fringes that obey the classical interference formula:

$$I_i(x, y) = I_t(x, y) + I_r(x, y) + 2\sqrt{I_t I_r}(\cos(\phi(x, y) + \alpha_i) \quad [1]$$

In the above $I_i(x, y)$ is the irradiance observed at any one point of the image, $I_t(x, y)$ and $I_r(x, y)$ are the irradiances of the test and object beams, $\phi(x, y)$ is the optical path difference between the two beams, and α_i is a phase offset introduced by the interferometric instrument. By varying α_i in a known manner and collecting at least three irradiance patterns, any of the three unknowns in the above equation can be determined; specifically, the optical path difference may be solved for every point. Assuming measurements are in a medium of known and constant refractive index, such that the only contributions to the OPD are height variations rather than index variations, surface height at each point is simply:

$$H(x, y) = \frac{\lambda}{4\pi}\phi(x, y) \quad [2]$$

Phase shifting techniques produce sub-nm repeatability and very high accuracy, but suffer from a 2π phase ambiguity from the calculation of phase from irradiance data. Thus, this technique cannot handle surfaces with point-to-point steps greater than one quarter of the wavelength of the light used. Calibration methods and further information may be found in the article on **Interferometry:** Phase Measurement Interferometry.

In order to extend interference microscopy to cover large steps, steep slopes, and vertical ranges greater than the depth-of-focus of the microscope objectives, a second technique of fringe analysis may be employed, that of coherence sensing. In this method an incoherent light source is used, such as a tungsten–halogen bulb. Due to the wide spectral range of the source, as interference patterns from different wavelengths add in the observation plane, the observed interference fringes will rapidly wash out for nonzero optical path difference between the two beams. See the article on **Interferometry:** White Light Interferometry, for further details and illustrations of this effect.

With the localization of fringes only at the point of zero optical path, the instrument scans the interference objective over the desired range and then calculates the position at which maximum fringe contrast occurs for each point. Calculating this position across all points within the field creates a height map of the surface; the interference fringes serve as a tool for depth discrimination in much the same way that the pinhole in a confocal microscope does (*see* **Microscopy:** Confocal Microscopy), another technique finding wide use in surface profiling. No phase unwrapping is required, with each pixel height calculated independently of all others, avoiding ambiguities associated with phase shifting methods. The vertical range is limited only by the scan range of the measurement and the working distance of the microscope objectives; current commercial instruments can achieve vertical ranges of over 8 mm. Figure 4 shows the fringes one would observe at a given scan location during measurement of a test piece with three steps. As the OPD between the test and reference arms varies, the fringes will

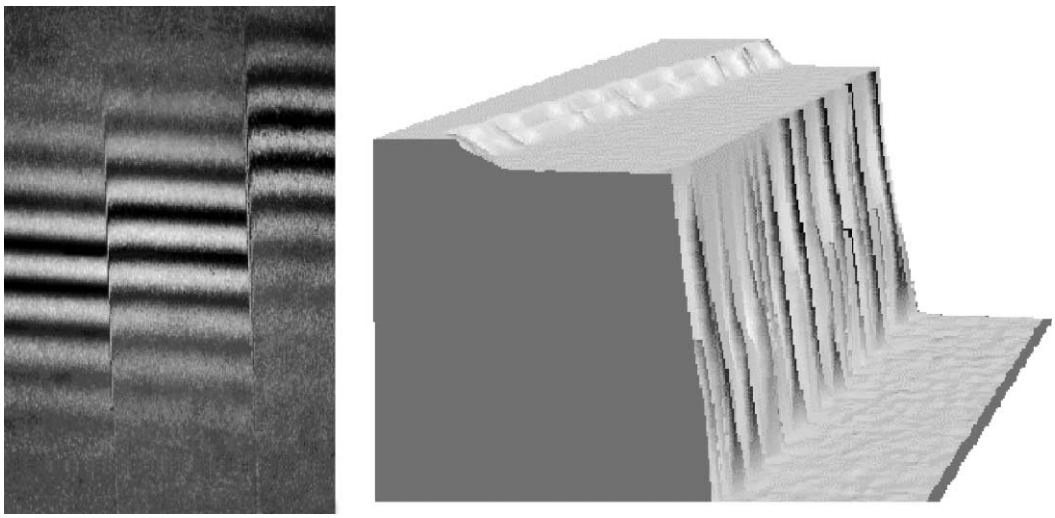


Figure 4 Localized fringes (left) of a 3-tiered surface structure (right). As the OPD is varied, the fringes will scan through the field, with the point of maximum contrast corresponding to the relative surface height of each pixel.

translate across the steps, and the surface plot at the right can be generated by calculating the scanner location of the peak contrast condition for each pixel.

There are a variety of methods for calculating the position of maximum fringe contrast using the coherence-sensing method. These include center-of-mass calculations of the peak fringe position, various frequency-space techniques and correlation of the measured fringe pattern with a theoretical pattern. Calibration of these systems generally involves repeated measurement of a known step height standard and varying the scanner motion until the step height is accurately calculated, though some systems employ separate distance-measuring interferometers to perform on-the-fly calibration for each scan. Generally speaking, coherence sensing techniques have repeatability in the range of a few nm, as opposed to the sub-nm repeatability of phase shifting techniques. This reduced sensitivity is the trade-off for avoiding the height and slope limitations of phase-shifting techniques. More details can again be found in the article on **Interferometry:** White Light Interferometry.

Recent advances in computing power have enabled an additional method of calculating surface height by combining coherence sensing with phase shifting. Coherence sensing methods are used to determine the point of best fringe contrast, and uses this information combined with the phase calculations to overcome ambiguities caused by large steps or steep slopes. Thus, one can obtain the sub-nm repeatability of phase-shifting methods without limitations on the vertical range or slope of the sample to be measured.

New fringe analysis techniques are extending the utility of these systems even further. Strobed sources are now available which effectively freeze the periodic motion for the instrument; this allows dynamic properties of samples to be measured, from surface deformation under motion to resonance frequency and switching time calculations. By analyzing the multiple interference patterns generated by transparent films deposited on a surface, film thickness, homogeneity, and differences between substrate and film surface properties may be determined.

Interferometric Microscope Objectives

One of the most critical elements in white light interferometers is the interferometric objective which splits and recombines the wavefront to form an interference pattern. Three primary interferometric configurations are used to divide and recombine the wavefront: Michelson, Mirau, and Linnik. Each of the three types of objectives is described below.

For low-magnification objectives, typically 5X or below, a Michelson interferometer configuration is typically used. A schematic of such an objective is shown in Figure 4. The illuminating beam passes through the objective and is divided by a beamsplitter placed just underneath the objective. A portion of the light travels to the test sample and the rest to a high-quality reference surface. The light reflected from each beam path is recombined by the beamsplitter and propagates to the observation plane. Note that it is important that the distance between the beamsplitter and reference surface match the distance between the beamsplitter and focus position of the sample. If the reference surface is defocused, fringe contrast will be reduced, aberrations increased, and overall system accuracy and repeatability will suffer.

This objective configuration has several advantages. First, one can add a Michelson housing to almost any standard objective with sufficient working distance for the beamsplitter, making these objectives easy to produce and relatively inexpensive. Second, the reference surface can be easily changed by the user to closely match the reflectivity of the test piece, ensuring maximum fringe contrast in the interfered beam (*see* **Interferometry:** Phase Measurement Interferometry). Third, provided there is distance between the beamsplitter and reference, one can insert a dispersion-compensating plate into the reference arm to allow measurement of parts beneath a cover glass or other dispersive medium; without such compensation, fringe contrast is generally insufficient for accurate measurements for transmissive media beyond a few hundred micrometers thickness.

As one moves towards higher magnifications, however, the working distance of the objective is typically insufficient to accommodate a large beamsplitter and still leave a reasonable distance between the objective and sample. In addition, a large cube beamsplitter will generally introduce unacceptably high aberrations in these higher numerical aperture objectives. For this reason, objectives from 10X magnification to 50X magnification are typically of the Mirau configuration, shown schematically in Figure 5.

Mirau objectives contain two additional plates compared to a standard bright-field objective: one is coated and acts as a beamsplitter, splitting the incident beam, while a second holds a small, high-quality reference surface and acts as a compensating plate so that optical paths in both arms of the interferometer are balanced (Figure 6). Such objectives are more expensive than Michelson designs, but maintain long working distances even at higher magnifications. Thermal stability becomes

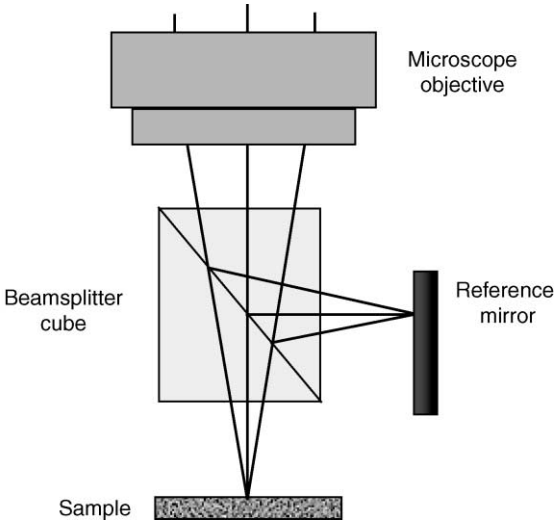


Figure 5 Schematic of a Michelson interference objective used for low-magnification applications.

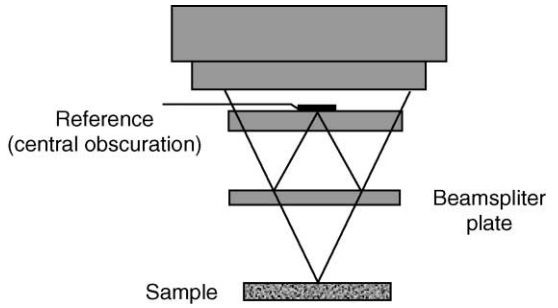


Figure 6 Schematic of a Mirau interference objective used for medium-magnification applications.

very important, as depth of focus is much reduced over lower-magnification objectives and the path from beamsplitter to reference still must be the same as from beamsplitter to in-focus test object, and most Mirau designs are therefore thermally compensated.

As one moves to even higher magnifications and numerical apertures, such as 75X or 100X, the microscope objectives do not have sufficient working distance to allow for any reasonable beamsplitter after the objective unless the numerical aperture is greatly reduced. To maintain high numerical aperture and high lateral resolution, a Linnik configuration must be used. A schematic of this configuration is shown in Figure 5. In this configuration, two separate microscope objectives are used after a beamsplitter, with one beam again reflecting from a high-quality reference and the other from the sample.

A Linnik objective is generally the most difficult to use and expensive interferometric configuration

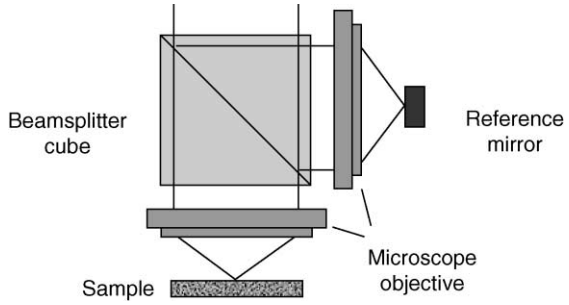


Figure 7 Schematic of a Linnik interference objective used for high-magnification applications.

(Figure 7). High numerical aperture typically limits working distance to a few hundred micrometers. In addition, depth of focus is very small, and thus these objectives must often actively adjust the reference mirror position to maintain focus over long periods of time and environmental conditions. Also, the dispersion and aberrations in the two objectives must be closely matched to ensure good fringe contrast, as there are many noncommon-path optical elements in this design; typically many objectives must be evaluated to find two that suitably match, another contributor to cost. However, these objectives do have the highest lateral resolution, and with an NA of up to 0.95, can measure the steepest slopes due to a large collection angle.

Applications

White-light interferometers are noncontact, and thus will not damage samples under test. They provide true three-dimensional surface characterization with nm-level accuracy and repeatability, and can be employed in production and laboratory environments. The various interferometric objectives allow for lateral resolutions and ranges to meet most needs; fields of view may even be stitched together to create surface maps whose extent is limited only by the range of the translation stage employed. In addition, each scan is typically quite fast, from under one second in phase-shifting mode to vertical scan rates of up to 100 μm/sec in coherence-sensing mode, with no sample preparation time. While the systems must be properly calibrated, vibration minimized, contrast maximized, and potential error sources such as dispersion and phase change on reflection effects must be managed, they remain one of the most accurate surface characterization tools available.

All of the above factors have enabled the use of white light interferometers in a variety of fields. The semiconductor and data storage industries

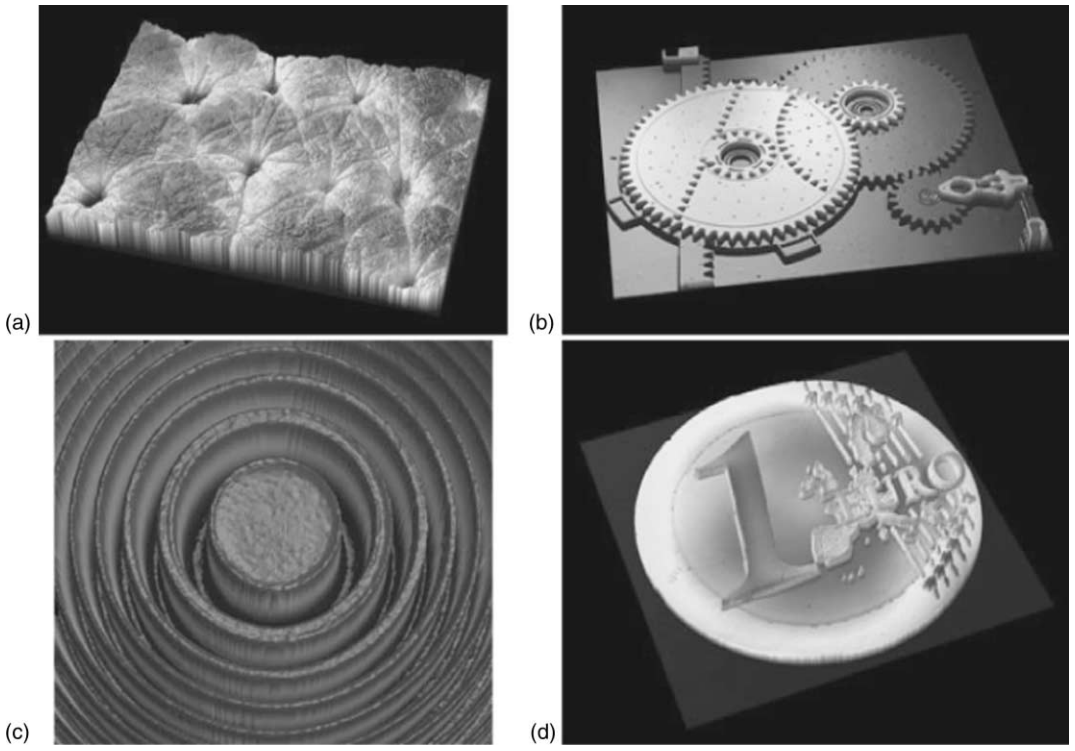


Figure 8 Surface plots from a white light optical profiler. (a) Human skin sample, (b) MEMS microgears, (c) binary optic, and (d) 1 Euro coin. Pictures courtesy of Veeco Instruments.

perform critical dimension and surface characterization of various components using these systems, some of which are configured for full wafer-handling capabilities. Systems are also used to measure precision machine parts, optics, and medical instrumentation and implants. The high sensitivity of the coherence sensing techniques allows surfaces with less than 0.1% reflectivity to be measured, so that characterization of paper, abrasives, and even different foods has been measured with these instruments. With the addition of the advanced technologies mentioned above, these systems are also now doing dynamic measurements of MEMS and thermal response characterization of precision surfaces, and also are measuring thicknesses of adhesives, drug coatings, photoresist, and a variety of other semi-transparent layers in various applications. Figure 8 shows several surface measurements, including human skin, MEMS microgears, a binary optic, and 1 Euro coin.

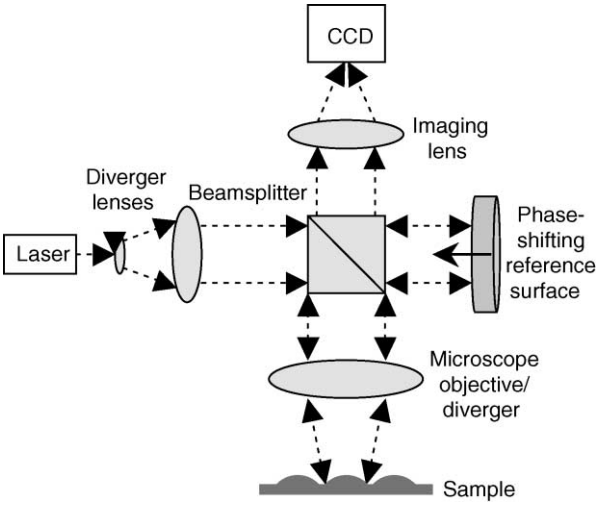


Figure 9 Schematic of a micro-LUPI interferometer operating in reflection mode.

Laser-based Interference Microscopes

Another, less common type of amplitude-division interference microscope has recently been commercialized, primarily for use in examining micro-optics. These laser-based systems employ a Tywman–Green interferometric configuration combined with a microscope, and are generally called micro-LUPI (laser unequal path interferometer) systems. A schematic of a micro-LUPI interferometer is shown in Figure 9. A beamsplitter divides the incident laser beam, half of which travels to a generally flat reference surface, while the other half typically travels through a

microscope objective that acts as a diverger lens and is incident on the test sample.

Due to the long-coherence length of the laser source, the two optical paths do not have to be equal in order to observe interference fringes, and samples may be measured either in reflection or transmission, depending on the desired test result. Collimated, diverging or converging illumination may be brought incident on the sample, depending on the optics in the test beam and their placement. The diverger elements are noncommon-path optics, and reduce the ultimate accuracy of the system to approximately lambda/10, similar to laser-based Fizeau interferometers.

Micro-LUPI-based systems have several advantages for testing small optics over other interference microscopes. First, the ability to measure both transmission and reflection properties of the optics, as well as different incident illumination schemes, means they can test the optics in their use configuration rather than inferring optical properties based solely on surface measurements. They can determine imaging properties, inhomogeneities and thickness variations in addition to characterizing the surfaces under test. Also, the curved test wavefront better matches spherical test surfaces, minimizing the number of observed fringes and making testing of small aspheres easier. Lastly, these systems can have curved reference surfaces as well as flat, and with the proper configuration aberrations in measurements due to the interferometric microscope can be minimized, ensuring high-quality measurements of the optic under test.

These systems have only phase-shifting measurement capability, as the coherence length of the source is too long for white-light interferometric techniques. Transmission tests necessitate a more complex part handling structure, with both the test piece and reflective return optic needing adjustment for tip/tilt and vertical and lateral translation. Therefore, such systems are more costly and not as versatile as a white light optical profiler, but do provide a valuable capability to the micro-optics community.

See also

Coherence: Coherence and Imaging. **Imaging:** Interferometric Imaging. **Interferometry:** Overview; Phase Measurement Interferometry; White Light Interferometry. **Microscopy:** Confocal Microscopy; Overview; Phase Contrast Microscopy.

Further Reading

Advanced Light Microscopy: Measuring Techniques; *Maksymilian Pluta*; North-Holland (1993).

Born M and Wolf E (eds) (1999) *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, 7th edn. Cambridge University Press.

Creath K (1988) Phase-measurement interferometry techniques. In: Wolf E (ed.) *Progress in Optics*, vol. XXVI, pp. 349–393. North-Holland, Amsterdam.

Hariharan P (2003) *Optical Interferometry*, 2nd edn. Academic Press.

Hariharan P and Malacara D (eds) (1995). Selected Papers on Interference, Interferometry, and Interferometric Metrology. *SPIE Milestone*. MS 110.

Krug W, Rienitz J and Shulz G (1964) *Contributions to Interference Microscopy* (translated from the German edition of 1961 by J. Home Dickson): Hilger and Watts Ltd.

Malacara D (ed.) (1992) *Optical Shop Testing*. John Wiley & Sons.

Massie NA (ed.) (1988). Interferometric Metrology. *SPIE* 816.

Polarvarapu PL (ed.) (1997) *Principles and Applications of Polarization-Division Interferometry*. John Wiley & Sons.

Tolansky S (1970) *Multiple-beam Interferometry of Surfaces and Films*. Dover Publications.

Nonlinear Microscopy

S Lévêque-Fort and P Georges, Université Paris Sud, Orsay, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Nonlinear microscopy has become an important tool for the investigation of biological applications where high resolution and three-dimensional imaging are essential for understanding the underlying biological function. Two-photon excitation fluorescence (TPEF) microscopy has opened the way to nonlinear techniques and is now widely available. As this new technology enables noninvasive study of biological specimens with high resolution, the number of applications in biology and medicine is still increasing. By exploiting autofluorescence of cells or tissues components, or by coupling with genetically modified fluorescent indicators or small extrinsic probes, high

contrast and localization can be obtained. Different detection techniques (fluorescence lifetime imaging (FLIM), fluorescence correlation spectroscope (FCS), intensity, and spectral imaging) permits the analyze of several fluorescence properties and so obtains complementary informations.

The successful development of two-photon techniques has generated substantial interest in the scientific imaging community in further developing and exploiting optical nonlinearities for generating new image contrast. All these new developments have strongly benefited from the major advances in ultrafast lasers which have become more compact and reliable. Second- and third-harmonic generation (SHG, THG) as well as coherent anti-Stokes Raman scattering microscopy (CARS) have thus benefited from the fact that it is now easier to generate ultrashort pulses at almost any desired wavelengths. Association of these new nonlinear techniques in the same instrument (TPEF-SHG/TPEF-CARS, TPEF/THG) is commonly encountered and provides a powerful instrumental tool for probing biological samples.

Basic Principles of Two-Photon Microscopy (TPM)

Two photon microscopy is preferred over single photon method in many biological applications due to several intrinsic advantages which arise from having a nonlinear intensity depend absorption. In general, the nonlinear polarization for a material can be expressed as

$$P = \chi^{(1)}E^1 + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \cdots \qquad [1]$$

where P is induced polarization, $\chi(n)$ is the nth order nonlinear susceptibility, and E the electric field vector of the incident light. The first term describes normal absorption and reflection of light; the second term describes SHG, hyper-Rayleigh scattering, sum and difference frequency generation, and the third terms covers multiphoton absorption, third-harmonic generation and stimulated Raman processes.

In 1931 Maria Goppert Mayer predicted that photons of lesser energy can together cause an excitation 'usually' produced by the absorption of a single photon of higher energy, in a process call multiphoton excitation. The hypothesis could not be confirmed until the invention of pulsed ruby lasers, in the 1970s, and was first applied to microscopy with the work of Denk *et al*. Two-photon microscopy used the simplest version of her theory: two photons of equal energy (from the same laser) can excite a molecule, and the excitation is similar to that obtained with a single photon possessing twice the energy (Figure 1). Thus, the molecule can emit a fluorescent photon just as if it was excited with a single higher energy photon (Figures 1 and 2). Of course, the absorption of the two photons must occurs nearly simultaneously ($\sim 10^{-16}$ s), resulting in a quadratic dependence on the light intensity rather than the linear dependence of conventional one photon fluorescence.

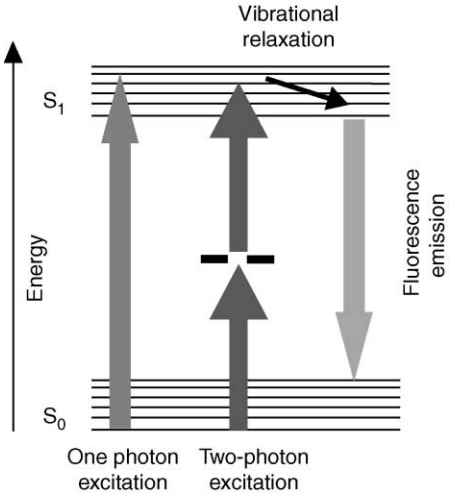


Figure 1 Jablonski diagram for one and two-photon excitation.

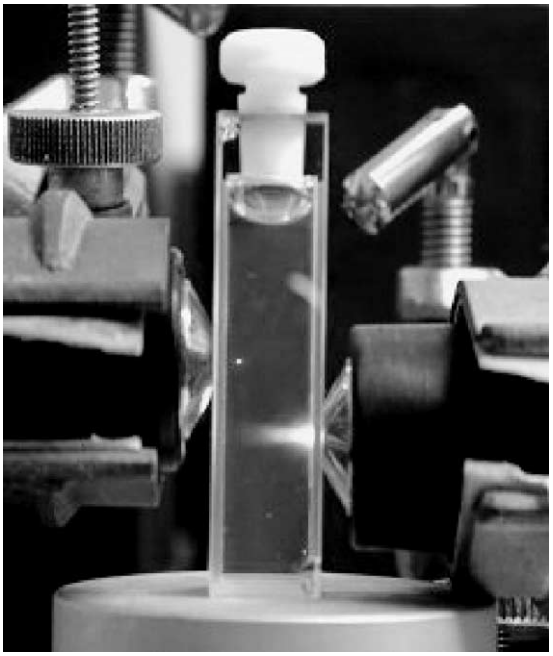


Figure 2 Localization of excitation under two photon excitation (left) compared to one photon excitation (right).

The nonlinear optical absorption property of two-photon excitation, limits the fluorochrome excitation to the point of focus. This localized excitation volume results in greatly improved axial resolution and image contrast compared with conventional microscopy, but also significantly decreases sample photobleaching and photodamage by reducing the region of photointeraction (Figure 2).

Another major advantage of TPEF is its ability to image deeper inside biological samples, owing to the reduced scattering and absorption of near infrared wavelength compared to UV and visible light. Near infrared light (700–1000 nm) is located within the 'therapeutic optical window' where the absorption of the main cells and tissue components is low.

TPEF also permits a high rejection of residual excitation light: in standard one-photon microscopy, the excitation wavelength can be spectrally close to the fluorescence emission band. Filters used to eliminate the leakage of excitation light into the detection channel often cuts off a part of the fluorescence emission. For TPEF excitation and emission band are well separated, and highly efficient filters can remove the residual excitation light without compromising the microscope sensitivity.

Under two-photon excitation, the fluorescence photon flux depends on the square of incident photon flux:

$$F \propto \sigma_{2P} C I^2 \qquad [2]$$

where σ_{2P} is the two-photon cross-section (cm^4 s) with 10^{-50} cm^4 s, called a Goppert–Mayer (GM), C the fluorophore concentration, and I^2 the incident photon flux. The two-photon cross-section is a quantitative measure of the probability of two-photon absorption and as a second-order process is low. As it is difficult to directly measure two-photon cross-sections, we usually use the cross-section of two photon-induced fluorescence (σ_{TPEF}) as the product of two photon absorption cross-section (σ_{2P}) with the fluorescence quantum yield $\phi_F \sigma_{TPEF} = \sigma_{2P} \phi_F$.

Intrinsic fluorophore, such as nicotinamide adenine dinucleotide (NADH), have low cross-sections ($<10^{-4}$ GM), usual fluorescent dye in biology ranging between 1 and 300 GM and quantum dots approaching 50 000 GM. As the probability of two-photon absorption is quiet low, high incident photon flux needs to be delivered to the sample to generate efficient absorption:

$$I \propto E \frac{1}{\Delta\tau \omega^2} \qquad [3]$$

I depends on the energy per impulsion E (mean power/pulse frequency), the pulse duration $\Delta\tau$, and the beam waist ω. The expression of I clearly demonstrates that by using ultrashort pulses and by reducing the beam waist, one can maximize the incident flux. Currently the most commonly used laser source for multiphoton microscopy is the femtosecond titanium-sapphire (Ti:Sa) system. These pulsed femtosecond lasers deliver a 100 fs pulse train at a repetition rate of ~80 MHz with a tuning range from 700 to 1000 nm. The beam waist depends on the objective numerical aperture. If we consider a Gaussian beam profile, with radial and axial parameters of σ_r and σ_z respectively and for a high numerical aperture NA > 0.8:

$$V = \left(\frac{\pi}{2}\right)^{3/2} \sigma_r^2 \sigma_z \qquad [4]$$

with $\sigma_r \approx 0.52\lambda/\sin\alpha$ and $\sigma_z \approx 0.76\lambda/(1-\cos\alpha)$. where λ is the excitation wavelength, α is the one-half of the objective angular aperture ($NA = n \sin\alpha$, n index of the immersion medium). So for a wavelength of 780 nm and an objective working with water immersion with a numerical aperture of 0.9, the excitation volume can be described as an obloidal shape with a radial axis of typically 500 nm and an axial axis of 1.90 μm. It corresponds to an excitation volume of ≈1 fL ou 1 μm^3. Higher resolution can be achieved using an immersion oil objective (NA = 1.3) at the cost of a higher cost and a shorter working distance.

Typical powers used to generate two-photon excitation without damaging the sample are between 1 and 10 mW for the mean incident power. For a laser beam of 1 mW, at a rate of 76 MHz with 100 fs pulse duration, and with an objective leading to a surface in the focal plane of 1 μm^2, the peak power is 0.13×10^9 W/cm^2 at the focal plane of the sample. Of course each samples has its own limits regarding incident intensity, and the key to use TPEF safely is always to reduce the excitation intensity in order to avoid photodamage.

Fluorescent Probes

Biological systems often possess endogenous fluorescent molecules that can be imaged, revealing important sample characteristics without the need of labeling (Figure 3a,b). Cells or tissues can be imaged through their autofluorescent, with the emission of several components, NADH, flavins, collagen fibers, etc. Pharmacology studies also exploit the natural autofluorescence of aromatic drugs, to

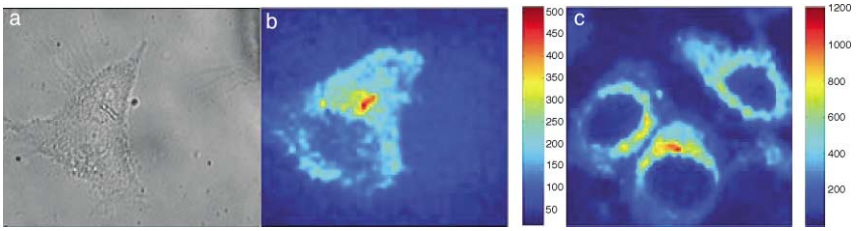


Figure 3 Applications showing various capabilities of TPEF: (a) Hela cells in transmitted image and corresponding autofluorescence image (b) under TPE (780 nm). (c) After adding 40 μM of an anti-HIV drug on similar cells, intrinsic fluorescence of the drug allows us to find its preferable localization inside the cytoplasm.

localize drug inside cells and study their reactivity without any labeling.

However, for some specific applications where the autofluorescence is too low, or if one needs to specifically identify different cells components (Figure 3g), or follow protein mobility, the samples need to be labeled your with extrinsic fluorophores. Extrinsic fluorophores are organic molecules, first developed for one-photon excitation, and are used under two-photon excitation with more or less efficiency as OPE and TPEF have different quantum mechanical selection rules, and the TPE spectrum is not always scaled to the half of OPE spectrum. Current works are focused on the development of specific markers for TPE or SHG to enhance contrast and also reduce incident laser power.

Two new generations of marker have been developed: the development of the green fluorescent protein (GFP) and its mutant introduces a convenient method for monitoring gene expression in cells and tissues; quantum dots are the latest focus in fluorescence labeling. They are promising reporters as their size (1–100 nm) induces a specific narrow emission spectrum in the visible. Absorption can occur over a wide range of short wavelengths because of the existence of multiple electronic states. Time-domain measurement has been performed with CdSe quantum dots which shows multiexponential decay with fluorescence lifetime of 3.4 ns, 16.1 ns, and 35.6 ns, much longer than autofluorescence lifetimes, hence providing greater sensitivity than intensity-based measurement. However, like GFPs, quantum dots display blinking behavior.

Two-photon Applications

Two-photon excitation has become the technique of choice for fluorescence microscopy and is applied to a variety of imaging tasks on samples ranging from single molecules to thick tissues. Owing to the high NIR penetration depth, biopsies and living tissue can be noninvasively investigated by morphological and functional fluorescence imaging of endogenous (Figure 3b) or exogenous fluorophores to depths of more than 300 μm. In genomics, multiphoton multicolour fluorescence *in situ* hybridization (FISH) are used to label various specific regions of DNA, and identified different genes and chromosome regions. In pharmacology, a higher penetration and localization of photodynamic therapy is observed, detection of low drug concentration inside cells facilitates the reactivity comprehension (Figure 3b,c). Neurobiologists investigate calcium dynamics inside brain slices and live animals, or follow neurotransmitter diffusion. Of course, these are only a few examples among the increasing number of applications, and some of them are illustrated in Figure 3.

Two-photon Instrumentation

Two-photon microscopes are now commercially available, however a large number of groups have built their own from separate components or by modifying an existing confocal microscope. Homemade microscopes present a higher configuration flexibility for excitation/detection at a much lower price.

Femtosecond, picosecond, and continuous wave (CW) laser sources have been used for TPM. By using a cw laser sources such as ArKr and NdYag, the average power increases by ~200-fold to achieve the same excitation rate obtained with a femtosecond laser, which is not highly desirable for biological samples. So as previously shown, the TPE efficiency is higher with shorter pulses, so Ti:Sa femtosecond lasers are most commonly used.

One of the drawbacks of TPM is that it is a laser scanning technique. To reconstruct an image you need either to move the sample or the laser beam. Most commonly $x-y$ mirrors driven by a galvanometer scanner are used to scan the laser at the focal plane of the sample. Lenses are used to expand the beam to overfill the objective back aperture and so obtain a diffraction limited focal volume. The excitation light enters the modified inverted epi-fluorescence microscope (Figure 4).

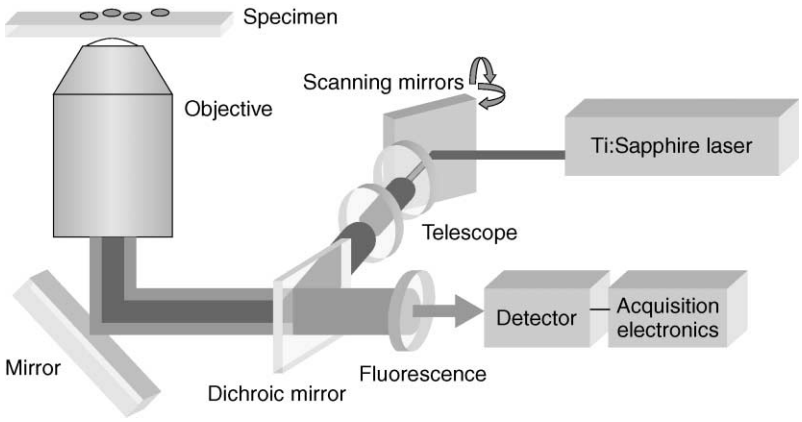


Figure 4 Schematic of a typical two-photon microscope.

A dichroic mirror transmits the excitation light and reflects fluorescence. High numerical aperture objectives are used to provide a high resolution but also maximize fluorescence collection. They are usually mounted on piezo-electric translators to provide a fast Z-scan. The fluorescence is collected by the same objective, reflected by the dichroic to the detector. A filter is placed in front of the detector to remove any residual excitation light. Typical detectors used are the photomultiplier tube (PMT), microchannel plate-PMT (MCP-PMT), avalanche photodiode or charge-coupled device (CCD) cameras. PMTs are the most commonly implemented because they are robust, low-cost, have large, active areas, and good sensitivity.

As TPE is intrinsically confined to the focal volume, two-photon microscopy unlike confocal microscopy, does not need emission pinholes and descanning optics to achieve axial depth discrimination Some authors have, however, added a pinhole to enhance resolution but at the cost of important sign al loss. Concerning the design of the detection pathway, depending on the size of the sample scanned, the objective magnification and the detector area, one may need to use a lens to collect the fluorescence onto the detector.

Fluorescence Detection

Contrast in fluorescence imaging can be provided through several complementary parameters such as steady-state intensity, spectral dependence (at absorption and emission), or fluorescence lifetime.

Intensity Imaging

Intensity images are obtained by summing all the incoming photons during an acquisition time (typically $\sim$1 to 10 ms) which strongly depends on the fluorescence quantum yield of the sample. The whole samples intensity image is obtain by scanning the beam on the focal plane with the $X-Y$ scanner, leading to an acquisition time of 25 s for a 50×50 μm image (1 μm step).

Spectral Imaging

Spectral imaging can be a powerful way to discriminate different molecules, and to identify cellular components or perform co-localization of different proteins. It can be performed by either using an interference filter in front of the detector to extract the contribution in a specific band, or more precisely with a spectrometer. However, autofluorescent spectral bands are usually broad and sometimes it is difficult to separate the different emissions. Extrinsic trackers are thus used to labels specifically, for example, the nucleus, the plasma membrane, and mitochondria, and each tracker usually emits in a specific and well-known spectral band, allowing to easily discriminate and localize the different cellular components. Quantum dots offer narrow emission spectra and allow high spatial resolution, and with the rapid development of their functionalization, they should lead to rapid progress in intracellular imaging.

Fluorescence Lifetime Imaging (FLIM)

Imaging of fluorescence lifetimes is a useful contrast-enhancing method that provides information related to the microenvironment with spatial resolution. The fluorescence lifetime τ corresponds to the average time a fluorophore stays in its excited state is given by $\tau = 1/(k_f + k_{nr})$ with k_f the radiative decay and k_{nr} the nonradiative decay rate.

The radiative decay rate is a characteristic property of a molecule and therefore may be used to contrast different chemical species or fluorophores.

The nonradiative decay rate is a function of the local environment of an excited molecule, and so fluorescence lifetime measurements can provide information about environmental changes (e.g., pH, pO_2, $[Ca^{2+}]$). Furthermore, different cellular probes or tissue components with similar emission spectra but different lifetimes can be distinguished. Another advantage of lifetime measurement is that as a relative measurement of the fluorescence intensity, it does not depend on a non-uniform intensity excitation of the sample, and also does not depend on molecular concentration, allowing one to detect very low quantity of molecules.

Following a delta excitation, the fluorescence intensity de-excitation is given by$I(t) = I_0\, e^{(-t/\tau)}$, with I_0 the initial amplitude. Fluorescence lifetime measurement has been demonstrated either in frequency or time-domains. Time-resolved measurements in the frequency domain are based on the determination of demodulation M and phase shift $\Delta\varphi$ and require modulated light sources (modulation frequency $\omega = 2\pi f$).

The excitation source $I_E(t)$ and induced fluorescence $I_F(t)$ can be written as:

$$\begin{aligned} I_E(t) &= I_{E0} + I_{E\omega}\cos(\omega t + \varphi_E) \\ I_F(t) &= I_{F0} + I_{F\omega}\cos(\omega t + \varphi_F) \end{aligned} \qquad [5]$$

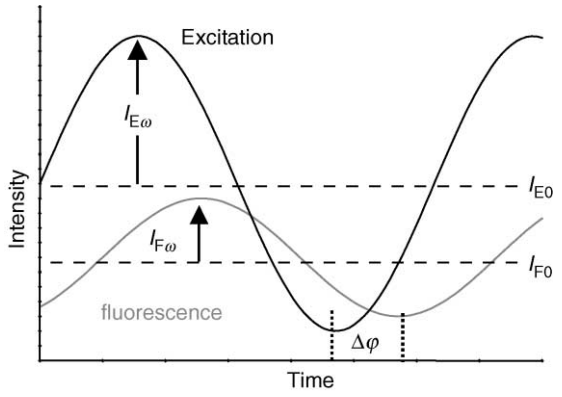


Figure 5 Time course of excitation and emission light in frequency domain fluorescence lifetime measurements techniques.

In the case of fluorescence with mono-exponential decay, the lifetime τ can deduced from the relations (Figure 5):

$$M = \frac{I_{F\omega}/I_{F0}}{I_{E\omega}/I_{E0}}. \qquad \omega\tau = \tan(\Delta\varphi) = \sqrt{\frac{1}{M^2} - 1} \qquad [6]$$

In order to translate the phase and demodulation information from the 80 MHz frequency to a lower kHz frequency, a modulated gain $G(t)$ at the frequency ω' can be applied to the detector (heterodyne detection):

$$G(t) = G_0 + G_{\omega'}\cos(\omega' t + \varphi_G) \qquad [7]$$

The detector response $S(t)$ is thus proportional to the fluorescence intensity:

$$\begin{aligned} S(t) = {} & G_0 I_{F0} + \frac{G\omega' I_{F\omega}}{2}\frac{1}{\sqrt{1+(\omega\tau)^2}} \\ & \times\cos[(\omega - \omega')t + \varphi_G - \varphi_E + \Delta\varphi] \end{aligned} \qquad [8]$$

In the time-domain, the single point detection system is called time-correlated single-photon counting (TCSPC). The principle of the TSCPC system is shown in Figure 6. For each excitation pulse a trigger signal is sent to a time-to-amplitude converter (TAC) within which a capacitor starts charging linearly with time. Upon arrival of each fluorescence photon on the detector a second trigger pulse stops the capacitor charging. The charge is proportional to the delay between the excitation pulse and the fluorescence photon. By repeating the measurement over many thousands of excitation pulse, a histogram of photon arrival can be build up. As each photon must be record individually, the incident power is reduce to avoid the detection of only photons from the beginning of the decay (pile-up) which would lead to shorter lifetimes. To match detector dead-time and TAC counting rate, the TAC is oftenly used in reverse mode: the fluorescence photon starts the TAC and the following pulse stops it.

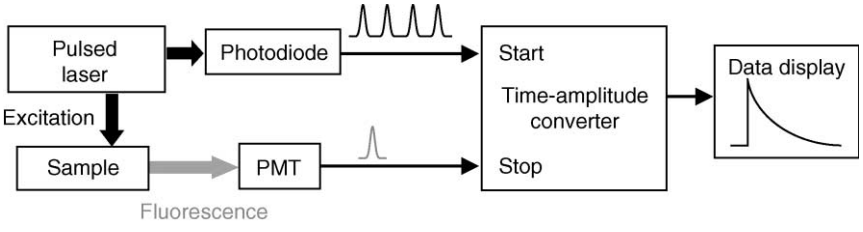


Figure 6 Schematic of time correlated single photon counting system.

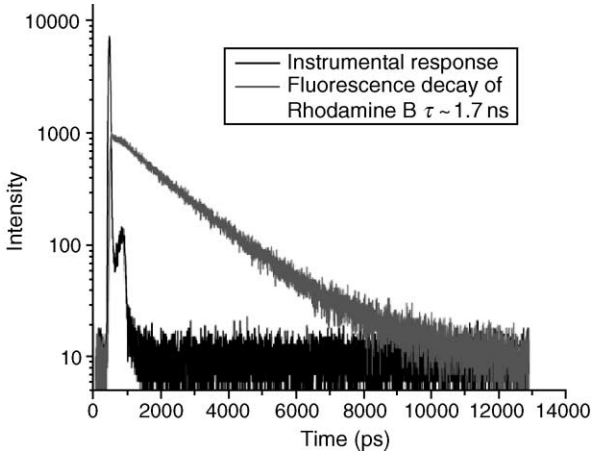


Figure 7 Typical instrumental response of MCP-PMT (black curve), and fluorescence decay curve of rhodamine B (grey curve) with $\tau \sim 1.7$ ns.

Typical detectors used are the photomultiplier tube (PMT), microchannel plate-PMT (MCP-PMT), and avalanche photodiode.

Typical instrumental response of our MCP-PMT is represented in Figure 7 (black curve) leading to a resolution of 30 ps. Depending on the fluorescence quantum yield of the sample, in order to have a good signal to noise ration, acquisition time is ranging between 1 and 30 s per point. Fluorescence decay of rhodamine B with a lifetime $\tau \sim 1.7$ ns is also presented in Figure 7.

For both detection approaches, single point lifetime measurements can be extended to lifetime imaging by synchronizing the sample scanning and the lifetime detection. However, due to the long time acquisition needed, images are only limited to small interesting areas rather than the whole sample. We will see in the recent development section below, new ways to overcome this issue.

Fluorescence Correlation Spectroscopy (FCS)

FCS experiments are performed by recording fluorescence intensity fluctuations (i.e., fluorescence bursts (Figure 8)) from the small focal volume (<1 fl). Applications include diffusion, chemical reactions, molecule concentration, or hydrodynamic flow–FCS under TPE benefits from all the advantages of TPE versus OPE.

Various experimental parameters can be extracted through temporal analysis of fluorescence intensity fluctuations by calculating the autocorrelation or cross-correlation of the fluorescence signal defined as:

$$g(\tau) = \frac{\langle I_F(t) I_F(t+\tau)\rangle}{\langle I_F^2(t)\rangle} = \frac{\int I_F(t) I_F(t+\tau)\mathrm{d}t}{\int I_F^2(t)\mathrm{d}t} \qquad [9]$$

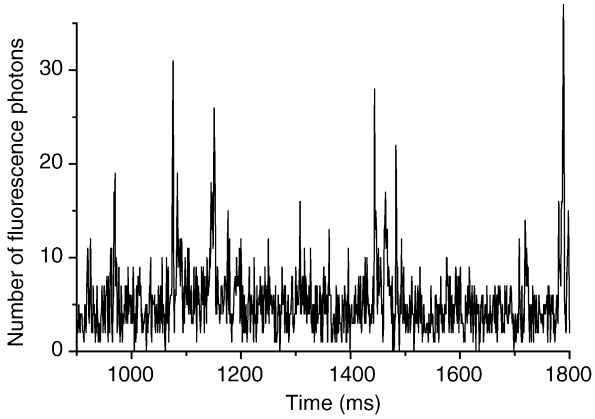


Figure 8 Fluorescence intensity raw data along the time. Bursts are observed when molecules cross the focal volume.

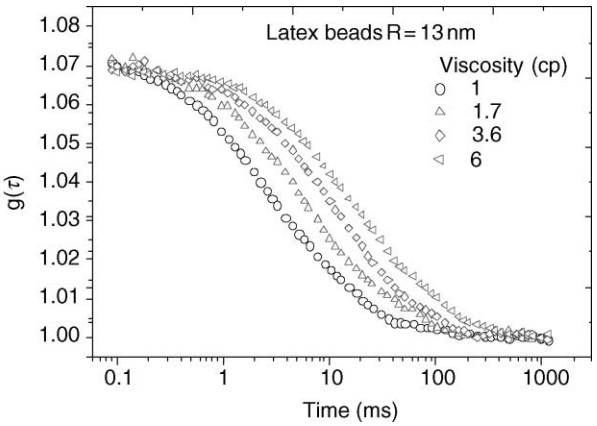


Figure 9 Correlation function observed for microspheres placed in different viscous solution.

where I_F is the time-dependent fluorescence signal. By considering that the fluorescence intensity fluctuates around a mean value:

$$I_F(t) = \langle I_F\rangle + \delta I_F(t) \qquad [10]$$

the autocorrelation function can be written as:

$$g(\tau) = 1 + \frac{\langle \delta I_F(t)\delta I_F(t+\tau)\rangle}{\langle I_F^2\rangle} \qquad [11]$$

FCS have been extensively used to extract diffusion coefficient of small fluorescent latex bead in solution, showing the influence of the bead radius or solvent properties. Figure 9 clearly exhibits the influence of the viscosity of the solvent on the diffusion of microspheres. Single molecule detection has also been reached. FCS as also been extended to biological applications where the low background associated with TPE and the reduced autofluorescence is critical. Diffusion of microbeads inside cells cytoplasm

or bacterial biofilms have been successfully demonstrated.

Recent Developments in Two-Photon Microscopy

Towards video microscopy

All the nonlinear microscopies share an important limitation: the image build up pixel-by-pixel by scanning across the sample, leading to long-time acquisition. Depending on the sample's emission and the area studied, scanning rates between 0.5 Hz and 30 Hz can be reached for intensity images. To speed up this acquisition time, a first approach is based on a line excitation, in combination with widefield CCD detection. This provides real-time image acquisition rates at the expense of a reduced axial resolution (5 μm instead of 1 μm). A more ideal way of spreading the available excitation power over the sample and speed up image buildup, is the simultaneous use of several well separated foci, each of them generating a diffraction limited spot. In combination with whole field detector, as with a CCD camera, image acquisition rates can be strongly enhanced without compromising resolution. The spacing between the foci must be carefully chosen in order to avoid out-of-focus interference. By introducing time-delay between the different foci, larger than the pulse duration, each foci of a microlens array can be position very closed without loss of resolution. An alternative way has been proposed; a special beamsplitter has been developed to create multiple beams with a higher efficiency, furthermore this setup inherently induced time-delay between each foci. Time-domain has been demonstrated in both approach, and FLIM maps on cells imaging have been obtained in only few second with 200 ps time resolution.

Sharpen the focal point

Nonlinear microscopy facilitates 3D-imaging but the resolution is still limited to 180 nm in the focal plane and 500–800 nm along the optical axis. According to the diffraction theory, the resolution of a focusing light microscope is related to the size of its focal point. In the mid-1990s, new concepts appeared to break this diffraction theory and sharpen the focal volume. In 4-pi microscopy the sample is placed between two large NA objective lenses in a single axis at their common focus. Coherent interference of beams passing through the two lenses generate a point spread function whose central maximum is 3–4 times narrower than in the normal confocal case, but which also has side lobes. This last issue can be overcome using two-photon excitation. A two-photon 4-pi microscopy can thus result in a resolution of ~100 nm in all directions, and has been demonstrated successfully on filamentous actin and immunofluorescently labeled microtubules. However, a 4-pi microscopy requires the sample to be thin and mounted between two coverslips, unless one of the lenses is a dipping lens. Recent development on sample chambers allows one to overcome this issue, and results in live mammalian cells have been shown. A compact 4-pi commercial version has recently been proposed (Leica), achieving a seven-fold-improved axial resolution (80 nm) compared to the classical two-photon microscope.

Another technique, also developed by Professor Hell's group, is called stimulated emission depletion (STED) microscopy. The fluorophore in the fluorescent state S1 is stimulated to the ground state S0 with a doughnut-shaped beam. The depletion of the S1 saturates with increased I_{STED} intensity, and therefore establishes a nonlinear relationship between the fluorescence and the intensity applied to the STED. The saturation of S1 depletion is the essential element for breaking the diffraction limit as it confines fluorescence to the center. Spot size is not limited, but in practise depends on the quality and the saturation factor limited depletion. Using STED microscopy with wavelengths of 750–800 nm leads to a lateral resolution of up to 28 nm on single molecules experiments.

STED microscope has been successfully combined to 4-pi microscopy, and lateral resolution of 30–40 nm has been demonstrated.

Others Nonlinear Imaging Techniques

The success of two-photon microscopy has opened the way for other nonlinear microscopy. As opposite to multiphoton fluorescence absorption techniques where the induced signal is incoherent fluorescence, the signal generated by harmonic generation and stimulated Raman scattering is coherent with specific directional and polarization properties, as such fluorescence and harmonic images are derived from fundamentally different contrast mechanisms. However, both techniques share many features like intrinsic three-dimensionality, reduce photobleaching or deep sectioning, and require essentially identical equipment, allowing one to combine these two techniques as discussed further below.

Second-Harmonic Generation Microscopy

SHG is a second-order nonlinear optical process that can only arise from a medium lacking a center of symmetry, for example, an anisotropic crystal or at an

interface of a membrane. SHG can be used to image tissue, muscles, cells membranes, or protein array, and has been recently reported to enhance membrane potential by using voltage sensitive dye.

SHG does not arise from an absorptive process, as the intense laser field induces a nonlinear, second-order, polarization in the assembly of molecules, resulting in the production of a coherent wave at exactly twice the incident frequency (half the wavelength). As represented in Figure 10, for an excitation at 800 nm, the SHG is observed at 400 nm and depending on the fluorescence properties of the molecule, a broadband fluorescent spectra is emitted in the visible.

Contrary to TPEF, where the emission properties depend only on the molecular properties and are independent of the laser source, the spectral and temporal characteristics of SHG are derived from the laser source. The SHG is temporally synchronous with the excitation pulse and its bandwidth is $1/\sqrt{2}$ of the laser's bandwidth. As SHG is an instantaneous process, the signal is only generated during the duration of the laser pulse (Figure 11), in contrary to TPEF where de-excitation of the molecules can occur over several nanoseconds after the pulse excitation.

A major difference between SHG and TPEF is that SHG requires a noncentrosymmetric environment to produce a signal. By examining the second term of eqn [1] $P = \chi^{(2)}E^2$ where the polarization and electric field are vector quantities and $\chi^{(2)}$ is a tensor, this condition is only satisfied for a birefringent crystal, an order array, or an interface. For example, the induced polarization from molecules in a dye cuvette (centrosymmetric sample) would be equal and opposite in all directions, leading to a vector sum zero. The SHG propagation is essentially in the forward direction, and the emission pattern is confined to twin lobes in the case of surface generation, or a cone in the case of volume generation. The off-axis of the two lobes critically depend on the tight focus of the driving field.

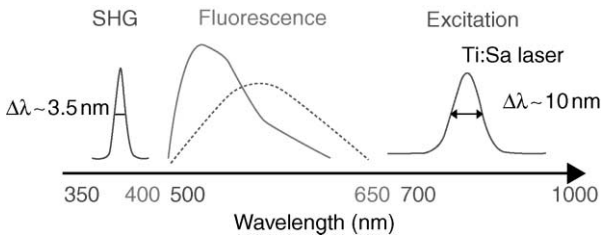


Figure 10 Wavelength domain for two-photon induced fluorescence and second-harmonic generation: excitation meets the same criteria and Ti: Sa are typically used in the range 700–1000 nm. For a typical excitation at 800 nm, fluorescence will be observed over ~450 nm, and SHG, at exactly 400 nm.

The total radiated SHG power can be derived by integrating the radiation profile from all the emission directions: $P_{SHG} = \frac{1}{2}\theta_2 N^2 \sigma_2 I_0^2$ with $\sigma_2 \propto |\chi_2|^2$, I_0 the driving field intensity at the focal center, N the effective number of radiative molecules, and $\theta_2 \approx (0.1\lambda/w_0)^2$.

A schematic of a typical SHG/TPEF microscope is shown in Figure 12. The excitation source is a mode locked Ti:Sa laser which delivers ~120 fs at 76 MHz. The laser light is focused into the sample with a water immersion microscope objective (Olympus, LUM-planFl 60x) and the resultant SHG is collected in a forward direction, while TPEF is collected in the backward direction.

Figure 13 illustrates SHG and TPEF images of Di-6-ASPBS molecules under 880 nm excitation.

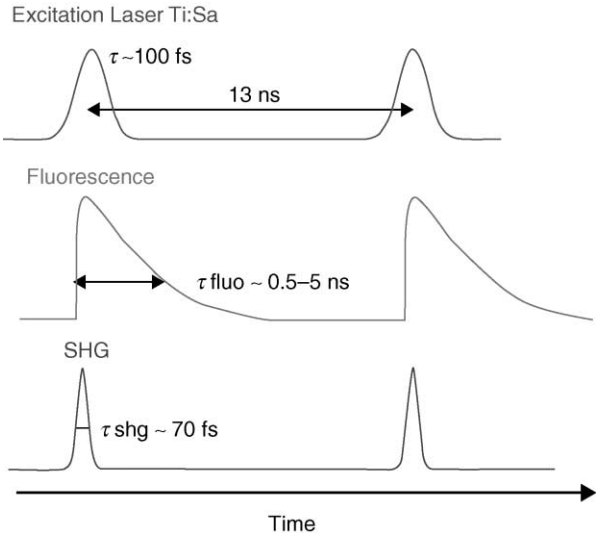


Figure 11 Typical time-response for two-photon induced fluorescence and SHG. SHG signal is an instantaneous process whereas fluorescence emission can occur several nanoseconds after the pulse excitation.

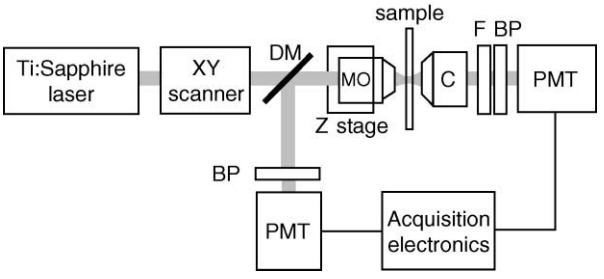


Figure 12 Schematic of SHG setup combined with TPEF extracted from (DM dichroic mirror, MO microscope objective NA 0.9 water immersion, C condenser, F colored glass filter, BP bandpass filter, PMT photomultiplier tube). Reproduced with permission from Moreaux L, Sandre O, Blanchard-Desce M and Mertz J (2000) Membrane imaging by simultaneous second-harmonic generation and two-photon microscopy. *Optics Letters* 25: 320–322, © 2000 Optical Society of America.

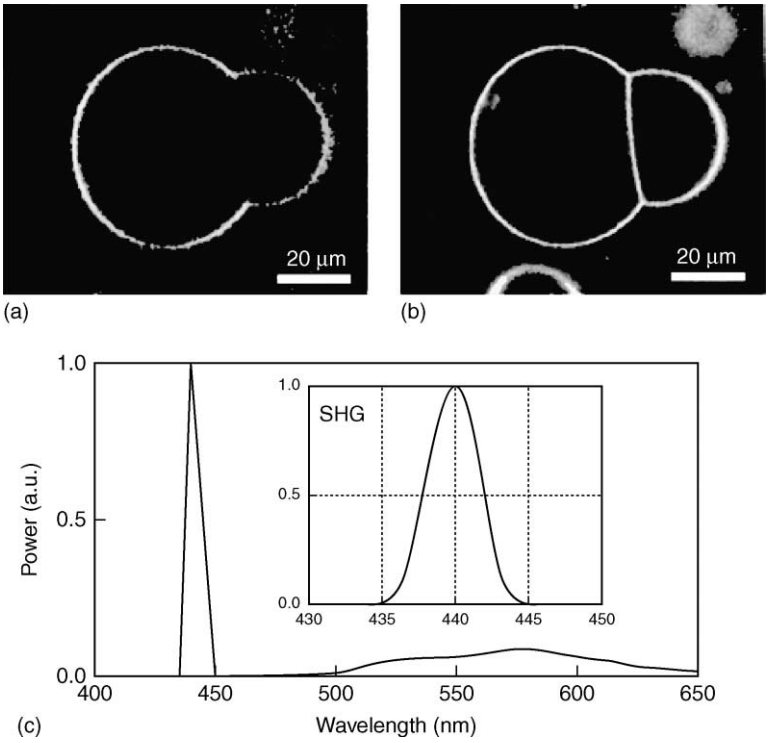


Figure 13 (a) SHG and (b) TPEF of two adhering vesicles labeled with Di-6-ASPBS (equatorial slice) excited at 880 nm. (c) The forward-detected SHG spectra (left peak) and the backward broad band fluorescence spectra (right peak). The Stokes shift in fluorescence leads to an important spectral separation between the two peaks. An expanded view of the SHG peak is presented in the inset. Reproduced with permission from Moreaux L, Sandre O, Blanchard-Desce M and Mertz J (2000) Membrane imaging by simultaneous second-harmonic generation and two-photon microscopy. *Optics Letters* 25: 320–322, © 2000 Optical Society of America.

The total acquisition time for the images was 1.5 s for an excitation power less than 1 mW. The adhesion area where the membranes are fused exhibits a centrosymmetric molecular distribution wherein TPEF is allowed, but not SHG. The corresponding forward-detected emission spectra of SHG (left peak) and TPEF (right peak) is shown in Figure 13c.

Third-Harmonic Generation (THG)

Third harmonic generation takes advantage of the discontinuity of third-order nonlinear susceptibility $\chi^{(3)}$ at material interfaces. At the focus point of an input laser beam with wavelength λ, this discontinuity gives rise to a new wave in the forward direction at the third-harmonic, $\lambda/3$. This signal can be analyzed in the transmission mode either with a single pixel detector or with a CCD camera. As with the other nonlinear techniques, THG offers a high 3D sectioning capability. Since all materials have nonvanishing third-order susceptibilities, THG microscopy can utilized as a general-purpose microscopy technique, with no need for fluorescence labeling or staining. Experimental setup for THG are very closed to SHG microscope. The need for efficient THG in the sample, together with low average laser power to avoid heating, dictates a laser source that operates with femtosecond pulses. The lasers that are often used THG microscopy are a synchronously pumped OPO at the wavelength of 1500 nm, a low repetition-rate OPA at 1200 nm and a Cr:Forsterite laser at 1240 nm. All these lasers are in the infrared above 1200 nm, thus generating THG signals in the visible range. These visible signals are compatible with the collection optics. A infrared laser undergoes low scattering, however absorption in water strongly increases.

THG have been demonstrated from subcellular to millimeter-size organisms with imaging of nerve cells, chloroplasts of a leaf, or Ca^{2+} intracellular dynamics.

Coherent Anti-Raman Stokes (CARS)

Chemical imaging by use of inherent molecular vibration signals avoids the photobleaching problem and perturbations to cell function induced by fluorophore labeling. Confocal Raman microscopy can provide 3D spatial resolution but it requires high average power due to small cross-section of Raman scattering, or long-time exposure which is not desirable in studying dynamic living system.

The above difficulties can be circumvented by multiphoton vibrational microscopy based on coherent anti-stokes Raman scattering (CARS). CARS is a four-wave mixing process in which a pump field $E_p(\omega_p)$, a Stokes field E_S (ω_S), and a probe field $E_{p'}$ $(\omega_{p'})$. Interact with a sample and generate an anti-Stokes field E_{AS}, at the frequency $\omega_{AS} = \omega_p - \omega_S + \omega_{p'}$. In practise, the pump and the probe fields derive from the same laser $\omega_p = \omega_{p'}$ and $\omega_{AS} = 2\omega_p - \omega_S$. The energy diagram of CARS is shown on Figure 14.

The CARS intensity is the square modulus of the induced polarization, $P^3 = \chi^{(3)} E_p^2 E_s$, with $\chi^{(3)}$ the third-order susceptibility.

The application of the CARS process to microscopy opens up a new method for chemical imaging. In CARS microscopy, the signal is also dependent on the square of the pump field intensity, so the temporally and spatially overlapped pump and Stokes laser pulses tightly focused into a sample, generate a signal of a small excitation volume ($<1\ \mu m^3$). As with all other nonlinear imaging techniques, the image is obtained by raster scanning the sample or the laser beams. The vibrational contrast in CARS microscopy arises from signal enhancement when $\omega_p - \omega_s$ is tuned to Raman-active vibrational frequency Ω. CARS signal generation needs to fulfill the phase-matching condition, $l < lc = \pi/|\Delta\mathbf{k}| = \pi/|\mathbf{k}_{as} - (2\mathbf{k}_p - \mathbf{k}_s)|$, where $\mathbf{k}_p$, $\mathbf{k}_s$, and $\mathbf{k}_{as}$ are wavevectors of the pump, Stoke and CARS fields, respectively, and $\Delta\mathbf{k}$ is the wavevector mismatch.

CARS presents several advantages similar to TPEF: high 3D sectioning capabilities, a low scattering and absorption of the excitation beams. CARS signal is higher in frequency than one-photon fluorescence, so it can be detected in the presence of a strong fluorescent background. Vibrational contrast is intrinsic to biological samples. Fluorescent probes are not necessary and photobleaching is thus circumvented. The coherent signal accumulation in the CARS process produces a strong and directional signal, making CARS microscopy much more sensitive than the conventional Raman microscopy. Consequently, CARS microscopy requires only a moderate average power tolerable by biological samples and has a high data acquisition speed. However, CARS microscopy is not without shortcomings. First, the intrinsically weak induced nonlinear polarizability requires sophisticated laser excitation sources with high peak power and moderate average power, and with low jitter. Second, unlike fluorescence imaging, CARS measurement is not background free, and the nonresonant signal limits the image contrast.

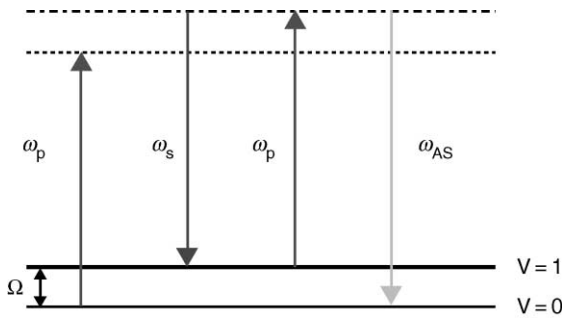


Figure 14 Energy diagram for Coherent anti-Raman spectroscopy (CARS).

The typical setup is very similar equipment compared to TPEF and SHG. The main difference arise from the excitation, two laser beams, the pump, and Stokes laser beams overlapping (in space and time) and are tightly focused into the sample using a high NA objective. The CARS signal can be detected in the forward (F-CARS) and the epi-detection (E-CARS). Unlike fluorescence emission and spontaneous Raman scattering, the radiation pattern of CARS is dependent on both the size and shape of a scatterer. The two detection geometry are thus complementary, the forward detection is suitable for imaging objects of a size comparable to or larger than the excitation wavelength.

For smaller objects, the F-CARS contrast is limited by the large nonresonant background from the solvent. Epi-detected CARS (E-CARS) microscopy provides a sensitive means of imaging objects having an axial length much smaller than the excitation wavelength because it avoids the large background from the solvent.

On the application side, CARS microscopy has been used for mapping the distribution, orientation, and diffusion of specific molecules in dynamic samples, for imaging lipids, proteins, and chromosomes in unstained live cells, and for monitoring cellular processes. CARS is a very promising technique to image small molecules such as lipids, hormones, and drug molecules, for which fluorescence labeling is prone to alter their function in cells and tissues. As for SHG, CARS microscopy is often combined with TPEF in order to obtain simultaneously the complementary contrasts.

Conclusion

Since its introduction more than a decade ago, two-photon fluorescence microscopy has quickly become a standard tool in biophotonic with applications ranging from the study of single molecules to thick tissue imaging. TPEF has opened the way to new coherent nonlinear imaging techniques such as SHG,

THG, or CARS which all shared a high spatial resolution. Based on different contrasts these different approaches are often combined for a more complete diagnosis. With developments of new low cost, compact, and reliable lasers with tuning wavelength covering the spectra of a wide range of endogenous or extrinsic markers, nonlinear imaging will continue its successful implementation outside physics labs. Development of endoscopy systems in which ultrafast pulse duration is preserved, will open a new range of *in vivo* studies. Nonlinear imaging still has more to offer and will definitely be a powerful tool for biomedical applications.

See also

Microscopy: Overview. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation; $\chi^{(3)}$–Third-Harmonic Generation. **Nonlinear Sources:** Harmonic Generation in Gases. **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Cheng J and Xie XS (2004) Coherent anti-stokes Raman scattering microscopy: instrumentation, theory and applications. *Journal of Physical chemistry B* 108: 827–840.

Gratton E, Breusegem S, Sutin J, Ruan Q and Barry N (2003) Fluorescence lifetime imaging for the two-photon microscope: time-domain and frequency-domain methods. *Journal of Biomedical Optics* 8: 381–390.

Hell SW (2003) Towards fluorescence nanoscopy. *Nature Biotechnology* 21: 1347–1355.

Konig K (2000) Multiphoton microscopy in life sciences. *Journal of microscopy* 200: 83–104.

Moreaux L, Sandre O, Blanchard-Desce M and Mertz J (2000) Membrane imaging by simultaneous second-harmonic generation and two-photon microscopy. *Optics Letters* 25: 320–322.

Schwille P, Haupts U, Maiti S and Webb WW (1999) Molecular dynamics in living cells observed by fluorescence correlation spectroscopy with one- and two-photon excitation. *Biophysics Journal* 77: 2251–2265.

So PTC, Dong CY, Masters BR and Berland KM (2000) Two photon micrscopy excitation fluorescence microscopy. *Annual Review of Biomedical Engineering* 399–429.

Squier J and Muller M (2001) High resolution nonlinear microscopy: a review of sources and methods for achieving optimal imaging. *Review of Science Instruments* 72: 7.

Zipfel WR, Williams RM and Webb WW (2003) Nonlinear magic: multiphoton microscopy in the biosciences. *Nature Biotechnology* 21: 1369–1377.

Zipfel WR, Williams RM, Christie R, Nikitin AY, Hyman BT and Webb WW (2003) Live tissue intrinsic emission microscopy using multiphoton-excited native fluorescence and second harmonic generation. *Proceedings of the National Academy of Science of USA* 100: 7075–7080.

Phase Contrast Microscopy

C J R Sheppard, National University of Singapore, Singapore

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Many biological specimens behave as phase objects, that is, they alter the phase of the light transmitted through them, rather than its intensity. Similarly, reflection from a surface with height variations alters the phase of the reflected light. Observation of phase objects requires special imaging techniques. These include defocusing, dark field, Zernike phase contrast, Schlieren imaging, Hoffmann modulation contrast, differential phase contrast, interference microscopy, shearing interferometry, Nomarski differential interference microscopy, and digital phase retrieval. The imaging performance of these different methods are most easily appreciated and compared by consideration of the transfer function approach. In particular, we can consider weakly scattering objects and slowly varying phase objects as special cases.

When a thin unstained biological specimen is illuminated, in general the transmitted light suffers changes in amplitude, phase, and polarization. The change in polarization is related to the birefringence of the sample, which is neglected in the following. Thus the object (complex) amplitude transmittance can be written:

$$t(x, y) = a(x, y)\mathrm{e}^{\mathrm{i}\phi(x,y)} \qquad [1]$$

where $a(x, y)$ is real and represents the modulus (also called amplitude) of the light wave on exiting from the sample, and $\phi(x, y)$, also real, represents the phase. In a perfect microscope, we observe the

intensity of the transmitted light, so the image is given by

$$I(x, y) = |a(x, y)e^{i\phi(x,y)}|^2 = a^2(x, y) \quad [2]$$

Thus the phase information is invisible in the image. A perfect microscope does not image phase variations in the object, but results in an image that depends on the amplitude of the transmittance, called amplitude contrast. Usually the change in modulus is also small, so that the image will exhibit weak contrast, which may not be discernable. In order to render the phase variations visible it is necessary to modify the microscope in one of a number of different ways.

This conclusion is also relevant in incident light microscopy of surface structures. Then phase changes result from variations in surface height, and can be made visible by phase contrast microscopy. Quantitative measurement of the phase allows quantitative measurement of the surface profile. But care must be taken to avoid errors caused by phase change on reflection from a conducting material, or from effects of the numerical aperture of the optical system. In transmission microscopy quantitative phase measurement can be useful too, for example giving the concentration of a solute.

Some of these phase contrast methods are easily understood using the simple imaging model described above. Others are more easily appreciated by considering imaging in terms of Fourier optics, equivalent to Abbe's theory of microscope imaging. We introduce the angular spectrum of the object amplitude transmittance:

$$T(m, n) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} t(x, y)\exp[-i2\pi(mx + ny)]dx\ dy \quad [3]$$

where m, n are spatial frequencies in the x, y directions. We can write for the object amplitude transmittance:

$$t(x, y) = e^{b(x,y)} \quad [4]$$

where $b(x, y)$ is in general complex. For a weak object we have

$$t(x, y) \approx 1 + b(x, y) \quad [5]$$

and the spectrum is

$$T(m, n) = \delta(m)\delta(n) + B(m, n) \quad [6]$$

where δ is a Dirac delta function and $|B| \ll 1$. The problem with imaging of phase arises because a change in the imaginary part of b produces only a small change in the modulus of t.

For a coherent imaging system, with coherent transfer function $c(m, n)$, the image intensity is

$$I(x, y) = \left|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c(m, n)T(m, n) \times \exp\left[\frac{i2\pi}{M}(mx + ny)\right]dm\ dn\right|^2 \quad [7]$$

where M is the magnification. The coherent transfer function $c(m, n)$, is equal to the scaled pupil function $P(\lambda fm, \lambda fn)$, where f is the focal length of the objective. Substituting eqn [6] into eqn [7] and dropping terms of second order we obtain:

$$I(x, y) = |c(0, 0)|^2 + 2\Re\left\{c^*(0, 0)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \times c(m, n)B(m, n)\exp\left[\frac{i2\pi}{M}(mx + ny)\right]dm\ dn\right\} \quad [8]$$

Normalizing the transfer function to unity at the origin:

$$I(x, y) = 1 + 2\Re\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c(m, n)B(m, n) \times \exp\left[\frac{i2\pi}{M}(mx + ny)\right]dm\ dn\right\} \quad [9]$$

For an axially symmetric aberration-free system, the coherent transfer function is real and even, so that a pure phase object is not imaged (only the Hermitian component of B is imaged). Many phase contrast methods in principle rely on appropriate modification of the coherent transfer function, either by making it complex or asymmetric.

However, this theory is an approximation as microscope systems are in practice partially coherent. For a partially coherent system the image intensity is

$$I(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C(m,n;p,q)T(m,n) \times T^*(p,q)\exp\left\{\frac{i2\pi}{M}[(m-p)x+(n-q)y]\right\}dm\,dn \quad [10]$$

where $C(m,n;p,q)$ is the partially coherent transfer function, sometimes called the transmission cross-coefficient, which can be calculated as an integral over the objective and condenser pupils. The partially coherent transfer function is nonzero over a finite region of $m,n;p,q$ space, thus limiting the resolution of the system. Then eqn [8] for a weak object

becomes, normalizing so that $C(0,0;0,0)=1$:

$$I(x,y)=1+2\Re\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}C(m,n;0,0)B(m,n)\times\exp\left[\frac{\mathrm{i}2\pi}{M}(mx+ny)\right]\mathrm{d}m\,\mathrm{d}n\right\} \quad [11]$$

where $C(m,n;0,0)$ is called the weak object transfer function.

For the image of a general object that is not necessarily weak, eqn [10] is often difficult to interpret. We can gain much insight into the imaging process by considering another special form of object, this time a slowly varying object:

$$t(x,y)=a(x,y)\exp\left[\mathrm{i}\left(x\frac{\partial\phi}{\partial x}+y\frac{\partial\phi}{\partial y}\right)\right] \quad [12]$$

where a and ϕ are assumed slowly varying. Thus, the object spectrum, assumed now as slowly spatially varying, is

$$T(m,n)\approx a\delta\left(m-\frac{1}{2\pi}\frac{\partial\phi}{\partial x}\right)\delta\left(n-\frac{1}{2\pi}\frac{\partial\phi}{\partial y}\right) \quad [13]$$

Substituting in eqn [10], we then find:

$$I(x,y)=a^2(x,y)C(m,n;m,n)$$
$$m=\frac{1}{2\pi}\frac{\partial\phi}{\partial x},\ n=\frac{1}{2\pi}\frac{\partial\phi}{\partial y} \quad [14]$$

in which C is called the phase gradient transfer function, and is real. For an ordinary bright field system, C is an even function. Thus a phase gradient results in a change in image intensity, but for an axially symmetric system the sign of the change is independent of the sign of the phase change.

Defocused Imaging

Historically, the oldest method of phase contrast is simply to defocus the object. The transfer function is then complex, and phase detail is imaged. If the system is aberration-free, the contrast of the phase information reverses on the opposite side of focus. The strength of the contrast increases with defocus, until eventually artifacts are introduced when the imaginary part of the transfer function becomes negative. If the relative condenser aperture is too large, contrast is reduced.

Assume a pure weak phase object:

$$t(x,y)=1+\mathrm{i}d\cos(2\pi\nu x) \quad [15]$$

so that

$$T(m,n)=\left[\delta(m)+\mathrm{i}\frac{d}{2}\delta(m-\nu)+\mathrm{i}\frac{d}{2}\delta(m+\nu)\right]\delta(n) \quad [16]$$

Then:

$$I(x,y)=1-dC_{\mathrm{i}}(\nu,0;0,0)\cos\left(\frac{2\pi\nu x}{M}\right) \quad [17]$$

where C_{i} is the imaginary part of C. Thus the phase information is imaged by the imaginary part of the weak object transfer function.

Dark Field Microscopy

Dark field microscopy can be used to visualize weak amplitude and phase information. The constant background in eqn [5] is eliminated, so we obtain an image of b. Note that eqns [8] and [11] are no longer valid because we cannot neglect the second-order term. For a coherent system, the background can be eliminated by use of a central stop to intercept the nondiffracted light. For a partially coherent microscope, an annular condenser aperture (larger in diameter than the objective pupil) is used (Figure 1). In both cases, in practice a range of low spatial frequencies are eliminated, resulting in the appearance of halo artifacts.

For the pure weak phase object of eqn [12], the image in a coherent microscope is

$$I(x,y)=\frac{1}{2}d^2c^2(\nu,0)\left[1+\cos\left(\frac{4\pi\nu x}{M}\right)\right] \quad [18]$$

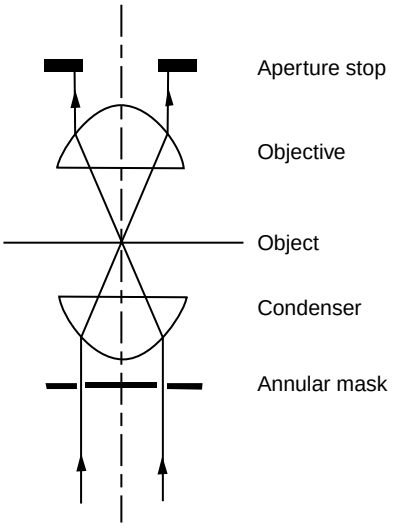


Figure 1 Schematic diagram of dark field microscopy.

so that the spatial frequency is double that in the original object. For the partially coherent microscope, we have

$$I(x,y) = \frac{1}{2}d^2\Big[\, C(\nu,\nu;0,0) + C(\nu,-\nu;0,0)\cos\left(\frac{4\pi\nu x}{M}\right)\Big] \qquad [19]$$

in which the two terms represent difference and sum frequency components. However, it can be shown that for the annular dark field system the transfer function for sum frequencies is zero, so the image intensity is a constant. The system images only difference frequencies of pairs of spatial frequencies, and not sum frequencies.

For a slowly varying phase object, zero intensity results from a region of constant phase, but the intensity rises quickly for small values of phase gradient, thus giving strong contrast. For larger phase gradients, the intensity falls off slowly.

Zernike Phase Contrast

The Zernike phase contrast method is similar to dark field except that the direct (nondiffracted) light has its relative phase changed by 90°, rather being eliminated. In practice this is achieved using an annular condenser aperture and an objective with a phase ring (Figure 2). The phase change can be ±90°, giving positive or negative phase contrast. The phase ring is usually only partially transmitting, which has the effect of enhancing the sensitivity. From eqns [8] or [11], imaging of a weak object is linear in phase. The halo artifact is present, as in dark field microscopy. Weak amplitude information is imaged as in dark field microscopy.

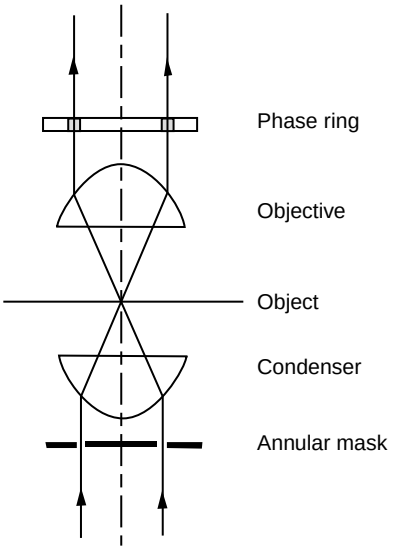


Figure 2 Schematic diagram of Zernike phase contrast microscopy.

For the weak phase object eqn [12], we then have

$$I(x,y) = 1 \pm \frac{2d}{g}C(\nu,0;0,0)\cos\left(\frac{2\pi\nu x}{M}\right) \qquad [20]$$

for positive or negative phase contrast, respectively, where g is the amplitude transmittance of the phase ring. The response of the weak object transfer function is improved as a result of the annular condenser aperture.

Schlieren Imaging, Hoffman Modulation Contrast, and Differential Phase Contrast (DPC)

In Schlieren imaging with coherent illumination, a half plane mask (a Foucault knife edge) is inserted to eliminate say the negative spatial frequencies in x. Taking now $c(0,0) = 1/2$, eqn [8] becomes, after normalization:

$$I(x,y) = 1 + 4\Re\left\{\int_{-\infty}^{\infty}\int_{0}^{\infty} c(m,n)B(m,n) \times \exp\left[\frac{\mathrm{i}2\pi}{M}(mx+ny)\right]\mathrm{d}m\,\mathrm{d}n\right\} \qquad [21]$$

For the weak phase object eqn [15], we then have:

$$I(x,y) = 1 - dc(\nu)\sin\left(\frac{2\pi\nu x}{M}\right) \qquad [22]$$

The image exhibits a variation in phase quadrature to the original phase variation, that is, it is related to the x-derivative of the phase. The effect occurs because the coherent transfer function contains an odd (but real) part.

For a partially coherent microscope a similar effect can be achieved simply by offsetting the illumination system, but now the transfer function is no longer single-sided. It can be resolved into odd and even parts that result in differential phase contrast and amplitude contrast, respectively. The amplitude contrast component merely contributes to the background for a phase-only object, and can be removed simply by contrast enhancement of a digital image. The offset can be introduced in practice by inserting a half-plane mask into the condenser aperture stop. Imaging of a slowly varying object can be explained by eqn [14], in which the transfer function is not symmetrical. The image shows a pseudo-three-dimensional bas relief effect. A commercial implementation of an asymmetric transfer function is Hoffman modulation contrast, which uses offset

illumination together with an amplitude mask in the objective pupil (Figure 3). Advantages of Hoffman contrast are that it uses standard bright-field objectives and avoids problems from birefringent specimens.

A similar method can be used in scanning laser microscopy. In a scanning microscope the effect of the illumination and detection systems are interchanged. Thus DPC can be achieved in a scanning laser microscope by offsetting the detector. In practice this is conveniently done by using a split detector (Figure 4), a detector divided into two semicircular elements. If the signal of one half is subtracted from that of the other, the amplitude contrast component cancels out, so that pure differential phase contrast results. For weak phase gradients, the intensity varies linearly with phase gradient, again showing the bas relief effect. It is thus bi-directional, and the difference signal can become negative. A constant value is, therefore, usually added to it before display. The sum of the two detector elements results in a conventional bright-field image. From eqn [14], both difference and sum signals are proportional to $a^2(x, y)$, so that if the ratio of these two signals is extracted, an image of the phase gradient in the x direction results. The response to either weak phase gradients or fine detail can be enhanced by modifications using annular split detectors. Use of a quadrant detector allows the phase gradients in the x and y directions to be imaged. DPC microscopy has been demonstrated to be sensitive enough to observe monomolecular films and atomic surface steps.

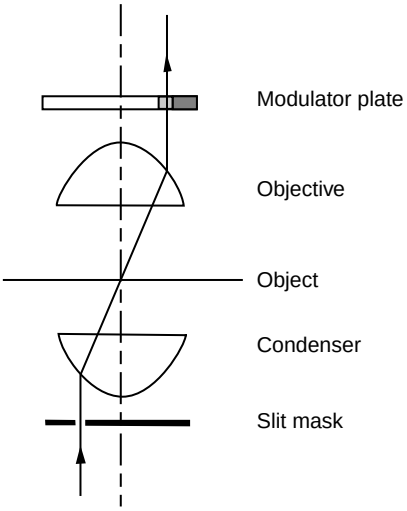


Figure 3 Schematic diagram of Hoffmann modulation contrast microscopy.

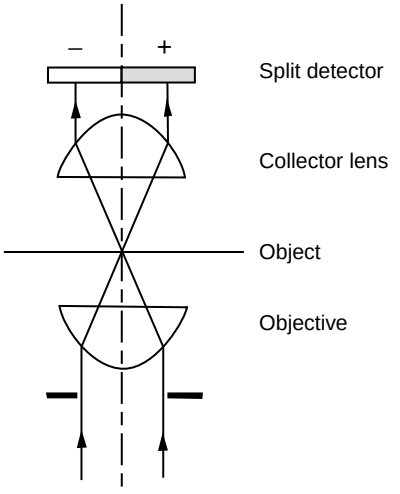


Figure 4 Schematic diagram of differential phase contrast (DPC) microscopy.

Interference Microscopy

Interference microscopy allows quantitative measurement of the object phase by interference of the object beam with a reference beam. In the transmission geometry, a Mach–Zehnder interferometer is used, which is effectively two parallel microscopes, each with matched condenser and objective lenses. In the reflection geometry we can use a Michelson, a Mirau, or a Linnik interferometer. The Michelson interferometer has an inclined beamsplitter and the reference mirror situated between the objective and object, and hence can be used only with low numerical aperture objectives. The Linnik interferometer has matched objectives in the object and reference beam paths. There is no limit to the numerical aperture that can be used, but it is not a common-path interferometer and is hence sensitive to vibrations and air currents. Probably the most practicable system for interference microscopy in the reflection geometry is based on the Mirau interferometer that uses a beamsplitter with its normal parallel to the optic axis. Until recently, interference microscopes used a small condenser aperture giving nearly coherent illumination. The image consists of three components: an object beam term (noninterference image) a reference beam term, and the interference term. The phase (and amplitude) of the object beam can be extracted using phase stepping or heterodyning techniques. The measured phase is wrapped, and requires unwrapping to obtain the absolute phase.

If the condenser aperture is opened up, the behavior is modified. The interference term is

given by

$$I_{\text{int}}(x,y) = 2\Re\left\{ r^* \int\int c_{\text{int}}(m,n)T(m,n) \times \exp\left[\frac{\text{i}2\pi(mx+ny)}{M}\right] \text{d}m\,\text{d}n \right\} \quad [23]$$

where $c_{\text{int}}(m,n)$ is the coherent transfer function for interference imaging, given by a convolution integral of the object path objective pupil with the product of the reference pupil and the source. The first thing to note is that for a high numerical aperture, the spacing of the interference fringes is increased by 20–40%, and depends on defocus, tilt, and central obscuration. Thus accurate measurement of phase requires careful calibration. The objective pupil exhibits a phase variation when defocused that reduces the modulus of the coherent transfer function, resulting in the fringes being modulated by an envelope the width of which becomes smaller as the aperture becomes larger. The envelope is similar to that of the optical sectioning effect in confocal microscopy. This envelope allows the zero-order fringe to be identified, although simulations have suggested that fringe hopping could occur at phase jumps as a result of diffraction effects.

Multiple Beam Interferometry

The topography of a surface can be investigated by using a Fizeau interferometer formed by bringing a coated optical flat near to the surface. Then if the surface and flat have high reflectivity, multiple beam interference occurs and the width of the bright fringes can be much reduced. The system can be used with a tilted flat, to give fringes that are substantially straight, or with an almost parallel flat, so that the entire field of view is occupied by a single fringe. In the latter case, the sensitivity can be in the subnanometer range. Multiple beam interference also enhances the contrast of weak amplitude information.

Shearing Interferometry

In conventional interferometry, the object beam is arranged to interfere with a constant reference beam. In shearing interferometry, on the other hand, it interferes with a shifted version of itself.

Lateral Shearing Interferometry

An approach that is suitable for quantitative optical path measurements is to use an interferometer that shears the microscope image. The Interphako system uses a Mach–Zehnder interferometer that can measure optical path differences as low as ~3 nm in either the reflection or transmission geometry. The interferometer allows the shear distance to be varied.

Differential Interference Microscopy (DIC)

The most common form of shearing interferometry is differential interference microscopy, usually associated with the name of Nomarski. DIC microscopy can be performed in either the transmission or the reflection geometry. A plane polarized illumination beam is split into two orthogonally polarized beams traveling at slightly differing angles by a Wollaston prism in the front focal plane of the condenser lens (Figure 5). The beams are thus sheared laterally in the object plane. The light from the sample travels through the objective and a second Wollaston prism, which recombines the two beams into one. In the reflection geometry, a single lens and Wollaston prism are used. An analyzer set at 45° allows the two beams to interfere. The shear is arranged to be smaller than the width of the point spread function. Different prisms are supplied for use with different magnification objectives. The system results in

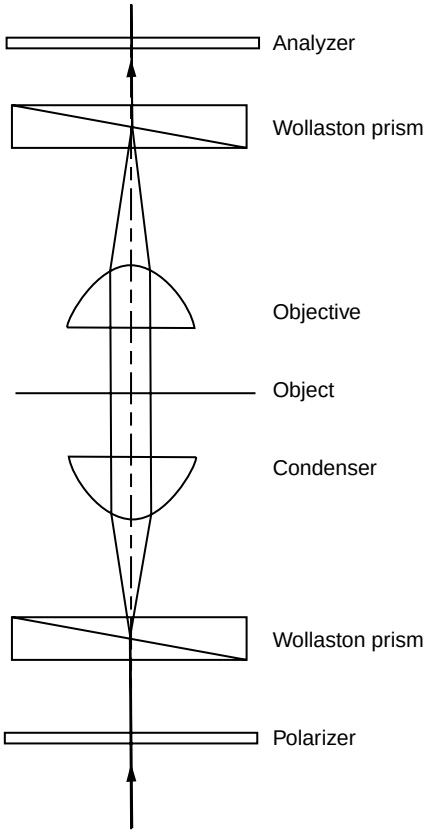


Figure 5 Nomarski differential interference contrast (DIC) microscopy.

differential phase contrast, but there are differences as compared with the DPC microscope described above.

There are two different implementations of DIC in use. In one implementation, the illumination polarizer and analyzer are crossed. The relative phase difference between the two beams (the bias retardation) is adjusted by moving the Wollaston prism laterally. It is claimed that rotation of the analyzer can be useful when observing mixed phase/amplitude objects. In the second arrangement (the Sénarmont compensator) the position of the Wollaston prism is fixed, and the analyzer rotated to change the relative phase.

A recent variation of DIC uses circularly polarized illumination. The Wollaston prism still produces two plane polarized components, but with an additional phase difference of $\pi/2$. The advantage of using circular polarization is that the prism can then be rotated to change the direction of shear.

If the shear is 2Δ, and the phase difference is $\pi - 2\phi_0$ (i.e., if $\phi_0 = 0$ then the two beams are in anti-phase), then the effective object spectrum is:

$$T_{\text{eff}}(m,n) = -2\sin(2\pi m\Delta - \phi_0)T(m,n) \qquad [24]$$

Thus the effective transfer function can be written:

$$\begin{aligned} C_{\text{eff}}(m,n;p,q) &= 4\sin(2\pi m\Delta - \phi_0)\sin(2\pi p\Delta - \phi_0) \\ &\quad \times C(m,n;p,q) \\ &= 2\{\cos[2\pi(m-p)\Delta] \\ &\quad - \cos 2\phi_0 \cos[2\pi(m+p)\Delta] \\ &\quad - \sin 2\phi_0 \sin[2\pi(m+p)\Delta]\} \\ &\quad \times C(m,n;p,q) \end{aligned} \qquad [25]$$

It is the final term of the three in braces that, because it is an odd function, results in differential phase contrast. We can now examine imaging of special forms of object. For a weak object, putting $p = 0$ as in eqn [11]:

$$\begin{aligned} C_{\text{eff}}(m,n;0,0) &= 2\{(1 - \cos 2\phi_0)\cos[2\pi m\Delta] \\ &\quad - \sin 2\phi_0 \sin[2\pi m\Delta]\}C(m,n;0,0) \end{aligned} \qquad [26]$$

The first term in braces corresponds to an amplitude contrast background component because it is real and even. The second term results in DPC, because it is real and odd. For zero bias retardation, dark-field imaging results. As the bias retardation is increased from zero, the strength of both the amplitude contrast and the DPC components increase, but that of the DPC component increases more rapidly. For visual observation we require good contrast, and hence we choose a small value of bias retardation, just large enough so that the background is large enough to avoid artifacts in the DPC image. In digital microscopy, contrast is not so important as it can be enhanced digitally, and hence we can increase ϕ_0 to a value of $\pi/4$, which maximizes the strength of the DPC component. Further increase in bias retardation decreases the DPC signal, until for $\pi/2$ pure amplitude contrast results. The actual form of the weak object transfer function depends on the magnitude of the shear. The condition $m_0\Delta = 1/4$, where m_0 is the cut-off in spatial frequency, gives optimum performance for the imaging of weak objects.

The phase gradient transfer function is:

$$C_{\text{eff}}(m,n;m,n) = 4\sin^2(2\pi m\Delta - \phi_0)C(m,n;m,n) \qquad [27]$$

For the bright-field microscope the phase gradient transfer function falls off symmetrically from its value at zero phase gradient. The effect of the pre-multiplying factor in eqn [27] is to increase the fall-off for positive phase gradients and reduce it for negative ones. A bas-relief effect occurs, with highlighting for negative phase gradients for some values of the parameters. Increasing the value of the shear increases the highlighting effect, but, unlike in DPC, the behavior is not anti-symmetric.

Because interference is dependent on the wavelength of the light, useful and beautiful effects can be seen with DIC microscopy using white light. Usually it is regarded as a qualitative, rather than a quantitative, imaging method. However, as from eqn [14], the intensity can be written:

$$I(x,y) = 2a^2(x,y)[1 - \cos(2\pi m\Delta - \phi_0)]C(m,n;m,n) \qquad [28]$$

By recording images at different bias retardations, conventional phase stepping methods can be used to extract the phase gradient $2\pi m = \partial\phi/\partial x$.

Axial Shearing Interferometry

Instead of shearing in the transverse direction, a shear in the axial direction can be used. In the reflection geometry, the complex amplitude from the point of observation is compared with the average over a defocused spot. This method has proved to be extremely sensitive, and can detect height changes smaller than 0.1 nm.

Phase Retrieval

Digital Phase Retrieval

The phase of a single two-dimensional (2D) measurement of intensity can be retrieved by an iterative

procedure, if constraints such as nonnegativity and band-limitedness can be assumed. If the modulus of both an object and its Fourier transform are known, then the phases can also be recovered. In this case, the two 2D real data sets have information content sufficient to reconstruct the 2D complex object. This approach has been extended to the case of measurement of two defocused images, using the method of phase diversity.

Imaging Using the Transport of Intensity Equation

The transport equation of intensity can be used to recover the phase of a partially coherent wavefield directly (i.e., noniteratively) from two 2D intensity measurements. If the wavefield is represented in terms of its intensity I and phase ϕ in a plane, the transport equation can be written in the paraxial approximation and in the absence of phase singularities:

$$-k\frac{\partial I}{\partial z} = \nabla_T \cdot [I\nabla_T \phi] \qquad [29]$$

where ∇_T is a 2D operator acting in the plane. This expression can be inverted to give ϕ quantitatively, if I and $\partial I/\partial z$ are known. However, it gives the phase of the image field, rather than the phase of the object, so that obtaining quantitative values for a strong object may be difficult. The transport of intensity method is valid with partially coherent illumination: if the relative condenser aperture is too great then contrast becomes weak. Further, an effect similar to the variation in fringe spacing with aperture in interference microscopy is introduced because oblique illuminating rays traverse the sample at an angle to the axis. In practice, a series of three images are usually taken at different focus positions. Then, from the quantitative phase data, various different contrast mechanisms such as DIC, Zernike phase contrast, and so on, can be simulated.

Surface Profiling Methods: Triangulation and Coherence Domain Methods

Although strictly they are not phase imaging methods, other surface profiling methods can be used to measure optical path. These could also be adapted to measure optical path in transmission. It is important also to be aware of these methods, as they could affect measurements made using other phase contrast techniques.

Apart from the phase measurement techniques described above, surface profiling methods fall into one of two categories: triangulation and coherence domain methods. An example of a technique that relies on triangulation is confocal microscopy. Confocal microscopy exhibits an optical sectioning effect similar to that described for interference microscopy. The effect is stronger for higher numerical aperture. As a reflector is scanned through focus the intensity goes through a maximum. By locating the position of the maximum, the local surface height can be obtained. In a similar way, in a confocal transmission system, introduction of an object with some optical thickness has the effect of altering the axial position of the confocal pinhole, with a resultant drop in intensity. This effect can be used to image the optical thickness of the sample.

Coherence domain methods rely on the finite coherence length of the light. In an interferometer, this results in an envelope that multiplies that arising from the nonzero relative condenser aperture. Measurement of the peak of the envelope allows local surface height to be measured. Unlike the confocal approach, the width of this coherence envelope is independent of the numerical aperture of the system.

In reflection, both the confocal and the coherence domain methods have the advantage over direct phase imaging techniques that they allow measurement of surface height changes without errors caused by the phase change on reflection from a conducting medium.

See also

Interferometry: Phase Measurement Interferometry. **Microscopy:** Interference Microscopy. **Phase Control:** Phase Conjugation and Image Correction.

Further Reading

Bennett AH, Jupnik H, Osterberg H and Richards OW (1952) *Phase Microscopy*. New York: Wiley.

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge: Cambridge University Press.

Bradbury S and Evenett PJ (1996) *Contrast Techniques in Light Microscopy*. Oxford: Bios Scientific Publishers.

Françon M (1950) *Le Contraste de Phase en Optique et en Microscopie*. Paris: Revue d'Optique.

Françon M (ed.) (1952) *Le Contraste de Phase et le Contraste par Interférences*. Paris: Revue d'Optique.

Inoue S and Spring KR (1997) *Video Microscopy: The Fundamentals*, 2nd edn. New York: Plenum.

Lessor DL, Hartman JS and Gordon RL (1979) Quantitative surface topography determination by Nomarski reflection microscopy. *Journal of the Optical Society of America* 69: 357–366.

Pluta M (1989) *Advanced Light Microscopy*. Specialized Methods, vol. 2. Amsterdam: Elsevier.
Sheppard CJR and Wilson T (1980) Fourier imaging of phase information in conventional and scanning microscopes. *Philosophical Transactions of Royal Society A* 295: 513–536.
Wilson T and Sheppard CJR (1984) *Theory and Practice of Scanning Optical Microscopy*. London: Academic Press.

MODULATORS

Contents

Acousto-Optics
Electro-Optics
Modulation and Demodulation of Optical Signals

Acousto-Optics

M Gottlieb and D Suhre, Carnegie Mellon University, Pittsburgh, PA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Acousto-optics deals with the interaction of light waves with sound waves and has given rise to a large number of devices related to various laser systems for display, information handling, optical signal processing, and numerous other applications requiring the spatial or temporal modulation of coherent light. The basic principles and phenomena controlling these interactions were largely understood by the mid-1930s, but had little practical significance because very high acoustic power levels were required to attain good optical efficiency and there were few good sources of well-collimated monochromatic light. During the period from 1960 to the present, several key technologies were developing rapidly, at the same time that many applications of the laser were being suggested which require high-speed, high-resolution scanning methods. These new technologies gave rise to high-efficiency, wideband acoustic transducers capable of operation to several gigahertz, high-power wideband solid-state amplifiers to drive such transducers, and the development of a number of new, synthetic acousto-optic crystals with low-drive-power requirements and low acoustic losses at high frequencies. This combination of properties makes acousto-optics feasible for many systems, and for several is the method of choice to satisfy demanding requirements. This chapter will summarize the basic features of acousto-optic interactions and the operating principles of the most common acousto-optic devices.

The Photo-Elastic Effect

The change induced in the refractive index of a transparent material by the pressure change produced by an acoustic disturbance is the underlying mechanism of all acousto-optic interactions. An acoustic wave produces periodic regions of compression and rarefaction in the material, which modulates the density. The Lorentz–Lorenz relation relates the refractive index to the density, for the simplest case of an ideal gas

$$\frac{(n^2-1)}{(n^2+2)} \propto \rho \qquad [1]$$

where n is the refractive index and ρ is the density. This relation is followed to a good approximation for most simple solid materials as well. The elastooptic coefficient is

$$\frac{\rho\, \mathrm{d}n}{\mathrm{d}\rho} = \frac{(n^2-1)/(n^2+2)}{6n} = p \qquad [2]$$

The fundamental quantity given by eqn [2], also known as the photo-elastic constant p, is related to the pressure applied by

$$p = \frac{1}{\beta}\left(\frac{\mathrm{d}n}{\mathrm{d}p}\right) \qquad [3]$$

where P is the applied pressure and β is the compressibility of the material. The photo-elastic

constant of an ideal material with refractive index of 1.5 is 0.59. The photo-elastic constants of a wide variety of materials lie in the range from about 0.1 to 0.6.

The relation in eqn [3] follows from the usual definition of the photo-elastic constant:

$$\Delta(1/\varepsilon) = \Delta(1/n^2) = pe \qquad [4]$$

where ε is the dielectric constant, $\varepsilon = n^2$, and e is the acoustic strain amplitude. The induced change in refractive index, Δn, is

$$\Delta n = -n^3 pe \qquad [5]$$

Strain amplitudes typically lie in the range of 10^{-8} to 10^{-5}, with Δn in the range of about 10^{-8} to 10^{-5} (for $n = 1.5$). Devices based upon such a small change in refractive index are capable of generating large effects because these devices are configured in a way that can produce large phase changes at optical wavelengths.

The more complete relation defining the photo-elastic interaction is more complex than the simple scalar eqn [5] in which the photoelastic constant is independent of the directional properties of the material. In fact, even for an isotropic material such as glass, longitudinal acoustic waves and transverse (shear) acoustic waves cause the photoelastic interaction to assume different parameters. A tensor relation between the dielectric properties, the elastic strain, and the photo-elastic coefficient describes the interaction, particularly for anisotropic materials. The tensor equation is

$$\Delta(1/n^2) = \sum_{kl} p_{ijkl}\, e_{kl} = \Delta(1/\varepsilon)_{ij} \qquad [6]$$

where $(1/\varepsilon)_{ij}$ is a component of the optical index ellipsoid, e_{kl} are the Cartesian strain components, and p_{ijkl} are the components of the photo-elastic tensor. The crystal symmetry of the photoelastic material places limits on the possible configurations of interaction geometry.

Diffraction by Acoustic Waves

For typical acousto-optic devices the acoustic wave acts as a diffraction grating, made up of periodic changes in optical phase, moving at sound velocity. These features determine the properties of acousto-optic diffraction. Using a quantum-based model, the light and sound may be thought of as particles, photons and phonons, undergoing collisions in which energy and momentum are conserved. Either of these descriptions may be used to obtain all the important diffraction effects, but some are more easily understood on the basis of one or the other. A description of both is given here. Consider Figure 1 in which the light wave, of frequency ω and wavelength λ, is incident into the material with an acoustic wave of frequency f and wavelength Λ. If the refractive index of the medium is $n + \Delta n$ in the presence of the acoustic wave, the phase of the optical wave will be changed by an amount

$$\Delta\phi = 2\pi(L/\lambda)\Delta n \qquad [7]$$

where L is the length of the interaction region. Some typical values of $\Delta\phi$ can be obtained by assuming $L = 2.5$ cm and $\lambda = 0.5$ μm, with Δn reaching a peak value of 10^{-5}. This yields a phase change of π rad, which is, of course, quite large. It is large because L/λ, the number of optical wavelengths, is 50 000, so that a very small Δn can still produce a sizable $\Delta\phi$. If the electric field incident on the delay line is represented by

$$E = E_0\, e^{i\omega t} \qquad [8]$$

then the field of the phase-modulated emerging light will be

$$E = E_0\, e^{(i\omega t + \Delta\phi)} = e^{i\omega t}\, e^{2\pi i(L/\lambda)\sin(ft)} \qquad [9]$$

where f is the acoustic frequency.

A detailed derivation of the resulting temporal and spatial distribution of the light field is mathematically complex, but, analogy with radio-wave modulation can be used to arrive at the resultant fields. The spectrum of a phase-modulated carrier of frequency ω consists of components separated by multiples of the modulation frequency f, as shown in Figure 2. Sidebands are produced about the carrier frequency, such that the frequency of the mth sideband is

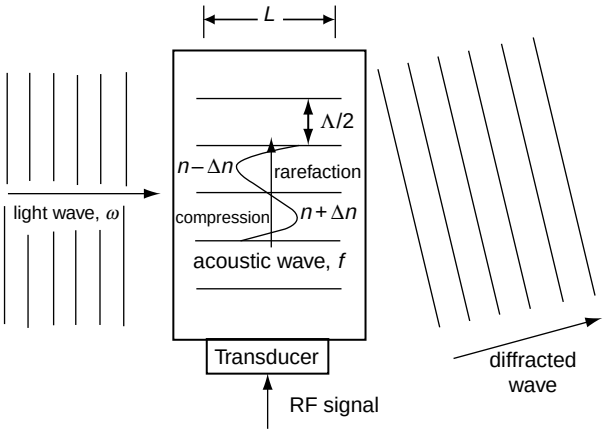


Figure 1 Interaction of light waves with acoustic waves.

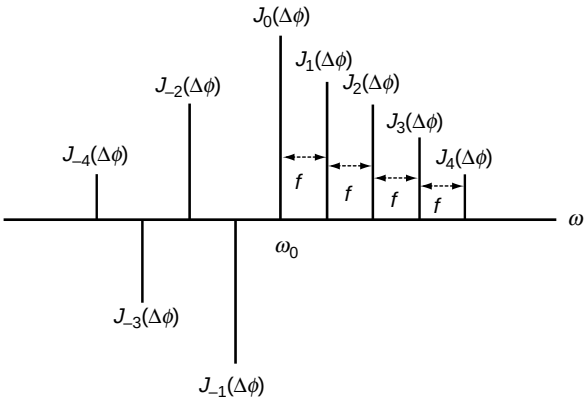


Figure 2 Spectrum of a phase-modulated wave of carrier frequency ω_0, and modulation index $\Delta\phi$.

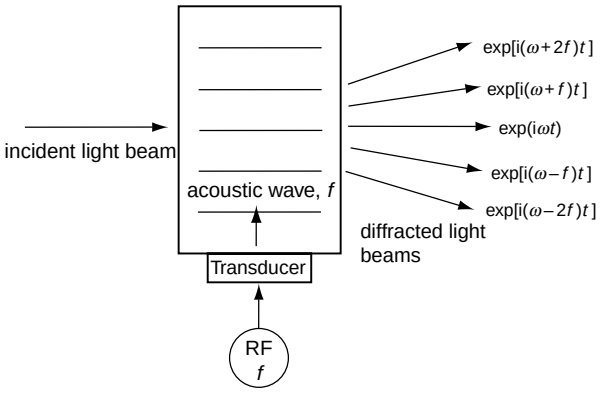


Figure 3 Diffraction of light in the Raman–Nath limit.

$\omega + mf$, where m may be positive or negative. The amplitude of each of the sidebands is proportional to the Bessel function of order equal to the sideband number, and whose argument is the modulation index $\Delta\phi$. The odd-numbered negative orders are 180° out of phase with the even-numbered ones, a feature that may be useful for certain signal processing applications. The light emerging from the delay line is composed of a number of light waves whose frequencies have been shifted by mf from the frequency ω of the incident light. The relative amplitudes will be determined by the peak change in the refractive index. The two extremes of the diffraction process are simplified in the 'thin grating' and the 'thick grating' models. The ratio of the interaction length, L, to the acoustic wavelength, Λ, will determine the character of the diffraction process. The plane wave approximation in which the acoustic energy propagates as a plane wave is valid when this ratio is very large. However, when this ratio is small, the acoustic propagation can be described in terms of a sum of plane waves, the angular spectrum of which increases as the ratio decreases. In this phenomenological model the partial wave which is propagating at an angle λ/Λ to the forward direction may diffract the light second time into an angle $2\theta = 2\lambda/\Lambda$, and the frequency of this light will once again be up-shifted, for a total frequency shift of $2f$. If the spectrum of acoustic waves contains sufficient power of still higher orders, then this rediffraction process can be repeated, so that light will be multiply diffracted m times into higher-order angles, $m\theta = m\lambda/\Lambda$ each with a frequency shift mf. A similar argument holds for the negative orders, so that a complete set of diffracted light beams will appear as shown in Figure 3, where the deflection angle corresponding to the mth order is given by $\sin\theta_m = m\lambda/\Lambda$ and the frequency of the light deflected into the mth order is $\omega_m = \omega + mf$. The intensity of the carrier wave, or zero order, will be zero when the modulation index $\Delta\phi$ is equal to 2.4. The first order will be a maximum for $\Delta\phi = 1.8$, corresponding to the maximum of the first-order Bessel function B_1, and decreasing for larger modulation indices. These phenomena were described by Debye and Sears and so are often referred to as Debye–Sears diffraction. An extensive theoretical analysis of the effect was given by Raman and Nath and is alternatively referred to as Raman–Nath diffraction. As the interaction length is increased the Raman–Nath diffraction gradually weakens. The weakening begins around an interaction length $L \sim \Lambda^2/4\lambda$. This value of L is expressed in the Q parameter

$$Q = 4L\lambda/\Lambda^2 \qquad [10]$$

which is known as the Raman–Nath parameter. A different regime of diffraction takes effect for values of the interaction length $Q \gg 1$.

The phenomena in this regime can be more easily understood in terms of the quantum description of the light and sound waves as collisions between photons and phonons. In this model, the dynamics of the collisions of the light and sound are governed by the laws of conservation of energy and momentum. The momentum vectors magnitudes of the light and sound, $\mathbf{k}$ and $\mathbf{K}$, are given by the well-known expressions

$$|\mathbf{k}| = 2\pi n/\lambda \text{ and } |\mathbf{K}| = 2\pi/\Lambda \qquad [11]$$

where the acoustic wavelength, Λ, is related to the acoustic velocity, V, by $\Lambda = V/f$.

Conservation of momentum is expressed by the vector relation

$$\mathbf{k}_\mathrm{i} + \mathbf{K} = \mathbf{k}_\mathrm{d} \qquad [12]$$

the diagram for which is shown in Figure 3, where $\mathbf{k}_i$, and $\mathbf{k}_d$ represent the momentum of the incident photon and the diffracted photon, respectively. In this notation

$$\mathbf{k}_i = 2\pi n_i/\lambda, \qquad [13a]$$

and

$$\mathbf{k}_d = 2\pi n_d/\lambda \qquad [13b]$$

If the material is anisotropic (birefringent), n_i and n_d may be different.

Conservation of energy requires that

$$h\omega_d = h\omega_i + hf \qquad [14]$$

in which h is Planck's constant. Since ω_i lies in the optical frequency range, $\sim 10^{13}$ Hz, and f lies in the RF or microwave range, 10^6–10^9 Hz, then $\omega_d \sim \omega_i$. This results, for isotropic materials, in the magnitudes of $\mathbf{k}_i$ and $\mathbf{k}_d$ being equal and resulting in the isosceles momentum triangle of Figure 4. The angles of incidence and diffraction (with respect to the planar acoustic wavefronts), called the Bragg angles, are equal for this case, and are

$$\theta_B = (1/2)\lambda/\Lambda \qquad [15]$$

A schematic diagram for this process is shown in Figure 5. The interaction will be large only if the light is well-collimated, and incident at this angle. Unlike the Debye–Sears regime, there is only a single diffracted beam. The energy of the phonon may either be added to that of the incident photon, increasing its frequency to $\omega_d = \omega_i + f$, or the reverse, resulting in $\omega_d = \omega_i - f$. The sense of the momentum vectors determines which of these occurs. The Debye–Sears

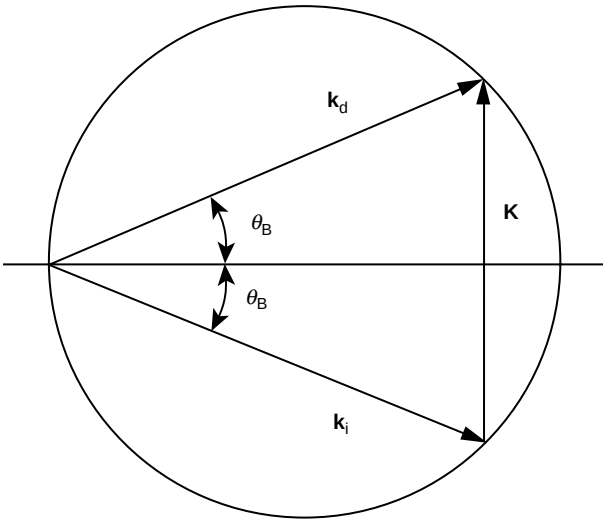


Figure 4 Momentum diagram for diffraction of light.

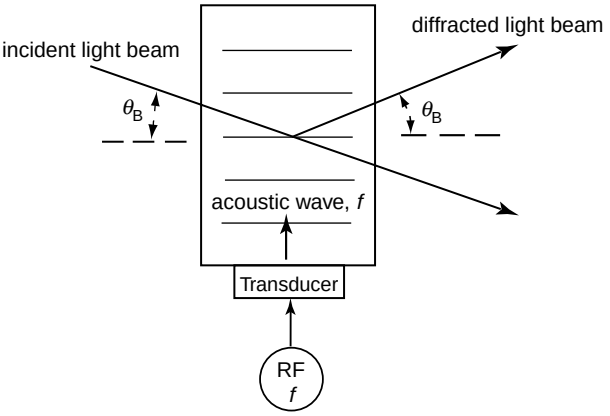


Figure 5 Acousto-optic diffraction cell.

effect and Bragg diffraction are not different phenomena, but are the limits of the same mechanism. The Raman–Nath parameter Q determines which is the appropriate limit for a given set of values λ, Λ, and L.

Anisotropic Bragg Diffraction

In optically anisotropic materials, acousto-optic Bragg diffraction can occur between an ordinary and extraordinary optical beam, and vice versa. This is generally referred to as birefringent diffraction, and it offers additional design capabilities compared to the isotropic case, such as enhancing the angular aperture, and extending the aperture–bandwidth product.

The theory of diffraction of light so far described has assumed that the optical medium is isotropic. A number of important acousto-optic devices make use of the properties of birefringent materials. It is different from diffraction in isotropic media in that the magnitude of the momentum vector of the light will in general be different for different light polarization directions. Thus, the vector diagram representing conservation of momentum will no longer be the simple isosceles triangle of Figure 4. The momentum vector for ordinary polarized light will, in general, be different from the momentum vector for light that is extraordinary polarized. To understand the effect of anisotropy on diffraction, it is necessary to mention another phenomenon which occurs when light interacts with shear acoustic waves, i.e., waves in which the displacement of matter is perpendicular to the direction of propagation of the acoustic wave. A shear acoustic wave may cause the direction of polarization of the diffracted light to be rotated by 90°. The underlying reason for this is that the shear

disturbance induces a birefringence which causes the plane of polarization to be rotated. The acousto-optic tunable filter (AOTF) is a particularly interesting birefringent device. It was first developed for the case of collinear optical and acoustic beams, and was used for wavelength selection in a dye laser.

Diffraction Efficiency

The acoustic power required to diffract light will be determined by the geometric parameters as well as by the properties of the medium. A simplified calculation leads to results that are useful. Referring to the spectrum of a phase-modulated wave shown in Figure 2, we can see that the ratio of the intensity in the first order to that in the zero order is

$$I_1/I_0 = \left[B_1(\Delta\phi)/B_0(\Delta\phi)\right]^2 \qquad [16]$$

By analogy this same result comes about for acousto-optically diffracted light. The acoustic power flow is given by

$$P_a = \frac{1}{2} cVe^2 \qquad [17]$$

where c is the elastic stiffness constant. The elastic stiffness constant is related to the bulk modulus β, the density ρ, and acoustic velocity v through the expression

$$c = \frac{1}{\beta} = \rho V^2 \qquad [18]$$

and the acoustic power density is

$$P_a = \frac{1}{2}\rho V^3 e^2 \qquad [19]$$

The optical phase modulation of the medium resulting from the change in refractive index, Δn, is

$$\Delta\phi = 2\pi(L/\lambda)\Delta n \qquad [20]$$

Using eqn [5] for Δn, the phase modulation is related to the acoustic power density by

$$\Delta\phi = -\pi(L/\lambda)\, n^3 p\, (2P_a/\rho V^3)^{1/2} \qquad [21]$$

The diffraction efficiency can now be calculated by using this value for the optical phase change in eqn [16]. The first- and second-order Bessel functions can be approximated for small modulation index by

$$\begin{aligned} &B_0(\Delta\phi) \sim \cos(\Delta\phi) \sim 1 - (\Delta\phi)^2, \quad \text{and} \\ &B_1(\Delta\phi) \sim \sin(\Delta\phi) \sim \Delta\phi \end{aligned} \qquad [22]$$

The small signal approximation to the diffraction efficiency is then

$$I_1/I_0 \sim (\Delta\phi)^2 = \left(\pi^2 L^2/2\lambda^2\right)\left(n^6 p^2/\rho V^3\right) P_a \qquad [23]$$

The expression $n^6 p^2/\rho V^3$ is known as the figure of merit of the material, and is designated as M_2; it is comprised entirely of intrinsic material properties.

Acousto-optic Materials

Acousto-optic device technology has matured to the point that performance is chiefly limited by material parameters, particularly the figure of merit and acoustic attenuation. Nature has arranged that materials with high figures of merit usually have high attenuation and vice versa. The widely used acousto-optic materials are fused quartz, tellurium dioxide, and lithium niobate. Development work on new infrared materials has been reported recently. A list of commonly used acousto-optic materials is given in Table 1. For materials with a low figure of merit, a higher drive power can be used to obtain the required efficiency.

Experience has indicated that the upper limit for very small devices (active area $\sim 0.1\ \text{mm}^2$) is a drive power density of 100–500 mW/mm^2, provided there is proper heat sinking to transfer the heat energy. For larger devices, sizes greater than $\sim 1\ \text{cm}^2$, the limit is closer to a few W/cm^2. At the high drive power levels, the acoustic attenuation may cause significant optical distortion.

AO Devices

Resolution, bandwidth, and speed are the important characteristics of acousto-optic scanners, shared by all types of scanning devices. Depending upon the application, only one, or all, of these characteristics may have to be optimized. Consider the acousto-optic scanner in Figure 6 with a collimated incident beam of width D, diffracted to an angle θ_B at its RF bandcenter, and whose bandwidth is Δf. If the diffracted beam is focused onto a plane by a lens, or lens combination, at the scanner, the diffraction spread of the optical beam will be

$$\Delta x = F\Delta\phi \sim F - \lambda/D \qquad [24]$$

Table 1 Selected acousto-optic materials

Material	*Transmission range* (μm) mission	*Acoustic mode & propagation direction*	*v* (*cm/s* × 10^5)	*Acoustic attenuation* (*dB/cm GHz*^{2})[c]	*n*	M_2 ($s^3/g \times 10^{-18}$)
Visible–near-infrared (VIS-NIR)						
$LiNb0_3$	0.04–4.5	L[100][a]	6.57	0.15	2.20	7.0
		S[001][b]	3.59	2.6	2.29	2.92
TiO_2,	0.45–6	L[001]	10.3	0.55	2.58	1.52
α-SiO_2;	0.12–4.5	L[001]	6.32	2.1	1.54	1.48
		L[100]	5.72	3.0	1.55	2.38
TeO_2.	0.35–5	L[001]	4.20	15	2.26	34.5
		S[110]	0.616	90	2.26	793
Far IR						
Ge	2–20	L[111]	5.50	30	4.00	840
$T1_3AsS_4$ (TAS)	0.6–12	L[001]	2.5	29	2.63	510
GaAs	1–11	L[110]	5.15	30	3.37	104
ZnTe	0.55–20	L[110]	3.37	130	2.77	18
GaP	0.6–10	L[110]	6.32	60	3.31	30
$T1_3PSe_4$	0.85–9	L[100]	2.0	150	2.9	2069
Te	5–20	L[100]	2.2	60	4.8	4400
CdS	0.5–11	L[100]	4.17	90	2.44	12
GaP	0.6–10	L[110]	6.32	60	3.31	30
ZnS	0.4–12	L[001]	5.82	27	2.35	3.4
		S[001]	2.63	130	2.35	8.4

[a]Longitudinal acoustic waves propagating along the [100] crystallographic direction.
[b]Shear acoustic waves propagating along the [001] crystallographic direction.
[c]Attenuation constant at 1 GHz; the frequency dependence of the attenuation for most crystals is quadratic.

where F is the focal length of the lens. The light intensity will be distributed in the focal plane with a sinc^2 distribution by aperture diffraction. The number of resolvable spots, N, will be the angular scan range divided by the angular diffraction spread,

$$N = \frac{\Delta\theta}{\Delta\phi} \qquad [25]$$

where $\Delta\theta$ is the range of the angular scan. Differentiating the Bragg angle formula yields

$$\Delta\theta = (\lambda/V \cos\theta_0)\Delta f \qquad [26]$$

and

$$N = \Delta f(D/V \cos\theta_0) = \Delta f \tau \qquad [27]$$

where θ_0 is the Bragg angle at band center, and τ is the time that it takes the acoustic wave to cross the optical aperture. The resulting expression is the time–bandwidth product of the acousto-optic scanner, a concept applied to a variety of electronic devices as a measure of information handling capacity. The time–bandwidth product of an acousto-optic Bragg cell is equivalent to the number of bits of information which may be instantaneously processed by the system. In order to maximize the number of resolution elements, it is desirable to have as large a bandwidth as possible (i.e., large frequency range) and also as large an aperture delay time as possible. There are two factors limiting the bandwidth of an acousto-optic device: the bandwidth of the transducer structure (discussed later), and acoustic absorption in the delay medium. The acoustic absorption increases with increasing frequency; for high-purity single crystals the increase generally goes with the square of the frequency. For glassy materials, on the other hand, the attenuation will increase more slowly with frequency, often approaching a linear function. The maximum frequency is generally taken as that for which the attenuation of the acoustic wave across the optical aperture is equal to

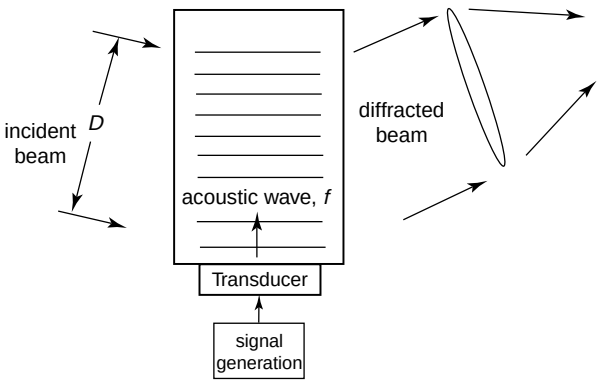


Figure 6 Acousto-optic scanner.

3 dB. A reasonable approximation of the maximum attainable bandwidth is

$$\Delta f = 0.7 f_{\max} \qquad [28]$$

Birefringent Scanners

The birefringent scanner is a significant use for anisotropic Bragg diffraction. There are a number of advantages with the birefringent scanner over the anisotropic scanner, such as a larger angular scan range along with lower frequencies. There are also disadvantages, such as lower speed due to generally lower shear wave velocities, and the particular application will dictate whether an isotropic or birefringent scanner is better. Scanners represent a fairly important aspect of acousto-optic devices due to the widespread commercial use of laser beam deflectors for displays and laser printers. For applications where the required speed is beyond that of mechanical scanners, the acousto-optic scanner is an ideal candidate. However, unlike mirrors, acousto-optic deflectors are wavelength sensitive, and can only be used with single-wavelength laser beams.

The birefringent scanner can be described with the wavevector diagram, as shown in Figure 7 for a positive uniaxial crystal where the extraordinary index of refraction is larger than the ordinary. The acoustic wavevector is tangent to the diffracted surface, which produces the largest scan angle for a given acoustic bandwidth. This is also the degenerate case, where only a single diffracted beam results, whereas two beams at two different acoustic frequencies will generally result from arbitrary input and acoustic propagation directions. The azimuthal acceptance angles (angles normal to the plane of incidence) of the acoustic and optical wavevectors can be different, although propagation and polarization directions must be selected that will produce high efficiency.

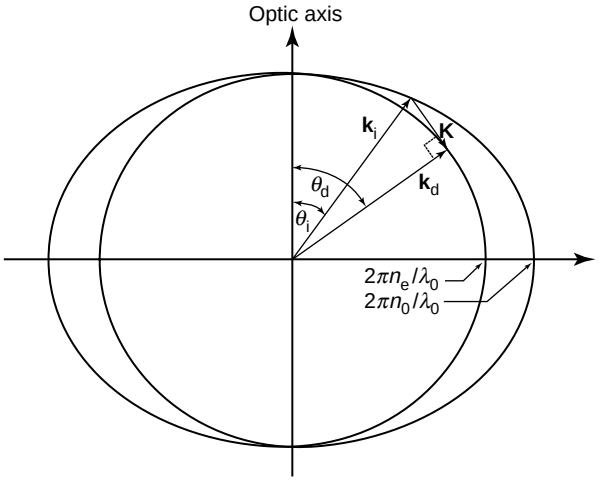


Figure 7 Wavevector diagram for birefringent scanner using a positive uniaxial crystal.

For the positive uniaxial case, the acoustic and diffracted wavevectors will be perpendicular at the design point, which allows the center frequency to be calculated from geometry as

$$f_0 = (V/\lambda)(n_i^2 - n_0^2)^{1/2} \qquad [29]$$

where n_i is the index of refraction at the incidence angle. Due to the typically low velocity of shear waves, this frequency can be much lower than an isotropic scanner, and the incidence angle can also be chosen to adjust the frequency. It is also possible to use the optical activity of certain materials such as TeO_2, along with circularly polarized light to further reduce the acoustic frequency. It is important to maintain as low a frequency as possible due to acoustic attenuation, which is especially high with the soft materials typically used for birefringent applications.

The bandwidth over which the scanner can efficiently operate is fairly large due to having the acoustic and diffracted wavevectors perpendicular, and is approximately $\Delta f = 2f_0$, assuming an octave of bandwidth for the isotropic scanner. This will produce a larger scan angle than an isotropic scanner, and more spots of resolution. The number of resolution spots is determined through diffraction by the aperture size and diffraction, along with the scan angle. Since a larger aperture requires a longer time for the acoustic waves to propagate across the aperture, the response time τ to access the scanner will increase, and the product of τ and f is related to the number of spots by eqn [27]. The advantage of the birefringent scanner is that it can operate at a lower frequency f_0 with better performance. However, since other factors such as acoustic attenuation are important in designing a scanner, for some applications it might be better to operate an isotropic scanner at a higher frequency.

AOTFs

With anisotropic Bragg diffraction, the magnitude of the diffracted wavevector $\mathbf{k}_d$ will differ from the incident wavevector $\mathbf{k}_i$, which cannot occur for the isotropic case. This is illustrated in Figure 8 for an AOTF utilizing a negative uniaxial crystal where an extraordinary input wave propagating at an incident angle θ_i relative to the crystal axis is diffracted into an ordinary output wave at an angle θ_d. This occurs through an acoustic wave propagating at an angle θ_a

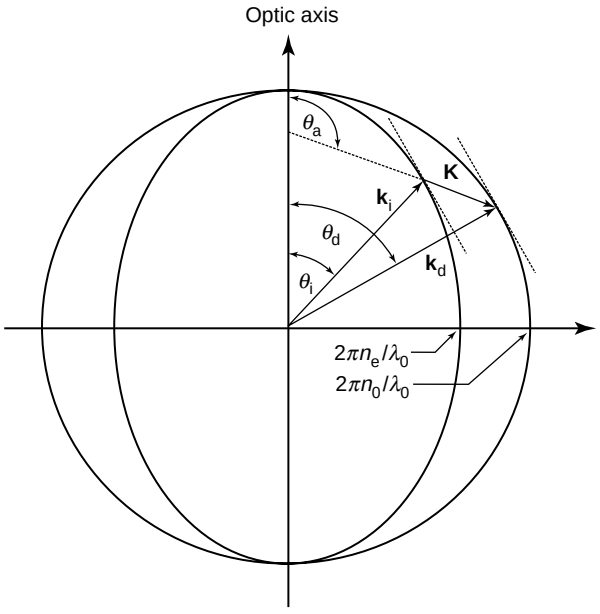


Figure 8 Wavevector diagram for an AOTF using a negative uniaxial crystal.

with wavevector $\mathbf{K}_a$, where all the wavevectors lie in a plane through the optic axis of the crystal. This diagram is identical to the index ellipsoid for the crystal, scaled by $2\pi/\lambda$, where λ is the vacuum wavelength. The acoustic wavevector is shown as adding to the incident wavevector, which increases the frequency of the optical wave by the acoustic frequency. It can also be represented in the reverse direction, which would reduce the optical frequency by the same amount.

An AOTF spectrally filters the optical input, and maintains its spectral purity over a large angular aperture. These conditions are maximized when the power flow or ray directions of the input and output beams are collinear. This produces a parallel-tangents condition, where lines drawn tangent to the wavevector surfaces and connecting the input and diffracted wavevectors are parallel. For beams at 90° to the optical axis, it is referred to as the collinear case, whereas all other parallel-tangents conditions are noncollinear. The diffracted beam is rotated by 90° during anisotropic Bragg interaction, so that crossed polarizers can be used to separate the input and diffracted beams. For the collinear case polarization separation must be used, whereas angular separation can also be used for the noncollinear case.

The AOTF design equations can be derived from the geometrical conditions imposed by the wavevector matching condition

$$\mathbf{k}_d = \mathbf{k}_i + \mathbf{K} \qquad [30]$$

along with decomposing the acoustic wave into its Fourier components which result from the finite interaction length produced by the transducer. These components allow for phase matching over a range of input angles, and they also allow for a spectral spread of the interaction. The full width half maximum (FWHM) resolution is given by

$$\Delta\lambda = \frac{1.8\lambda^2}{bL \sin 2\theta_i} \qquad [31]$$

where λ is the vacuum wavelength, b is the dispersion term (essentially $2\pi\Delta n$, where the birefringence Δn is the difference between the ordinary index n_0 and extraordinary index n_e of refraction), L is the interaction length of the input beam defined geometrically by the acoustic beam, and θ_i refers to the angle of the central or cardinal ray of the input beam, and the AOTF is designed around this angle. It is therefore a sensitive function of the incidence design angle, with the highest resolution occurring for the collinear case. The resolution is also narrower for larger birefringence.

The geometry of a noncollinear AOTF is shown schematically in Figure 9. The acoustic waves propagate in the correct direction and are generated by a transducer of length L_t, which is related geometrically to the interaction length. These waves are generally absorbed at the other side of the AOTF to prevent interfering reflections, and the sides may also be wedged to eliminate reflections into the interaction region. The optical beam enters the AOTF at the correct angle to the crystal optic axis, and either diffracts or passes through as the zero order. The input beam will have an angular spread, producing an acceptance angle that is a function of the interaction length, wavelength, and crystal parameters, and the external solid angular aperture

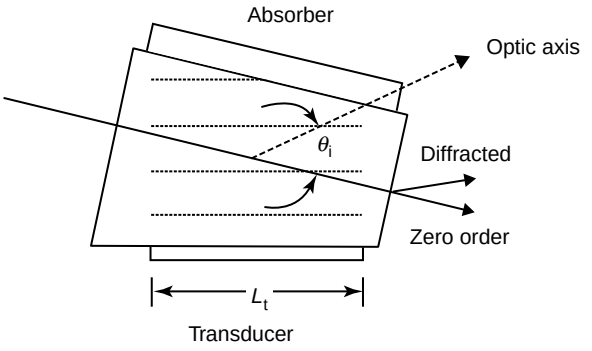


Figure 9 Noncollinear AOTF orientation of optical and acoustic beams.

is roughly given by

$$\Delta\Omega = \frac{\pi n^2 \lambda}{\Delta n L} \qquad [32]$$

where n refers to the diffracted beam index of refraction, which can be either n_0 or n_e depending upon the design. The resolution and solid angle are therefore related through the transducer length, and the product of the resolving power times the solid angle is given by

$$(\lambda/\Delta\lambda)(\Delta\Omega) = 2\pi n^2 \sin^2\theta_i / |F_\theta F_\varphi|^{1/2} \qquad [33]$$

where

$$F_\theta = 2\cos^2\theta_i - \sin^2\theta_i \qquad [34a]$$

$$F_\varphi = 2\cos^2\theta_i + \sin^2\theta_i \qquad [34b]$$

are form factors for the polar and azimuthal components of the optical field. Since all the wavectors lie in a plane containing the optic axis, the azimuthal components of the incident, diffracted, and acoustic wavevectors must always be identical under the parallel-tangents condition.

The product $(\lambda/\Delta\lambda)(\Delta\Omega)$ forms a figure of merit for spectrometers, and for the collinear case, the product is identical to a Fabry–Perot etalon having an index of n, and at other angles it is more complicated due to the angular dependence. The angular field is also symmetrical for the collinear AOTF, but nowhere else other than along the optic axis, where the resolving power becomes zero.

The great advantage of the AOTF designed under the parallel-tangents condition is the large angular aperture compared to the general case. Under this condition, the dependence of the angular aperture on resolving power is second order rather than first order, and the angular aperture can be tens of degrees for a typical resolution, which is useful for imaging applications. At the special angle of 54.7°, the second-order dependence is also zero, and the aperture can be extremely large relative to the resolving power, although this requires a high-resolution device with a corresponding narrow field, since the condition would soon be violated at larger field angles.

The efficiency of the AOTF under phase matching is given by

$$\eta = \sin^2[\pi^2 M_2 P_a L^2/(2\lambda^2)]^{1/2} \qquad [35]$$

which depends on θ_i and various material properties. Since the interaction is anisotropic, M_2 must be taken as a tensor quantity, in which both the polarization and propagation directions of the acoustic and optical waves must be accounted for. For acoustic waves, the polarization direction is the particle motion, which is perpendicular to the propagation direction for shear waves. In general, M_2 is much larger for a specific configuration of a particular material, and AOTFs are designed around this condition. The range of useful design angles is also usually limited since M_2 is generally a sensitive function of θ_i. The angular dependence on M_2 is important in designing an efficient device since it can be near zero at specific design angles for some materials, such as with TeO_2 for the collinear case.

The AOTF must be designed to optimize performance. The resolution is generally given as a system requirement, and the design angle along with the transducer length must be adjusted to optimize the throughput, or total optical power through the AOTF. This requires maximizing the efficiency, angular aperture, and aperture dimension. The maximum crystal size that can be grown ultimately limits these parameters, both in the interaction length and in the aperture size.

AOTFs have been used for a wide range of spectral filtering applications for spectroscopy and laser applications. Both collinear and noncollinear devices have electronically tuned a variety of lasers, including dye, semiconductor, and Ti:sapphire lasers. The AOTF is placed within the laser cavity, which requires high transmission on the laser line, with enough resolution to separate nearby laser transitions. A related application is the multiplexing of optical communications systems, where the AOTF is used to separate the various laser diode wavelengths. Perhaps the most widely used application is spectroscopy, and single-point detection systems have been used for a variety of biological and chemical applications in the visible and infrared. These techniques have been extended to spectral imaging, again in the visible and the infrared. Due to a variety of AOTF aberrations the image quality will be somewhat degraded. Since the AOTF is electronically tunable, it can be rapidly modulated both in amplitude and wavelength, which allows for modulation spectroscopy to detect small signals in a large background. By applying multiple frequencies, the AOTF has also been used as a rejection filter, in which all wavelengths other than that selected are allowed to pass.

Modulators

Acousto-optic modulators are used to vary and control laser beam intensity. A Bragg configuration gives a single first-order output beam, where intensity is directly linked to the power of RF control signal. The rise time of the modulator is the time required for

the acoustic wave to traverse the laser beam. For minimum rise time the laser beam will be focused down, forming a beam waist as it passes through the modulator.

A Bragg configuration gives a single first-order output beam, the intensity of which is controlled by the RF signal power, and diffaction angle controlled by the drive frequency. Two deflectors can be used in series and at right angles to each other to give full two-dimensional scanning.

One requirement in the design of acousto-optic modulators is that the diffracted beam and undiffracted beam must be well separated. For an adequate extinction ratio, the Bragg angle should be at least as large as the divergence of the optical beam. This condition puts a minimum value on the center frequency. Equating the Bragg angle

$$\theta_B = \frac{\lambda f_0}{2nV} \qquad [36]$$

and the diffraction angle of the Gaussian beam

$$\delta\theta_0 = \frac{4\lambda}{\pi(nd)} \qquad [37]$$

where d is the beam diameter at the minimum, it follows that the lower limit of the acoustic frequency is given by

$$f_0 = \frac{8}{\pi\tau} \qquad [38]$$

We may combine these expressions to determine a limit of modulator bandwidth of acousto-optic modulators, with the result

$$\Delta f \sim \frac{1}{4} f_m \qquad [39]$$

i.e., the modulation bandwidth is approximately equal to 25% of the midband acoustic frequency, f_m. In view of the present status of transducer technology, the modulation bandwidth of acousto-optic modulators is limited to several hundred MHz.

A frequency shifter uses the shift inherent in the acousto-optic interaction to up- or down-shift a laser's frequency. Two kinds of shifters can be distinguished: the fixed frequency shifter, and the variable frequency shifter. The frequency shift is equal to the signal frequency applied to the transducer. Various device configurations can be used to shift the laser beam, such as multiple travels inside the shifter to double or triple the frequency shift, or a combination of two frequency shifters in series. Acousto-optic frequency shifters can be used, for example, in optical heterodyning and interferometry, or in laser Doppler velocimetry for particle velocity analysis.

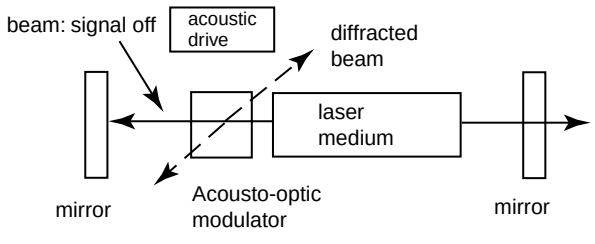


Figure 10 An acousto-optic Q-switch in a laser cavity.

Q-Switches

A Q-switch is a special modulator which introduces high repetition rate losses inside a laser cavity (typically 1 to 100 KHz). They are designed for minimum insertion loss and to be able to withstand very high laser powers. In normal use an RF signal is applied to diffract a portion of the laser cavity flux out of the cavity. This increases the cavity losses and prevents oscillation. When the RF signal is switched off, the cavity losses decrease rapidly and an intense laser pulse evolves. A diagram of a typical acousto-optic Q-switched laser cavity is shown in Figure 10. Q-switches are the preferred method of intracavity modulation where high speed, high stability, and modest cost are important design factors.

List of Units and Nomenclature

b	dispersion constant	–
c	elastic stiffness	dynes cm^{-2}
D, d	optical beam diameters	cm
e	elastic strain	–
f	acoustic frequency	hertz
F	optical focal length	cm
F_ϕ, F_θ	angular form factors	–
$\mathbf{k}$	optical momentum wavevector	cm^{-1}
$\mathbf{K}$	acoustic momentum wavevector	cm^{-1}
M_2	acousto-optic figure of merit	$sec^3\,gram^{-1}$
n	refractive index	
p	photo-elastic constant	–
P	pressure	dynes cm^{-2}
P_a	acoustic power	watts cm^{-2}
V	acoustic velocity	cm sec^{-1}
β	compressibility	$cm^2\,dyne^{-1}$
η	diffraction efficiency	–
ε	dielectric constant	–
ω	optical frequency	hertz
Ω	solid angle	steradians
ρ	density	grams cm^{-3}
θ	Bragg angles	radians
τ	acoustic wave travel time	sec

See also

Modulators: Acousto-Optics. **Optical Amplifiers:** Raman, Brillouin and Parametric Amplifiers. **Scattering:** Raman Scattering. **Spectroscopy:** Raman Spectroscopy.

Further Reading

Chang IC (1976) Acousto-optic devices and applications. *IEEE Transactions: Sonics and Ultrasonics* SU-23: 2–22.

Chang IC (1981) Acousto-optic tunable filter. *Optical Engineering* 20: 824–829.

Dixon RW (1967) Photo-elastic properties of selected materials and their relevance for applications to acoustic light modulators and scanners. *Journal of Applied Physics* 38: 5149–5153.

Gordon EI (1966) A review of acousto-optical deflection and modulation devices. *Proceedings of the IEEE* 54: 1391–1401.

Gottlieb M, Ireland CLM and Ley JM (1983) *Electro-optic and Acousto-optic Scanning and Deflection*. New York: Marcel Dekker.

Goutzoulis AP and Pape DR (1994) *Design and Fabrication of Acousto-optic Devices*. New York: Marcel Dekker.

Uchida N and Niizeki N (1973) Acousto-optic deflection materials and techniques. *Proceedings of the IEEE* 61: 1073–1092.

Xu J and Stroud R (1992) *Acousto-optic Devices: Principles, Design, and Applications*. New York: Wiley.

Electro-Optics

L R Dalton, University of Washington, Seattle, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

The electro-optic (EO) effect can be described as any one of a number of phenomena that occur when an electromagnetic wave in the optical spectrum (e.g., characterized by wavelengths in the range 200 to 2000 nm) interacts with an electric field, or with matter under the influence of an electric field. Two of the most important electro-optic phenomena are the Kerr effect (discovered by John Kerr in 1875) and the Pockels effect (discovered by Friedrich Pockels in 1893), in which birefringence is induced or modified in a liquid (the Kerr effect) or a solid (the Pockels effect). Birefringence is the difference in refractive indices for light of orthogonal line polarizations, one of which is parallel to the induced optical axis. The Kerr effect involves creation of birefringence in a liquid that is otherwise not birefringent. The degree of birefringence is quadratically proportional to the applied electric field strength. Hence, the Kerr effect is frequently referred to as the quadratic electro-optic effect while the Pockels effect involves a linear dependence on the applied electric field and is referred to as the linear electro-optic effect. In considering practical applications of the electro-optic effect, a commonly encountered term is that of 'electro-optic modulator', which can be described as a device wherein a signal control element is used to modulate a beam of light. The control element is typically an electric field with frequency components in the zero (DC) to hundreds of gigahertz (1 GHz = 10^9 Hz) range (or even tens of terahertz (1 THz = 10^{12} Hz). The modulation may be imposed on the phase, frequency, amplitude, or direction of the modulated optical beam.

Electro-optic effects are one class of second-order nonlinear optical phenomena. Other important second-order nonlinear optical phenomena include sum (e.g., second-harmonic) and difference frequency generation. Such phenomena derive from the second-order term in the power series expansion of macroscopic (material) polarization, P, in terms of applied electric fields:

$$P_i = \chi_{ij}^{(1)} E_j + \chi_{ijk}^{(2)} E_j E_k + \chi_{ijkl}^{(3)} E_j E_k E_l + \cdots \quad [1]$$

where $\chi_{ij}^{(1)}$, $\chi_{ijk}^{(2)}$, and $\chi_{ijkl}^{(3)}$ are the linear, second-, and third-order optical susceptibilities, respectively. Second-harmonic generation (SHG) or frequency doubling (2ω) effects, where a beam at twice the frequency of the input beam is generated, can be seen by substituting a sinusoidal field, $E_\omega \cos(\omega t - kz)$; for E_j and E_k. After using a well-known trigonometric identity, the equation for macroscopic polarization becomes (through second-order and dropping the subscript i):

$$P = \chi^{(1)} E_\omega \cos(\omega t - kz) + (1/2)\chi^{(2)} E_\omega^2 \times [1 + \cos(2\omega t - 2kz)] + \cdots \quad [2]$$

The electro-optic effect can be appreciated by considering interaction of the medium with an optical field $E_\omega \cos(\omega t - kz)$ and a low-frequency field E_0. Note that although we have used the symbol zero to

denote the frequency of the low-frequency electrical field, actual frequencies can extend to hundreds of gigahertz or even tens of terahertz. Again substituting (and now keeping terms to third order), the equation for polarization can be written as:

$$\begin{aligned} P &= \chi^{(1)}E_\omega\cos(\omega t-kz)+2\chi^{(2)}E_0E_\omega\cos(\omega t-kz) \\ &\quad +3\chi^{(3)}E_0^2E_\omega\cos(\omega t-kz)+(3/4)\chi^{(3)}E_\omega^3\cos(\omega t-kz) \\ &= \chi_{\text{eff}}E_\omega\cos(\omega t-kz) \end{aligned} \quad [3]$$

The equation for the nonlinear index of refraction, n, can be expressed as

$$\begin{aligned} n^2 &= 1+4\pi\chi_{\text{eff}} \\ &= 1+4\pi[\chi^{(1)}+2\chi^{(2)}E_0+3\chi^{(3)}E_0^2+(3/4)\chi^{(3)}E_\omega^2] \end{aligned} \quad [4]$$

Denoting the linear index of refraction as n_0, the above equation can be rewritten as

$$n^2-n_0^2=8\pi\chi^{(2)}E_0+12\pi\chi^{(3)}E_0^2+3\pi\chi^{(3)}E_\omega^2 \quad [5]$$

which in turn leads to

$$\begin{aligned} n &= n_0+(4\pi/n_0)\chi^{(2)}E_0+(6\pi/n_0)\chi^{(3)}E_0^2 \\ &\quad +(3\pi/2n_0)\chi^{(3)}E_\omega^2 \end{aligned} \quad [6]$$

The definition of light intensity in cgs units is $E_\omega^2=(8\pi/cn_0)I(\omega)$, which when substituted into the above equation gives

$$\begin{aligned} n &= n_0+(4\pi/n_0)\chi^{(2)}E_0+(6\pi/n_0)\chi^{(3)}E_0^2 \\ &\quad +(12\pi/cn_0^2)\chi^{(3)}I(\omega) \end{aligned} \quad [7]$$

The index of refraction can now be written as

$$n(\omega)=n_0(\omega)+n_1(\omega)E_0+n_2(0)E_0^2+n_3(\omega)I(\omega) \quad [8]$$

where the terms $n_1(\omega)$, $n_2(0)$, $n_3(\omega)$ correspond to the linear electro-optic effect, the quadratic electro-optic effect, and the optical Kerr effect, respectively. The above equation can also be expressed as

$$n=n_0-(1/2)n^3rE_0-(1/2)n^3sE_0^2-\cdots \quad [9]$$

where r and s are the linear and quadratic electro-optic coefficients, respectively.

Unfortunately, a great deal of confusion exists concerning the use of the terms electro-optic effect and electro-optic modulator. First of all, the terms 'electro-optic' and 'opto-electronic' are frequently confused and used interchangeably. Opto-electronic devices function as electrical-to-optical or optical-to-electrical signal transducers. Examples of commonly used opto-electronic devices include light-emitting diodes (LEDs), photodiodes (PDs), injection laser diodes (ILDs), and integrated optical circuit (IOC) elements. An electro-optic device can also function as an electrical-to-optical signal transducer (e.g., a Mach–Zehnder interferometer or a birefringent modulator employing polarizers at the input and the output); however, the mode of operation is entirely different for electro-optic and opto-electronic devices. For example, with LEDs or modulated lasers, the applied electric field produces modulation of transmission of light by altering the excited state population of the light emitting (lasing) state of the material. With electro-optic materials, no actual excited state population is generated. Electro-optic devices typically operate in regions of the optical spectrum removed from resonant transitions, i.e., regions of relatively high transparency. The applied electric field acts only to perturb the charge distribution of the material (the spatial positions of the electrons and nuclei). The altered charge distribution interacts with the optical beam transmitting the material with the result that the speed of light in the material (i.e., the index of refraction or birefringence) is altered. The response times of opto-electronic and electro-optic phenomena (and hence bandwidths of devices exploiting these phenomena) are quite different. In the former case, response will be limited by the lifetime of the relevant (emitting) excited state while in the latter case, response will be defined by the relaxation time of the material (e.g., reorientation time of a liquid or lattice relaxation time of a solid) following removal of the perturbing electric field.

An even greater confusion can arise due to the jargon used in particular technologies, such as telecommunications. A telecommunications engineer frequently refers to 'electro-optic switching' when describing opto-electronic transduction of an optical signal to an electrical signal, followed by re-routing in the electrical domain, and finally opto-electronic transduction of the electrical signal back to the optical domain. Electro-optic switching (in the sense used in this article) involves quite a different operation and would be described by that same telecommunications engineer as 'all-optical switching' – a term reserved by physicists for the optical Kerr effect (control of one light beam by another light beam).

A second point of confusion involves distinguishing between the terms 'electro-optic' and 'electro-absorptive' modulation. Again, the terms are frequently used interchangeably. However, they involve quite different physical mechanisms. An electro-absorptive

modulator is, like a modulated laser or LED, a 'resonant' device. In a device such as a gallium arsenide (GaAs) or indium phosphide (InP) quantum dot electro-absorptive modulator, the position and width of an optical absorption (resonant transition) are defined by the physical dimensions of the quantum dot. Application of an electric field shifts the spectral wavelength of the quantum dot absorption. This phenomenon can be used to dramatically change optical transmission through (or equivalently, absorption by) the material by choosing the wavelength of the propagating beam of light to be near the resonant absorption. With the electric field off, the material is reasonably transparent but becomes strongly absorbing when the electric field is applied, shifting the resonant absorption to the wavelength of the propagating beam. Obviously, the voltage required to achieve a desired change in optical transmission will depend on control of quantum dot dimensions in fabrication, which has been and continues to be a topic of research focus for electro-absorptive materials. Likewise, control of insertion losses of such devices is a concern as the wavelength used for the transmissive state of the device must be sufficiently close to the optical resonance to achieve the desired attenuation with application of a modest control voltage. Another issue to be faced with electro-absorptive devices is that of 'chirp' (optical signal distortion arising from the fact that both absorption and index of refraction – the imaginary and real parts of the optical susceptibility – change with application of the electric control field). On the positive side, electro-absorptive modulators are 'polarization-insensitive' modulators and are thus conveniently used with multimode (as well as single mode) optical transmission.

Further confusion exists because widely different types of materials can be used in electro-optic and electro-absorptive devices. As these materials are frequently competing for the same applications (signal transduction, optical switching, etc.), the advantages and disadvantages of different materials are often compared without maintaining the context that quite different phenomena are involved. Even restricting our discussion to electro-optic materials, we note that materials can range from organic liquid crystalline materials (nematic, smetic, ferroelectric, etc.) to organic electro-optic materials (both crystalline and 'macromolecular') to inorganic crystalline materials (lithium niobate, $LiNbO_3$; barium titanate, $BaTiO_3$; barium borate, BBO; potassium dihydrogen phosphate, KDP; lithium tantalate, $LiTaO_3$; zinc telluride, ZnTe, etc.). The response times of these materials to transient application of an applied electric field will be quite different, reflecting the different types of lattice motion involved. With liquid crystalline materials, particularly nematic materials, considerable molecular reorientation is involved and due to the mass that must be moved, response times are quite slow although the index change will be relatively large. With use of liquid crystalline materials, device bandwidths will typically be limited to tens of megahertz (MHz) or less. In contrast, conjugated π-electron organic materials typically exhibit response times of tens to hundreds of femtoseconds, which translates to potential device bandwidths of tens of terahertz (actual device bandwidths may be less due to factors other than material response). For conjugated π-electron organics, the response time is the phase relaxation time of the π-electron system. Because only a slight change in bond length alternation of the conjugated π-electron system occurs with application of an applied electric field and because strong electron–phonon coupling and resonance stabilization of the π-electron system act to reduce the barrier to lattice relaxation, the response times of π-electron organic materials are among the fastest observed in nature. With crystalline inorganic materials, the ionic constituents move to new locations with application of an electric field with the exact movement determined by the field strength, the charge on the ions, and the restoring force. Unequal restoring forces along the three mutually orthogonal (perpendicular) axes of the crystal lead to anisotropy in the optical properties of the material. The applied electric field will induce a change in the anisotropy (the principal refractive indices). The symmetry of the electro-optic tensor will be closely related to the symmetry of the piezoelectric tensor. The linear electro-optic effect requires that the crystal exhibit noncentrosymmetric (acentric or ferroelectric) symmetry. A centrosymmetric crystal possesses a center of symmetry defined by identical particles existing in the lattice at vectors $\mathbf{r}$ and $-\mathbf{r}$, where $\mathbf{r}$ is a position vector measured from an arbitrary origin. A centrosymmetric crystal, like an isotropic liquid or gas, can exhibit a quadratic electro-optic effect.

For the sake of completeness, it can also be noted that index of refraction changes can be induced by acoustic waves (acousto-optic modulation) and by heating (thermo-optic modulation). Also, elasto-optic and photo-elastic effects can produce index of refraction changes.

To keep this article to a manageable length, discussion is limited to solid-state electro-optic materials. Design principles being used to produce new organic electro-optic materials will be

illustrated. Since inorganic materials exist as crystals, limited chemical modification of such materials is possible, although processing techniques for fabricating thin films of such crystalline materials have been developed in a number of cases. For organic macromolecular materials, development of new materials is an on-going activity. Several representative devices being fabricated, using electro-optic materials, will be discussed. Four of the most commonly encountered types of EO devices include: (i) stripline devices; (ii) prism, cascaded prism, and superprism devices; (iii) resonated devices, such as ring microresonator and photonic bandgap devices, and (iv) waveplates. Of course, devices such as stripline devices can be configured in a variety of ways to produce polarization-sensitive and polarization-insensitive electrical-to-optical signal transducers, optical switches either using simple 1×2 or 2×2 switch architectures or multimode interference (MMI) switches, optical gyroscopes, photonically controlled phased array radar systems, spectrum analyzers, optical digital signal processors, analog-to-digital (A/D) and digital-to-analog (D/A) signal converters, electromagnetic (radiofrequency to millimeter wave) signal generators, voltage sensors, etc. Prism-type devices are typically used for spatial light modulation or laser beam deflection. Ring microresonator devices afford a rich array of applications ranging from active wavelength division multiplexing (WDM) to active optical interconnect reconfiguration, to voltage-controlled wavelength tuning of laser outputs.

Basic Relationships

The effect of the applied field is to deform the index of refraction ellipsoid, which can be represented as

$$\left(\frac{1}{n^2}\right)_1 X^2 + \left(\frac{1}{n^2}\right)_2 Y^2 + \left(\frac{1}{n^2}\right)_3 Z^2 + 2\left(\frac{1}{n^2}\right)_4 YZ + 2\left(\frac{1}{n^2}\right)_5 XZ + 2\left(\frac{1}{n^2}\right)_6 XY = 1 \qquad [10]$$

leading to the corrections

$$\Delta\left(\frac{1}{n^2}\right)_i = \sum_{j=1}^{3} r_{ij} E_{0j} \qquad [11]$$

where the electro-optic tensor components are related to the tensor components of the second-order nonlinear optical susceptibility by $r_{ij} = -(8\pi/n_{0i}^2 n_{0j}^2)\chi_{ji}^{(2)}$. The full tensorial equation for change of the 'indicatrix' (index of refraction ellipsoid) can be expressed as

$$\begin{pmatrix} \Delta\left(\frac{1}{n^2}\right)_1 \\ \Delta\left(\frac{1}{n^2}\right)_2 \\ \Delta\left(\frac{1}{n^2}\right)_3 \\ \Delta\left(\frac{1}{n^2}\right)_4 \\ \Delta\left(\frac{1}{n^2}\right)_5 \\ \Delta\left(\frac{1}{n^2}\right)_6 \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \\ r_{41} & r_{42} & r_{43} \\ r_{51} & r_{52} & r_{53} \\ r_{61} & r_{62} & r_{63} \end{pmatrix} \begin{pmatrix} E_{0X} \\ E_{0Y} \\ E_{0Z} \end{pmatrix} \qquad [12]$$

The electro-optic tensor reduces considerably for specific materials reflecting the symmetry of the particular material. For inorganic crystalline materials, the electro-optic tensor can be reasonably complex (containing many nonzero elements). It is also helpful to note that for crystalline inorganic materials, the electric-field-induced change in crystal shape leads to strain and the orientation of the indicatrix is altered, leading to an additional contribution to the change in index coefficients:

$$\Delta(1/n^2)_i = \sum_k p_{ik} S_k + \sum_j r_{ij} E_j \qquad [13]$$

where S_k is a component of the strain and p_{ik} is the elasto-optic tensor. At high frequencies, the inertia of the crystal prevents macroscopic straining so that the first term of the above equation vanishes. At low frequencies, elasto-optic effects cannot be ignored. Because the deformation leading to strain is generally caused by the inverse piezoelectric effect, it is linearly related to the applied electric field (the same dependence as the linear electro-optic effect). This can lead to a frequency dependence of the 'effective' electro-optic coefficient for crystalline materials.

A brief comment on the photo-elastic effect (the change in index coefficients produced directly by applied stress) is warranted. This effect has the form:

$$\Delta(1/n^2)_i = \sum_l \pi_{il} \sigma_1 \qquad [14]$$

where σ_l are the components of the stress and π_{il} are the piezo-optical coefficients. Note that this effect is independent of the applied electric field.

For axially symmetric 'charge-transfer type' organic chromophores prepared by deposition or

electric field poling, only two unique tensor elements, r_{33} and r_{13}, survive:

$$\tilde{r} = \begin{pmatrix} 0 & 0 & r_{13} \\ 0 & 0 & r_{13} \\ 0 & 0 & r_{33} \\ 0 & r_{13} & 0 \\ r_{13} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad [15]$$

The electro-optic effect for organic macromolecular materials derives from molecules with conjugated π-electrons confined within molecules (chromophores). The electro-optic tensor elements can be related to molecular first hyperpolarizability, β, and the chromophore number density, N (molecules/cm^3), by

$$\begin{aligned} r_{33} &= -2Nf(0)\beta\langle\cos^3\theta\rangle/n_0e^4 \\ r_{13} &= -Nf(0)\beta\langle\sin^2\theta\cos\theta\rangle/n_{0o}^2n_{0e}^2 \end{aligned} \qquad [16]$$

The expressions in brackets are 'order parameters' describing the degree of noncentrosymmetric order. $\langle\cos^3\theta\rangle = 1$ corresponds to all 'dipolar' (charge transfer type) molecules pointing in the same direction (perfect noncentrosymmetric or ferroelectric order) while $\langle\cos^3\theta\rangle = 0$ corresponds to complete disorder and no electro-optic effect. The reason that three angles appear in the order parameter is that an angle is required to represent the principal symmetry axis for the chromophore, the principal symmetry axis of the optical field, and the principal symmetry axis of the electric control field. Thus, averaging (denoted by brackets) must be carried out over each of these three angles. The tensor element r_{33} corresponds to the electric control field applied along the principal axis of the ordered chromophores and the principal axis of the optical field vector, transverse magnetic (TM) polarized light. The tensor element r_{13} corresponds to the control electric field applied along the principal axis of the optical field, transverse electric (TE) polarized light, and orthogonal to the principal chromophore axis. The notation r_{33} is contracted notation for r_{333}. The factor $f(0)$ takes into account the dielectric nature of the medium into which the chromophores (molecules with large first hyperpolarizability) are embedded.

For organic electro-optic materials, elasto-optic effects do not appear to make significant contributions. Moreover, electro-optic activity can be systematically improved by improving β and by optimizing the product of chromophore number density and order parameter. Due to the presence of intermolecular electrostatic interactions, W, among chromophores, order parameters and number density are not independent, e.g., for materials that are prepared by electric field poling the order parameter relevant to the principal component of the electro-optic tensor can be expressed approximately as

$$\langle\cos^3\theta\rangle = L_3(\mu f(0)E_p/kT)[1 - L_1(W/kT)^2] \qquad [17]$$

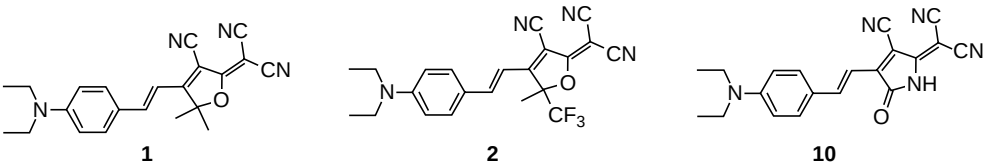
where L_3, and L_1 are Langevin functions, μ is the chromophore dipole moment, E_p is the strength of the electric poling field, k is the Boltzmann constant, and T is the Kelvin poling temperature. W is a quadratic function of chromophore number density, N. The above equation indicates that a maximum will be observed in the graph of r_{33} versus N.

Materials

With inorganic and organic crystalline materials, very little can be done to optimize electro-optic activity other than discovering new EO crystalline materials or employing isotopic (e.g., deuterium for hydrogen) or ion substitution with existing materials.

As with organic liquid crystalline materials, the electro-optic activity of macromolecular organic second-order nonlinear materials can be systematically improved by molecular modification. As shown in **Figure 1**, quantum mechanical calculations can provide useful guidance as to which structural modifications will lead to improvements in molecular (first) hyperpolarizability and ultimately electro-optic activity. The molecular hyperpolarizability (long wavelength limit) can be increased by a factor of two by simple variation of the acceptor structure. The calculated values of wavelengths for the interband charge transfer transitions are 390 nm, 403 nm, and 430 nm. Since dipole moments do not change significantly with structure, intermolecular electrostatic interactions should be similar for these three chromophores and the improvement in β should translate to an improvement in electro-optic activity. This theoretical prediction has been experimentally verified. This is just one example of use of quantum mechanics to guide the improvement of electro-optic activity for organic materials. This figure also illustrates the typical structure of an organic electro-optic chromophore, which consists of an electron donor region (an amine donor in the example shown), a connective

conjugated π-electron bridge, and an electron deficient electron acceptor region. The theoretically predicted (by density functional theory, DFT, calculations) variation of molecular first hyperpolarizability with molecular structure has been verified for the structures shown. However, the critical point is that electro-optic activity of organic materials (currently exhibiting a maximum value of 182 pm/V at 1.3 microns wavelength) is likely to be improved for some time to come. This value is to be compared to a value of 32 pm/V for lithium niobate. Another factor influencing the value of macroscopic electro-optic activity realized for electrically poled organic materials is the effect of intermolecular electrostatic interactions on optimization of $N\langle\cos^3\theta\rangle$. Statistical mechanical calculations can provide guidance as to how this is achieved (see **Figure 2** where the role of chromophore shape is illustrated). Approximating the chromophore shape by a rigid prolate ellipsoid with embedded dipole moment yields nearly quantitative reproduction of the experimental data. If the π-electron structure (dipole moment) is left unchanged but the shape of the chromophore is altered to a spherical shape, a significant improvement in electro-optic activity is predicted. The shape effect arises because there are two components to the electronic dipole–dipole intermolecular electrostatic potential function describing the many body interaction of chromophores.



	1	2	10	
Dipole moment:	11.8	11.9	11.7	Debye
Hyperpolarizability	23.8	34.4	43.0	$*10^{-30}$ esu

Figure 1 Density functional calculations (DFT) of molecular first hyperpolarizability (β) and dipole moment (μ) for three chromophore structures.

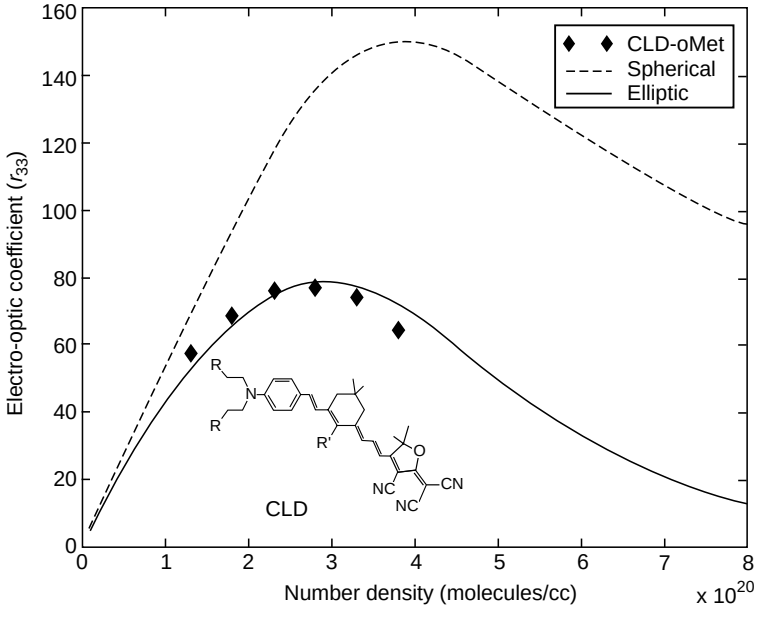


Figure 2 Two theoretical (statistical mechanical) simulations of the variation of electro-optic activity with chromophore number density are shown and compared with experimentally observed results for the CLD chromophore (structure shown) dissolved in an amorphous polycarbonate host polymer.

One component favors centrosymmetric ordering while the second component favors noncentrosymmetric ordering. Their relative weighting in defining macroscopic order depends on chromophore shape. Theoretical calculations have also shown that embedding chromophores in nanoscopic macromolecular objects (such as dendrimers and dendronized polymers) can dramatically enhance electro-optic activity. This type of engineering can also be used to control auxiliary properties such as optical loss and stability. The exceptional processability of organic electro-optic materials has permitted conformal and flexible devices to be fabricated and has permitted electro-optic circuitry to be integrated vertically on top of semiconductor VLSI (very large-scale integration) electronics. Mach–Zehnder modulators, with 3 dB operational bandwidths of 200 GHz, have been fabricated from organic electro-optic materials as have Mach–Zehnder devices with operational drive voltages of less than one volt. The major advantage that crystalline inorganic materials possess, relative to macromolecular organic EO materials, is superior thermal and photostability. A comparison of lithium niobate and the best organic EO materials, is given in Table 1. Cost is a significant factor if electro-optic materials are to compete with modulated lasers and other opto-electronic devices for applications in communications, computing, transportation, and defense. Currently, manufacturing costs for both crystalline and macromolecular materials are very high (several thousand dollars per modulator). Recently, it has been demonstrated that organic macromolecular materials can be used with soft lithography techniques to mass-produce electro-optic circuitry. This could conceivably dramatically reduce manufacturing costs. Also, the fact that organic macromolecular materials can be directly integrated with silicon photonics and VLSI semiconductor electronics, may influence future commercial utilization of electro-optic materials, which is currently dominated by lithium niobate.

Table 1 Comparison of electro-optic performance for inorganic and organic EO materials at 1.3 microns wavelength

Parameter	*$LiNbO_3$*	*NLO polymer*
Effective EO coefficient, r_{eff}, pm/V	32	182
Index of refraction (n)	2.2	1.6–1.7
Dielectric permittivity (ε)	30	3
Material Figure of Merit, $n^3(r_{eff})/\varepsilon$	10	270
Bandwidth × length product, GHz cm	7	>100
$V_\pi L$, V cm	5	1.5
Optical loss, dB/cm	0.2	0.2–1.0
Processing temperature (°C)	1000	<200
Multiple layers possible	No	Yes

Devices

Representative simple device structures are shown in Figure 3. For a simple stripline phase (birefringence) modulator (Figure 3, upper left), the phase retardation produced by application of an electric field is given by

$$\Delta\phi = 2\pi\Delta nL/\lambda = \pi n^3 rEL/\lambda = \pi n^3 rVL/\lambda d \qquad [18]$$

where L is the electrode length (the distance over which the electrical and optical fields co-propagate),

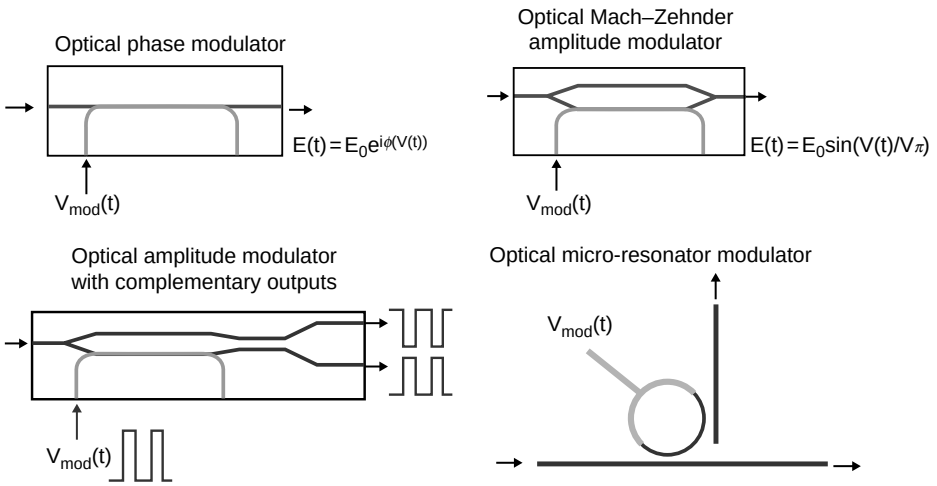


Figure 3 Schematic representation of simple electro-optic device structures are shown: Upper left (birefringent modulator), upper right (Mach–Zehnder interferometer), lower left (1 × 2 optical switch), lower right (ring microresonator).

λ is the optical wavelength, $V = E/d$ is the electric field felt by the material, and d is the electrode spacing. Electro-optic device performance is frequently characterized by the voltage, V_π, required to produce a phase shift of π radians. For a TM polarized light propagating through a birefringent modulator:

$$V_\pi = \lambda d/(n_{0e}^3 r_{33} L\Gamma) \qquad [19]$$

where Γ characterizes the overlap of the optical and electrical waves.

A simple Mach–Zehnder interferometer (electro-to-optical signal transducer) is shown in the upper right of Figure 3. When such a device is constructed from electrically poled organic electro-optic chromophores and a voltage is applied to one arm, a retardation of light propagating in that arm will be effected. This, in turn, will result in a change in the constructive/destructive interference of the beams at the output of the device. The consequence of applying an electrical signal to one arm of a Mach–Zehnder interferometer is that the applied voltage is transduced onto the optical output beam as an amplitude modulation of that beam. The voltage required to achieve full wave modulation for an optical beam of TM polarization is

$$V_\pi = \lambda d/(n_{0e}^3 r_{33} L\Gamma) \qquad [20]$$

Amplitude modulation can also be achieved using a birefringent modulator by placing polarizers at the input and output of the device. If an input polarizer is used to launch TM and TE modes, the birefringent modulator functions to produce a rotation of light, which is turned into an amplitude modulation by inserting a polarizer at the output. For such a device, V_π depends upon the difference in r_{33} and r_{13}. Note that for poled organic chromophores, $r_{33} \approx 3r_{13}$, $(V_\pi)_{\mathrm{BRM}} \approx 1.5(V_\pi)_{\mathrm{MZM}}$. For the above devices, it is clear that V_π will depend on the length over which the electrical and optical fields co-propagate in phase. For low electrical field frequencies, L will be the electrode length. The electrode spacing d, that can be used, will depend on the index of refraction difference between the core (electro-optic material) and a cladding material used to prevent the optical field experiencing the metal electrodes (an event that would dramatically increase propagation loss). For poled organic macromolecular materials and standard commercial claddings, d is typically in the order of 10 microns. Electrode lengths usually range from 1 to 3 cm. Two factors limit bandwidth in stripline devices: (i) velocity mismatch between the electrical and optical waves (which relates to the difference between n_0^2 and ε, where ε is the dielectric permittivity of the material – see Table 1); and (ii) microwave and millimeter wave loss in the metal electrodes. For organic electro-optic materials, the latter defines operational 3 dB bandwidth of devices.

Optical routing switches (see Figure 3, lower left) are more complicated to understand and typically require somewhat larger V_π voltages to effect a switching operation (e.g., $(V_\pi)_{\mathrm{simple2\times2crossbar}} \approx 1.7(V_\pi)_{\mathrm{MZM}}$).

Two conditions must be satisfied for light to be coupled into a ring microresonator device (see Figure 3, lower right): (i) the circumference of the ring must be a multiple of the wavelength of light, and (ii) the velocity of light in the ring must equal the velocity of light at the input. If a ring microresonator is fabricated from an electro-optic material, then an applied electrical voltage can be used to control coupling of light into and out of the ring microresonator. The same electrode structure, that is used to achieve voltage-controlled wavelength selective filtering, can also be used to transduce electrical information onto the optical beam while it is in the ring resonator. The quality (Q) factor of a ring microresonator defines both the wavelength selectivity and the bandwidth of the device. Ring microresonator devices also have the advantage of permitting a long interaction length to be achieved in a very compact device structure. The dimensions (radii) of ring microresonator structures are defined by bending loss and thus by the difference between core and cladding materials. Typical radii can range from hundreds of nanometers to hundreds of microns. With ring microresonators, a reduction in drive voltage requirements can be achieved by trading off bandwidth. Another price paid in the use of ring microresonator device structures is that of bending loss, which is not present in the other device structure shown. Ring microresonators can be used for active wavelength division multiplexing (WDM), voltage-controlled optical signal routing, and wavelength tuning of optical sources.

The interaction length of a prism optical beam deflector can be increased by cascading prisms together. For such a device structure, the angle of deflection becomes

$$\theta = n_0^3 r_{33}(V/d)(L/h) \qquad [21]$$

where h is the height of an individual prism and L is the length of the prism array. For currently available electro-optic materials only small deflection angles (a few degrees or less) can be achieved with practical

voltage levels. Superprism structures may permit large deflection angles to be achieved but such devices have yet to be demonstrated in a convincing manner.

See also

Interferometry: Overview. **Modulators:** Acousto-Optics. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. New York: Cambridge University Press.

Dagli N (1999) Wide-bandwidth lasers and modulators for RF photonics. *IEEE Trans. Microwave Theor. Techniques* 47: 1151–1171.

Dalton LR (2003) Rational design of organic electro-optic materials. *Journal of Physical Condensed Materials* 15: R897–R934.

Dalton L, Harper A, Ren A, *et al.* (1999) Polymeric electro-optic modulators: from chromophore design to integration with semiconductor very large scale integration electronics and silica fiber optics. *Ind. Eng. Chem. Res.* 38: 8–33.

Hornak LA (1992) *Polymers for Lightwave and Integrated Optics*. New York: Marcel Dekker.

Khazaei HR, Wang WJ, Berolo E and Ghannouchi F (1998) High speed GaAs/AlGaAs traveling wave electro-optic modulators. *Proceedings of SPIE* 3278: 94–100.

Lawetz C, Cartledge JC, Rolland C and Yu J (1997) Modulation characteristics of semiconductor mach-zehnder optical modulators. *IEEE Journal of Lightwave Technology* 15: 697–702.

Lee KS (2002). *Advances in Polymer Science*, vol. 158. Heidelberg: Springer.

Lee M, Katz HE, Erben C, *et al.* (2002) Broadband modulation of light using an electro-optic polymer. *Science* 298: 1401–1403.

Nalwa HS and Miyata S (1997) *Nonlinear Optics of Organic Molecules and Polymers*. Boca Raton, FL: CRC Press.

Noguchi K, Mitomi O and Miyazawa H (1998) Millimeter-wave Ti:$LiNbO_3$ optical modulators. *IEEE Journal of Lightwave Technology* 16: 615–619.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: John Wiley & Sons.

Shi Y, Zhang C, Zhang H, *et al.* (2000) Low (sub-1 volt) halfwave voltage polymeric electro-optic modulators achieved by controlling chromophore shape. *Science* 288: 119–122.

Wise DL, Wnek GE, Trantolo DJ, Cooper TM and Gresser JD (1998) *Electrical and Optical Polymer Systems. Fundamentals, Methods and Applications*. New York: Marcel Dekker.

Yariv A and Yeh P (1984) *Optical Waves in Crystals*. New York: John Wiley & Sons.

Modulation and Demodulation of Optical Signals

R A Minasian, University of Sydney, Sydney, NSW, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since earliest times, light has been used for communication purposes. In order to convey information, a physical variable of light, such as the optical intensity or field amplitude, must be modulated to encode a signal onto the optical carrier, and it must be demodulated or decoded after transmission. Laser sources provide the capability of extremely high modulation speeds, which result in high information-carrying capacity in optical communication systems. After transmission, semiconductor photodetectors demodulate the optical signals to recover the original information (*see* **Optical Communication Systems:** Basic Concepts).

The modulation of optical signals can be classified into two types: field modulation and intensity modulation, which is illustrated in Figure 1. In field modulation, the physical variable that is modulated is the frequency, phase, or amplitude of the optical field. These three techniques are sometimes referred to as frequency shift keying (FSK), phase shift keying (PSK) and amplitude shift keying (ASK). Figure 1b shows an example of frequency modulation of light. Field modulation requires keeping track of the phase of the carrier, which imposes strict requirements on the optical coherence (*see* **Coherent Lightwave Systems**). Because the optical carrier is at an extremely high frequency of around 2×10^{14} Hz at fiber-optic communication wavelengths of 1.5 μm, the required optical coherence is very difficult to realize, and moreover the receiver requires complex optical heterodyne methods to detect the phase of the optical field.

In intensity modulation, the physical variable that is modulated is the optical power. Figure 1c shows that it is the envelope of the light power that contains the information. The optical field oscillations at 10^{14}–10^{16} Hz in this case are not related to the

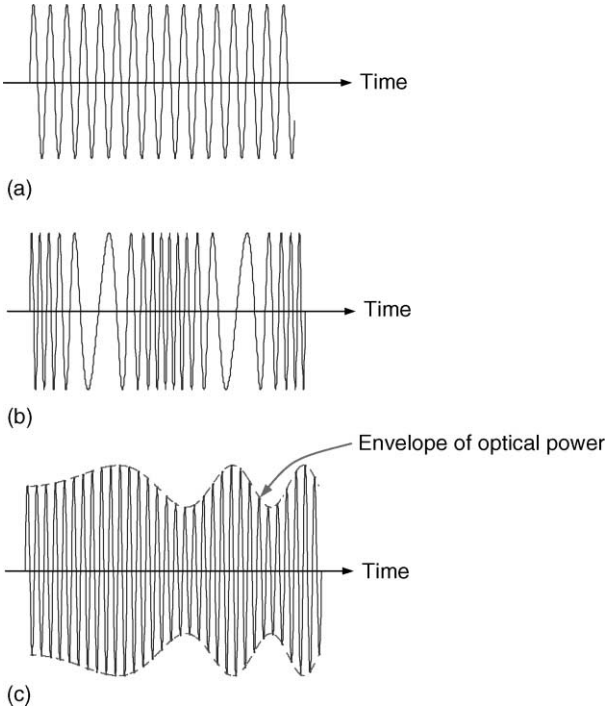


Figure 1 Types of modulation of optical signals. (a) unmodulated optical field; (b) frequency-modulated field; (c) intensity or envelope modulation of the optical power.

modulation and demodulation process, and only the optical power variation is important. This is a key advantage, and as a result intensity modulation systems are prevalent for optical communications. Here, we discuss how varying the injected current in a semiconductor laser or light emitting diode, or by operating the source with constant output and introducing an external modulator, can realize modulation of the optical power of the source. The process by which the optical power variations can be detected at the receiver with semiconductor photodiodes are also discussed, including detection noise characteristics and means of enhancing the responsivity by introducing internal gain mechanisms or wavelength selectivity. The speed capabilities of optical modulation and demodulation in semiconductor lasers and photodetectors are discussed, based on the rate equations for carrier and photon dynamics.

The Modulation of Semiconductor Lasers

Townes, Basov and Prokhorov, who shared the Nobel Prize in 1964, developed the principle of lasers. Semiconductor lasers emit light through a stimulated emission mechanism. The required optical gain for

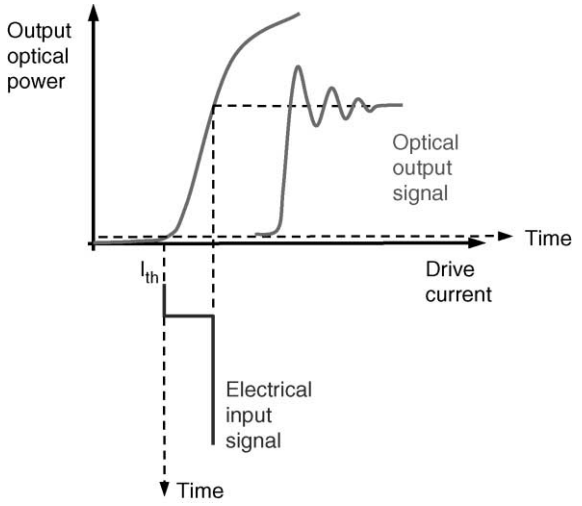


Figure 2 Light-current properties and the dynamic modulation characteristics of semiconductor lasers.

the lasing action is obtained by injecting a forward bias current into a heavily doped p-n junction, thereby injecting carriers into the active junction layer and creating a population inversion. Here we limit our discussion to the optical modulation characteristics of the laser in response to changes in the injection current.

Figure 2 shows the typical light-current characteristics of semiconductor lasers. For drive currents above the threshold value I_{th}, the emitted optical power is predominantly by stimulated emission and it increases linearly with the drive current up to moderate currents. The differential responsivity of the laser, defined by the ratio of optical power increase to electric current increase above threshold, is typically around 0.5 mW/mA. At high current levels, the light output deviates from a proportional dependence on the current because of saturation effects and the differential responsivity decreases.

The dynamic response of the semiconductor laser to modulation of the injection drive current is governed by the rate equations for the carrier density and the photon density in the active junction region. Figure 3 shows the active layer under consideration (*see* **Lasers:** Edge Emitters). The electron concentration in the active layer is denoted N and the photon concentration in the active layer is denoted S. The dynamic behavior of the electron concentration and photon density is governed by the following rate equations:

$$\frac{\partial N(x)}{\partial t} = \frac{I(x)}{q\nu_a} + D_n \frac{\partial^2 N(x)}{\partial x^2} - \frac{N(x)}{\tau_{ns}} - G(x)\cdot S(x) \qquad [1]$$

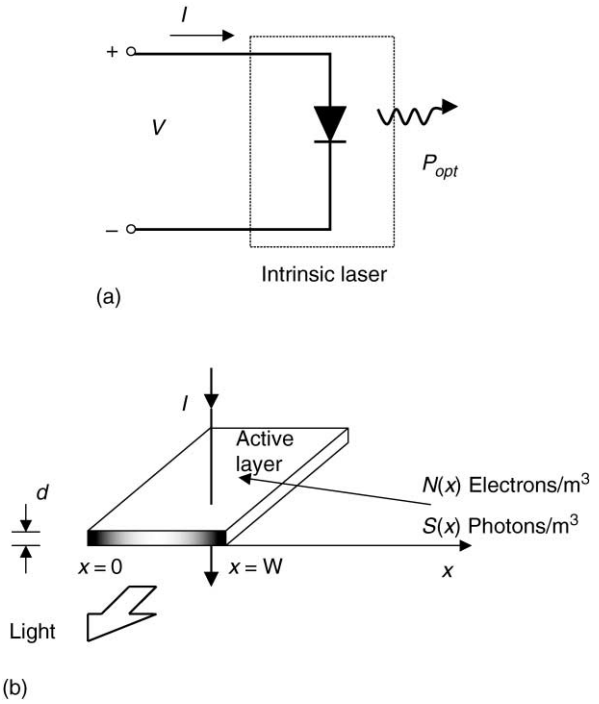


Figure 3 Semiconductor laser. (a) intrinsic device; (b) active region with injected current I, electron concentration N, and photon concentration S.

$$\frac{\partial S(x)}{\partial t} = \Gamma \cdot \beta \frac{N(x)}{\tau_{ns}} + \Gamma \cdot G(x) \cdot S(x) - \frac{S(x)}{\tau_p} \qquad [2]$$

where I is the injected current, q is the electronic charge, ν_a is the volume of the active layer, D_n is the diffusion constant, τ_{ns} is the carrier lifetime, Γ is the confinement factor (defined as the fraction of the optical power in the active region), β is the fraction of the spontaneous emission power coupled into the lasing mode, τ_p is the cavity or photon lifetime, and

$$G(x) = \gamma[N(x) - N_t] \qquad [3]$$

where G is the net rate of stimulated emission, and N_t is the electron concentration at transparency.

Equations [1] and [2] are statements of the fundamental continuity equations for electrons and photons in the active layer. Hence, eqn [1] expresses the fact that the rate of increase of electron concentration in the active layer is given by the rate of increase of electron concentration due to injected current, the rate of increase of electron concentration due to diffusion current, the rate of decrease of electron concentration due to spontaneous recombination, and the rate of decrease in electron concentration due to stimulated photon emission. Equation [2] expresses the fact that the rate of increase of photon density in the active layer is given by the rate of increase in photon density due to spontaneous emission coupled into the lasing mode, the rate of increase of photon density due to stimulated emission coupled into the active layer, and the rate of decrease of photon density due to cavity loss mechanisms.

The actual electron and photon distribution varies across the active layer. However, for simplicity, it is convenient to use average values of N and S, and the simplified rate equations in terms of position independent variables are given by:

$$\frac{dN}{dt} = \frac{I_i}{q\nu_a} - \frac{N}{\tau_{ns}} - G(1 - \varepsilon S)S \qquad [4]$$

$$\frac{dS}{dt} = \Gamma \cdot \beta \frac{N}{\tau_{ns}} + \Gamma \cdot G(1 - \varepsilon S)S - \frac{S}{\tau_p} \qquad [5]$$

where the scaling factor $(1 - \varepsilon S)$ accounts for gain saturation arising from several phenomena such as spatial hole burning, spectral hole burning, carrier heating and two-photon absorption, and ε is a nonlinear gain parameter.

The modulation response of the laser is obtained by solving the rate eqns [4] and [5] for a time dependent input current of the form

$$I(t) = I_b + I_m x(t) \qquad [6]$$

where I_b is the steady-state bias current, I_m is the modulation current amplitude and $x(t)$ represents the normalized shape of the modulation waveform.

The small-signal modulation bandwidth of the laser is obtained by considering the response to small amplitude sinusoidal current modulation at frequency ω, so that $x(t) = \sin(\omega t)$, around a steady-state bias current above lasing threshold. Under the small signal condition, the rate equations can be linearized and combined to result in a second order differential equation. The solution for the photon output exhibits a damped resonant type response. Hence the modulation response displays some peaking at the relaxation oscillation resonance frequency ω_r, and is flat for frequencies much lower than ω_r. The resonance frequency $\omega_r \propto \sqrt{S_o}$, where S_o is the photon density at the laser bias level, so the bandwidth can be increased by biasing the laser further above its threshold current. Semiconductor laser bandwidth to 30 GHz has been observed; however, it is often limited to around 10 GHz due to electrical parasitics in the current drive circuit and interface.

The large-signal modulation response of the laser is obtained by solving the differential eqns [4] and [5]. The response of the carrier-photon coupled system to

a step change of drive current has the potential for damped oscillation. This is caused by a resonant type of behavior in which, for example, an increase in photon density causes a depletion in population inversion which leads to a decrease in photon density enabling the carrier density to recover, thus producing a time lag between cause and effect resulting in damped oscillations before settling to the steady state value. These dynamic characteristics are illustrated in Figure 2. For a step modulation of the laser drive current, the optical output response exhibits a delay time, a risetime, an overshoot, and damped oscillations. The risetime of the laser output is typically in the range of 100–300 ps. The turn-on time of the laser and the overshoot are improved by ensuring that the minimum drive current does not fall below threshold. The damped oscillations in the optical output occur at the relaxation oscillation frequency of the laser.

These laser pulse waveform imperfections cause deteriorations in the eye diagram of the laser transmitter that is modulated by very high speed digital bit streams (*see* **Optical Communication Systems:** Lightwave Transmitters). Figure 4 shows the eye diagram for the optical output of a semiconductor laser directly modulated by a 5 Gbit/s pseudo-random bit stream of length $2^{15}-1$. The output waveform imperfections cause a partial closure of the eye, which ultimately limits the modulation speed that can be obtained by directly modulating the laser drive current. Although the semiconductor laser optical output does not exactly replicate the applied electrical drive waveform, the deviations are small enough so that these lasers can be used for data transmission up to around 10 Gbit/s.

Another inherent consequence of direct modulation of semiconductor lasers is the production of laser frequency chirp. This arises from the fact that modulation induced changes in the carrier density in the active region of the laser cavity cause a change in the refractive index of the active layer, which results in a chirping in the output wavelength during modulation. Hence amplitude modulation of semiconductor lasers is fundamentally accompanied by phase modulation from the basic physical mechanism of carrier induced changes in the refractive index. The resulting transient changes in the output wavelength of the laser causes a chirping phenomenon as it is being modulated. The order of chirp magnitude is about 0.1–1 GHz/mA of drive current change. During digital pulse modulation, the laser wavelength shifts to the blue side near the leading edge and towards the red side near the trailing edge of the pulse. Chirp results in a broadening of the laser linewidth. This wavelength change effect in combination with the dispersion characteristics of optical fiber (*see* **Fiber and Guided Wave Optics:** Dispersion; Light Propagation) causes an important limitation to the transmission capacity of fiber optic communication systems. The chirp in semiconductor lasers can be reduced, but not entirely eliminated, by using modulation-doped strained multiquantum-well lasers, which reduce the linewidth enhancement factor.

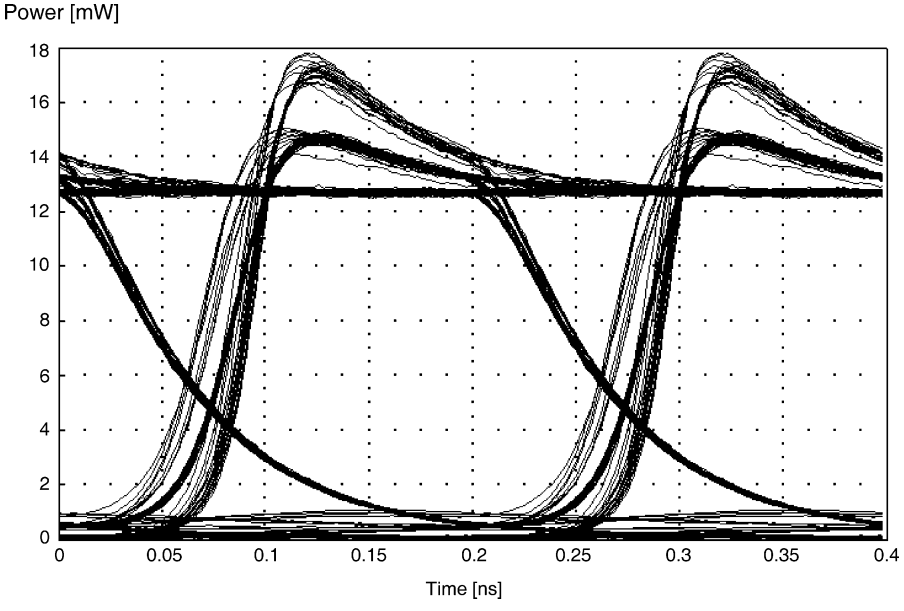


Figure 4 Eye diagram for the optical output of a semiconductor laser directly modulated by a 5 Gbit/s pseudo-random bit stream.

The Modulation of Light Emitting Diodes

The light emitting diode (LED) generates light through spontaneous emission. A semiconductor p-n junction is pumped with a forward current, so that electrons are provided in the upper energy state that can recombine radiatively to produce incoherent light by a spontaneous emission mechanism. The rate of photon generation is directly proportional to the carrier injection rate, and hence the modulating drive current.

Figure 5 shows the typical light-current properties and the modulation characteristics of a LED. The responsivity of the LED, defined as the ratio of the emitted optical power to the injected current, is typically in the range of 10 to 50 μW/mA. The optical output power generated is linearly related to the injection (drive) current up to moderate currents, and is given by

$$P = \eta \cdot h\nu \cdot \frac{I}{q} \qquad [7]$$

where η is the emission quantum efficiency (defined as the ratio of the radiative recombination rate to the total recombination rate), h is Planck's constant, ν is the optical frequency, I is the injected current, and q is the electronic charge. For larger currents, the light output deviates from a proportional dependence on the current because of saturation effects and the responsivity decreases.

The modulation response of the LED is governed by the rate equation for the carrier density N

$$\frac{dN}{dt} = \frac{I}{qv_a} - \frac{N}{\tau_c} \qquad [8]$$

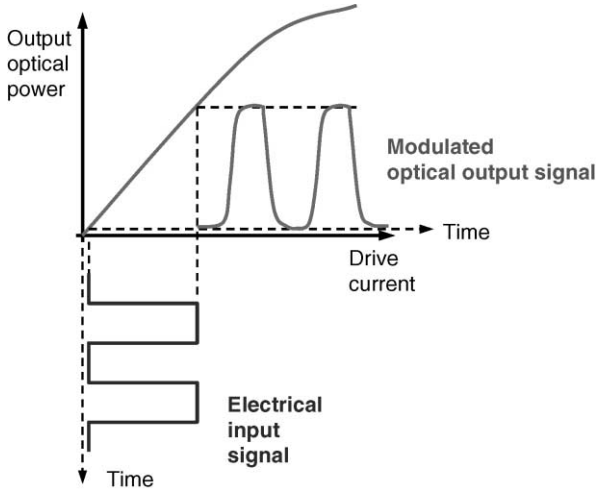


Figure 5 Light-current properties and the modulation characteristics of light emitting diodes.

where I is the injected current, q is the electronic charge, v_a is the volume of the active region, and τ_c is the carrier lifetime. Equation [8] expresses the fact that the rate of increase of carrier concentration is given by the rate of increase of carriers through current injection and the rate of carrier loss through radiative and nonradiative recombination. The small signal response arising from this first order differential equation, exhibits a simple single pole response and the resulting modulation bandwidth of the LED is given by

$$B = 1/(2\pi\tau_c) \qquad [9]$$

The pulse modulation response of the LED is shown in Figure 5. The output modulated optical waveform is not square but exhibits risetimes and falltimes determined by the carrier lifetime τ_c. This sets a limit to the maximum modulation rate that the LED can achieve. The bandwidth can be increased by decreasing the radiative lifetime, which can be controlled to a limited extent by the choice of doping level. Typical lifetimes for InGaAsP LEDs are in the range of 1–5 ns, which corresponds to attainable modulation data rates of several hundred Mbit/s.

The External Modulation of Lasers

Modulation of lasers realized by changing the injection current, as described above, is classed as direct modulation. This is the simplest method of modulation; however, it is limited by several physical phenomena. These include bandwidth limitations due to carrier dynamics in the laser active region, waveform limitations due to relaxation oscillations and overshoot effects, and chirping which causes transmission limitations in optical fibers arising from dispersion effects.

A superior modulation technique, classed as external modulation, operates the laser with continuous wave (CW) output and impresses the modulation by means of a device external to the laser. Although more complex, this approach has many modulation performance advantages. Because the laser is operated CW, this technique removes chirping effects and also enables higher power lasers to be used. Moreover, it is capable of much higher modulation bandwidths, larger on–off extinction ratios, and superior modulation spectral purity.

Electro-optic Modulation of Light

The linear electro-optic effect in materials was first described by Friedrich Pockels in 1893. The Pockels effect is the change in refractive index of the material

resulting from the application of an electric field, and is given by

$$n(E) = n - \frac{1}{2} r n^3 E \qquad [10]$$

where n is the refractive index, E is the applied electric field, and r is the electro-optic coefficient. Typical values of r lie in the range 10^{-12} to 10^{-10} m/V, and the most common electro-optic materials used include $LiNbO_3$, GaAs, InP, InGaAsP, and synthetic polymers such as polymethylmethacrylate (*see* **Modulators:** Electro-optics).

If light traverses an optical waveguide of length L, to which an electric field E is applied, it undergoes a phase shift given by

$$\Delta\phi = -\frac{\pi r n^3 E L}{\lambda} \qquad [11]$$

The electric field can be produced by applying a voltage V across the waveguide of width d, hence $E = V/d$. It is therefore possible to modulate the phase of the light by varying the voltage V that is applied to the material through which the light passes. The voltage required to produce a phase shift of π is given by

$$V_\pi = \frac{\lambda d}{n^3 r L} \qquad [12]$$

The parameter V_π, known as the half-wave voltage, is an important characteristic that expresses the applied voltage required to give a phase shift of π.

Intensity modulation of light can be obtained by placing the phase modulator in one arm of an optical interferometer. The structure, known as a Mach-Zehnder modulator, is shown in Figure 6a. The optical output is the phasor sum of two lightwaves that travel through the two arms, resulting in constructive or destructive interference depending on the optical phase difference. The electric field is applied to the optical waveguide in a direction perpendicular to the direction of light propagation using the metal electrodes, see Figure 6a. Hence the refractive index of the waveguide and thus the phase of the light can be modulated by varying the voltage V. When no voltage is applied, the optical fields in the two arms of the interferometer experience identical phase shifts, and hence interfere constructively at the output. When a voltage is applied, the waveguide index changes, and the light in that arm experiences a phase shift with respect to the light in the other arm. This destroys the constructive nature of the interference at the output, and reduces the transmitted light emitted at the output. The extreme

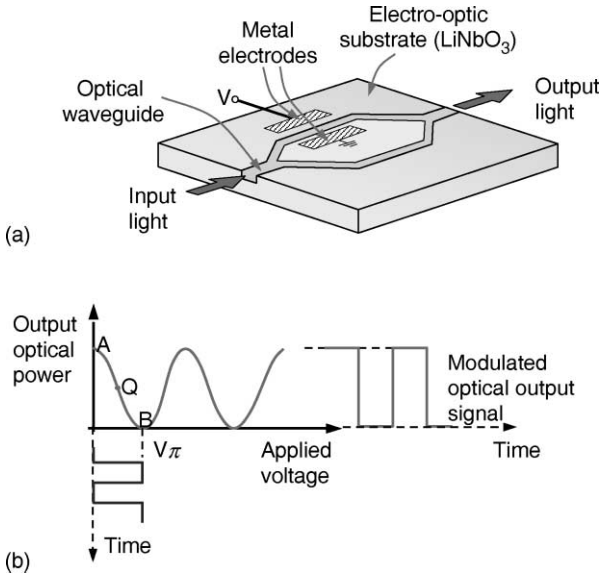


Figure 6 Mach-Zehnder electro-optic intensity modulator. (a) Structure; (b) transfer function and modulation characteristics.

case is when the phase difference between the two arms equals π, which causes complete destructive interference resulting in no light transmission.

The modulation transfer function is given by

$$P_{out} = \frac{1}{2} P_{in}[1 + \cos(\Delta\phi)] \qquad [13]$$

where $\Delta\phi$ is the phase difference encountered by light as it travels through the two arms. In terms of the applied voltage, the optical power transmitted through the modulator is given by

$$P_{out} = \frac{1}{2} P_{in}\left[1 + \cos\left(\pi \frac{V}{V_\pi}\right)\right] \qquad [14]$$

This transfer characteristic is shown in Figure 6b. The modulator may be operated as an analog optical modulator by operating in the nearly linear region around point Q. Alternately, it may be operated as a digital modulator by operating between points A and B, thus switching the light on and off as V is switched between 0 and V_π.

Improved efficiency of modulation, and hence a reduction in the value of V_π (typically a few volts) can be obtained by placing a second electrode on the other arm of the interferometer, and using a push-pull configuration to drive the interferometer. Hence the phase modulators in both arms are driven with modulating voltages of opposite polarities so the phase is advanced in one arm and is retarded in the other arm. This produces a net phase difference, which is double that produced in either arm.

The Pockels effect is an extremely fast physical mechanism, hence in principle external modulation is capable of achieving extremely high speeds. In practice, the modulation speed of an electro-optic modulator is limited by the capacitive effects at the electrodes and by the transit time of the light through the material. If the electric field varies significantly within the time it takes for light to transit through the waveguide, then the travelling optical wave will be subjected to different electric fields as it traverses the waveguide, and this results in a transit time limitation to the modulation bandwidth. A technique to avoid this limitation is to apply the voltage at one end of the waveguide and to design the electrodes to form a transmission line. By this means it is possible to make the velocity of the electrical wave match that of the optical wave to essentially eliminate the transit time effects. These external modulators have demonstrated speeds well in excess of 100 GHz.

Electro-absorption Modulation of Light

Electro-absorption external modulators are based on the principle of semiconductor bandgap decrease that occurs upon application of an electric field (the Franz-Keldysh and Stark effect). With no applied voltage, the semiconductor waveguide is transparent to light having a photon energy smaller than the bandgap, and such wavelengths are transmitted with very little attenuation. When a voltage is applied, the bandgap of the semiconductor decreases and the material absorbs the light once the photon energy exceeds the bandgap energy, and this produces optical attenuation. Hence by applying a modulation voltage to the semiconductor waveguide, the absorption property of the waveguide and hence the output light intensity can be modulated.

Improved modulation efficiency, and hence a reduction in the voltage required to decrease the bandgap energy, is obtained by using multiquantum-well semiconductor structures, which exhibit enhanced electro-absorption effects arising from the quantum confined Stark effect. The conventional technique of applying a transverse field across a multiquantum-well is to apply a reverse voltage across it. Modulator drive voltages of a few volts can be obtained. The on/off extinction ratio of modulation depends on the value of the field-dependent absorption coefficient at the operating voltage, and typical on/off extinction ratios are 10–15 dB.

The modulation bandwidth is determined by intrinsic factors related to carrier dynamics, such as exciton formation and dissociation times and the time taken to remove electrons and holes from the wells by thermionic emission and tunelling. It is also determined by extrinsic factors related to the capacitance of the junction diode. These modulators can have bandwidths of around 50 GHz. They also enable low chirp modulation to be obtained. The optical power handling capability of electro-absorption modulators is lower than for electro-optic modulators. However, a principal advantage of the electro-absorption modulator is that the semiconductor waveguide can be fabricated onto one substrate with the semiconductor laser, enabling the realization of monolithic integration on the same chip.

The Demodulation of Optical Signals

The process of demodulation converts the modulated intensity of an optical wave into an electrical signal to recover the information content. Semiconductor photodetectors are based on the principle of absorption of photons to generate mobile electron-hole pairs, and the transport of these mobile carriers via an applied electric field to create a flow of current.

Demodulation using PIN Photodiodes

The PIN photodiode comprises a P-doped/intrinsic/N-doped semiconductor structure. An applied reverse voltage bias creates a depletion region with a high electric field in the intrinsic layer. When the intrinsic layer is illuminated with light, photons are absorbed and electron-hole pairs are created, which are swept by the built-in electric field. The resulting flow of current is proportional to the incident optical power.

The quantum efficiency of the photodetector is defined as the ratio of electron generation rate to the photon incidence rate, and is given by

$$\eta = \frac{h\nu}{q} \cdot \frac{I_p}{P} \qquad [15]$$

where I_p is the photocurrent, P is the incident optical power, ν is the optical frequency, h is Planck's constant and q is the electron charge. If the width of the absorbing intrinsic region is large enough to capture all the light, quantum efficiency values approaching 100% can be achieved.

The current-voltage characteristics of the photodiode under optical illumination are shown in Figure 7a. The dark current I_d, which is the current generated in the photodetector in the absence of light originates from thermally generated electron-hole pairs, and is typically extremely small (<5 nA). Figure 7b shows the transfer characteristic of the

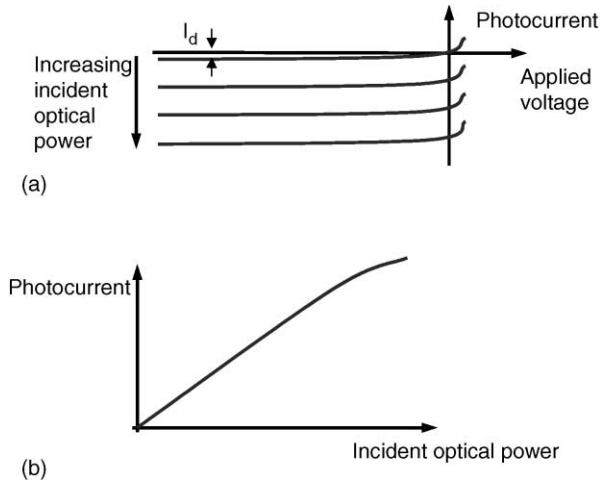


Figure 7 Semiconductor photodiode. (a) Current-light characteristics under optical illumination; (b) transfer function and demodulation photocurrent versus incident optical power.

photodetector. The current is linearly proportional to the optical power up to moderate power levels, when saturation mechanisms occur. The responsivity of the photodiode $\Re$, defined as the ratio of the generated electric current to the incident optical power, is given by

$$\Re = \frac{\eta q}{h\nu} \qquad [16]$$

The responsivity is an important parameter, and provides a measure of the photodetection efficiency of the device for demodulating optical signals. Typical responsivity values for PIN photodiodes are around 1 A/W.

The bandwidth B of demodulation of the photodetector is determined by the speed with which it responds to variations in the incident optical power. This is limited by the transit time τ_{tr} and the RC time constant τ_{RC} and is given by

$$B = \frac{2\pi}{\tau_{tr} + \tau_{\mathrm{RC}}} \qquad [17]$$

The transit time limitation arises because of the time it takes for the generated carriers to traverse the absorption region and be collected. To maximize the bandwidth, the transit time needs to be reduced, which can be achieved by decreasing the width of the absorption region. However, this requires a trade-off with the responsivity, which drops because a smaller fraction of the light can be absorbed under these conditions. The capacitance C and the resistance R of the photodetector and its circuitry form the RC time constant. To maximize the bandwidth, the photodiode capacitance needs to be reduced by decreasing the area of the photodiode. The demodulation bandwidth of PIN photodiodes can exceed 100 GHz.

Photodetection shot noise is a fundamental noise mechanism that introduces noise in the demodulation process of optical signals. This leads to fluctuations in the photocurrent, even for a constant incident optical power, because the photocurrent actually consists of a stream of electrons that are generated at random times. The photodiode current generated by a constant incident optical power is given by

$$I(t) = I_{\mathrm{avg}} + i_{\mathrm{shot}}(t) \qquad [18]$$

where I_{avg} is the average photocurrent generated, and $i_{\mathrm{shot}}(t)$ is the current fluctuation due to shot noise and which is described by a stationary random process with Poisson statistics. The mean-square noise current is given by

$$< i^2_{\mathrm{shot}}(t) >= 2qI_{\mathrm{avg}}\Delta f \qquad [19]$$

where Δf is the effective noise bandwidth of the receiver. The shot noise degrades the signal to noise ratio of the demodulated signal, and causes a fundamental limit to the sensitivity of the PIN photodetection.

Demodulation using Avalanche Photodiodes

Demodulation using avalanche photodiodes detectors (APD) has the advantage of significantly increased responsivity. This is achieved by providing an internal current gain mechanism within the device. The APD operates by generating more than one electron-hole pair per absorbed photon. This is achieved by strongly reverse biasing the diode to create a large junction electric field, so photoelectrons accelerate and acquire enough energy to produce new carriers by the process of impact ionization. These devices incorporate a modification to the PIN structure, by including an additional doped layer known as the multiplication layer, which under reverse bias provides a region of very high electric field that can produce avalanche breakdown. Hence a single primary electron generated through absorption of a photon, generates many secondary electrons and holes in the avalanche multiplication layer, and all these secondary carriers contribute to the photocurrent. As a result of this current gain mechanism, the responsivity of the APD is increased by a multiplication factor M, and is given by

$$\Re_{ADP} = \frac{M\eta q}{h\nu} \qquad [20]$$

where M can range between 5–50 depending on the bias voltage.

The bandwidth of the APD depends on the multiplication factor M. This is because the avalanche process takes additional time. The frequency response can be written

$$M(\omega) = \frac{M_0}{[1 + (\omega\tau_e M_0)^2]^{\frac{1}{2}}} \qquad [21]$$

where M_0 is the low frequency gain and τ_e is the effective transit time that depends on the ionization coefficient. The intrinsic APD bandwidth (assuming no RC limitations) is approximately given by

$$B_{\text{APD}} = \frac{1}{2\pi\tau_e M_0} \qquad [22]$$

Hence the bandwidth decreases as the multiplication factor increases, and there exists a trade-off between the bandwidth and the responsivity. The demodulation characteristics of the avalanche photodiode can be expressed by means of a gain-bandwidth product $B_{\text{APD}} \cdot M_0$, and this can exceed 150 GHz.

Photodetection noise in APDs has an additional contribution because the avalanche gain itself is a random variable. Hence the secondary electron-hole pairs are generated at random times and this adds a noise component to the basic shot noise associated with the generation of primary electro-hole pairs. The total shot noise is given by

$$< i^2_{\text{shot,APD}}(t) >= 2qM^2F(M)I_{\text{avg}}\Delta f \qquad [23]$$

where $I_{\text{avg}} = \Re P$, and F is the excess noise factor expressed by

$$F = \frac{< M^2 >}{< M >^2} \qquad [24]$$

and is approximately given by $F \approx M^x$, where $0.2 < x < 1$. Hence the excess noise increases with the multiplication ratio, so there exists an optimum avalanche photodiode gain that maximizes the signal to noise ratio of the demodulated signal. Despite the fact that additional noise is introduced, because of the higher responsivity, APDs can provide higher sensitivity in demodulating optical signals than PIN photodiodes.

Demodulation using Metal-Semiconductor-Metal (MSM) Photodiodes

Demodulation using MSM photodiodes enables extremely high bandwidths to be achieved. The MSM photodiode comprises two reverse-biased metal-to-semiconductor Schottky contacts formed on a semiconductor layer that prevents flow of electrons from metal to semiconductor. Photons incident between the contacts create electron-hole pairs that are collected by the electric field to generate a photocurrent proportional to the incident optical power. The metal contacts are usually fabricated on the same side of the semiconductor to create a planar structure, which is suitable for monolithic integration. However, a drawback of the MSM device is that the electrodes block some of the light from reaching the absorption region, which results in a reduced responsivity.

The bandwidth of demodulation of the MSM photodiode is determined by the transit time of the photogenerated carriers and the RC time constant. Since the electrode spacing can be around 1 μm, the transit time is extremely small. Moreover, since in the planar structure the parasitic capacitance is inherently low, the RC time constant is also extremely small. This results in demodulation bandwidths of up to 300 GHz.

Wavelength Selective Demodulation

A technique to obtain both wavelength selective demodulation and high-speed, high-responsivity demodulation involves the use of microcavity resonant photodetectors. This is useful for very high-speed detection in wavelength multiplexed systems.

The microcavity photodiode comprises a Fabry-Perot cavity placed around the photodetector absorption region, to enhance the responsivity. The mirrors are typically formed using Bragg reflectors or dielectric mirrors implemented in the semiconductor. A resonance is built up in the cavity at the wavelengths of incoming light for which the round-trip phase shift is 2π, which corresponds to constructive interference. Hence at the resonant wavelength, the incoming light is reflected at the two mirrors and permits the absorbing layer to be reused many times by effectively recirculating the optical signal, thus enhancing the absorption and the responsivity. However, the transit length can remain extremely small, because the photogenerated carriers only have to traverse the thickness of the absorption region to be collected. This solves the problem of the fundamental responsivity-bandwidth trade-off in normal photodetectors, in which increasing the absorption layer width to increase responsivity increases the transit time thus reducing the bandwidth. Microcavity resonant photodiodes enable the simultaneous achievement of large bandwidth and high responsivity possible. Also, the resonant

behavior leads to wavelength selectivity, which can be used to advantage in some wavelength division multiplexed WDM applications (*see* **Optical Communication Systems**: Wavelength Division Multiplexing). This approach can provide demodulation of optical signals with quantum efficiencies close to 100% and with bandwidths in excess of 100 GHz.

Optical Receivers – Preamplification and Clock Recovery

The front end of an optical receiver comprises the photodiode and a preamplifier to amplify the electrical signal for further processing. In order to obtain a high demodulation bandwidth, it is important to minimize the photodetector capacitance and to minimize the input resistance of the preamplifier that acts as the load resistance. However, because the thermal noise current is inversely proportional to the load resistance, minimizing the load resistance causes excessive noise and degraded demodulation sensitivity. An alternative approach is to use a high-impedance preamplifier. This reduces the noise at the price of a lower bandwidth. Thus there is a trade-off between the bandwidth and sensitivity. The trans-impedance front-end, in which the load resistor is connected as a feedback resistor around an inverting amplifier, is often used as an optimum compromise that combines high sensitivity together with large bandwidth. This enables the use of a large resistance, which minimizes noise, but at the same time the effect of negative feedback reduces the input impedance by a factor equal to the amplifier gain, thus increasing the bandwidth.

In digital optical transmission systems the data recovery section of the optical receiver comprises a clock recovery circuit and a decision circuit. Accurate timing (clock) information about the bit slot, in order to synchronize the decision process, is obtained from the amplified front-end waveform using a timing extraction circuit (*see* **All-Optical Signal Regeneration**). Typical clock recovery circuits comprise a high-Q resonant circuit which rings at the bit rate frequency or a phase locked loop. This enables precise data time sampling in the following decision circuit within the bit intervals of the original pulse train.

See also

All-Optical Signal Regeneration. **Fiber and Guided Wave Optics:** Dispersion; Light Propagation. **Lasers:** Edge Emitters. **Modulators:** Electro-Optics. **Optical Communication Systems:** Basic Concepts; Lightwave Transmitters; Wavelength Division Multiplexing.

Further Reading

Agrawal GP (1997) *Fiber-Optic Communication Systems*. New York: Wiley.

Coldren L and Corzine S (1995) *Diode Lasers and Photonic Integrated Circuits*. New York: Wiley.

Gowar J (1993) *Optical Communication Systems*. Englewood Cliffs: Prentice Hall.

Kazovsky L, Bendetto S and Willner AE (1996) *Optical Fiber Communication Systems*. Boston: Artech House.

Keiser GE (1991) *Optical Fiber Communications*. New York: McGraw-Hill.

Miller SE and Kaminow I (eds) (1988) *Optical Fiber Telecommunications II*. New York: Academic Press.

Personick SD (1985) *Fiber Optics: Technology and Applications*. New York: Plenum Press.

Petermann K (1991) *Laser Diode Modulation and Noise*. Boston: Kluwer Academic.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: Wiley.

Senior J (1992) *Optical Fiber Communications*. NJ: Prentice Hall.

Siegman AE (1986) *Lasers*. Sausalito, CA: University Science Books.

N

NONCLASSICAL LIGHT

H Walther, University of Munich and Max-Planck-Institute for Quantum Optics, Garching, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The pioneering experiments by Hanbury-Brown and Twiss, and the subsequent development of the quantum theory of electromagnetic radiation by Glauber and Sudarshan, raised the interest in the question of photon statistics. It followed from their theories that states of the electromagnetic field which can only exist in a quantum treatment. These nonclassical states show up in photon statistics through the phenomenon of antibunching. Later, further nonclassical states of the radiation field were found in the so-called squeezed fields, which are fields with phase-sensitive quantum fluctuations being at certain phase angles smaller than those of a perfect classical coherent field. In order to achieve squeezing of laser radiation, various forms of nonlinear optical interactions can be used.

Classical Versus Quantum Waves

In classical electrodynamics, a monochromatic wave (e.g., an extended light beam) is often described using the complex representation:

$$E(r,t) = E^{-}(r,t) + E^{+}(r,t) = \frac{1}{2}[\varepsilon(r,t)\mathrm{e}^{-\mathrm{i}\omega t} + \varepsilon^{*}(r,t)\mathrm{e}^{\mathrm{i}\omega t}] \qquad [1]$$

In classical theory, this complex description is merely a tool for the simplification of some calculations. However, there is some physical significance when using the quantum description, where ε and ε^{*} lead to the lowering and raising operators of the field. The above field can also be represented in the following way:

$$E(r,t) = \frac{\varepsilon + \varepsilon^{*}}{2}\cos\omega t + \frac{\varepsilon - \varepsilon^{*}}{2}\sin\omega t \qquad [2]$$

with real coefficients $\varepsilon_1 = (\varepsilon + \varepsilon^{*})/2$ and $\varepsilon_2 = (\varepsilon - \varepsilon^{*})/2\mathrm{i}$. This description gives the state of the field in a two-dimensional phase diagram where the sin-component is plotted versus the cos-component. In this plot, the wave above is represented by one point giving phase and amplitude of the field. These phase diagrams will play an important role in characterizing the nonclassical fields in a quantum phase diagram, as discussed below.

The properties of a field are described by the first- and second-order correlation functions. The first-order correlation function is given by

$$G^{(1)}(r_1,t_1,r_2,t_2) = \langle \varepsilon^{*}(r_1,t_1)\varepsilon(r_2,t_2)\rangle \qquad [3]$$

It is obtained by multiplying the electric field amplitudes at two spacetime points and taking the ensemble average. For $r_1 = r_2$ and $t_1 = t_2$, $G^{(1)}$ is proportional to the intensity at this point and is, therefore, called the first-order in intensity. The first-order correlation function is connected to the spectrum $I(\omega)$ of the field through its Fourier transform.

The second-order correlation function is necessary to characterize the radiation field completely. It describes the intensity fluctuations, being the classical equivalent to photon statistics, and is given in the general form:

$$G^{(2)}(r_1,t_1,r_2,t_2,r_3,t_3,r_4,t_4) = \langle \varepsilon^{*}(r_1,t_1)\varepsilon^{*}(r_2,t_2)\varepsilon(r_3,t_3)\varepsilon(r_4,t_4)\rangle \qquad [4]$$

If $G^{(2)}$ is considered for two spacetime points $r_1 = r_2$, the connection to intensity fluctuations becomes obvious:

$$G_{\mathrm{I}}^{(2)}(\tau) = 4\langle I(t_1)I(t_2)\rangle = \langle \varepsilon^{*}(t_1)\varepsilon(t_1)\varepsilon^{*}(t_2)\varepsilon(t_2)\rangle \qquad [5]$$

The subscript I indicates that we have to deal with intensity correlations; τ stands for $t_2 - t_1$.

In connection with the Hanbury-Brown, Twiss experiment, in general, the normalized intensity correlation is given by

$$g_{\mathrm{I}}^{(2)}(\tau) = \langle I(t)I(t+\tau)\rangle / \langle I(t)\rangle^2 \qquad [6]$$

$I(t)$ stands for the intensity which is given by the mean value $\langle I(t)\rangle$ plus fluctuations $\Delta I(t)$ for which the mean value is zero. The mean value of the squared intensity is then $\langle I^2\rangle = \langle I\rangle^2 + \langle \Delta I^2\rangle$. It follows for the normalized intensity correlation function:

$$g_{\mathrm{I}}^{(2)}(\tau) = \langle I(t)I(t+\tau)\rangle / \langle I(t)\rangle^2 = \begin{cases} 1 + \langle \Delta I^2\rangle / \langle I\rangle^2 & \text{for } \tau = 0 \\ 1 & \text{for } \tau \to \infty \end{cases} \qquad [7]$$

For $\tau \to \infty$, the intensities are not correlated and $g_{\mathrm{I}}^{(2)}(\tau)$ becomes 1. Otherwise $\langle \Delta I^2\rangle / \langle I\rangle^2 \geq 0$; this means that $g_{\mathrm{I}}^{(2)}(\tau)$ is always equal to or larger than 1, and $g_{\mathrm{I}}^{(2)}(\tau)$ smaller than 1 is not possible classically. The photon correlation is usually measured by the classical Hanbury-Brown, Twiss setup shown in Figure 1.

The maximum of $g_{\mathrm{I}}^{(2)}(\tau)$ at $\tau = 0$ is called intensity bunching and was first observed by Hanbury-Brown and Twiss. The decrease of $g_{\mathrm{I}}^{(2)}(\tau)$ as a function of τ is a measure of the coherence length of the light source.

When the experiment in Figure 1 is considered in the photon picture there is a principal difference compared to the classical intensity. In the classical case, the intensity on the two detectors is the same if the beamsplitter is dividing the light beam in equal portions. Both detectors, therefore, measure the same fluctuations. In the quantum picture, however, a photon can only be detected in one of the two detectors; this leads to the result that values for $g_{\mathrm{I}}^{(2)}(\tau)$ smaller than 1 can be found. The effect of photon bunching was initially thought to be an inherent property of light owing to the boson nature of photons. Therefore, it came as a big surprise when Glauber showed theoretically that for the quantized radiation field $g_{\mathrm{I}}^{(2)}(\tau)$, <1 is possible. This effect, known as photon antibunching, and its experimental observation constitutes an unambiguous proof of the quantum nature of light.

In the quantum description of the electromagnetic field, the electric field amplitudes ε and ε^* are replaced by $\hat{a}$ and $\hat{a}^+$, the photon annihilation and creation operators.

In this case, the electric field operator is written as

$$\hat{E} \propto \frac{\hat{a} + \hat{a}^+}{2} \cos\omega t \, \frac{\hat{a} - \hat{a}^+}{2\mathrm{i}} \sin\omega t \qquad [8]$$

In analogy, the operators $\hat{a}_1 = (\hat{a} + \hat{a}^+)/2$ and $\hat{a}_2 = (\hat{a} - \hat{a}^+)/2\mathrm{i}$, can be defined describing the amplitude of the cosine and sine parts of the wave. $\hat{a}_1$ and $\hat{a}_2$ do not commute $[\hat{a}_1, \hat{a}_2] = (\mathrm{i}/2)1$, therefore, the expectation values of the two operators $\hat{a}_1$ and $\hat{a}_2$ underlie an uncertainty relation given by

$$\langle \Delta\hat{a}_1^2\rangle = \langle \hat{a}_2^2\rangle - \langle \Delta\hat{a}_i\rangle^2 \geq 1/16 \qquad [9]$$

with $\langle \Delta\hat{a}_i^2\rangle = \langle \hat{a}_i^2\rangle - \langle \Delta\hat{a}_i\rangle^2$. In the corresponding phase diagram, the expectation values for $\hat{a}_1$ and $\hat{a}_2$ are represented by a shaded area defined by the uncertainty (Figure 2). The state of the radiation field, where both phase and amplitude fluctuations are symmetrically minimized, $\langle \Delta\hat{a}_1^2\rangle = \langle \Delta\hat{a}_2^2\rangle = \frac{1}{4}$, best approximates the classical wave with well-defined phase and amplitude. The state with symmetrical minimum uncertainty is the coherent state describing laser radiation. It is an eigenstate of $\hat{a}$ with a photon number fluctuation given by a Poissonian distribution with $\langle \Delta n^2\rangle = \langle n\rangle$.

There are other minimum uncertainty states, however, with unsymmetrical uncertainty regions

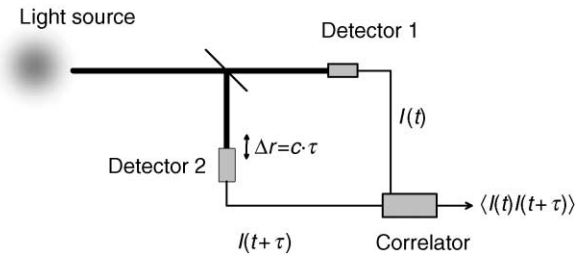


Figure 1 Experimental setup for the Hanbury-Brown, Twiss experiment.

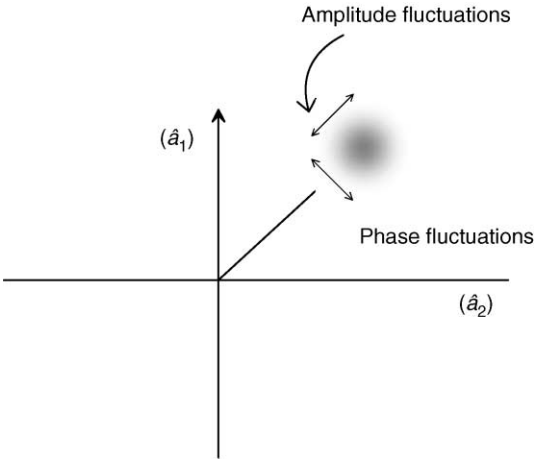


Figure 2 Phase diagram of the quantum mechanical field amplitudes $\langle \hat{a}_1\rangle$ and $\langle \hat{a}_2\rangle$ for coherent radiation.

with the uncertainty squeezed into mainly one component. These states, which are called squeezed states, are defined by $\langle \Delta \hat{a}_1^2 \rangle < 0.25$ or $\langle \Delta \hat{a}_2 \rangle < 0.25$. Squeezed state may also refer to a state which is not a minimum uncertainty state but where, nevertheless, the mean square fluctuation in one of the two components is less than 0.25.

A classical electromagnetic field consists of waves with well-defined amplitude and phase. This is not the case in a quantum treatment: fluctuations are associated with both conjugate variables. The case of a coherent state that most nearly describes a classical electromagnetic field has an equal amount of uncertainty in the two variables (when normalized to the field of a single photon). Equivalently, the field can be described in two conjugate quadrature components and the uncertainties in the two conjugate variables satisfy an uncertainty relation. The coherent state represents a minimum uncertainty state with equal uncertainties in the two quadrature components. This case is usually called the shot-noise limit; it is represented in Figure 2 as a phase diagram.

In a quantum treatment of radiation, it is also possible to generate states that are not present in the classical limit. They can show fluctuations reduced below the classical limit in one of the quadrature components, while the canonically conjugate quadrature component must display enhanced fluctuations in order to fulfill the uncertainty relation. Those states are called 'squeezed states.' An electromagnetic field with fluctuations below the standard quantum limit in one of the quadrature components has, in principle, many attractive applications, for example, in optical communication, in precision and sensitive measurements such as gravitational wave detection, or in noise-free amplification. Therefore, there has been great interest in generating squeezed radiation. Optical measurements have three characteristics that enable them to reach the quantum noise level more readily than in other fields of physics: (i) optical signals are naturally immune to external sources of noise; (ii) thermal noise at room temperature is negligible in the optical domain; and (iii) the outstanding equalities of the optical sources and detectors allow a very-low-level instrumental noise.

Squeezed states with reduced photon number fluctuations (sub-Poissonian photon statistics) correspond to a light beam in which the photons arrive more equally spaced than in a coherent beam for which $g_I^{(2)}(0) = 1$. Therefore, sub-Poissonian number fluctuations correspond to photon antibunching. This shows that photon antibunching and squeezing are related. The state with the largest possible antibunching is the number state, represented by a circle in the phase diagram. It has large fluctuations in both components a_1 and a_2 and, therefore, is not a squeezed state. Thus, it is obvious that antibunching does not imply squeezing. On the other hand, weak squeezing of a coherent state may result in antibunching. If the state is squeezed harder the a_1 component will be better defined. The amplitude fluctuations, however, will start increasing and there will be a transition from antibunching to bunching.

The quantum calculation of the normalized second-order correlation function leads to the following result:

$$g_I^{(2)}(0) = 1 + (\langle \Delta n^2 \rangle - \langle n \rangle / \langle n \rangle^2) \qquad [10]$$

A pure number state of the radiation field with $\langle \Delta n^2 \rangle = 0$, leads to $g_I^{(2)}(0) < 1$, as discussed above. For a coherent state, $\langle \Delta n^2 \rangle = \langle n \rangle$ yielding $g_I^{(2)}(0) = 1$, corresponding to the classical calculations.

It can be concluded that although antibunching and squeezing are related and nonclassical phenomena, one does not imply the other. Thus, the detection schemes necessary to determine the existence of one or the other are quite different. Antibunching or sub-Poissonian statistics can be detected via intensity correlations. Squeezing, on the other hand, is a property of either $\langle a_1 \rangle$ or $\langle a_2 \rangle$, and phase sensitive homo- or heterodyne detection is needed.

Nonclassical states of the radiation field are very sensitive to linear attenuation. Coherent and thermal light fields are the only ones that do not change their photon statistics when attenuated. This is different for nonclassical light. The change depends on the particular case and has to be carefully evaluated.

The first observation of squeezing was achieved in 1985, in an experiment of parametric generation involving four-wave mixing in sodium vapor.

Nonclassical Light from a Single Atom

Resonance fluorescence of atoms is a basic process in radiation–atom interactions, and has therefore always generated considerable interest. The methods of experimental investigation have changed continuously due to the availability of new experimental tools. A considerable step forward occurred when tunable and narrowband dye laser radiation became available. These laser sources are sufficiently intense to easily saturate an atomic transition. In addition, the lasers provide highly monochromatic light with coherence times much longer than typical natural lifetimes of excited atomic states. Excitation spectra with laser light using well-collimated atomic

beams lead to a width, being practically the natural width of the resonance transition, therefore it became possible to investigate the frequency spectrum of the fluorescence radiation with high resolution and to also study the photon statistics of the fluorescent light.

Previous experiments to investigate antibunching in resonance fluorescence have been performed by means of laser-excited collimated atomic beams. The initial results obtained in 1977 showed for the second-order correlation function $g^{(2)}(t)$, a positive slope characteristic of photon antibunching, but $g^2(0)$ was larger than $g^2(t)$ for $t \to \infty$. This was due to number fluctuations in the atomic beam and to the finite interaction time of the atoms. In 1978, the analysis of the experiment was refined and in 1982, another experiment with a longer interaction time was performed. In the latter experiment, the photon correlation was also measured for very low laser intensities. The ideal experiment to study the photon statistics in resonance fluorescence is the investigation of a trapped ion since, in such an experiment, the fluctuations of the atomic number are excluded.

The fluorescence of a single ion should also display the following property. The probability distribution of the photon number recorded in a finite time interval t is narrower than Poissonian, which means, that the variance is smaller than the mean value of the photon number. This is because the single ion can only emit a single photon. Antibunching and sub-Poissonian statistics are often associated. They are, nevertheless, distinct properties and need not necessarily be simultaneously observed, as is the case in the experiment described here. Although there is evidence of antibunching in the atomic-beam experiments, the photon counts are not sub-Poissonian as a result of fluctuations in the number of atoms. In further experiments this effect was excluded by use of a special trigger scheme for the single-atom event. In the setup with a trapped ion these precautions are not necessary since there are no fluctuations in the atomic number. Figure 3 shows the photon correlation of a single trapped ion. The measurements are performed using the $3^2S_{1/2} - 3^2P_{3/2}$ transition of the $^{24}Mg^+$-ion at a wavelength of 280 nm. The laser is tuned slightly below resonance in order to Doppler-cool the secular motion of the ion.

The investigation of the photon correlation employed the ordinary Hanbury-Brown and Twiss setup with two photomultipliers and a beamsplitter (Figure 1). The time delay τ between the photomultiplier signals was converted by a time-to-amplitude converter into a voltage amplitude proportional to the time delay. A delay line of 100 ns

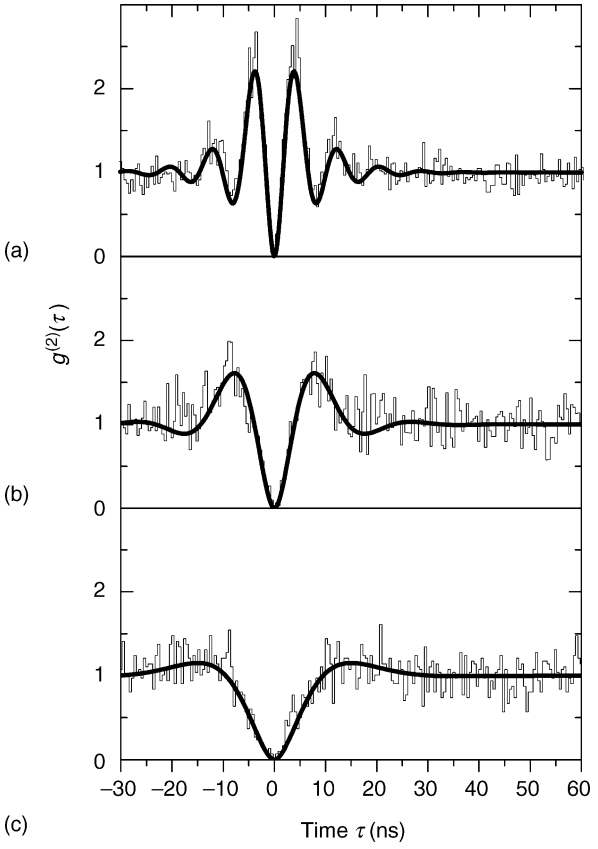


Figure 3 Antibunching signals of a single $^{24}Mg^+$-ion for different laser intensities (decreasing from (a) to (c)). At larger intensities (a) and (b) the Rabi oscillations owing to a strong coherent coupling between ground and excited state are observed.

in the stop channel allowed for the measurement of $g^{(2)}(\tau)$ for both positive and negative τ in order to check the symmetry of the measured signal. The output of the time-to-amplitude converter was accumulated by a multichannel analyzer in pulse height analyzing mode. Three typical measurements, each at small values of the laser intensity but with different detunings, are shown in Figure 3. The solid curve is a theoretical fit to the measurements. For small time delays (<20 ns), the nonclassical antibunching effect is observed, superimposed with Rabi oscillations being damped out with a time constant corresponding to the lifetime of the excited state.

Recently it also became possible to combine a trapped ion with a cavity. In this way the emission of a photon in the cavity mode can be achieved. This leads to a directed emission of the photon being more efficient than the fluorescence occurring in the entire 4π geometry. In this way, a deterministic single photon source is obtained.

Survey on the Experiments on Squeezing

In order to produce squeezed light from nonlinear optical phenomena, a variety of methods has been used, depending on the type of nonlinearity and the number of interacting modes involved. The choice of optical nonlinearity reduces to either three-wave or four-wave mixing, leading to quadratic or cubic nonlinear response functions, respectively. Each of these methods has advantages and disadvantages. Three-wave mixing requires a pump field at twice the frequency of the field-mode being squeezed. However, it has the advantage as any sidebands induced on the pump, owing to phase noise, have a large frequency offset from the squeezed fields. The other method, the much simpler four-wave mixing, does not require any frequency-doubling step. This was, therefore, the first method used to obtain squeezing with quadrature noise below the vacuum level.

Most of the early experiments on squeezing were performed with one or two cavity modes with the squeezed modes being transferred to propagating external modes via beamsplitters for measurement. This narrowband procedure is followed in more recent experiments commonly used for three-wave (parametric) squeezing. Also squeezing in waveguides was achieved leading to a relatively simple and broadband implementation of squeezing, with, mostly a four-wave mixing with short pulses or solitons being performed.

The main limiting factors that remain include control over the driving laser phase and intensity noise, absorption in the dielectric material and associated devices, and phase fluctuations caused both by low-frequency technical noise in components, and by intrinsic material properties of phonon interactions and refractive index fluctuations. It appears not impossible to reach squeezing levels of 10 dB or more below shot noise, over bandwidths of up to 1 THz. At this level of noise reduction, the main problem is the issue of improving detector technology to allow such low noise levels to be investigated.

Squeezing in Lasers

The quantum noise properties of the laser were studied intensively after its invention in the early 1960s. The conclusion of these studies is that the laser output approaches quite closely the noise properties of a coherent state, provided the laser is operated well above threshold. That is, the spectral variance of the photon number and phase fluctuations at high enough Fourier frequencies will be at the Poissonian or quantum noise limit (also referred to as the shot noise level). On the other hand, the phase of the laser is unconstrained and drifts as a function of time, so-called laser phase diffusion. As a result, the phase fluctuations show excess noise at low frequencies.

In the early 1980s, it was realized that if the pumping mechanism for the laser could be made sub-Poissonian then amplitude squeezing could be obtained at low frequency. A few years later it was shown that sub-Poissonian pumping could be realized in semi-conductor lasers by regularizing the pump current. An experimental demonstration of amplitude squeezing from semi-conductor lasers was soon achieved. Although still requiring rather specialized lasers, many groups now routinely obtain squeezing from semi-conductor lasers.

There are a number of disadvantages to semi-conductor lasers, including their poor beam quality and short coherence length. As a result, interest in other possible squeezing mechanisms continues. It has been found that intrinsic dynamical effects can also lead to amplitude squeezing in the laser output.

Laser squeezing is of a different character to that produced by more 'traditional squeezers.' For example, in parametric amplification, the de-amplification of phase fluctuations is accompanied by an equal and opposite amplification of amplitude fluctuations which preserves the uncertainty relationship between them. However, in laser squeezing the reduction in amplitude fluctuations is not linked to any increase in phase fluctuations. Thus, in laser squeezing we take advantage of the fact that the laser output is not in a minimum uncertainty state to suppress amplitude noise. This means that laser squeezing is always amplitude or intensity squeezing.

Application of Nonclassical Light

The applications of nonclassical light are numerous. The single photon sources discussed above have their main applications in quantum communication with single photons where they may be used to establish secure signal transmission. There is also a decisive advance in the field of quantum information expected from the combination of local atomic quantum processing with photonic transmission of quantum states. Conceptually, new possibilities would arise to enhance quantum computational capabilities. For example, separate local quantum processors of moderate size could be linked in a quantum network, with quantum information being exchanged by means of photons.

Although a large number of phenomena have been predicted in which the quantum nature of the squeezed vacuum plays an essential role in modifying

atomic properties, only the linear dependence of the two-photon absorption rate of a three-level ladder atom has been confirmed experimentally but classical sources show a quadratic dependence. In 1998 an important series of experiments was reported which investigated the interaction of a beam of Cs atoms with squeezed light in the Fabry–Perot cavity environment. The principal phenomena examined were the modification of the transmission spectrum of a weak, tunable probe beam due to the presence of the squeezed light, and the variation of the transmission modulation with the phase of the squeezed vacuum relative to that of an applied, saturating, coherent field.

There have been many proposed practical applications of squeezed light, and some experiments have already demonstrated its potential in precision measurements and improvement of spectroscopic sensitivity. Squeezed light may be used in optical communications to enhance signal-to-noise ratio, gravitational wave detection, and quantum nondemolition measurements.

The wider applications of squeezed light are numerous. An idea of the range and variety of applications is provided by the following selection: the protective measurement of a single squeezed harmonic oscillator state, coherent nonlinear spectroscopy using number squeezed light, the improvement of quantum nondemolition measurements using a squeezed meter input, sub-shot-noise Doppler anemometry with amplitude squeezed light, noiseless transfer of nonclassical light through bistable systems, generation of few photon states from squeezed atoms, and spin squeezing in a collection of atoms illuminated with squeezed light.

Einstein–Podolsky–Rosen (EPR) Correlations and Entanglement of Photons

In 1986, a quantum nondemolition experiment confirming the quantum correlation of the quadrature phase amplitudes of two spatially separated fields, was reported. In 1987, two-mode squeezing with two spatially separated detectors was experimentally demonstrated using parametric oscillation. With the experimental realization of such two-mode squeezing and quantum nondemolition measurements, the possibility of testing for continuous variable EPR-type correlations became apparent. In these proposals, the conjugate 'position' and 'momentum' observables would be the two orthogonal quadrature phase amplitudes of the field. The high efficiency of detectors used in the quadrature phase amplitude squeezing experiments was an important issue in this connection. Experimental tests of EPR correlations at the time were based on Bohm's version of the EPR paradox, and in fact, focused on Bell's extension of this work. These tests involved measurements performed on individual photon pairs, and used very inefficient detectors.

A specific proposal and a criterion to demonstrate the continuous variable EPR correlations for real experiments by way of a realization of the 1935 EPR paradox, was put forward in 1989. This proposal employed the two-mode squeezed state as the EPR source. The first experimental achievement, using the parametric oscillator, of the 1989 EPR criterion, for efficient measurements with continuous variable outcomes, was reported in 1992.

The EPR fields have proved significant in enabling the experimental realization of continuous variable quantum teleportation, and may have application also to quantum cryptography. Again, because of the efficient nature of the measurement of field quadrature phase amplitudes, and the intrinsic macroscopic nature of the correlated fields, continuous variable teleportation and cryptography may provide large advantages over protocols using a finite number state basis, where signals of individual photons are sent. At the moment, several groups are working in this promising direction.

Conclusions

In this contribution the generation of nonclassical light was reviewed. The generation of nonclassical light became possible after stable and reliable laser sources were available. There are many sources of nonclassical light available now. It is clear that we will see many new applications in the future. Especially in connection with optical communication, rather interesting improvements can be expected.

See also

Coherent Lightwave Systems. **Photonic Crystals:** Electromagnetic Theory. **Quantum Optics:** Atomic Coherence Effects.

Further Reading

Caves CM and Drummond PD (1994) Quantum limits on Bosonic communication rates. *Review of Modern Physics* 44: 481–537.

Cresser JD, Häger J, Leuchs G, Rateike FM and Walther H (1982) Resonance fluorescence of atoms in strong monochromatic laser fields. *Topics in Current Physics* 27: 21–59.

Dagenais M and Mandel L (1978) Investigation of two-time correlations in photon emissions from a single atom. *Physical Review A* 18: 2217–2228.

Diedrich F and Walther H (1987) Nonclassical radiation of a single stored ion. *Physical Review Letters* 58: 203–206.

Drummond PD and Ficek Z (eds) (2004) Springer Series on Atomic, Optical, and Plasma Physics, vol. 27. Berlin, Heidelberg, New York: Springer Verlag.

Gatti A, Lugiato LA and Wiedemann H (1997) In: Giacobino E, Reynaud S, and Zinn-Justin J (eds). *Quantum Fluctuations*, p. 431. Amsterdam: Elsevier.

Georgiades NP, Polzik ES, Edamatsu K, Kimble HJ and Parkins AS (1995) Nonclassical excitation for atoms in a squeezed vacuum. *Physical Review Letters* 75: 3426–3429.

Glauber R (1963) The quantum theory of optical coherence. *Physical Review* 130: 2529–2539.

Glauber R (1963) Coherent and incoherent states of the radiation field. *Physical Review* 130: 2766–2788.

Guthöhrlein GR, Keller M, Hayasaka K and Lange W (2001) A single ion as a nanoscopic probe of an optical field. *Nature* 414: 49–51.

Heidmann A, Horowicz RJ, Reynaud S, Giacobino E, Fabre C and Camy G (1987) Observation of quantum noise reduction on twin laser beams. *Physical Review Letters* 59: 2555–2557.

Höffges JT, Baldauf HW, Eichler T, Helmfrid SR and Walther H (1997) Heterodyne measurement of the fluorescent radiation of a single trapped ion. *Optical Communications* 133: 170–174.

Hood CJ, Kimble HJ, Turchette QA, Georgiades NPh and Parkins AS (1998) Squeezed excitation in cavity QED: experiment and theory. *Physical Review A* 58: 4056–4077.

Jakeman E, Pike ER, Pusey PN and Vaugham JM (1977) Effect of atomic number fluctuations on photon antibunching in resonance fluorescence. *Journal of Physics A: Mathematical and General* 10: L257–L259.

Kimble HJ, Dagenais M and Mandel L (1977) Photon antibunching in resonance fluorescence. *Physical Review Letters* 39: 691–695.

Kimble HJ, Dagenais M and Mandel L (1978) Multiatom and transit-time effects on photon-correlation measurements in resonance fluorescence. *Physical Review A* 18: 201–207.

Kimble HJ, Carnal O, Hu Z, Mabuchi H, Polzik ES, Thompson RJ and Turchette QA (1995) *Annals of the New York Academy of Sciences* 755: 87.

Levenson MD, Shelby RM, Reid MD and Walls DF (1986) Quantum nondemolition detection of optical quadrature amplitudes. *Physical Review Letters* 57: 2473–2476.

Loudon R (2000) *The Quantum Theory of Light*, 3rd edn. Oxford, New York: Oxford University Press.

Ou ZY, Pereira SF, Kimble HJ and Peng KC (1992) Realization of the Einstein–Podolsky–Rosen paradox for continuous variables. *Physical Review Letters* 68: 3663–3666.

Reid MD (1989) Demonstration of the Einstein–Podolsky–Rosen paradox using nondegenerate parametric amplification. *Physical Review A* 40: 913–923.

Short R and Mandel L (1983) Observation of sub-Poissonian photon statistics. *Physical Review Letters* 51: 384–387.

Short R and Mandel L (1984) Sub-Poissonian photon statistics in resonance fluorescence. In: Mandel L and Wolf E (eds). *Coherence and Quantum Optics*, vol. V, pp. 671–678. New York: Plenum.

Slusher RE, Hollberg LW, Yurke B, Mertz JC and Vally JF (1985) Observation of squeezed states generated by four-wave mixing in an optical cavity. *Physical Review Letters* 55: 2409–2412.

Walls DF (1979) Evidence for the quantum nature of light. *Nature* 280: 451–453.

Wu LA, Kimble HJ, Hall J and Wu H (1986) Generation of squeezed states by parametric down conversion. *Physical Review Letters* 57: 2520–2523.

Yamamoto Y and Machida S (1987) High-impedance suppression of pump fluctuation and amplitude squeezing in semiconductor lasers. *Physical Review A* 35: 5114–5130.

NONLINEAR OPTICS AT THE CRITICAL FIELD LIMIT

G A Mourou, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

It is now possible to produce, with compact lasers, intensities so large that the relativistic character of the electrons dominates laser–matter interactions. This is a fundamentally new regime in optics where optics, confined to the study of eV characteristic phenomena, are now addressing phenomena with characteristic energies vastly greater in the MeV–GeV (TeV in the near future) range. With relativistic optics, the field of nonlinear optics found a number of novel applications, including fast ignition of inertially

compressed fusion target, bright sources of energetic electrons, protons, ions, positrons, pions, etc., but also nuclear physics, high energy physics, astrophysics, and general relativity.

Relative optics has been one of the most active fields of optics and it is impossible to describe all the applications here. So, this new regime will be illustrated with few examples of the key relativistic mechanisms – reminiscent of conventional nonlinear optics, such as relativistic harmonics, relativistic optical rectification (wake-field acceleration), and relativistic self-focusing in refraction and reflection. We finally show that relativistic optics offers a new and efficient gateway to attosecond physics and nonlinear quantum electrodynamics (QED).

Optics is the study of how electrons respond to light. All material optical properties are the consequence of how light interacts with electrons in materials. Because, in the visible regime the photon energy is of the order of 1 eV, optics deals with phenomena with eV characteristic energy. Light is a wave composed of coupled electric and magnetic fields oscillating in synchrony at high frequencies. The electric and magnetic field oscillate perpendicular to each other, and perpendicular to the direction of propagation, k. Because of the electric field E, electrons under the action of the Lorentz force, $F = eE$, moves a distance $x = a_1 E$ transversally to the propagation direction. In classical optics, the oscillation amplitude x is small compared to the wavelength, and the electrons oscillate and radiate at the laser frequency. The radiated field amplitude and phase depend on how the electron is bound to the atom of the material. The invention of the laser in 1960 gave access to much larger intensities and electromagnetic fields. The laser field became strong enough that the electron displacement, x, in this field could no longer be considered as strictly proportional to the laser field. This departure from linearity is addressed by the addition of new terms in the Taylor series expansion of the displacement:

$$x = a_1 E + a_2 E^2 + a_3 E^3 + \cdots \qquad [1]$$

These additional terms are responsible for all the effects that form the field of nonlinear optics. Note here that the field is still not strong enough to ionize an electron from its parent atom and is therefore below the field that could irreversibly damage the material. Thus, the typical intensity must be kept to a value of less than $\sim 10^{14}\,\mathrm{W/cm^2}$, where the quiver energy E_Q starts to challenge the ionization energy of a few eV, depending on the material:

$$E_Q = \left\langle \frac{e^2 E^2}{4m\omega^2} \right\rangle \qquad [2]$$

The bracket denotes the time average over one cycle and ω is the laser angular frequency. Above this limit, the material response is dominated by multiphoton ionization or tunneling, as shown in Figure 1. During the ionization event, high harmonics are produced. Note here again that because the electron is still bound to its nucleus during the emission, we can consider the process of high harmonic generation as bound-electron nonlinear optics. As the intensity increases from 10^{14} to $10^{18}\,\mathrm{W/cm^2}$ (Figure 1), we are entering the regime of relativistic intensity or relativistic nonlinear optics. This is a regime that was predicted and discussed in the 1960s but that could be accessed only after the technique of chirped pulse amplification (CPA) was demonstrated, followed by significant progress in the areas of short pulse generation and pulse focusing.

Relativistic Optics

In classical optics one always assumes that during a laser oscillation period the electron mass m_0 stays constant. It is a reasonable assumption since the quiver velocity v of the electron is small compared to the speed of light c. As the intensity is increased a point is reached where the electron quiver velocity approaches the speed of light. This has several implications; first, the mass of the electron changes to become

$$m = \gamma m_0 \qquad [3]$$

where γ is the Lorentz factor given by

$$\gamma = \left(1 - \frac{v^2}{c^2}\right)^{-1} \qquad [4]$$

Second, because v is close to c, the magnetic force component needs to be added and the Lorentz force becomes:

$$\vec{F} = e\vec{E} + \frac{\vec{v}}{c} \times \vec{B} \qquad [5]$$

Until recently, in classical optics and laser physics, the $v \times B$ component was totally ignored. However, in the relativistic regime, this component is as large as

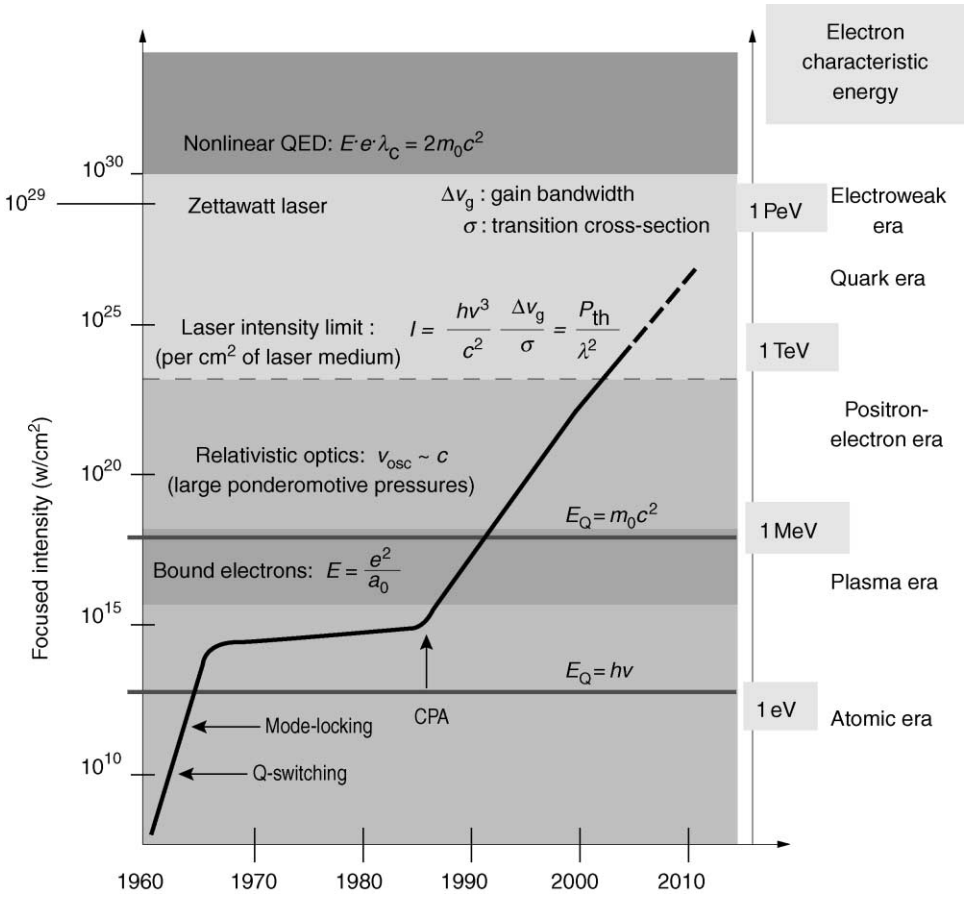


Figure 1 Laser intensity versus years. Note the very steep slope in intensities that occurs during the 1960s. This period corresponds to the discovery of most nonlinear optics effects. Since 1985 after the CPA invention followed by the successive laser short pulse improvements, we are experiencing a similar increase in intensity opening a new regime in optics at 10^{18} W/cm^2 dominated by the relativistic character of the electrons.

the electric one. It pushes the electron forward and is the tenet of the relativistic regime.

At this point, it is convenient to introduce the normalized vector potential:

$$a_0 = \frac{e\lambda E}{m_0 c^2} \qquad [6]$$

which represents the energy gain by the electron over one period (λ) normalized to the electron rest mass energy. At $a_0 = 1$, corresponding to 10^{18} W/cm^2 for 1 μm wavelength light, the laser–matter interaction is strongly dominated by the relativistic character of the electrons. At these intensities, the motion of the electrons becomes relativistic and the trajectory of the electron dramatically changes. Figure 2 depicts the difference between the two regimes. It can be shown that independently of the intensity regime, the electron transverse motion is always proportional to a_0, while the longitudinal motion is always proportional to a_0^2. This means that for classical

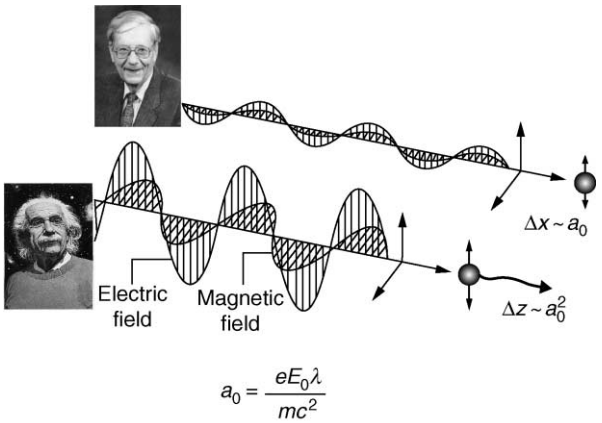


Figure 2 A comparison between classical optics and relativistic optics. The transverse electron displacement scales as a_0 and the longitudinal scales as a_0^2. (a) In classical optics $a_0 < 1$ the transverse electron displacement is proportional to a_0. It is larger than the longitudinal proportional to a_0^2. (b) In relativistic optics it is the opposite, $a_0 > 1$ the transverse displacement becomes smaller than a_0^2.

optics or even bound-electron nonlinear optics with a_0 varying from 10^{-10} to 10^{-4}, the longitudinal motion is $\sim 10^{-10}$ to 10^{-4} smaller than the transverse motion. Of course, the situation reverses completely for $a_0 > 1$, where the longitudinal motion becomes larger than the transverse motion. The departure of the electron motion from the linear regime, due to its relativistic character, gives rise to a new kind of nonlinear optical effects that constitutes the field of relativistic nonlinear optics. This area has become one of the most active in laser physics. In many cases, there are strong similarities between conventional and relativistic optics, and it is the purpose of this article to illustrate the field of relativistic optics by using some of the effects well established in bound electron nonlinear optics.

Considering the rapid progress in laser intensities that the field experienced in the 1960s, as indicated in Figure 1, scientists expected to observe the manifestation of relativistic effects in the 1970s. Formidable intensities on the order of 10^{18} W/cm^2 for 1 μm wavelength light were necessary, however, and could not be readily obtained. Such intensity levels could only be produced in the early 1980s with large-scale sources such as the Helios laser at Los Alamos National Laboratory (LANL) (Figure 3a). Work in relativistic optics virtually could not be conducted in university research laboratories, but rather only in government laboratories. The first relativistic harmonics produced by the critical surface nonlinear motion were observed in 1980 with a kJ-pulse-energy laser at mHz repetition rates (one shot/20 m).

The advent of CPA (Figure 4) revolutionized the field of relativistic optics, placing it within reach of more researchers through enormous reductions in laser size and cost. Further advances in the ultrashort-pulse duration of lasers and in laser focusing made relativistic optics more accessible, especially to universities. Relativistic intensities can be achieved with mJ energy levels at kHz repetition rates by university-size lasers (Figure 3b) instead of with large, kJ, sub-Hz-repetition-rate government-type lasers. Figures 3a and b illustrate the revolution in the relativistic laser evolution representing the Helios (LANL) system. Figure 3b depicts the relativistic λ^3 laser target chamber. The λ^3 laser reaches its relativistic intensity (10^{18} W/cm^2) with pulse energy only on the order of 1 mJ and pulse duration on the order of 10 fs, i.e., a few light periods. Its energy is concentrated on a single wavelength spot size using an f#1 paraboloid corrected by a deformable mirror. Note that all the energy is contained in a focal volume of a few λ^3, hence its name.



(a)



(b)

Figure 3 Two relativistic lasers. (a) Helios at LANL, was the first relativistic laser with $a_0 \sim 1$ at mHz repetition rate. (b) The λ^3 at the University of Michigan has an $a_0 \sim 1$ at kHz repetition rate.

Some Examples of Relativistic Nonlinear Optics

We will illustrate the field of relativistic optics by comparing relativistic effects with some well-known nonlinear effects.

(a) Relativistic optical rectification (laser wake-field acceleration of electrons and ions)

One well-known bound-electron optical effect coming from the second-order term in eqn [1] is optical rectification. This effect, which takes place in noncentrosymmetric crystals, gives rise to a DC field transverse to the propagation direction. In the relativistic regime, a similar effect known as wake-field acceleration will also contribute to rectification of the optical field. We have seen that in vacuum, electrons are strongly pushed forward by $v \times B$ force. In a plasma, the electromagnetic wave will also push

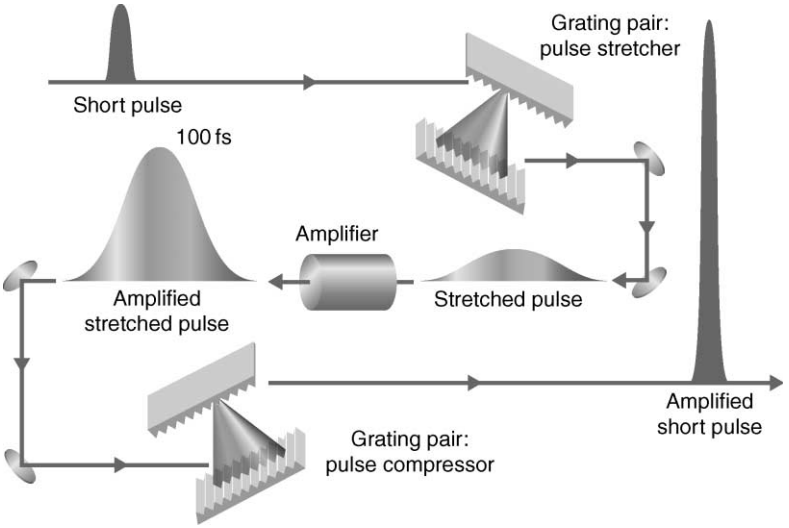


Figure 4 The concept of CPA is the key technique to get relativistic intensity. An initial short pulse is first stretched out (chirped) by a factor $\sim 10^4$, for instance by a pair of diffraction gratings. The stretched pulse intensity becomes lowered by the same factor of $\sim 10^4$. The stretched pulse intensity can then be amplified to a level 10^4 times higher. Once the laser energy has been extracted from the amplifier the stretched pulse is recompressed by the same factor ($\sim 10^4$) to its initial value. This way the laser intensity has been boosted to 10^4 times the peak intensity that the amplifier could have withstood.

forward the electrons, dragging the much heavier positive ions behind, as shown in Figures 5a and b. The charge separation that ensues will produce a strong electrostatic field that can be of the order of the transverse light electric field, which is huge. Unlike in classical optical rectification, in the relativistic regime the effect can be very efficient. Of the order of 10% or more of the transverse field may be rectified. For an intensity of 10^{18} W/cm^2, the electric field can be on the order of 10^{12} V/m (CERN on one meter), or 10^5 times larger than the field that can be obtained by more conventional technology. In addition, the rectified field is flipped along the longitudinal direction. Note also that since the field is created within a plasma, the medium cannot be broken down further and can support these extremely high fields. This effect had to wait for the relativistic-intensity lasers to become a reality before being demonstrated. Today energies up to 200 MeV have been demonstrated by using a 50 TW laser focused on a gas jet. Because the electrons drag the ions, a beam of ions is also observed concomitant with the electron beam. The generation of protons with energy up to 56 MeV has been reported by LLNL.

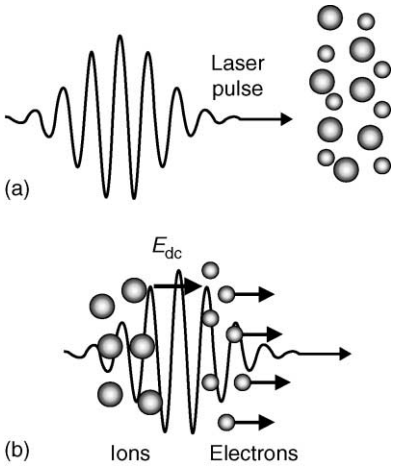


Figure 5 Relativistic rectification in plasma. (a) The high intensity pulse before entering the plasma. (b) Illustrate the $v \times B$ pushing first the plasma electrons. The electrons then drag the heavy ions behind them like a horse pulling a cart. The longitudinal electrostatic field that is generated is almost as large as the transverse laser field.

(b) Relativistic harmonic generation

Harmonic generation in the relativistic regime can be produced in two ways, as a result of the laser plasma interaction with electron density n_e, either below or above the critical density n_c. The first is for $n_e < n_c$ and corresponds to an underdense plasma where light can propagate. Harmonics generation was predicted in the late 1960s. Even-order harmonics are observed with linear and circular polarizations. This is the signature of relativistic effects since even-harmonics cannot occur in conventional nonlinear optics in centrosymmetric media.

Harmonic generation can also be produced in over-critical plasmas, where $n > n_c$. In this case, the light cannot propagate and be reflected. The strong $v \times B$ oscillatory force applied to the critical surface will produce a self-modulated reflected beam. This effect was first observed using the Helios laser. The idea of the moving-mirror model was first put forward in the early 1980s. Generation of frequencies up to the 70th harmonic has been demonstrated so this method could provide an efficient way to produce high harmonics. This technique has the advantage over conventional high harmonic generation in that it is capable of accommodating a pulse of large energy, since one can focus the beam over a small area, limited only by the laser wavelength, without fear of disrupting the plasma.

(c) Relativistic self-focusing in refraction

In bound-electron nonlinear optics, the third-order term in eqn [1] is responsible for self-focusing. A similar effect also exists in relativistic optics; it is described in Figure 6a. In the relativistic regime, the electron driven by the laser field changes its mass. In turn, this mass change will modify the index of refraction according to:

$$n^2 = 1 - \left(\frac{\omega_p}{\omega}\right)^2 \qquad [7]$$

where the plasma frequency is given by

$$\omega_p^2 = \frac{e^2}{m} n_e \qquad [8]$$

where n_e is the plasma electron density, and m is the intensity-dependent mass according to eqn [1].

For $\omega > \omega_p$, the index of refraction is positive and the pulse can propagate in the plasma. For the higher-beam-intensity part of the beam, i.e., on the beam axis, the plasma frequency will be less than the frequencies corresponding to the pulse wing. This will cause a modification of the laser wavefront and the beam to self-focus.

(d) Relativistic self-focusing in reflection and isolated attosecond pulse generation

For $\omega < \omega_p$, (eqn [7]) is negative and the pulse cannot propagate and so is reflected at the critical surface. However, the large light pressure will push the plasma critical-density surface at twice the laser frequency, making it cancel out the beam spatial intensity profile. In the relativistic λ^3 regime, the laser is focused at one wavelength. If the light is linearly polarized, the electron density will be collectively moved back and forth along the critical surface. This time-dependent deformation of the critical surface will produce a deflection of each electromagnetic half-cycle in a specific direction (Figure 7), providing a simple means for generating isolated attosecond pulses efficiently. The relativistic motion of the critical surface compresses (relativistic Doppler) the incident femtosecond pulse to the attosecond regime. To understand this compression, recall that the incident pulse travels opposite to the mirror motion formed by the critical surface (for a fraction of the cycle). As we have seen in this regime the critical surface moves also at close to the speed of light, v. Since the reflected light must also travel at c, it has no choice but to be compressed, producing an electromagnetic accumulation along the critical surface. The compression factor will

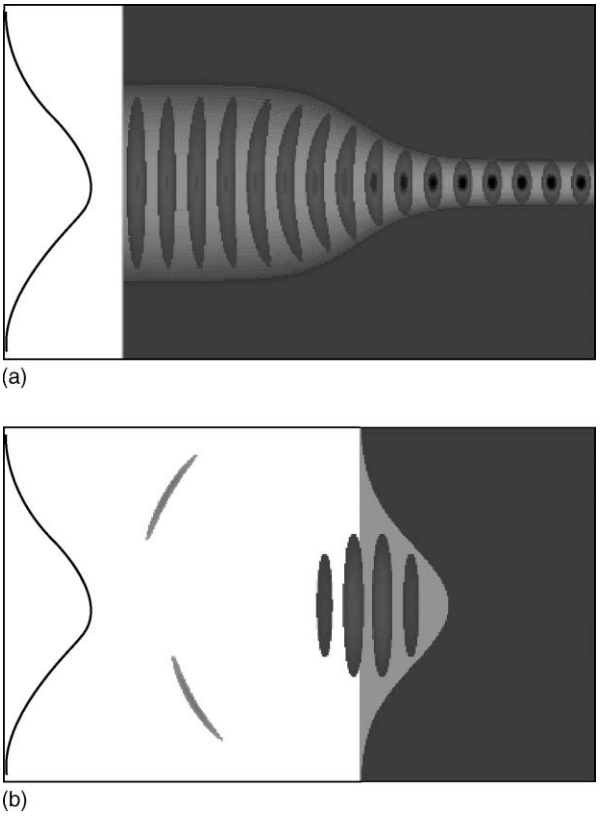


Figure 6 Relativistic self-focusing. (a) Relativistic self-focusing in refraction. Here the index of refraction is modified by the change in the electron mass. The electron mass change will in turn modify the plasma frequency across the beam and consequently the index of refraction. (b) Relativistic self-focusing in reflection showing the deformation of the critical surface by the light pressure from a strongly focused beam. The light is reflected in a series of ultrashort pulses in the attosecond domain.

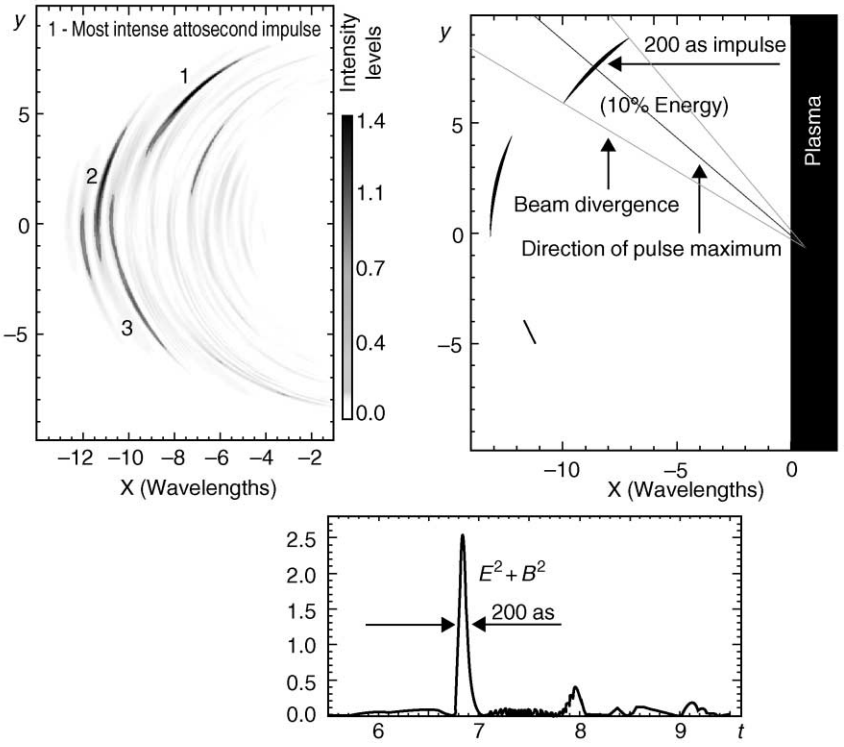


Figure 7 Compressed and deflected attosecond pulses. The combination of relativistic deflection and compression leads to 200 as isolated pulses.

depend on the critical surface velocity and is dependent on the plasma stiffness, i.e., the plasma density. Low-density plasmas are softer than high-density ones; here, n_c is the plasma critical density. The compressed pulse duration has been determined by a particles-in-cell (PIC) simulation to be given by $\tau = 600/a_0$. From the PIC simulation, it can be determined that the compressed pulse optimum duration is obtained for $n_e = a_0 n_c/2$. This technique is an interesting alternative to the one based on high-harmonic generation due to bound electrons, which has been recently demonstrated. The advantage of the relativistic optics approach is that it can be scaled up to large energies, from the mJ to the several-joule level, in contrast with demonstrated attosecond schemes that can only compress mJ pulse energies with an efficiency of 10^{-6}. For $a_0 \sim 100$, corresponding to incident intensities of $\sim 10^{22}$ W/cm^2, a 10-as pulse can be produced by relativistic compression and deflection.

(e) Nonlinear QED on a table top

Looking at Figure 1, we see that there is a regime around 10^{29} W/cm^2 where it becomes possible to produce electron–positron pairs. The condition to fulfill this is given by

$$E_s \lambda_c e = 2 m_0 c^2 \qquad [9]$$

Here $\lambda_c = h/m_0 c$ is the Compton wavelength, E_s the Schwinger field, and h the Planck constant. This expression states that for pair creation the work of the laser field over a Compton wavelength must be equal to the electron–positron rest-mass energy. This field E_s, also known as the critical field, $E_s \sim 10^{16}$ V/cm, corresponds to a laser intensity of 10^{29} W/cm^2.

An alternative way to depict the process of electron–positron generation from vacuum is by considering the vacuum as a dielectric medium with a gap $E_G = 2 m_0 c^2$. The field to produce an electron–positron pair is the one necessary to bend the bands so that an electron can tunnel from the Dirac sea into the vacuum states (Figure 8). The probability for such an event is given by

$$W \propto \exp\left(-\frac{\pi E_s}{E}\right) \qquad [10]$$

Today, the intensity necessary to reach pair creation is more than six orders of magnitude higher

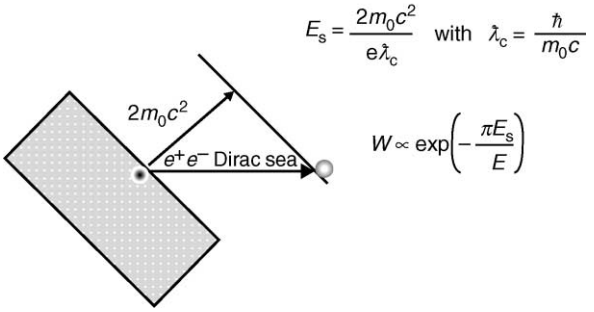


Figure 8 Pair creation. Electron tunneling from the Dirac sea into the vacuum state.

than what we could produce, $\sim 10^{22}$ W/cm^2. To bridge this enormous gap the electric field has been enhanced by colliding a short and intense pulse with a Stanford Linear Accelerator Center (SLAC) super-relativistic, 50 GeV electron beam. The Lorentz factor is $\gamma = 10^5$, and the laser field of 210^{10} V/cm (corresponding to an intensity of 10^{18} W/cm^2) can be enhanced by 10^5 to reach the critical field.

The progress in relativistic compression described above could make possible the generation of electron–positron pairs directly from the laser. We are on the verge of having a tabletop laser that could produce an intensity close to 10^{23} W/cm^2, corresponding to $a_0 > 100$. From the scaling law described previously, we should get $\tau < 6$, where the wavelength of such a compressed pulse would also be reduced by $\sim$100. The focal volume would then be reduced by six orders of magnitude. If we assume an efficiency of 10^{-1}, an intensity close to the critical intensity of 10^{29} W/cm could be reached with a tabletop laser and boil vacuum.

Conclusion

Today, laser intensities are such that during light–matter interaction, the electron motion is dominated by relativistic effects responsible for a new type of nonlinear optics called relativistic nonlinear optics. Relativistic nonlinear optics is characterized by extreme field gradients (TV/m) and light pressure in the Gbar regime. Relativistic nonlinear optics offers an attractive and novel way for efficient attosecond pulse generation with intensities near the critical (Schwinger) field, opening optics to vast new horizons.

See also

Relativistic Nonlinear Optics.

Further Reading

Agostini P, Barjot G, Bonnal JF, Mainfray G and Manus C (1968) *IEEE Journal of Quantum Electronics* QE-4: 667.

Albert O, Wang H, Liu D, Chang Z and Mourou G (2000) *Optics Letters* 125: 1125.

Bloembergen N (1965) *Nonlinear Optics*. New York: Benjamin.

Boyd RW (1992) *Nonlinear Optics*. San Diego: Academic.

Brabec T and Krausz F (2000) *Review of Modern Physics* 72: 545.

Brezin E and Itzykson C (1970) *Physics Review D* 2: 1191.

Bula C, McDonald KT, Prebys EI, *et al.* (1996) *Physics Review Letters* 76: 3116.

Bulanov SV, Naumova NM and Pegoraro F (1994) *Physics Plasmas* 1: 745.

Carman RL, Benjamin RF and Rhodes CK (1981) *Physics Review A* 24: 2649.

Chen S-Y, Maksimchuk A, Esarey E and Umstadter D (2000) *Physics Review Letters* 84: 5528.

Eberly J (1969) *Progress in Optics* 7: 361.

Franken PA, Hill A, Peters C and Weinreich G (1961) *Physics Review Letters* 7: 118.

Hentschel M, Kienberger R, Spielmann Ch, *et al.* (2001) *Nature* 414: 509.

Joshi C and Katsouleas T (2003) Plasmas accelerators at the energy frontier and on tabletops. *Physics Today* 56: 47.

Litvak AG (1969) *Soviet Physics JETP* 30: 344.

Malka V, Fritzler S, Lefebvre E, *et al.* (2002) Electron acceleration by a wakefield forced by an intense ultrashort laser pulse. *Science* 22: 298.

Max C, Arons J and Langdon A (1974) *Physics Review Letters* 33: 209.

Mourou GA, Barty CPJ and Perry MD (1998) *Physics Today* 51(1): 22.

Naumova NM, Nees JA, Sokolov IV, Hou B and Mourou GA (2004) Relativistic generation of isolated attosecond pulses in a λ^3 focal volume. *Physics Review Letters* 92: 063902.

Norreys PA, *et al.* (1996) *Physics Review Letters* 76: 1832.

Sarachik E and Schappert G (1970) *Physics Review D* 1: 2738.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: Wiley.

Snavely RA, *et al.* (2000) *Physics Review Letters* 85: 2945.

Strickland D and Mourou G (1985) *Optics Communications* 55: 219 and 447.

Tajima T and Dawson JM (1979) *Physics Review Letters* 43: 262.

Tarasevitch A, Orisch A, von der Linde D, *et al.* (2000) *Physics Review A* 62: 023816.

Umstadter D, Chen S-Y, Maksimchuk A, Mourou G and Wagner R (1996) *Science* 273: 472.

NONLINEAR OPTICS, APPLICATIONS

Contents

Phase Matching
Pulse Compression via Nonlinear Optics
Raman Lasers
Self-Focusing and Related Effects (Solitons and Multiphoton Absorption)
Three-Dimensional Microfabrication

Phase Matching

A V Smith, Sandia National Laboratories, Albuquerque, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Phase matching is important in any process that coherently transfers energy from one optical mode to another over a long interaction length. We will start by illustrating the principle, by imagining that the original mode is a single transverse mode of a waveguide, and that microscopic scattering centers distributed uniformly throughout the waveguide weakly scatter light into a second mode with orthogonal polarization. Then suppose that the scattering process preserves phase so the phase of the wavelet scattered at each point into mode two is the same as that of the original wave in mode one. Such a process is shown in Figure 1 where the propagation constant of the initially excited mode is k_1 and the propagation constant of the second mode is k_2. At each z location the light scattered into the second mode has the phase $\phi = k_1 z$, and after propagating to the end of the waveguide ($z = L$) with the phase velocity characteristic of the second mode, it has the phase $\phi_z = [k_1 z + k_2(L - z)] = [k_2 L + z(k_1 - k_2)]$. Coherently summing the fields scattered at all of the z positions, gives the net output field in the second mode. Figure 2 shows how this field can be constructed graphically. The rotation angle of each phasor represents its phase, ϕ_z, while the length of each phasor represents the magnitude of the field contribution from the corresponding z segment. It is apparent that the energy transfer is maximized if the contributions from all z positions have the same phase, in which case the phasors form a straight line, and the power in the second mode is proportional to L^2. This corresponds to $(k_1 - k_2 = 0)$. When this condition is met the waves in the two

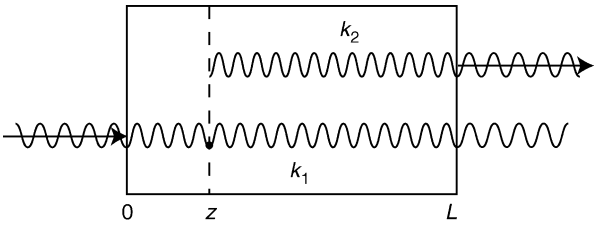


Figure 1 An input wave with propagation constant k_1 in a waveguide mode is weakly scattered into a second waveguide mode with propagation constant k_2. The phase mismatch is $\Delta k \equiv k_2 - k_1$.

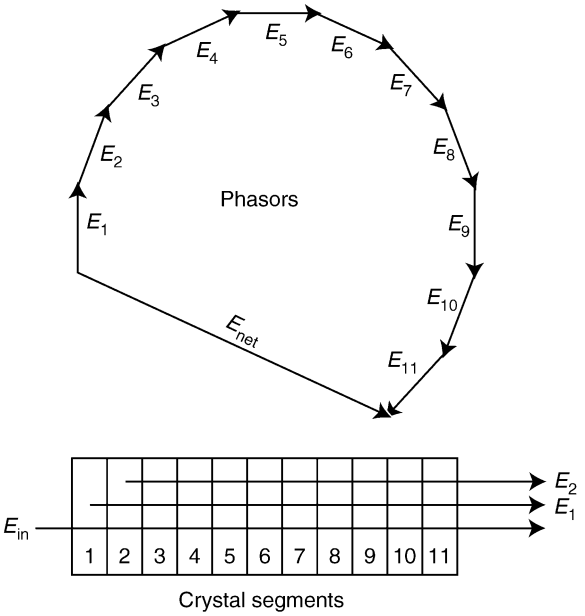


Figure 2 Phasor diagram showing the net output field contributions from segments 1–11. The net field at the output end, E_{net}, is the vector sum of the phasors. The phasors are complex quantities whose angles represent their relative phases according to $E_n = E_o \exp(i\phi_n)$ where $\phi_n = \phi_o - \Delta k(z_n - z_1)$.

modes have equal phase velocities, and the process is phase matched.

The quantity $(k_2 - k_1)$ is called Δk, the phase mismatch. If $\Delta k \neq 0$ the phasors curl to form a section of a circle, reducing the net field in mode two. If all the phasors have the same length, as they would if the wave in mode one is undepleted, the size of the mode two field at position z is

$$E(z) = Kz\frac{\sin(\Delta kz/2)}{\Delta kz/2} = Kz\,\mathrm{sinc}\left(\frac{\Delta kz}{2}\right) \qquad [1]$$

where K is a constant describing the scattering strength. The corresponding power in mode two is plotted in Figure 3 for various values of Δk. When $\Delta k = 0$ the power grows quadratically with z, but for $\Delta k \neq 0$ it is proportional to $\sin^2(\Delta kz/2)/\Delta k^2$. If we plot the normalized output power in mode two ($z = L$) versus Δk, we have the curve shown in Figure 4 which is $\mathrm{sinc}^2(\Delta kL/2)$.

When $\Delta k = 2\pi/L$ the phasors form a complete circle so the power in mode two first grows with z but then shrinks back to zero at $z = L$. For a given value of Δk the length for which the phasors form a complete circle is known as the coherence length, $L_{\mathrm{coh}} = 2\pi/\Delta k$.

Reducing the z increments to infinitesimals, we can write the output field for mode two in integral form:

$$E_{\mathrm{net}} = \int_0^L S(z)e^{-i\Delta kz}\,\mathrm{d}z \qquad [2]$$

This shows that when Δk is independent of z, the output field is the Fourier transform of a source term, $S(z)$, transformed into Δk space. Where $S(z)$ has a constant value from $z = 0$ to $z = L$ and is zero outside that range, the Fourier transform has the form given by eqn [1], with $z = L$.

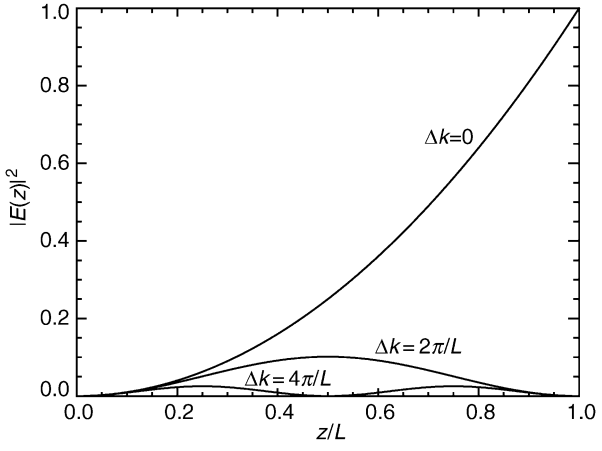


Figure 3 Plot of power in second mode versus z for different values of Δk. For $\Delta k = 0$ the power is proportional to z^2. For $\Delta k = 2\pi$/L the phasors form a closed circle so the net field is zero at $z = L$. For $\Delta k = 4\pi$/L the phasors trace two complete overlapped circles so again the field is zero at the $z = L/2$ and at $z = L$.

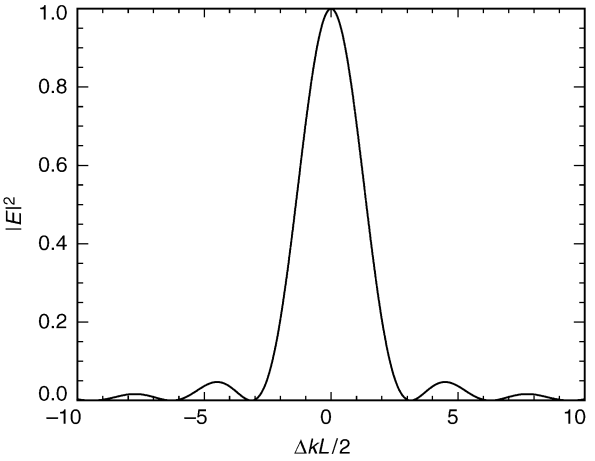


Figure 4 The plotted function indicates the relative irradiance of the output field, $|E|^2$, versus the phase mismatch Δk. The full width at half maximum is $\Delta k = 1.77\pi/L = 5.57/L$.

The $L^2\,\mathrm{sinc}^2(\Delta kL/2)$ form for the output power is ubiquitous in functions that describe the efficiency of nonlinear optical interactions, as well as acousto-optic and electro-optic modulation of light waves. However, it should be kept in mind that $L^2\,\mathrm{sinc}^2(\Delta kL/2)$ holds only when $S(z)$ is independent of z. If the scattering is strong enough to alter the phase or amplitude of the driving wave, or if the coupling strength varies with z for any other reason, $S(z)$ is not independent of z and the $L^2\,\mathrm{sinc}^2(\Delta kL/2)$ dependence is modified.

Our simple example relied on scattering in a waveguide, with a source term that was linear in the field in mode one. In practice, most optical devices that couple light between modes, rely on nonlinear interactions among multiple optical fields, or between an optical field and an acoustic or an rf electric wave. These nonlinear interactions produce a polarization wave in a suitable material at a sum or difference frequency. Each point of this polarization radiates a wavelet in all directions. These wavelets interfere constructively only in directions that are nearly phase matched. Examples include traveling wave acousto-optic filters, in which an acoustic wave traveling parallel to an input light beam of one polarization in a birefringent material, couples light into the orthogonal polarization via the elasto-optic effect. Traveling wave electro-optic modulators are similar but rely on the linear electro-optic effect for the coupling. Optical frequency mixing uses the nonlinearity of various

materials to mix multiple optical waves, creating polarization waves at sum and difference frequencies. A wide variety of such nonlinear interactions is discussed in textbooks on nonlinear optics, electro-optics, and acousto-optics. In all of these cases the phase velocity mismatch of interest is that between an induced polarization wave and a radiated mode, and the phase mismatch is $\Delta k = k_{\text{radiation}} - k_{\text{polarization}}$.

To find $S(z)$ and Δk for these nonlinear mixing processes, we must construct the polarization wave. We consider only monochromatic plane waves here, considering beams with spatial or temporal structure in a later section. The electric field of a monochromatic plane optical wave propagating in direction $\mathbf{r}_o$ can be described by

$$\begin{aligned} E &= E_o \cos(\omega_o t - \mathbf{k}_o \cdot \mathbf{r}_o + \phi_o) \\ &= \tfrac{1}{2}[E\, e^{-i(\omega_o t - \mathbf{k}_o \cdot \mathbf{r}_o)} + E^*\, e^{i(\omega_o t - \mathbf{k}_o \cdot \mathbf{r}_o)}] \end{aligned} \quad [3]$$

The complex notation is a convenience that simplifies algebraic book-keeping. Similarly, an acoustic field propagating in direction $\mathbf{r}_a$ can be written:

$$\begin{aligned} A &= A_o \cos(\omega_a t - \mathbf{k}_a \cdot \mathbf{r}_a + \phi_a) \\ &= \tfrac{1}{2}[A\, e^{-i(\omega_a t - \mathbf{k}_a \cdot \mathbf{r}_a)} + A^*\, e^{i(\omega_a t - \mathbf{k}_a \cdot \mathbf{r}_a)}] \end{aligned} \quad [4]$$

In a suitable material the polarization induced by the product of the optical and acoustic waves has a term of the form:

$$P = sEA \quad [5]$$

P has components at both the sum and difference frequency. The sum frequency component is:

$$P_{\omega_o+\omega_a} = \frac{1}{4} s[EA\, e^{-i(\omega_o+\omega_a)t + i(\mathbf{k}_o+\mathbf{k}_a)\cdot\mathbf{r}} + \text{cc}] \quad [6]$$

where 'cc' indicates the complex conjugate of the first term. The component at the difference frequency is

$$P_{\omega_o-\omega_a} = \frac{1}{4} s[EA^*\, e^{-i(\omega_o-\omega_a)t + i(\mathbf{k}_o-\mathbf{k}_a)\cdot\mathbf{r}} + \text{cc}] \quad [7]$$

An important point is that the k vector of the sum-frequency polarization is $(\mathbf{k}_o + \mathbf{k}_a)$. Phase matching requires that the radiated sum-frequency wave has the same propagation vector. The radiated difference frequency wave must have the propagation vector $(\mathbf{k}_o - \mathbf{k}_a)$ to be phase matched. Phase matching for these two cases is shown in Figure 5. Note the general rule that if the frequency of a wave appears as $+\omega$ in the induced polarization, the associated k vector contribution to the polarization wave is $+\mathbf{k}$, while if it appears as $-\omega$, the k vector contribution is $-\mathbf{k}$. For example, the acoustic wave in $P_{\omega_o+\omega_a}$ is associated with $+\mathbf{k}_a$ contribution to the polarization wave, while in $P_{\omega_o-\omega_a}$ it is associated with a $-\mathbf{k}_a$ contribution.

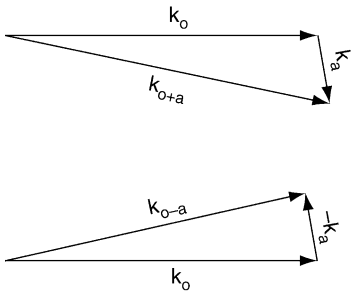


Figure 5 An optical wave with propagation vector k_o and an acoustic wave with propagation vector k_a produce a polarization wave at the sum frequency with the propagation vector k_{o+a} (top diagram), and a polarization wave at the difference frequency with the propagation vector k_{o-a} (bottom diagram).

Using the definition of k:

$$\mathbf{k} = \frac{2\pi}{\lambda}\hat{\mathbf{k}} = \frac{\omega}{v}\hat{\mathbf{k}} \quad [8]$$

where v is the phase velocity, we can write Δk as

$$\begin{aligned} \Delta\mathbf{k} &= \mathbf{k}_{o+a} - \mathbf{k}_o - \mathbf{k}_a \\ &= \left(\frac{\omega_{o+a}}{v_{o+a}}\hat{\mathbf{k}}_{o+a} - \frac{\omega_o}{v_o}\hat{\mathbf{k}}_o - \frac{\omega_a}{v_a}\hat{\mathbf{k}}_a\right) \end{aligned} \quad [9]$$

Clearly, if all of the waves had equal phase velocities, phase matching would be achieved for collinear interactions, where all of the propagation vectors are parallel, because the frequencies inside the parentheses sum to zero. This is seldom the case as for example, the phase velocity of an acoustic wave is much less than that of an optical wave, making an acoustic wave's k-vector disproportionately long compared with that of an optical wave. On the other hand, an rf wave usually has a disproportionately short k-vector compared with an optical wave. Even when all of the waves are optical, dispersion of the refractive index usually makes the k-vector of the bluer waves disproportionately long compared with those of the red waves. A wide variety of methods are used separately or in combinations to compensate for these effects in order to achieve phase matching and these are discussed below.

Phase Matching Techniques

Noncollinear Phase Matching

If one of the low frequency waves in a three wave mixing process, such as the acousto-optic interaction discussed above, has a lower phase velocity than the bluest wave, then noncollinear phase matching may be possible. Acousto-optic deflectors are often phase matched with a noncollinear geometry, such as that shown in Figure 6. An input optical wave of frequency ω_o reflects from a traveling acoustic wave with frequency ω_a in an optically isotropic material to produce an output optical wave of frequency $\omega_o + \omega_a$. The Bragg condition for efficient reflection is

$$\sin\theta = \frac{k_a}{2k_o} \qquad [10]$$

which may be seen from the figure to be the same as the phase matching condition, assuming k_{out} and k_{in} have the same magnitude, which is a good approximation for typical acoustic frequencies less than a few GHz.

Stimulated Brillouin scattering is another example of noncollinear mixing of acoustic and optical waves. One possible geometry has the input optical and acoustic waves traveling in opposite directions to create a reflected optical wave at the sum frequency. The phase matching condition is then

$$k_{acoustic} \approx 2k_{optical} \qquad [11]$$

which implies

$$\omega_{acoustic} = \frac{2n\omega_{optical}}{c} v_{acoustic} \qquad [12]$$

The acoustic frequency that phase matches such counter propagating Brillouin scattering is usually a few GHz in liquids and solids.

If noncollinear phase matching in an isotropic material is not possible, as is usually the case for electro-optic modulation and for all-optical sum and difference frequency mixing, it is sometimes possible to use the properties of nonisotropic media to adjust the phase velocity of one or more of the waves. There are several possibilities, such as natural birefringence, form birefringence, photonic band edge tuning, wave guide modal propagation, and various forms of quasi phase matching. Each is briefly described below.

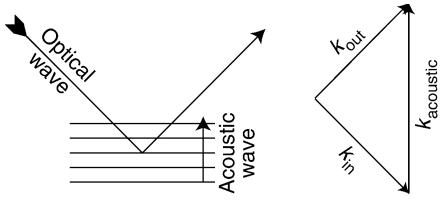


Figure 6 Phase matching in an acousto-optic deflector.

Birefringent Phase Matching

Natural birefringence

Naturally birefringent crystals can be categorized as either uniaxial or biaxial. In either case, light propagating along any direction (except along an optical axis) has two orthogonal eigen polarizations with different refractive indices. In uniaxial crystals, one of these indices is independent of the propagation direction and is called the ordinary index, n_o. The refractive index for the other polarization, known as the extraordinary polarization, is given by

$$\frac{1}{n_e^2(\theta)} = \frac{\sin^2\theta}{n_e^2} + \frac{\cos^2\theta}{n_o^2} \qquad [13]$$

where θ specifies the propagation direction measured from the z, or optical, axis, and n_e and n_o are the refractive indices for light polarized parallel and perpendicular to z. The two eigen polarizations are aligned with the lines of constant or latitude (o-wave) or constant longitude (e-wave). The variation of the two refractive indices with propagation angle θ is shown in Figure 7 for a negative uniaxial crystal which is characterized by $n_e < n_o$, and for a positive uniaxial crystal, characterized by $n_e > n_o$.

The acousto-optic tunable filter relies on natural birefringence to achieve a phase match. An optical and an acoustic wave propagate normal to the optical axis of a uniaxial crystal. If the crystal is negative uniaxial, with an input optical polarization of 'e', and an output optical polarization of 'o', the k vector of the acoustic wave must make up the difference

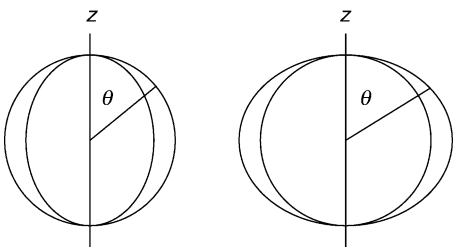


Figure 7 The circle and ellipse represent the two refractive indices for a negative uniaxial crystal (left diagram) and a positive uniaxial crystal (right diagram) for different propagation directions specified by the angles (θ, ϕ) with ϕ fixed. These curves are independent of the azimuthal angle ϕ, so in three dimensions the ordinary refractive index, n_o, forms a sphere in both cases, while the extraordinary index, n_e, forms a prolate ellipsoid for a negative uniaxial crystal, or an oblate ellipsoid for a positive uniaxial crystal.

$k_o - k_e$. This may be expressed as

$$k_a = \frac{\omega_a}{v_a} = \frac{(\omega + \omega_a)n_o}{c} - \frac{\omega n_e}{c} \qquad [14]$$

or, ignoring the difference between ω and $\omega + \omega_a$, the optical frequency that is phase matched is

$$\omega = \frac{\omega_a c}{v_a \Delta n} \qquad [15]$$

where v_a is the phase velocity of the acoustic wave, and $\Delta n = n_o - n_e$. The phase matched optical frequency is seen to be tunable with ω_a. A typical acoustic frequency for a visible filter is a few hundred MHz. The phase matching bandwidth set by the $\text{sinc}^2(\Delta kL/2)$ function can be quite narrow, with a subnanometer optical bandpass in a crystal a few cm long.

Natural birefringence is also widely used to phase match all-optical sum and difference frequency mixing. To counter normal optical dispersion, the bluest wave is polarized in the direction associated with the smaller refractive index, while at least one of the two redder waves has the opposite polarization. By adjusting the propagation angle, θ, the refractive index of the extraordinary waves can be adjusted until exact phase matching is achieved. If the two redder waves both have the high index polarization the process is sometimes called type I, whereas if they have opposite polarizations it is called type II.

Figure 8 illustrates type I phase matched second harmonic generation in a negative uniaxial crystal. The fundamental wave with frequency ω is o-polarized, while the second harmonic wave at frequency 2ω is e-polarized. The circle drawn in dashes represents the line swept out by the polarization wave vector $2k_\omega$ as the propagation direction θ is swept, while the ellipse represents the loci of $k_{2\omega}$ for the same angle. Collinear phase matching is achieved at the point of intersection of the two curves. This sets the angle θ_{pm}. For angles greater than θ_{pm}, dispersion is over-compensated and $\Delta k < 0$ so noncollinear phase matching is possible. All ϕ angles have equivalent propagation properties, so the loci of phase matched propagation directions form a cone at $\theta = \theta_{pm}$. Usually the propagation angle ϕ is chosen to maximize the nonlinear response of the crystal.

Biaxial crystals share with uniaxial crystals the property of two orthogonal eigen polarizations with different refractive indices, but the angular dependence of the indices is more complex. The two indices associated with propagation along direction (θ, ϕ) can be found by solving

$$\frac{\sin^2\theta\,\cos^2\phi}{1/n^2 - 1/n_x^2} + \frac{\sin^2\theta\,\sin^2\phi}{1/n^2 - 1/n_y^2} + \frac{\cos^2\phi}{1/n^2 - 1/n_z^2} = 0 \qquad [16]$$

for the two possible values of n. Here n_x, n_y, and n_z are the refractive indices for light polarized along the respective principal axes, where the convention is $n_x < n_y < n_z$. There are two optical axes lying in the x–z plane at an angle determined by the values of n_x, n_y, and n_z. The phase matching loci are no longer lines of constant latitude as they were for uniaxial crystals, but instead are curves that encircle either the optical axes or one of the principal axes. Figure 9 shows an

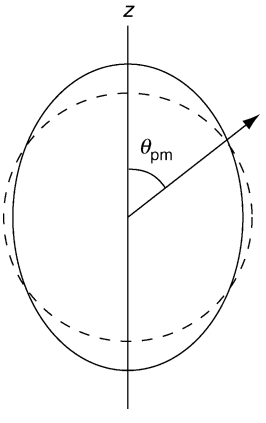


Figure 8 Collinear phase matching of second harmonic generation in a negative uniaxial crystal. The second harmonic light is e-polarized with a refractive index indicated by the solid line while the fundamental light is o-polarized as indicated by the dashed line. The crossing point of the two refractive indices determines the phase matching angle θ_{pm}.

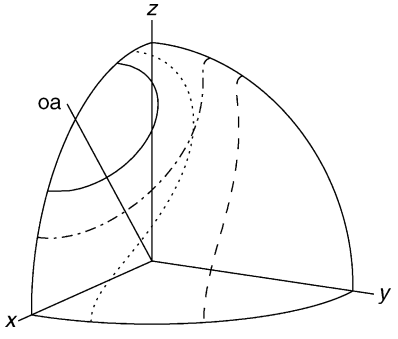


Figure 9 Calculated loci of propagation directions that achieve birefringent phase matching for type I collinear second harmonic generation of different wavelengths in the crystal CsB_3O_5. The fundamental wavelengths are 2,700 nm (dotted curve looping around the x axis), 3,000 nm (dashed curve looping around the y axis), 1,000 nm (solid curve looping around the optical axis 'oa'), and 800 nm (chained curve looping around the z axis). For clarity only a single octant of the sphere of propagation directions is shown. The loci in the other octants of the sphere are formed by reflections in the three principal planes.

example of each type of phase matching curve associated with second harmonic generation of different wavelengths in the crystal CsB_3O_5.

Note that because these curves always intersect at least one of the principal planes, if phase matching in biaxial crystals is possible at all, it is also possible in a principal plane. Biaxial crystals are usually phase matched in a principal plane, but phase matching outside the principal planes can in some cases maximize the nonlinear response, minimize the temperature sensitivity, or adjust the relative group velocities.

Form birefringence

Some materials, such as GaAs, have a strong nonlinear optical response but are not birefringent, so optical frequency mixing cannot be phase matched using natural birefringence. However, form birefringence can be engineered by layered growth, and it can be used to phase match. If many alternating layers of two materials with different refractive indices are stacked together with a layer thickness comparable to the wavelength of light, the material is birefringent for light propagating parallel to the layers. Light polarized with its E field parallel to the layers will have a higher refractive index than light polarized with its field perpendicular to the layers. The strength of birefringence, $(n_{\parallel} - n_{\perp})$, can be engineered by varying the relative thicknesses and refractive indices of the two layered materials. This has been demonstrated in GaAs/AlAs waveguide structures in which the AlAs layers were oxidized to form Al_2O_3, which has a refractive index of 1.6, compared with an index of 3.4 for GaAs. The resulting large birefringence was used to phase match infrared difference frequency mixing.

Photonic Lattice Phase Matching

The propagation constant of a light wave can be controlled by tuning its frequency near the bandgap of a Bragg structure. The Bragg structure can be a stack of layers with alternating refractive index. If the stack consists of layers of thickness and refractive index l_1, n_1 and l_2, n_2, the Bragg condition at normal incidence is

$$\phi_1 + \phi_2 = \frac{\omega_{\text{Bragg}}}{c}(n_1 l_1 + n_2 l_2) = m\pi \qquad [17]$$

where m is an integer. Frequencies near ω_{Bragg} are strongly reflected. The width of the reflective band of a structure with many layers is proportional to the refractive index contrast, $(n_1 - n_2)/(n_1 + n_2)$. At certain frequencies, on both sides of the bandgap, the reflectivity vanishes. The first transmission peak on the red side of the bandgap has an effective refractive index that is greater than the average $(n_1 + n_2)/2$, while the first peak on the blue side has an effective index less than the average. This effect can be used to adjust the effective value of Δk in the structure. As a bonus, operating on one of the transmission peaks near the bandgap enhances the optical field in the material, increasing the strength of the nonlinear interaction.

It is also possible to use a Bragg structure in reflection mode. Tuning to the red side of the band center increases the effective refractive index of the reflected wave, while tuning to the blue side decreases it. The internal fields are not enhanced in this case and the interaction length may be quite short.

Quasi-Phase Matching (QPM)

Quasi-phase matching uses spatial modulation of the polarization source with a modulation period of $L_{\text{mod}} = 2\pi/\Delta k$ to boost the efficiency of the mixing process even though $\Delta k \neq 0$. We can consider the periodic modulation as another interacting wave that is included in the creation of the polarization wave. The modulation wave has $\omega = 0$ and is stationary with period L_{mod}. Because the modulation is usually a square wave rather than sinusoidal, it is made up of multiple sinusoidal waves with periods L_{mod}/m, all of which can participate in generating a polarization wave. Like traveling waves, the stationary modulation waves can be oriented in any direction, and any Fourier component of the modulation wave with wave vector $\Lambda_m = 2\pi m/L_{\text{mod}}$ can participate in (quasi) phase matching, as shown in Figure 10.

Usually the spatial modulation is achieved by periodically reversing the sign of the nonlinear coupling coefficient. Figure 11 shows the associated phasor diagram for perfect first-order QPM. The net fields from the first two modulation periods are labeled E_1 and E_2. If the sign of the nonlinearity in the first half of the modulation period were not reversed in the second half, the phasors would form a complete circle resulting in a null net field. All following periods would also produce null net fields. However, if the sign is reversed midway through the first period, the phasors from the second half are sign reversed and the net field is E_1 as shown. The magnitude of E_1 is smaller than for a truly phase matched segment by the

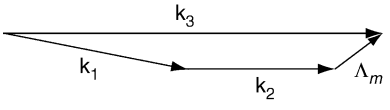


Figure 10 Quasi-phase matching vector diagram for sum frequency mixing. The mth Fourier component of the grating vector is indicated by Λm.

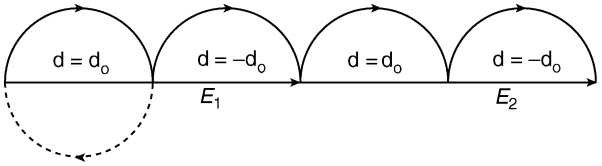


Figure 11 Phasor diagram for a quasi-phase matched interaction in a material with alternating signs of the nonlinear coefficient. The contributions from four segments, each of length $L_{coh}/2$ are shown. The net field from the first QPM period is labeled E_1, and that from the second period is E_2. The dashed line shows the path the phasors would trace if the nonlinear coefficient were not sign reversed in the even numbered segments.

ratio of the diameter of the circle to its circumference, or by $2/\pi$. Coherence lengths in crystals are typically 5–50 μm so crystals that are several mm long have a large number of periods, in which case the net phasors $E_1, E_2, \ldots$ can themselves be considered elementary phasors that form the net field for the full crystal. For example, if the modulation period L_{mod} is not exactly equal to $2\pi/\Delta k$, QPM is imperfect and phasors $E_1, E_2, \ldots$ curl up to give a net field of $\text{sinc}(\Delta k' L/2)$ where $\Delta k' = \Delta k - 2\pi/L_{\text{mod}}$. For all practical purposes, a QPM crystal containing many modulation periods performs the same as a phase matched crystal with its nonlinear coupling reduced by $2/\pi$.

Referring to eqn [2], we note that periodic modulation of the source term $S(z)$ produces peaks in the phase matching curves at $\Delta k = m2\pi/L_{\text{mod}}$ where m is an integer. If $S(z)$ is a square wave of alternating sign over $0 < z < L$, its Fourier transform is the convolution of $\text{sinc}(\Delta kL/2)$ with a picket fence function with spikes at $\Delta k = m2\pi/L_{\text{mod}}$. If $L_{\text{mod}} \ll L$, the crystal contains many periods of the square wave and the spikes are separated by many times the width of the $\text{sinc}(\Delta kL/2)$ function. This results in well separated sinc functions centered at $\Delta k = 2\pi m/L_{\text{mod}}$, and any process with a coherence length of L_{mod}/m will be quasi-phase matched. That is, one of the Fourier peaks in Δk space will be centered on the corresponding value of Δk.

The magnitudes of the QPM peaks are maximized if the modulation of $S(z)$ is a sign reversal, as described above. If the positive and negative segments have unequal lengths, we can define the duty cycle as $D = l_+/(l_+ + l_-)$. The heights of the pickets in the picket fence are then

$$h_m = \frac{2}{m\pi}\sin(\pi m D) \qquad [18]$$

The largest magnitude peak is for $m = 1$ and $D = 0.5$ in which case $h_1 = 2/\pi$.

Unlike birefringent phase matching, QPM can, in principle, phase match any set of polarizations for the interacting waves. This makes it possible to use a polarization set that maximizes the nonlinear response of the crystal.

Ferroelectric poling

There are various ways to achieve QPM. By far the most popular and versatile is by ferroelectric poling. Ferroelectric crystals have an internal electric field that can be reversed by the application of a sufficiently strong external poling field. Reversing the internal field also reverses the sign of the nonlinearity for some light polarizations but does not alter the refractive indices. Crystals of $LiNbO_3$ and $KTiOPO_4$, with thickness up to 1 mm, can be poled with periods as short as 5 μm. This is sufficient to quasi-phase match frequency mixing of visible or near infrared light. Figure 12 shows an example of the orientation of the internal fields of patterned ferroelectric domains.

It is possible to pole a crystal aperiodically so that multiple mixing processes are quasi-phase matched simultaneously. To design the poling pattern for such a device, the Fourier transform of an initial trial pattern can be calculated and compared with the target spectrum in Δk space. Iterative algorithms can be used to modify the poling pattern, zeroing in on the target spectrum.

Ferroelectric poling can also form two-dimensional poling patterns, corresponding to a two-dimensional set of Fourier components in Δk space. It can also produce chirped gratings which have a broad phase matching peak. Chirped gratings can be used to simultaneously frequency convert and tailor the temporal shapes of the output pulse when the input pulses are frequency chirped. Poling can also yield fanned poling patterns so the poling period and thus the phase matching wavelength can be shifted by translating the grating.

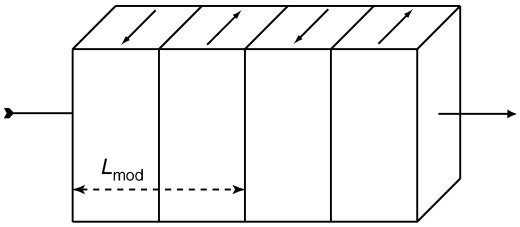


Figure 12 Diagram of a poled ferroelectric crystal. The four crystal segments represent four ferroelectric domains with internal crystal field directions indicated by the arrows. Reversal of the internal field is accompanied by reversal of the nonlinear coefficient, so this is a quasi-phase matched structure with alternating signs of d_{eff}.

Wafer stacking and patterned crystal growth

Another method of quasi-phase matching relies on stacks of thin plates oriented so the sign of the nonlinearity reverses in adjacent plates. Each plate has a thickness of $mL_{coh}/2$ where m is an odd integer. If all the plates are identically prepared crystals, it is often possible to reverse the sign of their nonlinearity by reorienting them. It is sometimes possible to optically contact the plates or even to diffusion bond them, avoiding reflective losses at the interfaces. This technique is conducive to large aperture QPM crystals. However, the plates are thin so fabrication is a challenge.

Quasi-phase matched crystals of GaAs have also been epitaxially grown on patterned substrates. The patterning serves as a template for growth of zones with alternating crystal orientation and sign of the nonlinear coefficient.

Total internal reflection

Total internal reflection in a thin slab has also been exploited for QPM (Figure 13). Two effects contribute to the source modulation, one is the rephasing of the waves on total internal reflection; the other is the change in nonlinearity that may accompany the change in propagation direction. Usually plates much thicker than one coherence length are used, so the quasi-phase matching is of high order ($m \gg 1$).

Wave Guide Modal Phase Matching

In waveguides the size and shape of the guide influence the transverse field distributions and the propagation constants for the transverse modes. Generally the higher order transverse modes have higher phase velocities than the lower order modes. This can be used to counteract index dispersion in three wave mixing if the bluest wave propagates in a higher order mode than the redder waves. A problem frequently encountered in using this effect is that the transverse modes are orthogonal, or nearly so. Because the nonlinear response is proportional to the overlap integral

$$d_{eff} \propto \int\int_A d(x,y)E_1(x,y)E_2(x,y)E_3(x,y)dx\,dy \qquad [19]$$

where the E's are the transverse field distributions of the three waves, the nonlinear response is often quite small. If the local nonlinearity $d(x,y)$ can be engineered to maximize the integral, this problem can be alleviated. This may be thought of as transverse quasi-phase matching.

Waveguides can also be designed for Cerenkov phase matching where the phase velocity of the radiated wave is less than that of the polarization wave. Figure 14 shows an example. The emission angle is

$$\cos\theta = \frac{2k_\omega}{k_{2\omega}} = \frac{v_{2\omega}}{v_\omega} \qquad [20]$$

where the v's are phase velocities. The transverse momentum Λ must be supplied by the waveguide. In other words, the source harmonic polarization must have a transverse Fourier component with spatial frequency Λ. This is another example of transverse quasi-phase matching.

Another application of waveguide phase matching is in traveling wave electro-optic modulation, where the rf wave travels along an rf waveguide parallel to the light path. For maximum modulation bandwidth, the phase velocity of the rf wave should equal that of the light wave over the desired modulation bandwidth. Because the phase velocity of the rf wave in the bulk medium is usually faster than that of the light wave, the rf guide must be designed for a slow phase velocity.

Compensated Phase Matching (CPM)

Compensated phase matching is somewhat like QPM in that multiple crystal segments are used, but it differs by requiring that sequential crystal segments have ΔkL's of equal magnitude but opposite sign. Ideally the nonlinearity would be the same in all segments. Figure 15 shows the phasor diagram for one such arrangement, in this case with unequal nonlinearity in alternating segments. CPM is realized

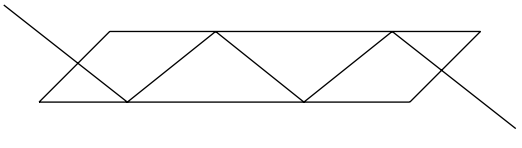


Figure 13 Diagram of quasi-phase matching based on total internal reflection. Relative phases acquired by the waves upon total internal reflection, combined with the path lengths and possible changes in the value of d_{eff} for alternating legs, can be combined to achieve quasi-phase matching.

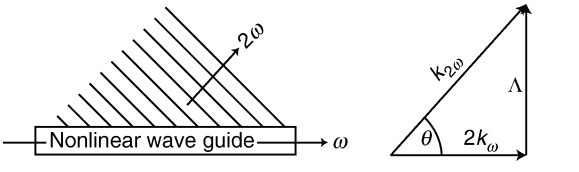


Figure 14 Cerenkov phase matched second harmonic generation. The phase velocity of the fundamental wave in the waveguide is greater than the phase velocity of the harmonic wave in the cladding.

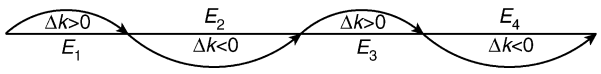


Figure 15 Phasor diagram for four segments of a device with compensated phase matching device. The magnitude of ΔkL is equal in each segment but the sign of Δk alternates. The length of the E_n field from segment n is proportional to $\text{sinc}(\Delta kL/2)$ so the net field ΣE_n must also be proportional to this function.

in walkoff compensated geometries where a sequence of crystals is arranged so $\Delta k = 0$ in each for a particular monochromatic plane wave. In the case of spatial walkoff compensation the walkoff direction in alternating crystals is reversed so the walkoff is exactly compensated in pairs of adjacent crystals. Tilting the input beams by a small amount changes the values of the Δk's in such a way that CPM is achieved. In the case of temporal walkoff compensation, temporal walkoff is reversed in alternating crystals, in which case tuning the input frequency leads to CPM. The advantage of CPM is that the acceptance bands, which will be discussed in the next section, are those of the individual crystal segments which are larger than those of a single crystal of equivalent length without walkoff compensation.

Spatial and Temporal Effects

Acceptance Bands

Our discussion to this point has been based on monochromatic plane waves. Real beams have spatial and temporal structure which means they can be decomposed into sets of monochromatic plane waves with tilted propagation directions relative to the nominal direction, and with detuned frequencies relative to the nominal frequency. As we showed above for monochromatic plane waves, there is a range of angles and frequencies, called the acceptance bands, over which $\Delta kL < 2\pi$. If the interacting beams' angular spectra and frequency spectra lie within these bands, the process is well phase matched for all angular and frequency components, and efficiency is maximized. This statement is equivalent to saying that if the spatial walkoff due to beam tilts and birefringence, and the temporal walkoff due to differing group velocities are small enough that the interacting beams stay well overlapped in space and time, the efficiency of energy transfer to the output channel does not suffer. Consider, for example, second harmonic generation of short pulses. The frequency acceptance bandwidth for the fundamental wave, $\Delta\omega_1$, is found using

$$\frac{\mathrm{d}(\Delta k)}{\mathrm{d}\omega_1}\Delta\omega_1 = \left(\frac{2\mathrm{d}k_2}{\mathrm{d}\omega_2} - \frac{2\mathrm{d}k_1}{\mathrm{d}\omega_1}\right)\Delta\omega_1 = \frac{2\pi}{L} \qquad [21]$$

Using the usual definition of the group velocity:

$$v_\mathrm{g} = \frac{\mathrm{d}\omega}{\mathrm{d}k} \qquad [22]$$

this can be written

$$\Delta\omega_1 = \frac{\pi}{L(1/v_{\mathrm{g}2} - 1/v_{\mathrm{g}1})} = \frac{\pi}{\tau_2 - \tau_1} = \frac{\pi}{\Delta\tau} \qquad [23]$$

where τ_1 and τ_2 are the passage times through a crystal of length L for the fundamental and harmonic pulses, and $\Delta\tau = (\tau_2 - \tau_1)$ is the temporal walkoff between the two. This shows that the acceptance bandwidth is approximately equal to the transform bandwidth of a pulse whose duration is equal to the temporal walkoff $\Delta\tau$. Input pulses shorter than $\Delta\tau$ will not convert efficiently because the fundamental and harmonic pulses do not remain overlapped as they propagate through the crystal. In the frequency domain it is said that pulses shorter than the $\Delta\tau$ have spectral widths larger than the acceptance bandwidth, meaning they are not well phase matched across their full spectra, and hence do not convert efficiently. Parallel arguments apply to spatial walkoff and angular acceptance bandwidths. Beams that do not remain spatially overlapped through the crystal do not convert efficiently. Equivalently, if the angular spectra of the beams exceed the acceptance angle, efficiency suffers.

Sometimes these notions are described in terms of an effective interaction length or an effective crystal length. For short pulses the effective interaction length is that for which temporal walkoff is equal to the pulse duration. Equivalently it is the length for which the acceptance bandwidth is equal to the spectral width of the pulse. Crystals shorter than this will have mixing efficiencies nearly equal to that expected for a phase matched process, while longer crystals have efficiencies that fall short of this expectation.

Focused beams

Focused Gaussian beams are a particularly important class of spatially structured beams. In passing through a focus, a lowest order Gaussian beam acquires a phase shift relative to an unfocused beam. The phase velocity near the focus is higher for the focused beam, leading to a phase shift of π through the focus. For second harmonic generation the phase shift of the second harmonic polarization is twice this, or 2π.

A second harmonic wave also acquires a phase shift of π, but this leaves a π phase shift between the harmonic polarization source wave and a freely propagating harmonic wave. The phasors representing harmonic radiation from symmetrically located points well before and well after the focus will therefore be 180° out of phase. Point pairs closer to the focus will have less severe cancelation, but it is not surprising that the maximum efficiency is achieved, not at $\Delta k = 0$, but for a slightly negative value Δk. The phase matching function $\mathrm{sinc}(\Delta kL/2)$ is shifted and distorted for focused beams. Its exact shape depends on the nature of the mixing process. For example, the phase matching peak is different for third harmonic generation and second harmonic generation, and it is different for mixing with small birefringence and mixing with large birefringence.

Phase Conjugation

Degenerate four-wave mixing

Some wave mixing processes are automatically phase matched. One example is phase conjugation by degenerate four wave mixing. This process creates a polarization wave that is the time reversed image of one of the input waves. Figure 16 compares the reflection from a conventional mirror with that from a phase conjugate mirror.

The mixing process shown in Figure 17 is one that can act as a phase conjugate mirror. Waves f and b are strong pump waves that are phase conjugates of one another. In their presence a weak signal field at the same frequency produces a polarization given by

$$P = sE_fE_bE_s^*\, e^{-i\omega t + i(\mathbf{k}_f - \mathbf{k}_b - \mathbf{k}_s)\cdot\mathbf{r}} \qquad [24]$$

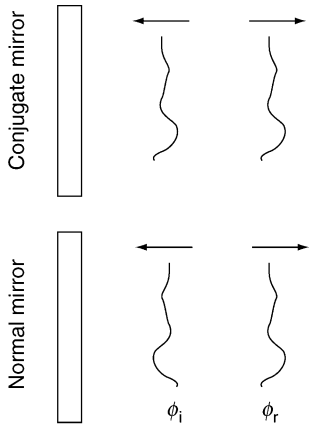


Figure 16 Comparison of a normal and a phase conjugate mirror. The wavy lines represent the phases of the incident and reflected waves.

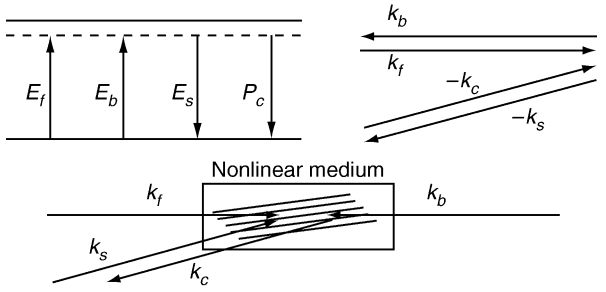


Figure 17 Phase conjugation via degenerate four wave mixing near a one photon resonance. Two counter propagating pump waves with fields E_f and E_b interact with a signal beam, E_s, to produce a polarization wave P_c where the frequencies obey $\omega_f - \omega_s + \omega_b = \omega_c$ (upper left), implying the k-vector of the polarization wave is given by $k_c = k_f - k_s + k_b$ (upper right). The lower diagram shows the grating induced by waves f and s that scatters wave b into the direction of c.

The two pump propagation vectors cancel one another leaving

$$P = sE_fE_b(E_s^*\, e^{-i\omega t - i\mathbf{k}_s\cdot\mathbf{r}}) \qquad [25]$$

The quantity in parentheses is the phase conjugate of the signal wave, so the polarization wave exactly phase matches a radiated conjugate wave.

Brillouin scattering

Other possibilities for automatic phase matching are processes that involve a stimulated wave in the nonlinear medium such as an acoustic wave (Brillouin scattering) or waves of coherent molecular vibration or rotation (Raman scattering). Phase conjugation, based on Brillouin scattering, is widely used to reduce phase distortions acquired on passing through an amplifier, for example. A single beam of light is focused into a material that has a high polarizability. Example media include high pressure gases Xe and SF_6, liquids acetone and freon, and fused silica. The conjugation process is not as simple to describe as it was in the case of four wave mixing because many acoustic waves can be generated. The key is that the phase conjugate wave has the highest gain under certain conditions and can dominate the Brillouin scattering process, so the reflected wave is nearly a phase conjugate of the incoming wave.

See also

Modulators: Acousto-Optics; Electro-Optics. **Nonlinear Optics, Basics:** Cascading; Four-Wave Mixing; Nonlinear Optical Phase Conjugation; $\chi^{(3)}$–Third-Harmonic Generation. **Nonlinear Sources**: Harmonic Generation in Gases. **Optical Parametric Devices:** Overview. **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed).

Phase Control: Phase Conjugation and Image Correction. **Photonic Crystals:** Nonlinear Optics in Photonic Crystal Fibers. **Scattering**: Raman Scattering.

Further Reading

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.

Boyd RW (1992) *Nonlinear Optics*. New York: Academic Press.

Dmitriev VG, Gurzadyan GG and Nikogosyan DN (1999) *Handbook of Nonlinear Optical Crystals*. Berlin: Springer-Verlag.

Hanna DC, Yuratich MA and Cotter D (1979) *Nonlinear Optics of Free Atoms and Molecules*. Berlin: Springer-Verlag.

Mills DL (1991) *Nonlinear Optics*. Berlin: Springer-Verlag.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: John Wiley & Sons.

Sutherland RL (1996) *Handbook of Nonlinear Optics*. New York: Marcel Dekker.

Yariv A and Yeh P (1984) *Optical Waves in Crystals*. New York: John Wiley & Sons.

Zel'dovich BY, Pilipetsky NF and Shkunov VV (1985) *Principles of Phase Conjugation*. Berlin: Springer-Verlag.

Zernike F and Midwinter JE (1973) *Applied Nonlinear Optics*. New York: John Wiley & Sons.

Pulse Compression via Nonlinear Optics

M F S Ferreira, University of Aveiro, Aveiro, Portugal

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Ultrashort pulses have a great potential in several areas of research and application, such as measurements of ultrafast physical processes, infrared time-resolved spectroscopy, and sampling systems, but they are also invaluable for communications. With increasing demands being put on the performance of optical fiber communications, the emphasis now is on producing a source of ultrashort optical pulses that can be used in high-bit rate and long-distance optical communication systems.

Several optical pulse compression techniques are described in the literature. These techniques utilize dispersion, possibly from gratings or prisms, simultaneously to compensate for chirp and achieve pulse compression. In fact, when a chirped optical pulse propagates in a linear dispersive medium, it acquires a dispersion-induced chirp. If the initial chirp is in the opposite direction to that imposed by the group-velocity dispersion (GVD), the two tend to cancel each other, resulting in the compression of the optical pulse.

Early work on optical pulse compression did not make use of any nonlinear optical effects. Only during the 1980s, when the use of single-mode silica fibers as nonlinear media became widespread, the self-phase modulation (SPM) effect was used to achieve pulse compression. It has led, in 1987, to the production of optical pulses as short as 6 fs in the 620 nm wavelength region. Such advances were possible, only after having understood the evolution of optical pulses in silica fibers.

We can consider two main categories of pulse compressors which make use of nonlinear effects in optical fibers: grating-fiber and soliton-effect compressors. In a grating-fiber compressor, the input pulse is propagated in the normal-dispersion regime of the fiber, which imposes a nearly linear, positive chirp on the pulse through a combination of SPM and GVD. The output pulse is then sent through a grating pair where it experiences anomalous GVD and gets compressed. In the case of a soliton-effect compressor, the fiber itself acts as a compressor without the need of an external grating pair. The input pulse propagates in the anomalous-GVD regime of the fiber and is compressed through an interplay between SPM and GVD. This compression mechanism is related to a fundamental property of the higher-order solitons. These solitons follow a periodic evolution pattern such that they go through an initial narrowing phase at the beginning of each period. If the fiber length is suitably chosen, the input pulses can be compressed by a factor that depends on the soliton order.

Grating-Fiber Compressors

The grating-fiber compressor is generally used at wavelengths $\lambda < 1.3\ \mu$m. At these wavelengths, optical pulses acquire a positive chirp across its entire width during their passage through conventional silica fibers. A schematic of a grating-fiber compressor in the double pass configuration is shown in Figure 1. After the propagation in the fiber, the optical pulse is sent through a grating pair, which provides the anomalous (or negative) GVD required to get the pulse compression.

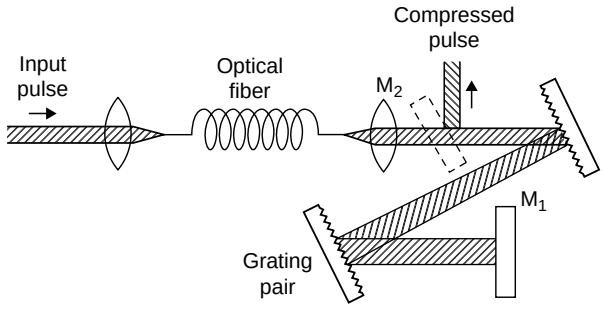


Figure 1 Schematic representation of a grating-fiber compressor in the double-pass configuration. (Reproduced with permission from Agarwal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego: Academic Press.)

Different frequency components of an optical pulse incident at one grating are diffracted at slightly different angles. As a consequence, they experience different time delays during their passage through the grating pair, the blue-shifted components arriving earlier than the red-shifted ones. In the case of an optical pulse with a positive chirp, the blue-shifted components occur near the trailing edge of the pulse and the passage through the grating pair provides its compression.

A limitation of the grating pair is that the spectral components of the pulse are dispersed not only temporally but also spatially. As a consequence, the optical beam becomes deformed, which is undesirable. This problem can be avoided simply by using a mirror to reflect the pulse back through the grating pair. Mirror M_1, in Figure 1, performs this function. Reversing the direction of propagation not only allows the beam to recover its original cross-section, but also doubles the amount of GVD, thereby reducing the grating separation by a factor of 2. The mirror M_2 in Figure 1 is used to deflect the compressed pulse out of the compressor.

The phase shift acquired by the spectral component of the pulse at the frequency ω passing through the grating pair can be written in the form:

$$\phi_g = \phi_0 + \phi_1(\omega - \omega_0) + \frac{1}{2}\phi_2(\omega - \omega_0)^2 + \frac{1}{6}\phi_3(\omega - \omega_0)^3 + \cdots \quad [1]$$

where ω_0 is the pulse center frequency, ϕ_0 is a constant, ϕ_1 is a constant delay, and ϕ_2 and ϕ_3 are parameters that take into account the GVD effects associated with the grating pairs. These parameters depend on the grating period (line spacing), as well as on the orientation and separation of the two gratings. In most cases of practical interest, the spectral width of the pulse satisfies the condition $\Delta\omega \ll \omega_0$ and the cubic and higher-order terms in eqn [1] can be neglected.

To achieve optimum performance from a grating-fiber compressor, it is necessary to optimize the fiber length, as well as the grating separation. Concerning the first aspect, the effects of both GVD and SPM during the propagation of the pulse inside the fiber must be considered. SPM alone determines a linear chirp only over the central part of an optical pulse. Since the grating pair compresses only this region, while a significant amount of energy remains in the wings, the compressed pulse is not of high quality in this case. The effect of GVD turns out to be important, since it broadens and reshapes the pulse, which develops a nearly linear chirp across its entire width. In these circumstances, the grating pair can compress most of the pulse energy into a narrow pulse.

The balanced effects of both GVD and SPM explain the existence of an optimum fiber length, z_{opt}, for maximum pulse compression. In fact, when the fiber length is less than z_{opt}, the SPM-induced chirp has not yet been linearized. On the other hand, when it is longer than z_{opt}, the SPM effects become negligible due to the GVD-induced pulse broadening.

If the input pulse is unchirped and presents a hyperbolic secant shape, the optimum fiber length for maximum pulse compression is well approximated by

$$z_{opt} \approx (6L_D L_{NL})^{1/2} \quad [2]$$

where

$$L_D = \frac{T_0^2}{|\beta_2|} \quad [3]$$

and

$$L_{NL} = \frac{1}{\gamma P_0} \quad [4]$$

are length scales associated with the GVD and the SMP effects, respectively. In eqns [3] and [4], T_0 is the input pulsewidth, β_2 is the fiber GVD parameter, γ is the nonlinearity parameter of the fiber, and P_0 is the peak power of the input pulse.

If the input pulse presents a negative chirp, this must be compensated for by the positive chirp provided by the fiber. As a result, the optimum fiber length increases. The opposite occurs in the case of an input pulse with a positive chirp.

The optimum compression factor, F_{opt}, in the case of a grating-pair compressor, is given by

$$F_{opt} = \frac{T_{FWHM}}{T_{comp}} \approx 0.625\left(\frac{L_D}{L_{NL}}\right)^{1/2} \quad [5]$$

where T_{FWHM} is the full width at half maximum (FWHM) of the input pulse ($T_{FWHM} \approx 1.76T_0$ for a hyperbolic secant pulse) and T_{comp} is the FWHM of the compressed pulse. Equations [2] and [5] provide a good estimate even for pulse shapes other than a hyperbolic secant as long as $\frac{L_D}{L_{NL}} > 100$.

The previous results were obtained neglecting the higher-order nonlinear and dispersive effects, which is acceptable for pulsewidths $T_0 > 0.1$ ps. However, when shorter pulses are used at the input, their spectral width $\Delta\omega$ is large enough that the cubic term in the expansion [1] must be taken into account. In this case, the compression factor turns out to be smaller than that given by eqn [5].

The third-order dispersion (TOD) of the grating pair, resulting from cubic term in eqn [1], has been compensated for in one experiment using a combination of gratings and prisms. Performing such compensation, a reduction of the pulsewidth to about 6 fs has been achieved. Such a pulse at 620 nm consists of only three optical cycles.

Even though the compression factor $F_c \propto \sqrt{P_0}$, the peak power of the input pulse must be kept below the Raman threshold to avoid the loss of pulse energy through the stimulated Raman scattering (SRS) process. If this condition is not then satisfied, besides the energy loss problem, the Raman pulse can interact with the input pulse through cross-phase modulation and deform the linear nature of the frequency chirp. As a result, the compression factor is reduced from the predictions of eqn [5].

Fiber Bragg Gratings

A uniform fiber Bragg grating (FBG) reflects light whose wavelength is within the stop-band centered at the Bragg wavelength. The GVD is anomalous on the high-frequency side of the stop-band and becomes normal on the low-frequency side. Near the stop-band edges, the grating exhibits large GVD. Since a FBG acts as a dispersive delay line, it can be used in place of the bulk-grating pair in a grating-fiber compressor, providing a compact all-fiber device. In fact, a typical 1 cm long fiber grating may provide as much dispersion as a bulk-grating pair with more than one meter spacing. However, the grating induced TOD can affect significantly the quality of the pulses when the optical frequency falls close to the edges of the stop-band.

The compression factor as well as the pulse quality can be significantly improved by using a FBG in which the grating optical period changes along its length – the so-called chirped fiber grating. In this case, different frequency components of the pulse are reflected from different regions of the grating. Such a device has been used since the mid-1990s to compensate for dispersion-induced broadening of pulses in fiber-optic communication systems, and compression factors above 100 have been achieved. In practice, the only disadvantage of a chirped fiber grating is that the compressed pulse is reflected rather than transmitted.

Soliton-Effect Compressors

Optical pulses at wavelengths exceeding 1.3 μm generally experience both SPM and anomalous GVD during their propagation in silica fibers. Such a fiber can act as a compressor by itself without the need of an external grating pair. In the case of a fundamental or first-order soliton, the effects of SPM cancel the effects of anomalous GVD perfectly, and the soliton propagates whilst preserving its hyperbolic secant shape in a lossless optical fiber. A higher-order soliton, on the other hand, changes its shape periodically as it propagates in a fiber. Such a soliton always experiences an initial pulse narrowing phase before recovering its original hyperbolic secant profile at integral multiples of the soliton period z_0, and this behavior is exploited to achieve pulse compression. Optimum compression is achieved when the higher-order soliton leaves the fiber just as it attains its narrowest width.

The evolution of a soliton of order N inside an optical fiber is governed by the nonlinear Schrödinger equation (NLSE), which can be written in the following form:

$$i\frac{\partial q}{\partial \xi} + \frac{1}{2}\frac{\partial^2 q}{\partial \tau^2} + N^2|q|^2 q = 0 \qquad [6]$$

where q is the normalized amplitude, $\xi = z/L_D$, $\tau = T/T_0$, and the parameter N is the soliton order, given by

$$N^2 = \frac{L_D}{L_{NL}} = \frac{\gamma P_0 T_0^2}{|\beta_2|} \qquad [7]$$

The soliton period, z_0, is given by

$$z_0 = \frac{\pi}{2}L_D = \frac{\pi T_0^2}{2|\beta_2|} \qquad [8]$$

In writing eqn [6], the fiber losses were neglected, since fiber lengths employed in practice are relatively small. Even though higher-order solitons follow an exact periodic evolution only for integer values of N, eqn [6] can be used to describe pulse evolution for arbitrary values of N.

In practice, soliton effect compression is carried out by initially amplifying optical pulses up to the power level required for the formation of higher-order solitons. The peak optical power of the initial pulse required for the formation of the *N*th-order soliton is given by

$$P_{\mathrm{N}} = 3.11\,\frac{|\beta_2|N^2}{\gamma T_{\mathrm{FWHM}}^2} \qquad [9]$$

These *N*th-order solitons are then passed through the appropriate length of optical fiber to achieve a highly compressed pulse. The optimum fiber length, z_{opt}, and the optimum pulse compression factor, F_{opt}, of a soliton-effect compressor can be estimated from the following empirical relations:

$$z_{\mathrm{opt}} \approx z_0\left(\frac{0.32}{N} + \frac{1.1}{N^2}\right) \qquad [10]$$

$$F_{\mathrm{opt}} \approx 4.1N \qquad [11]$$

These relations are accurate to within a few percent for $N > 10$.

Compression factors as large as 500 have been attained using soliton-effect compressors. In fact, considering the same input pulse, the soliton-effect compressor provides a pulse compression that is larger with a fiber that is shorter than a grating-fiber compressor. However, the pulse quality is poorer, since the compressed pulse carries only a fraction of the input energy, while the remaining energy appears in the form of a broad pedestal. From a practical point of view, the pedestal is deleterious since it causes the compressed pulse to be unstable, making it unsuitable for some applications. Despite this, soliton-effect compressed pulses can still be useful because there are some techniques that can eliminate the pedestal.

One difficulty faced, when using the soliton-effect compressor, is that pulses with high peak power are required for the formation of high-order solitons in conventional fibers. As suggested by eqn [9], the use of dispersion-shifted fibers (DSFs) with small values of β_2 at the operating wavelength can reduce the peak power required for soliton generation by an order of magnitude. However, because the soliton period z_0 is inversely proportional to $|\beta_2|$ (see eqn [8]), eqn [10], indicates that longer lengths of DSFs will be required for solitons to achieve optimum compression. As a result, the total fiber loss experienced by those solitons will be larger and the loss-induced pulse broadening will have a significant impact on the compressor global performance.

Another problem introduced by the use of DSFs in soliton-effect compressors is the TOD, which generally degrades the quality of the compressed pulse. In the case of soliton-effect compressors using conventional optical fibers, the effects of TOD become significant only when the widths of the compressed pulse become very short ($\ll$1 ps). In DSFs, however, the relative importance of TOD is increased and those effects become more pronounced because the GVD parameter, β_2, is small. In this case, TOD is detrimental in the compression, even of pulses with widths of a few picoseconds, resulting in serious degradation of the optimum compression factor F_{opt}.

The result given by eqn [11] for the optimum compression factor holds only for the case of an ideal soliton-effect compressor, when high-order nonlinear and dispersive effects, such as intrapulse Raman scattering (IRS), TOD and self-steepening, are neglected. However, these effects cannot be neglected for the highly compressed pulses because of their subpicosecond widths and high intensities.

For not-too-short optical pulses (widths >50 fs) propagating not too close to the zero-dispersion wavelength of the fiber, the dominant higher-order effect is IRS, which manifests as a shift of the pulse spectrum toward the red side – the so-called soliton self-frequency shift. As a consequence, the group velocity is reduced and there is a delay of the optical pulse. This delay affects substantially the interplay between GVD and SPM that is responsible for pulse compression.

Interestingly, the IRS effect can be used to improve the quality of the compressed pulse, by removing the pedestal mentioned above. In fact, as a consequence of the change in the group velocity induced by the IRS effect, the sharp narrow spike corresponding to the compressed pulse travels more slowly than the pedestal and separates from it. Moreover, the pedestal can be removed by spectral filtering and a red-shifted, pedestal-free, and highly compressed pulse is then produced.

Compression of Fundamental Solitons

As described in the last section, the propagation of higher-order solitons provide a rapid compression method, but it suffers from the existence of residual pedestals. Some techniques can help to reduce or even eliminate those pedestals, but energy is always wasted in this process.

A less rapid technique that provides better pulse quality is adiabatic amplification of fundamental or first-order solitons, which are of primary importance in the domain of optical communications. Solitons have fixed area, so the increased energy from amplification is accommodated by an increase in

power and a decrease in width. To avoid distortion, the amplification per soliton period cannot be too great.

The same effect as adiabatic amplification can be achieved using an optical fiber with dispersion that decreases along the length of the fiber. In fact, for optical solitons, a small variation in the dispersion has a similar perturbative effect as an amplification or loss: such a variation perturbs the equilibrium between the dispersion and nonlinearity in such a way that when, for example, the dispersion decreases, the soliton pulse is compressed. It can be seen from eqn [7] that if the value of $|\beta_2|$ decreases along the fiber and to keep the soliton order N ($N = 1$ for the fundamental soliton), the pulsewidth must indeed decrease as $|\beta_2|^{1/2}$. Hence, the use of fibers with variable dispersion is viewed as a passive and effective method to control optical solitons in soliton communication systems. Dispersion-decreasing fibers (DDFs), in particular, have been recognized to be very useful for high-quality, stable, polarization-insensitive, adiabatic soliton pulse compression and soliton train generation. These fibers can be made by tapering the core diameter of a single-mode fiber during the drawing process, and hence changing the waveguide contribution to the second-order dispersion.

In a DDF of certain length, the dispersion is monotonically and smoothly decreased from an initial value to a smaller value at the end of the length according to some specified profile. Provided the dispersion variation in the DDF is sufficiently gradual, soliton compression can be an adiabatic process where an input fundamental soliton pulse can be ideally compressed as it propagates, while retaining its soliton character and conserving the energy.

The evolution of the fundamental soliton in a DDF can be described by the following NLSE:

$$\mathrm{i}\frac{\partial q}{\partial \xi} + \frac{1}{2}p(\xi)\frac{\partial^2 q}{\partial \tau^2} + |q|^2 q = 0 \qquad [12]$$

where the variable coefficient $p(\xi) = |\beta_2(\xi)/\beta_2(0)|$ takes into account the variation of dispersion along the fiber. Using the transformations $s = \int_0^{\xi} p(y)\mathrm{d}y$ and $U = q/\sqrt{p}$, eqn [12] assumes the form:

$$\mathrm{i}\frac{\partial U}{\partial s} + \frac{1}{2}\frac{\partial^2 U}{\partial \tau^2} + |U|^2 U = \mathrm{i}GU \qquad [13]$$

where

$$G = -\frac{1}{2p}\frac{\mathrm{d}p}{\mathrm{d}s} \qquad [14]$$

Equation [13] shows clearly that the effect of decreasing dispersion is mathematically equivalent to the effect of an optical amplifier, adding a gain term to the NLSE. The effective gain coefficient G is related to the rate at which GVD decreases along the fiber.

Since decreasing dispersion is equivalent to an effective gain, a DDF can be used in place of a conventional fiber amplifier to generate a train of ultrashort pulses. To achieve such an objective, a continuous wave (CW) beam with a weak sinusoidal modulation imposed on it is injected into the DDF. The sinusoidal modulation can be imposed, for example, by beating two optical signals. As a result of the combined effect of GVD, SPM, and decreasing GVD, the nearly CW beam is converted into a high-quality train of ultrashort solitons, whose repetition rate is governed by the frequency of the initial sinusoidal modulation.

Different approaches have been developed to determine the optimum GVD profile and its dependence on the width and peak power of input pulses. In the case of picosecond soliton pulse compression, direct numerical simulations of the NLSE show that linear, Gaussian, and exponential dispersion profiles may all be used effectively to provide ideal, adiabatic compression, where the input pulse energy is conserved and remains localized within the pulse. In this case, the final pulse compression factor at the end of a certain length L of DDF is equal to the ratio of the initial to the final second-order dispersions:

$$F_c = \frac{T_{\mathrm{FWHM}}}{T_{\mathrm{comp}}} = \frac{\beta_2(0)}{\beta_2(L)} \qquad [15]$$

Compression factors larger than 50 are possible launching input pulses with peak powers corresponding to the fundamental soliton into a DDF whose length is about one soliton period. This technique takes advantage of soliton-effect compression but requires lower peak powers and produces compressed pulses of better quality.

In the case of subpicosecond soliton pulses, the influence of higher-order nonlinear and dispersive effects must be accounted for. Some studies show that, in this case, the linear and Gaussian dispersion profiles are the most suitable to achieve high-quality, pedestal-free, adiabatic compression of fundamental solitons. However, in the presence of higher-order effects, the final compression factor is generally lower, and after reaching a maximum at a particular length of DDF, it decreases steadily. This can be explained in terms of pulse compression stabilization, which originates from a competition between the rate of dispersion decrease in the DDF and the rate of dispersion increase due to the combined effects of soliton self-frequency shift and third-order dispersion.

List of Units and Nomenclature

F_c	compression factor
F_{opt}	optimum compression factor
G	amplifier gain
L	fiber length (m)
L_D	dispersion length (m)
L_{NL}	nonlinear length (m)
N	soliton order
P_N	peak power of the Nth-order soliton (W)
P_0	pulse peak power (W)
q	normalized pulse amplitude
T_0	input pulse (s)
T_{comp}	compressed pulse FWHM (s)
T_{FWHM}	pulse full width at half maximum (s)
z_0	soliton period (m)
z_{opt}	optimum fiber length (m)
β_2	group-velocity dispersion coefficient ($s^2\,m^{-1}$)
γ	fiber nonlinearity coefficient ($W^{-1}\,m^{-1}$)
ξ	normalized distance
τ	normalized time
ϕ_g	phase shift induced by the grating pair (rad)
ω_0	pulse center frequency (rad s^{-1})
ω	optical frequency (rad s^{-1})
$\Delta\omega$	pulse spectral width (rad s^{-1})

See also

Chirped Pulse Amplification. **Fiber Gratings**. **Scattering:** Raman Scattering. **Solitons:** Soliton Communication Systems. **Ultrafast Laser Techniques:** Pulse Characterization Techniques.

Further Reading

Agrawal GP (1995) *Nonlinear Fiber Optics*, 2nd edn. San Diego: Academic Press.

Agrawal GP (1997) *Fiber-Optic Communication Systems*, 2nd edn. New York: John Wiley & Sons.

Agrawal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego: Academic Press.

Barty C, White W, Sibbett W and Trebino R (eds) (1998) Special issue on ultrafast optics. *IEEE Journal of Selected Topics on Quantum Electronics* 4: 157.

Buck JA (1995) *Fundamentals of Optical Fibers*. New York: John Wiley & Sons.

Kashyap R (1999) *Fiber Bragg Gratings*. San Diego, CA: Academic Press.

Mamyshev PV (1992) Generation and compression of femtosecond solitons in optical fibers. In: Taylor JR (ed.) *Optical Solitons – Theory and Experiment*. Cambridge, UK: Cambridge University Press.

Raman Lasers

M Santagiustina, Istituto Nazionale di Fisica della Materia, Università degli Studi di Padova, Padova, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

The Raman effect bears the name of the Indian scientist Chandrasekhara V. Raman (1888–1970) who received the Nobel Prize in the year 1930 for the discovery, published in 1928, of this scattering phenomenon.

The Raman effect is an inelastic light scattering process, occurring in gaseous, liquid, and solid media. The frequency of the scattered photon can be either down-shifted or up-shifted and the result of the process is the generation of new radiation. Frequency down-shifted and up-shifted waves are respectively called Stokes and anti-Stokes waves. The energy difference between input and output photons is either absorbed by, or released from, the atoms or molecules of the medium, which are left in different energy states; jumps can be between electronic states, but more often they are among vibrational or rotational states. The upper energy state involved in the scattering process can be either a true energy state, and this case is referred to as resonant Raman scattering, or a virtual one. The off-resonance effect has a lower probability to occur and high-power beams are needed to observe it. The emission of anti-Stokes waves requires either an inverted Raman medium or particular four-wave-mixing phase-matching conditions to be satisfied in order to occur. The second condition is referred to as coherent anti-Stokes Raman scattering. Here we will consider mainly Stokes wave generation, which is the most common in Raman lasers.

When pumping is very hard, or a seed wave is provided, stimulated Raman scattering can also occur and a high percentage of the input pump radiation can be converted into Stokes, or anti-Stokes, radiation. The stimulated process efficiency, defined as the ratio between the output Raman wave energy

and the input pump wave energy, depends on the imaginary part of the third-order nonlinear susceptibility coefficient and on the pump intensity. Stimulated Raman scattering observation dates back to 1962 during the pioneering studies with lasers sources by Woodbury and Ng; their experiment with a ruby laser was actually the first example of an intracavity Raman laser.

Since its discovery, Raman scattering has become a valuable tool for spectroscopy and a practical technique to extend laser light emission to wavelengths not directly generated by laser active materials. This fundamental property stems from the fact that Raman scattering does not require an upper energy state to occur. Hence, the frequency of the emitted radiation can be any, if a suitable incident radiation wavelength is provided. New frequencies are constrained only by the available powerful pump laser sources and by Raman medium shifts.

The frequency shifts provided by some of the materials that are most often used in Raman lasers are presented in Table 1. Finally, it is important to note that Raman scattering can also occur for the newly generated radiation through an identical process in which a Stokes photon is Raman scattered. This radiation, which is generated if the Stokes radiation intensity grows large enough, is referred to as a second-order Stokes wave. For very large pump intensity, a cascading process to several Stokes orders can take place, thus extending the range of generated frequencies to the far-infrared portion of the spectrum.

We will describe the design of Raman lasers that are most commonly used, as well as their performances, also taking into consideration the Raman medium that sets the main constraint to the laser characteristics. Continuous wave as well as pulse Raman lasers will be described and the tuning capability will be discussed.

Raman Shifters

The simplest conversion scheme is obtained when a Raman medium is placed at the output end of a laser. This single-pass setup, shown in Figure 1, provides a fairly efficient frequency shifter.

Single-pass amplification of the spontaneous scattering or of an injected seed copropagating with an intense pump is usually modeled by a couple of equations whose derivation, under a plane, continuous wave approximation, can be found in the

Table 1 Measured Raman frequency shifts and steady-state gain for some of the most important Raman media

Medium	*Frequency shift [THz]*	*Gain [cm/GW]*
Calcite ($CaCO_3$)	32.58[1]	5.5[1]
Lithium iodate ($LiIO_3$)	23.1–24.66[2]	4.8[3]
Barium nitrate ($Ba(NO_3)_2$)	31.46[4]	11[5]
Potassium gadolinium tungstate ($KGd(WO_4)_2$)	27.03 or 23[6]	6[6]
Silica fiber (SiO_2)	13.2[7]	0.01[7]
Hydrogen gas (H_2) ($p = 10$ atm, $T = 293$ K)	124.65[8]	1.53[9]
Nitrobenzene	40.35[10]	3[10]

[1]Pump wavelength 694 nm; from Zverev PG, Basiev TT and Prokhorov AM (1999) Stimulated Raman scattering of laser radiation in Raman crystals. *Optical Materials* 11: 335–352.
[2]Values refer to a propagation direction respectively perpendicular or orthogonal to the crystal axis; from Guadalberto GM and Arguello CA (1974) Raman spectra from oblique phonons in powdered samples. *Solid State Communications* 14: 911–914.
[3]Pump wavelength 1064 nm; from Falk J and Moshrefzadeh (1985) Oblique Raman and polariton scattering in kithium iodate. *IEEE Journal of Quantum Electronics* QE-21: 110–113.
[4]Pump wavelength 1064 nm; from Basiev TT, Sobol AA, Zverev PG, Ivleva LI, and Osiko VV (1999) Raman spectroscopy of crystals for stimulated Raman scattering. *Optical Materials* 11: 307–314.
[5]Pump wavelength 1064 nm; from Basiev TT, Voitsekhovskii VN, Zverev PG, *et al.* (1987) Conversion of tunable radiation from a laser utilizing an LiF crystal containing F_2^- color centers by stimulated Raman scattering in $Ba(NO_3)_2$ and $KGd(WO_4)_2$ crystals. *Soviet Journal of Quantum Electronics* 17: 1560–1561.
[6]Values refer respectively to two linear and orthogonal pump beams at the wavelength 1064 nm; from Zverev PG, Basiev TT and Prokhorov AM (1999) Stimulated Raman scattering of laser radiation in Raman crystals. *Optical Materials* 11: 335–352.
[7]Pump wavelength 1064 nm; from Stolen RH (1978), Fibre Raman lasers. In Ostrowsky D (ed) *Fiber and Integrated Optics*, NATO ASI Series B41. New York, Plenum Press, 157–182.
[8]From Bischel WK and Dyer MJ (1986) Temperature dependence of the Raman linewidth and lineshift for the Q(0)–Q(1) transition in normal and para H_2. *Physical Review* A 33: 3113–3123.
[9]Pump wavelength 1064 nm; from Ottusch JJ and Rockwell DA (1988) Measurements of Raman gain coefficients in hydrogen, deuterium and methane. *IEEE Journal of Quantum Electronics* 24: 2076–2080.
[10]Pump wavelength 694 nm; from Bloembergen N (1967) The stimulated Raman effect. *American Journal of Physics* 35: 989–1023.

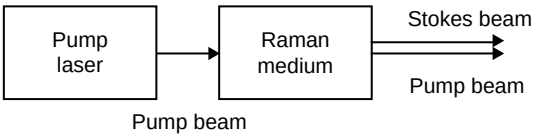


Figure 1 Experimental setup of a single-pass Raman shifter.

literature:

$$\frac{dI_S(z)}{dz} = g_R I_S(z) I_P(z) - \alpha_S I_S(z) \qquad [1]$$

I_S, I_P are the continuous wave intensities of the Stokes and pump waves, z is the propagation distance, g_R the Raman steady-state gain coefficient, and α_S, α_P are the attenuation coefficients at the Stokes and pump frequencies. The Raman gain coefficients, whose general expression as a function of the microscopic material properties can be found in many textbooks, of some of the most commonly used Raman laser media are listed in Table 1.

For short propagation lengths the medium loss can usually be neglected, and if the pump beam is well collimated, the pump intensity can be considered as a constant. For a long interaction length or very absorptive media, the evolution of the pump intensity must also be considered. Though the pump beam is depleted by the scattering process an undepleted pump approximation is often used until high conversion is reached. Under this approximation, the evolution of the pump intensity is mainly governed by diffraction spreading, by losses or by both. In bulk media, single-pass scattered radiation has a coaxial component whose divergence is dictated by the geometry of the pump beam, because Stokes radiation is close to the diffraction limit. Besides this radiation, a nonaxial component can also be observed; this phenomenon stems from the well-known effect of conical emission, which is a phase-matched parametric process. If the Raman medium is within a waveguide, losses are usually the dominating effect. In the undepleted approximation, the evolution of the pump beam intensity as a function of the propagation distance z, either dictated by diffraction or by losses, is known. Then the general solution of eqn [1] is:

$$I_S(z) = I_S(0) \exp\left[g_R \int_0^z I_P(z')dz' - \alpha_S z \right] \qquad [2]$$

where $I_S(0)$ is either the input Stokes signal intensity or an equivalent spontaneous scattering intensity. Then, the amplification gain G of the Stokes beam in the single-pass shifter, for a medium of length L, is given by

$$G = \frac{I_S(L)}{I_S(0)} = \exp\left[g_R \int_0^L I_P(z')dz' - \alpha_S L \right] \qquad [3]$$

Though single-pass Raman shifters are appealing for their simplicity, their application is hampered by the high power necessary for reaching a sufficient efficiency due to the short interaction length that can be obtained, in particular in bulk media, where diffraction dominates. Q-switched pulsed pump lasers are often required, but the propagation in the Raman medium of very powerful pulses requires special precautions to prevent other linear and nonlinear effects occurring. In fact, optical breakdown, thermal and Kerr self-focusing, four-wave mixing, and Brillouin scattering can take place, thus reducing the efficiency of the converter. Precautions may include the optimization of the pump pulse duration, of the focusing, of the interaction length, and the control of the input beam spot size. The multipass configuration has also been exploited to avoid many of the single-pass configuration shortcomings; the typical experimental setup is presented in Figure 2. Note that neither the pump nor the Stokes wave are resonant within the cavity.

The primary advantage of the multipass configuration is the lowering of the pump power required to realize an appreciable conversion. This is due to the increased total gain coefficient, that has been found to be equal to

$$G_n = G\frac{1 - R^n}{1 - R} + (n - 1)\log R \qquad [4]$$

where G is the single-pass gain of eqn [3], R the reflectivity of the mirrors at the Stokes wavelength, and n the number of passes through the Raman medium. There are two main drawbacks of multipass Raman shifters that finally limit the conversion efficiency of this setup. The first is the simultaneous generation of many undesired and uncontrolled Stokes orders, because the first-order Stokes intensity grows rapidly during propagation. The second is the

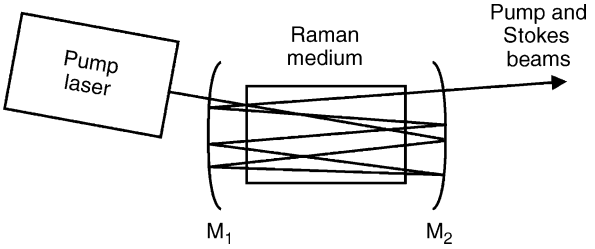


Figure 2 Experimental setup of a multipass Raman shifter. Mirrors M_1 and M_2 reflect both Stokes and pump waves.

intensity decrease associated with the dispersion of pulses, in particular if the Raman medium is a solid. These problems have limited the applications of multipass shifters to hydrogen gas.

Raman Lasers

By placing the Raman medium into an optical cavity, Stokes radiation can be selectively enhanced and Raman oscillation takes place. The two main experimental setups, either with the Raman medium placed in an external cavity or within the pump laser cavity (intracavity), are represented in Figures 3 and 4. In some cases, for the intracavity configuration, the Raman medium may also be the host of the doping laser active material.

For the external resonator Raman laser, the reflectivity of the mirrors is such that they are almost transparent for the pump but highly reflective (resonant) for the shifted radiation. The threshold condition of Raman lasers can be obtained by requiring that the net round trip gain equals the total cavity losses. The net round trip gain, under the plane wave, steady-state, undepleted pump approximation, is given by the single-pass gain G, multiplied by a factor of two to account for the forward and backward amplification. In fact, Raman scattering takes place also between the pump and the counter-propagating Stokes wave reflected by mirror M_2.

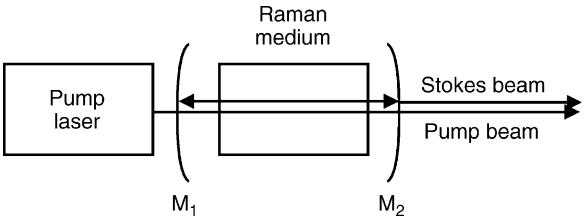


Figure 3 Experimental setup of an externally pumped Raman oscillator. Mirrors M_1 and M_2 are highly reflective for the Stokes wave but not for the pump wave.

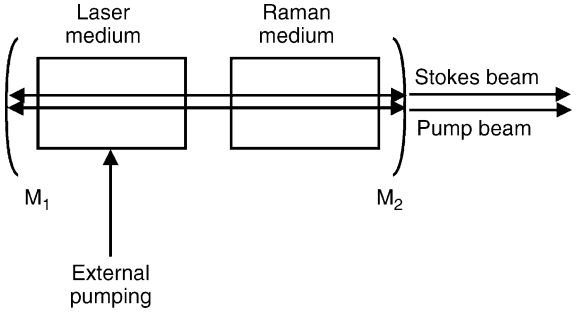


Figure 4 Experimental setup of an intracavity Raman oscillator. Mirrors M_1 and M_2 are highly reflective for both the Stokes and the pump wave.

Losses are given by the medium contribution, which is already accounted for in eqns [2] and [3], and by the energy leakage from the mirrors. Assuming that the mean reflectivities of the mirrors, at the Stokes wavelength, are respectively R_1, R_2, then the threshold condition is $GR_1R_2 = 1$. Through eqn [3] we obtain

$$g_R \int_0^L I_P(z')dz' - \alpha_S L = -\log(R_1R_2) \qquad [5]$$

Equation [5] can be used to evaluate the pump intensity because at threshold low depletion occurs and the approximation taken to obtain [5] is valid. Raman steady-state gain is not very different for gases, liquids and crystals, as shown in Table 1 and oscillation thresholds depend more on the interaction length and the focusing properties of the pump and Stokes beams. Nonetheless, the type of Raman medium imposes many practical constraints for laser design, which are addressed in the next paragraphs.

With gas media it is hard to maintain a high pump intensity over the entire length of the cell, which must be used to contain the gas. Usually gas cells can be from 5 to 30 cm long; this means that Q-switched pump lasers are required to set intensities above 10^4 GW m^{-2}, that is the typical threshold value of Raman oscillation. Continuous wave oscillation of gas Raman lasers is, nonetheless, possible when the pump wavelength is near a medium resonance. In fact, the imaginary part of the third-order non-linear polarization coefficient, to which the Raman gain is proportional, is enhanced near energy transitions because of the resonance in correspondence of that photon energy. The main limitations of this setup are that pump wavelengths are dictated by the Raman medium transitions, and thus also Stokes wavelengths are fixed, and finally conversion efficiency is usually low (about 5% is reported) possibly because of the high absorption. Off-resonance, continuous wave operation can also be attained by using high finesse cavities in a doubly resonant cavity. In fact, by resonating both Stokes and pump wave with high reflectivity mirrors ($R \geq 0.999$), the threshold intensity is lowered and conversion efficiency is high (about 35% has been demonstrated). In spite of the relatively high Raman gain and the large Stokes shifts obtainable, the exploitation of gas cell Raman oscillators is mainly hampered by the need of a high-pressure-resistant vessel and a constant gas flow. The latter is needed because of the low thermal conductivity of gases that might cause local temperature variations and thermal-optical distortions that diminish the output beam uniformity and stability. Moreover the repetition rate of pump pulses must be small (below 50 Hz) so as not to impose a too strong thermal load on the gas.

Liquid media suffer basically the same problems as gases and, moreover, they show higher non-linear refractive indices that result in strong self-focusing. This lens effect results in filamentation, optical breakdown, and high beam divergence. Often Raman liquids are also toxic and very volatile; their use has been slowly but steadily decreasing.

Solid media, in particular crystals, do not show many of the problems listed above and present the advantages of solid-state technology such as compactness, low maintenance, and opportunity of integration. The high density and the restricted atom motion result in high Raman steady-state gains and minimal line broadening. Precise resonance enhancement, as well as low threshold intensities, can be obtained by choosing an adequate mirror reflectivity spectral response. Pump intensities as low as 10^3 GW m^{-2} were reported for the oscillation of an external cavity Raman laser whose medium was a 5-cm-long, barium nitrate (see Table 1) crystal, with mirror reflectivity $R = 0.49$ at the Stokes wavelength. Though the intensity threshold is about an order of magnitude smaller than for gas Raman lasers, Q-switched input pump pulses are still required. Solid-state Raman lasers can be easily forced to oscillate on a $TEM_{0,0}$ mode with divergence close to the diffraction limit even with multimode pumps. Moreover the Stokes beam is subject to a very complex clean-up process; this can be basically explained by the fact that the Stokes beam, growing from the spontaneous emission, does not retain the pump spatial characteristics. Pulse duration is equal to or slightly less than the input one. Conversion efficiency, which depends strongly on mirror reflectivity, can reach 60%, where saturation occurs due to higher-order Stokes wave generation. Solid-state Raman lasers can be easily forced to oscillate also at these higher-order Stokes frequencies if mirror reflectivity is large enough at the relative wavelengths. Second- and third-order Stokes conversion efficiencies of 30% and 20% have been reported.

The theoretical determination of the cavity pump and Stokes fields does not have an analytical solution. For continuous wave or long Q-switched pulsed lasers, eqns [1] must be extended to include all forward and backward Stokes and pump waves and boundary conditions due to the mirrors must be imposed. This problem can only be solved numerically.

The experimental setup of intracavity Raman lasers is shown in Figure 4; both the Raman and the laser medium are placed within the cavity and sometimes the two media coincide. This is, for example, the case of potassium gadolinium tungstate which can be doped with neodymium ions (see Table 1). The cavity is usually resonant for both the pump and the Stokes waves; very high conversion efficiency (up to 80%) is attained when special dichroic coatings are provided such that pump output is suppressed (high-finesse cavity for the pump). Besides the nearly diffraction-limited property of output radiation and clean output beams that are common characteristics of all Raman lasers, enhanced brightness is also observed in this setup. This effect is probably due to the high thermal loading, the consequent linear and nonlinear lens effects, and also to a short pulse generation mechanism that increases output beam intensity. This regime can be reached by reducing the Stokes cavity length with respect to the pump cavity length. A first setup is to use an additional dichroic intracavity mirror that is transparent for the pump but highly reflective for the Stokes; an alternative setup exploits a double ring cavity configuration as depicted in Figure 5.

The principle governing the compression can be explained as follows. The optical paths of the pump and Stokes pulses are made different by an amount ΔL. Then the Stokes pulse generated within the Raman medium, once re-injected into the cavity from mirror M_{S1}, will overlap with an undepleted portion of the recirculating pump pulse. Hence, another Stokes pulse is generated at each round trip, until the pump pulse is completely depleted; the Stokes output finally consists of a train of pulses of duration $\Delta L/c$, at a repetition rate equal to the cavity round-trip. This mechanism leads to pulses as short as the limit imposed by the Raman gain bandwidth. In fact, when the pulse duration becomes comparable to the inverse of the gain bandwidth, transient effects take place because of the delay in the nonlinear medium response. Ring Raman lasers have been obtained with

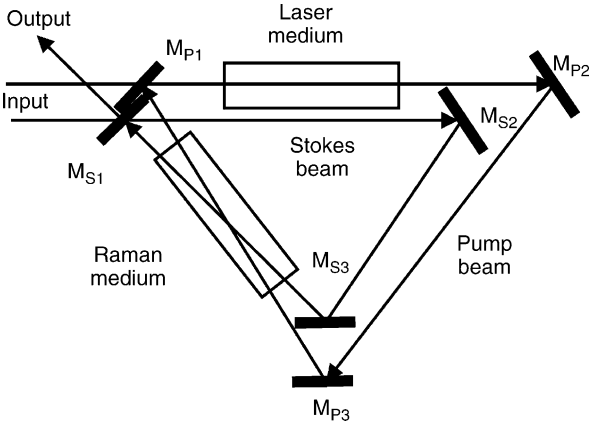


Figure 5 Experimental setup of an intracavity, double ring, Raman oscillator. Mirrors M_{S2}, M_{S3}, M_{P2}, and M_{P3} are highly reflective respectively for the Stokes and the pump wave. M_{S1}, M_{P1} are partially reflective respectively for the Stokes and pump wave.

gases (e.g., hydrogen) and crystal media (barium nitrate and lithium iodate).

Another commonly used experimental setup is that of synchronously pumped ring Raman lasers. In such devices, whose typical setup is sketched in Figure 6, a new input pump pulse is synchronized with the output Stokes pulse thus enhancing the conversion within the Raman medium.

Though electronic trigger detection is needed to lock the pump pulse time repetition rate to a multiple or submultiple of the cavity round-trip, synchronously pumped Raman lasers are relatively simple, stable, high repetition rate sources of radiation. Very compact and reliable synchronously pumped Raman lasers can be realized with optical fibers; in this case the feedback of the ring resonator can be realized by means of a fiber directional coupler. The device, besides the high integration achieved by the all-fiber technology, can also be continuously tuned over a large bandwidth.

Continuous tuning of the same laser source is a very desirable property; the Raman lasers so far described can generate radiation at wavelengths restricted to a few values. In fact, output spectral lines can be changed either by replacing the Raman medium, or the pump laser, or by exploiting cascade Raman scattering or finally resonating a different vibrational line of the same material (if there are many as for example in hydrogen gas); all these cases require substantial cavity changes. Frequency switching without setup changes is attained only in case of potassium gadolinium tungstate. For this medium there exist two independent directions, orthogonal to the crystal optical axis, for which the Raman shifts are different if the pump polarization changes. Then, the wavelength selection is just a matter of controlling the input pump polarization. Though Raman lasers can greatly expand the wavelength range of available laser sources, controlled, continuous tuning of the same Raman laser over a certain frequency band can be obtained only in a few particular cases.

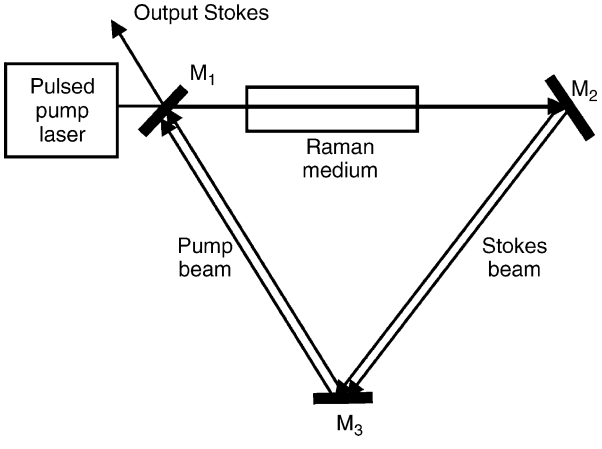


Figure 6 Experimental setup of a synchronously pumped, ring, Raman oscillator. Mirrors M_2 and M_3 are highly reflective for both the Stokes and the pump wave. M_1 is partially reflective for both the Stokes and pump wave.

Wavelength tuning can be attained by selecting the direction of propagation in polar crystals (like lithium niobate and lithium iodate). Different directions allow for wavevector matching of polariton scattering at different frequencies. Also phase matching of parametric interactions, like four-wave mixing, is a possible technique to attain wavelength tuning. However, such Raman lasers cannot operate at maximum gain, and this fact decreases the output energy, the conversion efficiency, and the stability, and increases the pump intensity threshold. Continuous tuning with an external control parameter is attained in the spin-flip Raman semiconductor laser. Semiconductors are Raman media that can be exploited to build far-infrared or even submillimeter Raman lasers, with pump intensity thresholds that are comparable with those of gas and liquid media. Though frequency shifts are fixed, under normal conditions, tuneable operation can be achieved if the semiconductor is placed in a magnetic field. Each energy level of an electron in the solid is, in fact, split into two, depending on the orientation of the electron spin with respect to the magnetic field. When there is a transition involving Raman scattering, the electron spin can change its direction. Then the wavelength of the laser can be continuously varied by changing the magnetic field applied. Tuning over a range of several micrometers in the infrared has been attained and spin-flip Raman lasers have been developed with semiconductors like InSb, $Hg_{0.77}Cd_{0.23}Te$, and $Pb_{0.88}Sn_{0.12}Te$. Use of this technique is severely limited by the need for magnetic field sources.

The simplest method to achieve tuneable radiation is through the use of a material that shows a broad Raman linewidth. Inhomogeneous broadening of the energy level is a well-known effect that takes place in noncrystalline solid media, such as glasses. Though line broadening is accompanied by a decrease in the steady-state Raman gain of about two orders of magnitude with respect to a crystalline medium, this disadvantage can be over-compensated by the increase in the Raman medium length that can be of 3 or 4 orders of magnitude. In fact, glasses are commonly exploited to make optical waveguides, in particular optical fibers, that present very low optical absorption coefficients, in the near-infrared part of the spectrum (1200–1600 nm).

Optical fiber Raman lasers based on silica telecommunication fibers are very versatile, compact,

and simple to implement. In principle the setup is not very different from the corresponding external cavity or intracavity Raman lasers so far presented; however, fiber lasers show the great advantage of an integrated, all-fiber technology.

Cavity mirrors can be replaced by fiber Bragg grating reflectors (Figure 7) that can be constructed with a very well-defined spectral reflectivity, thus allowing fine enhancement or suppression of the desired wavelengths. By splicing the Raman fiber, which can be several hundred meters long, with the gratings, the net losses are very low, finally decreasing the pump intensity threshold. The pump threshold for first-order Stokes generation can be easily as low as 200 mW; this means that diode lasers can be used for pumping. Moreover, the need for cavity alignment is eliminated, the laser stability increased and the construction simplified. This simplification allows us to build, much more easily than by using bulk optics, nested multiple resonant cavities, to generate higher-order Stokes beams (Figure 8).

Continuous-wave lasing of up to six Stokes orders has been reported for a fiber Raman laser of this type, with a total injected power of about 7 W. Another key advantage of Bragg gratings is that their reflectivity can be adjusted by means of a piezoelectric stretching of the fiber support; this allows one to control the output intensity of all waves. As previously mentioned, ring cavities can also be constructed by exploiting an all-fiber technology, if feedback is provided by a fiber directional coupler (Figure 9).

Pulsed operation of fiber Raman lasers has been obtained too; due to the long interaction length, Q-switched pump lasers are no longer needed but rather mode-locked sources can be used. This means that trains of very short laser pulses, down to the hundreds of femto-second regime, can be obtained from fiber Raman lasers, the limit being represented by the Raman bandwidth, which is as large as about 7 THz, at the half maximum, in silica fibers. Propagation of short and intense laser pulses in an optical fiber may induce several other nonlinear effects, like self-, cross-phase modulation, and four-wave mixing. Besides nonlinearity, linear dispersion and velocity mismatch become important too, due to the long propagation. All these linear and non-linear effects limit the Stokes pulse quality. In more detail: self- and cross-phase modulations lead to the spectral broadening of both pump and Stokes waves; broadening is asymmetric, because of the contemporary action of group velocity mismatch. In fact during the propagation the Raman pulse walks off the pump pulse, because of velocity mismatch, and this induces an asymmetry in gain depletion and in nonlinear index modulation. The final outcome is that Stokes waves can be asymmetric both in time and in frequency, and also highly chirped. Four-wave mixing is usually negligible, if phase-matching conditions are not imposed and Brillouin scattering is not an issue for short pulses. Very symmetric and low chirped pulses are attained as soon as the fiber Raman laser is working in the soliton regime, that is when compensation between the dispersion and the nonlinear phase shift occurs. Synchronously pumped, fiber Raman, soliton lasers are able to provide 200-fs pulses, at various wavelengths and with very high conversion efficiency (50%). Intracavity fiber Raman lasers have

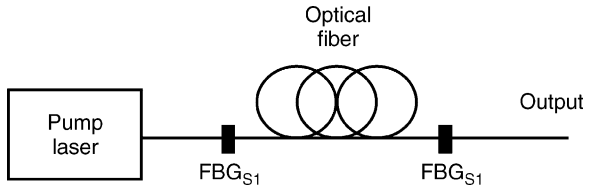


Figure 7 Experimental setup of a continuous-wave, all-fiber, Raman oscillator. FBG_{S1} are fiber Bragg gratings highly reflective for both the first Stokes wave.

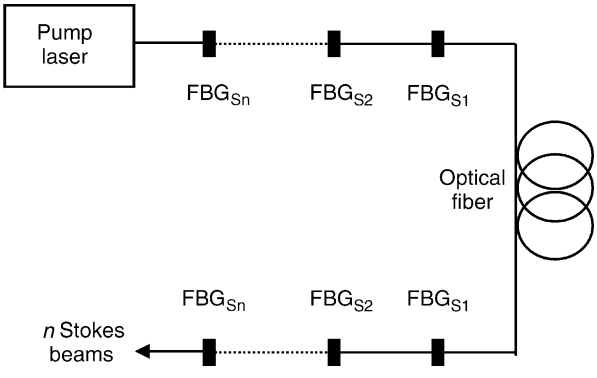


Figure 8 Experimental setup of a continuous-wave, all-fiber, *n*th-order Stokes, Raman oscillator. FBG_{Si} are fiber Bragg gratings highly reflective for the *i*th Stokes wave.

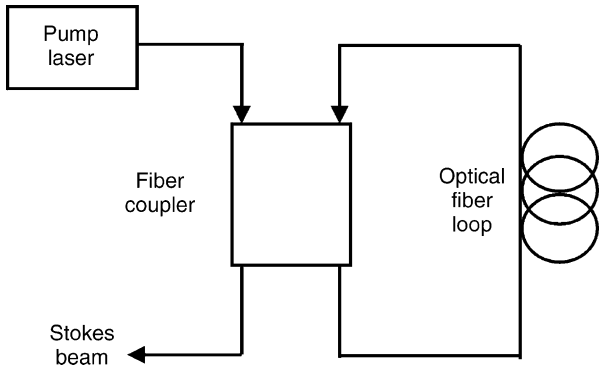


Figure 9 Experimental setup of an all-fiber, ring Raman oscillator.

also been realized by doping the fiber with neodymium, erbium, or ytterbium ions.

The extremely broad Raman spectral linewidth of the glass allows for broadband tuning of the emitted wavelength. For continuous and pulsed wave operation the change in the peak Stokes wavelength can be attained through an intracavity frequency selective element, like for example a prism or a grating. In this case tuning is not obtained in an all-fiber configuration but rather with bulk optical elements and thus additional lenses or objective are needed to collimate the Stokes wave at the fiber ends. The experimental setup is shown in Figure 10.

The laser can be continuously tuned, over about 6 THz, by adjusting the prism or grating rotation. All-fiber Raman laser tuning is possible in a double ring configuration with an inserted optical filter, whose experimental setup is shown in Figure 11. Stokes and pump waves are counterpropagating in the Raman fiber; the Stokes wave frequency is selected by the tuneable optical filter.

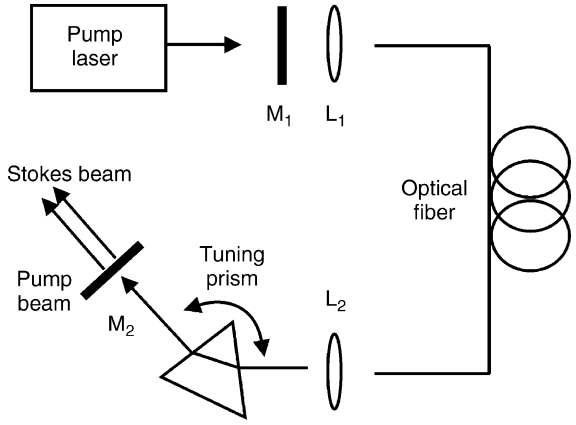


Figure 10 Experimental setup of a tuneable, fiber Raman oscillator. Mirrors M_1 and M_2 are highly reflective for the Stokes wave but not for the pump wave. L_1 and L_2 are collimating lenses or objectives. The tuning prism is rotated to control wavelength selection.

The devices of Figures 10 and 11 stop emitting Stokes radiation for frequencies at which the Raman gain has decreased to a level such that laser threshold is higher than the pump intensity. Due to the fact that Raman gain in fibers is almost constant over about 3 THz, tuning within this frequency range does not lead to any sensible decrease in conversion efficiency. For synchronously pumped ring fiber Raman lasers another technique, named time dispersion tuning, can be used. The experimental setup is shown in Figure 12.

The output fiber end is placed on a moveable mount, that controls the relative delay between the Stokes pulse and the new pump pulse at the fiber input. Since Stokes pulses are dispersed during fiber propagation, to each time slot within a pulse there corresponds a certain portion of the Stokes spectrum. By properly delaying the time of injection of the Stokes pulse back into the fiber the most powerful part of the pump pulse is overlapped to a different part of the Stokes pulse, and this causes a preferred amplification of certain frequencies with respect to others. The shift of the Stokes peak wavelength $\Delta\lambda$ is given by:

$$\Delta\lambda = \frac{\Delta l}{cLD(\lambda)} \qquad [6]$$

where Δl is the translation step, c the speed of light, L the fiber length, and $D(\lambda)$ the fiber dispersion at the Stokes wavelength λ. A total tuning range of 10 nm can be obtained with this technique that can also be used as a test measurement of fiber dispersion.

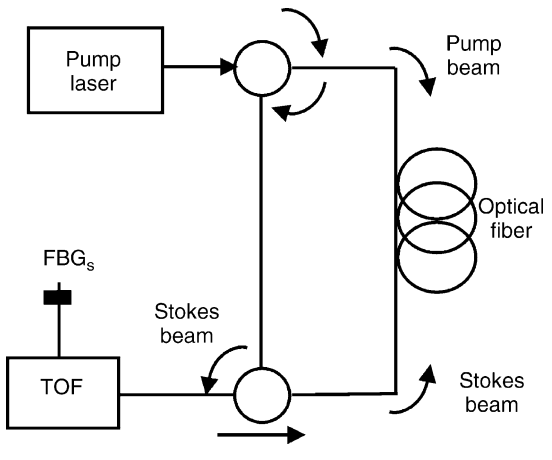


Figure 11 Experimental setup of a double-ring fiber Raman oscillator. OC_1 and OC_2 are fiber optic circulators, and TOF is a tuneable optical filter.

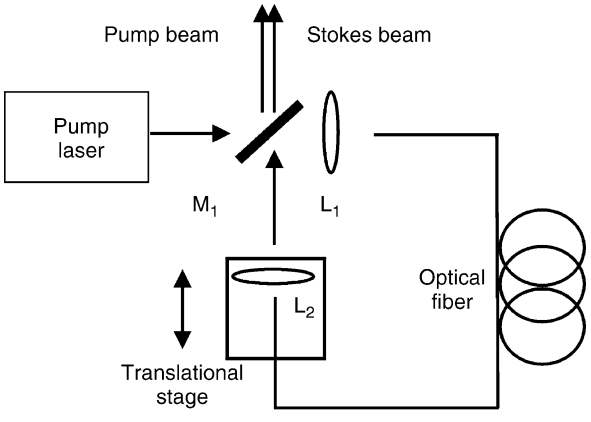


Figure 12 Experimental setup of a time dispersion tuned, ring fiber Raman oscillator. L_1 and L_2 are lenses or objectives. The output collimating lens L_2 and the fiber end are mounted on a translation stage to control wavelength selection by time delaying the re-entrant Stokes pulse.

Applications

Raman lasers have found a large variety of applications. Gas and liquid Raman lasers are fundamental tools for spectroscopic investigation. Solid-state Raman lasers have been shown to be very powerful, eye-safe sources for light detection and ranging (LIDAR) and medical applications. Fiber Raman lasers are exploited for long-haul fiber optics transmission. They can be either sources of short pulses and solitons or high-power continuous-wave sources at multiple wavelengths for subsequent link-distributed Raman broadband amplification.

List of Units and Nomenclature

Amplification gain	[dimensionless]
Efficiency	[dimensionless]
Energy	[J]
Frequency	[Hz]
Intensity	[W m^{-2}]
Length	[m]
Loss coefficient	[m^{-1}]
Power	[W]
Raman bandwidth	[Hz]
Raman gain	[m W^{-1}]
Reflectivity	[dimensionless]
Wavelength	[m]

See also

Fiber and Guided Wave Optics: Dispersion; Light Propagation; Measuring Fiber Characteristics; Nonlinear Effects (Basics); Overview. **Fiber Gratings**. **Optical Amplifiers:** Optical Amplifiers Long-Haul Transmission Systems. **Imaging:** Lidar. **Laser-Induced Damage of Optical Materials**. **Nonlinear Optics, Basics:** Four-Wave Mixing. **Magneto-Optics:** Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering. **Nonlinear Optics, Applications:** Pulse Compression via Nonlinear Optics; Phase Matching; Self-Focusing and Related Effects (Solitons and Multiphoton Absorption). **Optical Materials:** Optical Glasses. **Solitons:** Temporal Solitons. **Scattering:** Raman Scattering; Stimulated Scattering. **Spectroscopy:** Nonlinear Laser Spectroscopy; Raman Spectroscopy.

Further Reading

Band YB, Ackerhalt JR, Krasinski JS and Heller DF (1989) Intra-cavity Raman lasers. *IEEE Journal of Quantum Electronics* QE-25: 208–213.

Basiev TT and Powell RC (eds) (1999) Special issue on solid-state Raman lasers. *Optical Materials* 11: 301–390.

Bloembergen N (1967) The stimulated Raman effect. *American Journal of Physics* 35: 989–1023.

Eckhardt G (1966) Selection of Raman laser materials. *IEEE Journal of Quantum Electronics* QE-2: 1–8.

Grasiuk AZ and Zubarev IG (1978) High power tunable IR Raman lasers. *Applied Physics* 17: 211–232.

Kaiser W and Maier M (1972) Stimulated Rayleigh, Brillouin and Raman spectroscopy. In: Arecchi FT and Schulz-Dubois D (eds) *Laser Handbook*. Amsterdam: North Holland.

Koechner W (1999) *Solid-state Laser Engineering*. Berlin: Springer-Verlag.

Pask HM (2003) The design and operation of solid-state Raman lasers. *Progress in Quantum Electronics* 27: 3–56.

Stolen RH (1978) Fibre Raman lasers. In: Ostrowsky D (ed.) *Fiber and Integrated Optics*, NATO ASI Series B41, pp. 157–182. New York: Plenum Press.

Suto K and Nishizawa J (1994) *Semiconductor Raman Lasers*. Norwood, MA: Artech House.

Weber A (1979) *Raman Spectroscopy of Gases and Liquids*. Topics in Current Physics, vol. 11, Berlin: Springer.

Self-Focusing and Related Effects (Solitons and Multiphoton Absorption)

R L Sutherland, Science Applications International Corporation, Dayton, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

When an intense beam of light with axial symmetry propagates in a medium possessing a nonlinear index of refraction, it will induce its own lens in the medium and be focused. This phenomenon, first discovered just a few years after the invention of the laser, is called self-focusing. It is one of many phenomena under the general classification of self-action, which includes such effects as self-trapping, self-bending, self-phase modulation, and self-steepening. This general class of phenomena involves action on the beam of light induced by the beam itself: light causes a change in an optical property of the medium, which

in turn produces an effect back on the light beam. We will discuss the mechanisms responsible for self-focusing in nonlinear optical media and the resultant action on the optical beam. In this we will see how the topics of self-trapping and self-phase modulation are intimately connected to self-focusing. This will lead to a discussion of related effects as well as many potential applications.

It is instructive to first review the properties of a conventional thin lens, illustrated in Figure 1. By 'thin', we mean that just before and just after the lens, the beam size and shape are approximately the same. However, the phase of the optical wave is changed on passage through the lens. We apply the paraxial (aberrationless) approximation where only rays that lie near the lens axis are considered. In this treatment the spherical surfaces of the lens are approximated as parabolic surfaces. For a symmetric double-convex lens (as shown in Figure 1), this phase change is

$$\Delta\varphi(r) = -k(n-1)\frac{r^2}{R} \equiv -\frac{kr^2}{2f}, \qquad f^{-1} = (n-1)\frac{2}{R} \qquad [1]$$

where n is the index of refraction of the lens material, and R is the radius of curvature of the lens surfaces. The light beam has a wavevector $k = \omega/c$, where ω is the frequency and c the speed of light in a vacuum. The phase retardation is a maximum on the lens axis (z) and decreases quadratically with the radial coordinate r. In other words, the phase advances quadratically with distance from the axis and forms a spherical, converging wave that comes to focus on the z-axis a distance f (the focal length) from the lens.

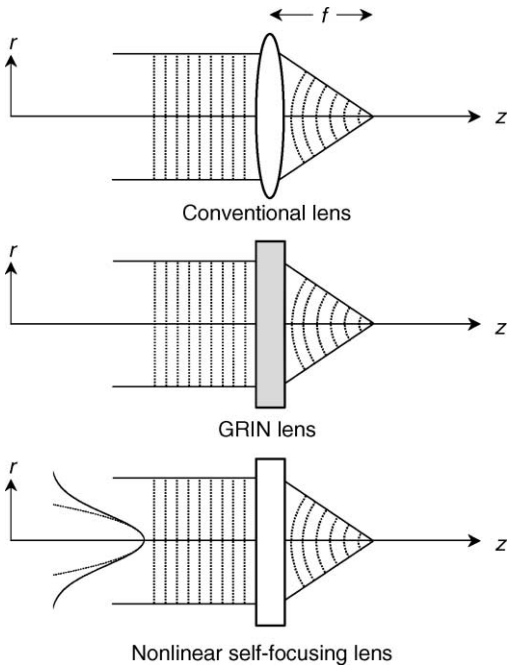


Figure 1 Focusing with a conventional lens, a graded-index (GRIN) lens, and a nonlinear self-focusing lens.

Next consider a graded-index (GRIN) lens, also illustrated in Figure 1. This is a flat slab of material with thickness L and an inhomogeneous index of refraction given by $n(r) = n_0(1 - 1/2\alpha^2 r^2)$. The index has a maximum (n_0) on the lens axis and decreases quadratically with radial distance from the axis. The constant α is a measure of the rate of the index change with r. Upon passage of light through a GRIN lens, the phase change is

$$\Delta\varphi(r) = -kn_0\alpha^2 L\frac{r^2}{2} \equiv -\frac{kr^2}{2f}, \quad f^{-1} = n_0\alpha^2 L \qquad [2]$$

Because of the parabolic nature of the index distribution, the phase retardation again decreases quadratically with r and thus forms a converging spherical wave. Notice that for both the conventional and GRIN lenses, the optical thickness of the lens is parabolic. In the conventional lens the index is constant, but the physical thickness varies continuously. In the GRIN lens, the physical thickness remains constant while the index varies continuously. Both produce the same kind of focusing action.

Finally, consider a flat slab of homogeneous material that has a nonlinear index of refraction. Light of intensity I induces a change in the refractive index given by $\Delta n = n_2 I$. The parameter n_2 is called the nonlinear refractive index coefficient and has physical units of $\text{cm}^2\ \text{W}^{-1}$. There are other definitions of n_2 given in the literature where the index change is given in terms of the square of the electric field, which would obviously have different physical units. We will adopt the intensity form of n_2 in this discussion. We now assume the light incident on this slab is a plane wave with a Gaussian intensity distribution given by $I(r) = I_0 \exp(-2r^2/w^2)$ as shown in Figure 1. I_0 is the peak intensity along the z-axis, and w is the radius of the beam at $1/e^2$ of the peak intensity. Consistent with the paraxial approximation, we expand the exponential function, keeping only the first two terms, to yield $I(r) \approx I_0(1 - 2r^2/w^2)$. This parabolic approximation to the intensity is sketched as the dotted curve in Figure 1. The index distribution induced in the medium by the intensity approximately follows this parabolic shape, and the resultant

phase change is

$$\Delta\varphi(r) = -k4n_2I_0\frac{L}{w^2}\frac{r^2}{2} \equiv -\frac{kr^2}{2f}, \qquad f^{-1} = \frac{4n_2I_0L}{w^2} \qquad [3]$$

Note the similarity to the GRIN lens, with the association of the peak index $n_0 \rightarrow n_2I_0$, and $\alpha^2 \rightarrow 4/w^2$. We can say that the nonlinear medium acts like a GRIN lens with an intensity-dependent focal length. However, this correspondence is only accurate within the parabolic approximation, and the nonlinear lens can introduce significant aberrations to the focal properties of the full beam.

In the paraxial approximation, the conventional, GRIN, and nonlinear lenses all have the same effect on an incident beam of light. Although we have considered only plane parallel incident waves, it should be obvious that the GRIN and nonlinear lenses will produce analogous effects on converging and diverging wavefronts as the conventional lens. The effect of a stack of nonlinear lenses in an optical train can also be analyzed in the same way as a stack of conventional lenses, with the exception that the focal length of each lens will depend on the local intensity and hence size of the beam. Finally, we note that if we change the sign of R, α^2, or n_2, we obtain a negative focal length. This is a diverging lens, and in the nonlinear optics realm the effect is called self-defocusing.

Nonlinear Index of Refraction

The concept of an intensity-dependent refractive index has its roots in the power series expression of the nonlinear dielectric polarization P in terms of the electric field E of the optical wave, which in its simplest form is

$$P = \varepsilon_0\left(\chi^{(1)}E + D^{(2)}\chi^{(2)}E^2 + D^{(3)}\chi^{(3)}E^3 + \cdots\right) = \varepsilon_0\left(\chi^{(1)} + D^{(2)}\chi^{(2)}E + D^{(3)}\chi^{(3)}E^2 + \cdots\right)E \qquad [4]$$

where $\chi^{(m)}$ $(m = 1, 2, 3\ldots)$ is called the m-th order susceptibility. The quantity $D^{(m)}$ is called the degeneracy factor and is 3 for $m = 3$ in the self-focusing case. This leads directly to the definition of a nonlinear susceptibility $\chi(E)$. All materials have first- and third-order susceptibilities. Centrosymmetric materials (e.g., gases, liquids, cubic crystals) do not possess a second-order susceptibility. Here, we will consider only centrosymmetric media and briefly discuss second-order materials later.

The refractive index is $n = [1 + \chi(E)]^{1/2}$. Consequently, the linear index (small electric field) is $n_0 = (1 + \chi^{(1)})^{1/2}$. The leading nonlinear term in $\chi(E)$ is generally small. Expanding the square root in a binomial series, the first nonlinear term in the refractive index can be approximated by $3\chi^{(3)}E^2/2n_0$. Since $I = 2\varepsilon_0 n_0 cE^2$, we see that

$$n = n_0 + n_2I \qquad [5]$$

where $n_2 = 3\chi^{(3)}/4\varepsilon_0 n_0^2 c$. Third-order (cubic) materials that exhibit an intensity-dependent refractive index described by eqn [5] are called Kerr or Kerr-like.

Several physical mechanisms produce a refractive index of the form given in eqn [5]. A list of the significant ones is presented in Table 1. We emphasize that these mechanisms lead to eqn [5] over a limited range of intensity. For example, in materials where the main mechanism is electronic state population redistribution, when the incident intensity becomes comparable with the saturation intensity, the index begins to saturate (i.e., the change of index with

Table 1 Physical mechanisms responsible for the nonlinear index of refraction

Mechanism	*Typical n_2 ($cm^2\ W^{-1}$)*	*Typical sign*	*Typical response time (seconds)*
Electronic polarization	$\sim 10^{-16}$–10^{-15}	+	$\sim 10^{-15}$
Molecular orientation			
Isotropic liquids	$\sim 10^{-15}$–10^{-14}	+	$\sim 10^{-12}$
Liquid crystals	$\sim 10^{-4}$–10^{-3}	+	$\sim 10^{-3}$–10^{-2}
Electrostriction	$\sim 10^{-13}$–10^{-12}	+	$\sim 10^{-9}$
Population redistribution			
Atomic vapors (ω below resonance)	$\sim 10^{-10}$–10^{-9}	−	$\sim 10^{-8}$
Atomic vapors (ω above resonance)	$\sim 10^{-10}$–10^{-9}	+	$\sim 10^{-8}$
Semiconductors	$\sim 10^{-6}$	+	$\sim 10^{-8}$
Band filling (semiconductors)	$\sim 10^{-6}$–10^{-4}	±	$\sim 10^{-8}$–10^{-7}
Exciton saturation (quantum wells)	$\sim 10^{-4}$	−	$\sim 10^{-8}$
Thermal	$\sim 10^{-8}$	±	$\sim 10^{-4}$

intensity becomes sublinear). The simple expression in eqn [5] no longer applies. These types of effects can be important and will be discussed later. Finally, we mention that other mechanisms exist that produce self-focusing related effects, but they are not Kerr-like and will be briefly discussed later.

We learn from Table 1 that values of n_2 range over several orders of magnitude. High-intensity lasers are required to observe the effects of small optical nonlinearities. Hence pulsed lasers are often used in self-focusing applications. Laser pulse widths can generally range from 100 s of nanoseconds to 10 s of femtoseconds. Consequently, one must keep in mind the response time of the mechanism/material when analyzing experiments. If the pulse width is long compared to the response time, a quasi-steady-state self-focusing regime results. If the pulse is too short, transient effects will appear. Also, as one goes to longer response time mechanisms, lower intensities can be used. In some cases it is even possible to use cw lasers.

Self-Lensing

Thin nonlinear media can be employed in self-lensing applications, where self-focusing is used to modify the far field properties of a beam, just as a conventional lens would do, but with effects dependent on the beam intensity. Gaussian beams are usually employed in the paraxial approximation.

A popular technique employing this effect to measure the magnitude and sign of n_2 is the Z-scan. A typical experiment is illustrated in Figure 2. A thin nonlinear medium is scanned axially between a focusing lens and an aperture in the far field. Consider a material with positive n_2. When the sample is close to the lens the intensity is too low to induce significant self-lensing. As the sample approaches the focal point of the lens, the intensity rises and a positive self-lens is induced that focuses the beam more strongly. This displaces the focus to the left, causing the beam to spread out more in the far field and less power to pass through the aperture. As the sample passes to the right of the focal point, the positive self-lens reduces the divergence of the beam so that more power is passed through the aperture. Near the focal point the self-lens is strongest, but a thin lens at this point has no effect on the beam passing through the far-field aperture. The measured aperture transmittance as a function of scan-position Z is sketched in Figure 2. The peak and valley of this dispersion-shaped curve will be interchanged for a negative n_2. For a small aperture and fixed power of the laser, the difference between the peak and valley transmittance is proportional $|n_2|$.

The same experimental configuration is also popular for measuring small absorption coefficients or thermo-optic coefficients ($\mathrm{d}n/\mathrm{d}T < 0$ for most materials). Under steady-state conditions (cw beam), the thermal n_2 is given by

$$n_2 = \frac{\alpha\tau}{\rho C}\frac{\mathrm{d}n}{\mathrm{d}T} \qquad [6]$$

where α is the absorption coefficient, τ is the thermal diffusion time, ρ is the mass density, and C is the specific heat. This technique is extremely sensitive and can easily be observed, for example, with a ~1 mW laser pointer and a 1 cm cuvette filled with minute quantities of ink or food coloring dissolved in methanol.

This combination of self-lensing and a hard aperture has also been put to use in devices. One such device is the optical power limiter. A typical configuration is similar to the Z-scan experiment in Figure 2, with the sample fixed near the lens focal point. Thick media in the focal region have also been used. At sufficiently low optical power, the medium acts like a flat window. As the input power rises, self-lensing sets in and spreads the beam out at the aperture, limiting the transmitted power. The higher the incident power gets, the more the beam spreads. The result is a transmitted power that is clamped below a preferred level. Thermal lensing has been used for cw beams, while positive Kerr

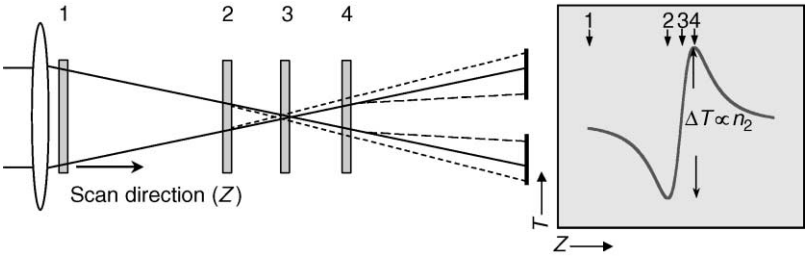


Figure 2 A Z-scan experiment. The transmittance through the aperture as a function of the position (Z) of the thin nonlinear medium is sketched at the right.

self-lensing has been employed for Q-switched lasers.

A less obvious but very important application of this configuration is for mode-locking lasers (see Figure 3). Mode-locking is a coherent phenomenon that locks the relative phase of multiple longitudinal modes in a laser having a broadband gain profile. These stabilized modes interfere to form a train of short pulses with pulsewidth inversely proportional to the gain bandwidth. A common technique for passively mode-locking a laser is to insert a saturable absorber into the laser cavity. A saturable absorber blocks low-power light, keeping the round-trip gain below the oscillation threshold, but transmits high-power light as the absorption coefficient saturates. Thus the mode-locked condition is favored for gain and oscillation since the peak power of the pulses is high. It turns out that transverse spatial modulation, due to self-focusing combined with an aperture, can produce the same type of effect in a laser cavity. The laser rod acts as both the gain medium and nonlinear lens. For example, titanium: sapphire ($Ti:Al_2O_3$) has $n_2 = 3 \times 10^{-16}\,cm^2\,W^{-1}$. Thus, in a $Ti:Al_2O_3$ laser the rod produces self-lensing that modifies the beam shape propagating through an intracavity aperture. The loss due to the aperture is intensity dependent, as in the Z-scan. Therefore, the net round-trip gain of the cavity is low for low-power but high for high-power pulses, and the mode-locking condition is fulfilled. These types of lasers are called Kerr lens mode-locked (KLM) lasers. KLM lasers are generally not self-starting, but can be self-sustaining. They, therefore, require some active means of starting the mode-locking. Ultrashort pulses (<100 fs) have been obtained with a cw KLM $Ti:Al_2O_3$ laser. For pulsewidths this short, the simple quasi-steady-state self-focusing picture is not entirely adequate to describe the dynamics. Other factors, such as group-velocity dispersion (GVD) and self-phase modulation (SPM), must be considered. Moreover, since the laser rod is typically several millimeters long, it is more appropriate to model it as a thick rather than a thin nonlinear lens.

Beam Trapping

When an intense beam propagates in a thick nonlinear medium, a question arises as to the ultimate fate of the beam. A possibility that has intrigued researchers for a long time is diffractionless propagation of a beam over long distances. In other words, self-focusing exactly balances diffraction to force the beam into a nonlinear waveguiding mode. This phenomenon is called beam trapping or self-trapping. A simplified analysis of a uniform beam captures the essential physics (see Figure 4). The refractive index inside the beam is higher than outside by eqn [5]. Therefore, light incident on the 'interface' at the edge of the beam can experience total internal reflection if its angle of incidence exceeds the critical angle, dependent on the value of the nonlinear index. If the complement of the critical angle matches the diffraction angle of the beam, the beam is trapped in its own self-induced waveguide. The critical power $\mathscr{P}_{cr2}$ at which this occurs is given in Table 2. The condition for beam trapping is independent of beam size since both diffraction and self-focusing depend on beam area, and these size effects cancel.

A detailed solution of the wave equation for a Gaussian beam using the parabolic approximation yields a slightly different critical power $\mathscr{P}_{cr1}$, also given in Table 2. Numerical analyses of the wave

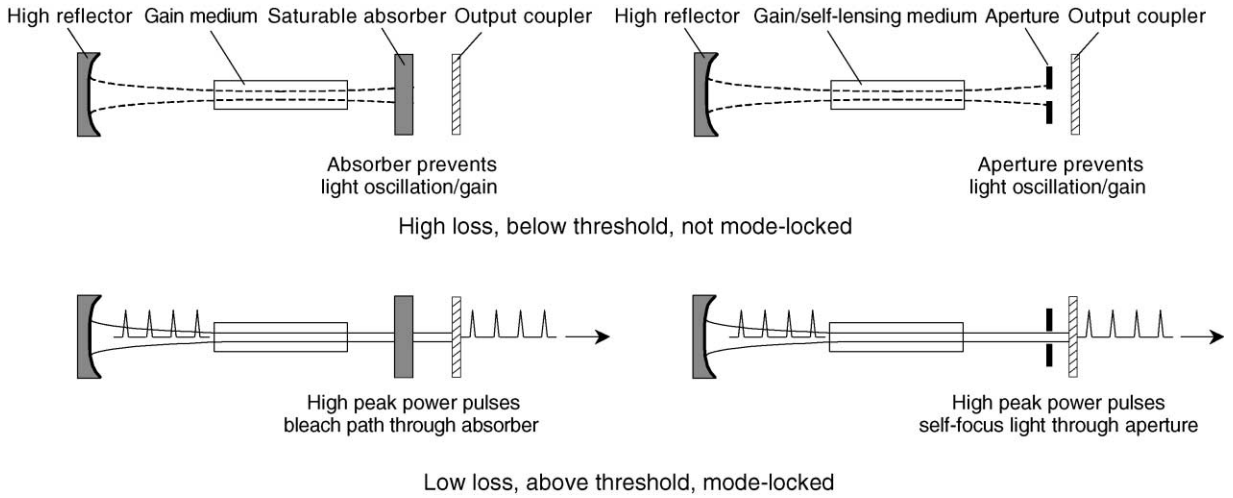


Figure 3 Laser mode-locking using a saturable absorber (left) and self-lensing in combination with an intracavity aperture (right).

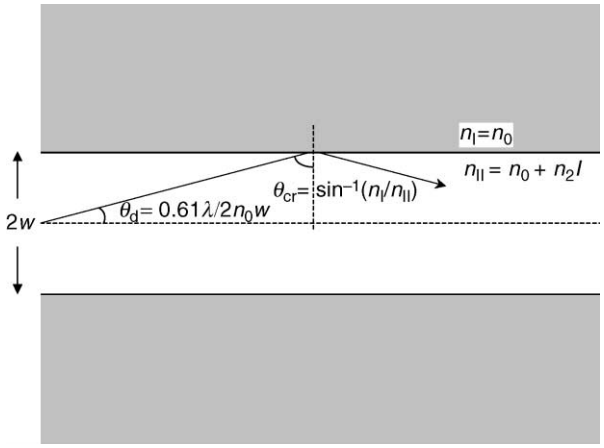


Figure 4 Self-trapping conditions for a uniform beam of intensity I and width $2w$ in a Kerr medium.

Table 2 Critical power for beam trapping (collapse)

Gaussian beam (parabolic approximation)	$\mathscr{P}_{cr1} = \frac{1}{2}\frac{\lambda^2}{4\pi n_0 n_2}$
Gaussian beam ('whole beam' collapse)	$\mathscr{P}_{cr2} \approx \frac{(1.22\pi)^2}{8}\frac{\lambda^2}{4\pi n_0 n_2}$
Townes soliton, lower bound for arbitrary beam shape	$\mathscr{P}_{crT} \approx 1.86225\frac{\lambda^2}{4\pi n_0 n_2}$
Arbitrary beam (upper bound)[a]	$\mathscr{P}_{cr3} = G[f]\frac{\lambda^2}{4\pi n_0 n_2}$

[a] $G[f]$ defined in text.

equation for a whole Gaussian beam have shown for incident power such that $\mathscr{P}_{cr1} < \mathscr{P} < \mathscr{P}_{cr2}$ the beam comes to a 'soft' focus and then diffracts (partial beam focusing), but for $\mathscr{P} > \mathscr{P}_{cr2}$ the whole beam catastrophically collapses (i.e., the beam approaches a singularity). Catastrophic self-focusing has been observed experimentally, but as the beam collapses the intensity becomes so large that additional nonlinear effects, such as multiphoton absorption and stimulated Raman scattering, come into play. These effects rob the beam of energy and terminate the collapse. At this stage the beam will then spread by diffraction.

A more rigorous analysis of this problem is obtained by examining the nonlinear wave equation in the slowly varying amplitude approximation (propagation along z), but including the transverse variation of the wave. If the electric field of the optical wave is of the form $E(\mathbf{r}) = A(\mathbf{r})\exp(ikz)$, where now $k = n_0\omega/c$, then the wave equation can be written as

$$2ik\frac{\partial A}{\partial z} + \nabla_\perp^2 A + 4\varepsilon_0 n_2 c k^2 |A|^2 A = 0 \qquad [7]$$

where the transverse Laplacian $\nabla_\perp^2$ is $\partial^2/\partial x^2 + \partial^2/\partial y^2$ in rectangular Cartesian coordinates, or $\partial^2/\partial r^2 + (1/r)\partial/\partial r$ in cylindrical coordinates (assuming azimuthal symmetry). This equation is known as the nonlinear Schrödinger equation (NLSE). In normalized coordinates $\xi = x/w$, $\eta = y/w$, $\rho = r/w$, $\zeta = z/2kw^2$, where w is the input beam radius, and $\psi = 2kw(c\varepsilon_0 n_2)^{1/2}A$, eqn [7] takes the normalized form of the NLSE:

$$i\frac{\partial\psi}{\partial\zeta} + \nabla_\perp^2\psi + |\psi|^2\psi = 0$$
$$\nabla_\perp^2 = \frac{\partial^2}{\partial\xi^2} + \frac{\partial^2}{\partial\eta^2} = \frac{\partial^2}{\partial\rho^2} + \frac{1}{\rho}\frac{\partial}{\partial\rho} \qquad [8]$$

A waveguiding (i.e., diffractionless) solution of the form $\psi(\zeta,\rho) = \exp(i\zeta)R(\rho)$ exists for eqn [8], where $R(\rho)$ is a solution of

$$\nabla_\perp^2 R - R + R^3 = 0, \quad R'(0) = 0, \quad R(\infty) = 0 \qquad [9]$$

$R(\rho)$ is known as the Townes soliton. It has a radial profile similar to a Gaussian but differing in fine detail. Since the Townes soliton exactly balances diffraction and self-focusing, it carries a critical power $\mathscr{P}_{crT}$ given by the expression in Table 2. For comparison, $(1.22\pi)^2/8 = 1.83624$. Therefore, $\mathscr{P}_{crT}$ and $\mathscr{P}_{cr2}$ differ by only about 1%. However, the Townes soliton is unstable. An analysis of the solutions of the NLSE shows that a beam will not catastrophically collapse (i.e., its intensity will not blow up) if its power is less than $\mathscr{P}_{crT}$. Thus $\mathscr{P}_{crT}$ forms a lower bound for the critical power. An upper bound can also be found from analysis of the NLSE. It depends on the input beam profile [$\psi(\rho,0) \propto f(\rho)$] and is given as $\mathscr{P}_{cr3}$ in Table 2. The quantity $G[f]$ characterizes the input beam profile and is given by

$$G[f] = \frac{2\int |f|^2\rho\mathrm{d}\rho \int |\nabla f|^2\rho\mathrm{d}\rho}{\int |f|^4\rho\mathrm{d}\rho} \qquad [10]$$

For example, a Townes soliton, a Gaussian beam, and a hyperbolic secant beam yield $G = 1.86225$, $G = 2$, and $G = 1.86257$, respectively. Thus, if the beam power is such that $\mathscr{P}_{crT} \le \mathscr{P} \le \mathscr{P}_{cr3}$, the beam may collapse, but it will definitely collapse if $\mathscr{P} > \mathscr{P}_{cr3}$. We note that if the input beam is elliptical, with ellipticity $e = b/a$ (where a and b are the semi-major and semi-minor widths, respectively), then

$$\mathscr{P}_{crT}(e) \approx \left(0.6 + 0.2\frac{e^2+1}{e}\right)\mathscr{P}_{crT} \qquad [11]$$

This expression is accurate to within 1.5% of numerical simulations of the NLSE, and also holds well for Gaussian input beams.

The conclusion is that there is no stable self-trapped (2 + 1)-D beam (two diffracting dimensions, one

propagating dimension) in a Kerr medium. The beam will diffract, blow up in a finite distance, or go through a single focusing–defocusing cycle. For an incident plane wave, the distance from the input plane to the blowup point is (assuming a Gaussian input)

$$z_f \approx \frac{0.369\pi w^2/\lambda}{(\mathscr{P}/\mathscr{P}_{cr2})^{1/2} - 0.858} \qquad [12]$$

where the approximation holds well for $\mathscr{P} > 1.2 \times \mathscr{P}_{cr2}$.

Why, then, does self-trapping appear to occur in Kerr media? Time integrated photographs show what appears to be the formation of a filament, and laser damage tracks in solids follow the shape of a self-trapped filament. This can be explained in terms of the moving focus model. In the quasi-steady-state approximation, each time slice of the pulse comes to focus independently at different times and positions. Figure 5 shows a series of snapshots of a Gaussian spatial and temporal pulse (assume axial symmetry), where the peak power of the pulse is greater than $\mathscr{P}_{cr2}$. Slices of the pulse with $\mathscr{P} > \mathscr{P}_{cr2}$ will come to a focus at a finite distance $z_f(\mathscr{P})$, which can be approximated by eqn [12]. The time at which that focus appears is determined by the time it takes for that slice of power $\mathscr{P}$ to propagate to the input plane ($z = 0$) plus propagate to $z_f(\mathscr{P})$. As illustrated in Figure 5, slice 1a takes the minimum time to come to a focus. At a later time slices 2a and 2′ come to focus simultaneously but at different positions, 2a to the left and 2′ to the right of 1a, since $\mathscr{P}(2a) > \mathscr{P}(1a)$ and $\mathscr{P}(2') < \mathscr{P}(1a)$. The solid curves in the left of Figure 5 are the trajectories of the beam rays showing where each slice will come to focus. Slice 2b, having the same power as 2a, comes to focus at the same position as 2a, but later in time, and so on. The original focus thus appears to split in two, with one branch propagating forward and the other backward. The backward branch comes to a turning point when slice 4 of maximum power in the pulse center comes to focus, then propagates forward as time slices in the back part of the pulse come to focus. The motion of the foci traces out the U-shaped curve shown to the right in Figure 5. The time-integrated picture that appears like a filament is thus just the tracks of moving foci.

The question remains, do conditions exist under which stable self-trapped beams can form? The answer is yes, and these types of beams are called spatial solitons. The $(1 + 1)$-dimensional NLSE admits the fundamental spatial soliton solution:

$$A(x, z) = A_0 \mathrm{sech}(x/w) \exp(iz/2kw^2)$$
$$|A_0|^2 = \frac{1}{2\varepsilon_0 c n_2 (kw)^2} \qquad [13]$$

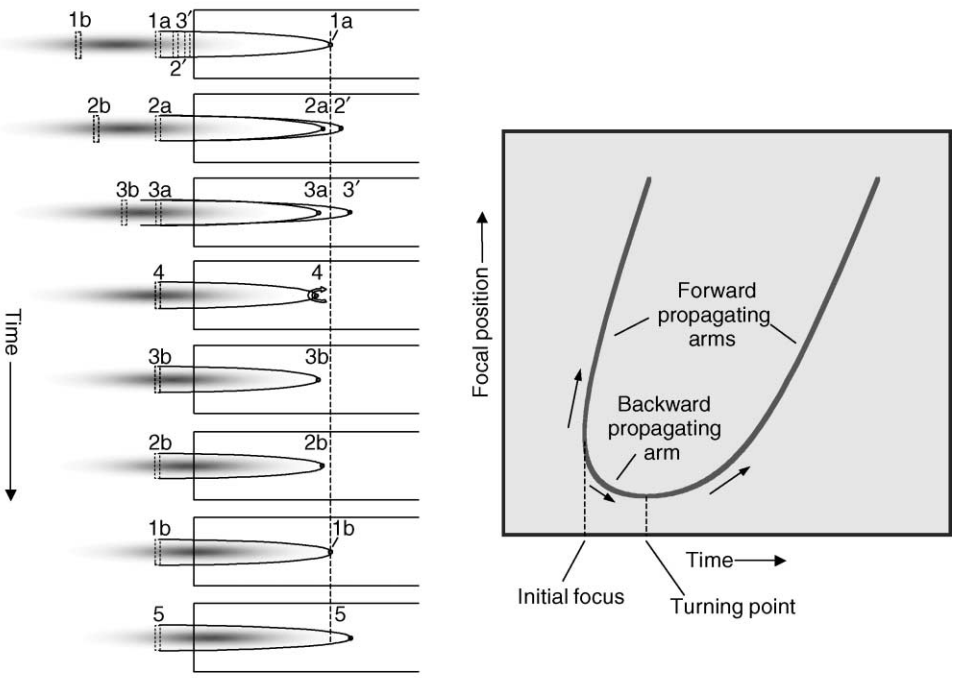


Figure 5 Moving focus model of apparent filament formation by quasi-steady-state self-focusing of a pulsed Gaussian beam. The series of snapshots on the left illustrate when and where different time slices of the pulse come to focus. The tracks of moving foci trace out the curve shown on the right.

The wave given by eqn [13] propagates along z with a constant width w along the x-axis. Such a $(1+1)$-D soliton can be formed in a slab waveguide, as illustrated in Figure 6, consisting of a nonlinear thin film deposited on a linear slab, where the wave is confined along y but can diffract along x. These types of solitons have been confirmed experimentally and shown to be quite robust. They are stable against perturbations and propagate as a self-induced waveguide. Since a soliton changes the refractive index where it propagates, it can also attract and guide other beams. The interactions or 'collisions' of solitons are an active area of current research. Intriguing potential applications include all-optical reconfigurable interconnects for telecommunications and optical computing based on the properties of soliton collisions.

Although $(2+1)$-D spatial solitons are not stable in ideal Kerr media, they can be in Kerr-like media with a saturating nonlinearity. For example, a two-level saturable absorber (e.g., an atomic vapor) at a wavelength tuned above the natural absorption frequency will exhibit a positive Kerr-like response at low intensity ($I \ll I_{sat}$), but then saturates as $(1+I/I_{sat})^{-1}$ for $I \sim I_{sat}$ or larger, where I_{sat} is the saturation intensity. This can form a stable trapped beam, as was first reported in 1974 for a cw laser in Na vapor, because as the beam self-focuses and the intensity rises, the nonlinear index levels off asymptotically until a balance between diffraction and focusing is reached. This type of situation is stable against perturbations, but may undergo almost periodic focusing–defocusing cycles.

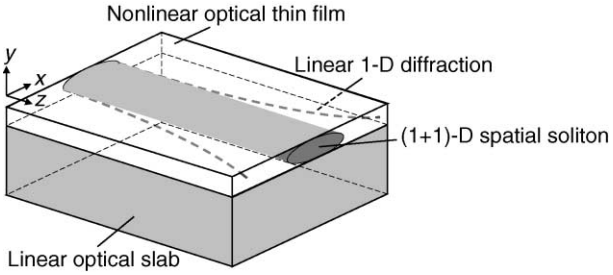


Figure 6 Spatial soliton propagation in a nonlinear slab waveguide.

Spatial solitons employing mechanisms other than the Kerr nonlinearity are also being vigorously studied. One of these is the photorefractive effect. An illustration of the generation of a slab soliton is shown in Figure 7. A Gaussian beam is focused to a line (e.g., by a cylindrical lens) and propagated as an extraordinary wave along the optic (z) axis of a uniaxial photorefractive crystal. A dc electric field is also applied along z. Charge carriers are generated in the slab beam by photo-ionization of dopant atoms and drift in the dc field. The lower resistivity in this region causes the voltage drop to be smaller and hence the field to be lower. The field induces a refractive index change by the Pockels effect ($\Delta n = -n_0^3 |r_{eff} E|/2$, where r_{eff} is the effective Pockels coefficient and n_0 is the background index). The index change is negative, but is reduced in magnitude where the field is lower. The index change depends on the local intensity, and the

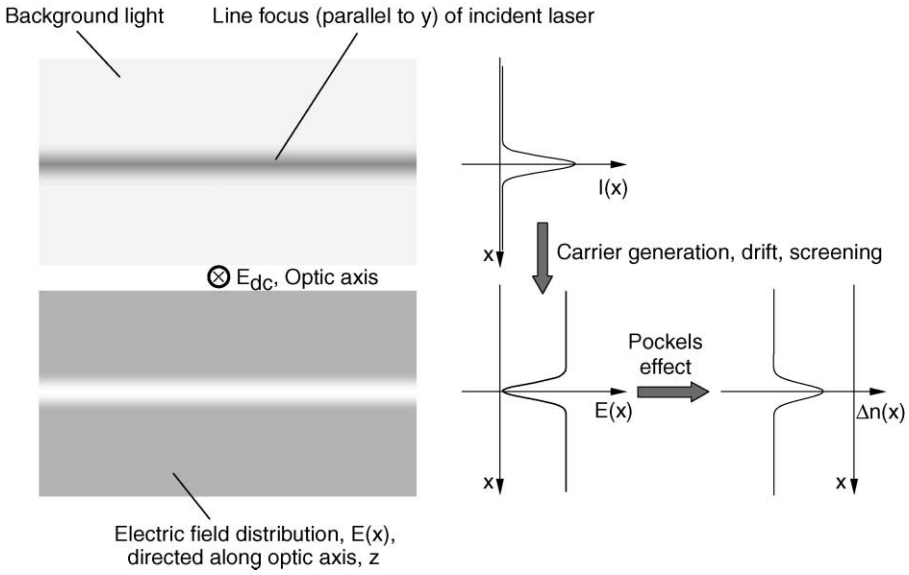


Figure 7 Mechanism of slab soliton formation in a photorefractive material.

photorefractive mechanism produces an effective saturable self-focusing nonlinearity. (There are a limited number of charge carriers, and the field cannot be reduced below zero.) This produces a waveguide that traps the slab beam, which then propagates without spreading. Thus the photorefractive soliton is stable. This effect can be observed with milliwatt-level cw lasers, but the space-charge field takes time to build up to a steady state (typically of the order of seconds). Needle (circularly symmetric) solitons have also been observed by launching a TEM_{00} Gaussian beam in a photorefractive crystal, but the theoretical description is not as simple. Other types of spatial solitons and other mechanisms have been explored using the photorefractive and photovoltaic effects. The reader is referred to the literature for a more in-depth discussion.

Another mechanism useful for soliton formation is cascaded second-order (quadratic) effects. This includes such phenomena as second-harmonic generation (SHG) and other frequency conversion processes. Consider SHG with imperfect phase matching. (The reader is referred to other articles discussing harmonic generation and phase matching.) A fundamental (F) wave generates its second harmonic (SH), but because of dispersion $n(\omega)$, the two waves travel at different phase velocities. The SH(2ω) beats with the F(ω) wave to generate, via the second-order nonlinear polarization, another wave at ω that adds to the incident F wave. This produces a net phase shift of the F wave that accumulates continuously as the waves propagate together and periodically exchange power; this is called cascading. In the nondepleted pump approximation, this phase accumulation can be expressed as

$$\Delta\varphi = -\frac{\Delta k z}{2}\left\{1 - \left[1 + \left(\frac{2\kappa\sqrt{I}}{\Delta k}\right)^2\right]^{1/2}\right\} \qquad [14]$$

$$\kappa = \frac{\omega\chi^{(2)}_{\text{eff}}}{\sqrt{2\varepsilon_0 n_0^3 c^3}}$$

where $\chi^{(2)}_{\text{eff}}$ is an effective second-order susceptibility (dependent on crystal symmetry, polarization vectors, and propagation direction), $\Delta k = 2k(\omega) - k(2\omega)$ is the phase mismatch, I is the intensity of the F wave, and n_0 is the approximate refractive index for both F and SH waves. Although not due to a refractive index change, this phase change produces a self-focusing effect for a bell-shaped intensity. For a small enough pump intensity or sufficiently large phase mismatch, an equivalent n_2 can be approximated by

$$n_2 = \frac{\omega(\chi^{(2)}_{\text{eff}})^2}{2\varepsilon_0 n_0^3 c^2 \Delta k} \qquad [15]$$

Note that the sign of n_2 depends on the sign of Δk. For $I \sim (\Delta k/2\kappa)^2$ or larger, eqn [15] is not valid and eqn [14] must be used. Note that the equivalent index change is then sublinear with respect to I, and soliton formation is stable. When pump depletion cannot be ignored, the F and SH coupled-wave equations must be solved numerically to analyze soliton dynamics. Figure 8 illustrates the difference between a beam diffracting normally and a diffractionless quadratic soliton. When an F beam is launched in a nonlinear medium, it does not immediately form a soliton, but evolves into one, shedding some energy in the process. Note that this soliton does not consist of a single frequency beam, but depends on the nonlinear coupling between the ω- and 2ω-waves that continuously exchange energy as the beams propagate. For this reason, these solitons are called simultons.

Self-Focusing in Space–Time

The moving focus model describing the self-focusing of optical pulses becomes inadequate when dispersion effects must be included. This is true in the case of ultrashort pulses (typically ~100 fs or less). To properly describe the propagation of a pulse with group velocity v_g, eqn [7] must be modified by $\partial A/\partial z \rightarrow [\partial/\partial z + (1/v_g)\partial/\partial t]A$. In addition, to include the effects of GVD, a term $-k_2 k(\partial^2 A/\partial t^2)$ is added, where $k_2 = (\partial^2 k/\partial\omega^2)$ is the GVD coefficient. For anomalous GVD, $k_2 < 0$. Taking this to be the case and defining another normalized coordinate $\tau = (t - z/v_g)/(|k_2|k)^{1/2}w$, the normalized NLSE (eqn [8]) becomes

$$i\frac{\partial\psi}{\partial\zeta} + \left(\frac{\partial^2}{\partial\xi^2} + \frac{\partial^2}{\partial\eta^2} + \frac{\partial^2}{\partial\tau^2}\right)\psi + |\psi|^2\psi = 0 \qquad [16]$$

Note the complete symmetry in eqn [16] between the transverse space (ξ, η) and time (τ) variables. There is thus a correlation between diffraction (space) and dispersion (time). The corresponding phenomenon to self-focusing is self-phase modulation (SPM). This effect denotes the additional phase retardation experienced by a pulse in a Kerr medium due to its time-dependent intensity: $\Delta\varphi(t) = -(\omega_0/c)n_2 I(t)z$, where ω_0 is the center (carrier) frequency of the pulse. SPM produces new frequencies ($d\varphi/dt$) that are red-shifted on the leading edge of the pulse and blue-shifted on the trailing edge. The frequency shifts in a self-focused filament can be quite sizeable and create a large continuum (a white light pulse) that

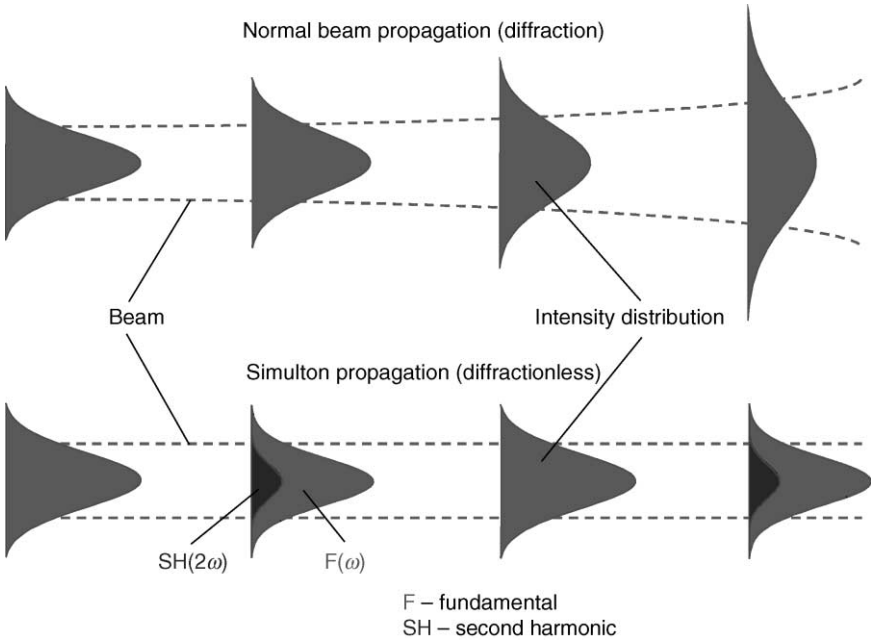


Figure 8 Comparison of normal beam propagation to simulton propagation in a quadratic nonlinear medium. In a simulton, the F and SH waves periodically exchange energy as they co-propagate in the medium.

can be used for ultrafast spectroscopy. In a medium with normal GVD the pulse spreads in time because blue frequencies in the trailing edge travel slower than red frequencies in the leading edge. However, for anomalous GVD the opposite is true, and the pulse can 'focus' or collapse in time.

The simultaneous spatiotemporal collapse of a beam of ultrashort pulses leads to the consideration of $(3 + 1)$-D soliton generation, where the third 'transverse' dimension is time. While the NLSE leads to unstable solitons for transverse dimensions >1, a saturating nonlinearity could possibly bound the collapse and lead to the formation of stable, diffractionless, dispersionless pulses. These types of pulses are called light bullets. Light bullets have a particle-like nature and display many interesting, exotic properties that could lead to ultrafast switching rates for optical digital logic systems in telecommunications and computing.

The saturable nature of self-focusing in quadratic nonlinear media makes them natural candidates in the search for light bullets, and $(2 + 1)$-D spatiotemporal simultons have been observed in crystals of lithium iodate. The drawback of ultrashort pulse simulton generation in quadratic media is the added complications of requiring anomalous GVD at both ω and 2ω, and dealing with the effects of group-velocity mismatch (GVM). The phenomenon of GVM involves the different group velocities of the F and SH pulses, which makes them walk away from one another. Without the physical overlap of these two pulses, the simulton will cease to exist. By tilting the amplitude front of the pulses with respect to their phase fronts, it is possible (over a limited distance) to compensate pulse walkoff due to GVM with spatial walkoff due to the birefringence of the quadratic medium. Accomplishing this, however, requires the beam to be in the form of a narrow stripe, and the propagation distance over which pulse overlap is good is greater for a longer stripe. This can be achieved by focusing a beam onto the medium with a cylindrical lens. The simulton then propagates without a change of width temporally or spatially (in 1-D). This is not a light bullet, however, because the confinement is not in all three dimensions.

The elliptic nature of the diffraction–dispersion operator in eqn [16] makes pulse collapse a possibility in nonlinear media with anomalous GVD. With normal GVD, the operator is hyperbolic. For such a medium, it has been predicted and observed that instead of collapsing, a self-focused pulse in a Kerr medium will split in time. Since the red-shifted frequencies generated at the front of the pulse travel faster than the blue-shifted frequencies produced at the back, the front of the pulse separates from the back part. Collapse of the pulse, as well as self-trapping, is not favored under conditions of pulse splitting.

Given this property of the NLSE, it is somewhat surprising that the reportedly first observation of light bullets was in a quadratic medium with normal GVD.

Focusing a beam of Gaussian profile in both space and time near the entrance to a lithium triborate crystal under the condition of a large positive phase mismatch, the pulses propagated without spread temporally or spatially (2D) over a distance ~20 mm (the length of the crystal). However, the pulse did not retain its shape, but evolved into a so-called 'X-wave', or X-shaped soliton. A generic X-shaped soliton (or, strictly speaking, simulton) is shown in Figure 9. The pulse has a conical (clepsydra, or hourglass-shaped) 3D structure with a central hump and slowly decaying conical tails. Interestingly, an X-shaped pulse has been shown to propagate without spreading in a linear optical medium exhibiting normal GVD. But such a pulse must be specially contrived at the input and may be impossible in many cases. In a quadratic nonlinear medium with normal GVD, the X-shaped light bullet appears to form spontaneously. The physics of this evolution lies buried in the coupled wave equations for the fundamental and second harmonic. In the strong phase-mismatched SHG process, weak perturbations in the waves at specific spatial and temporal frequencies can grow exponentially. The hyperbolic nature of the diffraction–dispersion operator with normal GVD leads to amplification of conical wave perturbations with frequencies approaching that of the X-wave. This amplification of the proper frequency components acts as a trigger that drives the evolution of the Gaussian spatiotemporal input pulse into the X-wave. This basic shape has been confirmed in experiments that outline the spatiotemporal intensity profile of the pulse exiting the rear of the crystal.

Other Related Effects (Multiphoton Nonlinearities)

At this point we should note that the susceptibility is frequency dependent. Optical waves couple principally to bound electrons, which can only occupy discrete energy levels. When the photon energy $\hbar\omega$ is nearly resonant with a difference between these energy levels, the dielectric response of the medium is enhanced. This well-known property of linear optics, accounting for absorption of light as well as normal and anomalous dispersion, is also applicable in nonlinear optics. The third-order susceptibility is actually a complex number. A sketch of its real and imaginary parts is given in Figure 10, the real part relating to n_2. The imaginary part is associated with optical loss, if positive, or gain when it is negative. In the present case the frequency dependence of the susceptibility displays a resonance (i.e., peak of the imaginary part) at a frequency $\omega = 2\omega_0$. Referring to Figure 10, when the optical frequency is significantly less than ω_0, the imaginary part of $\chi^{(3)}$ is small compared to the real part and nonlinear refraction dominates. However, when the frequency is comparable to ω_0, the resonance at $2\omega_0$ plays a major role. The real part of $\chi^{(3)}$ is resonantly enhanced, but the imaginary part signifies loss: two photons are simultaneously absorbed. This is called two-photon absorption (2PA). The intensity-dependent absorption loss per unit length is given by $\alpha_2 I$, where α_2 is called the two-photon absorption coefficient (often designated by β) and has physical units of cm W^{-1}.

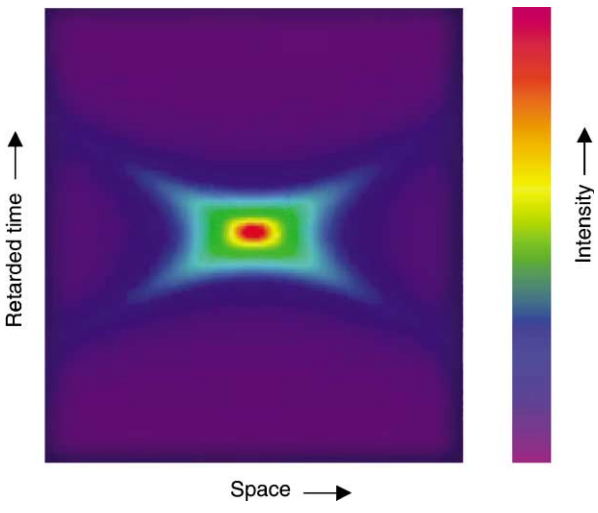


Figure 9 Generic shape of an X-wave light bullet as a function of space and retarded time ($t - z/v_g$) coordinates. The light bullet has radial symmetry.

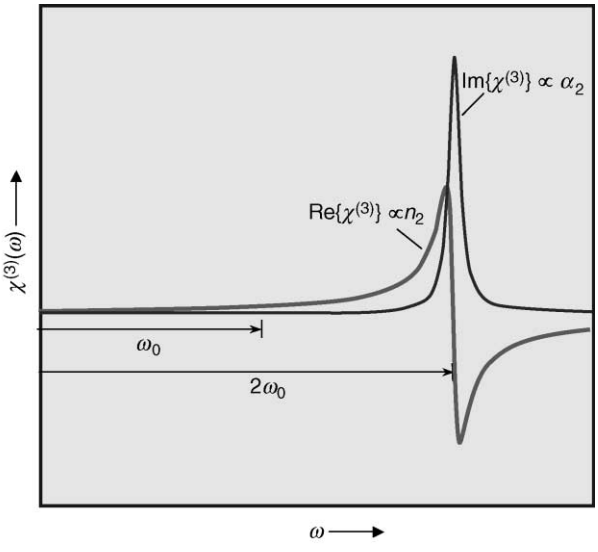


Figure 10 Sketch of the real and imaginary parts of the frequency-dependent third-order susceptibility $\chi^{(3)}$ illustrating a two-photon resonance.

In many materials at a variety of optical frequencies, 2PA will naturally accompany self-focusing. Since these two phenomena are both intensity dependent, while one decreases the intensity and the other increases it, there is an obvious interplay between the two. We illustrate the effect of 2PA in Figure 11 by comparison of the phase retardance of a Gaussian beam (in the parabolic approximation) with and without 2PA in a 'thin' medium, for which the size of an optical beam is nearly constant (diffraction ignored). Obviously, $\Delta\varphi(r)$ does not reproduce the parabolic shape when strong 2PA is present. Actually, it is possible to have multiphoton absorption (MPA) of m photons with corresponding coefficient α_m (related to $\chi^{(m)}$). The effects of 3PA and 4PA are also shown in Figure 11. 2PA, or more generally MPA, distorts the phase and modifies the focusing properties of the beam. This shows up in the Z-scan, for example, as a distortion of the dispersion-like transmittance curve: the valley is deeper and the peak is suppressed.

The reason for this phase distortion is that nonlinear absorption changes the shape of the beam: absorption is stronger where the intensity is higher. For MPA, the intensity at position z in a thin medium is

$$I(r,z) = \frac{I(r,0)}{\left[1 + (m-1)\alpha_m I^{m-1}(r,0)z\right]^{1/(m-1)}} \qquad [17]$$

The other obvious effect of MPA is that it reduces the strength of self-focusing since it robs power from the beam. The interplay of the two effects shows up more in thick media where the beam can diffract and change size. When self-focusing is dominant at the front of the medium, the intensity will start to blow up as the beam collapses, but the intensity becomes so large that MPA becomes much stronger. This robs energy from the beam and can stop or reverse the collapse. When MPA is dominant at the start, the beam will still self-focus, but MPA flattens the beam in the center causing the edges of the beam to experience dramatic diffraction effects; for example, a ring structure can develop.

The intensity of a self-focused filament can be quite high. At a level $\sim 10^{11}$ W cm^{-2} or higher, it is possible for MPA to lead to multiphoton ionization (MPI) of the medium. This is illustrated in Figure 12 where the absorption of m photons promotes an electron from the ground state to the continuum in a gas, or from the valence band to the conduction band in a solid. The ionization rate is $\sigma_m I^m$, where m depends on the optical frequency and the ionization potential of the material. For example, at a wavelength of 810 nm, $m = 11$ for nitrogen and $\sigma_{11} = 5.08 \times 10^{-144}$ s^{-1} cm^{22} W^{-11}. We note that for intensities larger than $\sim 10^{14}$ W cm^{-2}, the dominant ionization mechanism is tunneling, where a 'quivering' electron has sufficient energy to tunnel through the coulombic barrier. MPI can create a plasma whose optical properties depend on the density of free electrons. The refractive index of the plasma is nonuniform, and for $\omega > \omega_p = (N_e e^2/\varepsilon_0 m_e)^{1/2}$, the plasma frequency is <1; N_e, e, and m_e are the electron density, charge, and mass, respectively. Hence, plasma formation leads to defocusing. Consequently, self-focusing in air (which has a positive n_2) can lead to plasma formation by MPI, and the resultant defocusing can stabilize the formation of a self-trapped filament of light propagating over a distance of meters. These filaments have been called 'gas-induced solitons'. High peak power femtosecond pulses are typically used to generate these. The pulse width is short compared to the mean

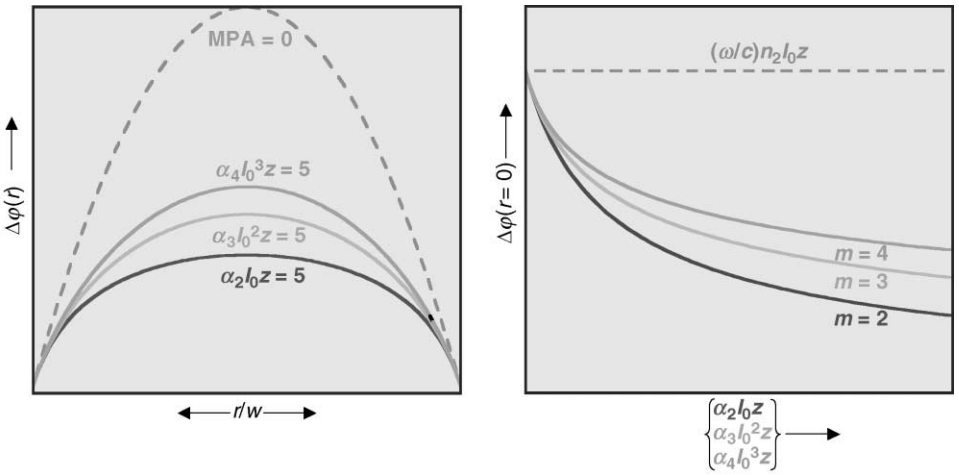


Figure 11 The effects of two-, three-, and four-photon absorption on the intensity-induced phase retardation due to a Kerr self-focusing nonlinearity. The radial dependence of $\Delta\varphi$ on MPA in the parabolic approximation is shown on the left, and the peak of $\Delta\varphi$ as a function of nonlinear absorption loss is plotted on the right.

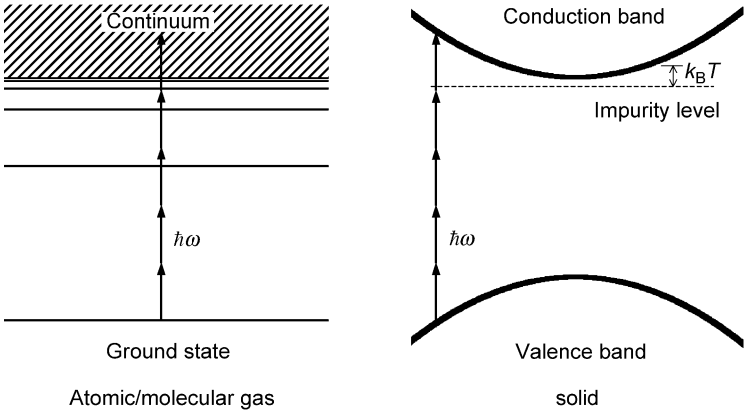


Figure 12 Multiphoton ionization in a gas (left) and a solid (right).

collision time of molecules, and the energy is low so that plasma heating is negligible.

The electron density in plasmas induced by MPI in gases is typically low enough that absorption by free electrons can be ignored. This is generally not the case in solids. MPI can provide enough 'seed' electrons to induce avalanche ionization. As illustrated in Figure 12, impurities may also supply these seed electrons by thermal ionization if they lie within an energy $\sim k_B T$ of the conduction band (k_B is the Boltzmann constant and T is the temperature). Free electrons can absorb electromagnetic energy while colliding with ions or atoms (a process called inverse bremsstrauhlung) and accelerate in a high-intensity field. They strip electrons from other atoms or molecules by impact ionization. The process multiplies free electrons and avalanche ionization ensues. A plasma forms, absorbs more energy and expands, developing a shock wave that can lead to a permanent crack in the solid. This phenomenon is called optical breakdown. Typically observed with nanosecond and picosecond pulses, permanent laser damage of the material also robs energy from the beam and generally terminates the collapse of a self-focused beam. When the electron density is sufficiently high so that $\omega_p > \omega$, the plasma absorbs and scatters light, a process that can be put to use in optical power limiters.

Many additional nonlinear phenomena play important roles when the intensity of a self-focused beam gets this large (e.g., stimulated Raman scattering, stimulated Brillouin scattering, etc.). The reader is referred to the literature on nonlinear optics for more in-depth discussions of these effects.

List of Units and Nomenclature

Critical power	[W]
Electric field	[V m^{-1}]
Four-photon absorption coefficient	[cm^5 W^{-3}]
Intensity	[W cm^{-2}]
Multiphoton (m-photon) absorption coefficient	[cm^{2m-3} W$^{-(m-1)}$]
Multiphoton (m-photon) ionization rate coefficient	[s^{-1} cm^{2m} W^{-m}]
Nonlinear refractive index coefficient	[cm^2 W^{-1}]
Pockels coefficient	[m V^{-1}]
Second-order susceptibility	[m V^{-1}]
Third-order susceptibility	[m^2 V^{-2}]
Three-photon absorption coefficient	[cm^3 W^{-2}]
Two-photon absorption coefficient	[cm W^{-1}]
Vacuum permittivity	[C V^{-1} m^{-1}]

See also

Laser-induced Damage of Optical Materials. **Materials Characterization Techniques:** $\chi^{(3)}$. **Nonlinear Optics, Applications:** Pulse Compression via Nonlinear Optics. **Nonlinear Optics, Basics:** Cascading; Kramers–Krönig Relations in Nonlinear Optics; Nomenclature and Units; Photorefraction; Ultrafast and Intense-Field Nonlinear Optics; $\chi^{(2)}$–Harmonic Generation. **Photonic Crystals**: Nonlinear Optics in Photonic Crystal Fibers. **Scattering:** Stimulated Scattering. **Solitons:** Bright Spatial Solitons.

Further Reading

Boyd RW (1992) *Nonlinear Optics*. Boston, MA: Academic Press.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. New York: Cambridge University Press.

Chiao RY, Garmire E and Townes CH (1964) Self-trapping of optical beams. *Physics Review Letters* 13: 479–482.

Crosignani B and Salamo G (2002) Photorefractive solitons. *Optics and Photonics News* 13: 38–41.
Kivshar YS and Stegeman GI (2002) Spatial optical solitons. *Optics and Photonics News* 13: 59–63.
Marburger JH (1975) Self-focusing: theory. *Progress in Quantum Electronics* 4: 35–110.
Sheik-bahae M, Said AA, Wei TH, Hagan DJ and Van Stryland EW (1989) Sensitive measurement of optical nonlinearities using a single beam. *IEEE Journal of Quantum Electronics* 26: 760–769.
Sheldon SJ, Knight LV and Thorne JM (1982) Laser-induced thermal lens effect: a new theoretical model. *Applied Optics* 21: 1663–1669.
Shen Y (1984) *The Principles of Nonlinear Optics*. New York: Wiley.
Silberberg Y (1990) Collapse of optical pulses. *Optics Letters* 15: 1282–1284.
Sutherland RL (2003) *Handbook of Nonlinear Optics*, 2nd edn. New York: Marcel Dekker.
Torner L and Sukhorukov AP (2002) Quadratic solitons. *Optics and Photonics News* 13: 42–47.
Trillo S and Torruellas WE (eds) (2001) *Spatial Solitons*. Berlin: Springer-Verlag.
Wise F and Di Trapani P (2002) The hunt for light bullets – spatiotemporal solitons. *Optics and Photonics News* 13: 28–32.

Three-Dimensional Microfabrication

S M Kuebler, University of Central Florida, Orlando, FL, USA

M Rumi, Georgia Institute of Technology, Atlanta, GA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Nonlinear optical three-dimensional microfabrication (3DM) is a photolithographic technique that enables topologically complex 3D microstructures with feature size as small as 1 μm or less to be generated in a single exposure step by nonlinear photopatterning in a material. In the broadest context, 3DM encompasses schemes based on single-beam serial patterning and parallel patterning achieved by multiple-beam interference exposure. In both cases, the photopatterning is initiated by coherent multiphoton excitation (MPE). The material may be a glass, a polymerizable resin, or even a heterogeneous composite, such as a resin containing dispersed nano-particles. The limiting resolution, or the dimensions of the smallest feature that can be generated, is determined by the optical parameters of the excitation geometry and the physical and chemical response of the material to MPE.

3DM offers great promise as a tool for generating complex microdevices, such as micro-electromechanical systems (MEMS), microfluidics, and micro-optical components. New frontiers in the technology of microdevices require that complex 3D structures can be produced in one or more materials that are optimized for a target application in terms of physical, chemical, and mechanical properties. Rapid and versatile fabrication is also essential. Extensive work has been done with conventional photolithography or electron-beam lithography, and more recently using soft-lithography approaches, such as microcontact printing. 3DM provides a new set of tools that overcomes some of the limitations inherent to other methods.

Single-Beam Three-Dimensional Microfabrication

3D Confinement

The simplest form of 3DM involves patterning a microstructure within a material using a single tightly focused laser beam that can activate the medium or one of its constituents by two-photon excitation (TPE) (Figure 1). The TPE initiates a photochemical or photophysical process (e.g., a polymerization reaction or a phase change) within the focal volume and results in an irreversible change in the material. The pattern of the target structure is impressed point by point as the focus is translated within the volume of the material. For some applications the final 3D microstructure is obtained when the unexposed material surrounding the photopatterned regions is removed. In the first report of this type of 3DM in 1990 two-photon-induced polymerization (TPIP) of an acrylate resin was used to fabricate free-standing 3D micron-scale letters on a substrate (Figure 2). The unexposed material does not have to be removed in a post-exposure process when the target structure consists of the phototransformed material supported within the matrix of the unexposed media, as in the case of 3DM of waveguide circuits embedded within a host slab.

3DM is possible because, under tight focusing, TPE and the subsequent material transformation are

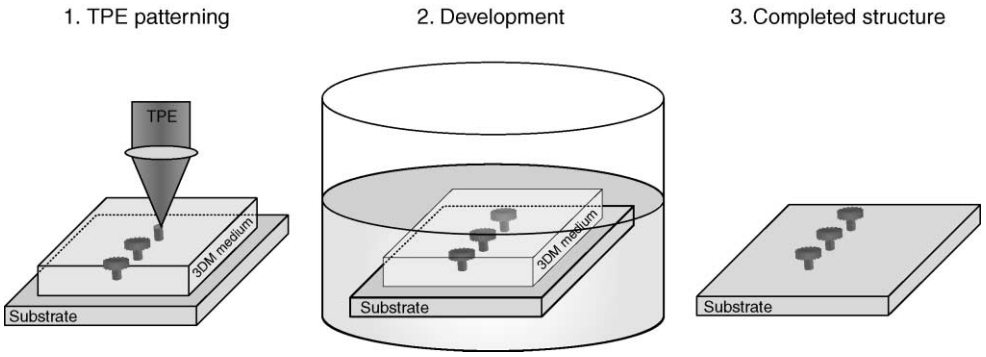


Figure 1 General scheme for single-beam 3DM based on TPE. (1. TPE patterning): The 3DM medium is supported on a substrate and translated in 3D-space relative to a tightly focused laser beam. The laser transforms the material only in the region around the focal point through a photochemical or photophysical process initiated by TPE. (2. Development): In cases where the final structure is needed free from the unexposed material, the sample is treated with a 'developer' that removes the unexposed material, for example by dissolving it in a solvent. (3. Completed structure): Following removal of the unexposed material, the final 3D structure is obtained free-standing on the substrate.

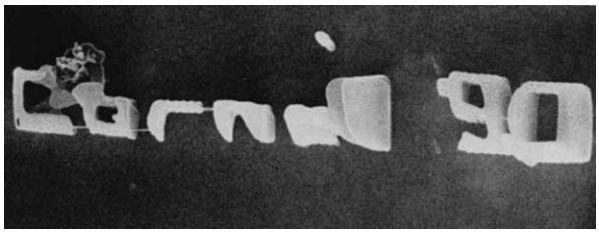


Figure 2 Scanning electron micrograph of a structure obtained by TPIP-3DM in an acrylate-ester resin. The excitation wavelength was 630 nm. The voxel dwell time was 10 ms along the letters. The thin filaments connecting some of the letters are approximately 100 nm wide. Reproduced from Strickler JH and Webb WW (1990) Two-photon excitation in laser scanning fluorscence microscopy. *Proceedings of the Society of Photo-Optical Instrumentation Engineers* 1398: 107–118. Copyright (1990), with permission of SPIE.

confined at the focus within a volume of $\sim(\lambda/n_0)^3$, where λ and n_0 are the vacuum wavelength of the exciting radiation and the refractive index of the material, respectively. Degenerate TPE is a resonant third-order nonlinear optical process in which a species is promoted to an excited electronic state by the simultaneous absorption of two photons of equal energy. Relative to conventional one-photon excitation (OPE), TPE is typically achieved using longer-wavelength radiation (often red or near infrared), for which the combined energy of two photons is sufficient to promote the species into one of its lower lying excited electronic states (Figure 3, left). The rate of TPE, that is the number of species excited per unit volume and time, $R_{2h\nu}$, is proportional to the square of the intensity, I. For a focused Gaussian beam, the on-axis intensity attains its maximum, I_0, at the focus and decreases with distance from the focal plane, z, approximately as $I \propto z^{-2}$. For a focused Gaussian beam, the on-axis intensity is given by $I = I_0/[1 + (z/z_R)^2]$, where z_R is the Rayleigh range of the focused wavefront. Thus, $I \propto z^{-2}$ in the limit of $z \gg z_R$. Consequently, the excitation rate decreases as $R_{2h\nu} \propto z^{-4}$. Similarly, the excitation rate decreases rapidly away from the propagation axis. This results in a region of high excitation, and thus material transformation, that is tightly confined both laterally and longitudinally in space about the focal point (Figure 3, right). This situation may be contrasted with that for OPE, for which the excitation rate, $R_{1h\nu}$, is linearly proportional to I, so $R_{1h\nu} \propto z^{-2}$. The weaker dependence of $R_{1h\nu}$ on z results in material being excited appreciably throughout the irradiation volume, with less longitudinal confinement at the focus than is achieved under TPE. Similar arguments can be used to describe 3DM involving higher-order excitation, for which the excitation is confined to an even smaller volume for a given λ.

The rate for molecular n-photon excitation, $R_{nh\nu}$, represents the average number of molecules that are excited per unit volume per unit time and is given by

$$R_{nh\nu} = \frac{1}{n(h\nu)^n}\sigma^{(n)}NI^n \qquad [1]$$

where n is the order of the excitation process, h is Planck's constant, ν is the frequency of the exciting radiation, $\sigma^{(n)}$ is a molecular n-photon absorption cross-section, which has units of $[\mathrm{cm}^{2n}\,\mathrm{s}^{n-1}\,\mathrm{photon}^{-(n-1)}]$, N is the number density of the photo-excitable species $[\mathrm{cm}^{-3}]$, and I is the intensity $[\mathrm{W\,cm}^{-2}]$. The factor of $1/n$ accounts for the fact that n photons must be absorbed to promote a single molecule to an excited state. Note that $\sigma^{(n)}$ is a

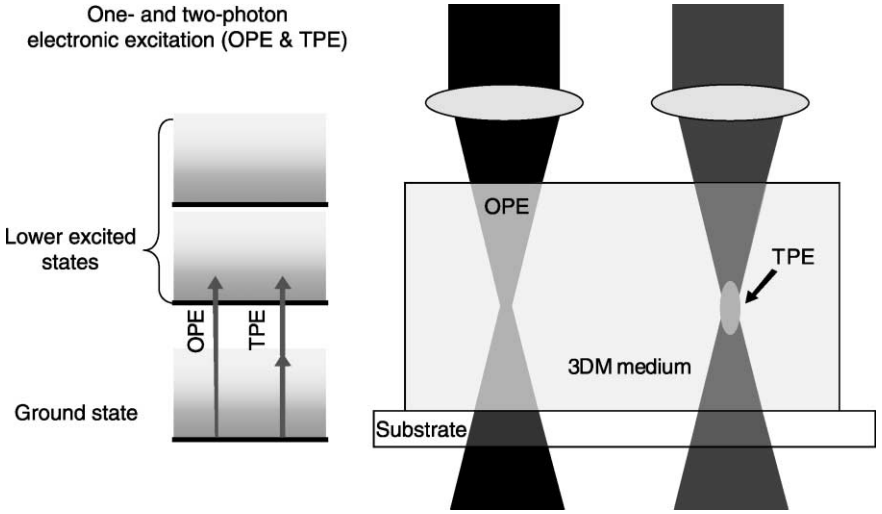


Figure 3 (Left) Jablonski representation of electronic one- and two-photon excitation (OPE and TPE, respectively). (Right) Illustration of a 3DM medium undergoing OPE and TPE using focused short- (black) and long-wavelength (dark gray) radiation, respectively. The regions that are appreciably excited are shown in light gray. Under OPE the medium is excited throughout the length of the interaction volume. Under TPE the excitation is confined in three dimensions to a small region around the focal point.

molecular parameter that is independent of the intensity. The molecular cross-sections for OPE and TPE, $\sigma^{(1)}$ and $\sigma^{(2)}$, are respectively equivalent to the parameters σ and δ that are commonly encountered in the literature.

For MPE to occur, the intensity must be sufficiently large that there is a high probability of more than one photon arriving simultaneously at the species to be excited. Typical magnitudes of the OPE and TPE cross-sections are $\sigma^{(1)} \approx 10^{-17}\,\mathrm{cm}^2$ and $\sigma^{(2)} \leq 10^{-47}\,\mathrm{cm}^4\,\mathrm{s\,photon}^{-1}$, and under conventional excitation conditions, $R_{2h\nu} \ll R_{1h\nu}$. To achieve $R2h\nu \approx R1h\nu$, $I \approx 2h\nu\sigma^{(1)}/\sigma^{(2)}$ is required. For the cross-sections given above, this condition corresponds to $I \approx 500\,\mathrm{GW\,cm}^{-2}$ at $\lambda = 800$ nm. Such peak powers can be achieved by focusing ultra-short laser pulses, so mode-locked femtosecond- or picosecond-pulse lasers are generally used for 3DM. Tight focusing in 3DM is desirable then, not only to provide a means for 3D confinement, but also to increase the intensity to a level that $R_{nh\nu}$ becomes large enough to produce a significant material change.

Opto-Mechanical Implementation

One method for implementing single-beam 3DM is illustrated in Figure 4. An excitation laser beam is directed onto a high numerical aperture (NA) objective. A beamsplitter, detector, attenuator, and shutter are placed upstream of the objective to measure and control the average power at the sample, $\langle P \rangle$. A substrate is coated with the photo-active medium and attached to a 3-axis nano/micro-positioner under the objective. The interior of the photo-active material is patterned during the exposure by translating the sample relative to the focus of the laser beam and shuttering the beam as needed. The process is automated as a computer controls the sample translation coordinates and the exposure conditions for each volume element (voxel). This configuration is well suited for patterning solid or semi-solid photo-media. An index-matching fluid must be used with high-NA objectives to achieve the smallest focal spot size. If the photo-medium is a liquid, a coverslip or other thin transparent barrier must be placed between the index matching fluid and the photomedium so that the latter does not become contaminated.

Alternatively, an inverted configuration may be used. Here, the laser is focused through the substrate into the medium. The objective and the index matching fluid are kept out of contact with the medium by the substrate. In this case, the maximum depth that can be patterned is limited by at least two conditions: (i) The working distance of high-NA objectives is small, often no more than 200 μm, so the substrate must be kept thin (~100 μm); and (ii) unless the viscosity of the medium is very high, patterning must begin at the medium/substrate interface and proceed layer by layer away from it, so that the structure remains anchored to the substrate. This could be disadvantageous, as the change in refractive index in the previously patterned layers could alter the intensity distribution

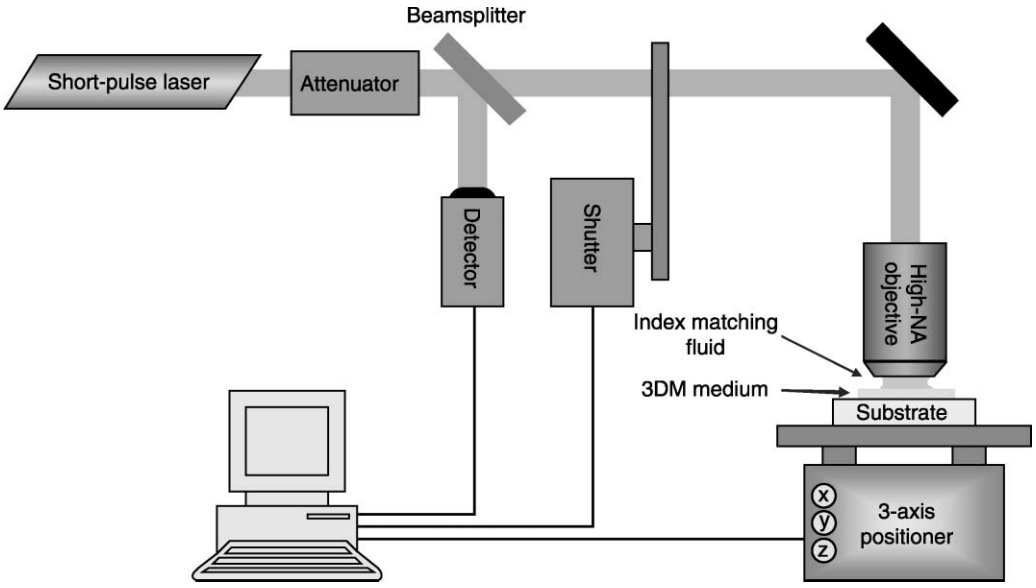


Figure 4 Optomechanical setup for single-beam 3DM.

at the focus. In all configurations, it is best to minimize the number of optical interfaces to maintain the quality of the focus.

A scanning confocal microscope system provides a convenient platform for single-beam 3DM. The nano/micro-positioner can be mounted directly on the sample stage of the microscope, and the optical components of the confocal system can be used directly for 3DM. The confocal system also enables the medium and the structure to be imaged *in situ*, using confocal reflectance or fluorescence imaging. If the material is fluorescent under MPE, a 3D image can be obtained by scanning multiphoton fluorescence microscopy. These imaging modes facilitate 3DM by providing a means for locating the focus at an absolute position within the photo-active medium.

3DM can also be realized by moving the focused beam relative to a fixed sample. One convenient implementation utilizes galvanometer-driven mirrors to scan the beam in planes parallel to the substrate (*xy*-plane). Motion in the *z*-direction is accomplished by raising and lowering the objective relative to the substrate. A galvo-based system typically enables faster fabrication, but may not afford motion as precise as that obtained using a multi-axis nano-positioner.

Excitation Sources

The most commonly used and arguably the most convenient and reliable laser for 3DM is the continuous-wave (CW) mode-locked titanium sapphire (Ti:S) laser. Ti:S lasers routinely produce sub-100 fs pulses at high repetition rates (~80 MHz) with single-pulse energies, E_p, of ~1 nJ. The Ti:S laser emission spans the range of 700–1050 nm, making it particularly well suited for two- and three-photon excitation. $\langle P \rangle$ can be as high as several watts at the peak of the gain curve ($\lambda \approx 800$ nm) with high pulse-to-pulse and time-averaged stability (RMS ≈ 1%). Such systems may also produce a nearly Gaussian TEM_{00} mode, which is helpful in achieving the tightest focus in the material. Note that focusing a 1 nJ, 100 fs pulse to a 1 μm-radius spot generates a focused average peak intensity of 320 GW cm^{-2}, as is needed to achieve photo-excitation rates of $R_{2h\nu} \approx R_{1h\nu}$.

For higher peak powers, amplified femtosecond (AFS) lasers can be used for 3DM. These systems can produce ~100 fs pulses with $E_p \approx 1$ mJ or higher at repetition rates of 1–100 kHz. Given that a continuous microfabricated structure should be generated from partly overlapping voxels, a low repetition rate reduces the speed with which the laser beam may be scanned within the material. An AFS laser can be used to pump an optical parametric generator to extend the output wavelength range.

Recently, some researchers have turned their attention toward identifying compact solid-state turn-key lasers for 3DM that could be more convenient and cheaper alternatives to the complex systems described above. A commercial Nd:YAG microlaser ($\lambda = 1064$ nm, $E_p = 5.4$ μJ, $\tau_p = 640$ ps, 13.6 kHz) and a mode-locked erbium-doped fiber laser ($\lambda = 780$ nm, $E_p \approx 0.5$ nJ, $\tau_p = 100$ fs, 50 MHz)

have both been shown to be satisfactory for 3DM in certain acrylate-based media.

Material Systems

Structural Materials

The most widely used materials for 3DM are based on acrylates that are patterned by multiphoton induced polymerization. The process begins when an initiator chromophore **C** is excited by n-photon absorption to $\mathbf{C}^*$, with an efficiency specified by the n-photon cross-section $\sigma^{(n)}$. $\mathbf{C}^*$ then generates an initiating radical **I·**. Acrylate monomers (**M**) can react with **I·** and undergo a free-radical chain reaction that leads to high molecular weight polymer. With sufficient exposure, the starting material is converted to less soluble high-molecular weight polymer chains or densely cross-linked polymer networks. The photo-patterned structure can then be 'developed' by immersing the sample into a solvent that removes the unexposed material, leaving behind a free-standing structure that is a replica of the photopattern.

Photo-polymerization starts when the MPE exposure generates enough **I·** to react with and consume all of the inhibitor in the local volume at point $\vec{r}$ within a homogeneous medium. In this discussion we assume that the local concentration of inhibitors is not changed by diffusion on the exposure timescale. The total number of **I·** generated during an exposure at $\vec{r}$ per unit volume, $N_{\mathrm{I}\cdot}(\vec{r},t)$, is obtained from the time-integrated MPE rate:

$$N_{\mathrm{I}\cdot}(\vec{r},t) = \frac{\phi_{\mathrm{c}}^{(n)}\sigma^{(n)}N_{\mathrm{C}}}{n(h\nu)^n}\int_{-\infty}^{t} I(\vec{r},t')^n \mathrm{d}t' \qquad [2]$$

Here, N_{C} is the number density of **C**, and $\phi_{\mathrm{c}}^{(n)}$ is the n-photon chemical quantum yield, defined as the number of **I·** generated divided by the number of $\mathbf{C}^*$ present after excitation. If we identify the initial number density of the inhibitors as N_{Inh}, then polymerization occurs when $N_{\mathrm{I}\cdot}(\vec{r},t) > N_{\mathrm{Inh}}$. In multiphoton-induced polymerization, $N_{\mathrm{I}\cdot}(\vec{r},t)$ is a nonlinear function of I, which generally involves a rapidly varying time envelope, as from a mode-locked ultrashort-pulse laser.

Several different commercial acrylate systems have been used successfully for 3DM, such as Nopcocure 800 (San Nopco), NOA 72 (urethane acrylate oligomer, Norland Products), and SCR 500 (blend of urethane acrylate monomers and oligomers, Japan Synthetic Rubber Co.). Researchers have also successfully developed custom formulations using blends of commercially available monomers, oligomers, and polymer additives to create resins with tailored physical, chemical, and optical properties. This aspect of acrylates, and more generally polymer-based systems, makes them one of the most attractive classes of materials for 3DM. Conventional one-photon UV- or blue-sensitive radical generators (e.g., benzil, benzoin methyl ether, 4,4′-bis(N,N′-dimethylamino)benzophenone, to name just a few) have most often been used, and these were activated by promoting the molecule into one of its low-lying electronic states by MPE over the range of 730–800 nm.

3DM based on epoxide polymerization has also been demonstrated using commercial resins including cross-linkable small molecular weight monomers (e.g., 4-vinyl-1-cyclohexene diepoxide) and epoxide oligomers (e.g., SU-8, first developed by IBM). Epoxide polymerization is most commonly initiated by molecules that, after excitation, generate a Brønsted acid, H^+. The propagating species is a carbocation, and bases present in the medium act as inhibitors. Relative to acrylates, epoxides shrink substantially less upon polymerization, leading to less distortion of the microstructure after development.

Acrylate chemistry has also been exploited for 3DM of functional composite materials. Acrylamide resin mixtures were used to fabricate bi-layer micro-cantilevers that deflect from the surface when illuminated with UV light. This type of microstructure could form the basis of a photo-activatable MEMS. 3D gratings were fabricated from an acrylate pre-polymer containing a dendrimer functionalized with cross-linkable acrylic L-phenylalanine groups. The dendrimer acts as a host encapsulation site that enables a laser dye to be loaded in the material at high weight-percent (see below). A variety of complex micro-structures have also been fabricated using 'Ormocer-I', an ORganically MOdified CERamic. Ormocer-I is an inorganic-organic hybrid polymer consisting of a Si–O–Si (siloxane) backbone that is functionalized with organic moieties, including cross-linkable methacryloxypropyl groups. Ormocers are extremely promising for 3DM because of the thermal stability and chemical inertness of the backbone and their attractive physical and optical properties.

3DM could be a powerful tool for fabricating biologically active structures and devices. In this context, hydrogels and Ormocers should be useful material systems as some formulations are known to be bio-compatible. Microstructures can also be fabricated from proteinaceous material. 3D microstructures were produced by patterned MPE of

solutions containing the dye rose bengal and the proteins bovine serum albumin or fibrinogen.

The material systems discussed thus far are 'negative-tone' media. This means that only the exposed regions of the material remain after the post-exposure development, so the final structure is a replica of the exposure pattern. In contrast, 'positive-tone' media are solid-state systems for which the exposed regions become soluble in a developer. The final structure is then the inverse of

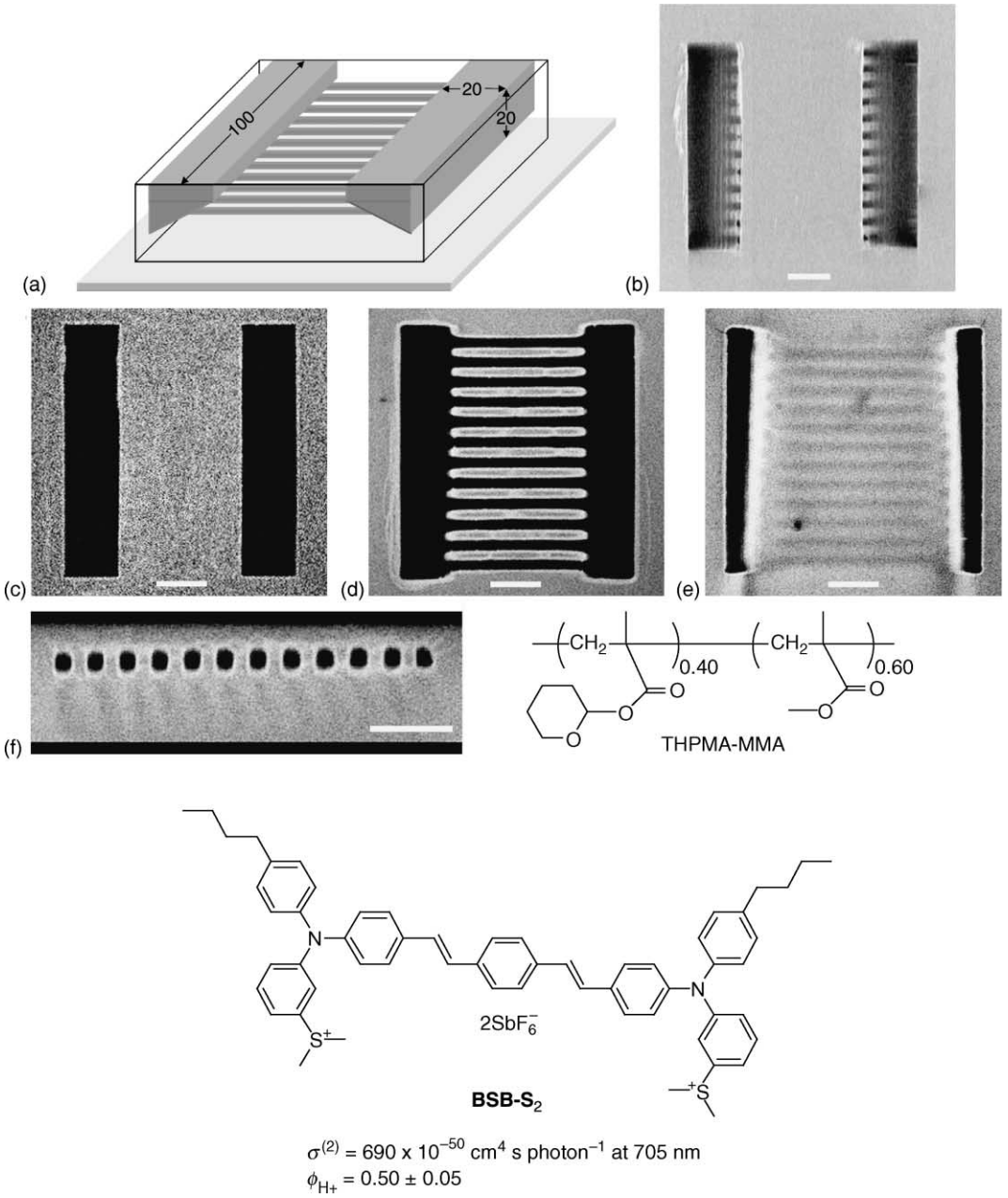


Figure 5 A 3D micro-channel structure fabricated by TPE-3DM using the positive-tone chemically amplified resist THPMA-MMA containing the two-photon PAG **BSB-S_2**. A 50 μm-thick film of the resist was exposed in the pattern of the target structure at 745 nm with tightly focused 80 fs pulses at $\langle P \rangle = 40$ μW and a linear scan speed of 50 μm s^{-1}. (a) Target structure consisting of two rectangular cavities (width: 100 μm, length: 20 μm, depth: 20 μm) with a sloped side-wall, that are connected by 12 channels (length: 50 μm; 4 μm by 4 μm cross-section) lying 10 μm below the surface and spaced apart by 8 μm (center-to-center). (b) Scanning electron micrograph of the final structure, viewed normal to the substrate. (c to e) Two-photon-fluorescence images of the final structure (viewed normal to the substrate): (c) at the surface of the film; (d) 10 μm below the surface; (e) 19 μm below the surface. (f) Two-photon-fluorescence cross-sectional image of the buried channels. The scale bar in b to f corresponds to 20 μm. Reproduced from Zhou W, *et al.* (2001) *Science* 296: 1106–1109. Copyright (2001), with permission of The American Association for the Advancement of Science.

the exposure pattern. Positive-tone 3DM was first demonstrated as early as 1992 with the generation of high aspect ratio trenches in thin layers of commercial photoresists. A positive-tone material system was designed for 3DM (Figure 5), is based on a chemically amplified resist and a high-sensitivity two-photon-activatable photo-acid generator (PAG) (discussed below). Using this medium, it was possible to fabricate subsurface microchannel structures and micrograting, using moderate laser scan speeds (50 $\mu m\ s^{-1}$) and exceptionally low $\langle P \rangle$ (40 μW).

Composites have been formulated that enable 3D-patterning of nanoscale metal features within a host matrix. A two-step process has been described that involves MPE of a Ag^+-containing sol–gel followed by thermal development of the latent image with a $AgClO_4$ solution. Using this system, 3D spirals of silver nano-particles embedded within the sol–gel matrix were generated. A $HAuCl_4$-doped poly(vinyl alcohol) composite was used to 3D-pattern gold nano-particles. It was shown that after TPE a particular sensitizer dye hosted in a poly(vinylcarbazole) matrix containing $AgBF_4$ and ligand-stabilized Ag nanoparticles, electron transfer from the dye to Ag^+ leads to nucleation of Ag^0 at the nanoparticles and the formation of conducting bulk metal features (Figure 6). Material systems like these should find wide application for generating micro- and nano-scale structures, having unusual and useful optical and electronic properties.

3DM in glasses is of great interest for photonic applications. Several groups have demonstrated that tightly focused infrared or visible amplified ultrashort laser pulses can be used to 3D-pattern pure and doped silica and other glasses. The patterned voxels can be generated as a result of photo-induced changes in the refractive index or extreme dielectric breakdown, resulting in pits within the material. The mechanisms for these two cases are not fully understood, but both are thought to be activated through an initial MPE event. The leading edge of the laser pulse generates a free-carrier plasma through MPE. The remainder of the pulse is increasingly absorbed by the plasma leading to avalanche ionization. The optical energy deposited into the local lattice results in cavitation, creating a pit at the focal point. Based on the electronic bandgap of silica, the process likely involves three- and six-photon absorption when 400- and 800-nm light is used, respectively. This approach has been used for the 3D-patterning fused silica and germanium doped glasses.

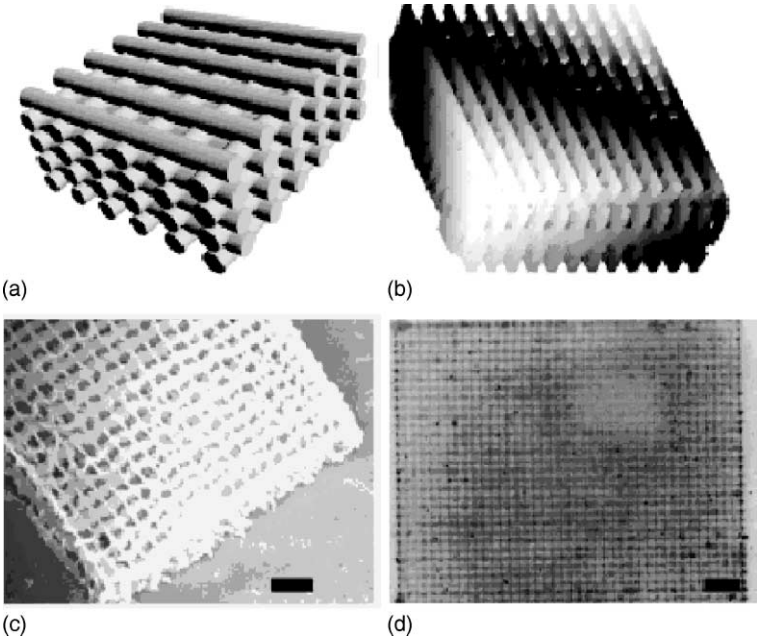


Figure 6 Metallic structures fabricated by TPE-3DM using a silver nanoparticle/polymer composite. (a) Model of the 'stack-of-logs' target structure. (b) 3D image constructed from a series of two-photon fluorescence microscopy images of the silver structure immediately following 3DM. (c) Scanning electron micrograph of the free-standing silver structure following removal of the surrounding unexposed composite. The scale bar corresponds to 10 μm. (d) Transmission optical microscopy image of the silver structure immediately following 3DM. The scale bar corresponds to 25 μm. Reproduced from Stellacci F, *et al.* (2001) *Advanced Materials* 14: 194–198. Copyright (2001), with permission of Wiley.

Multiphoton Initiators

For 3DM to develop into a widely used fabrication tool, material systems and patterning methods must be developed that enable rapid microfabrication at low cost. 'Soft materials', such as the polymer systems described above, are promising in this respect, as both the initiating species and the host resin can be engineered to optimize the multiphoton response and the physical and chemical properties of the patterned medium. It may also prove possible to create doped glasses that have physical properties close to that of neat silica and yet offer higher multiphoton sensitivity.

To develop this concept, let us consider 3DM in acrylates further. A high-sensitivity acrylate medium is one for which: (i) **C** strongly absorbs n-photons (large $\sigma^{(n)}$); (ii) $\mathbf{C}^*$ efficiently generates the initiating species $\mathbf{I}\cdot$ (large $\phi_c^{(n)}$); and (iii) $\mathbf{I}\cdot$ is an efficient initiator (high initiation rate). As criteria (i) and (ii) are predominantly determined by the molecular properties of **C**, it should be possible to engineer soft materials for 3DM by optimizing multiphoton initiators and incorporating these into pre-polymer resins that once patterned have the correct properties for the target application. The sensitivity of materials can be compared on the basis of the minimum average laser power, $\langle P\rangle_{\min}$, that is required to pattern a material using a fixed scan speed (or exposure time). Alternatively, we can use the product $\sigma(n)\phi c(n)$ as a multiphoton initiator figure-of-merit (FOM) that accounts for the strength of the multiphoton absorption and the efficiency with which $\mathbf{I}\cdot$ is generated.

The chromophores that have been used most frequently for polymerization-based 3DM do efficiently generate highly active initiating species once excited; however, they were developed for one-photon response to UV and blue radiation and generally have low multiphoton FOMs. The use of xanthene dyes was explored as initiators for free-radical polymerization of acrylamides and it was found that the multiphoton sensitivity was low. As a result, 3DM using conventional initiators requires high laser power (typically $\langle P\rangle_{\min} \approx 10-100$ mW) and slow laser scan speeds ($\sim 50\ \mu m\ s^{-1}$). Under tight-focusing conditions, the resulting peak powers are often near the damage threshold of the materials, rendering the process unreliable from a manufacturing stand-point.

A series of **D**–$\boldsymbol{\pi}$–**D** molecules was examined (Figure 7), where **D** is an electron-donating group and $\boldsymbol{\pi}$ is a conjugated π-electron bridge. This class of **D**–$\boldsymbol{\pi}$–**D** chromophores was found to have large TPE cross-sections ($\sigma^{(2)} > 200 \times 10^{-50}$ cm^4 s photon^{-1}) and to initiate the polymerization of acrylates following TPE in the visible and near infrared, with $\langle P\rangle_{\min}$ a factor of 30–50 times lower than for conventional one-photon initiators. The sensitivity achieved with **D**–$\boldsymbol{\pi}$–**D** initiators is so enhanced that TPIP-3DM was performed at laser scan-speeds as high as 9 cm s^{-1}, using $\langle P\rangle = 3$ mW, enabling a volume of 180 μm × 180 μm × 20 μm to be polymerized in 200 s.

D–$\boldsymbol{\pi}$–**A** molecules have also been examined for TPIP (Figure 7), where **A** is an electron accepting functionality. Such molecules can have large TPE cross-sections, and some have been shown to activate TPIP efficiently in acrylates at $\lambda \approx 800$ nm. A derivative of Michler's ketone was synthesized with an extended π-conjugated system (Figure 7). This molecule exhibited strong two-photon absorption over the broad range of 800–1100 nm and was also shown to be active toward acrylate polymerization at 1064 nm. A strategy was implemented for enhancing two-photon sensitivity based on up-converted fluorescence and used for 3DM of waveguide structures. The strongly two-photon-absorbing dye AF-183 was incorporated into a commercial acrylate photoresist containing a radical-generator that is activated by OPE at 315–450 nm. After AF-183 undergoes TPE at 800 nm, it fluoresces efficiently at ~465 nm. The up-converted emission then excites the one-photon initiator and radicals are generated.

The strategies discussed above are oriented toward enhancing multiphoton sensitivity by increasing $\sigma^{(2)}$ of the initiating chromophore. Little attention has been paid to the need of ensuring that the MPE couples to a mechanism for efficiently generating an active initiating species. This problem was addressed with the design of a two-photon PAG by covalently linking a **D**–$\boldsymbol{\pi}$–**D** chromophore and sulfonium moieties, which can generate H^+ when reduced by photo-induced electron transfer (**BSB-S$_2$** in Figure 5). **BSB-S$_2$** features strong two-photon absorption and its quantum yield for the generation of H^+ per photo-excited molecule (ϕ_{H^+}) exceeds that of some commercial one-photon PAGs. Based on $\langle P\rangle_{\min}$, **BSB-S$_2$** was found to be 10 to 100 times more sensitive toward multiphoton epoxide polymerization than one-photon UV/blue-sensitive PAGs.

It should be understood that this discussion of multiphoton initiators is not comprehensive of all work that has been done in the field. Rather, it is intended as a highlight of the many outstanding examples of progress in this area. The use of inexpensive low-power lasers for 3DM should become more practical as high-sensitivity

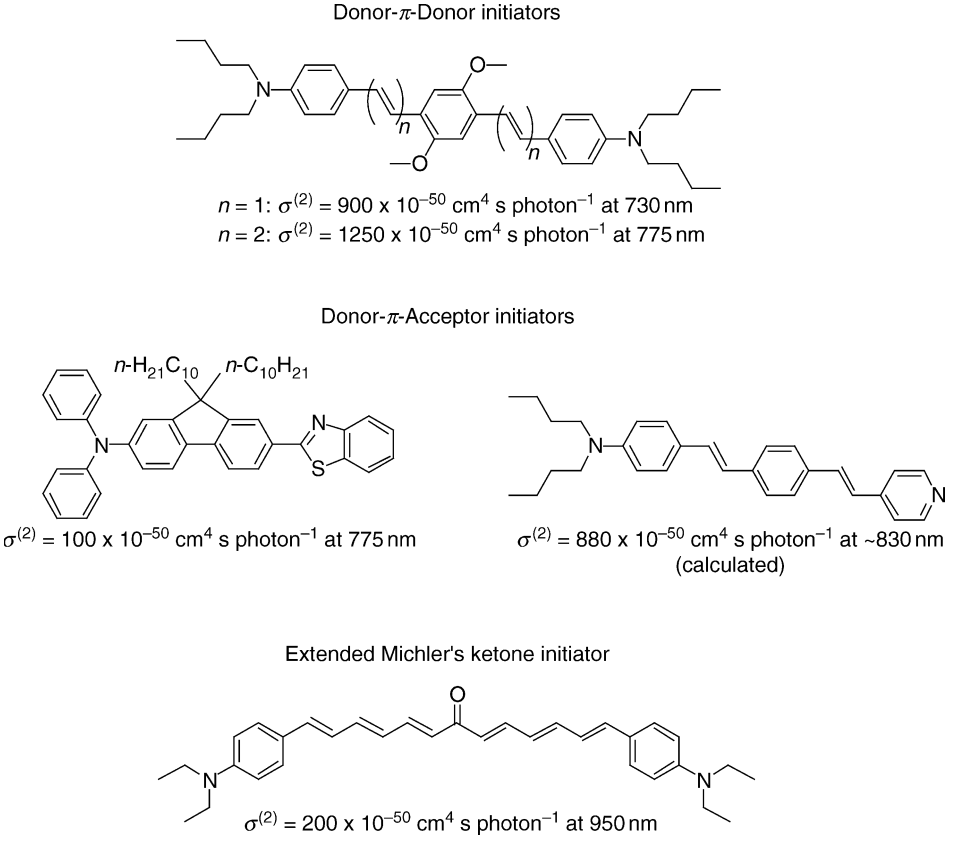


Figure 7 Examples of high-sensitivity initiators for 3DM based on TPIP of acrylates.

3DM media are developed. This should, in turn, facilitate the widespread use of 3DM as a microfabrication tool.

Resolution

The resolution of single-beam 3DM is determined by the size of the smallest voxel that can be reliably produced. Both the optical parameters and the material response under MPE affect the size of the voxel. Given the wide range of physical and chemical mechanisms that can be involved, no single set of guidelines can be outlined regarding the effect of the material response on voxel size. Additionally, this aspect of 3DM remains relatively unexplored. However, as 3DM based on TPIP of acrylates has been examined in some detail, an overview of this material response provides some valuable insight into the overall problem of resolution. The following simplified discussion is based on the local response of the material to photo-excitation.

TPIP initiated in the vicinity of the focus could, in principle, propagate throughout the entire volume of the medium, were it not for termination by inhibitors. Let us consider further the expression for $N_{\mathrm{I}\cdot}(\vec{r},t)$, expanding the intensity as:

$$I(z,r,t') = i(z,r)f(t') \qquad [3]$$

Here, $f(t')$ is a time-dependent amplitude function, $i(z,r)$ is the peak intensity, and the position vector $\vec{r}$ is expressed in cylindrical coordinates (z,r) with the origin at the focus. As before, z denotes the direction in which the beam propagates (longitudinal). The radial coordinate orthogonal to the z-axis (lateral) is denoted by r. The beam is taken to have cylindrical symmetry, so I does not depend on the angular coordinate. For a fixed exposure time $t \to \tau_{\mathrm{ex}}$, we can identify a polymerization-threshold peak intensity, I_{th}, for which $i(z,r) > I_{\mathrm{th}}$ results in $N_{\mathrm{I}\cdot}(z,r,\tau_{\mathrm{ex}}) > N_{\mathrm{Inh}}$ and polymerization occurs. The polymerized voxel consists then of all points within the volume centered at the focal point and bounded by the I_{th} isophote.

The shape of the I_{th} isophote depends upon the optical parameters of the system. The tightest focusing is achieved when the highest NA objective

is used to focus parallel incident rays that overfill the entrance aperture. This situation generates a diffraction limited intensity pattern about the focal point. Within the focal plane the lateral intensity distribution follows the Airy pattern:

$$i(0,r) = [2J_1(\rho)/\rho]^2 I_0 \qquad [4]$$

where J_1 is the first-order Bessel function:

$$\rho = 2\pi r \mathrm{NA}/\lambda \qquad [5]$$

and I_0 is the peak intensity at the origin. Near the focus, the longitudinal intensity distribution is given by

$$i(z,0) = [\sin(\zeta/4)/(\zeta/4)]^2 I_0 \qquad [6]$$

where

$$\zeta = 2\pi \mathrm{NA}^2 z/(n_0\lambda) \qquad [7]$$

Let us define the longitudinal and lateral widths of the focal spot, Z and R respectively, to be twice the distance along the corresponding z- and r-axes over which the peak intensity of the central bright fringe decreases to $I_0/2$. If $\langle P \rangle$ is adjusted so that $I_0/2 = I_{\mathrm{th}}$, then to first order we can expect the polymerization voxel to be a solid of rotation centered about the focus having a longitudinal and lateral width of Z and R, respectively. Numerical evaluation shows that the $I_0/2$-points occur at $\rho = \pm 1.6154$ and $\zeta = \pm 5.5661$. For $\lambda = 800$ nm, NA $= 1.4$, and $n_0 \approx 1.5$, we find $R = 0.29$ μm and $Z = 1.1$ μm.

Several groups have demonstrated that voxels generated under conditions similar to those discussed above do indeed have dimensions comparable to those calculated, based on diffraction limited propagation. As shown in Figure 8, voxels with lateral and longitudinal widths below the diffraction limit can be fabricated by steadily reducing τ_{ex} (or I_0) so that an increasingly smaller portion of the central bright fringe exceeds I_{th}. In this way, voxels having $R \approx 100$ nm and $Z < 800$ nm have been generated, which convincingly demonstrates the potential of this technique as a nano-fabrication tool.

Post-exposure material processing can also have a profound impact on the resolution. In the case of TPIP-3DM, material near the boundary of the I_{th} isophote may not have sufficient molecular weight or cross-linking density to resist removal in a development step. In this case, the final voxel will be smaller than the I_{th} isophote defined strictly in terms of a polymerization threshold. Additionally, a solvent-based developer can swell or otherwise distort the voxels. Overall, three important factors determine the final size of the voxel: (i) the exposure and focusing conditions; (ii) the response of the material to photo-excitation; and (iii) the behavior of the photoconverted material during the development process. It should be emphasized that 3D confinement is due to the combined effects of nonlinear photo-excitation and the nonlinearity inherent in the photoconversion and post-exposure development of the material. More advanced treatments have also attempted to address the effects of radical and inhibitor diffusion on the size of the voxel.

Previously, we assumed that the voxel is cylindrically symmetric. The intriguing result of some recent work has shown that the widths of the voxel in the transverse directions can actually differ by as much as 10% when linearly polarized radiation is used because the symmetry of the electric field vectors in the x- and y-directions is broken in high-NA focusing geometries. Cylindrically symmetric voxels can be obtained by using depolarized or circularly polarized radiation. Clearly, study of the combined effects of optical parameters and material response remains an important area for further research in this rapidly advancing field. Although this discussion has centered on TPIP, similar arguments can be developed for 3DM based on other photomaterials and higher-order processes.

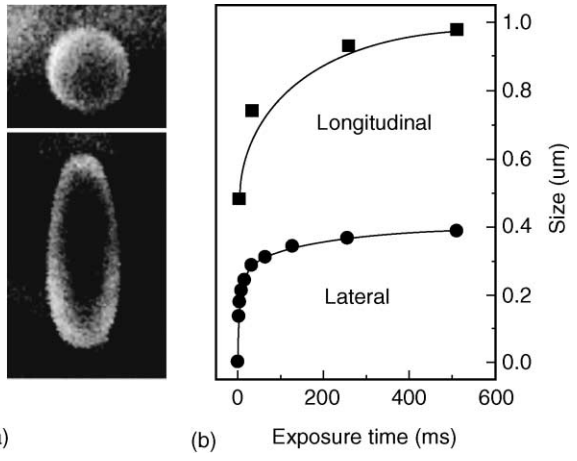


Figure 8 Single voxels produced by TPIP-3DM in an acrylate medium. (a) Scanning electron microscope (SEM) images of a single voxel showing its longitudinal (top) and lateral (bottom) width. (b) Dependence of the longitudinal and lateral widths of voxels on exposure time, as measured from SEM images following fabrication. Reproduced from Kawata S and Sun H-B (2003) Two-photon photopolymerization as a tool for making micro-devices. *Applied Surface Science* 208–209: 153–158. Copyright (2003), with permission of The American Institute of Physics.

Multibeam-Interference Three-Dimensional Microfabrication

Multibeam-interference (MBI) 3DM is an alternative approach to 3DM that enables large volumes of material to be simultaneously patterned into a 3D periodic structure. In MBI-3DM, multiple coherent laser beams interfere within a photo-active medium. A periodic structure results because the material undergoes photo-induced changes that occur maximally in the regions of peak intensity across the periodic interference pattern. Depending on the number of beams utilized and their reciprocal orientation, periodic structures in one, two, or three dimensions can be generated, and the periodicity is of the order of λ. This technique is often referred to as holographic lithography or recording.

MBI-3DM has most often been achieved using one-photon-induced polymerization; however, higher-order excitation may be used as well. In this discussion we will focus on multiphoton MBI-3DM, with the understanding that most of the concepts apply equally well to the case of linear photo-excitation. When the power of the incident beams is adjusted so that the intensity at the peaks of the interference pattern just exceeds I_{th}, polymerization only occurs in the local vicinity of these points, whereas the 3DM medium is not appreciably polymerized in regions where the intensity is below I_{th} and may be removed by dissolution in a solvent. The result is a periodic porous structure with highly interconnected parts, that extends over the whole volume where the beams overlapped.

Let us assume that m-beams are directed onto a photosensitive material. The intensity of light at a given point $\vec{r}$ in the region where the beams overlap is proportional to:

$$I(\vec{r}) \propto \vec{E}\cdot\vec{E}^{*} = \left(\sum_{j=1}^{m} \vec{E}_{0j}\, e^{i(\vec{K}_j\cdot\vec{r}-\omega t)}\right)\cdot\left(\sum_{l=1}^{m} \vec{E}^{*}_{0l}\, e^{-i(\vec{K}_l\cdot\vec{r}-\omega t)}\right)$$
$$\propto \sum_{j,l=1}^{m} \vec{E}_{0j}\cdot\vec{E}^{*}_{0l}\, e^{i(\vec{K}_j-\vec{K}_l)\cdot\vec{r}}$$
$$\propto \sum_{j=1}^{m} \vec{E}_{0j}\cdot\vec{E}^{*}_{0j} + \sum_{j\neq l}^{m} \vec{E}_{0j}\cdot\vec{E}^{*}_{0l}\, e^{i(\vec{K}_j-\vec{K}_l)\cdot\vec{r}} \qquad [8]$$

where $\vec{E}_{0j}$ and $\vec{K}_j$ are the electric field and wavevector of beam j, and ω is its angular frequency ($\omega = 2\pi\nu$). Equation [8] indicates that the intensity distribution depends on the difference between pairs of wavevectors of the interfering beams, $\vec{K}_j - \vec{K}_l$. This distribution is periodic in space and can be described as a crystal lattice with reciprocal lattice vectors equal to $\vec{K}_j - \vec{K}_l$ and with a unit cell structure determined by the pre-factors $\vec{E}_{0j}\cdot\vec{E}^{*}_{0l}$. By choosing the intensity of the incident beams and the exposure time appropriately, it is thus possible to generate a pattern in the medium that has the same periodicity and structure as the I_{th} isophote of the interference pattern.

The simplest implementation of MBI-3DM involves interfering two linearly polarized beams at angles $\pm\theta/2$ with respect to the normal to the film being patterned (θ is the angle between the two beams in vacuum). From eqn [8], it can be shown that the intensity is maximum if $(\vec{K}_1 - \vec{K}_2)\cdot\vec{r} = 2q\pi$, where q is an integer, or $K[\sin(\theta/2)]r' = q\pi$, where K is the magnitude of both $\vec{K}_1$ and $\vec{K}_2$, and r' is the component of $\vec{r}$ along $(\vec{K}_1 - \vec{K}_2)$. Thus, the two beams generate an interference pattern with periodicity (in one dimension) given by $\Lambda = \lambda/[2\sin(\theta/2)]$, if λ is the

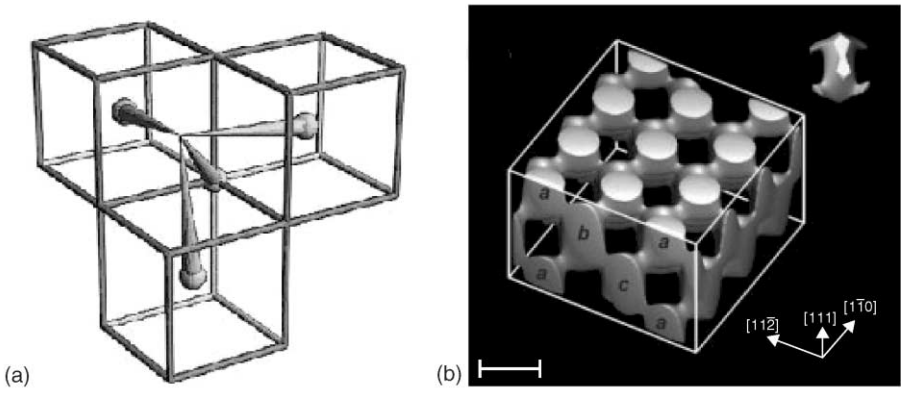


Figure 9 (a) Relative orientation of the beams in a four-beam interference scheme to obtain a face-centered cubic lattice. The beams have wave vectors along the directions [−3/2, −3/2, −3/2], [−5/2, −1/2, −1/2], [−1/2, −5/2, −1/2], and [−1/2, −1/2, −5/2]. (b) One of the isophotes generated by the interference of the four beams in (a). The inset represents the shape of the unit cell. Reproduced with permission from *Nature*, Campbell M, *et al.* (2000) *Nature* 404: 53–56. Copyright (2000), Macmillan Magazines Limited.

vacuum wavelength of the exciting beam ($\lambda = 2\pi/K$). When a film of a polymerizable material is placed at the intersection of the beams, a series of ridges is generated in the resist, spaced by Λ and with a width dependent on the beam intensity. For example, if a laser beam at 800 nm is split in two components forming an angle of 30°, $\Lambda = 1.5$ μm. The periodicity becomes smaller as the angle between the beams is increased.

By interfering three beams it is possible to obtain a two-dimensional periodic structure. For example, the interference of three beams forming an angle θ with respect to the medium normal and symmetrically arranged around it was used to generate an hexagonal lattice of rods that extends parallel to the normal. At least four beams must interfere to obtain a microstructure with 3D periodicity. The beam geometry shown in Figure 9a was used to produce a microstructure with the symmetry of a face-centered cubic lattice (Figure 9b). Beam geometries for generating interference patterns having other symmetries, such as body-centered cubic and body-centered tetragonal, have also been described.

An advantage of MBI-3DM is that it enables rapid patterning of large volumes during a single exposure. In fact, the polymerization takes place, in principle, throughout the region in which the beams overlap. When the process is activated by OPE, the beams can be attenuated significantly by absorption in the material. This can limit the thickness of the fabricated structure or result in spatial variation in the filling ratio. The use of TPIP can overcome this limitation and allow for thicker media to be patterned uniformly. Structures with the same symmetry are generated if either one- or two-photon processes are exploited to induce the material transformation. However, the structures will be different for at least two reasons: (i) if the same initiator is used in both cases, the wavelength used for TPE is approximately twice that for OPE, and

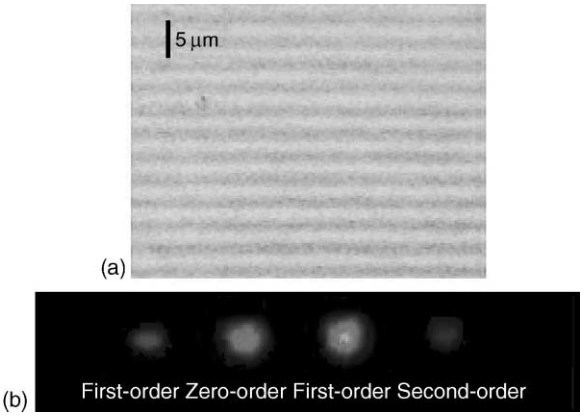


Figure 10 (a) Optical image of a diffraction grating obtained by TPIP using a two-beam interference scheme ($\lambda = 810$ nm). The spacing period is 3.4 μm and film thickness is 40 μm. The grating is made of an acrylate-based resin. (b) Diffraction pattern of the grating in (a) generated using a Helium–Neon laser (632.8 nm). The maximum diffraction efficiency was measured to be 57%. Reproduced from Guo H, *et al.* (2003) *Chemical Physics Letters* 374: 381–384. Copyright (2003), with permission from Elsevier.

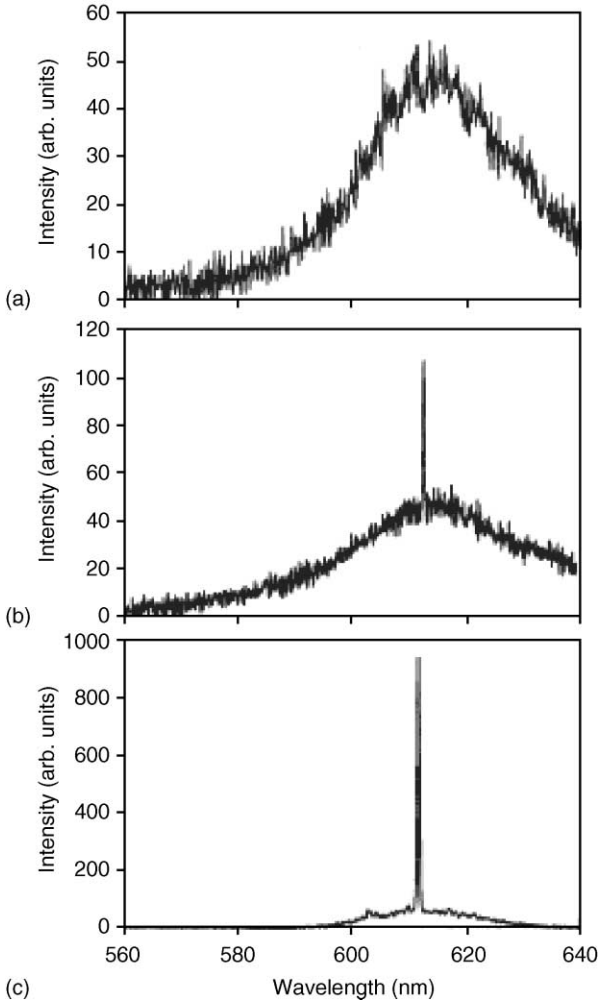


Figure 11 Emission spectra of the laser dye DCM encapsulated in a dendrimer-based resin that was patterned into a grating microstructure by single-beam 3DM. The grating was pumped at 532 nm (8 ns pulses) with excitation energies of (a) 0.03, (b) 0.05, and (c) 0.20 μJ. Starting at excitation energies as in (b), a narrow band at 612 nm appears over the spontaneous emission. The relative intensity of this lasing emission increases with excitation intensity with respect to the spontaneous emission. Reproduced from Yokoyama S, *et al.* (2003) *Thin Solid Films* 438–439: 452–456. Copyright (2003), with permission from Elsevier.

the periodicity of the structure is correspondingly different; and (ii) the I_{th} isophote, and thus the details of the polymerized structure inside each unit cell, is different under OPE and TPE because the excitation rate is proportional to I and I^2, respectively. In principle, better defined features can be obtained using two-photon or higher-order excitation because the transition between regions of low- and high-exposure is sharper than that under OPE, due to the nonlinear dependence on I. Interference fabrication techniques can enable the generation of features considerably smaller than the diffraction limit, when the beam geometry and excitation intensity are chosen appropriately.

The beams used to generate the interference pattern are usually obtained by splitting the output of a single laser source into the desired number of components. The intensity and polarization of each beam are then independently controlled. Temporal overlap of the beams at the sample can be ensured by introducing an optical delay into each of the beam paths. As fluctuations in the beam intensity can cause changes in the lattice unit cell or its filling ratio, the use of a laser with good stability is imperative in cases where exposure time extends over more than a single laser pulse.

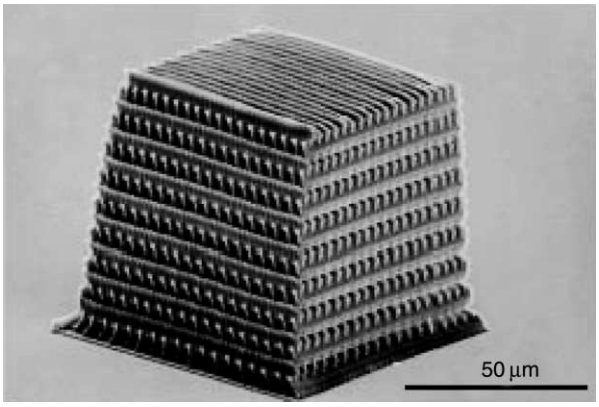


Figure 12 Example of a wood-pile PhC structure obtained by single-beam TPIP-3DM. The periodicity is 5 μm. The change in the cross-section of the structure from top to bottom, is due to shrinking of the material during development, whose effect is largest in the region farthest from the substrate. Reproduced with permission from *Nature*, Cumpston BH, Ananthavel SP, Barlow S, *et al.* (1999) Two-photon polymerization initiators for three-dimentional optical data storage and microfabrication. *Nature* 398: 51–54, Copyright (1999), Macmillan Magazines Limited.

Structures and Functional Devices

The versatility of 3DM is easily appreciated by surveying the variety of microstructures that have been fabricated over the years. These may be divided into two categories: (i) structures whose characteristics lay in their form and in the interconnectivity of their parts; and (ii) structures that incorporate an active functionality that can be exploited in a specific application. Some of these microstructures are described below.

Fabricating functional micro-optical devices is a promising application of 3DM. Figure 10 shows a grating fabricated by two-beam-interference TPIP in an acrylate with a period of 3.4 μm, whose diffraction efficiency is 28% as fabricated and 57% after annealing. The same method was used to generate a surface relief grating in a resin containing a polycationic peptide. In the presence of aqueous silicic acid, the peptide catalyzes the formation of ordered arrays of silica nanospheres in

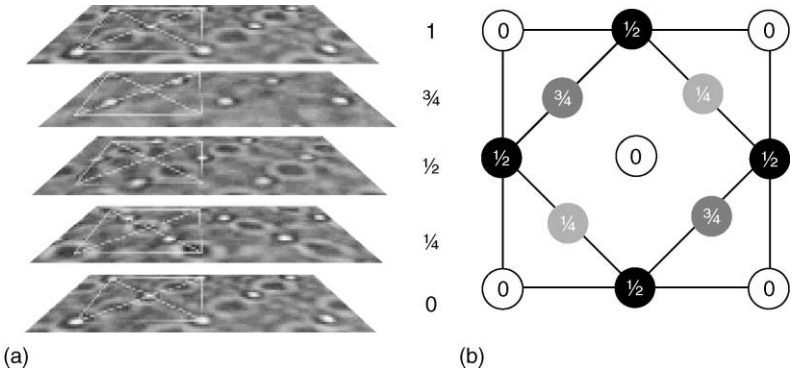


Figure 13 (a) Cross-sectional optical images of various planes of a diamond lattice fabricated in glass. The numbers to the right are the relative positions of the planes, in units of the lattice constant. (b) Diagram of the expected position of lattice points in a diamond structure. Points located in different planes have different shading. Reproduced from Mizeikis V, *et al.* (2001) *Journal of Photochemistry and Photobiology A* 145: 41–47. Copyright (2001), with permission from Elsevier.

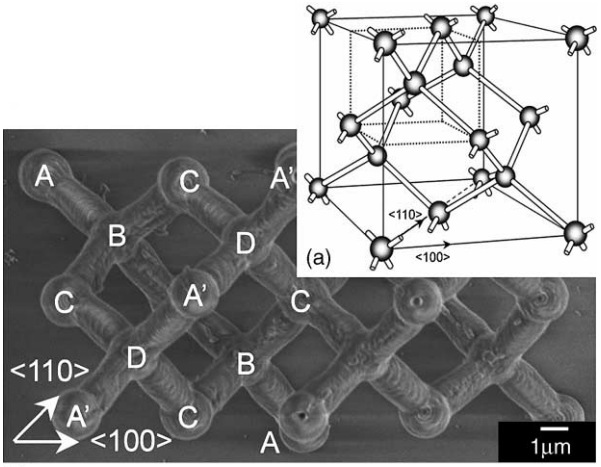


Figure 14 (a) Illustration of a diamond lattice. (b) Scanning electron micrograph of a polymeric microstructure with the symmetry of a diamond lattice, obtained by single-beam TPE-3DM. The letters A, B, C, D, A′ refer to the lattice planes at 0, 1/4, 1/2, 3/4, and 1 lattice units in the vertical direction. Reproduced from Kaneko K, *et al.* (2003) *Applied Physics Letters* 83: 2091–2093. Copyright (2003), with permission of The American Institute of Physics.

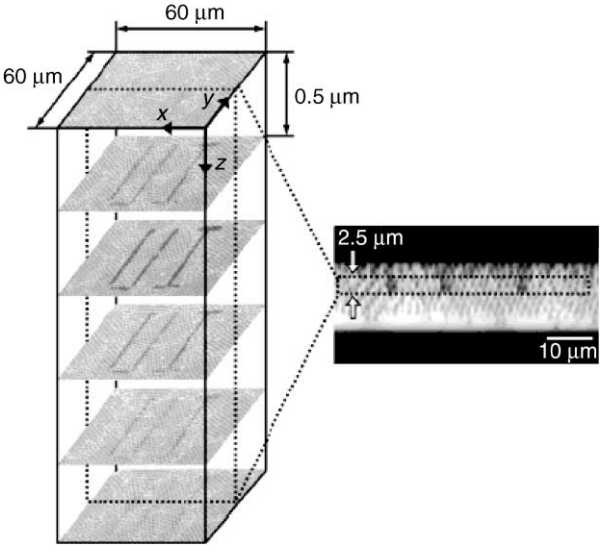


Figure 15 (Left) Images obtained by laser scanning confocal microscopy of linear defects generated by three-photon-induced polymerization of a triacrylate resin backfilled into a silica colloidal crystal. The spacing between the planes of the images is 0.5 μm. (Right) Cross-sectional image in the *xz*-plane of the same structure. Reproduced with permission from Lee W, *et al.* (2002) *Advanced Materials* 14: 271–274. Copyright (2002), with permission of Wiley.

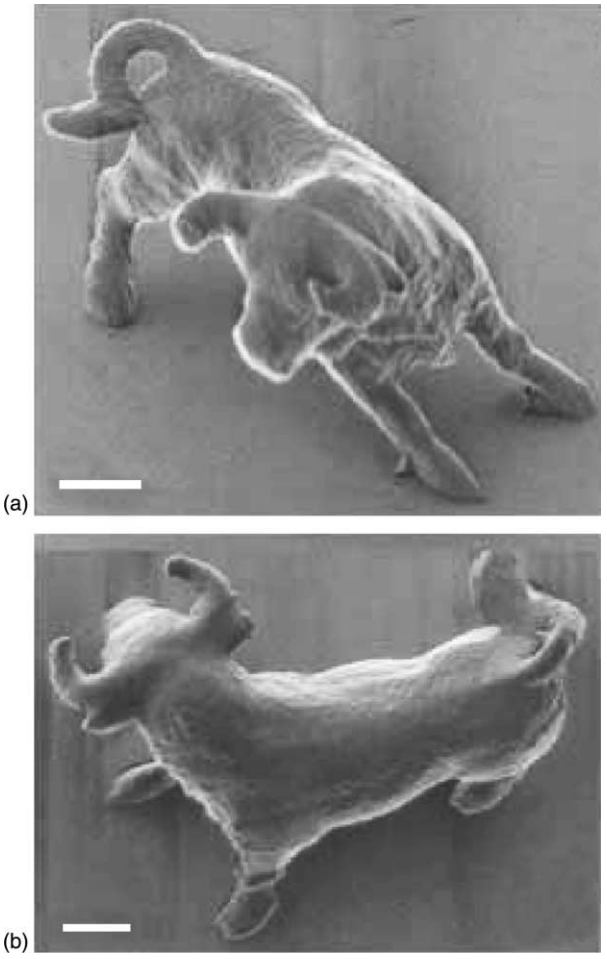


Figure 16 Microstructure of a bull fabricated by single-beam 3DM. The scale bar corresponds to 2 μm. Reproduced with permission from *Nature*, Kawata S, *et al.* (2001) *Nature* 412: 697–698. Copyright (2001), Macmillan Magazines Limited.

the troughs of the grating, which increases the diffraction efficiency 50-fold. 3DM was used to fabricate a grating structure in an acrylate resin containing the laser dye 4-(dicyanomethylene)-2-methyl-6-(4-dimethylaminostyryl)-4H-pyrane (DCM) encapsulated in a host dendrimer (see above) as an active optical medium. When pumped at 532 nm above a certain threshold, lasing emission from DCM was observed at 612 nm, with a bandwidth of 0.15 nm (Figure 11).

Several groups have explored the potential of 3DM as a tool for patterning waveguide-based optical circuitry. For example, a mixture of thermally and photocurable resins was used to generate embedded waveguides by photocross-linking within the thermally set medium. A Y-coupler was created in fused silica through the refractive index change generated by patterned MPE.

3DM provides a powerful route to 3D photonic crystals (PhC). PhCs are structures in which a basic motif, or 'unit cell', is repeated within the material with a given periodicity. PhCs have attracted great interest, because the modulation of the refractive

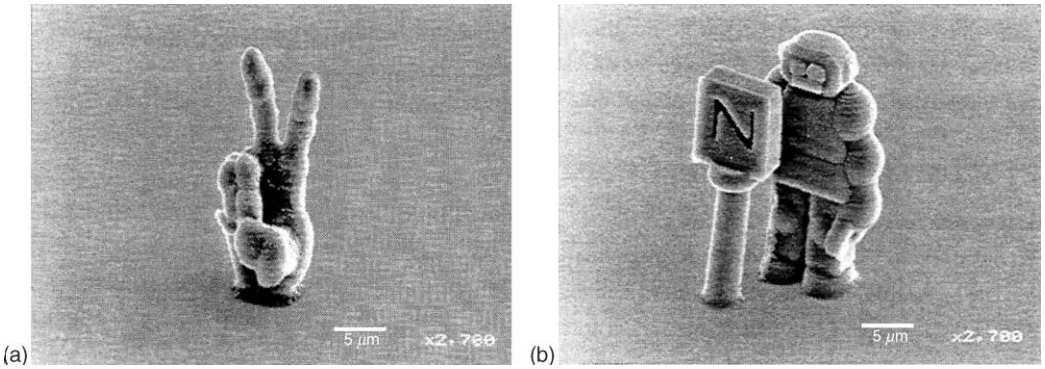


Figure 17 Scanning electron micrographs of two structures in the form of (a) a hand and (b) a robot which were fabricated by single-beam 3DM. Reproduced from Maruo S and Ikuta K (2000) *Proceedings of the Society of Photo-Optical Instrumentation Engineers* 3937: 106–112. Copyright (2000), with permission of SPIE.

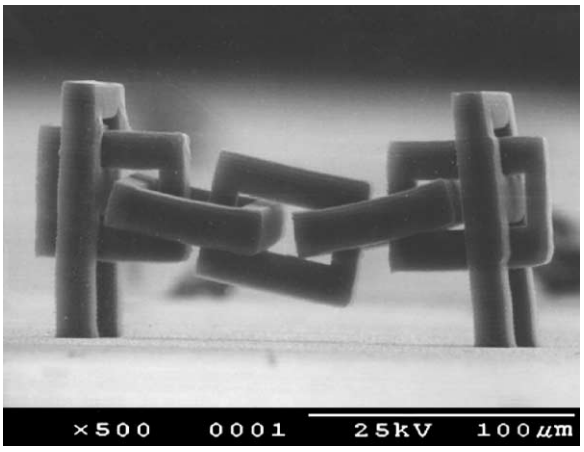


Figure 18 Scanning electron micrograph of a series of connected chain links obtained by single beam TPIP-3DM in an acrylate resin. Notice the tilt angle between the central links, indicating that each link can move with respect to the others. Reproduced from Kuebler SM, Rumi M, Watanabe T, *et al.* (2001) Optimizing two-photon initiators and exposure conditions for three-dimentional lithographic microfabrication. *Journal of Photopolymer Science and Technology* 14: 657–668. Copyright (2001), with permission of the Technical Association of Photopolymers Japan.

index between the 'filled' areas and the surrounding space provides a means for controlling the propagation of electromagnetic radiation within the structure. PhC with periodicities in the submicron and micron range exhibit their peculiar properties in the visible and near infrared spectral region. Optical wavelength PhCs can be fabricated in a straightforward way by 3DM, either by serial patterning using the single beam approach or by parallel patterning using the interference method, for which the number and orientation of the beams is dictated by the symmetry of the desired structure. Typical examples are woodpile structures (Figure 12), which consist of layers of rods of a material fabricated side by side with a fixed spacing, and alternate layers are rotated by 90° with respect to one another. If the rods in one layer are displaced by half the lattice constant with respect to the previous layer with the same orientation, the structure obtained has the symmetry of a face-center cubic lattice. PhCs with the symmetry of the diamond crystal have also been fabricated, either by inducing changes in a glass

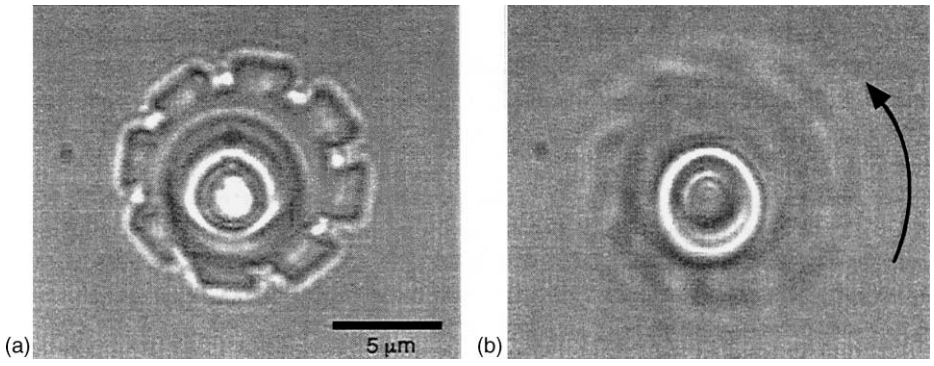


Figure 19 Optical image of a micro-wheel obtained by single-beam 3DM. (a) Structure before development. (b) When ethanol is used to remove the unpolymerized material, the wheel rotates around the axle in the solvent. Reproduced from Maruo S and Ikuta K (2000) *Proceedings of the Society of Photo-Optical Instrumentation Engineers* 3937: 106–112. With permission of the Institute of Electrical Engineers of Japan.

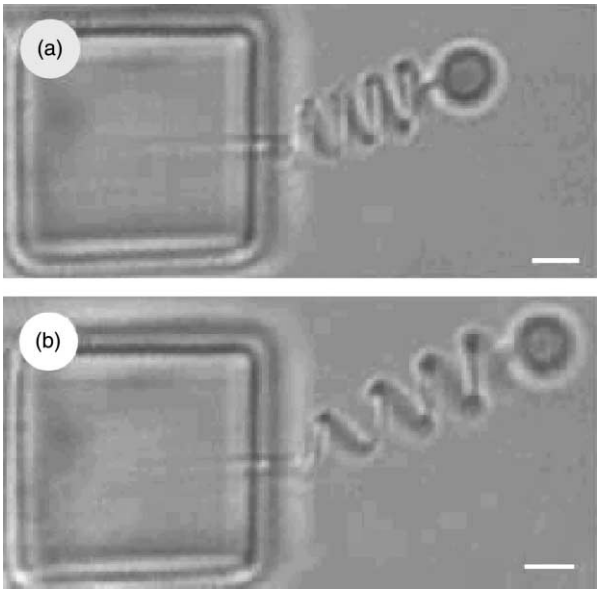


Figure 20 Optical image of a micro-spring tethered to a polymer block. The structure in (a) is at its resting point in ethanol. In (b) the spring was elongated by displacing the polymer bead at the free end using laser trapping. Reproduced with permission from *Nature*, Kawata S, *et al.* (2001) *Nature* 412: 697–698. Copyright (2001), Macmillan Magazines Limited.

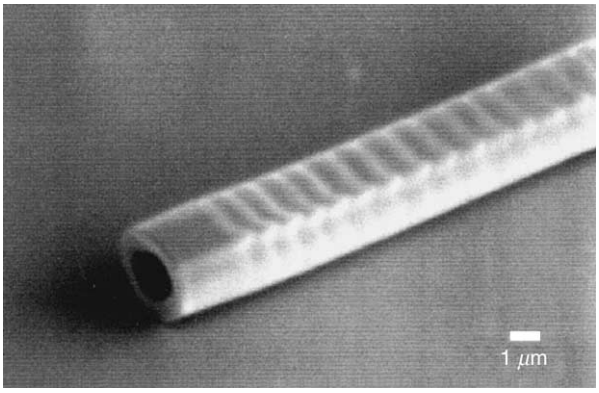


Figure 21 Hollow micro-tubule generated by single-beam TPIP-3DM in a urethane acrylate. The inner diameter is 1.8 μm. Reproduced from S Maruo and S Kawata (1998) *Journal of Microelectromechanical Systems* 7: 411–415. Copyright (1998), with permission of IEEE.

substrate at lattice points (Figure 13), or by constructing a network of polymeric rods connecting nearest neighbor points (Figure 14). It was showed that a PhC created by TPIP-3DM does exhibit a partial bandgap in the infrared spectral region. The position of the bandgap can be tuned by fabricating the structure with the appropriate periodicity.

The ability to introduce defects in PhCs in a controlled fashion is of key importance for the application of these structures in optical circuitry, as they can act as waveguides for channeling light into preferential directions. Point, line, and plane defects can be generated easily, by locally altering the fabrication pattern in the PhC. Point-by-point MPE of a Ge-doped glass was used to pattern a cubic lattice containing systematic defects that defined a Y-shaped waveguide splitter in a [100] plane. 3DM can also be used to modify a structure that is created using another fabrication method, and additional materials can be introduced into the structure in this second fabrication step. A silica colloidal crystal was back-filled with a polymerizable material and waveguides were generated by photocross-linking channels in the medium through three-photon-excitation 3DM (Figure 15).

The full potential of single-beam 3DM can be appreciated by surveying some of the structures that have been fabricated, featuring topologically complex shape, high interconnectivity, and extreme under-cut. Examples include those of Figures 16, 17, and the diamond-lattice PhC of Figure 14. Structures with freely movable parts, like those needed for MEMS, are difficult to obtain by other fabrication techniques, but can often be made by 3DM in a single exposure step. The ability to generate structures with independent but interconnected parts is exemplified by the microchain of Figure 18. A further example is the microgear in Figure 19, which is shown to spin around its axle when the sample is flooded with a liquid. Other functional microstructures have also been reported, such as the microspring in Figure 20, fabricated in a urethane acrylate resin and characterized by a core diameter of only 300 nm. When the sample is immersed in ethanol, the pendant mass can be captured and extended with optical tweezers. When the mass is released, the spring behaves like a damped oscillator.

Promising results have also been obtained in the production of structures to be used in micro-fluidics applications. Of key importance, in this case, is the ability to pattern channels, cavities, and other hollow structures that can be used to store or transfer small amounts of fluids between different parts of a device, as needed to perform, for example, analytical processes on a femtoliter scale. Examples include the hollow microtubule (Figure 21), and the submersed channels structure fabricated in a positive-tone resist (Figure 5).

List of Units and Nomenclature

A	Electron accepting molecular fragment
AFS	Amplified femtosecond
C	Chromophore
$\mathbf{C}^*$	Excited chromophore
CW	Continuous wave
D	Electron donating molecular fragment
3D	Three-dimensional
3DM	Three-dimensional microfabrication
$\vec{E}_0$	Spatial component of the electric field vector [V m^{-1}]
E_p	Laser pulse energy [J]
$f(t)$	Normalized temporal distribution of the beam intensity [dimensionless]
FOM	Figure of merit
h	Planck's constant [J s]
$i(z, r)$	Spatial distribution of the beam intensity [W cm^{-2}]
I	Intensity of a propagating beam [W cm^{-2}]
I_0	Maximum on-axis intensity for a Gaussian beam [W cm^{-2}]
I_{th}	Threshold intensity (minimum peak intensity at which polymerization occurs for a given exposure time) [W cm^{-2}]
I·	Initiating radical
j, l, q	Integers [dimensionless]
J_1	First order Bessel function [dimensionless]
K	Amplitude of a beam wave vector ($K = 2\pi/\lambda$) [m^{-1}]
$\vec{K}$	Wave vector of a beam [m^{-1}]
m	Number of beams in a multi-beam interference configuration [dimensionless]
M	Monomer
MBI	Multi-beam interference
MEMS	Micro-electromechanical systems
MPE	Multi-photon excitation
n	Number of photons involved in MPE [dimensionless]
n_0	Linear refractive index of the medium [dimensionless]
N	Number density (concentration) of a species [cm^{-3}]
NA	Numerical aperture of a lens [dimensionless]
N_C	Number density of chromophores present [cm^{-3}]
Nd:YAG	Neodymium:Yttrium aluminum garnet
$N_{I\cdot}$	Number density of initiating radicals generated [cm^{-3}]
N_{Inh}	Number density of inhibitors present [cm^{-3}]
OPE	One-photon excitation
Ormocer	Organically modified ceramic
$\langle P \rangle$	Average laser power [W]
$\langle P \rangle_{min}$	Minimum average laser power for patterning a material [W]
	Repetition rate (number of laser pulse per unit time) [Hz]
PAG	Photo-acid generator
PhC	Photonic crystal
r	Radial coordinate (distance from optical axis) [m]
r'	Component or $\vec{r}$ along a given wave vector [m]
$\vec{r}$	Position vector in three-dimensional space [m]
R	Twice the radial coordinate of the point at which the intensity is $I_0/2$ [m]
$R_{1h\nu}$	Molecular one-photon excitation rate (number of molecules excited per unit volume and time via OPE) [$cm^{-3}\,s^{-1}$]
$R_{2h\nu}$	Molecular two-photon excitation rate (number of molecules excited per unit volume and time via TPE) [$cm^{-3}\,s^{-1}$]
$R_{nh\nu}$	Molecular n-photon excitation rate (number of molecules excited per unit volume and time via n-order MPE) [$cm^{-3}\,s^{-1}$]
RMS	Root-mean-square
t	Time [s]
TEM_{00}	Lowest order transverse electromagnetic mode of a laser
Ti:S	Titanium:sapphire
TPE	Two-photon excitation
TPIP	Two-photon induced polymerization
z	Coordinate along the beam propagation direction [m]
z_R	Rayleigh range of a Gaussian beam; axial distance between the point of minimum beam waist and the point at which the waist has increased by a factor of $\sqrt{2}$ [m]
Z	Distance between the points of intensity $I_0/2$ along the coordinate z [m]
ζ	Longitudinal optical coordinate ($\zeta = 2\pi Z\,(\mathrm{NA})^2/n_0\lambda$) [dimensionless]
θ	Angle between a beam wave vector and the normal to a substrate [rad]
λ	Wavelength (in vacuum) [m]
Λ	Period of a one-dimensional grating [m]

ν	Frequency [s^{-1}]
$\boldsymbol{\pi}$	π-conjugated molecular fragment
ρ	Transverse optical coordinate ($\rho = 2\pi r \mathrm{NA}/\lambda$) [dimensionless]
$\sigma^{(1)}$	One-photon absorption cross-section [cm^2]
$\sigma^{(2)}$	Two-photon absorption cross-section [cm^4 s $photon^{-1}$]
$\sigma^{(n)}$	n-photon absorption cross-section [cm^{2n} s^{n-1} $photon^{-(n-1)}$]
τ_{ex}	Exposure time [s]
τ_p	Laser pulse duration [s]
$\phi_c^{(n)}$	n-photon chemical quantum yield [(# of chemical species generated)/(# of n-photon-excited species)] [dimensionless]
ϕ_{H+}	Chemical yield for the generation of H^+ (number of H^+ generated per excited molecule) [dimensionless]
ω	Angular frequency ($\omega = 2\pi\nu$) [Hz]

See also

Chemical Applications of Lasers: Nonlinear Spectroscopies. **Diffractive Systems:** Applications of Diffractive and Micro-Optics in Lithography; Design and Fabrication of Diffractive Optical Elements. **Holography, Techniques:** Overview. **Imaging:** Volume Holographic Imaging. **Laser-Induced Damage of Optical Materials**. **Materials for Nonlinear Optics:** Organic Nonlinear Materials. **Microscopy:** Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy; Nonlinear Microscopy; Overview. **Photonic Crystals:** Electromagnetic Theory; Nonlinear Optics in Photonic Crystal Fibers; Self-Assembled and Functionalized Photonic Crystals. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Belfield KD, Schafer KJ, Liu YU, Liu J, Ren XB and van Stryland EW (2000) Multiphoton-absorbing organic materials for microfabrication, emerging optical applications and non-destructive three-dimensional imaging. *Journal of Physical Organic Chemistry* 13: 837–849.

Born M and Wolf E (1980) *Principles of Optics*, 6th edn. Cambridge, UK: Cambridge University Press.

Cumpston BH, Ananthavel SP, Barlow S, *et al.* (1999) Two-photon polymerization initiators for three-dimensional optical data storage and microfabrication. *Nature* 398: 51–54.

Joshi MP, Pudavar HE, Swiatkiewicz J, Prasad PN and Reinhardt BA (1999) Three-dimensional optical circuitry using two-photon-assisted polymerization. *Applied Physics Letters* 74: 170–172.

Kawata S and Sun H-B (2003) Two-photon photopolymerization as a tool for making micro-devices. *Applied Surface Science* 208–209: 153–158.

Kuebler SM, Rumi M, Watanabe T, *et al.* (2001) Optimizing two-photon initiators and exposure conditions for three-dimensional lithographic microfabrication. *Journal of Photopolymer Science and Technology* 14: 657–668.

Kuebler SM, Braun KL, Zhou W, *et al.* (2003) Design and application of high-sensitivity two-photon initiators for three-dimensional microfabrication. *Journal of Photochemistry and Photobiology A: Chemistry* 158: 163–170.

Martineau C, Anémian R, Andraud C, Wang I, Bouriau M and Baldeck PL (2002) Efficient initiators for two-photon induced polymerization in the visible range. *Chemical Physics Letters* 362: 291–295.

Odian G (1981) *Principles of Polymerization*, 2nd edn. New York: Wiley.

Pitts JD, Campagnola PJ, Epling GA and Goodman SL (2000) Submicron multiphoton free-form fabrication of proteins and polymers: studies of reaction efficiencies and applications in sustained release. *Macromolecules* 33: 1514–1523.

Ren Y, Yu X-Q, Zhang D-J, *et al.* (2002) Synthesis, structure and properties of a new two-photon photopolymerization initiator. *Journal of Materials Chemistry* 12: 3431–3437.

Strickler JH and Webb WW (1990) Two-photon excitation in laser scanning fluorescence microscopy. *Proceedings of the Society of Photo-Optical Instrumental Engineers* 1398: 107–118.

Sun H-B and Kawata S (2003) Two-photon laser precision microfabrication and its applications to micro-nano devices and systems. *Journal of Lightwave Technology* 21: 624–633.

Sun H-B, Maeda M, Takada K, Chon JWM, Gu M and Kawata S (2003) Experimental investigation of single voxels for laser nanofabrication via two-photon photopolymerization. *Applied Physics Letters* 83: 819–821.

Wu ES, Strickler JH, Harrell WR and Webb WW (1992) Two-photon lithography for microelectronic application. *Proceedings of the Society of Photo-Optical Instrumental Engineering* 1674: 776–782.

Wu P-W, Cheng W, Martini IB, Dunn B, Schwartz BJ and Yablonovitch E (2000) Two-photon photographic production of three-dimensional metallic structures within a dielectric matrix. *Advanced Materials* 12: 1438–1441.

Xu C and Webb WW (1997) Multiphoton excitation of molecular fluorophores and nonlinear laser microscopy. In: Lakowicz JR (ed.) *Topics in Fluorescence Spectroscopy*, vol. 5, pp. 471–540. New York: Plenum.

NONLINEAR OPTICS, BASICS

Contents

Cascading
$\chi^{(2)}$–Harmonic Generation
$\chi^{(3)}$–Third-Harmonic Generation
Four-Wave Mixing
Kramers–Krönig Relations in Nonlinear Optics
Nomenclature and Units
Nonlinear Optical Phase Conjugation
Photorefraction
Ultrafast and Intense-Field Nonlinear Optics

Cascading

G Assanto, University 'Roma Tre', Rome, Italy

G I Stegeman, University of Central Florida, Orlando, FL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Historically the field of nonlinear optics has evolved along two main axes, depending on the order of the nonlinearity used. The phenomena associated with the second-order nonlinearity $\chi^{(2)}$ have dealt with frequency conversion requiring wavevector matching (i.e., momentum conservation), such as second harmonic and optical parametric generation. On the other hand, effects associated with the third-order susceptibility $\chi^{(3)}$ have focused on inherently phase-matched phenomena at a single frequency, for example an intensity-dependent refractive index or degenerate four-wave mixing.

In cascading, multiple sequential nonlinear events based on $\chi^{(2)}$ can mimic some well-known $\chi^{(3)}$ phenomena. For example, it leads to nonlinear phase shifts $\Delta\phi^{\mathrm{NL}}$ reminiscent of those obtained through self-phase modulation over a distance L, i.e.,

$$\Delta\phi^{\mathrm{NL}}(L, I) = k_0 L n_2 I \qquad [1]$$

where I is the irradiance, k_0 the vacuum wavevector, and n_2 the Kerr coefficient or nonlinear refractive index. These can be used for all-optical switching, the generation of spatio-temporal solitons, etc. Some of these cascading effects have been known from the very earliest days of nonlinear optics, but their effectiveness was not realized until recently, when it proved possible to phase match second-order interactions with usefully large nonlinearities. Note that the medium must be $\chi^{(2)}$-active, i.e., noncentrosymmetric, which usually requires crystals rather than amorphous materials (such as glass).

Here we review the basic concepts, the characteristic interactions, some of the numbers involved and, finally, a few applications.

Physics of Cascading

Nonlinear Phase Shift in Collinear Second Harmonic Generation

The most important cascading effect, the nonlinear phase shift, is rooted in the fundamentals of the parametric interaction when two or three waves of different frequencies ($\omega_3 = \omega_1 \pm \omega_2$) are coupled by the second-order susceptibility $\chi^{(2)}(-\omega_3;\ \omega_1,\ \pm\omega_2)$. The simplest case is 'Type I' second harmonic generation (SHG), in which one eigenwave at the fundamental (FF) and one at the harmonic frequency (SH) interact with $\omega_1 = \omega_2 = \frac{1}{2}\omega_3$. When the waves have different phase velocities, this process involves simultaneously up-conversion ($\omega_1 + \omega_1 \rightarrow \omega_3$) and down-conversion ($\omega_3 - \omega_1 \rightarrow \omega_1$). By the latter the FF is regenerated from the harmonic after a characteristic propagation distance called the coherence length. This phenomenon is cyclic (see Figure 1a).

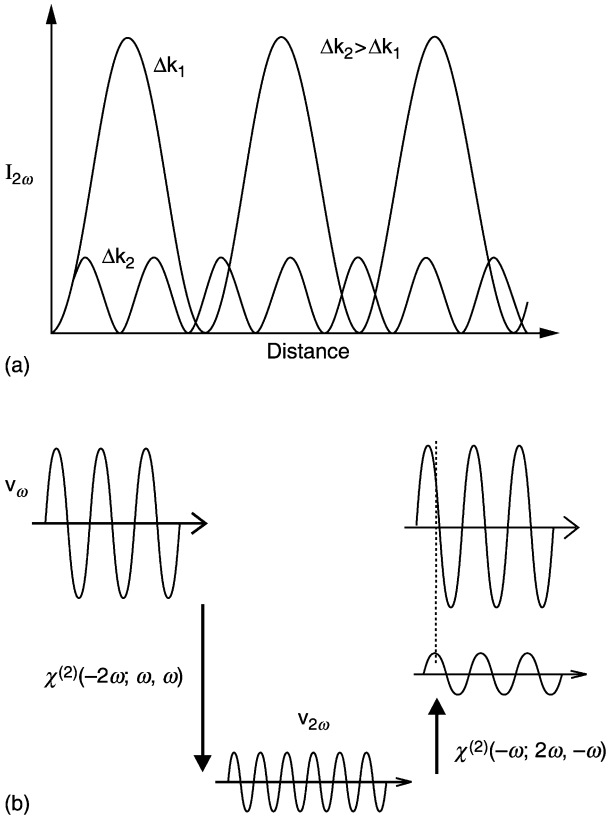


Figure 1 Phase-mismatched SHG. (a) Typical evolution of the generated second harmonic irradiance versus propagation for different mismatches. The period is twice the coherence length. (b) Intuitive picture of phase shift through cascading: the fraction of the initial energy which is first up-converted (SHG) and later down-converted (DFG), having traveled at a different phase velocity due to the wavevector mismatch, recombines with the through-FF causing an overall phase shift.

The field amplitudes in normalized form are

$$E_1(z,t) = \frac{1}{2}\sqrt{\frac{2}{c\varepsilon_0 n_1}}a_1(z)\mathrm{e}^{\mathrm{i}(\omega_1 t - k_1 z)} + \text{c.c.}$$
$$E_3(z,t) = \frac{1}{2}\sqrt{\frac{2}{c\varepsilon_0 n_3}}a_3(z)\mathrm{e}^{\mathrm{i}(\omega_3 t - k_3 z)} + \text{c.c.} \qquad [2]$$

where $|a_1|^2$ and $|a_3|^2$ are power densities, c is the speed of light in vacuum, and n_j and k_j are the refractive indices and wavevectors at angular frequencies $\omega_j\,(j = 1, 3)$, respectively. As is common in nonlinear optics, the efficiency of these processes depends on the irradiance of the different waves, and on the wavevector mismatch between them, i.e., $\Delta k = 2k_1 - k_3$ for collinear beams. Since the phase velocities of the two waves, $v_{\omega_1} = \omega_1/k_1$ and $v_{\omega_3} = \omega_3/k_3$, are generally different and $\Delta k \neq 0$, their phase fronts fall out of synchronism during propagation. This leads to a periodic oscillation with distance (coherence length $= \frac{1}{2}$period $= \pi/|\Delta k|$) in the irradiances of the fundamental and harmonic waves, conserving the total energy but giving rise to a cascading phase shift, as pictured in Figure 1b.

The equations that describe the evolution of the fields stem from coupled mode theory, and are given by

$$\frac{\mathrm{d}}{\mathrm{d}z}a_3(z) = -\mathrm{i}\Gamma a_1^2(z)\mathrm{e}^{-\mathrm{i}\Delta k z}$$
$$\frac{\mathrm{d}}{\mathrm{d}z}a_1(z) = -\mathrm{i}\Gamma a_3(z)a_1^*(z)\mathrm{e}^{\mathrm{i}\Delta k z} \qquad [3]$$

where

$$\Gamma = \frac{\omega_1 \chi^{(2)}_{\mathrm{eff}}}{\sqrt{2n^3c^3\varepsilon_0}} \qquad [4]$$

$\chi^{(2)}_{\mathrm{eff}}$ is the 'effective' nonlinearity (for a specific crystal orientation and field polarization) and we assumed $n = n_1 \cong n_3$. In the limit of a large mismatch, the energy conversion to the harmonic is negligibly small, and it is possible to find an analytical expression for an effective 'n_2', $n_{2,\mathrm{eff}}$. In this limit $|a_1(z)| = |a_1(0)| =$ constant, and, in the absence of loss, solving for the small SH wave and substituting back into the equation for $\mathrm{d}a_1(z)/\mathrm{d}z$ gives

$$\frac{\mathrm{d}}{\mathrm{d}z}a_1(z) = -\mathrm{i}\frac{\Gamma^2}{\Delta k}\Big\{1 - \cos(\Delta k z) - \mathrm{i}\sin(\Delta k z)\Big\}|a_1(0)|^2 a_1(z) \qquad [5]$$

By comparing this with the corresponding equation based on n_2 ($\propto\chi^{(3)}[-\omega;\omega,-\omega,\omega]$) and describing self-phase modulation, namely

$$\frac{\mathrm{d}}{\mathrm{d}z}a_1(z) = -\mathrm{i}n_2\frac{\omega_1}{c}|a_1(z)|^2 a_1(z) \qquad [6]$$

we can define an equivalent nonlinear refractive index

$$n_2(z) = \frac{\pi\{\chi^{(2)}_{\mathrm{eff}}\}^2}{2\varepsilon_0 c n^4 \lambda_1 \Delta k}[1 - \cos(\Delta k z)] \qquad [7]$$

which turns out to be strictly nonlocal because it relies on propagation. Since the $\cos(\Delta k z)$ term oscillates rapidly in space, the above averages to

$$n_{2,\mathrm{eff}} = \frac{\pi\{\chi^{(2)}_{\mathrm{eff}}\}^2}{2\varepsilon_0 c n^4 \lambda_1 \Delta k} \qquad [8]$$

In this so-called 'Kerr limit' cascading leads to a Kerr-like nonlinear response, with either a

self-focusing ($\Delta k > 0$, $n_{2,\text{eff}} > 0$) or a self-defocusing nonlinearity, depending on the sign of Δk.

It is worthwhile to link the above equations to the physics of cascading with a simple argument. The evolution of the waves on propagation depends on their fields (in both amplitude and phase) at the input. When only the fundamental beam at $\omega = \omega_1$ is incident (the classical case of SHG), the harmonic is initially up-converted $\pi/2$ out of phase with the input (first of eqns [3]) via $\chi^{(2)}(-2\omega; \omega, \omega)$. If $\Delta k = 0$, the phase fronts travel at the same velocity and the relative phase is locked in at $\pi/2$. Difference frequency generation (DFG, second of eqns [3]) results in a down-converted fundamental which is an additional $\pi/2$ out of phase, i.e., exactly out of phase with the original input field. This leads to the monotonic depletion of the input and the growth of the SH with distance. If $\Delta k \neq 0$, there is only partial and periodic conversion to the harmonic, the wavefronts travel at different velocities (v_ω and $v_{2\omega}$, respectively), and the relative phase between the two waves changes with distance. As a result, the down-converted FF is no longer π out of phase with the input, thereby changing the overall phase of the total input-frequency wave by a cascaded amount $\Delta\phi^{\text{NL}}$. This is the mechanism sketched in Figure 1b and buried in the better-known Maker-fringe effect represented in Figure 1a. Its magnitude depends on the irradiance and the mismatch, and its sign on that of Δk (see eqn [8]). The higher the input irradiance, the larger the harmonic conversion, and hence the larger $\Delta\phi^{\text{NL}}$.

Alternatively, with a harmonic input at high irradiance and a weak seed at the FF, parametric amplification makes the fundamental grow via down-conversion. If $\Delta k = k_3 - 2k_1 \neq 0$, self-focusing (-defocusing) is obtained on the strong harmonic for $\Delta k > 0$ ($\Delta k < 0$). These phenomena can be generalized to the nondegenerate cases of sum and difference frequency generation, keeping in mind that when using three input waves their mutual phase relationship is essential in determining the details of the interaction.

The solutions to eqns [3] can be obtained numerically or written in terms of Jacobi elliptic integrals subject to boundary conditions ($a_3(0) = 0$, $a_1(0) \neq 0$) at $z = 0$, providing additional insight into the cascading phenomenon. In contrast to the Kerr case given by eqn [1], the variation in n_2 with distance implies a staircase-like increase in $\Delta\phi^{\text{NL}}$ with distance for various detunings ΔkL, as shown in Figure 2a. This is consistent with our intuitive picture above, the increase in $\Delta\phi^{\text{NL}}$ being maximized where the fundamental is the smallest (see Figure 2b for the oscillations in the fundamental amplitude), with the step-period corresponding to twice the coherence length. Conversely, for a fixed input irradiance, the evolution of $\Delta\phi^{\text{NL}}$ exhibits two extrema with detuning ΔkL, as shown in Figure 3a. Note also that, for small ΔkL and large

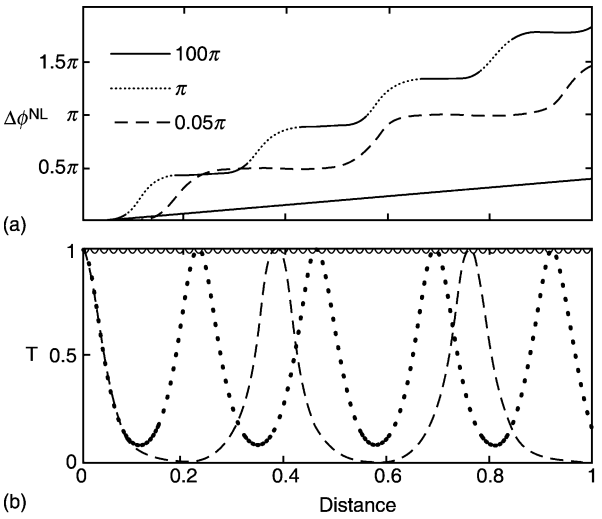


Figure 2 Calculated nonlinear phase (a) and throughput (b) of the fundamental field versus normalized propagation distance, for three values of phase mismatch.

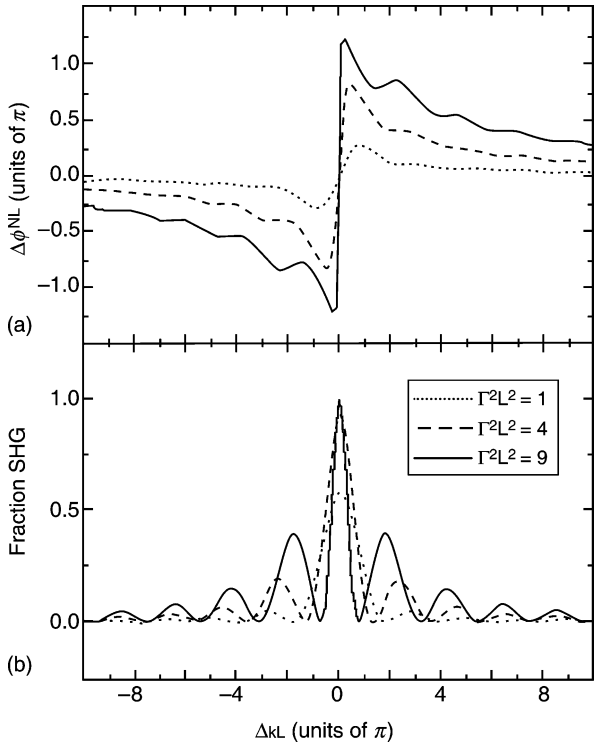


Figure 3 Calculated FF phase shift (a) and generated SH fraction (b) versus phase mismatch, i.e., the detuning curve, for three different nonlinear strengths (unitary FF irradiance).

(unrealistically high) irradiances, the increase in $\Delta\phi^{NL}$ tends to become proportional to the field amplitude, a clear indication that cascading depends on a different order of nonlinearity than $\chi^{(3)}$, namely $\chi^{(2)}$. On the other hand, for large ΔkL the growth of $\Delta\phi^{NL}$ is slower but linear with distance and irradiance, as in eqn [1].

There are trade-offs between the fundamental output and the net nonlinear phase shift. This is clearly visible in Figures 3a,b: the maxima in $\Delta\phi^{NL}$ correspond to a reduced FF throughput. Although it is possible to operate cw at the zeros of the SHG, their location changes with the irradiances required for useful phase shifts ($>\pi$), complicating the trade-offs and the response to temporal pulses (encompassing a continuous irradiance distribution). The problem can be alleviated at the expense of reducing the net phase shift for a given irradiance and propagation length, by introducing regions with a nonuniform matching condition at both ends of the sample. The goal at the input end is to first adiabatically generate the SH required for optimum $\Delta\phi^{NL}$. At the output, this harmonic is adiabatically converted back to the FF. The corresponding output for wavelengths larger than the phase-matching condition exhibits a $\Delta\phi^{NL}$ with negligible fundamental depletion. This approach has been used in temperature-tuned and quasi phase matched (QPM, where a periodic modulation of the nonlinearity provides phase matching via the additional grating momentum) lithium niobate waveguides to generate phase shifts of π with less than 10% net loss of the fundamental wave to the harmonic.

While the preceding discussion has focused on Type I SHG, it is also valid for Type II SHG (i.e., with two FF input waves), provided the two fundamental inputs have equal input energy. In fact, an input imbalance between the two fundamental polarizations could give rise to a phase shift on the weaker field even when $\Delta k = 0$, leading to all-optical modulation and transistor action, as demonstrated in potassium titanyl phosphate (KTP) crystals.

Finally, due to the nature of the quadratic nonlinearity, the phase and/or amplitude of a coherent SH seed can have a large effect on the cascaded outcome, both in the phase and amplitude of the fundamental, as shown in Figure 4 for a fixed irradiance seed 1000 times weaker than the FF input. This can be used in seed-controlled optical processors.

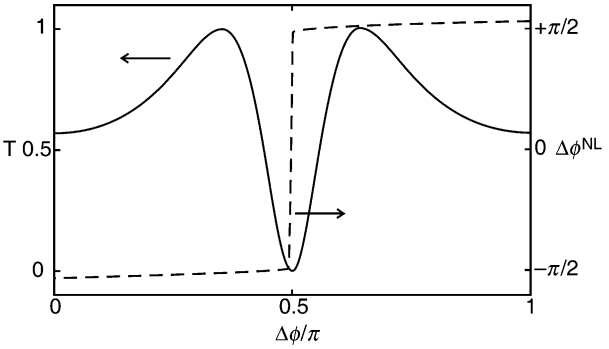


Figure 4 FF transmission (solid line) versus the relative phase of a much weaker (three orders of magnitude) SH seed. Here perfect phase matching was assumed.

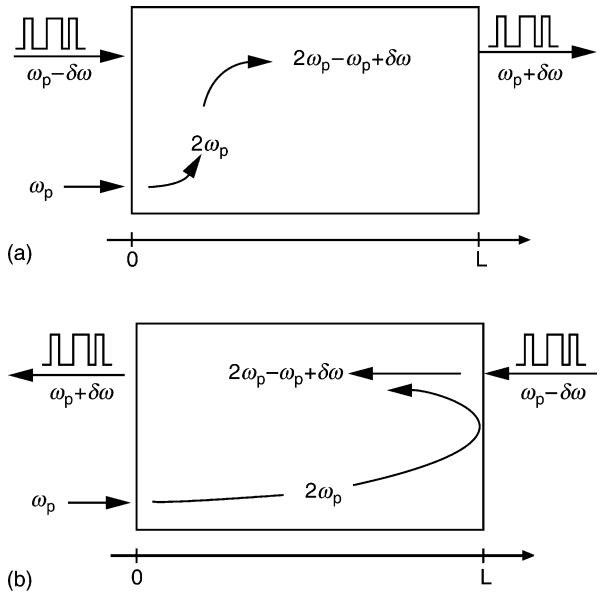


Figure 5 Wavelength shifter: (a) single-pass and (b) double-pass configurations. In (b) a dichroic mirror reflects the second harmonic of the pump.

Frequency Shifting

In the nondegenerate case of $\chi^{(2)}$ sum and difference frequency generation, two successive operations can perform the useful function of a frequency shifter (Figure 5). The idea is to shift the signal frequency ($\omega_p - \delta$) by first generating from the pump (ω_p) a second harmonic ($2\omega_p$), followed by DFG involving the latter and the signal. This leads to an output at $\omega_p + \delta$ through $\chi^{(2)}(-(\omega_p + \delta); 2\omega_p, -(\omega_p - \delta))$, i.e., a frequency shift of 2δ with an output spectrally mirror-imaged about the pump frequency. This feature is useful in WDM (wavelength division multiplexing) systems, and guarantees a protocol-insensitive wavelength shift with an additional phase conjugation useful in dispersion-compensating schemes. Notice that, since $\omega_p + \delta \cong \omega_p - \delta$ and $\delta << \omega_p$, the bandwidth is large because the interaction is near the degeneracy point.

Figure 5 shows two possible configurations for the same device length. The efficiency of the process can be increased in a double-pass geometry by using a back-propagating signal and a mirror-coated facet to reflect the SH, as sketched in Figure 5b. The best results to date were obtained at communication wavelengths with single-pass waveguides in periodically poled (QPM) $LiNbO_3$, including shift of multiple wavelength-channels with low cross-talk.

Multiple Beam Generation

Two successive $\chi^{(2)}$ processes can also be used to mimic three-wave mixing, a classic noncollinear $\chi^{(3)}$ interaction where two photons are taken out of one beam and one out of the second. In cascading, first a second harmonic is generated by one of the inputs, and then difference frequency mixing (DFM) between that SH and the second input leads to the signal of interest. In the simplest configuration, two equi-frequency (ω) beams of wavevectors $\mathbf{k}_1$ and $\mathbf{k}_2$ are incident in slightly different directions close to phase-matching and overlap inside the sample. When wave 1 is doubled (2ω) and wave 2 mixes with this harmonic via DFM, a new beam (at $2\omega - \omega \rightarrow \omega$) is generated with wavevector $2\mathbf{k}_1 - \mathbf{k}_2$. Of course the complementary output at $2\mathbf{k}_2 - \mathbf{k}_1$ is also generated via the SHG of wave 2. If these two new beams are strong enough, they can generate another pair via cascading, and so on. The generation of more than six extra beams has been demonstrated in β-barium borate with 1064 nm inputs.

This interaction can be applied to the control of the frequency chirp in pulses. For example, if wave 1 has a frequency chirp α_1 and wave 2 a chirp α_2, then the chirp of the beam scattered into the direction $2\mathbf{k}_1 - \mathbf{k}_2$ is $2\alpha_1 - \alpha_2$. This becomes especially interesting if $\alpha_2 = -\alpha_1$ so that the chirp in the new beam is $3\alpha_1$, as observed and subsequently employed in enhancing pulse compression. Note that, since a cascaded phase shift occurs on the down-conversion step back to the original fundamental (wave 1), it is not present in the scattered beam resulting in a clean transfer of the chirp.

Typical Numbers

From eqn [8] it is clear that the figure of merit (FOM) for cascaded processes is the same as for all second-order processes, namely $|\chi^{(2)}|^2/n^4$. The key is to find materials phase-matchable at the wavelengths of interest. Therefore, basically any material that is suited for frequency conversion, either SHG or parametric generation, will work. The most versatile material, which also has the largest easily usable nonlinear coefficient (18 pm/V), is lithium niobate when QPM is employed. As a reference, Table 1 gives the maximum $n_{2,\text{eff}}$ for 1-cm-long samples, obtained by extrapolating eqn [8] to the ΔkL which maximizes the nonlinear phase shift. This is not strictly valid for such small ΔkL, but it is a useful estimate.

Other Implications of Cascading

A straightforward conclusion that could be drawn is that cascading with its $n_{2,\text{eff}}$ and $\Delta\phi^{\text{NL}}$ can generally replace the Kerr n_2. This is true for self-phase modulation and the applications stemming from it, including all-optical guided-wave switches such as directional couplers, Mach–Zehnder interferometers, etc. However, this is not necessarily true for other applications, because all the waves have to be coupled through $\chi^{(2)}$ interactions (at most three at a time) and must be nearly phase-matched for efficient effects. For example, cross-phase modulation between two input FF polarization components is feasible via Type II SHG cascading, but specific wavelengths will require different crystals, or crystal cuts. The nonlinear phase shift can also be used in all-optical devices for analog processing, such as isolators, modulators, and transistors, in laser mode-locking, etc.

In the following subsections we discuss a few additional applications. Specifically, we describe a non-reciprocal device that operates as an isolator, and the cascading of beams of finite extent in space or time. The latter can counterbalance either spatial diffraction or temporal dispersion or both and gives rise to solitons of various dimensionalities.

An Optical Diode

A nonlinear response, when combined with a device structure that is asymmetric about its midpoint, can provide nonreciprocal behavior upon excitation of the structure (crystal, waveguide) from different ends.

Table 1 Figures for cascading in various materials. The effective nonlinear coefficient is estimated under optimum conditions and a 1 cm propagation length

Material L = *1 cm*	d_{ii} *(pm/V)*	d_{ij} *(pm/V)*	$n_{2,eff}$ *(cm^2/W)*
$LiNbO_3$	36		2×10^{-11}
$LiNbO_3$		5.8	5×10^{-13}
MNA(1)	165		7×10^{-10}
NPP(2)		84	2×10^{-10}
DAST(3)	600		6×10^{-9}

(1)MNA: 2-methyl-4-nitroaniline
(2)NPP: N-(4-nitrophenyl)-(L)-prolinol
(3)DAST: dimethyl amino stilbazolium tosylate

In analogy to the standard electronic element, we can refer to such device as a power-dependent optical diode. Dealing with quadratically nonlinear waveguides, simple ways to induce a directional asymmetry are variations in wavevector mismatch, or in QPM periodicity. A mismatch profile which depends on distance from one end, or a localized defect (i.e., an isolated domain in a QPM grating), will give rise to different throughputs when the waveguide is excited from opposite sides, up to complete FF depletion in one direction, and complete transmission in the opposite one. This can be explained by the fact that the output depends on the evolution and interaction of the FF field with its harmonic. The amount and phase of the latter, however, do depend on the details of the propagation path and can substantially change the FF throughput as in the case of seeded SHG.

The sketch of an all-optical diode in a QPM waveguide with its calculated response is shown in Figure 6a,b. Clearly, total depletion is obtained via SHG in backward propagation, whereas unity FF transmission is recovered by the insertion of a phase discontinuity (and seeded interaction) in the forward path. As demonstrated at 1.55 μm in lithium niobate, different degrees of isolation are available at various excitations, with a large extinction ratio at a given power when launching FF waves from opposite ends.

Quadratic Solitons

When beams propagate in a two- (planar waveguides) or three-dimensional (bulk) medium with a quadratic response, cascading induces a nonlinear phase-front distortion which, coupled to the inherent parametric gain, can effectively counteract diffraction and lead to diffractionless propagation or 'spatial solitons' (Figure 7). In contrast to the Kerr case, such solitons are multifrequency solutions or 'simultons' with field components (FF and SH in the degenerate SHG process) bound together in propagation despite possible walk-off or transverse velocity mismatch. Moreover, since the second-order nonlinearity is inherently saturable, the solitons are stable and robust even in bulk media, and provide 'clean' filtered beam profiles at FF and SH frequencies, as has been demonstrated in KTP. Just like other cascading phenomena, quadratic solitons can be controlled by a weak seed, by polarization imbalance at the input (Type II SHG), by direction of propagation (due to birefringence), by phase mismatch, etc. Employing quadratic spatial solitons, phenomena such as collisions, soliton fusion, spatial reshaping, and transverse light localization and instabilities have been studied and constitute an active field of investigation.

When employing pulses, the cascaded phase shift can balance dispersion (material, chromatic, or photonics bandgap) leading to temporal simultons with frequency components locked together despite their individual group-velocity differences and dispersion. This feature can also substantially broaden the bandwidth available for parametric generation. Finally, in analogy to the cubic case, a combination of space and time effects is expected to lead to the formation of 'light bullets' in (3 + 1) dimensions.

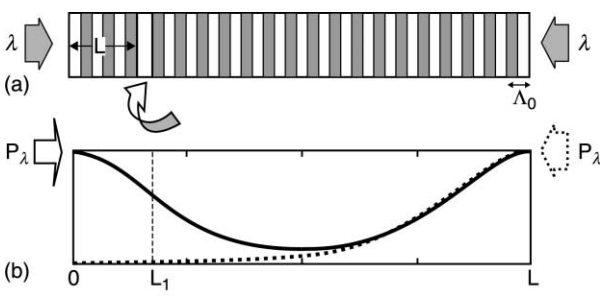


Figure 6 All-optical diode: (a) sketch of implementation with an engineered 'defect' in a QPM grating (*z*-cut periodically poled lithium niobate), and (b) example of FF transmittance *T* versus propagation for forward (solid line) and backward (dashed line) FF excitation of power P_λ.

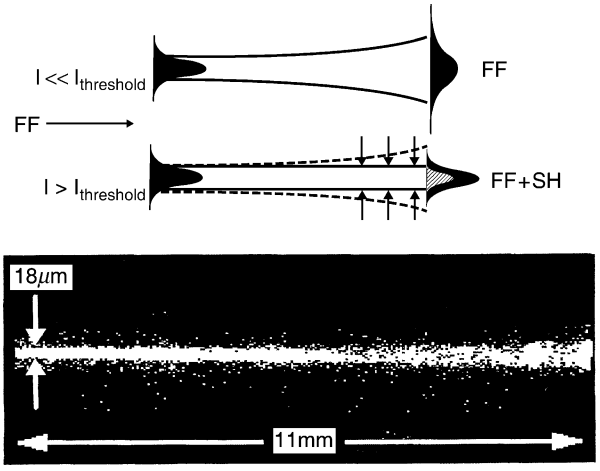


Figure 7 Top: sketch of a quadratic spatial soliton excited by an FF input: when the irradiance is high enough, linear diffraction is overcome by parametric self-trapping and a two-color simulton is obtained. Bottom: sample photograph (top view) of a quadratic soliton experimentally observed in a potassium niobate crystal excited by a 1064-nm laser beam.

List of Units and Nomenclature

Angular frequency	$[s^{-1}]$	ω
Dielectric constant of vacuum	$[FV^{-1}\,m^{-1}]$	ε_0

Electric field	[V m^{-1}]	E
Irradiance, intensity	[W m^{-2}]	I
Length	[m]	L
Refractive index	[ ]	n
Speed of light in vacuum	[m s^{-1}]	c
Second-order electronic susceptibility	[m V^{-1}]	$\chi^{(2)}$
Third-order electronic susceptibility	[m^2 V^{-2}]	$\chi^{(3)}$
Wavevector in medium	[m^{-1}]	$k = n\omega/c$
Wavevector in vacuum	[m^{-1}]	$k_0 = \omega/c$
Wavelength	[m]	$\lambda = 2\pi n/k = 2\pi/k_0$

See also

Solitons: Bright Spatial Solitons; Soliton Communication Systems. **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Assanto G (1995) All-optical switching in integrated structures. In: DeGiorgio V and Flytzanis C (eds) *Nonlinear Optical Materials: Principles and Applications*. Amsterdam: IOS Press.

Assanto G and Stegeman GI (2002) Simple physics of quadratic spatial solitons. *Optics Express* 10: 388–396.

Assanto G, Stegeman GI, Sheik-Bahae M and VanStryland E (1993) All optical switching devices based on large nonlinear phase shifts from second harmonic generation. *Applied Physics Letters* 62: 1323–1325.

Assanto G, Stegeman GI, Sheik-Bahae M and VanStryland EW (1995) Coherent interactions for all-optical signal processing via quadratic nonlinearities. *IEEE Journal of Quantum Electron* 31: 673.

Assanto G, Stegeman GI and Schiek R (1998) Thin film devices for all-optical switching and processing via quadratic nonlinearities. *Thin Solid Films* 331: 291–297.

Boardman AD (ed.) (1999) *Advanced Photonics with Second-Order Optically Nonlinear Processes*. Dordrecht, The Netherlands: Kluwer.

Boardman AD (ed.) (2001) *Soliton Driven Photonics*. Dordrecht, The Netherlands: Kluwer.

Gallo K and Assanto G (1999) Analysis of lithium niobate all-optical wavelength shifters for the third spectral window. *Journal of the Optical Society of America B* 16: 741–753.

Gallo K, Assanto G, Parameswaran KR and Fejer MM (2001) All-optical diode in a periodically poled lithium niobate waveguide. *Applied Physics Letters* 79: 314–316.

Kajzar F and Reinisch R (eds) (1997) *Beam Shaping and Control with Nonlinear Optics*. New York: Plenum Press.

Stegeman GI, Hagan DJ and Torner L (1996) $\chi^{(2)}$ cascading phenomena and their applications to all-optical signal processing, mode-locking, pulse compression and solitons. *Journal of Optics and Quantum Electronics* 28: 1691.

Stegeman GI and Assanto G (1999) Nonlinear integrated optical devices. In: Murphy EJ (ed.) *Design and Application of Integrated Optical Circuits and Components*. New York: Marcel Dekker.

$\chi^{(2)}$–Harmonic Generation

R C Eckardt, Cleveland Crystals, Inc., Highland Heights, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical harmonic generation is nonlinear conversion of laser output to light, of a frequency that is an integer multiple of that of the fundamental. It is used over a wide range of applications, from battery-operated, pencil-sized pointers emitting less than a milliwatt of light to installations that occupy large buildings and generate thousands of Joules of ultraviolet energy in nanosecond-duration pulses. The applications also range from frequency-stable continuous-wave light to generation of pulses a few femtoseconds in duration. Laser output with specific properties may only be available in narrow spectral regions, and harmonic generation is a method to convert the laser output to spectral regions of shorter wavelength. It is customary to call light of twice the frequency of the fundamental the second harmonic; three times the third harmonic, and so on.

The conversion of light by harmonic generation is a process that is driven by the nonlinear electric polarization. The total polarization $\mathbf{P}(t)$, including the linear term, is a vector quantity related to the electric field $\mathbf{E}(t)$ by the susceptibility tensors

χ of a material:

$$\mathbf{P}(t) = \varepsilon_0 \chi^{(1)} : \mathbf{E}(t) + \varepsilon_0 \chi^{(2)} : \mathbf{E}(t)\mathbf{E}(t) + \varepsilon_0 \chi^{(3)} : \mathbf{E}(t)\mathbf{E}(t)\mathbf{E}(t) + \ldots, \quad [1]$$

where $\varepsilon_0 \cong 8.854 \times 10^{-12}$ C/N·m^2 is the permittivity of free space. The first-order tensor describes linear optical properties on the material. The second-order tensor $\chi^{(2)}$ describes second-harmonic generation and sum- and difference-frequency generation. This is the term that is of interest here. The second-order nonlinear optical tensor, equal to half the second-order susceptibility, is used widely instead of the susceptibility:

$$d_{ijk} = \chi^{(2)}_{ijk}/2 \quad [2]$$

In calculations of harmonic generation and sum- and difference-frequency generation, the coefficient is usually reduced to a scalar quantity d_{eff}, the effective nonlinear coefficient that expresses the relationship between only the specific vector components involved in an interaction. The second-order optical nonlinearity, when present, is usually significantly larger than higher-order optical nonlinearities. When optical frequency conversion is possible, using the second-order nonlinearity, it is generally preferred. Any material that has inversion symmetry, however, will have vanishing second-order nonlinearity. Although there is continuing interest in developing other materials, such as poled polymers for future applications, materials for current practical applications are restricted to noncentrosymmetric crystals.

This article will provide a description of the process of optical harmonic generation and discuss some of the concerns in the design of harmonic generation systems. In many cases, harmonic generation can be modeled accurately by analytic or numerical techniques. Some of the theoretical development that is the basis for this modeling will be presented along with examples. There are several essential considerations that must be made in selecting a harmonic generation process. These involve both material properties and conditioning of the fundamental radiation.

General Considerations

The nonlinear optical material must be highly transmitting at the fundamental and harmonic wavelengths and must have a laser-induced-damage threshold high enough to allow optical intensities that will provide adequate conversion efficiency. Phase matching must be possible in the material. Other material characteristics must be considered for specific applications. The harmonic generation process requires that the phase velocities of the fundamental and harmonic radiation be matched. Without phase matching, power converted to the harmonic is back converted to the fundamental in a rapid oscillation, and harmonic conversion remains at low levels. The material also needs to be of good optical quality to provide optical transmission without distortion, which would destroy the phase relationship. Continuous-wave harmonic generation may involve local intensities of the order of 1 megawatt per square cm (1 MW/cm^2). Pulsed applications with Q-switched laser output of a few nanoseconds duration typically employ intensities of several hundred MW/cm^2, and harmonic-generation applications with mode-locked lasers may involve pulse durations of tens of femtoseconds with peak intensities of tens of gigawatts per square cm. Many other properties are desirable, but it is necessary to work within the limitations of available materials.

Phase matching is an essential requirement. When harmonic and fundamental waves of different eigen polarizations are coupled through the nonlinear optical tensor, it may be possible to use crystal birefringence to provide phase matching. The birefringence, if large enough, can compensate dispersion. The fundamental wave is introduced into the crystal with the polarization of the larger refractive index, and the harmonic is generated in the polarization of the smaller index. This is a description of type-I birefringent phase matching. Phase matching is achieved by selecting the direction of propagation in the crystal or by changing the temperature of the crystal. Type-II phase matching with the fundamental wave resolved into the two eigen polarizations is possible with large birefringence. Quasi phase matching involves periodically structuring the nonlinear material in a way that spatial modulation of the optical nonlinearity compensates for dispersion. Quasi phase matching is tuned by selection of the modulation period, and temperature control may provide a fine tuning. Periodic poling, one method of quasi phase matching, involves changing the orientation of ferroelectric domains after each coherence length of the harmonic generation process. These methods of phase matching are sketched in Figure 1.

The physical arrangement of the harmonic generation process depends on properties of the available fundamental light and the properties of the selected nonlinear material. The single-pass techniques, shown in Figure 1, may have the advantage of simplicity, but expanding or focusing the fundamental beam may be required. Crystal length and clear

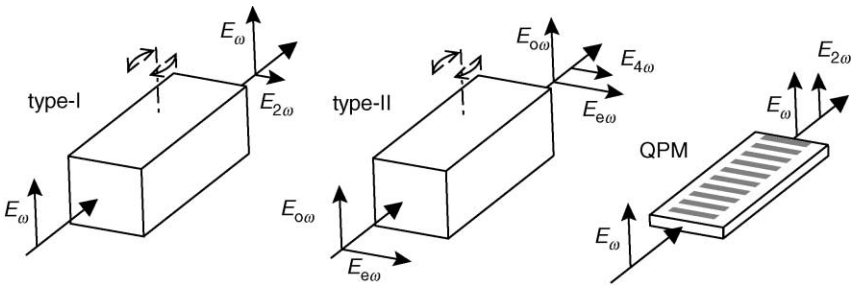


Figure 1 Different methods of phase matching optical harmonic generation are illustrated. The arrows indicate the directions of electric-field polarization and directions of propagation.

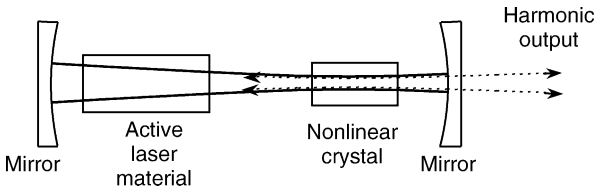


Figure 2 Placing a harmonic-generation crystal inside a laser cavity with mirrors highly reflecting at the laser wavelength is one technique of increasing the intensity of the fundamental beam.

aperture need to be selected. Intracavity harmonic generation (Figure 2), or external-resonant-cavity harmonic generation, may provide advantages in other cases. Analysis of harmonic generation provides the means to select among the options.

Basic Equations

Much of the analysis of second-harmonic generation follows from the differential equations that describe the interaction between the fundamental and harmonic fields, and it is appropriate to describe the development of these coupled equations. The wave equation:

$$\nabla^2\mathbf{E}(\mathbf{r},t) - \mu_0\sigma\frac{\partial\mathbf{E}(\mathbf{r},t)}{\partial t} - \mu_0\varepsilon_0\frac{\partial^2\mathbf{E}(\mathbf{r},t)}{\partial t^2} = \mu_0\frac{\partial^2\mathbf{P}(\mathbf{r},t)}{\partial t^2} \quad [3]$$

is obtained from Maxwell's equations. Here $\mu_0 = 4\pi\times10^{-7}$ is the permeability of free space, and σ is the conductivity, which later will be incorporated into the optical absorption coefficient $\alpha = \mu_0\sigma c/2$, and $(\varepsilon_0\mu_0)^{-1/2} = c \cong 2.998\times10^8$ m/s is the speed of light in vacuum. The electric field of a planewave, at angular frequency ω, is expressed in complex notation as the product of a slowly varying complex amplitude and a carrier wave:

$$\mathbf{E}_\omega(\mathbf{r},t) = (1/2)\{\mathbf{E}_\omega \exp i(\mathbf{k}_\omega\cdot\mathbf{r} - \omega t) + \text{c.c.}\} \quad [4]$$

The intensity of this plane wave is

$$I_\omega = (nc\varepsilon_0/2)|E_\omega|^2 \quad [5]$$

Here n is the index of refraction of the material in which E is specified. The nonlinear polarization resulting from the second-order susceptibility driven by the plane wave $\mathbf{E}_\omega(\mathbf{r},t)$ is expressed as

$$\mathbf{P}^{\text{NL}}_{2\omega}(\mathbf{r},t) = (1/2)\{\mathbf{P}^{\text{NL}}_{2\omega} \exp i(2\mathbf{k}_\omega\cdot\mathbf{r} - 2\omega t) + \text{c.c.}\} \quad [6]$$

The individual components of the polarization amplitude vector are

$$P^{\text{NL}}_{2\omega,i} = \varepsilon_0\sum_{j,k} d_{ijk}(-2\omega;\omega,\omega)E_{\omega,j}E_{\omega,k} \quad [7]$$

If the nonlinear polarization is produced by two distinguishable waves, such as for sum-frequency generation of the third harmonic by the second harmonic and the fundamental, the nonlinear polarization is

$$\mathbf{P}^{\text{NL}}_3(\mathbf{r},t) = (1/2)\{\mathbf{P}^{\text{NL}}_3 \exp i[(\mathbf{k}_1+\mathbf{k}_2)\cdot\mathbf{r} - (\omega_1+\omega_2)t] + \text{c.c.}\} \quad [8]$$

where

$$P^{\text{NL}}_{3,i} = 2\varepsilon_0\sum_{j,k} d_{ijk}(-\omega_3;\omega_1,\omega_2)E_{1,j}E_{2,k} \quad [9]$$

and

$$\omega_3 = \omega_1 + \omega_2 \quad [10]$$

A number of simplifications are used to arrive at the coupled equations for harmonic generation with monochromatic plane waves. All waves are assumed to be propagating collinearly in the z direction. Only the vector components involved in the

interaction are retained, and the nonlinear optical tensor is replaced by the scalar effective nonlinear coefficient d_{eff}. Finally, the slowly varying envelope approximation is invoked and the higher-order derivatives of the amplitudes are ignored, as are time derivatives of the envelopes. The wave equation simplifies to

$$\frac{\mathrm{d}E(z)}{\mathrm{d}z} + \alpha E(z) = \mathrm{i}\frac{\mu_0 c\omega}{2n} P(z) \qquad [11]$$

where n is the index of refraction.

With three distinguishable waves, as with sum-frequency generation and type-II second-harmonic generation, there are three coupled equations:

$$\frac{\mathrm{d}E_3(z)}{\mathrm{d}z} + \alpha_3 E_3(z) = \mathrm{i}\frac{\omega_3 d_{\text{eff}}}{n_3 c} E_1(z)E_2(z)\exp(-\mathrm{i}\Delta k z) \qquad [12]$$

$$\frac{\mathrm{d}E_1(z)}{\mathrm{d}z} + \alpha_1 E_1(z) = \mathrm{i}\frac{\omega_1 d_{\text{eff}}}{n_1 c} E_3(z)E_2^*(z)\exp(\mathrm{i}\Delta k z) \qquad [13]$$

$$\frac{\mathrm{d}E_2(z)}{\mathrm{d}z} + \alpha_2 E_2(z) = \mathrm{i}\frac{\omega_2 d_{\text{eff}}}{n_2 c} E_3(z)E_1^*(z)\exp(\mathrm{i}\Delta k z) \qquad [14]$$

Here the wave vector mismatch is

$$\Delta k = k_3 - k_1 - k_2 \qquad [15]$$

and

$$\omega_3 = \omega_1 + \omega_2 \qquad [16]$$

with the convention of $\omega_1 \leq \omega_2 < \omega_3$. The indices of refraction of the three waves are n_1, n_2, and n_3, respectively, and α_1, α_2, and α_3 are the respective amplitude absorption coefficients.

With type-I second-harmonic generation, the fundamental field consists of only a single eigen polarization, and there are two coupled equations:

$$\frac{\mathrm{d}E_{2\omega}(z)}{\mathrm{d}z} + \alpha_{2\omega} E_{2\omega}(z) = \mathrm{i}\frac{\omega d_{\text{eff}}}{n_{2\omega} c} E_\omega(z)E_\omega(z)\exp(-\mathrm{i}\Delta k z) \qquad [17]$$

$$\frac{\mathrm{d}E_\omega(z)}{\mathrm{d}z} + \alpha_\omega E_\omega(z) = \mathrm{i}\frac{\omega d_{\text{eff}}}{n_\omega c} E_\omega(z)E_\omega^*(z)\exp(\mathrm{i}\Delta k z) \qquad [18]$$

Indices of refraction at the harmonic and fundamental are $n_{2\omega}$ and n_ω respectively; $\alpha_{2\omega}$ and α_ω are the respective absorption coefficients; and the wavevector mismatch is

$$\Delta k = k_{2\omega} - 2k_\omega \qquad [19]$$

The equations for type-II second-harmonic generation, where $\omega_1 = \omega_2$ and $\omega_3 = 2\omega_1$, become equivalent for those for type-I second-harmonic generation, providing intensities of the two fundamental polarizations are equal, $I_1 = I_2$ and $\alpha_\omega = \alpha_{2\omega}$.

Steady-State Solutions

Three monochromatic planewave solutions for type-I second-harmonic generation are discussed. These solutions address type-I harmonic generation; first for low levels of conversion with arbitrary phase mismatch; next for high levels of conversion with perfect phase matching; and last for arbitrary levels of conversion with arbitrary amounts of phase mismatch. Near-field and focused-beam approximations also exist. Beyond that, numerical techniques, such as split-step Fourier transform methods can handle more complicated conditions. The simple approximations of monochromatic planewaves, or steady-state solutions, however, yield useful information.

The coupled equations for type-I second-harmonic generation can be solved by simple integration for low levels of fundamental depletion, where $E_\omega(z)$ is treated as constant and there is insignificant absorption. Initial conditions are $I_\omega(0)$, incident on the crystal and $I_{2\omega}(0) = 0$. Integration of the equation for the harmonic electric field yields an expression for electric field after propagating a distance of l in the crystal:

$$E_{2\omega}(l) = \frac{\omega d_{\text{eff}} E_\omega^2(0)}{n_{2\omega} c}\frac{1}{\Delta k}\left\{\exp\mathrm{i}(k_{2\omega} l) - \exp\mathrm{i}(2k_\omega l)\right\} \qquad [20]$$

Expressed as intensity the solution is

$$I_{2\omega}(l) = I_\omega(0)\Gamma^2 l^2 \,\mathrm{sinc}^2(\Delta k l/2) \qquad [21]$$

where $\mathrm{sinc}(x) = \sin(x)/x$ and

$$\Gamma^2 = 2\omega^2 d_{\text{eff}}^2 I_\omega(0)/(n_{2\omega} n_\omega^2 c^3 \varepsilon_0) \qquad [22]$$

When $\Delta k = 0$ and $\alpha_\omega = \alpha_{2\omega} = 0$, conditions of perfect phase matching and no absorption, the coupled equations for type-I second-harmonic generation can be combined into a single equation. This is done using conservation of energy, which requires $E_\omega^2(z) + E_{2\omega}^2(z) = E_\omega^2(0)$, in this case with no initial harmonic intensity. The equation can then be manipulated into a standard integral that yields the

inverse hyperbolic tangent. Rewritten and expressed in terms of intensity the result is

$$I_{2\omega}(l) = I_{\omega}(0)\tanh^2(\Gamma l) \qquad [23]$$

The two solutions for $I_{2\omega}(l)$ given above serve well for a number of applications. The $\mathrm{sinc}^2(\Delta k \cdot l/2)$ function is widely used to determine phase-matching acceptances. The $\tanh^2(\Gamma l)$ function is a good approximation for local conversion with large, well-collimated fundamental beams.

An exact solution to the coupled equations, however, gives insight to conditions such as high initial fundamental intensity with small amounts of phase mismatch. After extensive algebraic manipulation, a solution in the form of a Jacobian elliptic function $\mathrm{sn}(u|m)$ is obtained:

$$I_{2\omega}(l) = I_{\omega}(0)\nu_{\mathrm{b}}^2 \mathrm{sn}^2\left(\frac{\Gamma l}{\nu_{\mathrm{b}}}\middle|\nu_{\mathrm{b}}^4\right) \qquad [24]$$

where

$$1/\nu_{\mathrm{b}} = \Delta s/4 + \sqrt{1 + (\Delta s/4)^2} \qquad [25]$$

$$\Delta s = \Delta k/\Gamma$$

and Δk and Γ retain the earlier definitions. This solution can be extended to provide the phase of the fundamental and harmonic as the conversion process evolves. Expressing the complex harmonic and fundamental field amplitudes as the product of a real amplitude and a complex exponential:

$$E(z) = A(z)\exp \mathrm{i}\phi(z) \qquad [26]$$

the harmonic phase is

$$\phi_{2\omega}(z) = (\pi - \Delta k z)/2 \qquad [27]$$

The fundamental phase, however, is found to be dependent on the conversion process:

$$\phi_{\omega}(z) = (1/2)\arccos[A_{2\omega}(z)\Delta k/(2\kappa A_{\omega}^2(z))] - (\pi - \Delta k z)/z \qquad [28]$$

where $\kappa = \omega d_{\mathrm{eff}}/(n_{\omega} c)$. Small amounts of phase mismatch significantly change the phase of the fundamental at high levels of conversion. This phase change is used for switching applications in waveguide interferometers. It also makes phase-matching requirements more stringent at high levels of conversion. The three planewave solutions are illustrated in Figure 3. The narrowing of phase matching at higher conversion is illustrated in Figure 4.

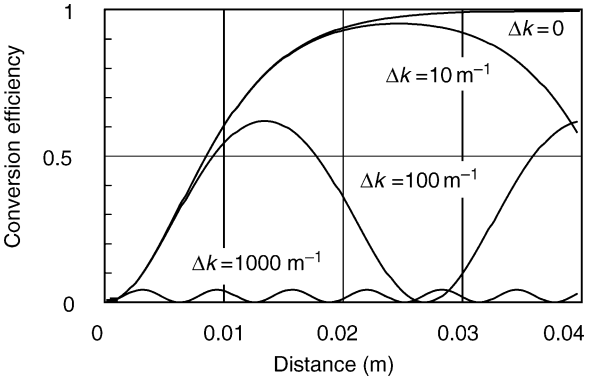


Figure 3 Monochromatic plane-wave solutions of the coupled equations for second-harmonic generation are illustrated for different amounts of wavevector mismatch. The solution takes the form $\tanh^2(\Gamma l)$ when $\Delta k = 0$ and $(\Gamma l)^2\mathrm{sinc}^2(\Delta k l/2)$ when conversion is small.

There is a Jacobian elliptic function solution for intensities with the coupled waves of type-II second-harmonic generation and sum-frequency generation. The phases of the individual waves can be recovered with elliptic integrals of the third kind. Both sets of equations, the two coupled equations for type-I phase matching and the three coupled equations for type-II phase matching, are amenable to numerical solution by Runge–Kutta methods.

Phase Matching

Phase matching is achieved in birefringent materials by balancing the dispersion of the material with the difference between the indices of refraction of the fast and slow waves. With adequate birefringence it may be possible to obtain phase matching by selecting the direction of propagation in the crystal. Equations describing the dispersion and birefringence of nonlinear crystals, necessary for calculating phase matching, along with the nonlinear optical coefficients and expressions for the effective nonlinear optical coefficients, are tabulated in handbooks. Values for a few examples are given here.

The negative uniaxial crystal potassium dihydrogen phosphate (KH_2PO_4 or KDP) is used as an illustrative example. The harmonic wave is chosen to have extraordinary polarization with the smaller index of refraction in this crystal with negative birefringence. With type-I phase matching, the fundamental consists of only an ordinary polarized wave. To have phase matching, $\Delta k = 0$, it is necessary to choose a direction of propagation that makes an angle θ with the optic axis that satisfies:

$$n_{\mathrm{e},2\omega}(\theta) = n_{\mathrm{o},\omega} \qquad [29]$$

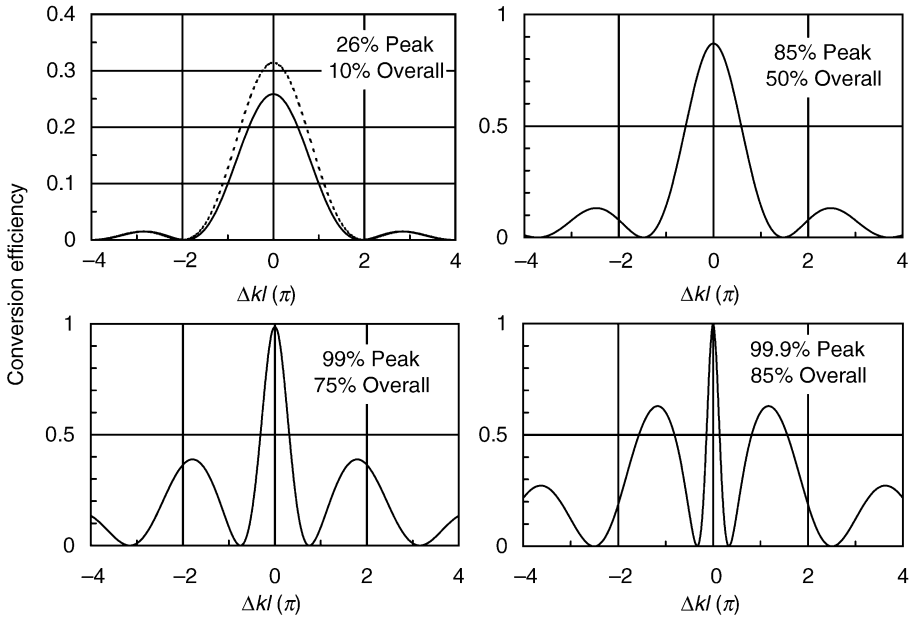


Figure 4 High levels of overall second-harmonic conversion require higher levels of conversion at peak pulse intensity. Phase-matching bandwidths narrow at the higher levels of peak conversion. The broken curve is the $(\Gamma l)^2\text{sinc}^2(\Delta kl/2)$ small conversion approximation.

The extraordinary index for propagation at an angle of θ to the optical axis in a uniaxial crystal $n_e(\theta)$ is given by

$$\frac{1}{n_e^2(\theta)} = \frac{\cos^2\theta}{n_o^2} + \frac{\sin^2\theta}{n_e^2} \qquad [30]$$

where n_o is the ordinary index and n_e is the extraordinary index for propagation at 90° to the optic axis. The phase-matching angle is obtained by using the identity $\cos^2\theta = 1 - \sin^2\theta$ and solving for θ:

$$\theta_{PM} = \sin^{-1}\left\{\frac{n_{e,2\omega}^2(n_{o,2\omega}^2 - n_{o,\omega}^2)}{n_{o,\omega}^2(n_{o,2\omega}^2 - n_{e,\omega}^2)}\right\}^{1/2} \qquad [31]$$

where the subscripts refer to ordinary and extraordinary indices at the fundamental and second-harmonic frequencies.

The effective nonlinear coefficient is determined by crystal symmetry, the type of phase matching, the birefringence of the crystal, and the direction of propagation. For crystals of point group $\bar{4}2m$ with negative birefringence, such as KDP, the effective nonlinear coefficient for type-I phase matching is given by

$$d_{eff} = -d_{36}\sin\theta\sin 2\phi \qquad [32]$$

The azimuthal angle ϕ is measured from the crystal x-axis. The x-axis and y-axis are specified by X-ray diffraction and piezoelectric coefficients and can be determined from the growth morphology. The azimuthal orientation of a finished crystal, however, is not obvious, and requires X-ray orientation, piezoelectric characterization, or nonlinear optical measurements, if the record of crystal orientation is lost. The nonlinear coefficient d_{36} is an abbreviation for the d_{312} and d_{321} tensor components, which are equal. For harmonic generation with a 1064 nm fundamental in KDP, $d_{36} = 0.39 \times 10^{-12}$ m/V. The unit of 10^{-12} m/V is usually expressed as picometers per volt (pm/V). The effective nonlinear coefficient for type-II phase matching for this class of crystals is

$$d_{eff} = d_{36}\sin 2\theta\cos 2\phi \qquad [33]$$

Type-I and type-II crystals cut from the same material are not interchangeable. The type-II phase-matching angle for KDP can be found by iterative solution of the equation:

$$n_{e,2\omega}(\theta) = \{n_{o,\omega} + n_{e,\omega}(\theta)\}/2 \qquad [34]$$

The birefringent walkoff angle, that is the angle the Poynting vector of an extraordinary wave makes with the wavevector, is another consideration for birefringent phase matching. There will be no further harmonic generation when the extraordinary and ordinary fundamental beams of type-II phase matching no longer overlap. For type-I phase matching, birefringent walkoff can limit the interaction length, but there can be several interaction lengths

in a crystal. The birefringent walkoff of an extraordinary wave propagating at an angle θ to the optic axis in a uniaxial crystal is given by

$$\rho = \tan^{-1}\left\{\frac{n_e^2(\theta)}{2}\left[\frac{1}{n_e^2} - \frac{1}{n_o^2}\right]\sin(2\theta)\right\} \qquad [35]$$

Refractive indices for 1064 nm to 532 nm second-harmonic generation in KDP are $n_{e,2\omega} = 1.4709$, $n_{o,2\omega} = 1.5129, n_{e,\omega} = 1.4603$, and $n_{o,\omega} = 1.4944$. The dispersion of KDP and indices at the phase-matching angles are shown in Figure 5. For type-I phase matching $d_{\text{eff}} = 0.26$ pm/V, $\theta_{\text{PM}} = 41.2°$, and $\rho_{2\omega} = 28$ milliradians (mr), and for type-II phase matching $d_{\text{eff}} = 0.34$ pm/V, $\theta_{\text{PM}} = 59.1°, \rho_{\omega} = 20$ mr, and $\rho_{2\omega} = 25$ mr. Type-II phase matching in KDP for this application offers an advantage of larger effective nonlinear coefficient. Also the angular acceptance is larger for type-II phase matching.

The angular acceptance is calculated from dispersion equations by first evaluating the derivative $\partial\Delta k/\partial\theta$ and then finding the value of $\Delta\theta$ corresponding to the full width at half maximum (FWHM) of the $\text{sinc}^2(\Delta k \cdot l/2)$ function:

$$\frac{\partial\Delta k}{\partial\theta}\Delta\theta_{\text{FWHM}} \cong 4 \times 1.39 \text{ radians}/l \qquad [36]$$

When θ is 0° or 90° for uniaxial crystals, or for propagation along one of the refractive index principal axes in a biaxial crystals, the first derivative $\partial\Delta k/\partial\theta$ becomes zero, and the angular acceptance must be calculated from the second derivative $\partial^2\Delta k/\partial\theta^2$. This case is called noncritical phase matching (NCPM), and the case for θ between 0° and 90° is called critical phase matching.

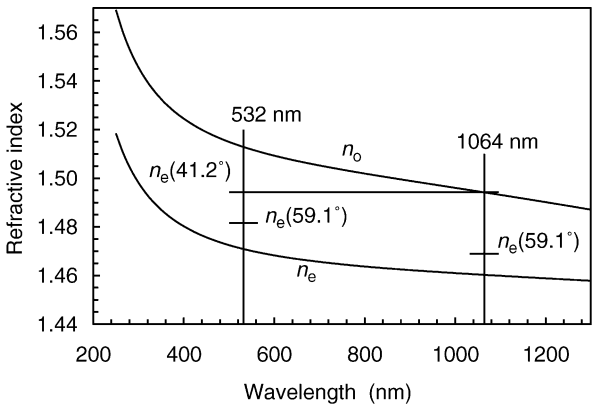


Figure 5 Dispersion of the birefringent nonlinear material KH_2PO_4 with refractive indices indicated for type-I and type-II 1064 nm to 532 nm second-harmonic generation. (Adapted from Byer RL (1977) Parametric oscillators and nonlinear materials. In: Harper PG and Wherrett BS (eds) *Nonlinear Optics*, pp. 47–159. San Francisco, CA: Academic Press.)

Second-harmonic generation of 1064 nm fundamental radiation in the crystal lithium triborate (LBO or LiB_3O_5) is an example of NCPM. At 149°C, this crystal will phase match for type-I SHG for propagation along the x-axis in the xyz frame determined by the indices of refraction of this biaxial crystal $n_x < n_y < n_z$. The following are calculated from dispersion equations for this case: $\partial\Delta k/\partial\theta = 0$; $\partial^2\Delta k/\partial\theta^2 = 2.04 \times 10^4$ (rad/cm)/rad^2; $\partial\Delta k/\partial\phi = 0$; and $\partial^2\Delta k/\partial\phi^2 = 6.45 \times 10^3$ (rad/cm)/rad^2. The FWHM angular acceptances for a 1.5 cm-long LBO crystal are $\Delta\theta_{\text{FWHM}} = 1.09°$ and $\Delta\phi_{\text{FWHM}} = 1.94°$. Angle-tuned CPM at 23°C in LBO occurs at $\theta = 90°$ and $\phi = 11°.5$. For this case $\partial\Delta k/\partial\phi = 1.32 \times 10^4$ (rad/cm)/rad, and the azimuthal angular acceptance decreases to $\Delta\phi_{\text{FWHM}} = 0.16°$. Also there is a birefringent walkoff angle of $\rho_{2\omega} = 19$ mr, which results in a displacement of 0.29 mm of the extraordinary-like harmonic beam for a 1.5 cm-long crystal. Both the larger angular acceptance and the absence of birefringent walkoff are advantages for harmonic generation. The effective nonlinear coefficient of LBO is reported as 0.85 pm/V and 1.04 pm/V by different sources.

Quasi phase matching offers the possibility, in principle, of noncritical phase matching at any wavelength in the transmission range of a material. However, it may be difficult producing the periodic structure that is required for phase matching in a particular application. The ferroelectric crystals lithium niobate ($LiNbO_3$), lithium tantalate ($LiTaO_3$), and potassium titanyl phosphate ($KTiOPO_4$ or KTP) can be processed by electric-field poling to produce domains of alternating polarity. Lithography techniques are used to apply a patterned electrode structure on the surface of a wafer or thin slab. Domains are reversed by the application of a carefully monitored and controlled electric field. Thicknesses of several mm have been achieved in periodically poled KTP (PPKTP). The higher coercive field of 21 kV/mm has limited the thicknesses that can be achieved in periodically poled lithium niobate (PPLN), particularly at shorter wavelengths for which poling periods become small. Progress is now being made with zincblende materials such as gallium arsenide. These materials have large linearity and good transmission in the infrared, but they are not birefringent. Epitaxial growth on patterned substrates is now being used to produce periodically structured GaAs on a research basis, and bonded stacks of wafers of alternating orientation also have been used. Harmonic conversion with quasi phase matching is compared with conventional phase matching and nonphase-matched harmonic generation in Figure 6.

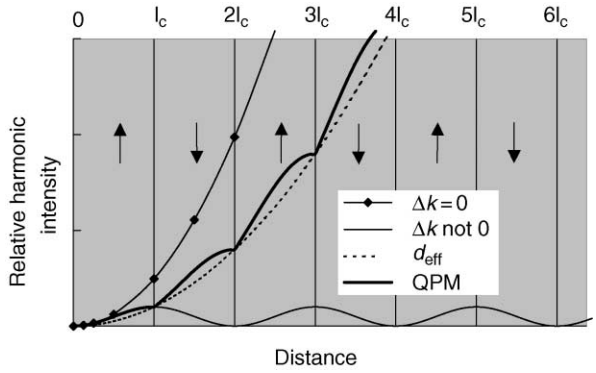


Figure 6 The growth of harmonic intensity is illustrated schematically for a quasi-phase-matched process in which the polarity of the nonlinearity is reversed after each coherence length. Conversion for $\Delta k = 0$, $\Delta k \neq 0$, and a uniform effective nonlinear coefficient are shown for comparison. (Adapted from Fejer MM (1992) Quasi-phase-matched second harmonic generation: tuning and tolerances. *IEEE Journal of Quantum Electronics* 28: 2631–2654.)

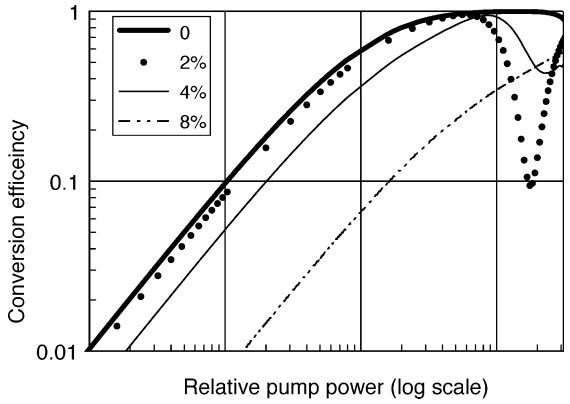


Figure 7 Random variations in domain length will lower harmonic conversion efficiency and have effect on back conversion. RMS variations in domain length are indicated in the legend for this specific example.

Quasi phase matching also provides access to tensor components that are not accessible with birefringent phase matching. This is the case for lithium niobate; the largest tensor component is $d_{333} = d_{33}$, which has a value of 25 pm/V. The effective nonlinear coefficient is the relevant Fourier component of the modulated nonlinearity:

$$d_{\text{eff}} = (2/\pi)d_{33} \approx 16 \text{ pm/V} \qquad [37]$$

This is a larger effective nonlinearity than is available with common birefringently phase matched materials in the visible and near-infrared. A more accurate treatment of second-harmonic generation involves application of the Jacobian-elliptic-function solution to each domain. After nearly complete conversion is reached, the phase perturbations on the individual domains have effect, and back conversion takes over (Figure 7). Conversion is lower and back conversion can set in sooner with random differences in the domain length. There is some experimental evidence that this type of back conversion has been observed in guided-wave harmonic conversion devices. Guided-wave nonlinear optical interaction can have long interaction lengths not limited by diffraction. Quasi phase matching is particularly useful in waveguide second-harmonic generation because the direction of propagation is fixed.

Gaussian Beams and Pulses

A radially symmetric Gaussian light beam has transverse intensity of

$$I(r) = I_0 \exp(-2r^2/w^2) \qquad [38]$$

where r is the radial coordinate and w is called the spot size. Harmonic generation in the low-conversion limit by a Gaussian beam, with a focus of spot size w_0 at the center of a crystal with optimized phase matching, is described by

$$p_{2\omega}(l)/p_{\omega}(0) = 2\omega^2 d_{\text{eff}}^2 p_{\omega}(0) l k h(B, \xi)/(\pi n^3 \varepsilon_0 c^3) \qquad [39]$$

The harmonic power after propagating distance l in the crystal is $p_{2\omega}(l)$, and $p_{\omega}(0)$ is the incident fundamental power. The function $h(B, \xi)$ is the Boyd and Kleinman focusing factor. It is a function of the walkoff parameter:

$$B = \rho\sqrt{lk}/2 \qquad [40]$$

and the focusing parameter:

$$\xi = l/(kw_0^2) \qquad [41]$$

The spot size w_0 of the fundamental beam at its waist is used to described the fundamental intensity distribution at the focus at the center of the crystal:

$$I_{\omega}(r) = (2p_{\omega}(0)/(\pi w_0^2))\exp(-2r^2/w_0^2) \qquad [42]$$

Values of the focusing factor for arbitrary focusing and walkoff are usually displayed in the form of a graph (Figure 8). In the case of weak focusing $\xi \ll 1$ and relatively small walkoff $w_0 > \rho l$, the focusing factor is given by

$$h(B, \xi) \approx \xi(1 - t^2/12 + t^4/120 - t^6/1344 + \ldots) \qquad [43]$$

where $t = 2B(2\xi)^{1/2}$.

An example of continuous-wave second harmonic generation is used. This calculation models a conversion experiment published in the literature. External-resonant-cavity second-harmonic generation of 6.5 W of 532 nm radiation was reported. The 1064 nm fundamental radiation with 378 W of circulating power in a resonant cavity external to the laser was focused to a 32 μm spot size in the center of a 6 mm-long LBO crystal (Figure 9). This corresponds to a focusing factor of $\xi = 0.62$ in the index $n = 1.6053$ material. The value of the focusing parameter is $h(0, 0.62) = 0.58$. The nonlinear optical coefficient for this process is reported as 0.85 pm/V and 1.04 pm/V by different sources. When these values are inserted into the earlier equation for $p_{2\omega}(l)/p_{\omega}(0)$, the higher coefficient yields a conversion efficiency of 0.0273 and 10.3 W of 532 nm output, whereas the lower coefficient yields a conversion efficiency of 0.0182 and 6.9 W of 532 nm radiation. Differences in published values of nonlinear optical coefficients are common. The reported precision of the experiment and the measured values tend to support the lower value. Practical applications are useful to check the accuracy of parameter values. With many variables and the added complication of operation at high intensity there is uncertainty that demands caution in accepting the derived parameter values. It is interesting to note that the power and beam size of the example result in a maximum intensity of 17 MW/cm^2 at the surface of this crystal. Many vendors will warranty antireflection coatings for only 1 MW/cm^2 maximum continuous-wave intensity.

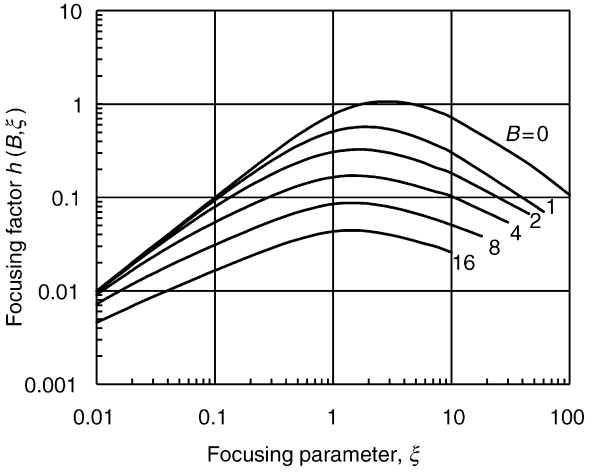


Figure 8 The Boyd and Kleinman focusing factor $h(B,\xi)$ for type-I second-harmonic generation is shown as a function of the focusing parameter ξ for different values of the walkoff parameter B. The curves shown were generated with a split-step Fourier transform calculation. (Adapted from Boyd GD (1968) Parametric interactions of focused Gaussian light beams. *Journal of Applied Physics* 39: 3597–3639.)

For a pulsed fundamental with temporal shape $p_{\omega}(t) = p_{\omega}(0)\exp(-t^2/\tau^2)$, integration over time gives a calculated energy conversion efficiency of:

$$\eta_{\text{calc}} = U_{2\omega}(l)/U_{\omega}(0)$$
$$= \sqrt{2}\omega^2 d_{\text{eff}}^2 U_{\omega}(0) l k h(B,\xi)/(\tau\sqrt{\pi^3} n^3 \varepsilon_0 c^3) \quad [44]$$

where $U_{2\omega}(l)$ is the harmonic energy generated after propagation distance l in the crystal and $U_{\omega}(0)$ is the incident fundamental energy. This is still a low conversion approximation. An estimate of conversion efficient at arbitrary levels of depletion η_{est} is obtained with the empirical expression:

$$\eta_{\text{est}} = \eta_{\text{calc}}/(1 + \eta_{\text{calc}}) \quad [45]$$

Returning to the example 1.064 nm to 532 nm, type-II SHG in KDP and using $l = 3$ cm, $U_{\omega}(0) = 0.3$ J, and $\tau = 3.6$ ns for a 6 ns FWHM pulse duration, the above two relationships predict a conversion efficiency of 50%. These parameters give a peak intensity of 330 MW/cm^2, a value typical for efficient harmonic conversion, while avoiding laser-induced damage in a long-term, repetitive-pulse application using KDP. Single-shot laser-induced-damage thresholds for KDP are about 10 times higher.

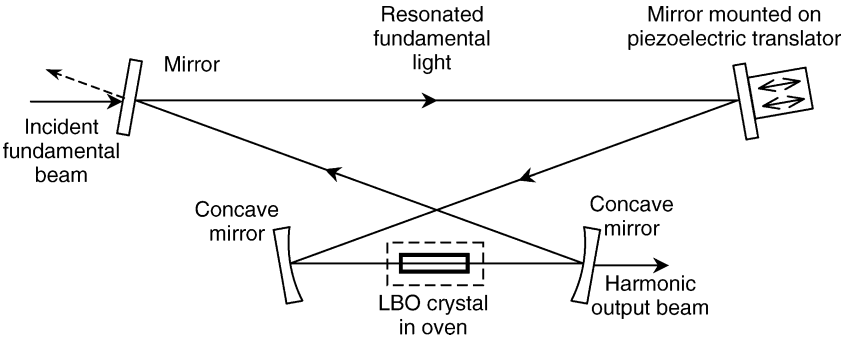


Figure 9 A bow-tie cavity for externally resonant second-harmonic generation is illustrated. The reflected fundamental beam is used to develop an error signal to drive the piezoelectric translator and hold the cavity in resonance.

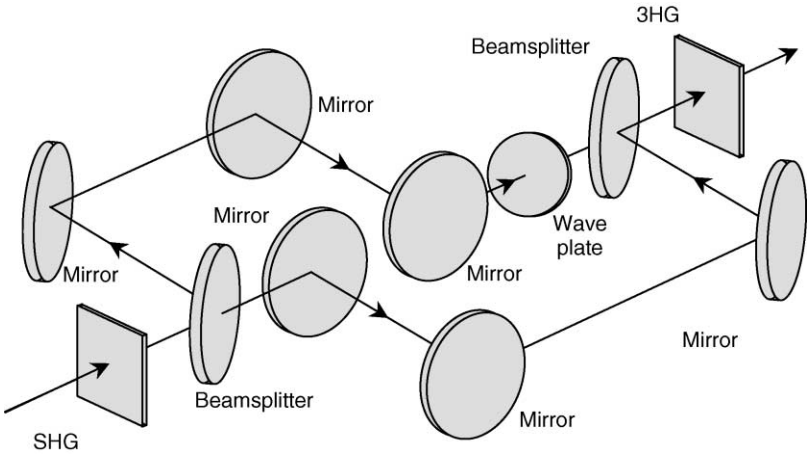


Figure 10 Short crystals less than 1 mm in length are typically used for second-harmonic generation with fundamental pulses that are only ten's of femtoseconds in duration. It is necessary to separate fundamental and second-harmonic pulses and recombine them with appropriate delay to compensate for group-velocity mismatch before sum-frequency generation to the third harmonic.

As a final example, consider second-harmonic generation of a 100 fs ($1\,\text{fs} = 10^{-15}$ s) duration pulse from a mode-locked Ti:Al_2O_3 laser operating at 800 nm. Typical laser output might consist of 10 nJ (10×10^{-9} Joule) pulses at 80 MHz repetition rate. Group velocity walkoff is a concern for this short pulse duration (Figure 10). Barium borate (BaB_2O_4 or BBO) is a common crystal for this application. The group velocity walkoff rate between the ordinary-polarization 800 nm fundamental pulse and the extraordinary-polarization 400 nm harmonic pulse is 190 fs/cm. A crystal length no longer than 0.5 mm is needed to avoid lengthening the harmonic pulse in time. Fortunately, the laser-induced-damage-threshold intensity increases approximately as the inverse of the square root of the pulse length between roughly 10 ns and 1 ps, and it is possible to focus the beam to achieve higher intensity without crystal damage and compensate for the short crystal length. However, the 68-milliradian birefringent walkoff of the harmonic will limit the size of the focus. A 20 μm spot size will allow an interaction length that roughly matches the crystal length and provides a peak fundamental intensity around 16 GW/cm^2 and a conversion efficiency of nearly 30%. Now suppose the 100 fs pulse is amplified to 100 mJ in a large chirped-pulse regenerative amplifier. A larger diameter crystal is required, and KDP is an appropriate choice. The group velocity walkoff rate in KDP is 78 fs/mm, and a crystal of 1.5 mm thickness could be used. With the beam spot size increased to 30 mm, the peak intensity will be 71 GW/cm^2, and the conversion efficiency will be about 40%. Material availability, such as the large sizes possible with KDP, often determines a choice.

Conclusion

Basic techniques for describing and modeling second-harmonic generation have been discussed. Birefringent phase matching and quasi phase matching were reviewed. Three examples were used to illustrate considerations for implementing harmonic-generation systems. There are many more parameters to be considered for other applications and for more detailed analysis. The examples presented illustrate the demands nonlinear optical frequency conversion places on materials. A few materials have been established as primary choices in common applications. But for many applications, available materials or laser system performance are limiting and it is necessary to use the materials and systems that are available. The importance of using the analysis methods is stressed. It is possible to sacrifice performance by improperly specifying a crystal. Choosing the best crystal for a harmonic-generation application with an appropriate optical design will optimize harmonic conversion.

See also

Nonlinear Optics, Applications: Phase Matching. **Nonlinear Sources:** Harmonic Generation in Gases. **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Akhmanov SA (1975) Optical harmonic generation and optical frequency multipliers. In: Rabin H and Tang CL (eds) *Quantum Electronics. A Treatise, vol. 1, part B*, pp. 475–583. New York: Academic Press.

Armstrong JA (1962) Interaction between lightwaves in a nonlinear dielectric. *Physics Review* 127: 1918–1962.
Boyd GD (1968) Parametric interactions of focused Gaussian light beams. *Journal of Applied Physics* 39: 3597–3639.
Boyd RW (2003) *Nonlinear Optics*, 2nd edn. New York: Academic Press.
Byer RL (1977) Parametric oscillators and nonlinear materials. In: Harper PG and Wherrett BS (eds) *Nonlinear Optics*, pp. 47–159. San Francisco, CA: Academic Press.
Dmitriev VG (1997) *Handbook of Nonlinear Optical Crystals*, 2nd edn. Berlin: Springer-Verlag.
Fejer MM (1992) Quasi-phase-matched second harmonic generation: tuning and tolerances. *IEEE Journal of Quantum Electronics* 28: 2631–2654.
Shen YR (1984) *The Principles of Nonlinear Optics*. New York: John Wiley & Sons.
Sutherland RL (2003) *Handbook of Nonlinear Optics*, 2nd edn. New York: Marcel Dekker.
Yang SL (1991) 6.5 W, 532 nm radiation by cw resonant external-cavity second-harmonic generation of an 18 W Nd:YAG laser in LiB_3O_5. *Optical Letters* 16: 1493–1495.
Yariv A (1989) *Quantum Electronics*, 3rd edn. New York: John Wiley & Sons.

$\chi^{(3)}$–Third-Harmonic Generation

B Y Soon, Thales Electro-Optics PTE Ltd., Singapore

J W Haus, University of Dayton, Dayton, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The invention of the ruby laser, in 1960, provided a revolutionary new light source, which emits an intense, highly collimated polarized beam of light. The new source opened the way for developments that have enabled new technologies, thus profoundly changing our way of life. One of these new frontiers opened by the laser was the field of nonlinear optics. In 1961, Franken *et al.* were the first to report the observation of second-harmonic generation from a quartz crystal, using ruby laser light. Third-harmonic generation (THG) was soon discovered by Terhune and co-workers in 1962.

THG has found applications as a spectroscopic tool to probe optical properties of materials, but recently it has been developed as a microscopy tool and a pulse measurement technique. More efficient THG is possible by engineering the materials; for instance, one way to improve conversion efficiency is to use periodic structured dielectric materials, called photonic bandgap structures.

Polarization and the Third-Order Susceptibility

THG is a nonlinear optical phenomenon, where light of angular frequency ω is converted into light at an angular frequency 3ω that is three times higher. Light is made of particles called photons and the nonlinear medium enables the process whereby three photons of frequency ω are annihilated to create one photon of frequency 3ω. The nonlinear interactions are strongly dependent on the polarization or the dipole moment per unit volume, which is written as a series

$$P(t) = \chi^{(1)} \times E(t) + \chi^{(2)} : E^2(t) + \chi^{(3)} \vdots E^3(t) + \cdots \quad [1]$$

where χ is the susceptibility of the optical material, the superscript of χ in parentheses denotes the order of nonlinearity, t is the time variable, and E is the electric field. Both the electric field and the polarization are vectors; therefore the susceptibilities are tensors. The first-order or linear susceptibility $\chi^{(1)}_{ij}$ is a second-order tensor. This term constitutes the linear polarization of the medium denoted a P^{L}. It has nine elements, but only up to three independent elements depending on the symmetry of the lattice. The other terms represent an expansion for the nonlinear polarization of the medium, P^{NL}. Note that the superscripts 'L' and 'NL' denote linear and nonlinear polarization contributions, respectively. The third-order susceptibility $\chi^{(3)}_{ijkl}$ is a four-rank tensor, which usually has 81 elements ($3 \times 3 \times 3 \times 3 = 81$). In asymmetric crystalline solids, all 81 elements can be of nonzero value and independent. With a higher degree of symmetry in other optical materials, the 81 elements are no longer all independent, and thus the number of independent elements can be reduced to a lower number. For example, isotropic material such as glass, liquid, or vapor has only three independent elements. For illustration, we consider a general case where the applied frequencies ω are arbitrary, and the third-susceptibility is given as

$$\chi_{ijkl} \equiv \chi^{(3)}_{ijkl}(\omega_4 = \omega_1 + \omega_2 + \omega_3) \quad [2]$$

Note that the superscript has been removed. In the isotropic material, the coordinate axes must be

equivalent. Thus, the susceptibility has symmetry given by

$$\chi_{1111} = \chi_{2222} = \chi_{3333} \quad [3a]$$

$$\chi_{1122} = \chi_{1133} = \chi_{2211} = \chi_{2233} = \chi_{3311} = \chi_{3322} \quad [3b]$$

$$\chi_{1212} = \chi_{1313} = \chi_{2323} = \chi_{2121} = \chi_{3131} = \chi_{3232} \quad [3c]$$

$$\chi_{1221} = \chi_{1331} = \chi_{2112} = \chi_{2332} = \chi_{3113} = \chi_{3223} \quad [3d]$$

For the isotropic material, there are only 21 nonzero elements, as given in eqns [3a]–[3d], because these elements have a Cartesian index (1, 2, or 3) that appears an even number of times. The reason for nonexistence of the elements with an index appearing an odd number of times is because, for example, χ_{1222} would give a response in the $\hat{x}_1$ direction due to a field applied in the $\hat{x}_2$ direction. This response must vanish in an isotropic material, because there is no reason why the response should be in the $+\hat{x}_1$ direction rather than in the $-\hat{x}_1$ direction. In summary, the elements in the third-order susceptibility of the isotropic material have the relationship given by

$$\chi_{ijkl} = \chi_{1122}\delta_{ij}\delta_{kl} + \chi_{1212}\delta_{ik}\delta_{jl} + \chi_{1221}\delta_{il}\delta_{jk} \quad [4]$$

where δ is the Kronecker delta function; it has a unit value, $\delta_{ij} = 1$, if and only if $i = j$, otherwise, it is zero, $\delta_{ij} = 0$. Thus, we see from eqn [4] that there are only three independent elements in the isotropic material. Crystalline materials have more independent components and they have been determined for the 32 crystals classes. The relationships between the elements for these crystalline classes are tabulated in books by Boyd and by Butcher and Cotter.

A block diagram of a material illustrates THG and its corresponding energy-level diagram is depicted in Figure 1. Figure 1a shows that fields of frequency ω excite the χ^3 optical material to generate the third-harmonic field with the residual of the fundamental harmonic field. Figure 1b shows that three photons of frequency ω are destroyed to generate a photon of frequency 3ω in the third-harmonic generation.

For third-harmonic generation, the frequency dependence of the susceptibility is expressed as $\chi_{ijkl}(3\omega = \omega + \omega + \omega)$. With the intrinsic permutation symmetry of the nonlinear susceptibility, the elements of the susceptibility tensor has a relationship such that $\chi_{1122} = \chi_{1212} = \chi_{1221}$. So we can further reduce eqn [4] for the third-harmonic generation as

$$\chi_{ijkl}(3\omega = \omega + \omega + \omega) = \chi_{1122}(3\omega = \omega + \omega + \omega) \times (\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) \quad [5]$$

Therefore, we see that the third-harmonic susceptibility tensor has only one independent element.

Wave Equation for χ^3 Third-Harmonic Generation

Theoretically, we have to use the wave equation to observe how an intense laser field interacts with the nonlinear material, so that the polarization of the medium could develop new frequency components that are not present in the incident field. Thus, these new frequency components of the polarization become the sources of the new frequency components of the electromagnetic field.

To formally describe THG, we begin with Maxwell's equations (in Gaussian units), and we have

$$\nabla \times D = 4\pi\rho \quad [6a]$$

$$\nabla \times B = 0 \quad [6b]$$

$$\nabla \times E = \frac{-1}{c}\frac{\partial B}{\partial t} \quad [6c]$$

$$\nabla \times H = \frac{-1}{c}\frac{\partial D}{\partial t} + \frac{4\pi}{c}J \quad [6d]$$

where H is the magnetic field, B is the magnetic flux density, D is the electric flux density, ρ is the charge density, and J is the current density. The desired solution is in the region of space in the nonlinear material that has no free charge ($\rho = 0$), and therefore, there is no free current ($J = 0$). We assume

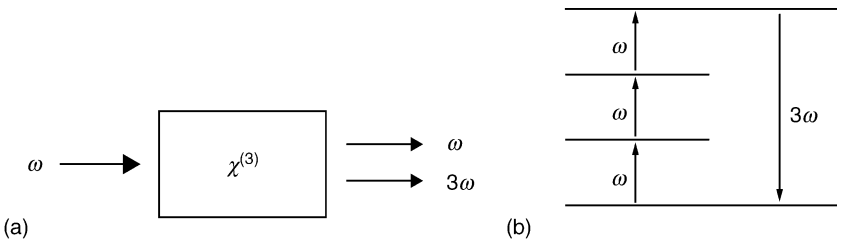


Figure 1 The third-harmonic (a) block diagram; and (b) energy level description.

that the nonlinear material is nonmagnetic, so that $B = \mu H = H$, where $\mu = 1$ in a nonmagnetic environment. In nonlinear material, the electric flux is defined as

$$D = E + 4\pi P \quad [7]$$

so that the polarization depends nonlinearly on the strength of E. Using eqns [6c] and [6d], we have

$$\nabla \times \nabla \times E + \frac{1}{c^2}\frac{\partial^2 D}{\partial t^2} = 0 \quad [8]$$

Substituting eqn [7] into [8], we have

$$\nabla \times \nabla \times E + \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} = \frac{-4\pi}{c^2}\frac{\partial^2 P}{\partial t^2} \quad [9]$$

Using an identity from the vector calculus, we have

$$\nabla \times \nabla \times E = \nabla(\nabla \times E) - \nabla^2 E \quad [10]$$

In linear optics, we see from eqn [6a] that when there is no free charge in source-free isotropic material, $\nabla \times D = \nabla \times E = 0$. Thus, the first term on the right-hand side of eqn [10] has no effect. Whereas in nonlinear optics, $\nabla \times E \neq 0$ in isotropic material, but the contribution from the first term on the right-hand side of eqn [10] is negligible, therefore, we can ignore it. Thus, eqn [9] can be simplified as

$$\nabla^2 E - \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} = \frac{4\pi}{c^2}\frac{\partial^2 P}{\partial t^2} \quad [11]$$

For distinction of the linear and nonlinear contributions, we modified eqn [7] to get

$$D = E + 4\pi P^{\mathrm{L}} + 4\pi P^{\mathrm{NL}}$$

where $P = P^{\mathrm{L}} + P^{\mathrm{NL}}$. Let $D^{\mathrm{L}} = E + 4\pi P^{\mathrm{L}}$, and we rewrite eqn [11] as

$$\nabla^2 E - \frac{1}{c^2}\frac{\partial^2 D^{\mathrm{L}}}{\partial t^2} = \frac{4\pi}{c^2}\frac{\partial^2 P^{\mathrm{NL}}}{\partial t^2} \quad [12]$$

Equation [12] is the most general form of the wave equation seen in nonlinear optics, and could be used as our starting point for our study of third-harmonic generation.

Methods of Third-Harmonic Generation

Since nonlinear optics is a popular research topic, there are a handful of existing methods that can be used for third-harmonic generation, and possibly new methods may be developed in the future. Here, we will discuss three methods using χ^3 material: (i) photonic crystal; (ii) defect mode using photonic crystal; and (iii) homogeneous nonlinear crystal.

Photonic Crystal

After the invention of electronic bandgap structure, the optical scientists and engineers saw the possibility to extend the operating waveband into the visible, and thus we have the photonic bandgap structure. The geometry and material choices of the photonic bandgap structure create a unique bandgap in frequency spectrum that has potential applications in laser optics, telecommunication, medical imaging, optical limiting, etc. A photonic bandgap structure is also commonly known as the photonic crystal. For a one-dimensional photonic crystal, two materials of certain thicknesses each, repeat themselves a number of times. For example, the first and second materials are of thickness 'a' and 'b', respectively, so that their combined thickness is 'd', and have a periodicity of 10. The schematic of such a one-dimensional photonic crystal, with its corresponding dielectric function as a function of axial distance z, is depicted in Figure 2.

By using χ^3 materials for photonic crystal, and using method of multiple scales (MMS), we have the nonlinear polarization given as

$$P^{\mathrm{NL}} = \lambda \chi^{(3)} E^3 \quad [13]$$

where λ is the perturbation parameter. With reference to Figure 2, assume that 'a' is equal to 'b' so that $d = 2a$, then the Fourier expansion of the dielectric

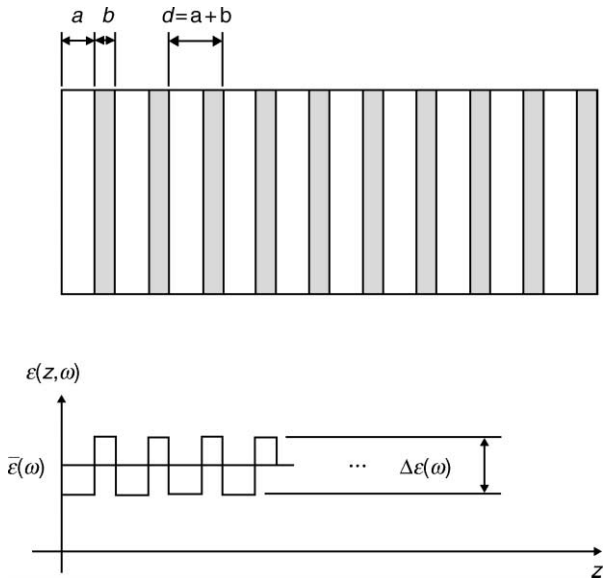


Figure 2 The schematic of the one-dimensional photonic crystal.

function is given as

$$\varepsilon(z,\omega) = \bar{\varepsilon}(\omega) + \lambda\Delta\varepsilon(\omega)\cos(2\pi z/d) - \frac{1}{3}\lambda\Delta\varepsilon(\omega)\cos(6\pi z/d) + \cdots \quad [14]$$

where $\bar{\varepsilon}(\omega)$ is the zeroth-order coefficient, and $\Delta\varepsilon(\omega)$ is first-order coefficient of the Fourier series expansion. Without going through the mathematics of MMS, we use eqns [12] and [13] to obtain the solution of the wave equation and two sets of coupled mode equations. The solution to the wave equation in eqn [12] is given by

$$E_0 = \left(\frac{1}{\sqrt{k_1}}A_{f1}(z_1,t_1)e^{ik_1z_0} + \frac{1}{\sqrt{k_1}}A_{b1}(z_1,t_1)e^{-ik_1z_0}\right)e^{-i\omega t_0} + \left(\frac{1}{\sqrt{k_3}}A_{f3}(z_1,t_1)e^{ik_3z_0} + \frac{1}{\sqrt{k_3}}A_{b3}(z_1,t_1)e^{-ik_3z_0}\right)e^{-3i\omega t_0} + \text{c.c.} \quad [15]$$

where A is the amplitude of the field, k is the wavenumber such that $k_1^2=\omega^2\varepsilon(\omega)/c^2$, and $k_3^2=(3\omega)^2\varepsilon(3\omega)/c^2$ where c is the speed of light, and 'c.c.' refers to the complex conjugate. Note that the subscript 'f' and 'b' refer to the forward and backward modes, respectively. Also, the subscript '0' refers to the slowest scale, and '3' refers to the fastest scale chosen in the analysis using MMS. Let $a_{f1}=NA_{f1}e^{i\pi\delta_1 z_1/d}$, $a_{b1}=NA_{b1}e^{-i\pi\delta_1 z_1/d}$, $a_{f3}=NA_{f3}e^{i\pi\delta_3 z_1/d}$, and $a_{b3}=NA_{b3}e^{-i\pi\delta_3 z_1/d}$, such that

$$N^2=\frac{2d\sqrt{3k_3}(3\omega)^2\chi^{(3)}}{k_3c^2k_1\sqrt{k_1}}$$

The coupled mode equations for the fundamental field are given as

$$\frac{\partial a_{f1}}{\partial z_1}+\frac{d}{\pi v_{g1}}\frac{\partial a_{f1}}{\partial t_1}=\frac{id}{2\pi k_1}\nabla_\perp^2 a_{f1}+i\delta_1 a_{f1}-i\kappa_1 a_{b1} -\frac{\alpha_1}{2}a_{f1}+ia_{f1}[2|a_{b1}|^2+|a_{f1}|^2] +ia_{f1}^{*2}a_{f3} \quad [16a]$$

$$\frac{-\partial a_{b1}}{\partial z_1}+\frac{d}{\pi v_{g1}}\frac{\partial a_{b1}}{\partial t_1}=\frac{id}{2\pi k_1}\nabla_\perp^2 a_{b1}+i\delta_1 a_{b1}-i\kappa_1 a_{f1} -\frac{\alpha_1}{2}a_{b1}+ia_{b1}[2|a_{f1}|^2+|a_{b1}|^2] +ia_{b1}^{*2}a_{b3} \quad [16b]$$

where, at the fundamental harmonic, v_{g1} is the group velocity of the homogeneous medium, δ_1 is the laser detuning, κ_1 is the coupling coefficient, α_1 is the absorption coefficient, $\nabla_\perp^2$ is the transverse Laplacian and the whole term ($\nabla_\perp^2 a_{f1}$ and $\nabla_\perp^2 a_{b1}$) accounts for diffraction effects, and '$*$' denote the complex conjugate of the corresponding term. Similarly, the coupled mode equations for the third-harmonic field are given as

$$\frac{\partial a_{f3}}{\partial z_1}+\frac{d}{\pi v_{g3}}\frac{\partial a_{f3}}{\partial t_1}=\frac{id}{2\pi k_3}\nabla_\perp^2 a_{f3} +ia_{f3}\left[\delta_3+6\sqrt{\frac{3k_1}{k_3}}(|a_{f1}|^2+|a_{b1}|^2)\right] -i\kappa_3 a_{b3}-\frac{\alpha_3}{2}a_{f3}+ia_{f1}^3 \quad [17a]$$

$$-\frac{\partial a_{b3}}{\partial z_1}+\frac{d}{\pi v_{g3}}\frac{\partial a_{b3}}{\partial t_1}=\frac{id}{2\pi k_3}\nabla_\perp^2 a_{b3} +ia_{b3}\left[\delta_3+6\sqrt{\frac{3k_1}{k_3}}(|a_{b1}|^2+|a_{f1}|^2)\right] -i\kappa_3 a_{f3}-\frac{\alpha_3}{2}a_{b3}+ia_{b1}^3 \quad [17b]$$

where at the third harmonic, v_{g3} is the group velocity of the homogeneous medium, δ_3 is the laser detuning, κ_3 is the coupling coefficient, and α_3 is the absorption coefficient. Note that in the limit when $k_3\approx 3k_1$, the coupling coefficient between the fundamental and third harmonics are related by

$$\kappa_3=3\left(\frac{d}{\pi}\right)\frac{\Delta\varepsilon\omega_0^2}{2k_3c^2}+3\delta_1=\kappa_1 \quad [18]$$

Also, note that the laser detuning between the fundamental and third harmonics are expressed as

$$\delta_3=\left(\frac{d}{\pi}\right)\Delta k+3\delta_1 \quad [19]$$

where $\Delta k=k_3-3k_1\times$ Equation [19] described the phase matching condition, which determines the efficiency of the third-harmonic generation. When the phase is completely matched ($\Delta k=0$), then we can achieved the most efficienct in third-harmonic generation when $\delta_3=3\delta_1$. Equations [16] and [17] are useful equations which are commonly used for numerically simulation of beam propagation.

Defect Mode Using Photonic Crystal

A defect mode can be generated in the photonic crystal if we introduce a defect in the geometry of the photonic crystal. The schematic of the photonic crystal with a defect of length d_d located at the center of the photonic crystal, where the subscript 'd' denotes defect, is depicted in Figure 3. For the convenience of the analysis, we divide the entire photonic crystal into three regions: (i) Region I; (ii) Region II; and (iii) Region III. Since we arbitrarily

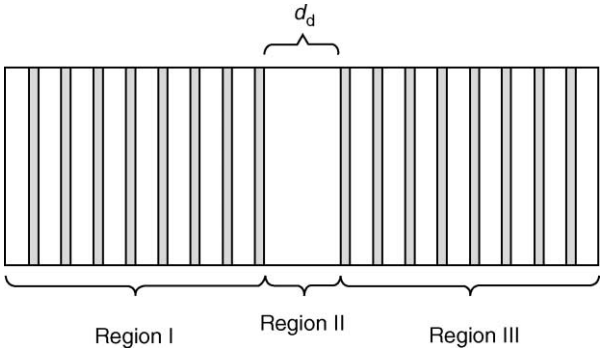


Figure 3 The schematic of the photonic crystal with a defect of length d_d.

put the defect in the center of the photonic crystal, Region I and II will be identical, thus it will be easier for us in our analysis. The strategy of our analysis is to solve the wave equation for Region I of the photonic crystal, and apply the result to the Region II analysis. Similarly, we apply the corresponding result of Region II to Region III. Thus, by breaking up the entire photonic crystal into three regions, we can then easily analyze the propagation of a field through the entire photonic crystal.

Assume that Region I and II are each of length $L/2$. Using the coupled mode equations without considering the effect of absorption, the fundamental fields in the forward and backward modes in Region I is given as

$$a_{f1(REG)}(z_1) = \left(\cos(\Delta_1 z_1) - \frac{i\delta_1}{\Delta_1}\sin(\Delta_1 z_1)\right)a_{f1(REG)}(0) + \frac{i\kappa}{\Delta_1}\sin(\Delta_1 z_1)a_{b1(REG)}(0) \quad [20a]$$

$$a_{b1(REG)}(z_1) = \left(\cos(\Delta_1 z_1) + \frac{i\delta_1}{\Delta_1}\sin(\Delta_1 z_1)\right)a_{b1(REG)}(0) - \frac{i\kappa}{\Delta_1}\sin(\Delta_1 z_1)a_{f1(REG)}(0) \quad [20b]$$

where

$$\Delta_1 = \sqrt{\delta_1^2 - \kappa^2}, \qquad \kappa = \frac{d\omega_0^2\Delta\varepsilon(\omega)}{2\pi k_1 c^2}$$

and the subscript '(REG)' denotes solution in either Region I, II, or III. The boundary conditions for the fundamental harmonic field, at the interface between Region I and II, are given as

$$a_{f1(II)}(d_d) = e^{ikd_d}a_{f1(I)}(L/2) \quad [21a]$$

$$a_{b1(II)}(d_d) = e^{-ikd_d}a_{b1(I)}(L/2) \quad [21b]$$

where $k = n\omega/c$. The field that propagates through Region II can be determined using eqns [20] and [21]. Also, the boundary conditions at the front and end faces of the Region III of the photonic crystal are given as

$$a_{f1(III)}(0) = a_{f1(II)}(d_d) \quad [22a]$$

$$a_{b1(III)}(L/2) = 0 \quad [22b]$$

Using eqns [20] and [22], we can determine the field that propagates through Region III. Similarly, we can use the same approach to analyze the third-harmonic field.

Homogeneous Nonlinear Crystal

If we examine Figure 2, we see that when $\Delta\varepsilon(\omega) = 0$, the photonic crystal becomes a homogeneous nonlinear crystal. In another word, $\kappa_3 = 0$ when we set the laser detuning to zero, thus there is no more coupling between the layers, and we are left with only the forward mode. Without going through the mathematics of four-wave mixing to describe the THG in a homogeneous nonlinear crystal, we can obtain the fundamental and third-harmonic coupled mode equations from eqns [16] and [17]. This will make it easier to make comparisons between the photonic crystal and the homogeneous nonlinear crystal, rather than comparing the coupled mode equation with the four-wave mixing equations among them. Since no backward wave is generated in the homogeneous nonlinear crystal when there is no coupling between the forward and backward modes, eqn [16] can remove the coupling term and the backward mode equation, and the modification is given as

$$\frac{\partial a_{f1}}{\partial z_1} + \frac{d}{\pi v_{g1}}\frac{\partial a_{f1}}{\partial t_1} = \frac{id}{2\pi k_1}\nabla_\perp^2 a_{f1} + i\delta_1 a_{f1} - \frac{\alpha_1}{2}a_{f1} + ia_{f1}|a_{f1}|^2 + ia_{f1}^{*2}a_{f3} \quad [23]$$

$$\frac{\partial a_{f3}}{\partial z_1} + \frac{d}{\pi v_{g3}}\frac{\partial a_{f3}}{\partial t_1} = \frac{id}{2\pi k_3}\nabla_\perp^2 a_{f3} + ia_{f3}\left[\delta_3 + 6\sqrt{\frac{3k_1}{k_3}}|a_{f1}|^2\right] - \frac{\alpha_3}{2}a_{f3} + ia_{f1}^3 \quad [24]$$

for the fundamental and third-harmonic fields, respectively.

Applications

Generation of short wavelength radiation is possible using THG. Therefore, it becomes less expensive to generate green, blue, or ultraviolet light with the

infrared and the red light. In order to improve THG conversion efficiency, a beam with a tight focus is used, which results in an electric field of highest intensity, and has the highest possible χ^3 polarization effect. Normally, due to the phenomenon of the Gouy phase shift at a tight focal point of the beam, third-harmonic generation is suppressed. However, in the presence of an interface within the focal volume, the phase is disturbed and a third-harmonic signal is generated. This phenomenon has been used to image cells in *in vivo* microscopy on a femtosecond time-scale. The ultrashort pulses have little energy and therefore little heat is deposited in the material. The cells can be continuously imaged without damage.

See also

Materials Characterization Techniques: $\chi^{(3)}$. **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguides.

Further Reading

Barad Y, Eisenberg H, Horovitz M and Silberberg Y (1997) Nonlinear scanning laser microscopy by third harmonic generation. *Applied Physics Letters* 70: 922–924.

Boyd RW (1992) *Nonlinear Optics*. Boston, MA: Academic Press.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge: Cambridge University Press.

Franken PA, Hill AE, Peters CW and Weinreich G (1961) Generation of optical harmonics. *Physics Review Letters* 7: 118.

Markowicz PP, Tiryaki H, Pudavar H, Prasad PN, Lepeshkin NN and Boyd RW (2004) Dramatic enhancement of third-harmonic generation in three-dimensional photonic crystals. *Physics Review Letters* 92: 083903.

Petrov GI, Shcheslavskiy V, Yakovlev VV, Ozerov I, Chelnokov E and Marine W (2003) Efficient third-harmonic generation in a thin nanocrystalline film of ZnO. *Applied Physics Letters* 83: 3993–3995.

Schins J, Schrama T, Brakenhoff G, Muller M and Squier J (2002) Determination of material properties using third harmonic generation microscopy. *Journal of the Optical Society of America B* 19: 1627.

Squier J, Mueller M, Brakenhoff GJ and Wilson KR (1998) Third harmonic generation microscopy. *Optics Express* 3: 315.

Terhune RW, Maker TD and Savage CM (1962) Optical harmonic generation in calcite. *Physics Review Letters* 8: 404.

Viswanathan R (1999). Enhanced Frequency Conversion in Periodic Structures. Ph.D. dissertation, Rensselaer Polytechnic Institute.

Four-Wave Mixing

L Canioni and L Sarger, University of Bordeaux, Talence, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Light waves propagate along straight lines in a vacuum and in materials. However, their amplitude and phase can only be affected by the intrinsic properties of the sample, such as absorption and refraction index. Experimental work from the early 1960s, has deviated from this simple approach, mainly through propagating intense light beams in crystals and glasses. Among many puzzling new effects, frequency conversion and beam coupling have triggered intense studies and opened a whole new field, theoretically and experimentally. These various phenomena have been classified according to the order of their optical nonlinearity. The second order mainly occurs in frequency generation and has already been discussed in previous articles of this encyclopedia. Therefore, this article will be devoted to the higher order, in which self focusing, coupling, and mixing of light beams will be explained. We will illustrate some simple ideas with a convenient model before describing a few applications.

The framework of electromagnetic wave interaction with matter is embedded in the Lorentz force. The oscillating electric field drives the components, ions, and electrons, but due to the large mass differences, only the electron movement is strongly perturbed. In dielectric media, these carriers are connected with static bonds such that the applied electric field modulates the spatial position of equilibrium of the charges. As a result, a dipolar moment is created and oscillates quasi-adiabatically at very high frequencies (between 10^{14} and 10^{17} Hz) for an applied optical wave. These collective oscillations of the induced dipoles compose the so-called matter polarization, proportional to the driving electric field amplitude. Magnetic field interaction is usually much weaker and so can be neglected in nonmagnetic media.

As an electromagnetic wave, lasers are a coherent source of very high intensity, i.e., large electric fields. For example, in standard pulsed

lasers focused on a target, the light intensity can easily reach 10 Tw/cm^2, corresponding to an electric field of the order of magnitude of the internal bonding field (10^{10} V/m). At this level of interaction, dipolar moments are no longer strictly proportional to the applied electric field, as depicted in Figure 1. In the weak field regime, the simple perturbation approach displaces the carriers in a power series of the applied electric field. The polarization is thus developed in series, in respect to this electric field.

Formalism

Relation Between Field and Polarization: The Response Function

Although numerous works have been published with a frequency description of the material polarization, we will discuss here the study of nonlinearity in the time domain in order to follow the physical reality. In most processes, both spaces are equal and can be selected at will, as long as the entire requirements are fully understood; the electromagnetic field is classically described.

In the time domain, the polarization can be developed as

$$P(r,t) = P_{\text{Linear}}(r,t) + P_{\text{Nonlinear}}(r,t) \qquad [1]$$

where the linear part of the polarization can be written as

$$P_{\text{Linear}}(r,t) = \varepsilon_0 \int_{R^3} \int_{-\infty}^{\infty} R^{(1)}(r_1,t_1) \times E(r-r_1,t-t_1)\mathrm{d}t_1\,\mathrm{d}r_1 \qquad [2]$$

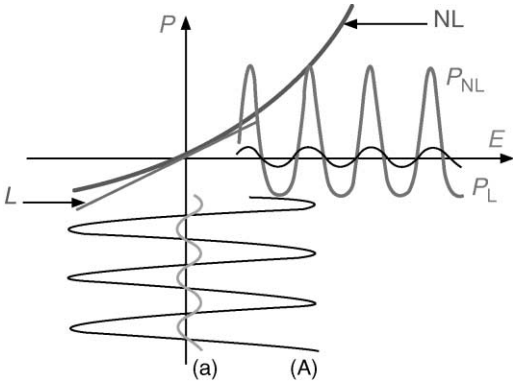


Figure 1 Polarization for various field amplitude: (a) Weak field case, linear response. (A) Strong field, distortion and nonlinearity.

This general expression is rather complex but shows that the polarization, due to collective movement of carriers in a macroscopic volume (thousands of dipoles) centered at position r, always depends on the applied electric field. This expression is also nonlocal as dipoles–dipoles interactions play an important role.

In general, for the nonlinear (NL) part of the polarization, an approximation of the local interaction is described for simplicity. For a second-order interaction, one can write:

$$P^{(2)}_{\text{NL}}(r,t) = \varepsilon_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} R^{(2)}(t_1,t_2) \times E(r,t-t_2)E(r,t-t_2-t_1)\mathrm{d}t_1\,\mathrm{d}t_2 \qquad [3]$$

For the other series expansion – at third order, one obtains

$$P^{(3)}_{\text{NL}}(r,t) = \varepsilon_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} R^{(3)}(t_1,t_2,t_3)E(r,t-t_3) \times E(r,t-t_3-t_2)E(r,t-t_3-t_2-t_1)\mathrm{d}t_1\,\mathrm{d}t_2\,\mathrm{d}t_3 \qquad [4]$$

This third-order polarization already combines three electric fields. This nonlinear polarization also radiates a fourth wave and is therefore the adopted framework for describing four-wave mixing experiments (Figure 2).

As the physical properties of materials depend only on time intervals between the different pulse interactions, eqn (4) can be rearranged as

$$P^{(3)}_{\text{NL}}(r,t) = \varepsilon_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} R^{(3)}(t-\tau_1,t-\tau_2,t-\tau_3) \times E(r,\tau_1)E(r,\tau_2)E(r,\tau_3)\mathrm{d}\tau_1\,\mathrm{d}\tau_2\,\mathrm{d}\tau_3 \qquad [5]$$

If the sample response time is much shorter than pulse duration, the so-called adiabatic limit (as for electronic cloud deformation excited distant from

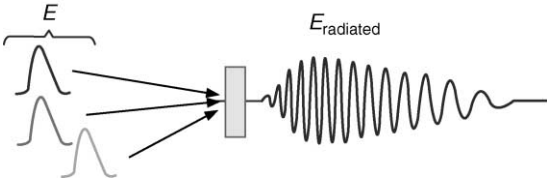


Figure 2 Three incoming beams E_1, E_2, E_3 create a polarization (nonlinear). The sample responds while radiating a fourth field E_4.

resonance) the polarization is written as

$$P_{\mathrm{NL}}^{(3)}(r,t)=\varepsilon_0\sigma^{(3)}E(r,t)E(r,t)E(r,t) \qquad [6]$$

We will see later how this expression is useful for third-harmonic generation and optical Kerr effects, etc. This expression is accurate for materials excited at a distance from resonances and care must be taken to reserve such a model for corresponding processes. In this case, the frequency approach is quite simple too and both spaces, time or Fourier, can be chosen at will.

For very long response time, such as orientational or structural molecular reorganization, one can use the Born–Oppenheimer approximation and present the nonlinear polarization as

$$P_{\mathrm{NL}}^{(3)}(r,t)=\varepsilon_0E(r,t)R'(t) \qquad [7]$$

Similar is linear polarization, where the response function takes into account the changes of the molecular structure under the light intensity stress:

$$R'(t)=\int_{-\infty}^{+\infty}d^{(3)}(t-t_1)E(r,t_1)E(r,t_1)\mathrm{d}t_1 \qquad [8]$$

The material response will then be mostly driven by the field amplitude rather than the frequency.

Properties of Nonlinear Susceptibilities

The third-order nonlinear susceptibility $\chi^{(3)}$ is the Fourier transform of the response function R and has proved to be a most practical tool in optical nonlinear theory. The symmetry properties of the material lead to simplification on the tensor of rank 4 (fourth rank tensor), involved in four-wave mixing:

$$P_i^{(3)}(\omega_m)=\varepsilon_0\sum_{n_1,\ldots,n_p}\sum_{i_1,\ldots,i_p}\chi_{i,i_1,\ldots,i_p}^{(3)}(\omega_m;\omega_{n_1},\ldots,\omega_{n_p})\times E_{i_1}(\omega_{n_1})\cdots E_{i_p}(\omega_{n_p}) \qquad [9]$$

The nonlinear susceptibility tensor must account for all the spatial and symmetrical order of the sample. This will eventually reduce the number of independent elements from a possible 81.

Wave Equation in Nonlinear Regime

From Maxwell's equations, with a sample without free carriers and current, one can deduce the wave equation in the time domain, as:

$$\Delta E-\frac{1}{c^2}\frac{\partial^2E}{\partial t^2}-\frac{1}{c^2}\frac{\partial^2P_{\mathrm{L}}}{\partial t^2}=\frac{1}{c^2}\frac{\partial^2P_{\mathrm{NL}}}{\partial t^2} \qquad [10]$$

In the frequency domain, this equation reduces to

$$\Delta\tilde{E}(r,\omega)+k(\omega)^2\tilde{E}(r,\omega)=-\mu_0\omega^2\tilde{P}_{\mathrm{NL}}(r,\omega) \qquad [11]$$

In order to solve these (coupled) wave equations in relation to four-wave mixing, we must develop the electric field and the polarization in a slowly varying envelope approximation, using the following conventions:

$$E_i(r,t)=\frac{1}{2}\left(A_i(r,t)\mathrm{e}^{\mathrm{i}(\omega_it-k(\omega_i)z)}+\bar{A}_i(r,t)\mathrm{e}^{-\mathrm{i}(\omega_it-k(\omega_i)z)}\right) \qquad [12]$$

$$P_{\mathrm{NL}}(r,t)=\frac{1}{2}\left(p_{\mathrm{nl}}(r,t)\mathrm{e}^{\mathrm{i}(\omega t-k(\omega)z)}+\bar{p}_{\mathrm{nl}}(r,t)\mathrm{e}^{-\mathrm{i}(\omega t-k(\omega)z)}\right) \qquad [13]$$

Within the framework of small perturbations, diffraction can be neglected and the wave equation can be easily reduced to one dimension:

$$\frac{\partial A}{\partial z}=\frac{\mathrm{i}\omega}{2n(\omega)\varepsilon oc}p^{\mathrm{NL}}(z,\omega)\exp(-\mathrm{i}k(\omega)z) \qquad [14]$$

where ω represents the sum or difference permutations over all the frequencies ω_i, thus reflecting the law of energy conservation.

In the case of four-wave mixing, one has to solve four coupled equations for the four interacting waves. Using the polarization eqn [13], one gets:

$$\frac{\partial A_j}{\partial z}=\frac{\mathrm{i}\omega_j}{2n(\omega_j)c}\chi^3A_1^{(*)}A_2^{(*)}A_3^{(*)}\times\exp\left(\sum_{i=1,2,3}\exp(\pm\mathrm{i}k_i(\omega_i)z)\right)\exp(-\mathrm{i}k(\omega_j)z) \qquad [15]$$

This set of generic equations respectively represent indifferently, either the three incoming beams ($j=1,2,3$), or the radiated field E_4 at ω. The algebra used here associates at each complex conjugate field (*) with the counter propagating wave ($-k$). The amplitude variation of the j field at frequency ω_j is due to the coupling of the three other fields within the sample. The accumulated field is nevertheless always strongly dependent of the phase term and will vanish except when this phase term is zero. This is known as the phase matching condition which has to be fulfilled to ensure an efficient energy conversion. A more general approach of this condition can be found in the corresponding article of this encyclopedia.

Example and Selected Applications

When inserting the full material polarization (linear and nonlinear) into the propagation equation, even

restricted at third order, a whole set of processes must be considered. All of them are described by the generic equation already presented, but only a few very important processes – with increasing complexity – will be described here.

Optical Kerr Effect

We consider one single beam incident at frequency ω on a sample. The non-linear polarization is given by

$$P_{\mathrm{NL}}^{(3)}(r,t)=\frac{\varepsilon_0}{8}\chi^{(3)}\left(E(r,t)^3+E^*(r,t)^3+3E(r,t)^2E^*(r,t)+3E(r,t)E^*(r,t)^2\right) \quad [16]$$

The polarization at frequency ω reduces to

$$P_{\mathrm{NL}}^{(3)}(r,t)=\frac{3\varepsilon_0}{4}\chi^{(3)}|E(r,t)|^2E(r,t) \quad [17]$$

This term has the same phase as the incident beam and, therefore, the phase matching is automatic. Using eqns [12] and [13], one can rewrite the wave equation for the optical Kerr effect in the thin sample approximation:

$$\frac{\partial A}{\partial z}=\frac{3\mathrm{i}\omega}{8nc}\chi^{(3)}|A|^2A \quad [18]$$

The solution of this equation is

$$A(z)=A(0)\exp\left(\frac{3\mathrm{i}}{8nc}\chi^{(3)}|A(0)|^2z\right) \quad [19]$$

Along the beam, the amplitude of the input pulse stays constant for this process but its phase is affected by a nonlinear shift proportional to power density I.

In literature, one can find the concept of n_2, the sample nonlinear index, in the following formula:

$$\psi_{\mathrm{NL}}(z)=\frac{\omega}{c}n_2Iz \quad [20]$$

In this way of describing the total phase along the beam, the index of refraction in the nonlinear regime can be expressed more conveniently as

$$n=n_0+n_2I \quad [21]$$

One of the most significant demonstrations of this effect is the self-focusing phenomenon. The usual Gaussian transverse structure of a laser beam presents high intensity at the center of the beam, therefore creating an index variation in the isotropic sample due to this optical Kerr effect. An induced lens can then be applied to this refractive index gradient, following the field distribution. As a result, the beam focuses itself along the propagation and can result in a complete breakdown.

Third Harmonic Generation

In the third harmonic generation, i.e., the third-order nonlinear process where the frequency of the generated wave is three times the frequency of the input wave at ω, only two waves are in the sample (at ω and 3ω). For the nonlinear wave, the master equation is then:

$$\frac{\partial A_{3\omega}}{\partial z}=\frac{3\mathrm{i}\omega}{8n(3\omega)c}\chi^{(3)}A_\omega^3\exp((3\mathrm{i}k(\omega)-\mathrm{i}k(3\omega))z) \quad [22]$$

Here, propagation effects have to be carefully considered as theses waves do travel at different velocities and will eventually be out of phase. Obviously this discrepancy of phase velocities precludes the co-propagation and strongly affects the global efficiency. The only way to ensure a coherent generation of the harmonic wave throughout the material is to fulfill the relation:

$$k(3\omega)=3k(\omega) \quad [23]$$

This is the so-called phase matching condition. Although difficult because the usual dispersion of materials is very large for such a frequency difference, this can be satisfied using either geometrical arrangement or, if possible, using the birefringent properties of the material. The efficiency of this method is poor and a better approach is implemented when two $\chi^{(2)}$ processes are cascaded (doubling and sum frequency mixing).

Degenerate Four-Wave Mixing and Phase Conjugation

One of the most popular and intriguing interactions is depicted in Figure 3. Where the sample is excited by three fields at the same frequency ω in this particular geometry: two-waves, named pump beams E_1 and E_2, are exactly counter-propagating with opposite wave vectors of $+k_1$ and $k_2=-k_1$, while the third wave (E_3) arrives at the sample with a given wave vector of k_3.

Of possible couplings, one contribution E_4 corresponds to a re-emitted beam along $k_4=-k_3$, i.e., opposite to the E_3 wave.

The third-order nonlinear polarization is given here by

$$P_{\mathrm{NL}}^{(3)}(r,t) = \frac{\varepsilon_0}{8}\chi^{(3)}(E_1(r,t) + E_1^*(r,t) + E_2(r,t) + E_2^*(r,t) + E_3(r,t) + E_3^*(r,t))^3 \quad [24]$$

This E_4 wave (along $-k_3$) originates in this product from source terms including:

$$E_1(r,t)E_2(r,t)E_3^*(r,t) \quad [25]$$

and thus present an evolution driven by:

$$\frac{\partial A_4}{\partial z} = \frac{3\mathrm{i}\omega}{4n(\omega)c}\chi^{(3)}A_1A_2A_3^* \quad [26]$$

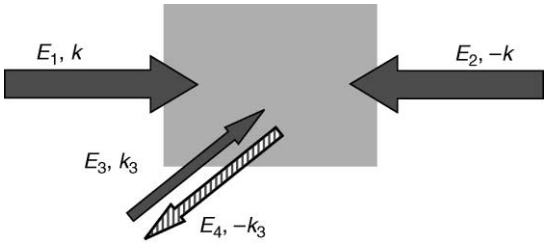


Figure 3 Two counter-propagating fields (E_1, E_2) and the third field E_3 create a polarization (nonlinear). The sample responds while radiating a fourth field E_4 opposite to the E_3 wave.

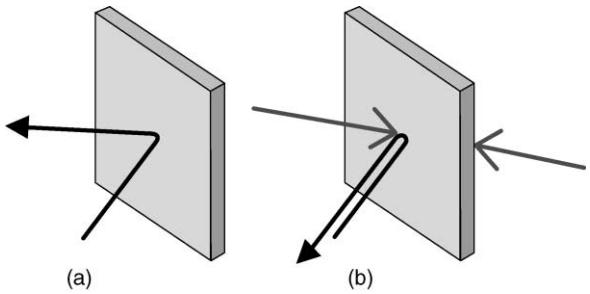


Figure 4 (a) Snell refraction in a conventional mirror; (b) Phase conjugate mirror.

where the phase matching condition is automatically fulfilled. Interesting facts can be described using this expression. First, the radiated amplitude is proportional to the complex conjugate of the E_3 field. This nonlinear process is not only able to reverse the direction of propagation but also the whole phase of an arbitrary incoming beam of light. This 'phase conjugator' can be considered as a kind of mirror with very unusual reflection properties. Unlike a conventional mirror, where a ray is redirected according to the ordinary law of reflection, a phase-conjugate mirror (PCM) retro-reflects all incoming rays back to their origin (Figure 4).

Second, the reflectivity of this mirror depends on the susceptibility and the amplitude of the first two fields, usually called pump beams. This will allow reflectivity that is much higher than the unity, unlike with conventional mirrors.

The remarkable reflection properties of the PCM have found many important applications. The most useful undoubtedly is related to distortion correction. If the image information has been distorted by the transmitting medium on its way to the PCM, then these aberrations will be corrected when the reflected signal retraces its original path through the medium. Through this amazing 'healing' property of optical phase conjugation, a high-quality optical beam can be double passed through a distorting medium, such as an imperfect imaging device or a turbulent atmosphere, without any loss of beam quality. Figure 5 shows the basic two-pass geometry for imaging and aberration correction using a PCM.

The spherical wave from the object point P is aberrated by the distorting medium. The wave-front is reversed by the PCM and a second passage through the same distorting path restores the initial field. The resulting spherical wave converges to the image point P', separated from P by the beamsplitter.

Numerous other applications for PCM and the various underlying nonlinear optical interactions have been proposed. These include effects as diverse

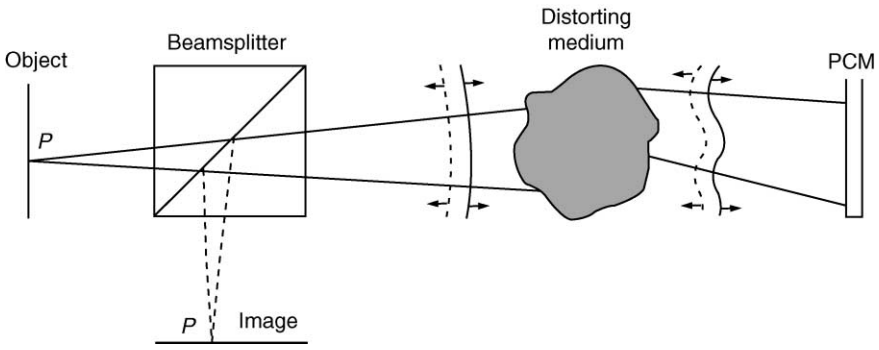


Figure 5 A basic two-pass system for imaging and distortion compensation using optical phase conjugation.

as parametric oscillation, optical tracking and pointing, spatial and temporal image processing, optical computing, optical filtering, etc. The PCM is also successfully implemented for aberration correction in high-power laser resonators.

Nondegenerate Four-Wave Mixing

If the frequencies of the impinging fields build a nonlinear polarization with an oscillating term close to absorption (or emission) frequencies of the sample, the nonlinear susceptibility increases tremendously. The radiated field amplitude changes by orders of magnitude and allows spectroscopic studies of the sample. The most popular process used to perform gas spectroscopy is CARS, an acronym for coherent anti-Stokes Raman scattering (or spectroscopy). In this case, the three incoming beams have different frequencies and the phase matching condition is eventually fulfilled in the 3D space.

Usually, two beams at the same frequency ω_P (pump beams), are mixed with a third beam at a tunable frequency ω_S, according to the energy scheme in Figure 6. As the radiated frequency ω_{CARS} is higher, the Raman spectroscopic notation is used and the subscripts 'S' hold for 'Stokes' and 'P' for 'anti-Stokes'.

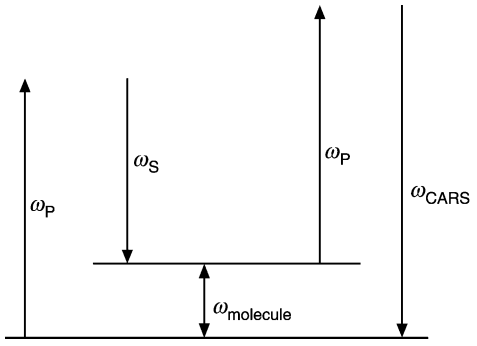


Figure 6 Energy diagram for a CARS experiment. Four-wave mixing efficiency increases when the frequency difference $\omega_P - \omega_S$ is close to a resonant frequency of the sample, $\omega_{molecule}$.

Whenever $\omega_P - \omega_S = \omega_{molecule}$, then an increase of the CARS intensity by many orders of magnitude is observed. Therefore, a Raman spectrum can be obtained by continuously changing, the Stokes laser and simultaneously recording the intensity of the CARS beam. This procedure is called scanning CARS, an interesting technique insofar as it allows taking high-resolution spectra, as well as measuring temperature of the sample while scanning over the Boltzmann distribution of the electronic ground state.

Figure 7 displays a classical geometrical arrangement where the phase matching condition fully determines the directionality of the radiated field at ω_{CARS}. Although the alignment of the setup can become tedious, this technique has been widely used, even in hostile industrial environments. The coherent aspect of this four-wave mixing process, imbedded in the phase matching condition, has a huge benefit over classical spectroscopic techniques and thus isotropic techniques, such as laser-induced fluorescence or Raman spectroscopy.

The CARS process is an interference process comparable to diffraction of a grating. The two fields at ω_P and ω_S form a laser-induced moving grating and the third field, at ω_P, undergoes a Doppler shift at ω_{CARS}. This scattered field undergoes a coherent amplitude addition in the Bragg direction.

Conclusion

In this article, we have discussed a set of optical nonlinear processes occurring at the third order of the development in series of the material polarization. At this order, wave mixing obviously shows most

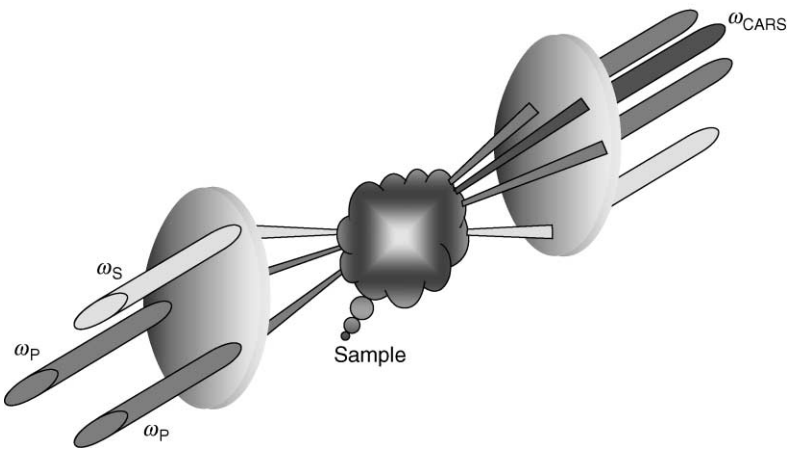


Figure 7 Layout of the CARS beams near a sample (a flame here). This particular arrangement is called folded (or 3D) Boxcars.

significant behavior and has been extensively studied and applied in many scientific areas. The general framework presented here can be further investigated using some books listed in Further Reading below.

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(3)}$–Third-Harmonic Generation. **Nonlinear Sources:** Harmonic Generation in Gases. **Optical Parametric Devices:** Optical Parametric Oscillators (Continuous Wave); Optical Parametric Oscillators (Pulsed).

Further Reading

Bloembergen N (1996) *Nonlinear Optics*. Singapore: World Scientific Pub Co.

Bloom DM and Bjorklund GC (1977) Conjugate wavefront generation and image reconstruction by four-wave mixing. *Applied Physics Letters* 31: 592–594.

Born M and Wolf E (1999) *Principles of Optics*. Cambridge, UK: Cambridge University Press.

Butcher PN and Cotter D (1991) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Eaton DF (1991) Nonlinear optical materials. *Science* 253: 281–287.

Fisher RA (1984) *Optical Phase Conjugation*. San Diego, CA: Academic.

Flytzanis C (1975) Theory of nonlinear optical susceptibilities. In: Rabin H and Tang CL (eds) *Quantum Electronics*, vol. 1, pp. 9–207. New York: Academic Press.

Hellwarth RW (1977) Generation of time-reversed wave fronts by nonlinear refraction. *Journal of the Optical Society of America* 67: 1–3.

Jackson JD (1998) *Classical Electrodynamics*. New York: Wiley.

Marburger JH (1975) Self-focusing: Theory. *Progress in Quantum Electronics* 4: 35–110.

Mittra R and Habashy TM (1984) Theory of wave-front-distortion correction by phase conjugation. *Journal of the Optical Society of America A* 1: 1103–1109.

Rockwell DA (1988) A review of phase-conjugate solid-state lasers. *IEEE Journal of Quantum Electronics* 24: 1124–1140.

Shen YR (2002) *The Principles of Nonlinear Optics*. New York: Wiley.

Yariv A (1977) Compensation for atmospheric degradation of optical beam transmission. *Optical Communication* 21: 49–50.

Yariv A (1989) *Quantum Electronics*, 3rd edn. New York: Wiley.

Yariv A and Pepper DM (1977) Amplified reflection, phase conjugation, and oscillation in degenerate four-wave mixing. *Optical Letters* 1: 16–18.

Zel'dovich BYa, Pilipetsky NF and Shkunov VV (1985) *Principles of Phase Conjugation*. Berlin: Springer-Verlag.

Kramers–Krönig Relations in Nonlinear Optics

M Sheik-Bahae, The University of New Mexico, Albuquerque, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Since their introduction nearly 75 years ago, the Kramers–Krönig (KK) dispersion relations have been widely appreciated and applied in the analysis of linear optical systems. Because they are a consequence of strict causality, the KK relations apply not only to optical systems, but also to any linear, causal system such as electrical networks and particle scattering. In this article, we review the formulation and application of these relations in nonlinear optical systems. Simple logical arguments are used to derive dispersion relations that relate the nonlinear absorption coefficient to the nonlinear refraction coefficient. More general formalisms are then derived that apply to all nonlinear susceptibilities including the harmonic generating cases. Examples of recent successful application of these dispersion relations in analyzing various nonlinear materials will be presented.

The mathematical formalism of the KK dispersion relations in nonlinear optics was studied in the formative days of the field. The great usefulness of these relations was appreciated only recently, however, when they were used to derive the dispersion of the optical Kerr effect in solids from the corresponding nonlinear absorption coefficients, including two-photon absorption.

Before examining the details of KK relations in nonlinear optical systems, it is instructive to revisit the linear dispersion relations and their derivation based on the logic of causality. We will begin this task by introducing the definition of the linear as well as nonlinear susceptibilities $\chi^{(n)}$. In most nonlinear optics texts, the total material polarization (P) that

drives the wave equation for the electric field (E) is expressed as

$$\begin{aligned}P_i(t) = \varepsilon_0 \int_{-\infty}^{\infty} R_{ij}^{(1)}(t-t_1)E_j(t_1)\mathrm{d}t_1 + \varepsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \\ \times R_{ijk}^{(2)}(t-t_1,t-t_2)E_j(t_1)E_k(t_2)\mathrm{d}t_1\,\mathrm{d}t_2 \\ + \varepsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R_{ijkl}^{(3)}(t-t_1,t-t_2,t-t_3) \\ \times E_j(t_1)E_k(t_2)E_l(t_3)\mathrm{d}t_1\,\mathrm{d}t_2\,\mathrm{d}t_3 + \cdots \end{aligned} \quad [1]$$

where $R^{(n)}$ is defined as the nth-order, time-dependent response function or time-dependent susceptibility. The subscripts are polarization indices indicating, in general, the tensor nature of the interactions. The summation over the various indices $j, k, l, \ldots$ is implied for the various tensor elements of $R^{(n)}$. Upon Fourier transformation, we obtain:

$$\begin{aligned}\vec{P}_i(\omega) = \varepsilon_0 \int_{\infty}^{-\infty} \mathrm{d}\omega_1\, \chi_{ij}^{(1)}(\omega_1)\vec{E}_j(\omega_1)\delta(\omega-\omega_1) \\ + \varepsilon_0 \int_{\infty}^{-\infty} \mathrm{d}\omega_1 \int_{\infty}^{-\infty} \mathrm{d}\omega_2 \\ \times \chi_{ijk}^{(2)}(\omega_1,\omega_2)\vec{E}_j(\omega_1)\vec{E}_k(\omega_2)\delta(\omega-\omega_1-\omega_2) \\ + \varepsilon_0 \int_{\infty}^{-\infty} \mathrm{d}\omega_1 \int_{\infty}^{-\infty} \mathrm{d}\omega_2 \int_{\infty}^{-\infty} \mathrm{d}\omega_3 \\ \times \chi_{ijkl}^{(3)}(\omega_1,\omega_2,\omega_3)\vec{E}_j(\omega_1)\vec{E}_k(\omega_2)\vec{E}_l(\omega_3) \\ \times \delta(\omega-\omega_1-\omega_2-\omega_3), \cdots \end{aligned} \quad [2]$$

where δ is the Dirac delta-function. Here the $E(\omega)$ are Fourier transforms of the corresponding electric field. The nth-order susceptibility is defined as the Fourier transform of the nth-order response function:

$$\begin{aligned}\chi_{ijk\ldots n}^{(n)}(\omega_1,\omega_2,\ldots,\omega_m) = \int_{-\infty}^{+\infty} \mathrm{d}\tau_1 \int_{-\infty}^{+\infty} \mathrm{d}\tau_2 \cdots \int_{-\infty}^{+\infty} \\ \times \mathrm{d}\tau_n R_{ijk\ldots m}^{(n)}(\tau_1,\tau_2,\ldots,\tau_m)\mathrm{e}^{\mathrm{i}(\omega_1\tau_1+\omega_2\tau_2+\cdots+\omega_m\tau_m)} \end{aligned} \quad [3]$$

For simplicity, we drop the polarization indices $i, j, \ldots$, and thus ignore the tensor properties of $\chi^{(n)}$ as well as the vector nature of the electric fields.

Let us for the moment concentrate on the linear polarization alone and derive the linear KK relations for the first-order susceptibility $\chi^{(1)}(\omega)$. For this, we rewrite eqn [3] for $n = 1$:

$$\chi^{(1)}(\omega) = \int_{-\infty}^{\infty} R^{(1)}(\tau)\mathrm{e}^{-\mathrm{i}\omega\tau}\,\mathrm{d}\tau \quad [4]$$

(As defined above, $\chi^{(1)}(\omega)$ and $R^{(1)}(\tau)$ are not a strict Fourier transform pair because of a missing factor of 2π.) Causality means that the effect cannot precede the cause. This can be restated mathematically as:

$$R^{(1)}(t) = R^{(1)}(t)\Theta(t) \quad [5]$$

i.e., the response to an impulse at $t = 0$ must be zero for $t < 0$. Here $\Theta(t)$ is the Heaviside step function defined as $\Theta(t) = 1$ for $t > 0$ and $\Theta(t) = 0$ for $t < 0$. Upon Fourier transforming this equation, the product in the time domain becomes a convolution integral in frequency space

$$\begin{aligned}\chi^{(1)}(\omega) &= \chi^{(1)}(\omega)\left[\frac{\delta(\omega)}{2} + \frac{\mathrm{i}}{2\pi\omega}\right] \\ &= \frac{\chi^{(1)}(\omega)}{2} + \frac{\mathrm{i}}{2\pi}\wp\int_{-\infty}^{\infty}\frac{\chi^{(1)}(\omega')}{\omega-\omega'}\mathrm{d}\omega' \\ &= \frac{1}{\mathrm{i}\pi}\wp\int_{-\infty}^{\infty}\frac{\chi^{(1)}(\omega')}{\omega'-\omega}\mathrm{d}\omega' \end{aligned} \quad [6]$$

which is the KK relation for the linear optical susceptibility. The symbol $\wp$ stands for the Cauchy principal value of the integral. The KK relation is thus a restatement of the causality condition [5] in the frequency domain. Taking the real part we have,

$$\Re e\{\chi^{(1)}(\omega)\} = \frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Im m\{\chi^{(1)}(\omega')\}}{\omega'-\omega}\mathrm{d}\omega' \quad [7]$$

Taking the imaginary part of eqn [6] leads to a similar relation relating the imaginary part to an integral involving the real part. It is conventional to write the optical dispersion relations in terms of the more familiar quantities of refractive index, $n(\omega)$, and absorption coefficient, $\alpha(\omega)$. For $|\chi^{(1)}| \ll 1$ then $n - 1 = \Re e\{\chi^{(1)}\}/2$ and $\alpha = \omega\Im m\{\chi^{(1)}\}/c$, and eqn [7] is transformed into

$$n(\omega) - 1 = \frac{c}{\pi}\wp\int_{0}^{\infty}\frac{\alpha(\omega')}{\omega' 2 - \omega^2}\mathrm{d}\omega' \quad [8]$$

where we additionally used the reality conditions of $n(\omega) = n(-\omega)$, and $\alpha(\omega) = \alpha(-\omega)$ to change the lower integral limit to 0. More rigorous analysis shows that eqn [8] is general and valid for any value of $|\chi^{(1)}|$. Although the KK dispersion relations and the extent of their applications in linear optics are well understood, some confusion sometimes exists about their applications to nonlinear optics. Causality clearly holds for both linear and nonlinear systems. The question is: what form do the resulting dispersion relations take in a nonlinear system? The linear Kramers–Krönig relations were derived from linear system theory, so it would appear to be impossible to apply the same logic to a nonlinear system. The key insight is that one can linearize the system. This is illustrated in Figure 1 where a linear

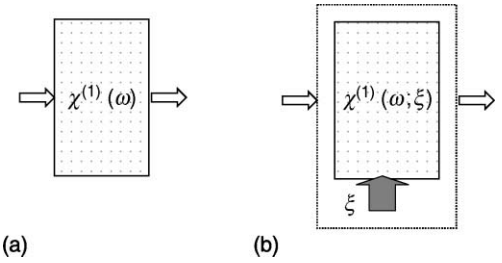


Figure 1 (a) A causal linear system obeying KK relations. (b) The system in (a) when externally perturbed by ξ. The dotted box now represents our new linear causal system whose altered $\chi^{(1)}$ obeys the KK relations.

(and of course, causal) optical material is transformed into a 'new' linear system that now contains the material and an external perturbation denoted by ξ. Although we are interested in perturbations of an optical nature, this formalism is general under any type of perturbation. It is important to appreciate the fact that our new system is causal even in the presence of the perturbation. This allows us to write down a modified form of the Kramers–Krönig relation linking the index of refraction to the absorption:

$$[n(\omega)+\Delta n(\omega;\zeta)]-1 = \frac{c}{\pi}\wp\int_0^\infty \frac{\alpha(\omega')+\Delta\alpha(\omega';\zeta)}{\omega' 2-\omega^2}\mathrm{d}\omega' \qquad [9]$$

which, after subtracting the linear relation between n and α leaves a relation between the changes in index and absorption:

$$\Delta n(\omega;\zeta)=\frac{c}{\pi}\wp\int_0^\infty \frac{\Delta\alpha(\omega';\zeta)}{\omega' 2-\omega^2}\mathrm{d}\omega' \qquad [10]$$

where ζ denotes a general perturbation. An equivalent relation also exists whereby the change in absorption coefficient can be calculated from the change in the refractive index. It is essential that the perturbation be independent of frequency of observation, ω', in the integral (i.e., the excitation ζ must be held constant as ω' is varied).

Equation [10] has been used to determine refractive changes due to 'real' excitations such as thermal and free-carrier nonlinearities in semiconductors. In those cases, ζ denotes either ΔT (change of temperature) or ΔN (change of free-carrier density), respectively. In the former case, one calculates the refractive index change resulting from a thermally excited electron–hole plasma and the temperature shift of the band edge. For cases where an electron–hole plasma is injected (e.g., optically), the change of absorption gives the plasma contribution to the refractive index. In this case, the ζ parameter in eqn [10] is taken as the change in plasma density regardless of the mechanism of generation or the optical frequency.

Let us now extend this formalism to the case where the perturbation is virtual, occurring at an excitation frequency Ω that is below any material resonance. To the lowest order in the excitation irradiance I_Ω, we write

$$\Delta\alpha(\omega;\zeta)=\Delta\alpha(\omega;\Omega)=2\alpha_2(\omega;\Omega)I_\Omega \qquad [11]$$

and

$$\Delta n(\omega;\zeta)=\Delta n(\omega;\Omega)=2n_2(\omega;\Omega)I_\Omega \qquad [12]$$

where n_2 and α_2 are the nonlinear refractive index and absorption coefficients of the material, respectively. By definition, these coefficients are related to the third-order nonlinear susceptibility $\chi^{(3)}(\omega_1,\omega_2,\omega_3)$ via (*see* **Nonlinear Optics, Basics**: Nomenclature and Units)

$$n_2(\omega;\Omega)=\frac{3}{4\varepsilon_0 n_0(\omega)n_0(\Omega)c}\Re e\{\chi^{(3)}(\omega,-\Omega,\Omega)\} \qquad [13]$$

and

$$\alpha_2(\omega;\Omega)=\frac{3\omega_a}{2\varepsilon_0 n_0(\omega)n_0(\Omega)c^2}\Im m\{\chi^{(3)}(\omega,-\Omega,\Omega)\} \qquad [14]$$

We can therefore write the dispersion relations between α_2 and n_2:

$$n_2(\omega;\Omega)=\frac{c}{\pi}\wp\int_0^\infty \frac{\alpha_2(\omega';\Omega)}{\omega' 2-\Omega^2}\mathrm{d}\omega' \qquad [15]$$

Note that even when the degenerate $n_2(\omega)=n_2(\omega;\omega)$ is desired (at a given ω), the dispersion relation requires that we should know the nondegenerate absorption spectrum $\alpha_2(\omega';\omega)$ at all frequencies ω'.

Let us pause here and discuss some physical mechanisms that can be involved for a given system of interest. Consider a material characterized by an optical resonance occurring at, say ω_0 (i.e., a degenerate two-level system). For a solid, this resonance can be regarded as that of the fundamental energy gap; $\omega_0=\omega_g=E_g/\hbar$ in a two-band system. Now, let us examine how the presence of an optical excitation at $\Omega<\omega_0$ can alter the absorption spectrum (at a variable probe ω'). In the quantum mechanical picture, this gives rise to a 'new' material whose perturbed wave functions are 'dressed' by the intensity and frequency of the applied optical field.

The lowest-order correction to the absorption is given by $\alpha_2(\omega';\Omega)$ which involves three major physical processes. Recalling that $\Omega < \omega_0$, these processes include (1) two-photon absorption (2PA) when $\omega' + \Omega \rightarrow \omega_0$ and (2) Raman-induced absorption when $\omega' - \Omega \rightarrow \omega_0$, both implying an absorption of a photon at the probe frequency ω' (i.e., $\alpha_2 > 0$). The third process can be identified as resulting from the blue-shift (for $\Omega < \omega_0$) of the resonance (known as the quadratic optical Stark effect) caused by the excitation field. For our two-level system, the latter results in a decrease followed by an increase in absorption in the vicinity of ω_0. An example of the overall absorption changes due to such processes is shown in Figure 2 where $\alpha_2(\omega';\Omega)$ is qualitatively plotted for a degenerate two-level system. We should note that the relative magnitude of each contribution as well as the width and shape of the resonances are chosen arbitrarily for the purpose of illustration. Using the KK relation in eqn [15], we can now arrive at the nonlinear index coefficient $n_2(\omega;\Omega)$. The result of this transformation is also given in Figure 2. The above simple example elucidates the key concepts involving the relationship between nonlinear absorption and refraction in materials for third-order processes. These concepts, when applied more rigorously to semiconductors, have been successful in predicting the sign, magnitude, and dispersion of n_2 due to the anharmonic motion of bound electrons. This will be briefly discussed later. Returning to the mathematical foundation of KK relations, we use eqns [13] and [14] to write eqn [15] in terms of the nonlinear susceptibility $\chi^{(3)}$:

$$\Re e\{\chi^{(3)}(\omega_1, \omega_2, -\omega_2)\} = \frac{1}{\pi}\wp\int_{-\infty}^{\infty} \frac{\Im m\{\chi^{(3)}(\omega', \omega_2, \omega_2)\}}{\omega' - \omega_1}\, d\omega' \qquad [16]$$

The above dispersion relation for $\chi^{(3)}$ was obtained using the physical and intuitive arguments that followed the linearization scheme depicted in Figure 1. General dispersion relations can be formulated following a mathematical procedure that is similar to the derivation of the linear KK relations. In this case we apply the causality condition directly to the nth-order nonlinear response $R^{(n)}$. For example, without loss of generality, we can write

$$R^{(n)}(\tau_1, \tau_2, \ldots, \tau_n) = R^{(n)}(\tau_1, \tau_2, \ldots, \tau_n)\Theta(\tau_j) \qquad [17]$$

and then calculate the Fourier transform of this equation. Here j can apply to any one of the indices $1, 2, \ldots, n$. Following the same procedure as for a linear response, we obtain

$$\chi^{(n)}(\omega_1, \omega_2, \ldots, \omega_j, \ldots, \omega_n) = \frac{-i}{\pi}\wp\int_{-\infty}^{\infty} \frac{\chi^{(n)}(\omega_1, \omega_2, \ldots, \omega', \ldots, \omega_n)}{\omega_j - \omega'} d\omega' \qquad [18]$$

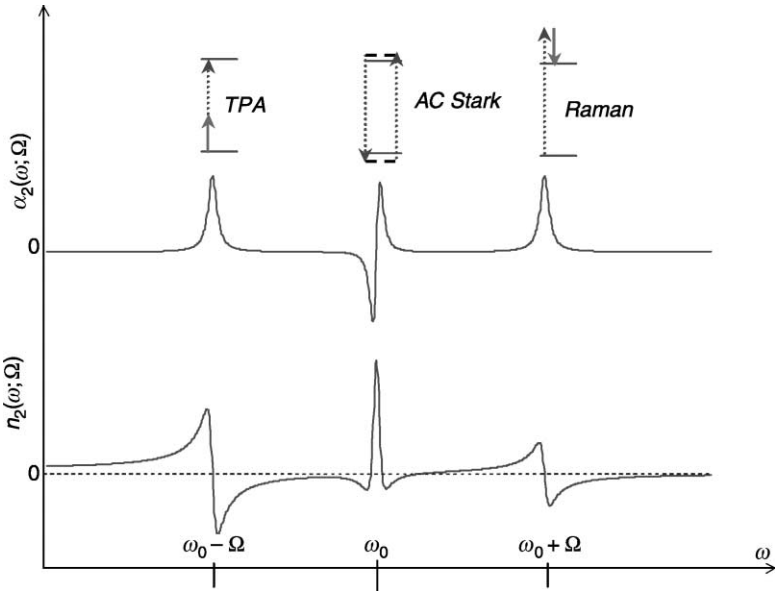


Figure 2 Upper trace: the nonlinear absorption coefficient in a fictitious 'degenerate' two-level system. Lower trace: the resulting nonlinear refractive index obtained using the KK relations. The insets show the three possible physical mechanisms involved.

By separating the real and imaginary parts of this equation, we get the generalized Kramers–Krönig relation pairs for a nondegenerate, nth-order nonlinear susceptibility:

$$\Re e\{\chi^{(n)}(\omega_1, \omega_2, \ldots, \omega_j, \ldots, \omega_n)\} = \frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Im m\{\chi^{(n)}(\omega_1, \omega_2, \ldots, \omega', \ldots, \omega_n)\}}{\omega' - \omega_j}\mathrm{d}\omega' \quad [19]$$

and

$$\Im m\{\chi^{(n)}(\omega_1, \omega_2, \ldots, \omega_j, \ldots, \omega_n)\} = -\frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Re e\{\chi^{(n)}(\omega_1, \omega_2, \ldots, \omega', \ldots, \omega_n)\}}{\omega' - \omega_j}\mathrm{d}\omega' \quad [20]$$

In particular, for $\chi^{(3)}$ processes having $\omega_1 = \omega_a, \omega_2\omega_b$, and ${}_3b$, this becomes identical to eqn [16].

Note that in describing the nonlinear susceptibilities, no special attention was given to the harmonic generating susceptibility $\chi^{(N)}(N\omega) \equiv \chi^{(N)}(\omega, \omega, \ldots\omega)$, i.e., the susceptibility generating the Nth harmonic at $N\omega$. It turns out that in addition to the KK relations given by eqns [19] and [20], the real and imaginary parts of $\chi^{(N)}(N\omega)$ can also be related in different sets of dispersion integrals that involve only the degenerate forms of the susceptibilities. A more general yet simple analysis gives the most general form of KK relations for any type of $\chi^{(n)}$:

$$\chi^{(n)}(\omega_1 + p_1\omega, \omega_2 + p_2\omega, \ldots, \omega_m + p_m\omega) = \frac{1}{\mathrm{i}\pi}\wp\int_{-\infty}^{\infty}\frac{\chi^{(n)}(\omega_1 + p_1\Omega, \omega_2 + p_2\Omega, \ldots, \omega_m + p_m\Omega)}{\Omega - \omega}\mathrm{d}\Omega \quad [21]$$

for all $p_1, p_2, \ldots, p_m \geq 0$. Setting $\omega_1 = \omega_2 = \cdots = \omega_m \equiv 0$, and $p_1 = p_2 = \cdots = p_m = 1$ in eqn [21] yields an interesting form of the KK relations for the Nth-harmonic susceptibilities:

$$\Re e\{\chi^{(N)}(N\omega)\} = \frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Im m\{\chi^{(N)}(N\omega')\}}{\omega' - \omega}\mathrm{d}\omega' \quad [22]$$

These dispersion relations have allowed calculations of $\chi^{(2)}(2\omega)$ and $\chi^{(3)}(3\omega)$ in semiconductors using full band structures.

At the beginning of this article, it was noted that all the KK relations for nonlinear optics were known in the early days of the field. Their application in unifying nonlinear absorption (in particular two-photon absorption) and the optical Kerr effect (n_2) in solids came only much later. More recent work demonstrated that the KK relations are a powerful analytical tool in nonlinear optics. Following the picture of a degenerate two-level system shown in Figure 2, a simple two-band model has been used to calculate the nonlinear absorption coefficient, $\alpha_2(\omega_1; \omega_2)$, resulting from three mechanisms: 2PA, the Raman absorption process, and the ac Stark effect. The optical Kerr coefficient $n_2(\omega_1; \omega_2)$ was then calculated using eqn [15]. Of particular practical interest is the degenerate case ($\omega_1 = \omega_2 = \omega$), from which the 2PA coefficient $\beta(\omega) = \alpha_2(\omega; \omega)$ can be extracted. Figure 3 depicts the calculated dispersion of n_2 and β as a function of $\hbar\omega/E_g$ where E_g is the bandgap energy of the solid. The dispersion of n_2 and its sign reversal shown in Figure 3 has been observed experimentally in many optical solids.

Finally, let us discuss a related implication of causality in nonlinear optics. The KK dispersion relations are traditionally derived in terms of internal material parameters such as susceptibility, absorption coefficient, and refractive index. Similar to the case of electrical circuits, one can obtain dispersion relations that apply to an external transfer function of the system that relates an input signal to an output signal. In this case, the dispersion of the transfer function includes system structure as well as the intrinsic dispersion of the material. As an optical (and linear) example, consider a Fabry–Perot etalon. The optical transmission of this system has well-known spectral

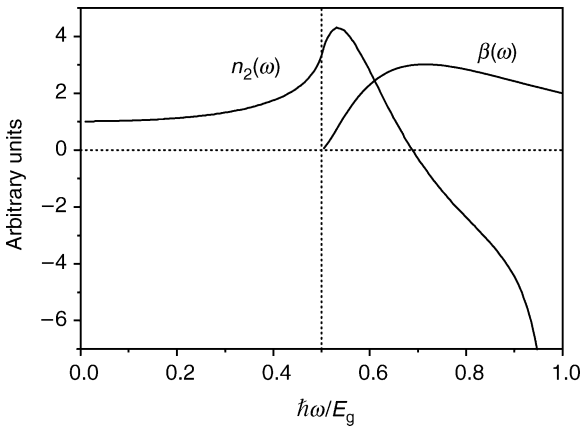


Figure 3 The two-photon absorption coefficient in semiconductors (β) calculated for a two-band model. The resultant nonlinear refractive index (n_2) obtained using a KK transformation of the calculated nondegenerate nonlinear absorption coefficient includes all major mechanisms.

features that are primarily caused by structural dispersion (i.e., interference) in addition to the intrinsic dispersion of the material. Causality still demands that the transmitted signal has a phase variation whose value and dispersion can be determined using a KK relation linking the real and imaginary parts of the transmission coefficient. In other words, the KK relations provide a spectral correlation between the real and imaginary components of the transfer function which in turn may translate to a spectral correlation between the phase and amplitude of the transmitted signal. However, the variations in phase do not necessarily imply the presence of a varying index of refraction, nor does an amplitude variation suggest the existence of material absorption (dissipation). Ultimately, this implies that any mechanism causing a variation in amplitude (including reflection, scattering, or absorption) must be accompanied by a phase variation. (One should note that the reverse of the previous statement is not necessarily true; i.e., a variation in phase does not have to be accompanied by an amplitude modulation.)

In nonlinear optics with the 'black box' approach of Figure 1, the optical perturbation ξ (with frequency Ω) can render an amplitude variation in the probe (at ω) using various frequency mixing schemes in a noncentrosymmetric material (i.e., with nonzero $\chi^{(2)}$). For instance, the probe at ω can be depleted by nonlinear conversion to $\omega_{\text{sum}} = \omega + \Omega$ via sum-frequency generation involving $\chi^{(2)}(\omega, \Omega)$ and/or to $\omega_{\text{diff}} = \omega - \Omega$ via difference-frequency generation involving $\chi^{(2)}(\omega, -\Omega)$. Such a conversion (or depletion) should be accompanied by a phase variation according to the KK dispersion relations. This type of nonlinear phase modulation is known as a $\chi^{(2)}{:}\chi^{(2)}$ cascaded nonlinearity. Such cascaded processes are routinely (and more simply) analyzed with Maxwell's equations governing the propagation of beams in a second-order nonlinear material. The KK relations, however, provide an interesting physical perspective of the process. We find that cascaded second-order nonlinearities are yet another manifestation of causality in nonlinear optics.

List of Units and Nomenclature

α	linear absorption coefficient
α_2	nonlinear absorption coefficient
β	two-photon absorption coefficient
$\chi^{(n)}$	nth-order nonlinear optical susceptibility
n	linear refractive index
n_2	nonlinear refractive index coefficient, coefficient of optical Kerr effect
$\Theta(t)$	step function
$\wp$	principal value
2PA	two-photon absorption
KK relations	Kramers–Krönig relations

See also

Materials for Nonlinear Optics: Liquid Crystals for NLO. **Nonlinear Optics, Basics:** Cascading; Nomenclature and Units.

Further Reading

Bassani F and Scandolo S (1991) Dispersion-relations and sum-rules in nonlinear optics. *Physical Review B–Condensed Matter* 44: 8446–8453.

Caspers PJ (1964) Dispersion relations for nonlinear optics. *Physical Review A* 133: 1249–1251.

Hutchings DC, Sheik-Bahae M, Hagan DJ and Van Stryland EW (1992) Kramers–Krönig relations in nonlinear optics. *Optical and Quantum Electronics* 24: 1–30.

Kogan SM (1963) On the electromagnetics of weakly nonlinear media. *Soviet Physics JETP* 16: 217–219.

Nussenzveig HM (1972) Causality and Dispersion Relations. New York: Academic Press.

Price PJ (1964) Theory of quadratic response functions. *Physical Review* 130: 1792–1797.

Ridener FLJ and Good RHJ (1975) Dispersion relations for nonlinear systems of arbitrary degree. *Physical Review B* 11: 2768–2770.

Sheik-Bahae M and Van Stryland EW (1999) Optical nonlinearities in the transparency region of bulk semiconductors. In: Garmire E and Kost A (eds) *Nonlinear Optics in Semiconductors I*, 58, pp. 257–318. Academic Press.

Sheik-Bahae M, Hutchings DC, Hagan DJ and Van Stryland EW (1991) Dispersion of bound electronic nonlinear refraction in solids. *IEEE Journal of Quantum Electronics* 27: 1296–1309.

Smet F and Vangroenendael A (1979) Dispersion-relations for N-order non-linear phenomena. *Physical Review A* 19: 334–337.

Toll JS (1956) Causality and the dispersion relation: logical foundations. *Physical Review* 104: 1760–1770.

Nomenclature and Units

M P Hasselbeck, The University of New Mexico, Albuquerque, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Nomenclature Associated with the Excitation Light

Particular care must be used when characterizing the excitation beam in nonlinear optical experiments compared to linear measurements. By definition, nonlinear optical phenomena depend on the electric field to high order. The higher the order, the more sensitive the observed behavior depends on the input. Extracting a representative nonlinear coefficient from an experiment, for example, becomes progressively more difficult as the order of the optical nonlinearity increases. In other words: the errors associated with optical beam characterization get magnified by the order of the nonlinearity under study.

There is an extensive nomenclature for characterizing the light (almost always laser light) interacting with the nonlinear optical medium. Different descriptions may be used depending on whether the excitation light is pulsed, continuous, or a continuous train of pulses. When a laser beam is constant or continuous, it is often described by its 'cw power'. 'Cw' stands for 'continuous wave', an acronym taken from the nomenclature of electronics. The cw power of a laser is determined by placing a power meter in the path of a beam. Repetitively pulsed lasers can be characterized in the same way, provided the response time of the power meter is slower than the pulse separation period. This will almost certainly be the case with a continuously pumped mode-locked laser such as a dye laser, fiber laser, or Ti:sapphire laser, which produce pulses at repetition frequencies of tens of megahertz. Cw power may also be suitable for describing pulsed flashlamp lasers, gas discharge lasers, or any laser that is excited in a periodic fashion.

When the optical output can be distinguished as individual pulses, additional metrics are used. A pulse, by definition, exists during a window of time, i.e., there is a time when light is present and a time when light is absent. How one characterizes the time light is present – the pulse duration – is critically important and not always obvious. There are a myriad of ways the pulse temporal envelope can manifest itself; common examples include square pulses, triangular pulses, Gaussian pulses, and hyperbolic secant pulses. These names refer to the mathematical waveforms that map the pulse envelope as a function of time.

A pulse waveform that exhibits symmetry about the peak in its temporal envelope is commonly characterized by its 'full-width, half-maximum', abbreviated FWHM. One obtains the full-width, half-maximum by locating the two points on the pulse profile that are at half the peak value. The temporal separation of these two points is the FWHM. One uses this measure because in a strict mathematical sense, waveforms such as the Gaussian or hyperbolic-secant exist even at times $t = \pm\infty$. A less common term is the 'half-width, half-maximum' or HWHM. As the name implies, the HWHM is exactly half the FWHM value. Not all light-pulses are temporally symmetric, however, so greater detail may be needed when giving a mathematical description of asymmetric pulses.

Light pulses are also characterized by their energy and peak power. The temporal envelope recorded by a so-called square-law detector (all laboratory detectors are square-law detectors) shows the power of the pulse as a function of time, provided the detector response time is sufficiently fast. Integrating this waveform gives the pulse energy. It makes no sense to talk about the energy of a cw beam of light, unless that beam is composed of repetitive, distinguishable pulses. Each pulse in the train carries a distinct amount of optical energy.

'Ultrafast' or 'ultrashort' laser pulses generally refer to light pulses that are of such small duration they cannot be measured directly with detectors and oscilloscopes because of bandwidth limitations. To infer the pulse duration, so-called intensity autocorrelation measurements are often made. The temporal power profile is then deduced from the autocorrelation signal with the appropriate mathematical conversion factor. Equally important is the spectrum of a short pulse. The optical bandwidth determines the lower limit of temporal compression; such ideally compressed pulses are said to be 'transform limited'. If different components of the spectrum can be distinguished at specific positions on the pulse temporal profile, then the pulse is 'chirped'. Information on the methods of ultrafast laser science can be found in **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Nonlinear optical effects take place when light is concentrated on a target – there is a cross-section or 'footprint' of the beam on the medium.

The interaction of the light and material takes place in a region called the beam area, beam cross-section, focal area, or focal volume. One must then define the appropriate interaction area or volume.

Specification of the beam area allows one to define two important quantities used frequently in nonlinear optics. The power/area is known as irradiance or power density. This quantity is commonly called the intensity, although the strict radiometric definition of intensity is power/(solid angle). The second important quantity is energy/area, which is called the energy density or fluence.

The convention in nonlinear optics is to define the optical electric field as:

$$\mathbf{E}(\mathbf{z}, t) = \frac{1}{2} E_0 \exp[\mathrm{i}(\mathbf{k}\cdot\mathbf{z} - \omega t)] + \text{c.c.} \qquad [1]$$

where E_0 is the peak field, $\mathbf{k}$ is the propagation vector, ω is the angular frequency of the light, and c.c. stands for complex conjugate. This represents a forward- and backward-propagating (in the z-direction) infinite, transverse plane wave. Other definitions are also used. In SI/mks units and using the convention of eqn [1], the irradiance (I in units of W/m^2) inside a material of refractive index n is related to the electric field vector of the light ($\mathbf{E}$ in units of V/m) as follows:

$$I = \frac{1}{2} cn\varepsilon_0 |\mathbf{E}|^2 \qquad [2]$$

Here c is the speed of light (2.998×10^8 m/s) and ε_0 is the permittivity of free space (8.854×10^{-12} F/m). Nonlinear optics has an unfortunate tradition of mixing mks and cgs units, resulting in a lot of confusion. Irradiance is usually expressed in units of W/cm^2. One can calculate the peak electric field (in units of V/cm) of a laser beam given its irradiance (in units of W/cm^2) by using this simple formula:

$$E_0 = 38.82\sqrt{\frac{I}{n}} \left(\frac{\mathrm{V}}{\mathrm{cm}}\right) \qquad [3]$$

The root-mean-square value of the electric field is obtained by replacing the factor 38.82 by 27.45. In Gaussian/cgs units, irradiance is expressed in units of ergs/(cm^2 sec) and is related to the field as:

$$I = \frac{cn}{8\pi} |\mathbf{E}|^2 \qquad [4]$$

where the field is in units of statvolt/cm and $c = 2.998 \times 10^{10}$ cm/sec. Conversion between mks and cgs for the electric field is 3×10^4 V/m = 1 statvolt/cm; irradiance is converted using 1 erg/(cm^{2}sec) = 1×10^{-3}W/m^2. The following point must be emphasized: depending on how the electric field is defined, there can be factors of 2 or even 4 discrepancies in the irradiance values quoted by different authors. The common definitions are used here, although they are certainly not universal.

A very common realization of the spatial irradiance profile of a laser beam is the radially symmetric Gaussian function:

$$I(r) = I_0 \exp\left(-\frac{2r^2}{w^2}\right) \qquad [5]$$

where I_0 is the peak irradiance at the center of the beam ($r = 0$) and w is the 'spot size'. The spot size changes continuously as the beam propagates and the minimum spot size is known as the waist. One often hears the spot size referred to as the radius, but the radius of a Gaussian beam traditionally denotes the curvature of the phase front. By definition, the phase front radius is infinite at the waist. The factor of 2 appearing in the argument of the exponential in eqn [5] stems from the fact that the spot size w is conventionally defined for the electric field of the Gaussian beam. When the field is squared to obtain the irradiance, the factor of 2 appears. Using this function for the irradiance spatial profile leads to the common parameter '1/e^2 diameter'. The irradiance falls to 1/e^2 (0.1353) of its peak value when $r = w$.

The Gaussian spatial profile or 'Gaussian beam' results when the laser has been designed to operate in the lowest-order transverse mode, usually denoted TEM$_{00}$. The TEM$_{00}$ Gaussian beam is particularly convenient because the same relative power profile is maintained in the near- and far-field, whether or not the beam is collimated or focused.

The cross-sectional area of a TEM$_{00}$ Gaussian beam normally incident on a surface is found as follows. The power in the beam is obtained by integrating the irradiance profile (I) over the surface:

$$P = \iint_{\pm\infty} I \mathrm{d}A \qquad [6]$$

Referring to eqn [5], the integral becomes:

$$P = \int_0^{2\pi} \mathrm{d}\theta \int_0^{\infty} r\,\mathrm{d}r\, I_0 \exp\left(-\frac{2r^2}{w^2}\right) = I_0 \frac{\pi w^2}{2} \qquad [7]$$

Hence the effective area of a TEM$_{00}$ Gaussian beam normally incident on a surface is:

$$A = \frac{\pi w^2}{2} \qquad [8]$$

There are many situations where the Gaussian formulation of eqn [5] is not suitable. Beams produced by unstable laser resonators, for example, are not Gaussian. A multitransverse mode beam profile obtained from a stable resonator does not usually lend itself to a simple mathematical characterization.

Nomenclature Associated with the Nonlinear Optical Medium

In nonlinear optics texts, the subject is often introduced by writing the macroscopic polarization in Maxwell's equations as a power series expansion in the electric field. This approach, proposed by Bloembergen and coworkers in the early 1960s, has been spectacularly successful for interpreting experiments, though it has also led to confusion in the definition of nonlinear optical coefficients. The confusion, which stems from arbitrary definitions and nomenclature used by different authors, cannot be resolved here. The reader should, however, be alert for these discrepancies. Comparison of results published by different laboratories requires that the fundamental equations for extracting the nonlinear coefficients from their data are known. As the discipline has matured, the nonlinear optics community largely recognized the ambiguities and confusion; precise definitions of terms and coefficients are now commonly provided in the research literature.

It should also be noted that the power-series framework may not always be the best formulation for modeling and characterizing nonlinear optical effects. A carefully designed optical limiter, for example, may exhibit an abrupt decrease of transmission at a specific 'threshold irradiance' or 'threshold fluence', perhaps at a given laser pulse duration. The physical phenomenon or phenomena driving this behavior may not readily succumb to characterization by an nth-order coefficient in a power series expansion. In this situation it may be appropriate to specify a threshold optical input parameter. Another example is a homogeneously broadened saturable absorber, in which the absorption of an optical medium decreases with increasing input irradiance. The irradiance-dependent absorption coefficient α (refer to eqn [26]) is characterized by a 'saturation irradiance' I_{sat}:

$$\alpha(I) = \frac{\alpha_0}{1 + \dfrac{I}{I_{\text{sat}}}} \qquad [9]$$

where α_0 is the linear absorption coefficient.

Nonlinear Susceptibility

The power-series expansion of the macroscopic polarization is the standard approach for modeling nonlinear optical behavior and categorizing the various phenomena. A common way to write the polarization, nonlinear in the electric field, is (mks units):

$$\mathbf{P} = \varepsilon_0\left(\chi^{(1)}\cdot\mathbf{E} + \chi^{(2)} : \mathbf{EE} + \chi^{(3)}\vdots\mathbf{EEE} + \ldots\right) \qquad [10]$$

where the polarization $\mathbf{P}$ is a time- and space-dependent vector and the terms $\chi^{(n)}$ are the various orders of the nonlinear susceptibility (the cgs equation is obtained by dropping the free-space permitivitty coefficient ε_0). A second-order effect is associated with $\chi^{(2)}$, a third-order with $\chi^{(3)}$, and so on. In general, there are distinguishable electric field vectors (in eqn [10], the fields have not been individually designated for clarity). The susceptibility terms are generally tensors, which means the medium is sensitive to the orientation of the input fields. The dots in eqn [10] indicate tensor products. The input field vectors can have different frequencies and the presence of complex conjugates in eqn [1] indicates that the frequency components will have both + and − signs, i.e. the jth electric field term will have associated frequency factors $\exp(\pm \mathrm{i}\omega_j t)$.

The nomenclature used here can be illustrated by example. In the case of the second-order nonlinear polarization (mks units):

$$\mathbf{P}_j = \varepsilon_0 \chi^{(2)}_{jkl} : \mathbf{E}_k\mathbf{E}_l \qquad [11]$$

Note that subscripts have been introduced. The indices correspond to Cartesian space vectors. This equation says that the polarization in the j-direction results from the tensor product of the appropriate $\chi^{(2)}$ with the input fields at $\mathbf{E}_k$ and $\mathbf{E}_l$. The jth component of polarization has an oscillation frequency that is determined by mixing of the input fields. In this example, the indices j, k and l can each take the value of x, y, and z. Let each input field be at one of two possible frequencies, call them ω_1 and ω_2. There are an enormous number of permutations (81 to be exact) of eqn [11]. One term is:

$$P_x(\omega_1 + \omega_2) = \varepsilon_0 \chi^{(2)}_{xzy}(\omega_1 + \omega_2)E_z(\omega_1)E_y(\omega_2) \times \exp[-\mathrm{i}(\omega_1 + \omega_2)t] \qquad [12]$$

Another possibility is:

$$P_z(\omega_2 - \omega_1) = \varepsilon_0 \chi^{(2)}_{zyx}(\omega_2 - \omega_1)E_y(\omega_2)E_x(-\omega_1) \times \exp[-\mathrm{i}(\omega_2 - \omega_1)t] \qquad [13]$$

Equation [12] corresponds to sum frequency generation ($\omega_1 + \omega_2$) and eqn [13] shows difference frequency generation ($\omega_2 - \omega_1$).

Not explicitly written on the right-hand side of the above equations is the exponential containing the mixing of the propagation vectors. To maximize the nonlinear polarization, which is the source of nonlinear behavior in Maxwell's equations, the vector sum $\sum \mathbf{k}_j$ for *all* the interacting fields should be close to zero. Arranging the propagation vectors to accomplish this is known as 'phase matching'.

It is important to point out that there are eight more terms describing the nonlinear polarization at $P_x(\omega_1 + \omega_2)$ in addition to eqn [12]; likewise for $P_z(\omega_2 - \omega_1)$. Also note that the indices k and l may be identical, hence a nonlinear polarization driven by $\chi^{(2)}_{xyy}$ for example, is allowed. Higher-order nonlinear polarizations get progressively more complicated. The general third-order nonlinear polarization is

$$\mathbf{P}_j = \varepsilon_0 \chi^{(3)}_{jklm} \vdots \mathbf{E}_k \mathbf{E}_l \mathbf{E}_m \qquad [14]$$

Sometimes the second-order nonlinear susceptibility is written with the coefficient d instead of $\chi^{(2)}$. The relation between the d-coefficient and $\chi^{(2)}$ depends arbitrarily on how it is defined; a common definition is

$$d_{jkl} = \frac{1}{2}\chi^{(2)}_{jkl} \qquad [15]$$

but the reader is cautioned that the factor $\frac{1}{2}$ is sometimes missing. In the preceding discussion it was pointed out how there can, in principle, be a large number of tensor components involved in the expansion of the nonlinear polarization. Simplification occurs when it is realized that there is no discernible physical difference between the frequency terms ω_j and $-\omega_j$ and that ordering of the fields in eqn [14] – which suggests a time order in the arrival of the fields – is irrelevant. These symmetry arguments show that $d_{jkl} = d_{jlk}$, which reduces the number of independent d-coefficients from 81 to 18. A simpler subscript notation is then introduced that uses the integers 1–6 to represent pairs of Cartesian components k and l. This reduces the number of subscripts from three to two. An example of this notation is $d_{xzz} = d_{14}$.

In certain situations, one can take advantage of crystal symmetry to further reduce the complicated summations to a single scalar coefficient for the nonlinear susceptibility, which is referred to as 'd-effective'. For these specialized cases, the nonlinear polarization is

$$P = \varepsilon_0 d_{\text{eff}} E_1 E_2 \qquad [16]$$

In SI/mks units, the polarization is in units of coul/m^2. The second-order susceptibility (i.e. $\chi^{(2)}$, d_{jkl}) must therefore have units of m/volt. The units of $\chi^{(3)}$ are m^2/volt^2, $\chi^{(4)}$ is m^3/volt^3, and so on. In Gaussian/cgs units, $\chi^{(2)}$ has dimensions cm/statvolt, $\chi^{(3)}$ will be cm^2/statvolt^2, etc. although sometimes in the Gaussian system *all* the coefficients $\chi^{(n)}$ are discussed with shorthand 'electrostatic units' or 'esu'.

Complex Quantities

In linear optics, the susceptiblity has real and imaginary parts

$$\chi^{(1)} = \chi^{(1)}_{\text{real}} + \mathrm{i}\chi^{(1)}_{\text{imaginary}} \qquad [17]$$

The complex linear index is, in turn, written as the sum of real and imaginary components, derived from the linear susceptibility as follows (mks units):

$$\sqrt{1 + \chi^{(1)}} = n_0 + \mathrm{i}\kappa \qquad [18]$$

where n_0 is the linear refractive index and κ is the imaginary term leading to absorption of light. In general, the nonlinear susceptibilty $\chi^{(n)}$ is also a complex number:

$$\chi^{(n)} = \chi^{(n)}_{\text{real}} + \mathrm{i}\chi^{(n)}_{\text{imaginary}} \qquad [19]$$

As in linear optics, the complex notation is a bookkeeping method that conveniently accounts for what is known as 'resonant enhancement'. Nonlinear optics research has revealed that the nonlinear susceptibility $\chi^{(n)}$ can be a strong function of frequency. Specifically, this quantity will be strongly enhanced when sums and/or differences of the photon energies in the interacting light beams coincide with quantum mechanical energy resonances in the material. When a resonance condition is achieved, $\chi^{(n)}$ will be dominated by its imaginary component. Far from the resonances, $\chi^{(n)}$ behaves more like a real quantity. These complex terms directly enter the wave equations describing how light propagates in a nonlinear medium and are particularly important for $\chi^{(3)}$; the complex quantity $\chi^{(3)}$ determines whether light will be refracted or absorbed and by how much. The real part of $\chi^{(3)}$ drives nonlinear refraction while the imaginary portion characterizes nonlinear absorption (e.g., two-photon absorption) and the inverse effect – gain. This rich subject – in particular the critical importance of the wavelength dependence of the nonlinear phenomena – is considered in **Nonlinear Optics, Basics:** Kramers–Krönig Relations in Nonlinear Optics.

Nonlinear Refraction

The most common manifestation of nonlinear refraction arises from the third-order nonlinear susceptibility, the so-called optical Kerr effect. In this case, the refractive index is linearly proportional to the irradiance (I) of a monochromatic light beam. The irradiance-dependent refractive index is

$$n(I) = n_0 + n_2 I \qquad [20]$$

where n_0 is the linear, irradiance-independent refractive index and n_2 is the nonlinear refractive index coefficient. Because n is a dimensionless quantity, n_2 must be in units of area/power.

The optical Kerr effect is usually written in mks units with the expression shown in eqn [20]. In cgs units, the common (but by no means universal) convention is to write:

$$n(E) = n_0 + \frac{1}{2}\tilde{n}_2|\mathbf{E}|^2 \qquad [21]$$

where the field is in units of statvolts/cm. The nonlinear coefficient $\tilde{n}_2$ (cgs) must therefore be in units of cm^2/statvolt^2, which is sometimes abbreviated to 'esu'. The conversion for n_2 between these two equations is:

$$\tilde{n}_2(\text{cgs}) = \frac{n_0 c}{40\pi} n_2(\text{mks}) \qquad [22]$$

where n_2(mks) is in units of m^2/W and c is in m/sec. Also useful is the relation between n_2 and $\chi^{(3)}$, written in mks units as:

$$\Delta n = n_2 I = \frac{\text{Re}\left(\chi^{(3)}\right)}{4\varepsilon_0 n^2 c} I \qquad [23]$$

where 'Re' denotes the real part of $\chi^{(3)}$. In cgs units we have:

$$\Delta n = \frac{1}{2}\tilde{n}_2|\mathbf{E}|^2 = \frac{4\pi^2 \text{Re}\left(\tilde{\chi}^{(3)}\right)}{n^2 c} I \qquad [24]$$

where I is in units of ergs/(cm^2 sec) defined by eqn [4]. The reader is again advised that different authors will show discrepancies of 2, 4, or even 8 when writing these equations. The susceptibilities $\chi^{(3)}$ are related as follows:

$$\tilde{\chi}^{(3)}(\text{cgs}) = \frac{9 \times 10^8}{4\pi} \chi^{(3)}(\text{mks}) \qquad [25]$$

If the coefficients n_0 and n_2 exist, does the nomenclature imply there are terms n_1, n_3, and others? These coefficients are certainly allowed, but not often seen in discussions of nonlinear refraction. When one refers to nonlinear refraction, the common understanding is that the index depends linearly on irradiance, which is conveniently modeled by eqn [20]. But there are other physically relevant situations to consider. The coefficient n_1, for example, describes the linear electro-optic effect in which the change in index is linearly proportional to the electric field (although it is called the linear electro-optic effect, it is actually derived from the second-order nonlinear susceptibility). The electric field can be the oscillating field of a laser beam, for example, or a dc field applied to an electro-optic crystal (e.g., Pockel's cell). When other terms are added to eqns [20] or [21], the units of the coefficients must be chosen to keep the equation dimensionless.

In the preceding discussion, there is an implication that the nonlinear index is an instantaneous function of the irradiance or field. Although the material system can never respond instantaneously, this is a good approximation in many situations. Sometimes it is not. Consider a pulsed laser beam that heats the material. When the local temperature increases, the linear refractive index can change. A short, Q-switched laser can have a pulse duration far less than a microsecond, while the temperature change it induces can last many orders of magnitude longer. This means optical modification of the refractive index may persist long after the exciting pulse has vanished, i.e., when the irradiance is at zero.

Another example is when a laser beam promotes electrons from their ground state to higher energy states of the material system. These excited electrons may modify the refractive index of the material. In general, the excited electrons remain in high-energy states for some period of time before relaxing to the ground level – if this time is longer than the excitation, the change of the refractive index persists beyond the time the laser beam is present. In these situations, the model of nonlinear refraction suggested by the above equations is inaccurate. One cannot obtain the nonlinear refraction from a simple algebraic analysis. A system of time-dependent equations – describing the dynamics of excitation and relaxation – must be solved using numerical procedures. Such phenomena are sometimes loosely categorized as 'dynamic linear optical effects' or 'effective third-order nonlinearities'.

Nonlinear Absorption

Consider a single-frequency light beam passing through an optically absorbing region of length L.

For simplicity, neglect reflections caused by surfaces that may define the region of interest, i.e., ignore reflections at surfaces that may be located on the optical axis z at points $z = 0$, $z = L$, or any other point in the path. Linear absorption means that the optical power extracted from the light beam as it traverses the absorbing medium is a direct function of the power at a given point. This is described mathematically by an elementary, linear differential equation known as Beer's law:

$$\frac{\mathrm{d}}{\mathrm{d}z} I(z) = -\alpha I(z) \qquad [26]$$

The constant of proportionality is α, which is the linear absorption coefficient. Equation [23] is solved by direct integration:

$$\int_{I(0)}^{I(L)} \frac{\mathrm{d}I}{I} = -\alpha \int_0^L \mathrm{d}z \qquad [27]$$

This has the solution:

$$\frac{I(L)}{I(0)} = \exp(-\alpha L) \qquad [28]$$

which means the irradiance decreases exponentially as a function of propagation distance in a linearly absorbing medium.

In the nonlinear regime, we don't expect a classical Beer's law model to hold. By definition, the absorption will be a nonlinear function of irradiance at a given point. One makes the following power-series expansion to describe nonlinear absorption of monochromatic light:

$$\frac{\mathrm{d}}{\mathrm{d}z} I = -\alpha I - \beta I^2 - \gamma I^3 - \cdots \qquad [29]$$

The coefficient β corresponds to a two-photon absorption process and γ is the coefficient of three-photon absorption. What about four-photon and even higher-order processes? These are rarely encountered, but certainly possible. To make these distinctions, eqn [29] is sometimes written with coefficients K_i or α_i instead of the sequential Greek alphabet:

$$\frac{\mathrm{d}}{\mathrm{d}z} I = -K_1 I - K_2 I^2 - K_3 I^3 - K_4 I^4 - \cdots \qquad [30]$$

The units of the various absorption coefficients must maintain the dimensional consistency of eqns [29] and [30], which is irradiance/length or power/(length)^3. These are shown in Table 1.

If there is linear absorption, the nonlinear processes at that wavelength are often (but not always) negligibly weak. This means we have the following inequalities:

$$K_1 I \gg K_i I^i, \qquad \text{where } i \geq 2 \qquad [31]$$

Nonlinear absorption is generally observed in the wavelength region where the medium is transparent to low-irradiance light, i.e., where linear absorption is negligible. It should be emphasized that there are situations where linear absorption is large and in fact a crucial component of the aggregate nonlinear effect. We will return to this point later in the discussion. For the moment, we neglect linear absorption. If the energy of the incident photons and energy levels of the system permit it, the lowest-order nonlinear process is two-photon absorption, described by the equation

$$\frac{\mathrm{d}}{\mathrm{d}z} I = -K_2 I^2 \qquad [32]$$

which can be solved by elementary integration:

$$\frac{I(L)}{I(0)} = \frac{1}{1 + I(0) K_2 L} \qquad [33]$$

Unfortunately, the situation is rarely this convenient. The above integration has ignored the fact that nonlinear absorption may enable significant linear absorption. This is possible because nonlinear absorption promotes electrons from low- to high-energy states in the medium. The change in excited electron density associated with absorption of light (both linearly and nonlinearly) is called 'photocarrier generation'. These photocarriers (e.g., photo-electrons) may modify the linear absorption as well as the refractive properties of the material. If two-photon absorption causes the linear absorption to increase, for example, the assumption that allowed us to ignore K_1 when writing eqn [32] ceases to be valid. While eqn [32] was appropriate at the very start of the light–matter interaction, the generation of photocarriers after the passage of time may change that condition. The behavior of the system is therefore time-dependent, i.e., it is dynamic. Simple analytic solutions are almost

Table 1 Dimensions of optical absorption coefficients

Process	*Coefficient*	*Units*
Linear absorption	α, K_1	(length)$^{-1}$
Two-photon absorption	β, K_2	length/power
Three-photon absorption	γ, K_3	length^3/power^2
Four-photon absorption	K_4	length^5/power^3

always not possible when dealing with nonlinear absorption.

The time rate of change of the photocarrier population (N) is modeled by the following equation:

$$\frac{\partial}{\partial t}N = \frac{K_1 I}{\hbar\omega} + \frac{K_2 I^2}{2\hbar\omega} + \frac{K_3 I^3}{3\hbar\omega} + \frac{K_4 I^4}{4\hbar\omega} + \cdots - \text{recombination} - \text{diffusion} \quad [34]$$

where ω is the angular frequency ($\omega = 2\pi f$) of the incident light. Also included are place-holders for loss processes such as recombination and diffusion, which may themselves have complicated mathematical descriptions. The coefficients K_i have exactly the same units, dimensions, and interpretation as in eqn [30] and Table 1; in the spirit of the preceding discussion, these coefficients may be time dependent. Note that the dimensions of eqn [30] describe absorption of light (irradiance/length), while eqn [34] models the photocarrier population density (length^{-3} time^{-1}). The units of these two equations are very different.

List of Units and Nomenclature

CARS	Coherent anti-Stokes Raman scattering
c.c.	Complex conjugate
CSRS	Coherent Stokes Raman scattering
DFG	Difference frequency generation
DFM	Difference frequency mixing
DFWM, FWM	(Degenerate) Four wave mixing
DRO	Doubly resonant OPO
EIT	Electromagnetically induced transparency
ESA	Excited state absorption
GVD	Group velocity dispersion
GVM	Group velocity mismatch
NLA	Nonlinear absorption
NLR	Nonlinear refraction
OFID	Optical free-induction decay
OPA	Optical parametric amplification/amplifier
OPG	Optical parametric generation/generator
OPL	Optical power limiter
OPO	Optical parametric oscillation/oscillator
2PA, TPA	Two-photon absorption
PCM	Phase conjugate mirror
QPM	Quasi phase matching
RIKES	Raman-induced Kerr effect spectroscopy
RSA	Reverse saturable absorption
SBS	Stimulated Brillouin scattering
SFG	Sum frequency generation
SHG	Second harmonic generation
SIT	Self-induced transparency
SRO	Singly resonant OPO
SRS	Stimulated Raman scattering
SRWS	Stimulated Rayleigh wing scattering
SVEA, SVAA	Slowly varying envelope (amplitude) approximation
THG	Third harmonic generation
TPF	Two-photon fluorescence

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Nonlinear Optics, Basics:** Ultrafast and Intense-Field Nonlinear Optics. **Spectroscopy:** Nonlinear Laser Spectroscopy.

Further Reading

Bloembergen N (1992) *Nonlinear Optics*. Redwood City, CA: Addison-Wesley.

Boyd RW (1991) *Nonlinear Optics*. Boston, MA: Academic Press.

Levenson MD and Kano S (1988) *Introduction to Nonlinear Laser Spectroscopy*. Boston, MA: Academic Press.

Miller A, Miller DAB and Smith SD (1981) Dynamic nonlinear optical processes in semiconductors. *Advances in Physics* 30: 697–800.

Mills DL (1998) *Nonlinear Optics: Basic Concepts*. New York: Springer.

Newell C and Moloney JV (1992) *Nonlinear Optics*. Redwood City, CA: Addison-Wesley.

Sauter EG (1996) *Nonlinear Optics*. New York: Wiley.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: Wiley.

Sutherland RL (1996) *Handbook of Nonlinear Optics*. New York: Marcel Dekker.

Yariv A (1991) *Optical Electronics*, 3rd edn. New York: Holt, Rinehart, and Winston.

Yariv A (1989) *Quantum Electronics*. New York: Wiley.

Zernike F and Midwinter JE (1973) *Applied Nonlinear Optics*. New York: Wiley.

Nonlinear Optical Phase Conjugation

B Y Zeldovich, University of Central Florida, Orlando, FL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Phase conjugation (PC) beams and their applications are best illustrated by Figure 1, which shows the passage of a coherent planewave incident upon a transparent optical element with inhomogeneities of refractive index (Figure 1a). Reversibility of light propagation implies the existence of such an 'anti-distorted' beam, which reverts back to a planewave after reversing through the same inhomogeneities (Figure 1b). A real-valued monochromatic optical wave, $E_{\text{real}}(\mathbf{R},t)$, may be represented by the complex amplitude $E(\mathbf{R})$ via $E_{\text{real}}(\mathbf{R},t) = 0.5[E(\mathbf{R})\exp(-i\omega t) + E^*(\mathbf{R})\exp(i\omega t)]$, where $E^*(\mathbf{R})$ represents the complex conjugate of $E(\mathbf{R})$:

$$E(\mathbf{R}) = |E(\mathbf{R})|\exp[i\varphi(\mathbf{R})]$$
$$E^*(\mathbf{R}) = |E(\mathbf{R})|\exp[-i\varphi(\mathbf{R})] \qquad [1]$$

Mathematical expression of time-reversal, $E_{\text{real}}(\mathbf{R},t) \rightarrow E_{\text{real}}(\mathbf{R},-t)$, becomes the exchange $E(\mathbf{R}) \leftrightarrow E^*(\mathbf{R})$ for monochromatic waves. The reversibility of propagation means that the complex propagating field $E(\mathbf{R})$ (for example, planewave $\exp(i\mathbf{k}\cdot\mathbf{R})$), is a wave equation solution equivalent to $E^*(\mathbf{R})$ (in this example, $\exp(-i\mathbf{k}\cdot\mathbf{R})$). Conjugation of complex amplitude means reversal of the sign of the phase, and a mixed label PC has been coined and nowadays is firmly established. Terms such as 'wave front reversal' and 'generation of time-reversed replica of the beam' are also used to describe PC.

Figure 1 also illustrates one of the most important applications of PC. If the element in question is a laser-type amplifier, then double-passage allows the extraction of energy from the optically inhomogeneous laser medium in the form of a perfectly collimated beam of diffraction-limited divergence. A PC device may also serve as a mirror of a laser resonator, resulting in the same beam-correction purpose.

One of the most practical and robust methods of PC is based on stimulated Brillouin back-scattering (SBS). The incident beam illuminates the SBS-active transparent medium, for example, liquids (CS_2, CCl_4, acetone, etc.), compressed gases (CH_4, SF_6, etc.), or solids (fused or crystalline quartz, glass, etc.). This 'pump' beam, $E(\mathbf{R}) \equiv E(\mathbf{r},z)$, must possess well-developed transverse inhomogeneities of intensity $|E(\mathbf{r},z)|^2$. Here, $\mathbf{r} = \{x,y\}$ is the part of coordinate vector transverse with respect to the central direction z of the beam in question, and $\mathbf{R} = \{\mathbf{r},z\}$. These inhomogeneities may constitute a narrow focal waist in the case of weakly distorted focused beams, speckle-inhomogeneities in cases of strong distortions, or a combination thereof; see Figure 2, where solid lines symbolize the 'rays' of the incident pump.

Spontaneous scattering of the pump results in a multitude of possible transverse profiles $S(\mathbf{r}, z=0)$ of the signal, whose 'rays' are depicted via the dotted lines on Figure 2. The signal is amplified exponentially due to SBS processes. This exponent in a simplified form may be represented as $|S(\mathbf{r},z)|^2 \propto \exp(\int g dz)$, with the gain $g(z)$ being the result of transverse overlapping of the intensity profiles of signal and pump:

$$g(z) \approx \text{const} \cdot \langle |E(\mathbf{r},z)|^2 \cdot |S(\mathbf{r},z)|^2 \rangle / \langle |S(\mathbf{r},z)|^2 \rangle \qquad [2]$$

Thus, the similarity of the intensity profile of the signal to that of the pump is rewarded by the exponential preference in the output signal level. However, this is not enough to guarantee that the output backward-scattered signal $S(\mathbf{r},z)$ is phase

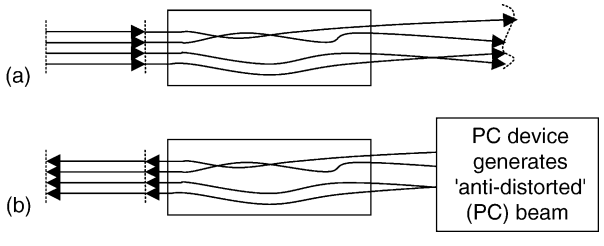


Figure 1 (a) Collimated beam is distorted by propagation through inhomogeneous medium. (b) Conjugate, a.k.a. antidistorted beam becomes collimated after backward passage through the same medium.

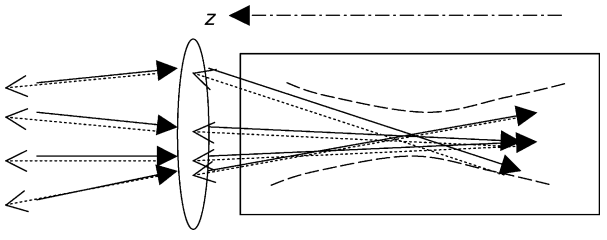


Figure 2 Stimulated Brillouin Scattering (SBS) method of PC. Solid lines symbolize the rays of incident 'beam', while dotted lines depict the rays of the signal amplified by the SBS process.

conjugate with respect to the pump. It is here that the mutual reversibility of propagation of strongly inhomogeneous signal and pump becomes crucial. Namely, good transverse overlapping is sustained along the entire interaction length z, only if $S(\mathbf{r}, z) \propto E^*(\mathbf{r}, z)$. This constitutes the 'discrimination mechanism' of SBS-PC: the 'conjugate mode' of the signal has a z-sustained advantage in exponential gain and, under appropriate conditions, suppresses all nonconjugate signal configurations in competition for pump power. Nonlinear-optical wave theory of this discrimination process has been developed further.

The holographic mechanism of PC has a different nature. It may take the form of static holography or dynamic holography; the latter essentially may be considered as a part of nonlinear optics. Here is a simplified description of one of the variants. Suppose one wants to obtain a conjugate replica of the incident monochromatic signal, whose complex profile is $S(\mathbf{R})$. A plane reference wave $A(\mathbf{R}) = \exp(i\mathbf{k}_A \cdot \mathbf{R})$ is also directed to the registering medium. If $S(\mathbf{R})$ and $A(\mathbf{R})$ are mutually coherent, then an interference pattern of intensity is produced $\{S^*(\mathbf{R})A(\mathbf{R}) + \text{compl. conj.}\}$. The medium records this pattern in the form of the modulation of refractive index and/or of the absorption coefficient, $\delta\varepsilon(\mathbf{R}) \propto (c_1 + ic_2)\{S^*(\mathbf{R})A\ (\mathbf{R}) + \text{c.c.}\}$ (Figure 3a). At the 'reconstruction' stage, the hologram is illuminated by another plane reference wave, $B(\mathbf{R}) = B_0 \exp(i\mathbf{k}_B \cdot \mathbf{R})$. In the specific case, when $\mathbf{k}_B = -\mathbf{k}_A$, the source δD_{conj} of the reconstructed wave becomes, as shown by Figure 3b:

$$\delta D_{\text{conj}} = (c_1 + ic_2)[S^*(\mathbf{R})A(\mathbf{R})] \cdot B(\mathbf{R}) \propto S^*(\mathbf{R}) \times (c_1 + ic_2)A_0B_0 \qquad [3]$$

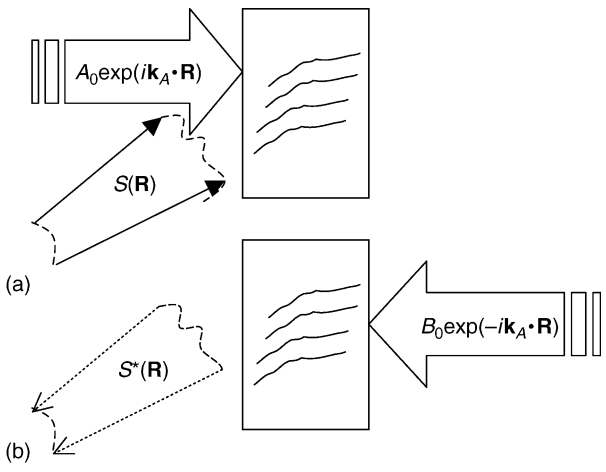


Figure 3 (a) Recording of a hologram in a photosensitive medium by a reference wave A and signal wave S. (b) Read-out of that hologram by the wave B counter-propagating to A results in generation of the conjugated signal S^*.

Reversibility of propagation laws guarantees that this source will indeed excite a conjugate wave with high efficiency. Both processes, recording of the interference pattern and reconstruction of the conjugate wave, may occur simultaneously, if one deals with the dynamic holography. The same process may also be described as nonlinear-optical four-wave mixing (FWM) via cubic nonlinearity $\chi^{(3)}$:

$$\delta D_{\text{conj}}(\mathbf{R}) = \chi^{(3)}S^*(\mathbf{R})A(\mathbf{R})B(\mathbf{R}) = \chi^{(3)}S^*(\mathbf{R})A_0B_0 \qquad [4]$$

with the four waves; A, B, S, S^*.

Important characteristics of a PC device are: fidelity of conjugation, efficiency of returning back-reflected power/energy, and reaction time or build-up time. The fidelity parameter must show how close the output of the device $E_{\text{out}}(\mathbf{r})$ is to the perfectly conjugate profile $S^*(\mathbf{R})$ of the incident signal (up to an arbitrary constant complex factor). Among other definitions, the square modulus of the normalized transverse overlapping integral of the fields is often discussed:

$$f = \frac{|\iint E_{\text{out}}(x, y)S(x, y)\mathrm{d}x\mathrm{d}y|^2}{\left(\iint |E_{\text{out}}(x, y)|^2\mathrm{d}x\mathrm{d}y \cdot \iint |S(x, y)|^2\mathrm{d}x\mathrm{d}y\right)} \qquad [5]$$

To achieve good (close to 1) fidelity of PC in the holographic or FWM scheme, one has to guarantee that the reference waves, $A_0 \exp(i\mathbf{k}_A \cdot \mathbf{R})$ and $B_0 \exp(i\mathbf{k}_B \cdot \mathbf{R})$, are exactly conjugate to each other, $\mathbf{k}_A = -\mathbf{k}_B$. That, in turn, requires the absence of any distortions in the holographic or nonlinear medium. To the contrary, there are very modest requirements on the phase inhomogeneities of the medium in the SBS scheme of PC, and for that reason it is usually labeled as a scheme of self-phase conjugation.

A hybrid scheme, Brillouin-assisted FWM, has been suggested and implemented, where mutually conjugate reference waves are produced via SBS-PC, while FWM, in a separate medium, exploits SBS nonlinearity. This scheme has produced extremely high, up to about 10^{10}, reflectivity, accompanied by an extremely low-noise detection of incoming signals.

A number of other PC schemes were first realized with the use of an important class of materials for nonlinear optics: photo-refractive crystals (PRC). These are crystals where ionization of the dopants by the incident light creates interference pattern of charge separation. The linear electro-optic (Pockels) effect transforms the resulting patterns of the electrostatic field into patterns of refractive index, thus creating dynamic holograms. A remarkable property of these crystals is that most of them act as very good electrical insulators. Therefore, the only source of conductivity, which tends to erase the

hologram, is photoconductivity, the value of the latter being proportional to the intensity of incident light itself. For this reason, the steady-state strength of the hologram turns out to be independent of the intensity. It is the build-up time that must be increased, if intensity is low. One can achieve various processes in PRC, including those similar to the $\chi^{(3)}$-type optical Kerr effect, stimulated scattering, etc.

The simplest use of PRC for generation of PC waves is to devise a FWM scheme (see Figure 3). However, a multitude of other, nontrivial schemes have been devised and implemented. The most impressive use of PRC for PC is based on interactions of the beams in the vicinity of a corner of a rectangular crystal, typically $BaTiO_3$, these interactions being of the stimulated scattering type. To honor J. Feinberg's cat, whose image was reconstructed in the first demonstration, this scheme is universally called 'cat conjugator'. It turned out to be a very reliable and robust scheme of self-phase-conjugation (SPC).

Another important group of the SPC scheme is called 'tail-biting'. One of the variants of the latter is shown in Figure 4. Incident 'pump' wave $A(\mathbf{R})$, whose transverse profile one wants to conjugate, enters the medium, and then is redirected by mirrors back into the same medium, to 'bite itself'. For the purpose of discussion, this redirected pump is labeled as wave $B(\mathbf{R})$ on its second arrival into the medium. Spontaneous scattering results in a seed for the amplification of the wave $C(\mathbf{R})$ via the process of stimulated scattering (SS) $A(\mathbf{R}) \rightarrow C(\mathbf{R})$.

Wave $C(\mathbf{R})$ is also redirected into the medium by the same mirrors, and is labeled as wave $D(\mathbf{R})$ on its second arrival into the medium. Both $C(\mathbf{R})$ and $D(\mathbf{R})$ are depicted in Figure 4 by white arrows. Wave $D(\mathbf{R})$ is also amplified due to the SS process $B(\mathbf{R}) \rightarrow D(\mathbf{R})$.

As it follows from the macroscopic description of SS process, volume gratings of refractive index are recorded in the medium, $\delta n(\mathbf{R}) \propto -i[A^*(\mathbf{R})C(\mathbf{R}) + B^*(\mathbf{R})D(\mathbf{R})]$. Among all the transverse profiles for the seed $C(\mathbf{R})$, the one that happens to be proportional to $B^*(\mathbf{R})$ will be reflected by the mirrors into the medium in the form $D(\mathbf{R}) \propto A^*(\mathbf{R})$. The latter case is again a consequence of the reversibility of propagation laws, where both gratings have the same profile and add coherently in the form $A^*(\mathbf{R})B^*(\mathbf{R})$. What is even more important, the grating $B^*(\mathbf{R})D(\mathbf{R}) \propto B^*(\mathbf{R})A^*(\mathbf{R})$ serves to close the feedback loop for seeding the appropriate $C(\mathbf{R})$. This, and other variants of tail-biting schemes, have also proved to be reliable and robust. Quite often a laser amplifier is inserted in the path $A \rightarrow B$ and $C \rightarrow D$, thus enhancing the feedback.

Another important scheme is traditionally called 'double PC'; albeit it may be better described as 'mutual PC'. Figure 5 should help in understanding that scheme. Consider the beam $A(\mathbf{R})$ incident to a medium active with respect to the stimulated-scattering-type process $A(\mathbf{R}) \rightarrow C(\mathbf{R})$. This process originates from a random seed and results in a fan of different waves $C(\mathbf{R})$. Such a process is characteristic of PRCs and is a consequence of recording dynamic gratings of refractive index $\delta n(\mathbf{R}) \propto -i[A^*(\mathbf{R})C(\mathbf{R})]$. Suppose another beam $B(\mathbf{R})$, typically incoherent with respect to $A(\mathbf{R})$, illuminates the medium from the other side, and also is engaged in the fanning process, this time $B(\mathbf{R}) \rightarrow D(\mathbf{R})$, with the grating of refractive index $\delta n(\mathbf{R}) \propto -i[B^*(\mathbf{R})D(\mathbf{R})]$ involved. Among all the transverse profiles for the seed grating $C(\mathbf{R})A^*(\mathbf{R})$, the one that happens to be proportional to $B^*(\mathbf{R})A^*(\mathbf{R})$ will be supported by similar profiles in the $B(\mathbf{R}) \rightarrow D(\mathbf{R})$ scattering. In this way, two fanning

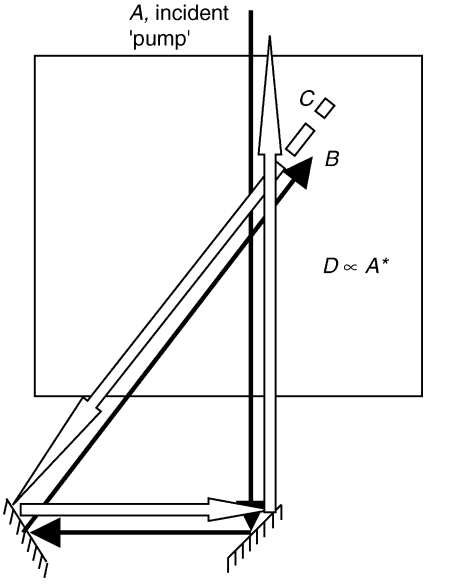


Figure 4 Tail-biting scheme of phase conjugation.

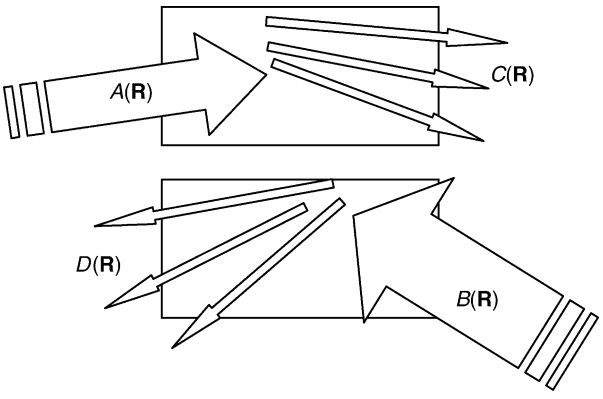


Figure 5 Towards the mechanism of double PC (mutual PC).

processes enhance each other and close the feedback loop. An important consequence is that one obtains the following output waves:

$$C(\mathbf{R}) = \text{const}_1 \times B^*(\mathbf{R})$$
$$D(\mathbf{R}) = \text{const}_2 \times A^*(\mathbf{R}) \qquad [6]$$

One can say that the output wave $C(\mathbf{R})$ presents a conjugate replica of $B(\mathbf{R})$-beam, while the output wave $D(\mathbf{R})$ presents a conjugate replica of $A(\mathbf{R})$-beam; hence the label 'mutual PC'. In some experiments the incident beams, $A(\mathbf{R})$ and $B(\mathbf{R})$, were of somewhat different colors (i.e., from different lasers), or their pulses did not overlap in time in the medium. It is the grating memory that 'informs' one beam about the presence of the other. The mechanism of cat conjugator is considered to be a combination of tail-biting and double PC schemes.

Pumped laser material with saturable gain often serves as an important nonlinear-optical medium for implementing PC. One of the evident advantages is that possible low efficiency of properly nonlinear process may be compensated by the gain in the same piece of material.

Considerable attention is paid nowadays to the chirp reversal of optical pulses that are used in fiber optical communications. Such chirped pulses, $S(t) = \exp(-i\omega_0 t + i\beta t^2/2 - t^2/\tau_1^2)$, with the value of the 'chirp' $\mathrm{d}\omega/\mathrm{d}t = -\beta$ and an increased pulse duration τ_1, arise as a result of propagation in a fiber which possesses group velocity dispersion (GVD). Pulse stretching and chirp are the consequences of different propagation times for the different frequency constituents of the pulse. If some device changes the sign of that chirp, then subsequent propagation of the pulse, through a piece of fiber with the same GVD, restores the duration of the pulse to original shorter value τ_0. A scheme has been suggested to use nonlinear optical FWM in a $\chi^{(3)}$-medium or three-wave mixing (ThWM) in a $\chi^{(2)}$-medium:

$$\delta D_{\text{conj}}(t) = \chi^{(3)} S^*(t) A(t) B(t)$$
$$\delta D_{\text{conj}}(t) = \chi^{(3)} S^*(t) C(t) \qquad [7]$$

Indeed, complex conjugation of the time dependence of the signal's field is equivalent to the change of the chirp sign. The second (ThWM) expression is written assuming that the reference wave $C(t) \propto \exp(-2i\omega_0 t)$, so that is has the frequency double that of the signal. The same process of ThWM has been shown to yield the conjugation of transverse phase profile of the beams, but for a number of reasons it was not applied.

As for the applications of PC, one of them was discussed above (**Figure 1**: double-pass scheme of a laser amplifier). Lasers with one or two PC mirrors, or with more complicated PC scheme, promise generation of high-transverse-quality beams with the use of realistic laser media, the latter inevitably possessing thermal and other distorting inhomogeneities. Another important potential application is free-space optical communications: Earth–Earth or Earth–satellite through atmosphere, or satellite–satellite communications. An 'interrogating' beam may be sent in one direction through an imperfect optical system and/or through turbulent atmosphere. Conjugation of transverse profile of the 'interrogating' beam at the other end of the communication link leads to an almost perfect redirection of the reflected beam towards the 'beacon' source, while time modulation may carry the information.

Readout of information data from a holographic storage is almost always performed in the regime of reading the conjugate wave, since it corrects for the most part of aberrations of optical systems in question.

To conclude, one may use a citation from the two consecutive *Scientific American* popular papers on PC: 'at present the number of applications would seem to be limited only by imagination.'

Reviews of various aspects of phase conjugation may be found in monographs and *Scientific American* papers listed below. The author of the present article was introduced to PC by his colleague V. V. Ragulskii, and has greatly benefited from collaboration with V. V. Ragulskii and V. V. Shkunov.

See also

Microscopy: Phase Contrast Microscopy. **Nonlinear Optics, Basics:** Four-Wave Mixing. **Phase Control:** Phase Conjugation and Image Correction. **Scattering:** Scattering Theory.

Further Reading

Bespalov VI and Pasmanik GA (1994) *Nonlinear Optics and Adaptive Laser Systems*. Commack, NY: Nova Science Publishers.

Feinberg J (1986) Self-pumped, continuous wave phase conjugator using internal reflection. *Optics Letters* 7: 486.

Fisher RA (ed.) (1983) *Optical Phase Conjugation*. New York: Academic Press.

Gower M and Proch D (eds) (1994) *Optical Phase Conjugation*. Berlin: Springer Verlag.

Hellwarth RW (1982) Optical beam conjugation by stimulated backscattering. *Optical Engineering* 21: 257.

Nosatch OYu, Popovichev VI, Ragulskii VV and Faizullov FS (1972) Compensation of phase distortions in an amplifying medium by a 'Brillouin mirror'. *JETP Letters* 16: 435.

Pepper DM (1986) Applications of optical phase conjugation. *Scientific American* 254: 74–83.

Shkunov VV and Zeldovich BYa (1985) Optical phase conjugation. *Scientific American* 253: 54–59.

Yariv A (1978) Phase conjugate optics and real-time holography. *IEEE Journal of Quantum Electronics* 14: 650.

Zeldovich BYa, Popovichev VI, Ragulskii VV and Faizullov FS (1972) On relation between wavefronts of reflected and exciting radiation in stimulated Brillouin scattering. *JETP Letters* 16: 109.

Zeldovich BYa, Shkunov VV and Pilipetsky NF (1985) *Principles of Phase Conjugation*. Berlin: Springer-Verlag.

Zeldovich BYa, Mamaev AV and Shkunov VV (1995) *Speckle-Wave Interactions in Application to Holography and Nonlinear Optics*. Boca Raton, FL: CRC Press.

Photorefraction

M Cronin-Golomb, Tufts University, Medford, MA, USA

B Kippelen, University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction and Basics of the Photorefractive Effect

In 1966 Ashkin and coworkers were pursuing research on optical devices using the electro-optic material lithium niobate. They noticed that the refractive index of lithium niobate would change when it was exposed to laser light, and upset the expected operation of their devices. They called the effect optical damage. Shortly thereafter, Chen, LaMacchia, and Fraser reported on the use of the optical damage effect for holographic data storage. Thus began the field of photorefractive nonlinear optics, which has been used in various applications such as real-time holography, optical data storage, optical image amplification, nondestructive testing, distortion compensation by phase conjugation, pattern recognition, and radar signal processing. Many inorganic and organic materials have been investigated for their photorefractive effects, including ferroelectrics, semiconductors, and sensitized polymers. The most well-known inorganic materials are lithium niobate, bismuth silicon oxide, barium titanate, and strontium barium niobate. Most organic photorefractive materials are based on polymeric photoconductors such as those used in xerography that are doped with electro-active molecules, some plasticizers, and sensitizers.

While in its broadest interpretation, the photorefractive effect occurs whenever light incident on a material causes a refractive index change, one usually applies the term to electro-optic index changes resulting from optically generated space charge fields. Materials that are photorefractive in this sense share the following properties

- high transmission at the operating wavelengths;
- linear electro-optic coefficients or orientational Kerr effects;
- charge carriers that become mobile when optically excited;
- trapping centers for these charge carriers to enable spatially non-uniform redistribution of charge.

Consider two beams from the same laser crossing inside a photorefractive material such as barium titanate. The interference pattern will have bright and dark fringes. Charge carriers are excited where the light is bright, then drift and diffuse to regions of relative darkness where they preferentially recombine into trapping centers. In this way, a net excess charge develops in the dark regions, and a net deficit of charge develops in the bright regions. The spatially varying charge distribution has an electric field associated with it and this electric field causes a spatially varying refractive index profile. Because the space charge, its field, and resulting refractive index have the same spatial periodicity as the original interference pattern we have a holographic phase grating. The diffraction efficiency of the hologram can easily approach 100% for materials such as barium titanate and strontium barium niobate which have high electro-optic coefficients. Likewise, such high diffraction efficiencies are easily obtained in 100 micrometer-thick photorefractive polymer films.

The Standard Rate Equation Model

The development of photorefractive gratings can be modeled using standard semiconductor rate equations. Figure 1 shows two beams incident

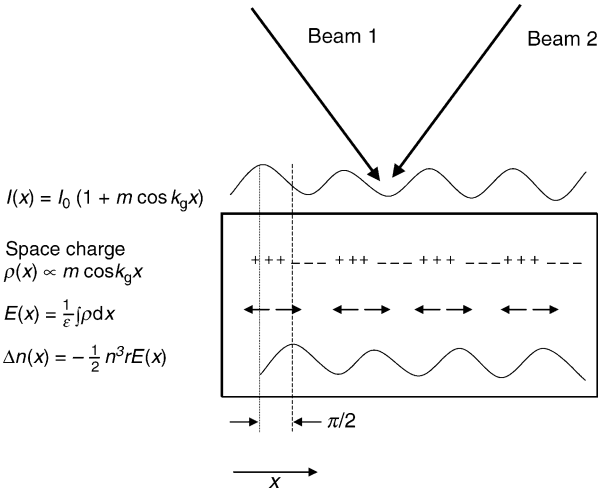


Figure 1 Diagram of two-beam coupling process showing optical interference pattern, space charge, resultant electric field, and $\pi/2$ phase shifted electro-optically induced refractive index grating.

symmetrically on a photorefractive crystal or polymer. They form an interference pattern whose intensity may be written:

$$I(x) = I_0(1 + m\cos(k_g x)) \qquad [1]$$

where x is the direction perpendicular to the interference fringes, I_0 is the average intensity, m is the modulation index, and k_g is the wavenumber of the interference pattern.

The crystal may be considered a wide bandgap semiconductor containing electron donors in the bandgap with density N_D and electron acceptors with density N_A. It is assumed that some electrons ionized from the donors permanently occupy these acceptors so that when charge in the crystal is distributed uniformly in the dark, the number density of ionized donors N_D^i is equal to N_A. Likewise, polymers contain donor and acceptor-like molecules that can be neutral or ionized. The spatially varying light distribution ionizes the donors at the following rate, assuming that $N_D^i \ll N_D$:

$$\frac{\partial N_D^i}{\partial t} = sIN_D - \gamma_R n_e N_D^i \qquad [2]$$

where s is a photoionization coefficient, γ_R is a recombination parameter, and n_e is the density of excited charge carriers, which we assume here to be electrons. The model can easily be generalized to include holes. We also use the equation of charge conservation:

$$\frac{\partial N_D^i}{\partial t} = \frac{\partial n_e}{\partial t} - \frac{1}{e}\nabla\cdot\mathbf{j} \qquad [3]$$

where e is the charge of an electron, and $\mathbf{j}$ is the current in the conduction band.

$$\mathbf{j} = \mu e n_e \mathbf{E} + k_B T \mu \nabla n_e \qquad [4]$$

where μ is the electron mobility, k_B is Boltzmann's constant, and T is the temperature. The electric field obeys Gauss's law:

$$\nabla\cdot\mathbf{E} = -e(n_e + N_A - N_D^i)/\varepsilon \qquad [5]$$

where N_A is the density of acceptors. These equations may be linearized by approximating the electron density, ionized donor density, and electric field with their first Fourier components:

$$\begin{aligned} E &= E_0 + \tfrac{1}{2}(E_1\exp(\mathrm{i}k_g x) + E_1^*\exp(-\mathrm{i}k_g x)) \\ N_D^i &= N_{D0}^i + \tfrac{1}{2}(N_{D1}^i\exp(\mathrm{i}k_g x) + N_{D1}^{i*}\exp(-\mathrm{i}k_g x)) \\ n_e &= n_{e0} + \tfrac{1}{2}(n_{e1}\exp(\mathrm{i}k_g x) + n_{e1}^*\exp(-\mathrm{i}k_g x)) \end{aligned} \qquad [6]$$

This assumption is strictly valid only when the modulation index m is much less than unity. Otherwise, a generalization to higher orders in the Fourier series is required. However, the linearized theory is sufficient to illustrate the most important features of the photorefractive effect. The solution for the space charge field E_1 for the case when the interference pattern is applied at time $t = 0$ is

$$E_1 = m\frac{-\mathrm{i}E_Q(E_0 + \mathrm{i}E_D)}{E_0 + \mathrm{i}(E_D + E_Q)}(1 - \exp(-t/\tau)) \qquad [7]$$

where E_0 is an externally applied or photovoltaic field (if any), E_D is the diffusion field

$$E_D = \frac{k_B T k_g}{e} \qquad [8]$$

and E_Q is the limiting space charge field

$$E_Q = \frac{eN_A}{\varepsilon k_g} \qquad [9]$$

Some photorefractive crystals, most notably $LiNbO_3$, exhibit the photovoltaic effect in which optical illumination induces a dc field across the crystal.

The sinusoidally varying space charge field E_1 operates through the linear electro-optic effect with effective coefficient r to produce a sinusoidal variation in the refractive index n of the crystal:

$$n(x) = n_0 + \tfrac{1}{2}(n_1\exp(\mathrm{i}k_g x) + n_1^*\exp(-\mathrm{i}k_g x)) \qquad [10]$$

where

$$n_1 = -\frac{1}{2} r n_0^3 E_1 \qquad [11]$$

The effective electro-optic coefficient r may be found from tensor analysis of the electro-optic tensor and

the vector space charge and optical fields. Notice that there is a 90-degree phase shift between the interference pattern and the refractive index grating when E_0 is zero. The time constant τ is given by

$$\tau = \frac{N_A}{sN_D I_0}\frac{E_0 + i(E_D + E_\mu)}{E_0 + i(E_D + E_Q)} \quad [12]$$

where E_μ is a mobility field

$$E_\mu = \frac{\gamma_R N_A}{\mu k_g} \quad [13]$$

When E_0 is nonzero, there is an oscillatory component to the time constant.

In contrast to the case of ordinary optical nonlinearities such as the optical Kerr effect, in which the nonlinear coupling strength is proportional to the optical intensity, the steady state strength of the photorefractive effect is independent of optical intensity while the time constant is inversely proportional to intensity in the basic model described above. The time constant varies with the photoconducting performance of a given material. In the fastest polymeric and inorganic materials it is, at the time of writing, of the order of a few milliseconds at 1 W cm^{-2} of incident optical intensity.

At low intensities (below the equivalent dark intensity, 1 W cm^{-2} in barium titanate), the above equations need to be modified to include the effect of dark conductivity. Even in the dark, there will be a few mobile charge carriers in the conduction band that tend to erase the grating. This will result in the effect appearing more Kerr-like, except still with the 90-degree phase shift between the index grating and the interference pattern.

Coupled Wave Equations

The change in refractive index n_1 can be large enough to produce substantial interactions between the writing beams. The writing beams generate a phase grating that diffracts the beams into each other. The grating influences the writing beams, which in turn influence the grating. In the cases where the diffusion field dominates, for example when the externally applied or photovoltaic field E_0 is absent, one beam is amplified by in-phase diffraction of the other beam from the grating. As shown in Figure 2, this amplification results from a 90-degree phase shift due to diffraction from a phase grating coupled with a -90-degree phase shift from the spatial phase shift between the interference pattern and the index grating. The second beam is attenuated by destructive interference with the first beam diffracted by the

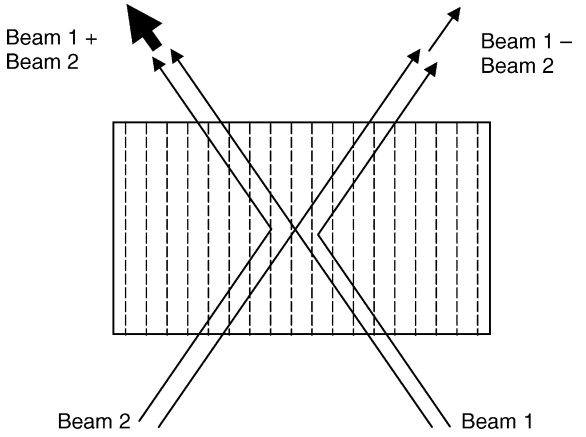


Figure 2 Two-beam coupling amplification. Beam 1 is amplified by constructive interference. Beam 2 is de-amplified by destructive interference.

grating. These interactions can be modeled by standard coupled wave theory. Let the electric field amplitude associated with the jth beam be

$$\mathbf{E}_j(\mathbf{r}, t) = \mathbf{e}\Big[A_j(\mathbf{r})\exp(i(\mathbf{k}_j \cdot \mathbf{r} - \omega t) + A_j^* \exp(-i(\mathbf{k}_j \cdot \mathbf{r} - \omega t))\Big] \quad [14]$$

where $\mathbf{e}$ is the polarization unit vector. Using the scalar wave equation

$$\nabla^2 \mathbf{E} + k^2 \mathbf{E} = 0 \quad [15]$$

and the slowly varying envelope approximation

$$\left|\frac{d^2 A_j}{dz^2}\right| \ll k\left|\frac{dA_j}{dz}\right| \quad [16]$$

we find

$$\frac{dA_1}{dz} = +\gamma\frac{A_1|A_2|^2}{I_0}$$
$$\frac{dA_2^*}{dz} = -\gamma\frac{A_2^*|A_1|^2}{I_0} \quad [17]$$

with the coupling constant γ given by

$$\gamma = -\frac{i\omega\Delta n_m}{c\cos\vartheta} \quad [18]$$

where ϑ is the half-angle between the writing beams and Δn_m is given by

$$\Delta n_m = -\frac{1}{2}n_0^3 r\frac{E_1}{m} \quad [19]$$

Equation [17] shows that beam 1 is amplified and beam 2 is attenuated. That beam 1 is amplified

instead of beam 2 is a result of the choice of crystal orientation and hence the sign of γ. These nonlinear equations can be solved for normalized intensity $I_j = |A_j|^2$:

$$I_1(z) = \frac{I_0}{1 + (I_2(0)/I_1(0))\exp(+\Gamma z)}$$
$$I_2(z) = \frac{I_0}{1 + (I_1(0)/I_2(0))\exp(-\Gamma z)} \quad [20]$$

where $\Gamma = \gamma + \gamma^*$ is the intensity coupling constant. In the limit where $I_1 \ll I_2$, the weak beam, beam 1, experiences exponential amplification. This amplification can be used to build optical oscillators with many of the same properties as conventional laser oscillators. The solution can be generalized to the case where γ has an imaginary component due to an applied or photovoltaic field E_0. Linear absorption can also be included.

Materials

Photorefractive materials may be separated into two broad classes: inorganic and organic. The first materials investigated were inorganic oxides and semiconductors. Their success led to efforts to endow more easily produced organic materials with the required photoconductivity, charge trapping centers, and electro-optic coefficients.

Inorganic Materials

The first requirement in a classic photorefractive material is a linear electro-optic coefficient, such as appears in sillenites such as bismuth silicon oxide $Bi_{12}SiO_{20}$ and bismuth germanium oxide $Bi_{12}GeO_{20}$. These were some of the first photorefractive materials used for image processing and phase conjugation applications. However, their electro-optic coefficients (a few pm/V) are not large enough to easily give rise to large diffraction efficiencies, or to photorefractive oscillators. Materials at a temperature near a phase transition generally have higher electro-optic coefficients because their crystal structures are on the verge of changing. They are extremely susceptible to the effect of any external influence such as the application of electric fields including photorefractive space charge fields. That is why ferroelectric materials are good candidates for photorefractive materials. These include barium titanate $BaTiO_3$, potassium niobate $KNbO_3$, and strontium barium niobate $Sr_xBa_{1-x}Nb_2O_6$. The mean free path of charge transport is less than that in sillenites, so they require more photons to reach a steady state. Therefore, ferroelectric materials are typically less sensitive than sillenites.

The second requirement is for photoconductivity. This requires the existence of photoexcitable charge carriers, either from valence band to conduction band or from intraband trapping centers. The latter source of photocarriers is used most often because the absorption length of light whose energy is greater than the bandgap is usually only a few micrometers. This places a severe restriction on the beams' interaction length ℓ. Thus there has been considerable research on suitable dopants, most commonly Fe^{2+} and Fe^{3+}. These dopants also act as trapping centers.

Lithium niobate is an example of a material with a large photovoltaic effect. When illuminated, this crystal develops a large dc field, which acts to enhance the grating strength. In those cases where the photorefractive effect is not wanted, such as in the design of lithium niobate waveguide devices, the photovoltaic effect can be greatly diminished by the addition of MgO to the crystal melt during growth.

For photorefractivity in the near infrared, semiconductors such as gallium arsenide and indium phosphide have been used with success. These materials also have the advantage that they can be grown in layered structures to produce, for example, multiple quantum wells that can be used to tailor the characteristics of optical excitation and charge transport.

Organic Materials

First-generation photorefractive polymers were designed to mimic the properties of their inorganic counterparts. Owing to their rich structural flexibility, organic synthetic materials with suitable charge transport, trapping, linear electro-optic effects and low optical absorption, were developed. This approach did build on the know-how developed previously in making photoconducting polymers for xerography and electro-optic polymers for optical modulation. Current polymers are based on an orientational photorefractive effect that leads to higher refractive index changes compared with traditional photorefractive materials. In materials with orientational photorefractivity, the refractive index change is produced by the field induced reorientation of anisotropic conjugated molecules that possess a permanent dipole moment and that have a high polarizability anisotropy. The photorefractive space charge field together with the applied field will periodically reorient these molecules and produce a periodic refractive index modulation through an orientational Kerr effect. The build-up and dynamics of this space-charge field are similar to

those of traditional photorefractive crystals and can be described by eqns [2]–[7]. The time constant of the hologram is a convolution of the build-up time of the space-charge and the orientational diffusion time of the dipolar molecules in the total electric field. In contrast to crystals, polymers are nearly amorphous and the transport properties are described by charge hopping processes instead of band-type transport in crystals. Consequently, the photoconducting properties of polymers are strongly field dependent and photorefractivity is mainly observed under strong applied voltages. Numerous polymer composites have been developed using the hole transport polymer poly(N-vinylcarbazole). Several materials with a refractive index modulation amplitude of the order of a percent and two-beam coupling constants $\gamma > 100\ \mathrm{cm}^{-1}$ have been reported. New polymer composites are continuously being tested. Photorefractivity is also studied in other organic materials including organic crystals, liquid crystals, nanocomposites such as polymer-dispersed liquid crystals, or hybrid materials such as sol-gels.

Applications

Holographic Data Storage

Holographic data storage takes advantage of the Bragg selectivity of thick gratings. This allows many holograms to be superimposed in the same small volume, typically of the order of one cubic centimeter. A page of data is displayed on a spatial light modulator and a laser beam passing through the modulator is holographically recorded in the crystal with a reference beam at a specific angle. Many pages of data can be superimposed by recording many holograms with angularly multiplexed reference beams. Other multiplexing schemes are implemented by changing the shape of the wavefront of the reference beam. In principle, the upper limit of recording density is determined by the wavelength of light: one bit per cubic wavelength. If the recording wavelength is 0.5 μm, then one cubic centimeter can contain 1000 gigabytes of data. In practical circumstances, when noise is taken into consideration, the capacity is more realistically of the order of one gigabyte if a 1000×1000 spatial light modulator is used.

Distortion Compensation by Phase Conjugation

Photorefractive materials can be used to make high-reflectivity phase conjugate mirrors. The phase conjugate of a laser beam is produced when a hologram of the beam is read by another beam traveling in the opposite direction to the original reference beam. The phase conjugate reconstruction is a time-reversed copy of the original beam. If the original beam has passed through distorting optics, then the phase conjugate beam will retrace the path of the original beam through the distortion and emerge in its undistorted original state. In the standard nomenclature of phase conjugation, the input beam is called the signal, or probe, and the two reference beams are called the pumps. The output beam at $z = 0$ is called the phase conjugate beam and has zero amplitude at its input at $z = \ell$, where ℓ is the interaction length. Applications for phase conjugation exist, for example, in signal transmission through distortions and laser cavity design. If a phase conjugate mirror is used as a cavity mirror, then the effects of intracavity distortions are removed.

Since the photorefractive gratings can be very strong, the diffraction efficiency of the counterpropagating reference beam can be so high that the phase conjugate reflectivity can exceed unity. The simplest generalization of eqn [17] to the four-wave mixing phase conjugation case is when only transmission gratings are important, as occurs in many circumstances depending on the mutual coherence properties of the beams, and the material's properties. The coupled wave equations are

$$\begin{aligned}
\frac{\mathrm{d}A_1}{\mathrm{d}z} &= \gamma\frac{(A_1A_4^* + A_2^*A_3)A_4}{I_0} \\
\frac{\mathrm{d}A_2^*}{\mathrm{d}z} &= \gamma\frac{(A_1A_4^* + A_2^*A_3)A_3^*}{I_0} \\
\frac{\mathrm{d}A_3}{\mathrm{d}z} &= -\gamma\frac{(A_1A_4^* + A_2^*A_3)A_2}{I_0} \\
\frac{\mathrm{d}A_4^*}{\mathrm{d}z} &= -\gamma\frac{(A_1A_4^* + A_2^*A_3)A_1^*}{I_0}
\end{aligned} \qquad [21]$$

In the undepleted pumps approximation, $(I_1, I_2 \gg I_3, I_4)$, the equations become linear and the solution for phase conjugate reflectivity $R = I_3(0)/I_4(0)$ is

$$R = \left|\frac{\sinh[\gamma\ell/2]}{\cosh[(\gamma\ell + \ln r)/2]}\right|^2 \qquad [22]$$

where $r = I_2/I_1$ is the ratio between the intensities of the pumping beams.

The fact that the reflectivity of the phase conjugate mirror can be greater than unity means that we can build an optical oscillator bounded by a regular mirror and a phase conjugate mirror only. Not only does it not require any additional optical gain, but it also compensates for intracavity distortions. The regular mirror can have any shape, provided that it is sufficiently reflective.

Self-pumped Phase Conjugate Mirrors

If a laser beam passes through a crystal placed inside an optical cavity formed by two facing mirrors, light scattered by imperfections in the crystal can be amplified through the photorefractive effect. The cavity provides feedback and optical oscillation can build up. The oscillation beams pump the crystal as a phase conjugate mirror for the incident beam, thus forming a self-pumped phase conjugate mirror. The feedback can even be provided by total internal reflection in the crystal in which case the crystal by itself can become a phase conjugate mirror.

Pattern Recognition and Image Filtering

Photorefractive wave mixing can be used to perform pattern recognition by matched filtering. One example would be to identify a tank in a battlefield scene, another would be to identify all of the occurrences of a particular word on a page of text. Suppose the input beams contain the corresponding pictorial information, such as might be obtained by passing the beams through an image-bearing transparency or spatial light modulator. Equation [21] shows that the source for beam 3 contains a term proportional to the product of the amplitudes of the three input beams. If lenses are placed in the input beams so that the crystal receives the Fourier transforms of those beams, then the output beam at the crystal, beam 3, will be proportional to the product of the Fourier transforms of the input beams 1, 2, and 4. A lens may then be used to perform the inverse Fourier transform of the product of the Fourier transforms of the input beams, producing an output proportional to beam 1 convolved with beam 2 correlated with beam 4. If beam 1 is a point source before its Fourier transforming lens, then it will be a plane wave at the crystal. Beam 3 after inverse Fourier transformation by its lens will be the correlation of beams 4 and 2. For example, suppose we want to find all the occurrences of a particular word, say 'optics' in a given page of text. Then we would make a transparency of the word 'optics' and place that in input beam 4. We would then place an image of the page of text in beam 2. Beam 3 would then contain a field with bright spots at the places containing the word 'optics' in the original text.

The real-time holographic recording properties of photorefractive materials can also be exploited in medical imaging applications by performing time-gated holography. In this method, a hologram is formed by the temporal overlap in the photorefractive sample of a reference pulse and the first-arriving (ballistic photons) light from a stretched image-bearing pulse that has propagated through a scattering medium. The filtering of the useful photons from the scattered ones is achieved by reconstructing the hologram formed with the ballistic photons in a four-wave mixing geometry.

Optical Limiting, the Novelty Filter, and Laser Ultrasonic Inspection

The attenuation of beam 2 in eqn [17] can be used in several applications including optical limiting and novelty filtering. If one wants to protect a sensor from high-intensity laser radiation, then one could split a small portion of the input beam directed at the sensor and use it as beam 1 in a photorefractive recording setup with the input beam acting as beam 2. If the laser intensity is above the equivalent dark intensity such that the optically excited charge density is greater than the thermally excited charge density, the photorefractive effect will be activated and the input beam will be de-amplified by destructive interference with beam 1, thus protecting the sensor. In materials with high gain–length products ($\gamma\ell > 1$), separate provision of beam 1 is unnecessary because light scattered from the input beam by imperfections in the material and other scattering centers will be greatly amplified, often to such an extent that the input beam is almost completely attenuated. This photorefractive amplification of scattered radiation is known as the fanning effect, because the amplified scattered light often appears as a fan, or brush of light, as can be seen in Figure 3. The effect can also be used to make a novelty filter, which transmits only the moving portion of a scene. The crystal is only fast enough to

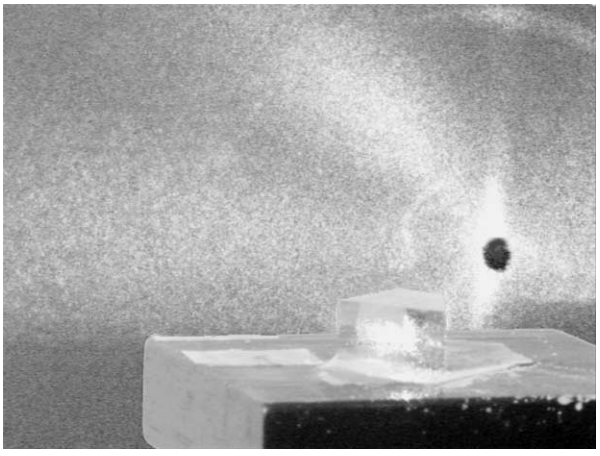


Figure 3 Photorefractive barium titanate exhibiting amplified scattering. A helium neon laser beam is incident from the lower left, passes through the crystal to a screen where it is blacked out to prevent saturation of the camera. The screen shows brushes of amplified scattering, or fanning.

respond to the slowly changing or static components of an image-bearing beam. Since the grating formed is Bragg-matched only to the slow component, the grating will only attenuate that slow component. Any rapidly changing parts of a scene pass through the crystal unattenuated. Such a filter is useful for picking out moving objects in a static cluttered background, for example a tank on a battlefield, or a micro-organism swimming against a stationary background.

A related application is in laser ultrasonic inspection. Defects in industrial material processing such as welding have characteristic ultrasonic signatures. The part under test is pinged by a pulsed laser and a probe laser is reflected from the part. As the part is shaken by the ultrasound, speckle in the reflected beam vibrates. A photorefractive recording is made of the speckle beam. Electrodes are placed on the photorefractive material, so that any optically induced currents can be detected. If the speckle pattern is not moving, or is moving more slowly than the speed of response of the photorefractive material, the photorefractive grating will be essentially at steady state: the drift currents balance the diffusion currents so there is no net current. There is no signal as the part moves through the process line. However, if the speckle beam is moving faster than the response time of the material, it will move photoexcited charge back and forth past the quasistatic grating and generate a net current for detection via the electrodes.

Adaptive Signal Processing

The relatively slow speed of photorefractive devices can be used to advantage in radio-frequency (RF) signal processing, such as signal extraction and coherent combination of signals from antenna arrays. The signal extraction application depends on grating competition in two-beam coupling. Suppose we wish to separate signals on two different RF carrier frequencies ω_1 and ω_2, respectively. The combined signal is applied to an optical carrier beam using a high-speed modulator. The resulting optical field may be represented as

$$S_1(t)\exp(\mathrm{i}\omega_1 t) + S_2(t)\exp(\mathrm{i}\omega_2 t) \qquad [23]$$

It is used to pump a unidirectional ring resonator so that grating competition allows oscillation only on the strongest component of the combined signal, say S_1. The output of the oscillator is proportional to the extracted component S_1. The effectiveness of the intersignal competition is enhanced by placing another photorefractive material in the cavity. A portion of the intracavity beam is picked off by a beamsplitter and used as a two-beam coupling pump in the second material. The crystal is oriented so that the photorefractive grating diffracts the picked-off beam back into the cavity. The return of the picked-off component is most effective for the stronger component S_1 thus decreasing its loss compared to that of the weaker component S_2. This beamsplitter/crystal combination is known as a reflexive coupler.

Photorefractive Solitons

Under favorable conditions, a single beam incident on a photorefractive crystal will induce a positive refractive index change at the center of the beam. This provides a self-focussing tendency that counteracts the beam's divergence due to diffraction. When these two effects balance each other, the beam can propagate with a constant diameter. Such a beam is known as a spatial soliton in analogy with temporal solitons in optical fibers and can occur when there is a component of the photorefractive response due to drift. The drift component of the photorefractive effect appears when a dc field E_0 is applied to the material. Potential applications are optically written waveguides and couplers.

List of Units and Nomenclature

Amplitude gain	$[\mathrm{cm}^{-1}]$	γ
Applied dc field	$[\mathrm{Vm}^{-1}]$	E_0
Current	$[\mathrm{Am}^{-2}]$	$\mathbf{j}$
Dielectric permittivity	[Farads m^{-1}]	ε
Diffusion field	$[\mathrm{Vm}^{-1}]$	E_D
Electron acceptor density	$[\mathrm{m}^{-3}]$	N_A
Electron charge	[C]	e
Electron donor density	$[\mathrm{m}^{-3}]$	N_D
Grating wavenumber	$[\mathrm{m}^{-1}]$	k_g
Intensity gain	$[\mathrm{cm}^{-1}]$	Γ
Interaction length	[cm]	ℓ
Ionized donor density	$[\mathrm{m}^{-3}]$	N_D^i
Limiting space charge field	$[\mathrm{Vm}^{-1}]$	E_Q
Mobility	$[\mathrm{m}^2\,\mathrm{V}^{-1}\,\mathrm{sec}^{-1}]$	μ
Mobility field	$[\mathrm{Vm}^{-1}]$	E_μ
Normalized intensity	$[\mathrm{V}^2\,\mathrm{m}^{-2}]$	I
Photo-ionization coefficient	$[\mathrm{kg}\,\mathrm{sec}^{-2}]$	s
Photorefractive time constant	[sec]	τ
Pump ratio	[1]	r
Radian frequency	$[\mathrm{sec}^{-1}]$	ω
Recombination coefficient	$[\mathrm{m}^3\,\mathrm{sec}^{-1}]$	γ_R
Slowly varying optical electric field	$[\mathrm{Vm}^{-1}]$	A

See also

Holography, Applications: Holographic Recording Materials and Their Processing. **Holography, Techniques:** Holographic Interferometry. **Imaging:** Volume Holographic Imaging. **Information Processing:** Free-Space Optical Computing; Optical Neural Networks. **Materials for Nonlinear Optics:** Organic Nonlinear Materials. **Modulators:** Electro-Optics. **Nonlinear Optics, Basics:** Four-Wave Mixing; Nonlinear Optical Phase Conjugation. **Phase Control:** Phase Conjugation and Image Correction. **Scattering:** Stimulated Scattering. **Solitons:** Bright Spatial Solitons.

Further Reading

Coufal HJ, Psaltis D and Sincerbox GT (eds) (2000) *Holographic Data Storage*. Berlin: Springer.

Gunter P and Huignard J-P (eds) (1988) *Photorefractive Materials and their Applications I*. Berlin: Springer-Verlag.

Gunter P and Huignard J-P (eds) (1989) *Photorefractive Materials and their Applications II*. Berlin: Springer-Verlag.

Nalwa HS and Miyata S (eds) (1997) *Nonlinear Optics of Organic Molecules and Polymers*. Boca Raton: CRC Press.

Nolte DD (ed.) (1995) *Photorefractive Effects and Materials*. Kluwer.

Pepper DM, Feinberg J and Kukhtarev NV (1990) The photorefractive effect. *Scientific American* 263: 62.

Solymar L, Webb DJ and Grunnet-Jepsen A (1996) The Physics and Application of Photorefractive Materials. Oxford: Clarendon Press.

Yeh P (1993) Introduction To Photorefractive Nonlinear Optics. New York: Wiley.

Yu F and Yin S (eds) (2000) *Photorefractive Optics: Materials, Properties, and Applications*. San Diego: Academic Press.

Ultrafast and Intense-Field Nonlinear Optics

A L Gaeta, Cornell University, Ithaca, NY, USA

R W Boyd, University of Rochester, Rochester, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The tremendous development of high-powered femtosecond laser systems since the 1980s, has opened up new areas of research in nonlinear optics, plasma physics, atomic and molecular dynamics, and intense-field physics. For many of these applications, it is important to understand how ultrashort light pulses propagate through a medium under conditions in one or more of the processes in which nonlinear optics play an important role.

The starting point for the modeling of light propagation under these conditions is Maxwell's wave equation for the electric field $E(\mathbf{r}, t)$ which is given in Gaussian units as

$$\nabla^2 E - \frac{\partial^2 E}{\partial t^2} = 4\pi \frac{\partial^2 P}{\partial t^2} \qquad [1]$$

where the $P(\mathbf{r}, t)$ is the polarization inside the medium. Typically, the polarization is separated into a term P_l that depends linearly on the field E and into a term P_{nl} that depends nonlinearly on the applied field. The Fourier transform of the linear polarization can be expressed as $\tilde{P}_l(\mathbf{r}, \omega) = \chi^{(1)}(\omega)\tilde{E}(\mathbf{r}, \omega)$, where $\tilde{E}(\mathbf{r}, \omega)$ is the Fourier transform of electric field $E(\mathbf{r}, t)$. For the laser–matter interactions that we consider here, we assume that the linear susceptibility $\chi^{(1)}(\omega)$ is real, in which case the wave equation can be expressed as

$$\nabla^2 \tilde{E} + \left[\frac{n_l(\omega)\omega}{c} \right]^2 \tilde{E} = -\frac{4\pi\omega^2}{c^2} \tilde{P}_{\mathrm{nl}} \qquad [2]$$

where $n_l(\omega) = \sqrt{1 + 4\pi\chi^{(1)}(\omega)}$ is the frequency-dependent linear refractive index of the medium.

For light pulses that are longer than an optical period, the envelope description is valid and the electric field can be described by an amplitude envelope $A(\mathbf{r}, t)$ and a carrier frequency ω_0 such that

$$E(\mathbf{r}, t) = A(\mathbf{r}, t)\mathrm{e}^{\mathrm{i}(k_0 z - \omega_0 t)} + c.c \qquad [3]$$

where $k_0 = k(\omega_0) = n_0\omega_0/c$ is the wavevector amplitude and $n_0 = n_l(\omega_0)$. This envelope description is advantageous for performing analytical and numerical studies of pulse propagation. However, for sufficiently short laser pulses, where the shape of the envelope function does not depend on the carrier phase of the carrier wave, such a description is no longer applicable. Nevertheless the envelope description can be made valid for pulses that are nearly as short as a single cycle or, alternatively, that have spectral bandwidths that are comparable to the central frequency ω_0. To derive an equation for the spatio-temporal evolution of the pulse envelope, the relation for the electric field is substituted into the

wave equation (eqn [2]) and the following two approximations are made: i) $k_0 \partial A/\partial z \ll \partial^2 A/\partial z^2$, which signifies that the envelope varies slowly in space over a distance compared to the central wavelength; ii) $v_{\mathrm{ph}} \sim v_{\mathrm{gr}}$ where $v_{\mathrm{ph}} = c/n_0$ is the phase velocity and $v_{\mathrm{gr}} = (\mathrm{d}k/\mathrm{d}\omega)^{-1}$ is the group velocity. This latter approximation is invariably well satisfied when the frequencies contained in the laser field are highly nonresonant with any transition frequencies of the medium.

For an input pulse at $z = 0$ with a peak amplitude A_0, and characteristic widths in space and time given by w_0 and τ_p, respectively, the equation for the normalized amplitude $u(r_\perp, z, t) = A(r_\perp, z, t)/A_0$ can be expressed as

$$\frac{\partial u}{\partial \xi} = \frac{i}{4}\left(1 + \frac{i}{\omega_0 \tau_p}\frac{\partial}{\partial \tau}\right)^{-1} \nabla_\perp^2 u - i \sum_{n=2} \frac{L_{\mathrm{ds}}}{n! L_{\mathrm{ds}}^{(n)}} \frac{\partial^n u}{\partial \tau^n} + i\left(1 + \frac{i}{\omega_0 \tau_p}\frac{\partial}{\partial \tau}\right) p^{\mathrm{nl}} \qquad [4]$$

where $\tau = (t - z/v_g)/\tau_p$ is the normalized retarded time for the pulse traveling at the group velocity v_g, $L_{\mathrm{ds}}^{(n)} = \tau_p^n/\beta_n$ is the nth-order dispersion length, $\beta_n (n \geq 2)$ is the nth-order dispersion constant, $L_{\mathrm{ds}} = |L_{\mathrm{ds}}^{(2)}|$ is the dispersion length, $\zeta = z/L_{\mathrm{df}}$ is the normalized distance, $L_{\mathrm{df}} = k w_0^2/2$ is the diffraction length, p_{nl} is the normalized nonlinear polarization, and $\nabla_\perp^2$ is the transverse Laplacian. The presence of the operator $(1 + i\partial/\omega_0 \partial\tau)$ in the diffraction term $(\nabla_\perp^2 u)$ of eqn [4] leads to space–time focusing, while its presence in the nonlinear term results in self-steepening and both these terms arise from not making the slowly varying envelope approximation in time (i.e., $k_0 \partial A/\partial t \ll \partial^2 A/\partial t^2$) in deriving eqn [4]. For a self-consistent analysis of pulse propagation in the nonlinear regime, both the effects of space–time focusing and self-steepening must be included.

Nonlinear Refractive Index

For an isotropic medium in the highly nonresonant limit, the third-order term is the lowest-order contribution to the nonlinear susceptibility. This term gives rise to the nonlinear refractive index, that is, the index of refraction depends on the local intensity of the laser field. For many materials there are two contributions to the nonlinear refractive index: i) an instantaneous part that arises from the electronic response of the medium; and ii) a noninstantaneous contribution due to the nuclear motion of the molecules (i.e, the Raman contribution). For such a medium, the nonlinear polarization may be expressed as

$$p_{\mathrm{nl}}(\zeta, \tau) = \frac{L_{\mathrm{ds}}}{L_{\mathrm{nl}}}\left[(1 - f)|u(\zeta, \tau)|^2 + f\gamma_R \int_{-\infty}^{\tau} \mathrm{d}\tau' R(\tau - \tau')|u(\zeta, \tau')|^2\right] u(\zeta, \tau) \qquad [5]$$

where $L_{\mathrm{nl}} = (c/\omega_0 n_2 I_0)$ is the nonlinear length, $I_0 = n_0 c|A_0|^2/2\pi$ is the peak input intensity, f is the fraction of the Raman contribution to the nonlinear refractive index, and $R(\tau) = \left\{\left[1 + (\Omega_R \tau_R)^2\right]/\Omega_R \tau_R\right\} \exp(-\tau/\tau_R) \sin(\Omega_R \tau)$ is the Raman response function, τ_R is the characteristic Raman response time, Ω_R is the characterstic Raman frequency, and $\gamma_R = \tau_p/\tau_R$. For example, for fused silica $f = 0.15$, $\tau_R = 50$ fs, and $\Omega_R \tau_R = 4.2$. For a noble gas such as Xe, there is no Raman contribution and $f = 0$.

Self Focusing, Supercontinuum Generation, and Filamentation

The presence of the nonlinear refractive index with $n_2 > 0$ can lead to self-focusing of a laser beam. For sufficiently long pulses such that dispersion, self-steepening, and space–time focusing effects can be neglected, it is found that laser beams with input powers greater than the critical power $P_{\mathrm{cr}} = \alpha\lambda^2/4\pi n_0 n_2$ will undergo catastrophic self-focusing collapse. The dimensionless parameter $\alpha \geq 1.86$ depends on the spatial profile of the input beam and for a Gaussian input beam is given by $\alpha = 1.9$, in which case the ratio of the input power P to the critical power satisfies the relation $P/P_{\mathrm{cr}} = 1.055 L_{\mathrm{df}}/L_{\mathrm{nl}}$. Extensive studies have been made on the dynamics of laser beams undergoing self-focusing. For example, it has been shown that the shape of the collapsing beam evolves to a radially symmetric profile as it approaches the collapse point and that the total power contained in the collapsing portion always corresponds to the minimum value (i.e., $\alpha \sim 1.86$) regardless of the initial power in the beam.

For light pulses shorter than a picosecond, the role of material dispersion plays an important role and completely alters the dynamics of the self-focusing process. These dispersive effects lead to a temporal splitting of the pulse into two pulses and the arrest of its collapse. At even higher powers, other phenomena can occur, such as 'optical shock' formation at the rear edge of the pulse, due to self-steepening and space–time focusing. Shock formation leads to the emission of a broad spectrum of radiation extending from the ultraviolet to the mid-infrared, known as supercontinuum generation (SCG). This phenomenon was first observed in 1970, and since then it has been

observed in many different solids, liquids, and gases, under a wide variety of experimental conditions.

If the peak intensity of the pulse approaches intensities of $> 10^{13}$ W/cm^2, either through self-focusing or external focusing, multiphoton ionization occurs and a plasma is formed in the medium. (See section below on Plasma Nonlinearities and Relativistic Effects.) The generation of the plasma lowers the refractive index and effectively counteracts the self-focusing process, resulting in the spatial confinement of the pulse for distances far beyond what would be allowed by ordinary linear diffraction and has been observed in gases, liquids, and solids. A striking example of this apparent self-waveguiding is the observation of 'light strings' in air which can extend more than 10 km into the atmosphere. This phenomenon was first observed with 100 fs laser pulses in the near infrared ($\lambda = 800$ nm). Researchers found that pulses with energies greater than 10 mJ undergo self-focusing collapse in air and produce a highly intense ($> 10^{13}$ W/cm^2) 100-micron-diameter light filament tens of meters long.

Multiphoton Absorption

Multiphoton absorption is a process in which an atom or molecule makes a transition from a ground state to an excited state by means of the simultaneous absorption of N photons. In the lowest order of perturbation theory, such a process can be described by means of a susceptibility of order $(2N-1)$, that is, by $\chi^{(2N-1)}$. Alternatively, this process can be described in terms of an N-photon cross-section $\sigma^{(N)}$ defined such that the transition rate per atom is given by

$$R^{(N)} = \sigma^{(N)} I^N \qquad [6]$$

where I is the intensity of the laser field. Quantum mechanical expressions for the N-photon cross-sections are readily obtained. One finds, for instance, that

$$R_{\mathrm{ng}}^{(2)} = \left| \sum_m \frac{\mu_{\mathrm{nm}}\mu_{\mathrm{mg}}E^2}{\hbar^2(\omega_{\mathrm{mg}} - \omega)} \right|^2 2\pi\rho_{\mathrm{f}}(\omega_{\mathrm{ng}} - 2\omega)$$

$$R_{\mathrm{og}}^{(3)} = \left| \sum_{mn} \frac{\mu_{\mathrm{on}}\mu_{\mathrm{nm}}\mu_{\mathrm{mg}}E^3}{\hbar^3(\omega_{\mathrm{ng}} - 2\omega)(\omega_{\mathrm{mg}} - \omega)} \right|^2 2\pi\rho_{\mathrm{f}}(\omega_{\mathrm{og}} - 3\omega) \qquad [7]$$

In each of these expressions, the quantity ρ_{f} represents the density of final states, or equivalently the atomic lineshape function, evaluated at the N-photon transition frequency.

Optical Damage

High-intensity laser fields can produce unwanted damage to optical materials. As a point of reference, the threshold for laser damage to fused silica at a wavelength of 1.05 micrometers for a pulse of 30 ps duration corresponds to an intensity of 230 GW/cm^2 or a fluence of 7 J/cm^2. Over a wide range of pulselengths (approximately 1 ps to 1 μs), the threshold intensity for laser damage decreases with pulse length T as $T^{-1/2}$ and correspondingly the threshold fluence for laser damage increases with pulse length $T^{1/2}$. In this range of pulse durations, the dominant mechanism of laser damage is avalanche breakdown. In this process, free electrons are accelerated by the laser field until they acquire sufficient energy to impact-ionize other atoms in the sample. These additional electrons are similarly accelerated and create still more free electrons. The combined action of the breaking of chemical bonds and the deposition of heat energy leads to the fracturing of the optical material. For pulses shorter than 1 ps, processes such as multiphoton absorption and multiphoton dissociation contribute to the mechanism of optical damage. For laser pulses longer than approximately 1 μs (including continuous wave laser beams), the dominant damage mechanism is direct heating of the optical material by linear absorption.

High-Harmonic Generation

Let us consider how nonlinear optical effects are modified when excited by a super-intense pulse. Nonlinear optical effects are traditionally modeled using a power-series expansion, such as

$$P = \chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \cdots \qquad [8]$$

but this series is not expected to converge if the laser field strength E exceeds the atomic unit of field strength $E_{\mathrm{at}} = e/a_0^2 = 2 \times 10^7$ statvolt/cm $= 6 \times 10^9$ V/cm. This field strength corresponds to a laser intensity of $I_{\mathrm{at}} = 4 \times 10^{16}$ W/cm^2, which constitutes the threshold intensity for exciting nonperturbative nonlinear optical response.

One of the consequences of excitation with intensities comparable to the atomic unit of intensity I_{at}, is the occurrence of high-harmonic generation. In a typical experimental arrangement, a gas jet is irradiated by high-intensity laser radiation, and all odd harmonics of the fundamental laser frequency, up to some maximum value N_{max}, are observed. The various harmonics below N_{max} are typically emitted with approximately equal intensity; such an

observation is incompatible with a perturbative explanation of this phenomenon. Recent work has demonstrated harmonic generation with N_{max} as large as 341.

The phenomenon of high-harmonic generation can be understood in terms of a simple physical model. One imagines an atomic electron that has received kinetic energy from the laser field and is excited to a highly elliptical orbit. The positively charged atomic nucleus is at one focus of this ellipse, and each time the electron passes near the nucleus it undergoes strong acceleration and emits a short pulse of radiation. This radiation will occur in the form of a train of short pulses; the spectrum of the radiation will be the square of the Fourier transform of this pulse train, which will contain the odd harmonics of the oscillation frequency up to some maximum frequency, that is approximately the inverse of the time the electron spends near the atomic core. This argument can be made quantitative to show that the maximum harmonic number is given by

$$N_{max}\hbar\omega = 3.17K + U_p \qquad [9]$$

where $K = e^2E^2/m\omega^2$ is the 'ponderomotive energy' (the kinetic energy of an electron in a laser field) and U_p is the ionization energy of the atom.

Plasma Nonlinearities and Relativistic Effects

The process of multiphoton ionization can liberate a sufficient number of free electrons to transform the optical medium into a plasma, that is, a fully or partially ionized gas. The process of plasma formation is described by the equation

$$\frac{dN_e}{dt} = \frac{dN_i}{dt} = (N_T - N_i)\sigma^{(N)}I^N - rN_eN_i \qquad [10]$$

where N_e is the number density of electons, N_i is the number density of ions, N_T is the total number of atoms (both ionized and un-ionized) in the material, and r is the electron–ion recombination coefficient. The optical properties of plasmas are very different from those of typical dielectric materials; the plasma contribution to the dielectric constant is given by

$$\epsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2} \qquad [11]$$

where $\omega_p = \sqrt{4\pi Ne^2/m}$ is known as the plasma frequency.

Nonlinear effects can occur in the propagation of light through a plasma. One example is the nonlinear response resulting from the relativistic change in mass of the electron due to the large velocity that it can attain in the field of an intense laser beam. Detailed consideration of this effect shows that the nonlinear change in refractive index can be described as $\Delta n = n_2 I$ where

$$n_2 = \frac{2\pi\omega_p^2e^2}{n_0^2m^2c^3\omega^4} = \frac{1}{2\pi n_0^2}\left(\frac{\omega_p}{\omega}\right)^2 1.1\times10^{-26}\,\frac{\text{cm}^2}{\text{W}} \qquad [12]$$

Nonlinear Quantum Electrodynamics

One can imagine an electric field so intense that it could lead to the spontaneous creation of electron–positron pairs. Such a field would have a strength of the order of $E_{QED} = mc^2/e\lambda\!\!\!^{-}_c$ where $\lambda\!\!\!^{-}_c = \hbar/mc$ is the reduced Compton wavelength of the electron. The intensity of a light beam with a peak field amplitude of E_{QED} is $I_{QED} = 4\times10^{29}$ W/cm^2. Detailed consideration shows that even for fields weaker than I_{QED} there will be a field-induced change in the dielectric tensor given by

$$\epsilon_{ik} = \delta_{ik} + \frac{e^4\hbar}{45\pi m^4c^7}\left[2(E^2 - B^2)\delta_{ik} + 7B_iB_k\right] \qquad [13]$$

Because of the unusual tensor properties of this relation, it displays a different polarization dependence than typical optical nonlinearities. Nonetheless, to an order of magnitude one can describe the strength of this response as

$$n_2 = \frac{7}{15c}\frac{e^2}{\hbar c}\frac{1}{E_{QED}^2} = 5.6\times10^{-34}\,\text{cm}^2/\text{W} \qquad [14]$$

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Quantum Electrodynamics:** Quantum Theory of the Electromagnetic Field.

Further Reading

Agostini P, Fabre F, Mainfray G, Petite G and Rahman NK (1979) Free-free transitions following six photon ionization of xenon. *Physics Review Letters* 42: 1127.

Alfano RR (ed.) (1989) *The Supercontinuum Laser Source*. New York: Springer-Verlag.

Bloembergen N (1997) A brief history of light breakdown. *Journal of Nonlinear Optical Physics and Materials* 6: 377.

Boyd RW (2003) *Nonlinear Optics, Second Edition*, Section 12.5. Amsterdam: Academic Press.

Braun A, Korn G, Liu X, Du D, Squier J and Mourou G (1995) Self-channeling of high-peak-power femtosecond laser pulses in air. *Optical Letters* 20: 73.

Brabec T and Krausz F (2000) Intense few-cycle laser fields: Frontiers of nonlinear optics. *Reviews of Modern Physics* 72: 545.

Chang Z, Rundquist A, Wang H, Murnane MM and Kapteyn HC (1997) Generation of coherent soft X-rays at 2.7 nm using high harmonics. *Physics Review Letters* 79: 2967.

Corkum PB (1993) Plasma perspective on strong field multiphoton ionization. *Physics Review Letters* 71: 1994.

Corkum PB, Rolland C and Srinivasanrao T (1986) Supercontinum generation in gases. *Physics Review Letters* 57: 2268.

Gaeta AL (2003) Collapsing light really shines. *Science* 301: 54.

Kasparian J, Rodriguez M, Mejean G, Yu J, Salmon E, Wille H, Bourayou R, Frey S, Andŕe Y-B, Mysyrowicz A, Sauerbrey R, Wolf J-P and Woste L (2003) White light filaments for atmospheric analysis. *Science* 301: 61.

Lewenstein M, Balcou Ph, Ivanov M Yu, L'Huillier A and Corkum PB (1994) Theory of high-harmonic generation by low-frequency laser fields. *Physics Review A* 49: 2117.

Max CE, Arons J and Langdon AB (1974) Self-modulation and self-focusing of electromagnetic waves in plasmas. *Physics Review Letters* 33: 209.

Moll KD, Fibich G and Gaeta AL (2003) Self-similar wave collapse: observation of the Townes profile. *Physics Review Letters* 90: 203902.

Monot P, Auguste T, Gibbon P, Jakober F, Mainfray G, Dulieu A, Louis-Jacquet M, Malka G and Miquel JL (1995) Experimental demonstration of relativistic self-channeling of a multiterawatt laser pulse in an underdense plasma. *Physics Review Letters* 74: 2953.

Rairoux P, Schillinger H, Niedermeier S, *et al.* (2000) *Applied Physics B—Lasers Optics* 71: 573.

Ranka JK, Schirmer RW and Gaeta AL (1996) Observation of pulse splitting in nonlinear dispersive media. *Physics Review Letters* 77: 3783.

Rothenberg JE (1992) Pulse splitting during self-focusing in normally dispersive media. *Optical Letters* 17: 583.

Sprangle P, Tang C-M and Esarey E (1987) Relativistic self-focusing of short-pulse radiation beams in plasmas. *IEEE Transactions on Plasma Science* 15: 145.

Stuart BC, Feit D, Rubenchik AM, Shore BW and Perry MD (1995) Laser-induced damage in dielectrics with nanosecond to subpicosecond pulses. *Physics Review Letters* 74: 2248.

Wagner R, Chen S-Y, Maksemchak A and Umstadter D (1997) Electron acceleration by a laser wakefield in a relativistically self guided channel. *Physics Review Letters* 78: 3125.

Wood RM (1986) *Laser Damage in Optical Materials*. Bristol, UK: Adam Hilger.

NONLINEAR SOURCES

Harmonic Generation in Gases

P Villoresi, University of Padova and INFM, Padova, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The generation of laser harmonics in gases provides high-brightness ultraviolet and soft X-ray radiation induced by an intense laser pulse. This phenomenon was discovered experimentally in the late 1980s, in a simple setup as that sketched in Figure 1. The laser pulse is focused in the interaction medium, which may be a gas jet formed by a pulsed electromagnetic valve, a capillary, or a small chamber filled with the gas. The beam leaving the interaction region includes a part of the laser pulse and a number of odd harmonics.

The harmonic orders generated in this way may reach and exceed an order of 300, so that this mechanism, often referred as high-order harmonics (HHs) generation (HHG), directly up-converts low frequency radiation, usually in the near-infrared spectrum, to extreme-ultraviolet (XUV) or soft X-rays. An example of a HH spectrum is shown in Figure 2, where several tens of discrete harmonic peaks are present. The striking difference with ordinary nonlinear processes is the flatness of the photon yield versus nonlinear order, up to a sharp cut-off point.

The HHG is based on the sequence of three processes: the electric field of an intense laser pulse ionizes atoms placed near the laser focus; the liberated electrons are driven by the laser electric field in a trajectory which returns on the parent ion; and the electron recombines releasing a photon whose energy is the ionization potential plus the electron's kinetic energy. The process is periodic in time, with the period of the laser pulse, and the emitted spectrum is then expressed by discrete components, multiples of the laser frequency. The emission is in the form of a traveling wave in the direction of the laser beam and so only odd harmonics are present.

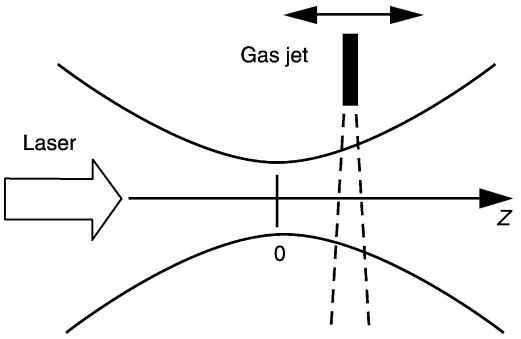


Figure 1 Scheme of the interaction between a focused laser pulse and a gas jet. *z* axis originates at laser focus.

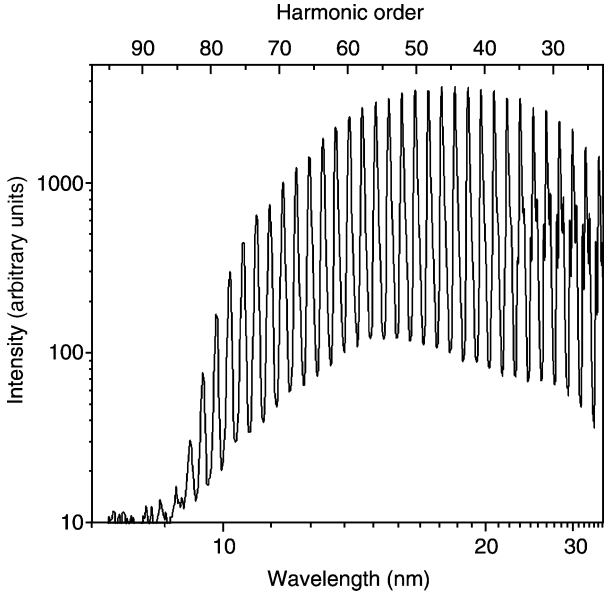


Figure 2 HHs experimental spectrum, obtained from helium gas, and a Ti:sapphire laser pulse at 790 nm, with 25 fs duration and 6×10^{14} W/cm^2 intensity.

The individual processes in HHG have highlighted the coherent response of electrons in the laser field, that show up interferences between quantum trajectories, as well as a nonlinear phase-matching issue – which have provided a simple method to investigate the interaction of single particles with the electromagnetic field, where plasma effects and bonds may be of minor influence.

Background

The interest to produce laser harmonics using gases instead of crystals was expressed by many. For the gas medium; the mass-density is three order of magnitude lower than for solids and the nonlinear susceptibility can be strongly enhanced for suitable resonance frequencies. Moreover, the phase-matching of the phase velocities may be achieved by tuning the gas pressure or by mixing diverse gases, so realizing a highly homogenous conversion region, free from walk-off and with an extension much longer than for crystals. In addition, gases have a high threshold for the intensity-induced damage, which is also non-permanent.

This research was made possible with the introduction of chirped-pulse laser amplifiers in the early 1980s, when picosecond lasers pulses reached intensities capable of directly ionizing solids and matter with their own electric field. This effect, known as optical ionization, was further enhanced when the laser pulses were reduced to subpicosecond duration, up to a few femtoseconds (10^{-15} s), close to their natural limit, which is the optical cycle. The experimental observation of HHs and of the ejection of electrons with unexpected high energy, the above threshold ionization (ATI), dramatically changed the scenario of laser matter interaction and has indicated a new regime in the coupling of laser radiation with the matter. The ordinary nonlinear optics, using the perturbative approach, was not capable of interpreting the findings and so the request was made for new models for the description of the interaction, which has been now named nonperturbative.

Typical Experimental Parameters

In the experimental investigation of the HHG, many laser wavelengths and pulse durations were used, including radiation from excimer, visible, Ti:sapphire, Nd:YAG (first and second harmonics) and CO_2 lasers, and pulses of duration from hundreds of picoseconds to a few femtoseconds. Among these, the more effective strategy in the generation of high harmonic order with high conversion efficiency, has been that of near infrared pulses with shortest duration. In order to induce optical single-ionization, the required laser intensity needs to span in the 10^{14} W/cm^2 range. Different elements are a possible choice for the gas in the interaction medium, though noble gases are used for the generation of the highest orders, since these have the higher ionization potential. Due to the very short duration of the laser pulse, the gas temperature is not relevant, while a pressure range of 10–500 mbar is usually chosen to balance, on one side, the needs of a large number of atoms that take part in the HHG and on the other side, the limit posed by the attenuation of the XUV and soft X-ray radiation after the generation, caused by the gas absorption. For the same reason, a powerful pumping system is needed, in order to reduce the ambient pressure in the apparatus below 10^{-5} mbar level.

The Models for the High Harmonics Generation

The physical processes involved in HHG are many: in the emission of the HH photons, the laser field interacts with a single atom of gas and induces the ionization, the electron's motion and its recollision with its parent ion; besides, the beam of HHs is the result of the superposition of the radiation from all the ionized atoms, as in an ordinary nonlinear medium, where the phase velocity of the laser field and of HHs is a function of the position and of the laser intensity and phase; therefore, the phase-matching issue has to be addressed in order to understand the global yield of HHG. The more relevant causes for the mismatch are the geometrical phase shift at the beam focus, which for Gaussian-profile beams is the Gouy phase, the nonlinear phase from the atomic polarization, and the effects of the plasma induced by the laser pulse itself. In addition, other effects have to be considered, such as the nonuniform density distribution of the target gas or the fact that the propagation of XUV radiation through a gas is usually influenced by a strong absorption, which is frequency dependent.

The study of HHG is first approached by introducing what is known as the simple man model, in which simplifications are introduced in order to treat the problem in a simple classical scheme, but from which the essential characteristics are derived. We then introduce the more complex quantum model based on the path integrals.

First Step: Ionization

The first step in the HHG model is the process with which the laser electric field applies to an electron, belonging to an atom or a molecule, a force that causes its detachment. This takes place only if such a field is strong enough to liberate the electron from its atomic bond. This may happen when the laser, which is taken here as linearly polarized and periodic in time, has modified the potential well so that the electrons may tunnel through it or pass over it. In fact, the sum of the atomic potential with that associated to the laser, which can be thought of as a plane that swings between a positive and negative slope at the laser frequency, gives rise to a finite sized well, through which the electron escapes by tunneling, or eventually to an aperture from which the electron is drawn away from the nucleus. In Figure 3, is shown such a well and the tunnel path for the electron.

Once freed, the electron responds to the periodic force from the laser which makes it oscillate. The

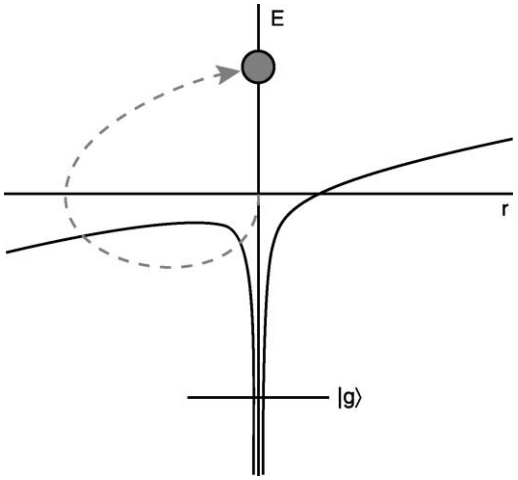


Figure 3 Electron potential when the laser pulse shines on the atom; |g⟩ indicated the level of the ground state; in red, a possible trajectory of the liberated electron after tunneling.

average kinetic energy on such oscillations is known as the ponderomotive potential U_p, and is related to the laser intensity and wavelength, according to the following: $U_p = e^2E_0^2/4m\omega_0^2$, which reads numerically as $U_p(\text{eV}) = 9.3I_0(10^{14}\ \text{W/cm}^2)\lambda^2(\mu\text{m})$, where λ is the laser wavelength, E_0 and I_0 the laser peak electric field and intensity.

Second Step: Free-electron Trajectory

On the freed electron are applied both the time-varying force due to the laser field and the ionic potential. Its motion can be approximated according to the simple man model as a nonrelativistic classical-mechanics trajectory; its initial velocity is set to zero, and its initial position is by the ion; finally, the laser exerts a force much stronger than that from the ion as well as from the laser magnetic field, thus these two latter forces are cancelled out.

As mentioned above, the electron trajectories that induce HHs are those where the electron returns by the ion with strong kinetic energy. If the laser is linearly polarized, the induced trajectory takes place along a line, at which also belongs to the parent ion. Moreover, the electron final velocity is related to the actual instant of ionization, or phase of the field, with respect to the laser peak. Therefore, electron final energies span from zero to a maximum value $K_{\text{max.}}$ By considering the first semi-cycle, K_{max} occurs at about 17° of field phase angle, while for the second semi-cycle it is at 197°.

In Figure 4 the electron trajectories for different ionization in the first semi-cycle are shown: the electron displacement in the laser-field direction with

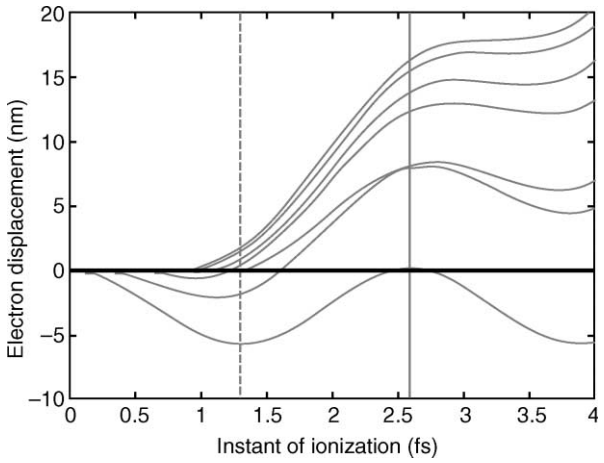


Figure 4 Electron trajectories relative to different ionization instant within the first semi-cycle; in dotted red, half-cycle, in continuous red, the whole optical cycle. Trajectories for later times do not return to the ion, do not contribute to HHG and correspond to ejected photoelectrons.

respect to the ion position, which is the initial as well as the final position, is represented versus time, where the time origin is at the laser peak, where field phase is equal to zero. The laser intensity directly influences the value of $K_{\max}$, according to the following expression: $K_{\max} = 3.17\,U_p$. In the case shown in the figure, the peak laser intensity is $I_0 = 8 \times 10^{14}$ W/cm^2 and the ponderomotive potential is $U_p = 45$ eV and $K_{\max} = 143$ eV.

Moreover, for a given final energy below $K_{\max}$, there are more than one value of the phase of the laser field where the trajectory results with the same final kinetic energy. This indicates that multiple trajectories may produce a given HH, and this is relevant in terms of the interference between them.

Third Step: Recombination

The final process which occurs in the HHG, is the recombination of the ionized electron with the parent ion. The energy conservation means that the kinetic energy and the ionization potential have to be delivered to the emitted photon. In the simple man model, the spectrum of HH then spans from I_p to the maximum values of $E_{\max} = I_p + 3.17\,K_{\max}$. In the case of Figure 2, obtained using helium gas at 6×10^{14} W/cm^2, the spectral position of cutoff can be estimated at about 9 nm, or 140 eV, which is in good agreement with the prediction of the relationship described above.

The electron recombination is not a unique process occurring at this point: the electron may be scattered by the ion and be emitted with a kinetic energy of up to 10 U_p, which is known as the rescattering plateau in ATI, or it may ionize the ion by collision, leaving a double ionized atom via sequential ionization, with a strong increase in the observed cross-section of this process.

Feynman Path-Integral Approach

The ideas in the simple man model have been expanded in order to include the quantum mechanical description of the laser-electron-ion interaction. By using the Feynman path-integral approach, the evolution of the electron driven by the laser force is associated to a probability which depends on the sum of an infinite number of amplitudes $\mathbf{f}_n$. These are complex numbers and correspond to different trajectories, or quantum orbits, $\mathbf{r}_n(t)$, followed by the electron. Their phase is proportional to the classical action of the electron motion, $S[\mathbf{r}_n(t)]$, calculated along $\mathbf{r}_n(t)$, which begins at time t_i and ends at a later time $t_f = t_i + \tau$. Here τ defines the duration of the liberated electron trajectory. In our case, the action is the sum of three terms:

$$S[\mathbf{r}_n(t)] = S_{\text{bound},n} + S_{\text{free},n} + S_{\text{final},n} \qquad [1]$$

that correspond respectively to the electron ionization:

$$S_{\text{bound},n} = -E_0 t_{i,n} \qquad [2]$$

where E_0 is the binding energy of electron in the atom ground state, $E_0 < 0$, to the free electron trajectory:

$$S_{\text{free},n} = \int_{t_i}^{t_f} [\mathbf{k} - e\mathbf{A}(t)^2]\mathrm{d}t \qquad [3]$$

where $\mathbf{k}$ is the electron momentum after ionization, directed parallel the laser electric field, and $\mathbf{A}(t)$ is the vector potential of the laser field, and to the final recombination with the parent ion:

$$S_{\text{final},n} = (\hbar\omega_q + E_0)t_{f,n} \qquad [4]$$

The condition that normally points out the trajectory made by the electron may be obtained from the principle of least action. Within the path integral approach, this principle can be used to deduce some equations which have to be obeyed by the quantum orbits, and that express the energy conservation in both the ionization and recombination process and the fulfilment of boundary conditions for electron trajectory. These are known as saddle-point equations, and reads as:

$$[\mathbf{k} - e\mathbf{A}(t_i)]^2 = 2mE_0 \qquad [5]$$

for the ionization process:

$$\mathbf{k}\tau = e\int_{t_i}^{t_f} \mathbf{A}(t)\mathrm{d}t \qquad [6]$$

which compares the initial momentum with that exchanged with the laser field during the trajectory, and

$$[\mathbf{k} - e\mathbf{A}(t_f)]^2 = 2m(\hbar\omega_q + E_0) \qquad [7]$$

for the recombination process which emits an harmonic of order q.

From eqn [5] we note that the solution for the ionization time t_i is a complex number, since the right hand side is negative. This suggests that the quantum orbits are complex as well. The reason for this is that the tunnel through the potential well is a nonclassical process.

The orbits solutions to eqns [5]–[7] have different durations, and for some of them τ is shorter than an optical cycle, while for the others the electron returns in a later cycle than the one where it was ionized. This latter solution is of minor influence, since the electron wavefunction spreads out along the trajectory, decreasing the recombination probability. Those within the initial optical cycle sum the amplitudes to give the probability of the generation of q-th harmonic, and their interference is evident from the irregular single-atom spectrum.

The time evolution of the laser pulse as well as the interaction geometry are experimental tools that influence the relevance of the different trajectories. In particular conditions it is possible to set the condition for HHG from a single trajectory, with a very simple phase structure and clear spectrum; some cases are described below.

HHs Generated by Few-Optical-Cycle Pulses

The laser electric field considered so far has been taken as stationary, relying on the hypothesis that the pulse envelope has a duration much longer than the laser optical cycle. For the analysis of the electron trajectories, that last a fraction of the laser optical cycle, this hypothesis is valid for pulses down to about 20 fs. For shorter pulses, as those generated with the hollow fiber technique, there is a noticeable variation of the electric field between a cycle and the next. The generation process is now aperiodic, and the characteristics of HH spectrum and conversion is strongly influenced by the pulse actual duration.

In order to show the striking changes in this regime, Figure 5 shows the laser electric field, the envelope and the ionization probability of helium gas in the case of a pulse of 5 fs of full width at half-maximum (FWHM) duration and a laser intensity of 9×10^{14} W/cm^2. The atom can be ionized during only three intervals, of sub-fs duration but where only one is dominant.

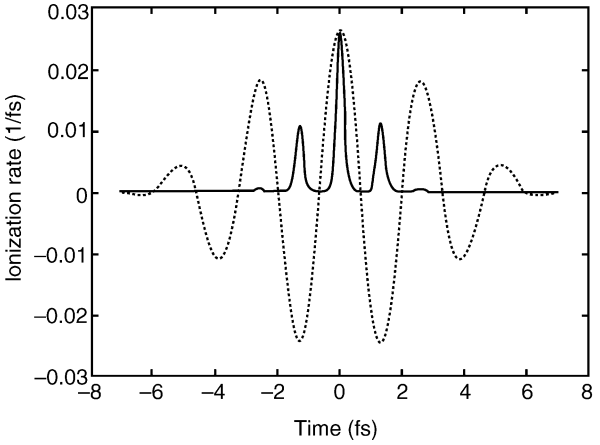


Figure 5 Ionization rate for helium gas shined by a 5 fs FWHM pulse of intensity 9×10^{14} W/cm^2. The dotted line shows the normalized evolution of electric field of the pulse. Only the three central half-cycles induce ionization.

The HHG experiment in this temporal regime have revealed that the influence of the plasma in the interaction region is progressively reduced and eventually vanishes for pulses of duration of about 5 fs. This is because the atoms are exposed directly to the maximum electric field, and so there is no pre-ionization of the gas caused by the initial part of the pulse. This effect, which is present in the case of many-cycle pulses, causes a reduction in the HHG efficiency, due to a depletion of the neutral density available at the time of maximum electric field, where the higher HHs are produced with a large dispersion, which deteriorates the phase-matching.

Moreover, the effect of the carrier-envelope-phase in the HHs spectrum has been demonstrated as an evident spectral shift of the cut-off harmonics generated with few-cycle pulses. The reason can be found in the variation of the phase of the emission from each electron after its trajectory. For the few-cycle pulses, this phase changes for every laser semi-cycle, due to the change in the actual electric field which drives the electron motion.

This portion of the spectrum is also relevant to its phase structure, since here the HHs are generated by a single electron trajectory. This condition sets a constant phase difference between the subsequent harmonic orders. In this way, similarly to the mode locking laser, this effect allows the synthesis of a pulse, from the combination of several harmonic orders, which has a duration much lower than the

individual harmonics. This process has been exploited to produce a soft X-ray pulse of duration of 650 as (1 as, attosecond $= 1 \times 10^{-18}$ s).

HHs Beam Characteristics

The HHG is a coherent process, in that the emitted radiation has a direct relationship with the laser pump. The emitted HH beam also reflects this fact in its good directional characteristics and in its spatial and temporal coherence, as different experiments have demonstrated.

Using a 110 fs laser pulse, the interference fringes produced in the far field of two spatially separated HH beams were measured as a function of the mutual delay. The HH coherence time t_c resulted in agreement with the expected duration of the HHs, which is a fraction of that of the laser pulse. Moreover, it has been observed at a different t_c in the central and external regions of the far field of harmonic 23rd, 50 fs and 20 fs respectively. This is because different phase-matching in the HHG contribute to the on-axis and off-axis part of the HHs beam, where the more prominent quantum orbit is also different.

The measurement of the beam divergence for different harmonic orders is relevant for both the estimate of the beam brightness and the irradiance available in the case of the HH applications. Moreover, it is useful for the understanding of the phase-matching which is realized in actual interaction geometry. The experimental measurements requires an astigmatic spectrometer, as is described later. In this kind of experiment, the focalplane images of the HH spectrum are taken for different generation conditions, usually moving the gas-jet position with respect to the laser focus, the z coordinate in Figure 1. The harmonic peaks appear as elongated spots, as shown in Figure 6 for helium gas and an intensity of 4×10^{14} W/cm^2 and duration of 20 fs, whose length in a transverse direction to the dispersion plane can be related, by a optical calculation, to the beam divergence.

A comparison between divergences for HHG driven by few- or many-optical-cycle pulses has revealed another aspect of the phase-matching condition. In this case, the laser pulse was compressed in time by means of the hollow fiber technique, and so the resulting phase front is of the Bessel-beam type, truncated at the first zero. The experiment indicates a divergence for plateau harmonics of about 3 mrad FWHM of the beam, in the case of the 7 fs pulse and larger values by more than 50% using the 20 fs pulse. In both cases, the beam divergence increase moving downstream the gas jet and with increasing harmonic order. The reduced divergence can be ascribed to a more regular wavefront of the harmonic beam, which can be related to the simpler phase structure of the few-cycle case.

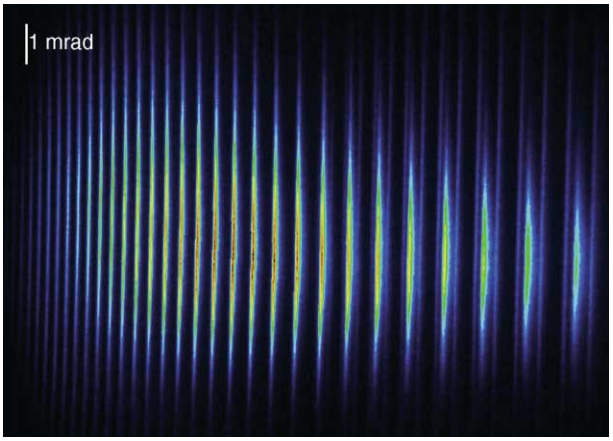


Figure 6 Focal plane image obtained from the spectrograph shown in Figure 9, relative to the HHs spectrum of helium with 20 fs and 4×10^{14} W/cm^2. The bar in upper-left corner indicates the scale-length of vertical axis corresponding to a divergence of the HH beam of 1 mrad.

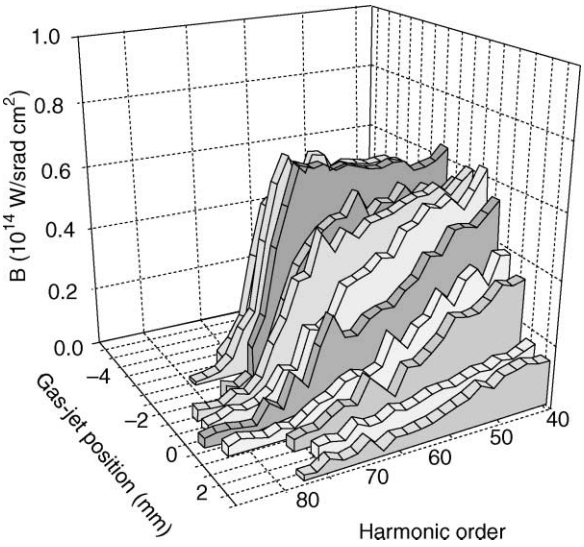


Figure 7 Peak brightness of the HHs beam determined from absolute-intensity and divergence measurements, using neon gas and a 790 nm laser pulse with 7 fs duration and 9×10^{14} W/cm^2 intensity.

By using an intensity-calibrated detector, the number of photons per harmonic order per laser shot can be measured. Combining this measure with that of the divergence, the HH beam brightness can be calculated. The source size and the HH pulse duration can be easily modeled from the experimental parameters. The result for the case of neon with 7 fs pump-pulse with intensity of 9×10^{14} W/cm^2 is shown in Figure 7, as a function of the gas-jet

position, with positive values when the jet is downstream with respect to laser focus.

The large divergence reduces the brightness of downstream positions. Maximum values of about 8×10^{13} W/srad cm^2 are found close to the focus, relatively flat with respect to harmonic order.

The conversion efficiency η_{HH} from the fundamental intensity to that of the HHs can also be derived from this kind of measurement. The more convenient parameters for HHG in the range 41–71 were found using neon gas jet from a pulsed valve and a pulse duration of 7 fs, which leads to η_{HH} in the scale of 10^{-7}, while for lower orders, from 17 to 31, a value about two orders of magnitude larger has been obtained using a short hollow fiber filled with argon gas. A remarkable intracycle optimization mechanism was introduced in this case, in order to maximize the given harmonic order of chosen. This was done by tailoring the shape of the pump pulse in spectrum and time, using an adaptive component in the laser compressor. In this case, the optimal pulse shape to maximize the efficiency in the driving of the electron on a trajectory producing the required harmonic order, is sought by means of a genetic algorithm.

A strong efficiency enhancement has also been demonstrated by manipulating the laser wavefront by means of a deformable mirror (DM). In this case, the DM was controlled by a genetic algorithm, aiming to increase the HHs intensity in a given spectral interval. The purpose of this technique is the correction of the optical aberrations on the beam, introduced by both the optical components in the beam focusing line and the phase distortion caused by nonlinear effects leading to an optimum laser front. The effect is shown in Figure 8, where the spectrum obtained with the original beam is compared with that optimized by the genetic algorithm. It can be noticed there is both a remarkable extension of the HHs spectrum, of about 20 orders, as well as a strong enhancement of the plateau intensity.

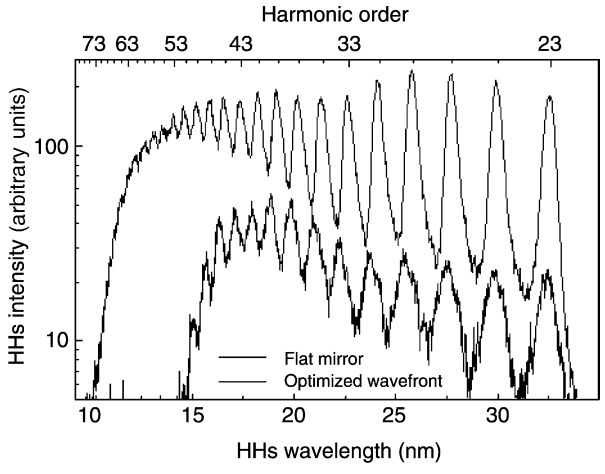


Figure 8 Comparison of the HHs spectrum obtained with and without the wavefront correction using a deformable mirror. Neon gas was used and a 790 nm laser pulse with 6 fs duration and 0.3 mJ energy.

Spectrometers and Monochromators for HHs

The extended spectrum of HHs reaches the optical domains of the extreme-ultraviolet (XUV, 10–100 nm) and soft X-ray (1–10 nm). In these regions, the optical materials have very poor or zero transmission, and so the optical design of instruments for HHs analysis uses reflective optics only. Exceptions sometimes used are the frequency-selective XUV transmission filters, whose thickness does not exceed a few tenths of a micrometer, and the diffracting zone-plates. Moreover, for the radiation of wavelength shorter than 30 nm, even the reflectivity of optics becomes poor, and this forces the use of grazing-incidence mountings or the narrow band XUV and soft X-ray multilayer mirrors.

On the other hand, HHs as optical sources have favorable characteristics, in that they are emitted in a region smaller than the laser waist focus, so usually of a few tens of micrometers, and their divergence is small, as seen above. In addition, the very short duration of the pump pulses broadens the HHs spectrum, due to time-bandwidth inequality, usually smearing details below 0.1 nm of width.

The spectrograph is then realized with the least number of components, usually combining in one or two optics of the spectral dispersion, the gathering of the HH beam and its focusing on the detector. Moreover, it is not feasible to collect with a single detector all the broad HH spectrum in once, so usually a spectral interval of interest is selected and acquired with a linear or bidimensional detector, or scanned sequentially by rotating the grating.

An example of spectrograph with 2D-detector, is shown in Figure 9, the layout of the instrument developed for the simultaneous measure of the divergence and the intensity of HHs generated from few-optical-cycle pulses.

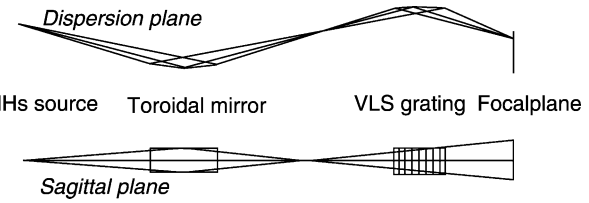


Figure 9 Optical layout of the astigmatic spectrometer for the divergence measurement.

The upper part of Figure 9 shows the optics in the dispersion plane, where the toroidal mirror gathers the HH radiation and relays it to the source point for the spherical grating. The latter has a varied line spacing (VLS) of the grooves, in order to image an erected spectrum in the focalplane. This feature is convenient in order to use a bidimensional detector. In the sagittal plane, the grating has negligible effect, and the toroidal mirror is designed to realize an astigmatic image of the source in the focalplane, from which the divergences can be extracted. The detector needs to be blind at the radiation of the laser and at the lower harmonics, and to have an extended linear response in the XUV and soft X-ray. A good solution can be obtained with an open micro-channel plate (MCP), with a suitable photocathode such as MgF_2, placed in the focalplane, coupled with a high-resolution and a high dynamic CCD camera.

For the selection of a single harmonic order or of a narrow spectral region, an XUV monochromator is used. This device is based on the spectral dispersion of the HHs beam by using a diffraction grating. The desired harmonic order is selected by scanning the wavelength which falls on the slit, usually by rotating the grating. This technique uses a single optical component, a concave grating, to reduce the reflection losses. On the other hand, the single-grating monochromator introduces a remarkable time broadening in the selected radiation. In fact, the grating diffraction is based on the optical path length difference for the rays diffracted by subsequent grating grooves. Therefore the selected beam, of wavelength λ, is composed of rays whose difference in length is $\Delta L = N_g m\lambda$, where m is the diffraction order and N_g is the number of illuminated grooves, given by the beam diameter on grating multiplied by the groove density. The time broadening of the selected pulse is then $\Delta\tau = \Delta L/c$. For a HH XUV pulse, the diffracted pulse results is 7 ps of duration, if beam size is of 3 mm and it is diffracted at 87° by a 1200 gr/mm grating.

The solution to this problem may be found in adopting a twin grating design. In this case, the dispersion induced by the first grating is used to select the desired spectral portion of the HH spectrum, and will be cancelled by the second grating. This latter grating has to be mounted in order to compensate for the different optical path length of different rays within the aperture of the HH beam. The layout of this optical scheme is reported in Figure 10. The ray tracing of this compensated monochromator indicates that for both the grazing incidence and the normal incidence regions, the residual broadening results is of the order of one femtosecond.

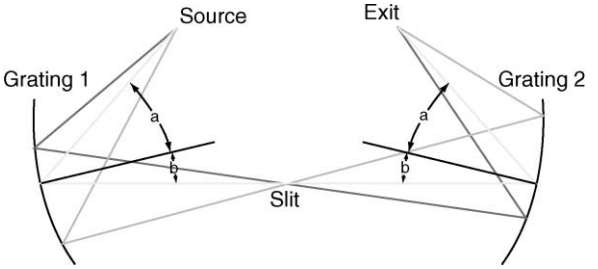


Figure 10 Scheme of the compensated XUV monochromator, for the selection of a spectral portion of the HHs spectrum without introducing temporal broadening. See also Villoresi P (1999) Compensation of optical path lengths in extreme-ultraviolet and soft X-ray monochromators for ultrafast pulses. *Applied Optics* 38: 6040.

HHs at Work: Applications and Perspectives

The advantages of HHs with respect to other XUV and soft X-ray sources, as laser-plasma, synchrotron radiation or X-ray lasers, are mainly the very short duration of their emission, their high brightness, and the source compactness and their intrinsic synchronization with a laser pulse. These aspects have been already exploited in a few experiments, based either on the detection of a photoelectron emitted in the time-dependent interaction of the matter with the HHs plus fundamental laser beams, or the observation of changes in the HHs XUV spectrum.

As examples of direct application, the interferometry of a laser-plasma has been obtained with HHs, to determine spatial distribution of the electron density. In this case, two beams of harmonic radiation were generated by splitting in equal parts the laser pulse, with one of them passing through a laser-generated plasma from a solid aluminum target. By means of a grating monochromator, a single harmonic order is selected, of both beams, and their interference pattern in the far field is acquired. From the shift in the fringes pattern, the map of the plasma electron density was estimated. Here, the broad spectrum of the HHs was exploited to operate the interference at the most convenient wavelength, far from atomic resonances and with fair fringe contrast.

The direct monitoring of a chemical reaction at the surface was achieved by pumping, with a part of the laser pulse, a platinum surface with some adsorbed molecular oxygen. By observing changes in the spectrum of the photoelectrons generated by the delayed HHs beam, the oxygen-platinum bond switch from a superoxo (O_2^-) to a peroxo (O_2^{2-}) state was observed in the time domain.

The electron bond in an atom may be dressed by a laser field, and the changes induced in its orbital are diagnosed by a HHs beam, which photo-ionizes the atom. This phenomenon have been exploited as a diagnostic of the temporal structure of a HHs beam: the sidebands in the photoelectron spectrum have been related to the phase of different harmonics. The findings indicate that HHs are locked in phase and their superposition produces a train of 250 as pulses.

By using the mixing scheme of some suitable orders of a HHs beam generated with few-optical-cycle pulses, mentioned already in section 'HHs beam characteristics,' a pulse with sub-femtosecond duration may be generated. This was used in the detailed analysis of the photoelectron spectrum in a pump-probe experiment, with time resolution of 150 as. Again using the IR and HHs radiation together, this technique has accessed the direct probe of the laser field oscillation, as demonstration by the use of HHs in the new domain of attosecond spectroscopy and metrology.

Finally, the HH radiation can also be considered as the seed for an amplifier of radiation in the soft X-ray region, as in the case of a free electron laser.

See also

Coherent Transients: Ultrafast Studies of Semiconductors. **Instrumentation:** Spectrometers. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Bartels R, Backus S, Zeek E, *et al.* (2000) Shaped-pulse optimization of coherent emission of high-harmonic soft x-ray. *Nature (London)* 406: 164.

Bauer M, Lei C, Read K, *et al.* (2001) Direct observation of surface chemistry using ultrafast soft-X-ray pulses. *Physics Review Letters* 87: 025501.

Brabec T and Krausz F (2000) Intense few-cycle laser fields: frontiers of nonlinear optics. *Review Modern Physics* 72: 545.

Corkum PB (1993) Plasma perspective on strong-field multiphoton ionization. *Physics Review Letters* 71: 1994.

Descamps D, Lynga L, Norin J, *et al.* (2000) Extreme ultraviolet interferometry measurements with high-order harmonics. *Optic Letters* 25: 135.

Hentschel M, Kienberger R, Spielmann CH, *et al.* (2001) Attosecond metrology. *Nature (London)* 414: 509.

Kulander KC, Schafer KJ and Krause JL (1993) Dynamics of short-pulse excitation, ionization and harmonic conversion. In: Piraux B, L'Huillier A and Rzazewski K (eds) *Super-Intense Laser-Atom Physics*, pp. 95–110. New York: Plenum.

Lewenstein M, Balcou Ph, Ivanov M, *et al.* (1994) Theory of high-harmonic generation by low-frequency laser fields. *Physics Review A* 49: 2117.

L'Huillier A and Balcou P (1993) High-order harmonic generation in rare gases with a 1-ps 1053-nm laser. *Physics Review Letters* 70: 774.

McPherson A, Gibson G, Jara H, *et al.* (1987) Studies of multiphoton production of vacuum-ultraviolet radiation in the rare gasses. *Journal of the Optical Society of America B* 4: 595–601.

Nisoli M, Priori E, Sansone G, *et al.* (2002) High-brightness high-order harmonic generation by truncated bessel beams in the sub-10-fs regime. *Physics Review Letters* 88: 033902.

Nisoli M, Sansone G, Stagira S, *et al.* (2003) Effects of carrier-envelope phase differences of few-optical-cycle light pulses in single-shot high-order-harmonic spectra. *Physics Review Letters* 91: 213905.

Paul PM, Toma ES, Breger P, *et al.* (2001) Observation of a train of attosecond pulses from high harmonic generation. *Science* 292: 1689.

Poletto L, Tondello G and Villoresi P (2001) High-order laser harmonics detection in the EUV and soft x-ray spectral regions. *Review Scientific Instruments* 72: 2868.

Salières P, L'Huillier P, Antoine A, *et al.* (1999) Study of spatial and temporal coherence of high order harmonics. In: Bederson B and Walther H (eds) *Advances in Atomic, Molecular, and Optical Physics*, vol. 41, pp. 83. New York: Academic Press.

Villoresi P, Ceccherini P, Poletto L and Tondello G (2000) Spectral features and modeling of high-order harmonics generated by sub-10-fs pulses. *Physics Review Letters* 85: 2494.

Villoresi P, Bonora S, Pascolini M, *et al.* (2004) Optimization of high-order harmonic generation by adaptive control of sub-10 fs pulse wavefront. *Optic Letters* 29: 207.

O

OPTICAL AMPLIFIERS

Contents

Basic Concepts
Erbrium Doped Fiber Amplifiers for Lightwave Systems
Optical Amplifiers in Long-Haul Transmission Systems
Raman, Brillouin and Parametric Amplifiers
Semiconductor Optical Amplifiers

Basic Concepts

M F S Ferreira, University of Aveiro, Aveiro, Portugal

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The transmission distance of a fiber-optic communication system is limited by fiber loss and dispersion. For long-haul lightwave systems, the loss limitations have traditionally been overcome by periodic regeneration of the optical signals at repeaters applying conversion to an intermediate electric signal. Because of the complexity and high cost of such regenerators, the need for optical amplifiers became obvious in the mid-1980s. The optical amplifier is ideally a transparent box that provides gain and is also insensitive to the bit rate, modulation format, power and wavelength of the signal passing through it.

Several means of obtaining optical amplification has been suggested since the 1970s, including direct use of the transmission fiber as gain medium through nonlinear effects, semiconductor amplifiers, or doping optical waveguides with an active material (rare-earth ions), that could provide gain. Due to the spectacular results on erbium-doped fiber amplifiers, which are particularly suitable in the third transmission window (around 1.5 μm), an intense worldwide research activity on optical amplifiers has developed. As a consequence, the development of erbium-doped fiber amplifiers has reached an industrial level, and commercial devices are now available.

Semiconductor amplifiers, on the other hand, have the same technical basis as semiconductor lasers. Although the strong nonlinearity of semiconductor amplifiers degrades the performances of transmission systems, the state-of-art semiconductor devices seem to be the most interesting amplifiers for transmission in the second transmission window (around 1.3 μm).

Amplifier Gain and Bandwidth

In a perfect amplifier, the amplification process would be insensitive to the bit rate, modulation format, power, state of polarization, wavelength, and optical bandwidth of the signal passing through it. On the other hand, no interaction would take place if more than one signal were amplified simultaneously. In practice, however, the optical gain depends not only on the wavelength (or frequency) of the incident signal, but also on the electromagnetic field intensity at any point inside the amplifier. Details of wavelength and intensity dependence of the optical signal depend on the amplifying medium.

We consider a case in which the gain medium is modeled as a homogeneously broadened two-level system. In such medium, the gain coefficient (i.e., the gain per unit length) can be written as:

$$g(\nu, P) = \frac{g_0}{1 + \dfrac{(\nu - \nu_0)^2}{\Delta\nu_0^2} + \dfrac{P}{P_{\text{sat}}}} \qquad [1]$$

where g_0 is the peak value of the gain coefficient determined by the pumping level of the amplifier, ν the optical frequency, ν_0 the atomic transition frequency, $\Delta\nu_0$ the 3 dB local gain bandwidth, P the optical power of the signal, and P_{sat} the saturation power, which depends on the gain medium parameters. It must be emphasized that $\Delta\nu_0$ and P_{sat} refer to the local gain. However, from the communication system point of view, it is more desirable to use the related concepts of amplifier bandwidth and amplifier saturation power that will be evaluated below.

The amplifier gain G is defined as

$$G = \frac{P_{\text{out}}}{P_{\text{in}}} \qquad [2]$$

where P_{in} is the input power and P_{out} the output power of a continuous wave (CW) signal being amplified. The amplifier gain G may be found by using the relation:

$$\frac{\mathrm{d}P}{\mathrm{d}z} = g(\nu, P)P \qquad [3]$$

where $P(z)$ is the optical power at a distance z from the amplifier input end.

If the signal power obeys the condition $P \ll P_{\text{sat}}$ throughout the amplifier, the gain coefficient given by eqn [1] can be considered independent of the signal power. In such a case, the amplifier is said to be operated in the unsaturated regime and works as a linear device. The gain coefficient presents in this situation a Lorentzian profile that is characteristic of homogeneously broadened two-level systems. However, the gain spectrum of actual amplifiers can deviate significantly from the Lorentzian profile.

The solution of eqn [3] in the unsaturated regime is an exponentially growing signal power, given by

$$P(z) = P_{\text{in}} \exp(gz) \qquad [4]$$

For an amplifier length L, we then find that the linear amplifier gain is

$$G(\nu) = \exp(gL) = \exp\left[\frac{g_0 L}{1 + (\nu - \nu_0)^2/\Delta\nu_0^2}\right] \qquad [5]$$

Both the amplifier gain $G(\nu)$ and the gain coefficient $g(\nu)$ are maximum when $\nu = \nu_0$. However, $G(\nu)$ decreases much faster than $g(\nu)$ with the signal detuning $\nu - \nu_0$, because of the exponential dependence of G on g. As a consequence, the amplifier bandwidth $\Delta\nu_{\text{A}}$, which is defined as the FWHM of $G(\nu)$, is much smaller than the gain bandwidth $\Delta\nu_0$ (Figure 1). This can result in signal distortion in the case where a broadband optical signal is transmitted through the amplifier. From eqn [5] we can obtain the following relation between the amplifier bandwidth and the gain bandwidth:

$$\Delta\nu_{\text{A}} = \Delta\nu_0 \sqrt{\frac{\ln 2}{g_0 L - \ln 2}} \qquad [6]$$

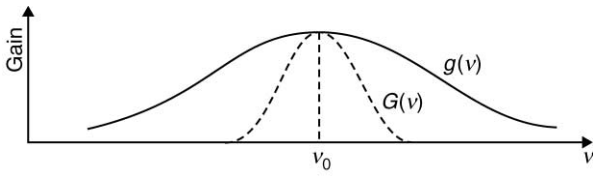


Figure 1 Gain coefficient profile $g(\nu)$ and the corresponding amplifier gain spectrum $G(\nu)$.

Gain Saturation

An important limitation of the nonideal amplifier is related with the power dependence of the gain coefficient given by eqn [1]. This property is known as gain saturation and it appears when the signal power ratio P/P_{sat} is non-negligible. Since the gain coefficient is reduced when the signal power P becomes comparable to the saturation power P_{sat}, the amplifier gain G will also decrease.

Assuming that $\nu = \nu_0$ and substituting g from eqn [1] in eqn [3] gives

$$\frac{\mathrm{d}P}{\mathrm{d}z} = \frac{g_0 P}{1 + P/P_{\text{sat}}} \qquad [7]$$

Considering the initial condition $P(0) = P_{\text{in}}$, we obtain from eqns [2] and [7] the following implicit relation for the amplifier gain:

$$(1 - G)\frac{P_{\text{in}}}{P_{\text{sat}}} = \ln\left(\frac{G}{G_0}\right) \qquad [8]$$

where $G_0 = \exp(g_0 L)$. The input saturation power P_{in}^{s} is defined as the input power for which the amplifier gain G is reduced by a factor of 2 from its unsaturated value G_0 (Figure 2). Indeed, it is obtained by using $G = G_0/2$ in eqn [8]:

$$P_{\text{in}}^{\text{s}} = \frac{2\ln(2) P_{\text{sat}}}{(G_0 - 2)} \qquad [9]$$

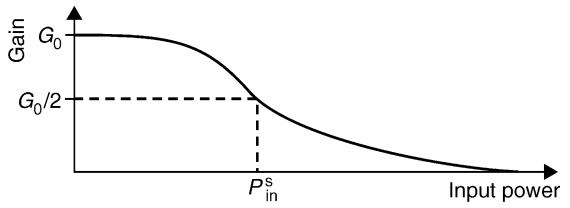


Figure 2 Saturated amplifier gain as a function of the input power.

As observed from eqn [9], the input saturation power P^s_{in} does not coincide with P_{sat}. The output saturation power is given by $P^s_{out} = G_0 P^s_{in}/2$. In practice, $G_0 \gg 2$ and P^s_{out} is found to be smaller than P_{sat}, by about 30%.

Gain saturation can be seen as a serious limitation, particularly for multichannel communication systems. However, the self-regulating effect of gain saturated amplifiers can be useful in long-haul lightwave communication systems, including many concatenated amplifiers. In fact, if the signal level in a chain of amplifiers is unexpectedly increased along the chain, the saturation effect causes a lower gain provided by the following amplifiers and vice versa for a sudden signal power decrease.

Amplifier Noise

Besides the bandwidth and gain saturation limitations, another property must be considered concerning practical optical amplifiers. In fact, optical amplifiers always add spontaneously emitted photons to the signal during the amplification process. Those photons are amplified besides the signal photons so that, at the amplifier output, an amplified spontaneous emission (ASE) noise is presented. Since spontaneous emission always takes place, ASE noise is unavoidable and does not depend on the amplifier temperature. This is one of the most important differences between optical and electrical amplifiers, where amplifier noise is of thermal origin and can be reduced by lowering the amplifier temperature.

The ASE determines a degradation of the signal-to-noise ratio (SNR). The SNR degradation is usually characterized by the amplifier noise figure, which is defined as the SNR ratio between input and output:

$$\mathrm{NF} = \frac{\mathrm{SNR}_{in}}{\mathrm{SNR}_{out}} \qquad [10]$$

The SNR is usually referred to the electrical power generated when the optical signal is converted to electrical current by using a photodetector. Therefore, the noise figure as defined in eqn [10] would usually depend on several detector parameters, which determine the shot noise and thermal noise associated with the practical detector. We will consider the case of an ideal detector, whose performance is limited by shot noise only.

The SNR of the input signal is simply determined by the detection shot noise and can be written as:

$$\mathrm{SNR}_{in} = \frac{P_{in}}{2h\nu\Delta f} \qquad [11]$$

where Δf is the detector bandwidth.

To evaluate the term SNR_{out}, one should add the contribution of spontaneous emission to the receiver noise. The ASE noise spectral density is assumed to be constant and can be written as

$$S_{sp}(\nu) = (G-1)n_{sp}h\nu \qquad [12]$$

where G the amplifier gain and

$$n_{sp} = \frac{N_1}{N_1 - N_0} \qquad [13]$$

is known as the spontaneous emission factor or the population inversion factor. In eqn [13] N_0 and N_1 are the atomic populations for the ground and excited states, respectively.

Considering a low noise amplifier, the signal power impinging the photodetector is larger than the optical noise power and the shot noise power. As a consequence, the electrical noise, due to the signal-ASE beating, is the dominant contribution and the SNR of the amplified signal is given by:

$$\mathrm{SNR}_{out} \approx \frac{GP_{in}}{4S_{sp}\Delta f} \qquad [14]$$

Using eqns [11]–[14], the amplifier noise figure defined by eqn [10] becomes:

$$\mathrm{NF} = 2n_{sp}\frac{G-1}{G} \approx 2n_{sp} \qquad [15]$$

where the approximation holds when the gain is much higher than one. In the case of an ideal amplifier, $n_{sp} = 1$ and eqn [15] show that the SNR is degraded by 3 dB. For most practical amplifiers, NF can be as large as 6–8 dB.

Basic Amplifier Configurations

The relative importance of the different limiting factors discussed above depends on the actual

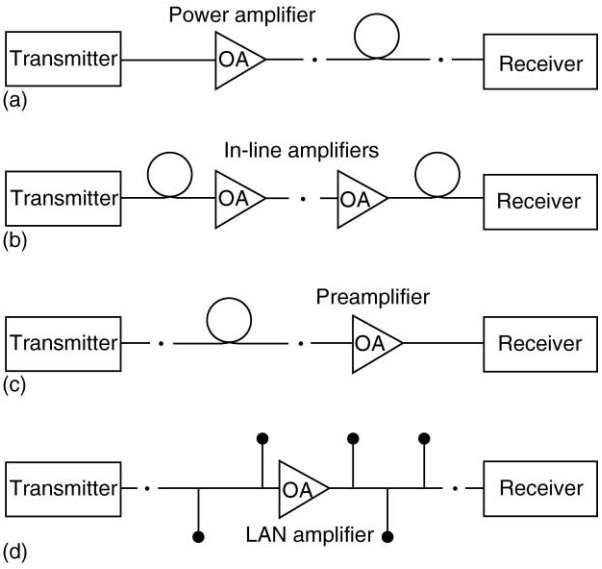


Figure 3 Four generic configurations for incorporating optical amplifiers into transmission systems: (a) as a power amplifier; (b) as in-line amplifiers; (c) as a preamplifier; (d) for compensation of distribution losses in local-area networks.

amplifier application. Figure 3 shows the four basic system configurations envisioned for the incorporation of optical amplifiers. The first configuration is to place the amplifier immediately following the laser transmitter to act as a power amplifier or booster (Figure 3a). The main purpose of such amplifiers is to boost the signal power, which can provide an increase of the transmission distance by 100 km or more. Since the signal input power is typically large (0.1–1.0 mW), the key parameter for the power amplifier will be to maximize the saturation output power and not necessarily the absolute gain.

The second configuration is to place the amplifier in-line and perhaps incorporated at one or more places along the transmission path (Figure 3b), replacing the electronic regenerators. The in-line amplifier corrects for periodic signal attenuation and may exist in a cascade form. The use of in-line optical amplifiers is particularly attractive for multichannel communication systems, since they can amplify all channels simultaneously.

The third configuration consists of using the amplifier immediately before the receiver, so it functions as a preamplifier (Figure 3c). The purpose of such an amplifier is to improve the receiver sensitivity. The main figures of merit are high gain and low amplifier noise, because the entire amplifier output is immediately detected.

In local-area networks (LANs), distribution losses often limit the number of possible nodes. The fourth application of optical amplifiers consists of using them for compensating such distribution losses (Figure 3d).

List of Units and Nomenclature

g_0	peak value of the gain coefficient at peak	m^{-1}
$g(\nu)$	gain coefficient	m^{-1}
G_0	unsaturated amplifier gain	
$G(\nu)$	amplifier gain	
L	amplifier length	m
n_{sp}	spontaneous emission factor	
P	signal power	W
P_{in}	input signal power	W
P_{in}^{s}	input saturation power	W
P_{out}	output signal power	W
P_{out}^{s}	output saturation power	W
P_{sat}	saturation power	W
$S_{sp}(\nu)$	ASE noise spectral density	J
Δf	detector bandwidth	Hz
$\Delta\nu_0$	bandwidth of the gain coefficient	Hz
ν	optical frequency	Hz
ν_0	atomic transition frequency	Hz

See also

Optical Amplifiers: Optical Amplifiers in Long-Haul Transmission Systems. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems.

Further Reading

Agrawal GP (1992) *Fiber-Optic Communication Systems*. New York: Wiley.

Bjarklev A (1993) *Optical Fiber Amplifiers: Design and System Applications*. Boston, MA: Artech House.

Desurvire E (1994) *Erbium-Doped Fiber Amplifiers: Principles and Applications*. New York: Wiley.

Green PE Jr (1993) *Fiber Optic Networks*. Englewood Cliffs, NJ: Prentice Hall.

Iannone E, Matera F, Mecozzi A and Settembre M (1998) *Nonlinear Optical Communication Networks*. New York: Wiley.

Kazovsky L, Benedetto S and Willner A (1996) *Optical Fiber Communication Systems*. Boston, MA: Artech House.

Shimada S and Ishio H (eds) (1994) *Optical Amplifiers and their Applications*. New York: Wiley.

Erbrium Doped Fiber Amplifiers for Lightwave Systems

P Bollond, JDS Uniphase Corporation, Ewing, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The telecommunications industry has undergone a revolution since the 1980s, by using glass optical fibers for the transmission of information encoded as pulses of light. A single telecommunications-grade optical fiber has been shown to support the propagation of more than 1 Tbit per second (1×10^{12} pulses per second) over distances comparable to typical city separations (>100 km). From this technology, optical fiber links have evolved to become a network on a planet-wide scale, and form the physical backbone of the information age.

Erbium-doped fiber amplifiers (EDFAs) are an enabling technology for optical fiber communication networks. They have several important properties that make them the amplification component of choice in long-distance commercial data transport networks. Erbium ions deposited in silica-based glass allow amplification in the lowest loss region of commercial-grade optical fiber (<0.25 dB/km from approximately 1530–1620 nm). Erbium-doped fiber (EDF) is manufactured in a form that allows for low-loss fusion splicing to standard communications fiber. Compact semiconductor laser diodes are available to excite the erbium ions into an amplification state. The long lifetime of the excited state of erbium provides the ability to simultaneously amplify multiple wavelength channels without significant cross-channel interference. Multiple channel systems have been deployed with more than 100 optical channels (or wavelengths) through each EDFA, and this has allowed network capacities to be dramatically increased. The low noise properties of the EDFA also allow networks to be constructed with many amplified spans before the optical signal has to be electronically regenerated. Practically unlimited transmission distance has been demonstrated using a small number of optical soliton channels through periodically amplified EDFA lightwave systems.

The development history of the EDFA can be traced back to the first optical amplifier. In 1962, a neodymium-based fiber amplifier was invented that operated at 1064 nm. During the 1980s, the need for an optical amplifier at telecommunications wavelengths initiated research at many locations throughout the world. In 1987, the University of Southampton (UK) was first to demonstrate an EDFA that had optical gain at 1550 nm, and during the following years the design of erbium-doped fiber was optimized for this application. In 1989, a practical semiconductor laser diode became available to pump EDF, and the first compact optical fiber amplifier modules soon appeared for commercial deployment. The traditional method of signal regeneration, prior to 1990, was to use electronics to detect the optical signal after each transmission span, recover the digital signal, and then retransmit using another laser diode. The EDFA allows practical wavelength division multiplexing (WDM) of multiple optical signals with all optical signal amplification, and provides significant performance and cost advantages over electronic regeneration.

EDFAs have emerged from the laboratory to be widely deployed in communication networks. EDFAs are used to boost transmitted power of the signal lasers (booster amplifier), amplify signals in transit to compensate losses sustained in the fiber (line amplifier), or amplify signals before a receiver (pre-amplifier). Typically, the amplifier module is specifically manufactured for particular systems that are mounted on electronic circuit boards. These circuit packs are then housed in central offices (local telephone exchanges), remote 'repeater huts', or even in undersea 'bottles' as part of a transoceanic cable. The high cost of network failure requires that the manufactured EDFA modules comply with stringent reliability criteria, to provide a useful operating lifetime greater than 25 years when subject to extreme environmental conditions.

Amplifiers are constructed for incorporation into either existing fiber links as part of an upgrade, or for newly planned systems. Because of the high cost of installing new fiber into the ground and securing property rights, it has become economically desirable to upgrade many existing fiber links rather than to build new systems. Transmission cables usually have many pairs of individual optical fibers, some of which will not initially be transporting data, and these 'unlit' or 'dark fibers' can be activated as consumer demand increases over time. Typically, each fiber of a pair is used to carry either 'east'- or 'west'-bound traffic. Around city areas metropolitan area networks can be expanded in this way, but for long-distance links (long haul networks with distances >1000 km) operation is designed for a larger number of channels (40 to 120 wavelengths) at higher data rates (10 or 40 GHz per channel) over specialized transmission

cables containing low numbers of fiber pairs. Typically communication systems are designed to meet certain cost targets, expressed as dollars per Gbit/s per km, for total transmission distances. The total transmission distance is limited by the optical signal to noise ratio (OSNR) degradation after each fiber-amplifier link, with a smaller permissible degradation at higher bit rates. The required gain and OSNR performance for the system is then translated to an EDFA module specification.

This article discusses EDFA design and applications, and shows the elements involved in producing reliable modules for commercial lightwave systems.

Components for EDFAs

EDFAs are comprised of passive optical components, erbium-doped fiber, and pump lasers. Passive components are chosen to meet optical and environmental specifications, while the erbium-doped fiber is selected based on the optical power, gain, and noise figure requirements. Pump lasers are a key influence on the price and performance of optical amplifiers.

The amplifier module is typically connected to the transmission fiber using fiber connectors. These polished fiber connectors have a higher insertion loss and reflectivity than fusion splicing, but allow for easy deployment in the network. Internal optical components are fusion spliced together to provide low loss, low back reflection, high strength, and high reliability joins. Fusion splicing is tailored to particular fiber types, so that optical components with dissimilar fiber types are joined with the lowest loss.

The signal and pump radiation is combined with low loss using optical components called wavelength division multiplexers (see **Figure 1**). These components are based on fused fiber or interference filter based technology. Fused fiber WDMs offer the lowest insertion loss (<0.1 dB is commercially available) but is restricted to widely spaced wavelengths (e.g., 980 nm pump and 1550 nm signal). Interference filter based WDMs are available for closely spaced wavelengths (e.g., 1480 nm pump and 1530 nm signal) and have a very low wavelength dependent insertion loss (flatness). Interference filters can be designed to produce sharp low-pass, high-pass, or bandpass type filters suitable for combining closely spaced wavelengths, as well as broader peaks suitable for lowering the gain in particular signal wavelength regions to produce gain flattened amplifiers.

Signal reflections can cause an amplifier to act as a laser, and this detrimental effect is eliminated in EDFAs by using optical isolators. In an isolator the signal light is coupled out of the single-mode fiber through a graded index (GRIN) lens and passes through a nonlinear crystal before being coupled back into the optical fiber. The isolator consists of a birefringent rutile (TiO_2) or Yttrium OrthoVanadate (YVO_4) wedge, followed by a Yttrium Iron Garnet (YIG) Faraday rotator, followed by another birefringent wedge. The YIG crystal is surrounded by a permanent magnet that rotates the light's polarization by 45 degrees. The 45-degree polarization rotation, coupled with the two birefringent wedges, ensures that light is efficiently coupled to the output fiber but not in the reverse direction. Commercial isolators are available with low insertion losses across the signal band, with some samples below 0.35 dB. Note that the YIG crystal works well for the 1480 nm pump and 1530–1620 nm signal bands, but currently there is no suitable isolator material that covers 980 to 1550 nm and this puts some limitations on certain EDFA designs.

The isolator design has been extended to make multiport circulators. A three-port circulator has the input into port 1 and output of port 2, light entering port 2 is directed to port 3, and light entering port 3 is blocked with an isolator. The circulator allows for adding and dropping of individual channels when a

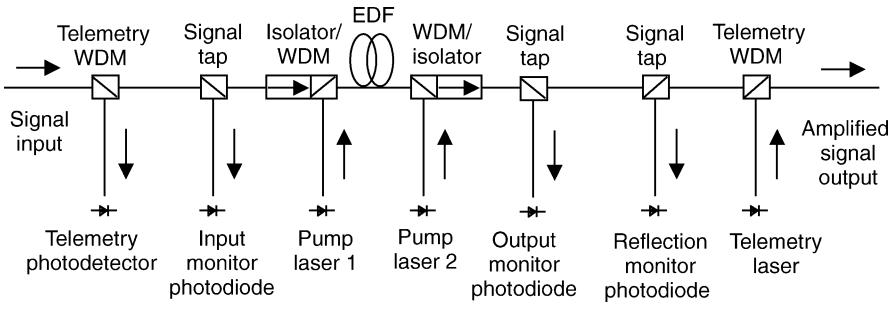


Figure 1 Single-stage EDFA with features.

narrow bandwidth reflective filter is placed on port 2. Circulators also can be used to separate co-propagating and counter-propagating traffic on a single transmission fiber before amplification is done.

Erbium-doped fiber has two strong absorption bands around 980 nm and 1480 nm that are suitable for achieving a population inversion in the erbium ion. The 980 nm wavelength allows low noise amplification, while 1480 nm lasers provide higher EDFA output. 980 nm pump lasers usually incorporate fiber Bragg gratings to stabilize the laser wavelength to match the narrow EDF absorption peak, and also have lower drive current requirements. A 1480 nm pump can provide more amplification, since there are more photons in each mW of laser power at 1480 nm than 980 nm. High-output power pump lasers incorporate thermo-electric coolers (TECs) within the pump laser package and can provide fiber coupled power greater than 500 mW. Both high-power lasers have been qualified to meet the most stringent reliability standards, with typical mean time before failure beyond 25 years. Low-cost 980 nm pump lasers, that do not have TECs, are also available in smaller packages and can supply up to $\sim$200 mW. By using wavelength division or polarization combining techniques, it is possible to further increase the available pump power in the erbium fiber.

The EDFA module is assembled into a package that may contain passive optical components, several meters of coiled EDF, pump lasers, photo-detectors, and electronic circuit boards. In some cases, it is advantageous to thermally stabilize the components so that the amplifier performance can be maintained when exposed to extreme environmental conditions. This may occur when the central office's environmental control is compromised (e.g., air conditioning failure). In particular, EDF can exhibit undesirable spectral gain changes if the ambient temperature changes by more than 5 °C.

Pump lasers with internal TECs can dissipate more than 5 Watts of heat per laser and since cooling fans usually do not have the required reliability, passive cooling is commonly used in the module in the form of a metal heat sink. The amplifier module's size can be compact, limited by the height of optical components or pump laser diodes, or by the size of a built-in heat sink. The module is designed to survive conditions of electrostatic discharge, humidity, temperature, thermal shock, extreme vibration, and other stresses that may be inadvertently present during operation in the field. In addition, all material in the EDFA module is also qualified against problems with out-gassing (e.g., hydrogen release), combustion and chemical or biological exposure.

The Single-Stage Amplifier

A simple single-stage EDFA consists of an erbium-doped fiber spool with signal and pump combining multiplexer. The fiber is optically pumped by 980 nm and/or 1480 nm laser(s), whose light is coupled into the signal fiber by a passive component called an optical multiplexer. The pump wavelengths are readily absorbed by erbium ions embedded in silica raising them to an excited state. Amplification occurs when, stimulated by a nearby signal photon, an excited erbium ion relaxes back to the ground state producing a second, identical signal photon. The erbium ion can be approximated as a three-level atomic system that can be completely inverted by a 980 nm photon, to provide the lowest noise amplification. In contrast, the 1480 nm pump will excite the erbium ion directly into the upper laser level as a two-level system, and because of rapid spontaneous emission from this level, the maximum inversion in this case cannot exceed approximately 75%. Note that as the pump photons are absorbed, the inversion will be nonuniform along the EDF length.

There are two signal wavelength regions commonly amplified by EDFAs, the C-band (conventional band) from approximately 1528 to 1565 nm, and the L-band (long band) from approximately 1570 to 1620 nm. Amplification in the C-band readily occurs when moderate pump power is available, and relies on the erbium ion's spectral absorption and emission wavelength window that is suited to high levels of atomic inversion. L-band amplification is also achieved with moderate pump powers, but because of the lower absorption and emission cross-sections, similar gain is reached using approximately five times more EDF with a lower average inversion. The C-band amplifier is typically less costly because less EDF is used, while high-concentration erbium fibers are available specifically for L-band EDFAs.

An EDFA's most critical performance parameters are its amplified signal output power (typically stated in dBm) and its noise figure (stated in dB). Output power is mainly determined by total pump power and the amplifier internal loss. The noise figure (NF) is defined as the ratio of the signal-to-noise ratio at the input to the signal-to-noise ratio at the output.

A single-stage amplifier typically has 1 or 2 pump lasers but can have more if polarization- or wavelength-pump-combining is implemented for higher power. When the pump radiation propagates in the same direction as the signal, the amplifier is co-pumped, while counter-pumping denotes the case when the pump laser propagates against the signal. For a single pump, a co-pumping 980 nm

laser minimizes the NF (suitable for a pre-amplifier) while counter-pumped 1480 nm architecture optimizes output power at some expense to the NF (suitable for a booster amplifier). Recent semiconductor pump laser improvements have enabled 980 nm pump lasers with output powers >500 mW to be commercially available, allowing most single- or dual-stage EDFAs to be energized by one laser.

Single-stage designs can be enhanced, as shown in Figure 1. To control the network an optical signal may be used as a telemetry or network supervisory channel, and this is removed with a filter. Telemetry wavelengths are usually outside of the useful EDF amplification window, and commonly range from 1500–1520 nm and from 1620–1640 nm. An isolator may be used at the input and/or output to prevent pump laser or amplified spontaneous emission from the erbium-doped fiber 'leaking' into the transmission path. Optical taps may be included to provide information about signal spectra at the input and output sides. Their feedback can be used to control pump laser biases for tuning output power, monitoring amplifier performance, or simply to trigger alarms. In addition, a reflection monitor is sometimes placed at the output to observe backwards propagating optical signals arising from reflections. Electronics in the module will continuously monitor the pump lasers and photodetectors and, for example, can place the module in an 'eye-safe' low-output power (<10 mW) mode within milliseconds if a transmission fiber break is detected through increased back-reflected optical power.

The amplifier has a signal input power of −30 to −10 dBm (1–100 μW) for each optical channel, and has a gain of 25 dB to compensate for a typical span loss of 100 km optical fiber link. This signal level allows high powers at the receiver photodetector for high-quality signal detection, yet is low enough to avoid nonlinear propagation effects in the transmission fiber. The typical total output signal power of an 80 channel (wavelength) C-band EDFA is less than 200 mW (+23 dBm) to avoid stimulated Raman scattering (SRS) in standard transmission fiber. The use of improved low nonlinearity transmission fibers and longer transmission distances can lead to specified total EDFA output powers to be greater than 400 mW (+26 dBm). To provide the best performance, the amplifier usually will allow only a single direction of propagation, with bi-directional communications systems using one transmission fiber and circulators to separate 'east' and 'west' traffic into individual EDFAs.

Figure 2 shows the results of a numerical simulation for a single-stage amplifier as a function of the input power and EDF length. The amplifier was assumed to have EDF with peak absorption of 5.5 dB/m near 1530 nm, and was pumped with 100 mW at 980 nm. The amplifiers signal loss before and after the EDF was taken to be 0.8 and 1.2 dB, respectively. As amplifier input power increases there is a decrease in gain provided by the medium since there is a fixed amount of pump power available. The maximum gain is usually achieved near 1530 nm where the difference between the EDF's emission and absorption cross-sections is greatest. Selecting the length of EDF is critical to achieving the desired performance, and this can be examined using numerical simulation for a wide range of design options.

Undersea systems, with their long distances and large costs of network failure, place stringent design requirements on EDFAs. These systems mostly use single-stage designs with emphasis on low noise operation. This can be achieved by eliminating most of the optical components prior to the EDF, and also by co-propagating a strong 980 nm pump using a low loss fused fiber WDM. Undersea repeaters are spaced by 30 to 80 km, shorter than terrestrial networks, with each channel operated at higher power to achieve multi-thousand kilometer distances. For example, the trans-Pacific TPC-5J cable spans 8600 km from Coos Bay, Oregon (USA) to Ninomiya (Japan), using EDFAs spaced every 33 km. The tight electrical power budgets available to each repeater (powered from land) necessitate using pump lasers without TECs. During installation, the EDFA will experience large mechanical shock, as the cable and metal repeater bottles are unwound aboard ship and dropped into the ocean. Other design considerations are done for operation at the constant ambient temperature of the ocean bottom (~+2 to 4 °C) or for seasonal temperature changes on the continental shelves, where the optics and EDF will operate at 15 to 30 degrees above ambient due to heating from the electronics.

Single-stage optical amplifiers are suited for a wide range of applications such as single-channel amplifiers, simple WDM amplifiers, and low-cost amplifiers. Using high-power pump lasers or combined pump laser schemes allows such amplifiers to deliver output powers >20 dBm. However, single-stage amplifiers cannot meet the requirements of all telecommunications architectures, leading to increasing demand for multiple-stage EDFAs that are discussed below.

Multiple-Stage EDFAs

The most common implementation of a multiple-stage EDFA is to improve the noise and output power

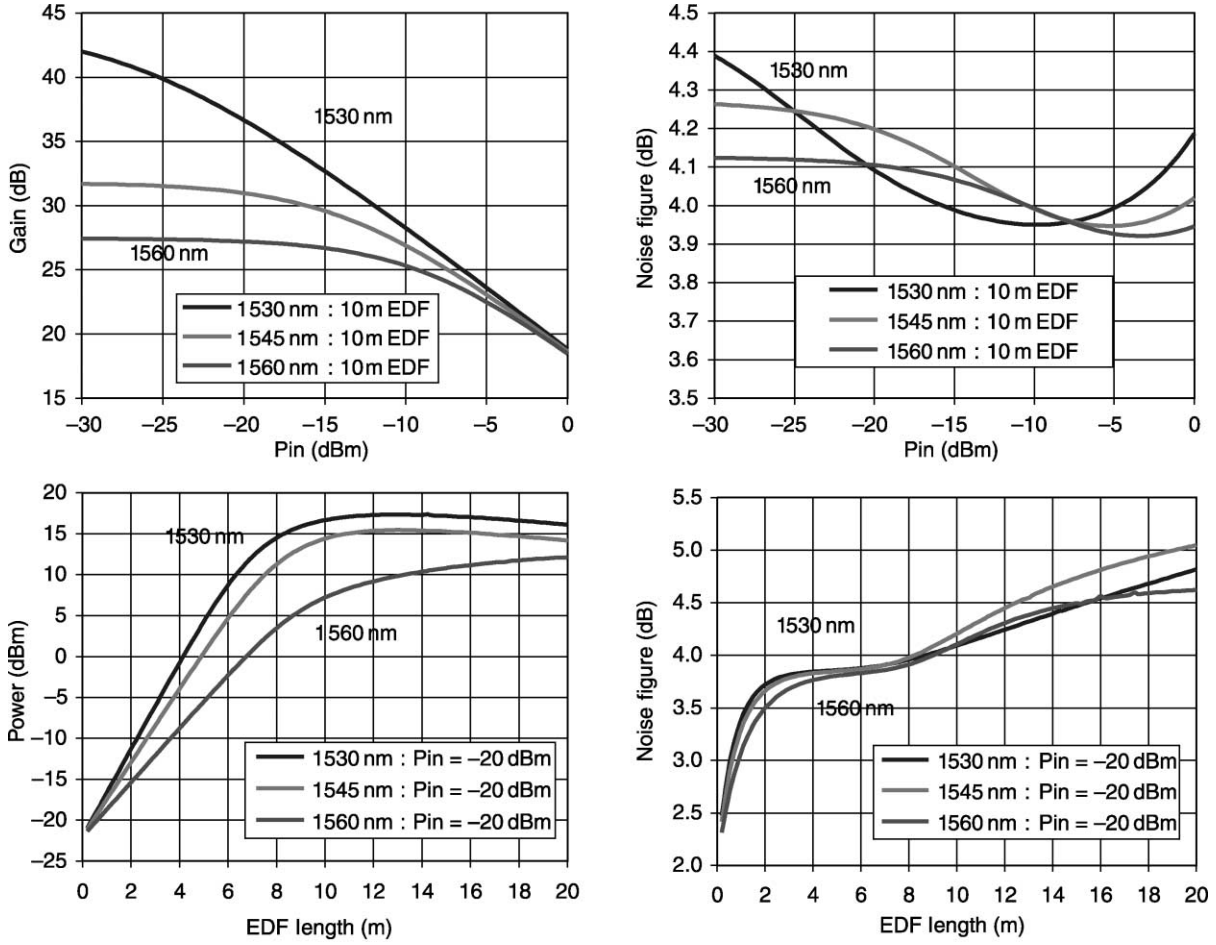


Figure 2 Typical single-channel performance of a single-stage EDFA.

characteristics by imbedding an optical isolator between two sections of EDF. The isolator blocks backward-traveling amplified spontaneous emission to improve the efficiency of the amplifier. This is shown in Figure 3 for two common single-pump designs. The first is the 'pump by-pass', design where the residual pump radiation from the first stage is redirected around the mid-stage isolator. This is often used when pumping with 980 nm since signal band optical isolators will not transmit 980 nm radiation. This design can introduce problems when using low-quality components as small signal levels can propagate around the pump by-pass fiber to create multiple path interference (MPI) effects. Although MPI effects can be eliminated by using a pump splitting coupler to directly pump each EDF section, the pump by-pass design makes the most efficient use of available pump power. The second design is called the 'pump through' design, and is used when pumping at 1480 nm since both the pump and signal band (e.g., 1550 nm) photons will transmit through commercial isolators with only small loss. Note that for cost and space constraints within the module, it is sometimes advantageous to use hybrid optical components, e.g., an isolator and WDM can be combined into one compact package.

An EDFA with multiple input signals will have a very nonuniform output gain profile. This is a consequence of the erbium ion's wavelength-dependent emission and absorption cross-sections in the host glass material. Figure 4 shows gain spectra for both C-band and L-band EDFAs with and without gain-flattening filters (GFFs). The gain spectrum is dependent on the EDF's chemical composition and EDFA design features. For an amplifier with 25 dB gain, a GFF with peak loss less than 10 dB is usually needed to correct for these gain deviations. Although many technologies are available for GFFs (e.g., thin-filter interference filters, Bragg gratings, tapered fiber filters, etc.) the basis function of these technologies is usually not identical to the EDF's gain profile, and this mismatch results in gain

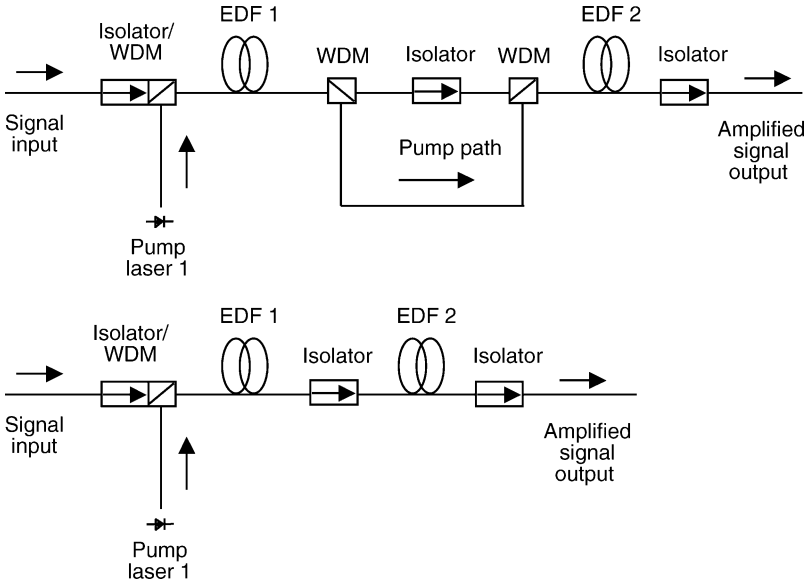


Figure 3 Imbedded Isolator EDFA designs.

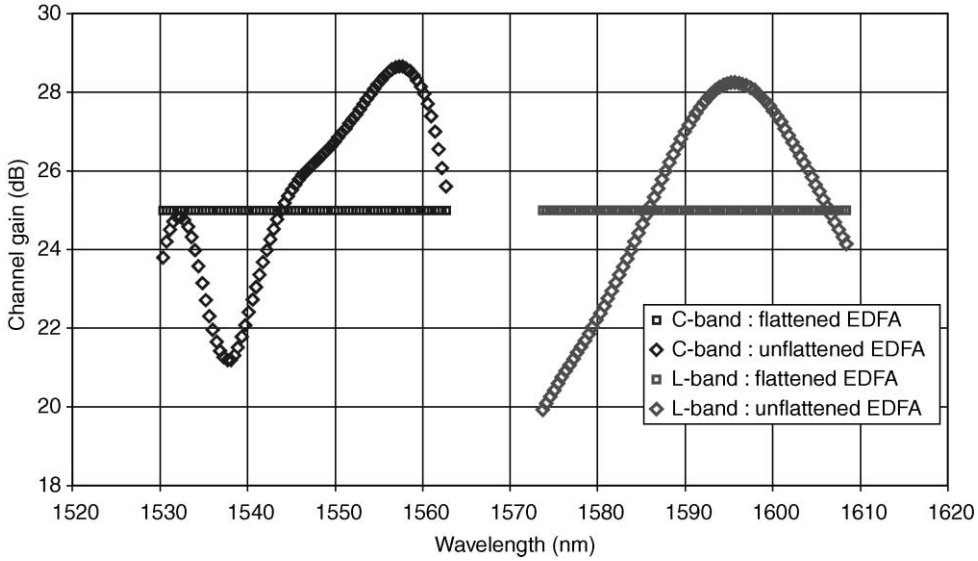


Figure 4 Gain spectrum of a multi-staged optical amplifier with and without gain flattening filters.

flatness error. However, when using these technologies, manufactured EDFAs can have a gain error (flatness) less than 1.0 dB over bandwidths greater than 35 nm.

Early EDFAs for WDM systems used 16 optical channels spaced by 200 GHz in the 1540 to 1560 nm part of the C-band. In this case, the EDF gain profile is relatively smooth and no GFFs are needed to achieve a 1.0 dB gain flatness specification. It is important to note that for a particular level of signal and pump power, a longer length of EDF in the EDFA will produce more gain at the long wavelength end of the spectrum, hence by shortening the EDF length in Figure 4 the result will be relatively flat gain from approximately 1540 to 1560 nm. As EDFA technology has matured, the EDF can be the bandwidth-limiting component in a system, and it has become necessary to use GFFs to realize useable bandwidths of up to 55 nm when using regular silica-based EDF. The most economic EDFAs operate in the C-band where relative to the L-band, pump laser efficiency is highest, EDF lengths are shortest,

and the chromatic dispersion of most installed transmission fiber limits nonlinear optical effects.

In single-stage amplifiers, the gain-flattening filter is at the amplifier output, and this results in lost output power. Wideband gain-flattened amplifiers usually have a multiple EDF stage design, with the GFF component between two EDF stages. The EDF sections that follow the attenuating GFF provide additional amplification, to give high power out of the EDFA module that has large internal losses.

In addition to GFF components inside the EDFA, an attenuator can be used to compensate for input signal power changes, which will produce a spectral gain rotation or tilt. The attenuator reduces the power on all optical channels equally and allows the amplifier to operate in a 'fixed gain' mode. In WDM systems, optical amplifiers may also need mid-stage access where the signal is fed into an external device between the amplifier stages. Reasons for doing this include monitoring, adding or dropping of individual channels, and dispersion compensation. Since additional losses up to 10 dB are introduced in the amplification path, the amplifier design has to be optimized for those losses.

Typical two-stage gain-flattened optical amplifier architecture is shown in Figure 5. Input and output couplers and telemetry WDMs can be included if required. A single pump can be split and shared between stages to save cost. When multiple pumps are needed, the most common configuration is a 980 nm pump laser co-pumping the first stage (EDF1) for low noise and a 1480 nm laser counter-pumping the second spool (EDF2) for high gain. A significant loss element, e.g., a gain-flattening filter, add/drop module or dispersion compensation module, is situated between the stages. Note that mid-stage access can be located before or after the gain flattening or between additional EDF stages. The gain spectrum of multistage gain-flattened EDFAs is shown in Figure 6, showing the gain equalization possible using a multistage amplifier design and gain-flattening filters. Using deeper GFFs can increase the usable bandwidth of the amplifiers, but this requires higher-power pump lasers to maintain the same gain and optical signal-to-noise ratio (OSNR) performance.

The two-stage EDFA shown in Figure 5 highlights a common problem for amplifier design. Given high-grade optical components and pump lasers, what

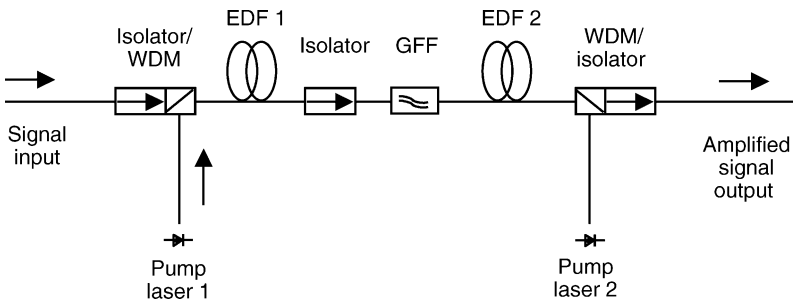


Figure 5 Typical multistage gain-flattened optical amplifier architecture.

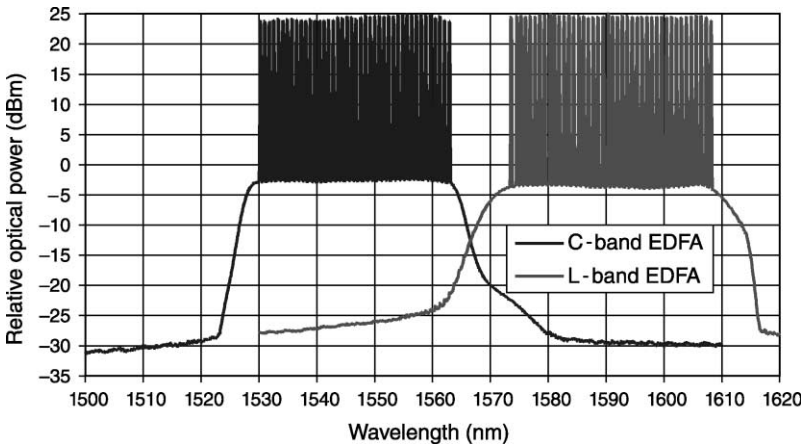


Figure 6 Measured gain and noise spectra of multistaged gain-flattened C-band and L-band EDFAs.

length of erbium-doped fiber should be used to give the best amplifier performance? This question is best answered by using the results of extensive numerical simulations of the optical amplifier.

Figure 7 shows the results of numerical simulation for a two-stage amplifier as a function of the first and second EDF stage lengths. The amplifier was assumed to have EDF with a peak absorption of 5.5 dB/m near 1530 nm, and was pumped with 130 mW at 980 nm in the first stage and 160 mW at 1480 nm in the second stage. The total input power was assumed to be −2.5 dBm for 80 channels distributed from 1530 nm to 1563 nm. Note that each of the 80 channels is separated by 100 GHz, and will generally support a long-distance communications system with each wavelength modulated at 10 GHz.

Figure 7a shows that the optical output power is largest when both EDF stages are approximately 20 meters. For this design, EDF lengths significantly longer than 20 meters will produce less output power, since the pump radiation will have been completely absorbed in the first few meters of EDF. The spectral gain flatness is shown in Figure 7b. For this design, approximately 17 meters of EDF was needed to produce a flat gain spectrum, i.e., each optical channel had the same gain. Other combinations of EDF lengths were not matched to the particular gain-flattening filter and produced large gain variations across the band. In general as the amplifier's total EDF length increases the gain spectrum exhibits a positive tilt (e.g., longer wavelength channels will experience more gain). Figures 7c, d show the average channel and maximum channel NF, respectively. From a system design perspective, a link can be limited by the optical channel with the lowest optical signal-to-noise ratio (OSNR) or highest noise figure, and this is a critical parameter of interest when designing optical amplifiers. In practice the best amplifier design is a compromise between high-gain, low-gain flatness, and low NF. From extensive numerical simulations, for the two-stage gain flattened amplifier example it was for

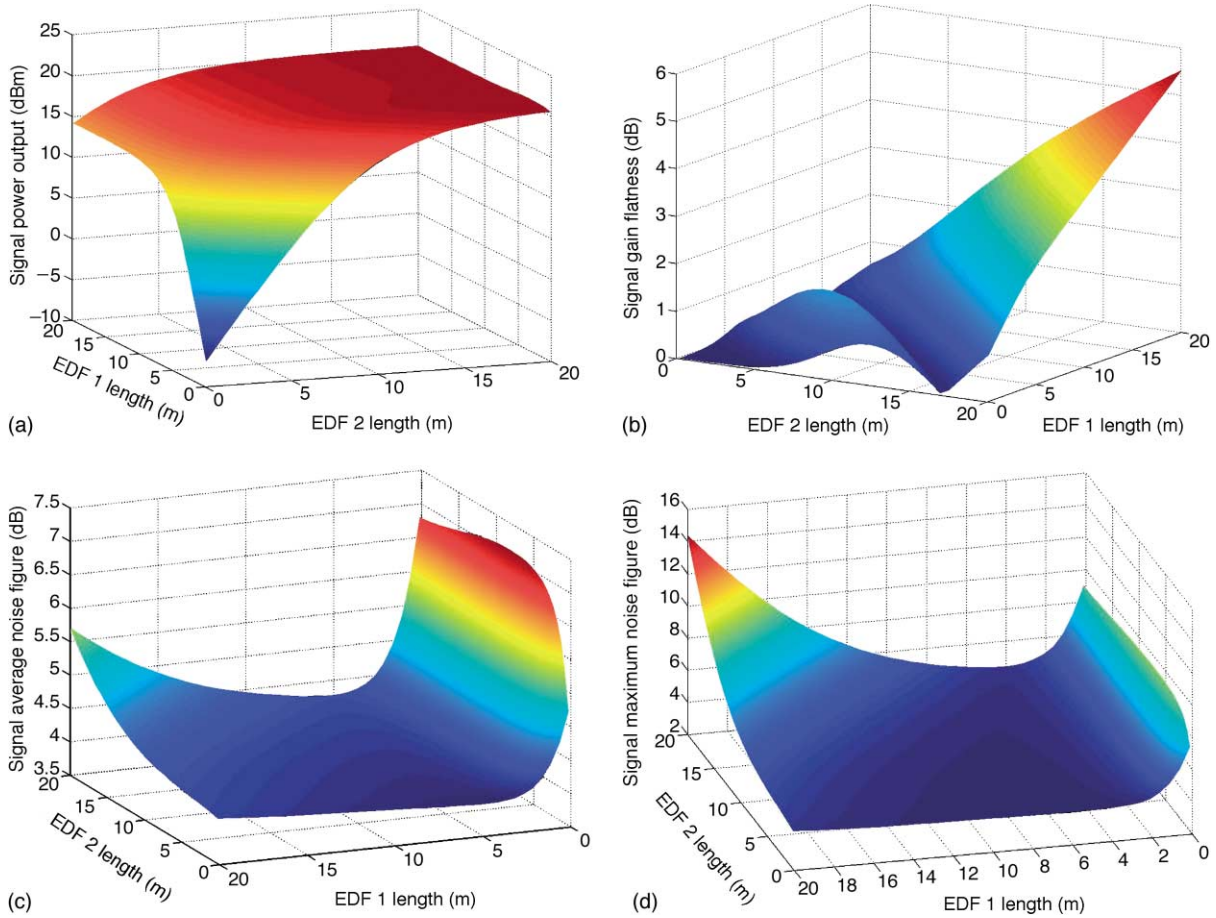


Figure 7 (a) Total output power, (b) spectral flatness, (c) average noise figure, and (d) maximum channel noise figure of a two-stage optical amplifier with a gain-flattening filter.

approximately 4 meters in EDF stage 1, and 13 meters in stage 2.

Successful commercial EDFA products are more than a single amplifier design that can accommodate all applications. In some cases an economical solution is a modular design approach, where smaller capacity, lower-cost amplifiers are initially installed in a network. Additional gain can be added to the basic amplifier for longer-distance spans or as network traffic increases over time. This 'pay as you grow' approach can be done using additional gain stages, or pump lasers, and can be upgraded without interruption of revenue-generating network traffic.

Low-Noise Design of EDFAs

A key design feature of commercial EDFAs is low-noise operation since the degradation of the OSNR limits the reach of the system. The NF of an EDFA can be defined as

$$\mathrm{NF} = \mathrm{SNR_{in}}/\mathrm{SNR_{out}} \qquad [1]$$

The NF can also be calculated from

$$\mathrm{NF} = \frac{1}{G} \times \left[\frac{P_{\mathrm{ASE}}(\nu_s, L)}{h\nu_s} + 1 \right] \qquad [2]$$

where P_{ASE} is the amplified spontaneous emission (ASE) noise power at the signal frequency ν_s for an EDFA of gain G (h = Plancks constant). An alternative equation for the NF for a single section of EDF is

$$\mathrm{NF} = \int_0^L \frac{\gamma_e(\nu_s, z) - \gamma_a(\nu_s, z)}{P_s(z)} \times P_s(0)\mathrm{d}z + 1 \qquad [3]$$

where γ_e and γ_a are the EDF emission and absorption factors at the signal frequency ν_s, at distance z along the amplifying fiber. Equation [3] shows that low NF can result from rapid signal gain in the initial fiber along the EDF. This result indicates that co-propagating pump and signal photons, both into the same end of the EDF, combined with the maximum inversion obtained by using a 980 nm pump, will result in low-noise amplification. The NF is typically expressed in dB units, and using a full quantum mechanical theory, the lowest possible NF for a fully inverted EDF with a large signal gain can be shown to be approximately 3 dB. Actual EDFAs will only approach the '3 dB quantum noise limit' when the signal is significantly lower than the available pump power (e.g., low channel count applications) so that high inversion can be maintained along the EDF length. Optical components in the EDFA prior to the first EDF will attenuate the signal but not the noise, since ASE is only generated in the EDF, and this input stage loss will directly add to the EDFA's NF.

The use of 1480 nm pump lasers can also result in a further NF penalty of 0.3 to 1.5 dB since complete inversion cannot be achieved when the erbium ions are operating as a two-level atomic system and more ASE is generated. The ASE, like the signal gain, has a wavelength-dependent spectral profile and this results in each channel having a different NF. Typically, the optical channel with the largest NF (lowest OSNR) is used to define the EDFA's noise performance. EDFAs with low-input signal power (e.g., −15 dBm) and high gain (e.g., 25 dB) may be described as low noise when the NF is <4 dB. For large total input signal power (e.g., +2.5 dBm for high channel count operation) low noise may be NF < 5 dB.

For an amplifier with multiple EDF sections, the noise figure can be calculated from

$$\mathrm{NF_{System}} = \frac{\mathrm{NF}_1}{L_1} + \frac{\mathrm{NF}_2}{L_1 G_1 L_2} + \ldots + \frac{\mathrm{NF}_N}{L_1 G_1 L_2 G_2 \ldots L_{N-1} G_{N-1}} \qquad [4]$$

where L_i is the signal loss prior to the gain G_i of the EDF stage i. From eqn [4], multistage amplifiers with large inter-stage losses will have numerically larger NFs, but the total module NF is dominated by the first EDF stage. Undersea EDFAs, with a single short EDF section, very small L_i and 980 nm pumping, can have NF <3.5 dB.

Advanced EDFA Functions

An important trend in optical amplifiers, and systems in general, is the move towards dynamic control. Several technologies are being considered and the primary idea is to dynamically control the intensity level of each optical channel or wavelength to correct for any gain/loss inequalities in the network. This is particularly important as channel counts and bit rates rise concurrently with the functionality demanded of optical communication networks. A common requirement in this category is fast amplifier response so that in a network where individual optical channels can be added or dropped there are no power distortions at other surviving wavelengths. In a typical drop event, in less than 1 microsecond, 31 of 32 wavelengths can be completely removed from the input to the EDFA by an add/drop module located elsewhere in the network. This event will cause a total input power decrease of 15 dB and without rapid pump power adjustment the remaining channel will be excessively amplified.

Erbium ions have an excited state lifetime of approximately 10 ms, and this sets the time-scale for the transient signal response. A network transient usually manifests as a sharp optical amplifier (OA) gain fluctuation lasting up to several tens of microseconds that will pulse large signal levels through the entire optical network. Dynamic network loading necessitates controlling temporal gain transients, and this may be achieved by using a microprocessor in the amplifier module. These intelligent EDFA modules can rapidly adjust the pump laser power based on feedback from internal photo detectors to control gain transients.

The need for dynamic control also exists in the frequency domain. Wideband amplifiers have a gain variation among the optical channels of typically up to 1 dB over all operating conditions, and since these EDFAs are concatenated over many spans the electronic receiver circuit at the end of the network will have to detect optical channels with widely varying power levels. This can limit the number of spans that the network can bridge before electronic regeneration becomes necessary. This problem can be corrected using a dynamic gain equalizing filter (DGEF) to actively attenuate individual optical channels levels. A DGEF can be embedded inside an amplifier to create a loss-less component, and can be used to equalize channel powers for an entire network.

The problems of accumulation of gain ripple become more apparent in systems that employ amplifiers based on stimulated Raman scattering. A conventional EDFA amplified system may have 10 spans transmitting 100 optical channels at 10 Gbit/s each over a distance of approximately 800 km before electronic regeneration becomes necessary. To span longer distances with higher capacity, Raman amplifiers are used to amplify the transmission fiber along with EDFAs. This system architecture has the advantage of a higher effective OSNR for the individual spans, and has been demonstrated for distances over 4000 km with 40 Gbit/s channels. However, the variability of the Raman amplifier gain, due to network loading and differing transmission fiber properties, necessitates the use of loss-less DGEFs.

High-capacity transmission uses individual channels that are each modulated up to typical 2.5, 10, or 40 GHz, that requires that the optical amplifiers have low polarization mode dispersion (PMD) and low chromatic dispersion (CD). Polarization mode dispersion distorts the temporal spread of an ultra-short pulse as different polarization states travel at different speeds through the transmission fiber and amplifier components. PMD is particularly a problem in the first generation of installed transmission fiber and through polarization components (e.g., isolators) that are not PMD compensated. Without PMD compensation, 40 Gbit/s transmission, for example, can be limited to very short distances (several km). Fortunately, PMD compensators have been developed for upgrading aged installed networks, and also amplifiers are available with very low PMD.

To achieve long-distance transmission, a chromatic dispersion map is produced for all the optical spans, and by using a dispersion compensating module in each amplifier, the net dispersion across the channel wavelength window is controlled. Although optical amplifiers have approximately 1000 times less optical fiber in them compared to the transmission span that it is amplifying, chromatic dispersion in the amplifier can also be of concern for very long-haul networks. The CD of the amplifier is primarily in the EDF, and is specific to the EDF's geometric and chemical composition. Furthermore, since the erbium ion is a resonant atomic system, there may be a pump and signal power contribution to the dispersion (resonant dispersion) that could degrade network performance. Characterization of these optical effects is typically done on manufactured 'field grade' amplifiers as part of a quality control process.

Conclusion

As optical networks evolve EDFAs will continue to be an enabling technology for higher capacity and more dynamic communications networks. EDFA technology has already advanced to accommodate multiple channels, to span several wavelength windows and to provide features such as dispersion compensation, gain-transient suppression, and dynamic gain flatness. As future network architectures are introduced, to incorporate Raman amplification and additional wavelength windows, the demands placed upon EDFA design will continue to expand.

See also

Optical Communication Systems: Architectures of Optical Fiber Communication Systems; Wavelength Division Multiplexing. **Scattering:** Raman Scattering.

Further Reading

Agrawal G (1989) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Agrawal G (1997) *Fiber Optic Communication Systems*. New York: John Wiley and Sons.

Becker PC, Olsson NA and Simpson JR (1999) *Erbium-doped Fiber Amplifiers – Fundamentals and Technology*. San Diego, CA: Academic Press.

Desurvire E (1995) *Erbium Doped Fiber Amplifiers*. New York: John Wiley and Sons.

Pedersen B, Bjarklev A, Povlsen JH, *et al.* (1991) The design of erbium-doped fiber amplifiers. *Journal of Lightwave Technology* 9(9): 1105.

Sudo S (1997) *Optical Fiber Amplifiers: Materials, Devices and Applications*. Boston, MA: Artech House.

Optical Amplifiers in Long-Haul Transmission Systems

B M Desthieux, Xtera Communications, Inc., Allen, TX, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical amplifiers have been a key element for enabling the development of optical transmission systems, which have reached total capacity in the order of a few terabits per second. The first type of optical amplifier to see application in the telecom industry has been the erbium-doped fiber amplifier. Following its invention in the late 1980s, the mid-1990s saw its industrial application in submarine and terrestrial transmission systems.

Since then, optical transmission systems have been the predominant source for transmitting information over distances ranging from a few 100 km to more than 10 000 km, connecting cities, countries, and continents.

This article will describe the applications of optical-fiber amplifiers in long-haul transmission systems, focusing on erbium-doped fiber amplifiers and Raman amplifiers, the most popular type of optical amplifiers used in modern transmission optical systems.

Fundamentals of Long-Haul Transmission Systems

This section explains the key elements and parameters of an optical transmission system.

Key Elements

An optical transmission system consists of a transmitter, a receiver, one or more optical amplifiers, and one or more spans of transmission fiber. Figure 1 shows a simplified schematic of a bidirectional transmission system using a fiber pair.

Optical amplifiers are used to compensate for the loss of the transmission fiber and the other optical elements placed along the signal path. Boosters and pre-amplifiers refer to optical amplifiers which are located at, respectively, the input end and output end of the system. In-line amplifiers refer to optical amplifiers located between the transmission fiber spans.

The transmitter includes laser diodes to generate the signal power and optical modulators to transfer the data stream from the electrical input signals to optical signals. The optical signal is transmitted to the receiver after traveling through the transmission fiber. At the receiver are located the photodiodes, which transfer back the information into electrical output signals. In order to increase the total bandwidth of the transmission system, time-division multiplexing (TDM) and wavelength-division multiplexing (WDM) can be used. TDM techniques combine the electrical domain data stream of several electrical channels, which is carried by a single optical channel. An optical channel refers to an optical signal located at a particular wavelength and modulated with the transmission data. TDM increases, therefore, the bit rate supported by the optical channel. Typical bit rates used in modern optical transmission systems are 2.5 Gbit/s and 10 Gbit/s. With WDM, several optical channels are combined, in the optical domain, into a single optical beam, which is launched into the transmission fiber. For this technique, optical multiplexers (Mux) and de-multiplexers (Demux) are required at, respectively, the transmitter and the receiver.

Long-haul transmission systems typically operate at wavelengths around 1.5 μm, where the attenuation of fibers is the lowest. We label C-band the region located between 1530 nm and 1565 nm. Its lower bound is labeled S-band; the upper bound the L-band.

Apart from attenuation, several other effects existing in fiber can be responsible for degrading the quality of the transmission. Among those is chromatic dispersion, which distorts the optical signal as it travels along the transmission fiber. Several types of fiber have been deployed with different chromatic dispersion characteristics. The most common one is the standard single-mode fiber (SSMF), which has a zero-dispersion wavelength – wavelength for which no chromatic dispersion is observed – near 1.3 μm. The dispersion-shifted fiber (DSF) has a zero-dispersion wavelength near 1.55 μm, and the nonzero

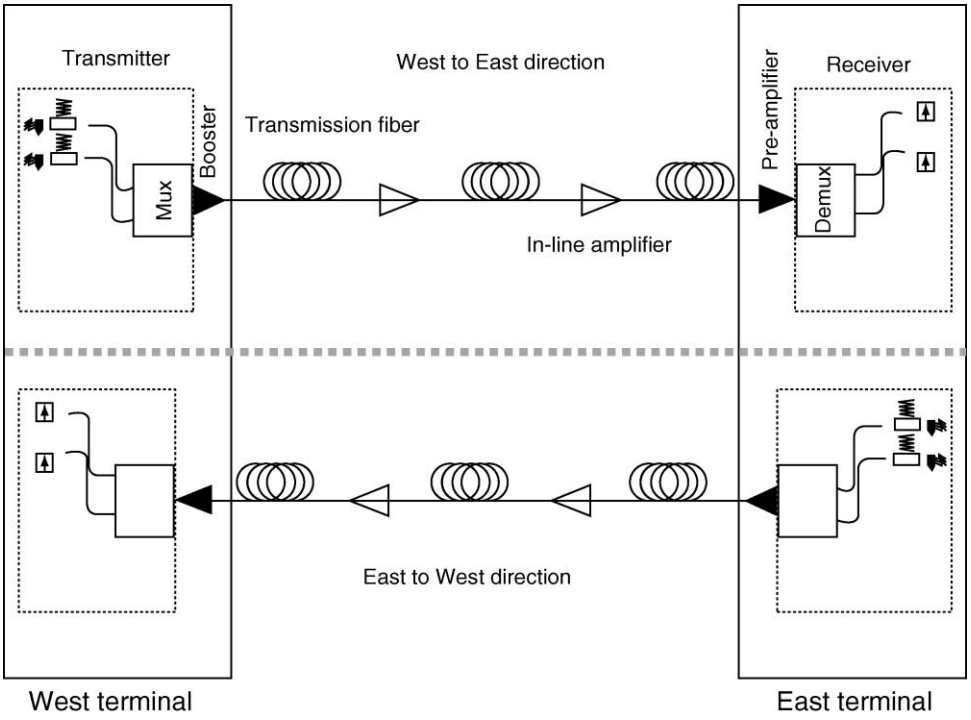


Figure 1 Schematic of a bidirectional long-haul transmission system.

dispersion shifted fiber (NZDSF) has a zero-dispersion wavelength close to 1.5 μm. To remedy this effect, dispersion compensating modules (DCM) are placed along the link. These modules – generally a special type of fiber – have chromatic dispersion with inverted characteristics with respect to those of the transmission fiber.

Key Optical-Amplifier Parameters

Optical amplifiers spontaneously generate optical noise – referred to as amplified spontaneous emission (ASE) – that is added to the modulated signal and consequently degrades the quality of the transmission. It is essential to quantify the level of ASE for a given optical transmission system.

We define the optical signal-to-noise ratio (OSNR) in the bandwidth $\Delta\nu$ as

$$\mathrm{OSNR} = \frac{P_{\mathrm{Signal}}}{P_{\mathrm{ASE}}} \qquad [1]$$

where P_{signal} is the signal power and P_{ASE} is the ASE power in the bandwidth $\Delta\nu$.

The required OSNR at a receiver, to detect the information with high quality, depends on the modulation format and on the other sources of noise or signal degradation existing in the transmission system. If we assume a nonreturn to zero (NRZ) modulation format, and signal detection limited only by ASE noise, an OSNR, in 0.1 nm bandwidth, of about 18 dB, is required at the receiver for a bit-error rate – number of erroneously transmitted bits divided by number of transmitted bits – of 10^{-12}. Introduction of other transmission degradation sources requires a higher OSNR to maintain the same bit-error rate.

The decibel unit (dB) for a given ratio, R, is defined by

$$R_{\mathrm{dB}} = 10\log_{10}R_{\mathrm{linear}} \qquad [2]$$

Similarly we defined the power decibel unit (dBm) as

$$P_{\mathrm{dBm}} = 10\log_{10}P_{\mathrm{mW}} \qquad [3]$$

where P_{mW} is the power in mW.

In order to quantify the amount of ASE generated by an optical amplifier, we define the equivalent input noise factor (n_{eq}), in a linear unit, by

$$n_{\mathrm{eq}} = \frac{P_{\mathrm{ASE}}}{h\nu\Delta\nu G} \qquad [4]$$

where P_{ASE} is the output ASE power in one polarization state in the bandwidth $\Delta\nu$, h the Planck's constant, ν the frequency of the ASE, and G the gain of the amplifier.

We also define the noise figure of the amplifier (NF) – a universal parameter to characterize

amplifier noise – as

$$\mathrm{NF} = 2n_{\mathrm{eq}} + \frac{1}{G} \qquad [5]$$

If an optical component, having a loss T_{in} ($T_{\mathrm{in}} < 1$), is inserted at the input of the amplifier, the total NF of the amplifier becomes $\mathrm{NF}_{\mathrm{total}} = \mathrm{NF}/T_{\mathrm{in}}$. If an optical component, having a loss T_{out} ($T_{\mathrm{out}} < 1$), is inserted at the output of the amplifier, $\mathrm{NF}_{\mathrm{total}} = \mathrm{NF} + 1/G(1/T_{\mathrm{out}} - 1)$. When $G \gg 1$, we have $\mathrm{NF}_{\mathrm{total}} \approx \mathrm{NF}$.

The OSNR in 0.1 nm bandwidth obtained after N fiber spans of loss, $\mathrm{Loss}_{\mathrm{Fiber}}$ (in dB), each having an amplifier, is in decibels:

$$\mathrm{OSNR} = P_{\mathrm{Signal}} - \mathrm{Loss}_{\mathrm{Fiber}} - \mathrm{NF} - 10\log_{10} N + 58 \qquad [6]$$

where P_{Signal} is the signal power at the output of the amplifier.

When the optical signal power is sufficiently high, nonlinear effects also become responsible for the degradation of the system performance. Therefore, choosing the value of the signal power launched into the transmission fiber results in a trade-off between increasing the OSNR, and reducing nonlinear effects.

Optical Fiber Amplifier

An optical fiber amplifier consists of fiber with an active medium and at least one optical pump source, which is launched simultaneously with the signal into the fiber by means of a dichroic coupler.

We define as forward and backward pumping, the case where the pump is launched, respectively, in the same direction as the signal and in the opposite direction to the signal. We define as bidirectional pumping, the case where the fiber is pumped from both ends. The most commonly used pump sources are the semiconductor laser diodes. These lasers may use different technologies to meet the various performance requirements with respect to the amplifier types and applications. The most popular pump diodes are the distributed Bragg reflector (DBR), distributed feedback (DFB), or external-grating stabilized lasers.

Erbium Fiber Amplifier

Fiber doped with the rare-earth Er^{3+} ion can amplify a signal located at 1.5 μm with the appropriate fiber host material and if pumped near 0.98 μm or near 1.48 μm. Silica is the most common material used for the fiber host. Er^{3+} has three energy levels, as described in Figure 2.

0.98 μm pumping occurs between the upper level $^4I_{11/2}$ and the ground level $^4I_{15/2}$, and 1.48 μm pumping occurs between the metastable level $^4I_{13/2}$ and the ground level.

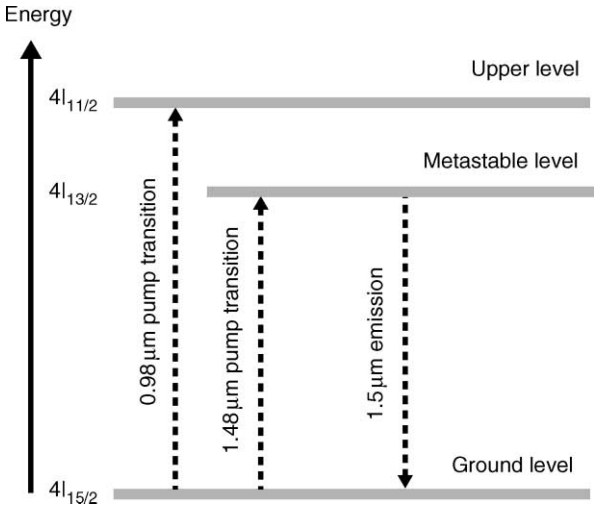


Figure 2 Energy level diagram for Er^{3+} ions.

The signal amplification takes place between the metastable level and the ground level. With 0.98 μm pumping, a rapid energy transfer occurs between the upper level and the metastable level via a phonon-assisted process. The lifetime of the upper level is much lower than the lifetime of the metastable level, about 10 ms. Therefore, with 0.98 μm pumping we can assume a two-level energy transfer, between the metastable and the ground levels, for the description of the amplification process in the Er^{3+} medium.

We define the normalized medium inversion (D) by

$$D = \frac{N_2 - N_1}{N_t} \qquad [7]$$

where N_2 is the population of the metastable level, N_1 the population of the ground level, and $N_t = N_2 + N_1$ the total population. $D = +1$ when the metastable level is fully populated and $D = -1$ when only the ground level is populated.

A simple approach for calculating the gain (G) with erbium-doped fiber amplifiers (EDFA) is given, in decibel units, by

$$G(\lambda) = \big((\alpha(\lambda) + g(\lambda)\big)\frac{D+1}{2} - \alpha(\lambda)L \qquad [8]$$

where λ is the wavelength of the signal, α the absorption coefficient in dB unit length at the signal wavelength, g the emission coefficient in dB unit length at the signal wavelength, and L the length of the erbium-doped fiber (EDF). Here D is the normalized path-average medium inversion of the EDF.

The gain per unit length, $G(\lambda)/L$, is equal to $g(\lambda)$ for $D = +1$ and to $-\alpha(\lambda)$ for $D = -1$.

The absorption and the emission coefficients depend on the EDF host material and the codoping element used for the realization of the EDF. With silicate EDF the most commonly used codoping elements are germanium, aluminum, and phosphorus.

In Figure 3 we show the dependence of the gain per unit length on the normalized path-average medium inversion, for a signal wavelength ranging from 1500 to 1560 nm.

For a more comprehensive description of the EDFA characteristics we use the standard confined-doping model. This model, which is also based on the two-level system, uses the following differential equations:

$$\frac{dP_S^+(z)}{dz} = \left(\frac{N_2}{N_t}(\alpha_S + g_S) - \alpha_s\right)P_S^+(z) \qquad [9]$$

$$\frac{dP_{ASE}^+(z)}{dz} = \left(\frac{N_2}{N_t}(\alpha_S + g_S) - \alpha_s\right) \times P_{ASE}^+(z) + \frac{N_2}{N_t} g_S 2h\nu_S \Delta\nu \qquad [10]$$

$$\frac{dP_{ASE}^-(z)}{dz} = \left(\alpha_s - \frac{N_2}{N_t}(\alpha_S + g_S)\right) \times P_{ASE}^-(z) - \frac{N_2}{N_t} g_S 2h\nu_S \Delta\nu \qquad [11]$$

$$\frac{dP_P^+(z)}{dz} = \left(\frac{N_2}{N_t}(\alpha_P + g_P) - \alpha_P\right)P_P^+(z) \qquad [12]$$

$$\frac{dP_P^-(z)}{dz} = \left(\alpha_P - \frac{N_2}{N_t}(\alpha_P + g_P)\right)P_P^-(z) \qquad [13]$$

where $P_S(z)$, $P_{ASE}(z)$, and $P_P(z)$ in W are, respectively, the signal power, the ASE power, and the pump power along the length (z-axis) of the EDF; $^+$ and $^-$ indicate, respectively, the backward-traveling and the forward-traveling directions; α_s and α_P in m^{-1} are the absorption coefficients at, respectively, the signal and pump wavelengths, g_s and g_p in m^{-1} are the emission coefficients at, respectively, the signal and pump wavelengths, ν_s and ν_p are the frequencies of, respectively, the signal and the pump beams; $\Delta\nu$ is the frequency bandwidth of the ASE; and h is the Planck's constant. N_2/N_1 is the medium inversion along the z-axis of the EDF, and is given by

$$\frac{N_2}{N_t} = \frac{\sum_k \dfrac{P_k(z)\alpha_k\tau}{h\nu_k B_{eff}^2 N_t}}{1 + \sum_k \dfrac{P_k(z)(\alpha_k + g_k)\tau}{h\nu_k B_{eff}^2 N_t}} \qquad [14]$$

where the index k represents all the beams, i.e., signal, ASE, and pump, traveling backward and forward along the EDF, τ is the lifetime of the metastable level, and B_{eff} is the effective erbium-doping area. For a typical silicate EDF, having a peak absorption coefficient of $1\ dBm^{-1}$, we have $\tau = 10$ ms, $B_{eff} = 12\ \mu m^2$, and $N_t = 5 \times 10^{23}\ m^{-3}$.

Using the standard confined-doping model we calculate the dependence of the EDFA noise figure on the normalized path-average medium inversion, as shown in Figure 4. We vary the path-average normalized medium inversion by changing the length of the EDFA, while adjusting the pump power so that the signal gain is kept constant.

The noise figure decreases when the medium inversion increases, and approaches the quantum limit value of 3 dB when the path-average normalized medium inversion is close to unity. Near full inversion

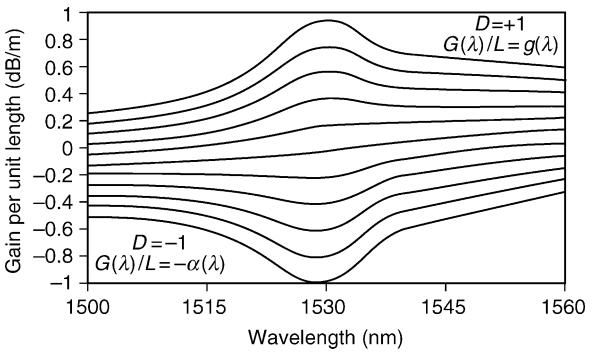


Figure 3 Dependence of the gain per unit length on the normalized path-average medium inversion, D. D ranging from -1 to $+1$ with a 0.2 step (alumino-silicate EDF; peak absorption of 1 dB m^{-1}; peak emission of 0.93 dB m^{-1}).

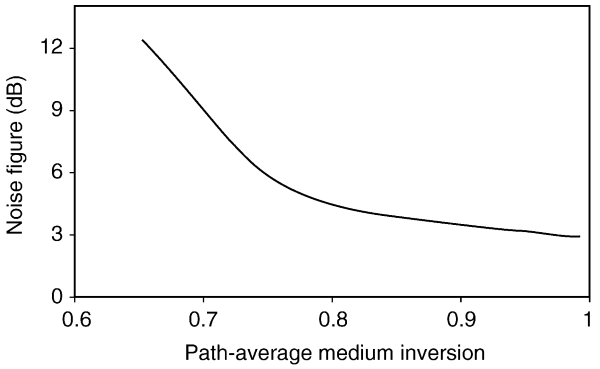


Figure 4 Dependence of the noise figure on the normalized medium inversion (alumino-silicate EDF with peak absorption of 1 dB m^{-1}; EDF lengths ranging from 16 m to 60 m; gain equals to 15 dB; input signal power of -40 dBm at 1530 nm; backward pumping at 980 nm).

may be obtained with a pump located at 0.98 μm because the emission coefficient is null at this pump wavelength. However, this is not the case at 1.48 μm where the ratio g_p/α_p is about 0.3. The maximum medium inversion that can be obtained with 1480 nm is $D = 0.8$. Therefore, higher noise figures are obtained with 1.48 μm pumping than with 0.98 μm pumping. The pumping direction also has an effect on the noise figure, in particular with 0.98 μm. Since at 0.98 μm the emission coefficient is null, when backward pumping configuration is used the 0.98 μm pump is quickly absorbed along the EDF length. This leads to a low-medium inversion at the input end of the EDF, resulting in high noise figures. However, in order to achieve forward pumping, a dichroic coupler must be placed at the input end of the EDF, thus inducing an additional loss at the input of the amplifier. This increases the noise figure and, therefore, to some extent, counterbalances the NF improvement associated with forward pumping.

Raman Fiber Amplifier

Stimulated Raman scattering, an optical-phonon based nonlinear effect present in silica fiber, is used to convert energy from a pump located at a frequency ν_p, to a signal located at a lower frequency ν_s. The efficiency of the so-called Raman amplification depends on the frequency shift $\Delta\nu_r = \nu_p - \nu_s$. Figure 5 shows the dependence of the Raman gain on the frequency shift in standard single-mode fibers. The exact Raman gain spectrum depends on the doping elements used during the fabrication process of the fiber. With standard telecommunications fibers the Raman gain is maximum for $\Delta\nu_r$ close to 13 THz.

Raman amplification can be achieved within the transmission fiber of the system – distributed amplification – or within a specific fiber, such as a dispersion compensation fiber – discrete amplification. Contrary to EDFA, Raman amplification has a very fast response time – in the order of a few hundred femtoseconds – and is strongly dependent on the polarization state of the signal and the pump, the highest gain efficiency being obtained when the polarization states of the signal and the pump are parallel. Because of the nonpolarization-maintaining nature of standard telecommunications fibers, depolarized pump sources are preferable for Raman amplification. To avoid pump fluctuation being transferred to the signal, Raman amplifiers generally use backward pumping, which virtually averages the fluctuation. Forward pumping can also be employed with Raman amplifiers but with careful control of the pump characteristics.

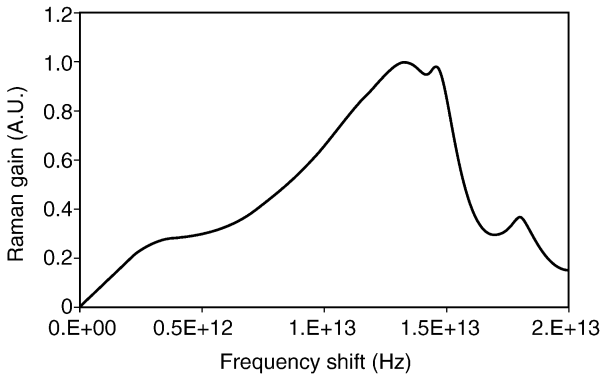


Figure 5 Raman gain spectrum in standard single mode fiber.

If the signal is sufficiently low the pump depletion can be neglected. In that case the gain of the Raman amplifier is given by

$$G = \exp\left(\frac{g_R P_P}{A_{eff}\alpha_p}(1 - \exp^{-\alpha_p L}) - \alpha_S L\right) \qquad [15]$$

where P_p is the launched pump power, L is the length of the fiber, g_R is the Raman gain coefficient, and A_{eff} is the effective area of the fiber. We assume for the simulation examples given in this article that the peak value of g_R is equal to 6×10^{-14} m/W.

A more comprehensive model for Raman amplification is given by the following set of differential equations:

$$\frac{dP_S^+(z)}{dz} = -\alpha_s P_S^+(z) + \left(\frac{g_R}{A_{eff}}\right)(P_p^+(z) + P_p^-(z))P_S^+(z) + r_S P_S^-(z) \qquad [16]$$

$$\frac{dP_S^-(z)}{dz} = \alpha_s P_S^-(z) - \left(\frac{g_R}{A_{eff}}\right)(P_p^+(z) + P_p^-(z))P_S^-(z) - r_S P_S^+(z) \qquad [17]$$

$$\frac{dP_{ASE}^+(z)}{dz} = -\alpha_s P_{ASE}^+(z) + \left(\frac{g_R}{A_{eff}}\right)(P_p^+(z) + P_p^-(z)) \times [P_{ASE}^+(z) + 2h\nu_S\Delta\nu] + r_S P_{ASE}^-(z) \qquad [18]$$

$$\frac{dP_{ASE}^-(z)}{dz} = \alpha_s P_{ASE}^-(z) - \left(\frac{g_R}{A_{eff}}\right)(P_p^+(z) + P_p^-(z)) \times [P_{ASE}^-(z) + 2h\nu_S\Delta\nu] - r_S P_{ASE}^+(z) \qquad [19]$$

$$\frac{dP_p^+(z)}{dz} = -\alpha_p P_p^+(z) - \left(\frac{\nu_p}{\nu_S}\right)\left(\frac{g_R}{A_{eff}}\right)P_p^+(z) \times (P_S^+(z) + P_S^-(z) + P_{ASE}^+(z) + P_{ASE}^-(z)) + r_P P_p^-(z) \qquad [20]$$

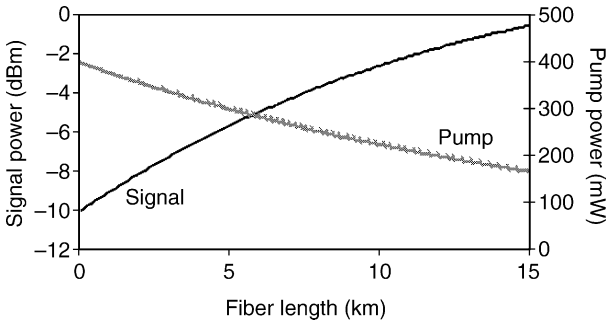


Figure 6 Evolution of the signal power and pump power along fiber length with Raman amplification (standard single-mode fiber; fiber length of 15 km; input signal power of −10 dBm at 1550 nm; forward pumping with 400 mW at 1450 nm; total amplifier gain of 9.5 dB).

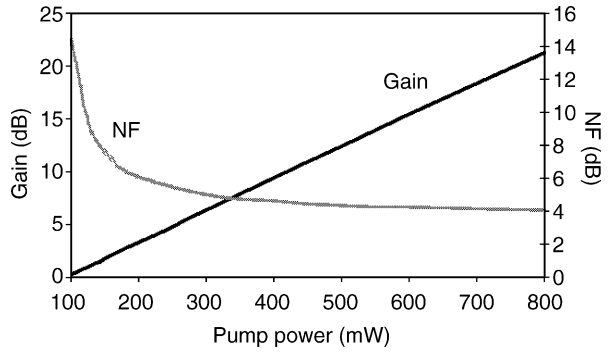


Figure 7 Dependence of the Raman amplifier gain and the noise figure on the pump power (standard single-mode fiber; fiber length of 15 km; input signal power of −10 dBm at 1550 nm; forward pumping at 1450 nm with power ranging from 100 mW to 800 mW).

$$\frac{dP_p^-(z)}{dz} = \alpha_p P_p^-(z) + \left(\frac{\nu_p}{\nu_S}\right)\left(\frac{g_R}{A_{eff}}\right) P_p^-(z) \times \left(P_S^+(z) + P_S^-(z) + P_{ASE}^+(z) + P_{ASE}^-(z)\right) - r_p P_p^+(z) \quad [21]$$

where r_S and r_p are the Rayleigh backscattering coefficients at, respectively, the signal wavelength and the pump wavelength.

An example of the evolution of the signal power and pump power along fiber length is given by Figure 6.

Figure 7 shows the dependence of the Raman amplifier gain and noise figure on the pump power.

Distributed and Discrete Amplification

A system uses distributed amplification when the transmission fiber is used as the medium for generating the optical signal amplification. It uses discrete amplification when an optical amplifier is located between the transmission fiber spans.

Discrete Amplification

Both EDFA and Raman amplifiers can be used to provide discrete amplification. Discrete amplifiers can be used as booster, pre-amplifier, or in-line amplifier. They can also be used to compensate for the loss of additional elements such as DCM, which may be located between the fiber spans. In Figure 8 we show the path of the signal when a discrete amplification is used to compensate for the loss of the transmission fiber and for the loss of a DCM. In this example, we assume that the discrete amplifier is made of two amplifiers. The first-stage amplifier compensates mainly for the loss of the transmission fiber, while the second-stage amplifier compensates mainly for the loss of the DCM.

If a 'lossy' element is inserted between two amplifiers, the noise figure of the composite amplifier increases the total noise figure of the dual-stage amplifier.

The total noise figure of a dual-stage amplifier, NF_{total}, is given by

$$NF_{total} = NF_1 - \frac{1}{G_1} + \frac{NF_2}{G_1 T_{mid}} \quad [22]$$

where NF_1 is the noise figure of the first-stage amplifier, G_1 is the gain of the first-stage amplifier, NF_2 is the noise figure of the second-stage amplifier, and T_{mid} is the loss of the optical element inserted between the two amplifiers ($T_{mid} < 1$).

Distributed Amplification

When fibers are deployed it is generally assumed, at least for terrestrial networks, that it can be used with different types of transmission systems, which may or

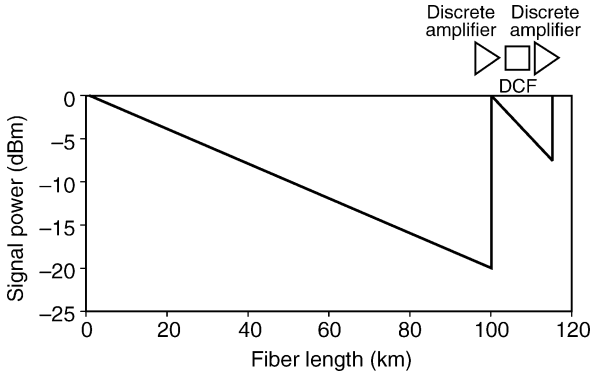


Figure 8 Schematic of a transmission system using discrete amplification to compensate for the loss of the transmission fiber and the dispersion compensating module.

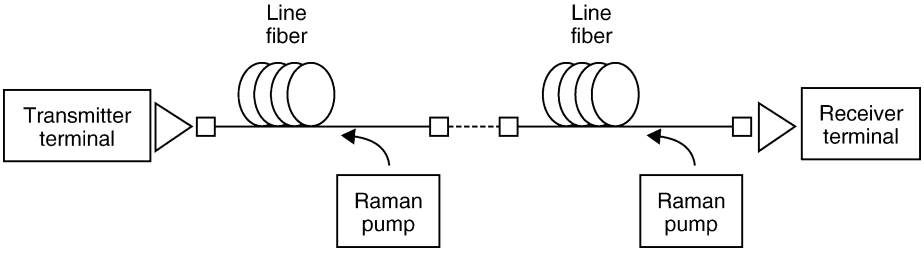


Figure 9 Schematic of a transmission system using distributed Raman amplification.

may not use distributed amplification. Doping the transmission fiber with erbium is not a practical solution since the absence of a pump would result in high fiber loss. On the other hand, Raman amplification can be achieved with conventional fibers, and is consequently the preferred solution to generate distributed amplification. Raman amplification is then obtained by launching a pump signal into the transmission fiber. Figure 9 shows the schematic of a transmission system using distributed Raman amplification with backward pumping.

Even if distributed amplification can be used to compensate fully for the loss of the transmission fiber, discrete amplifiers are generally required to compensate for the loss of the transmitter and receiver terminals. In order to compare distributed amplification with discrete amplification we define the equivalent noise figure as follows:

$$\mathrm{NF_{Equ}} = \frac{1}{G_{\mathrm{On/Off}}}\left(\frac{\mathrm{Pase}}{h\nu\Delta\nu} + 1\right) \qquad [23]$$

where Pase is the total ASE power at the output end of the transmission fiber and where the on/off gain, $G_{\mathrm{On/Off}}$ is given by

$$G_{\mathrm{On/Off}} = \frac{P^{\mathrm{On}}_{\mathrm{signal}}}{P^{\mathrm{Off}}_{\mathrm{signal}}} \qquad [24]$$

where $P^{\mathrm{On}}_{\mathrm{signal}}$ and $P^{\mathrm{Off}}_{\mathrm{signal}}$ are the signal power levels at the output end of the transmission fiber when the distributed amplifier is, respectively, turned on and off.

To illustrate the meaning of the on/off gain, we show in Figure 10 the evolution of the signal power along a fiber with and without Raman pump power. In this case, the on/off gain is 20 dB.

The equivalent noise figure, which has no real physical meaning, is equal to the noise figure of an imaginary discrete amplifier that would be located at the output end of the transmission fiber and would lead to the same OSNR. Figure 11 shows the dependence of the equivalent noise figure on the Raman gain, with no Rayleigh backscattering and with a Rayleigh backscattering level of $-74\ \mathrm{dB\ m^{-1}}$.

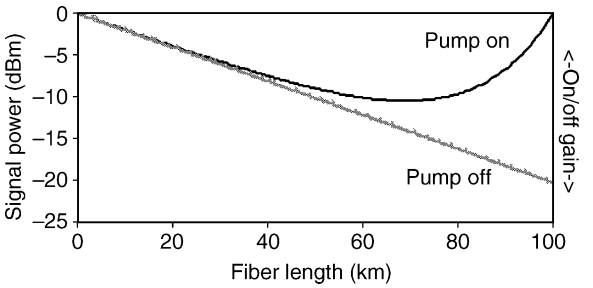


Figure 10 Evolution of signal power along the fiber length with and without pump power (standard single-mode fiber; fiber length of 100 km; fiber attenuation of 0.2 dB/km at signal wavelength; input signal power of 0 dBm at 1550 nm; backward pumping 400 mW at 1450 nm).

Once the on/off gain is higher than ~10 dB the equivalent noise figure becomes negative and it decreases when the on/off gain further increases. However, once the on/off is higher than 20 dB, Rayleigh backscattering limits the minimum equivalent noise figure that can be achieved. The lowest equivalent noise figure is observed for an on/off gain of ~33 dB. In addition, Rayleigh backscattering causes the reflected portion of the signal to be superimposed onto the transmitted signal, leading to multipath interference (MPI), which level increases with the gain. This results in an additional source of noise for the signal, and consequently degrades the system performance.

Hybrid Amplification

Discrete and distributed amplification can be used simultaneously in a same-transmission system. An example of the evolution of the signal power with hybrid amplification is represented in Figure 12, for the case where distributed Raman amplification and a dual-stage discrete amplifier are used to compensate for the loss of the transmission fiber and for the loss of a DCM.

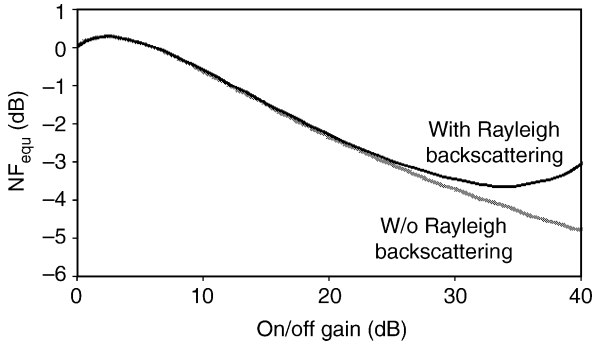


Figure 11 Equivalent noise figure versus distributed on/off gain (standard single-mode fiber; fiber length of 250 km; fiber attenuation of 0.2 dB/km at 1550 nm; input signal power of 0 dBm at 1550 nm; backward pumping ranging from 0 to 1 W at 1450 nm).

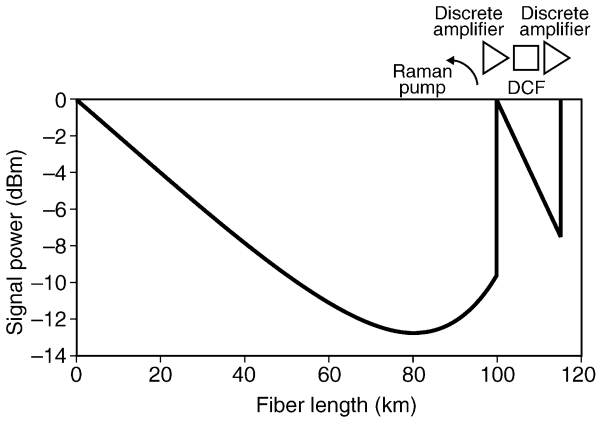


Figure 12 Evolution of the signal power using discrete and distributed amplification (standard single-mode fiber of 100 km; fiber loss of 20 dB at signal wavelength; DCM loss of 8 dB; input signal power of 0 dBm at 1550 nm; Raman backward pumping with 200 mW at 1450 nm; Raman on/off gain of 10 dB).

Similar to the case of distributed amplification, the total equivalent noise figure of a hybrid amplifier can be defined by

$$NF_{Equ}^{Total} = NF_{Equ}^{Distributed} + \frac{NF^{Discrete} - 1}{G_{On/Off}} \qquad [25]$$

where $NF_{Equ}^{Distributed}$ is the equivalent noise figure of the distributed amplifier and $NF^{Discrete}$ is the noise figure of the discrete amplifier.

In Figure 13 we show the dependence of the total equivalent noise figure on the on/off gain for a hybrid amplifier with fixed discrete noise figure. The dependence of the distributed equivalent noise figure on the on/off gain is also plotted on the same graph.

Since the equivalent noise figure of the distributed amplifier is relatively low, the total equivalent noise figure is lower than the noise figure of the discrete amplifier once distributed amplification is used. Moreover, as the on/off gain becomes large, the total equivalent noise figure of the hybrid amplifier converges toward the distributed equivalent noise figure.

Wavelength-Division Multiplexing

The transmission of WDM signals over long distances requires meticulous control of the spectral characteristics of the amplifier gain. The amplifier gain excursion induces a power excursion between the channels of the WDM signal, which increases with the number of spans. Consequently, the available bandwidth for WDM transmission narrows with the total distance of the link. This is illustrated in Figure 14, where the power excursion observed after an amplifier chain is plotted for different numbers of amplifiers, a fiber span being included between each amplifier. With this example, the 1 dB-bandwidth drops from 9 nm after one span to 1.5 nm after 50 spans.

For a comprehensive analysis of the spectral gain characteristic of EDFA and Raman amplifiers we generalize the single-channel model described in the amplifier section of this article to the multichannel case. The generalization of the models is obtained by adding a frequency dimension to each beam propagating in the amplifier, as described in Table 1, where each k element represents the signal, the ASE noise or the pump, λ is the wavelength of the kth beam, $\Delta\nu$ is the frequency separation between the pump beam p, and the signal beam s.

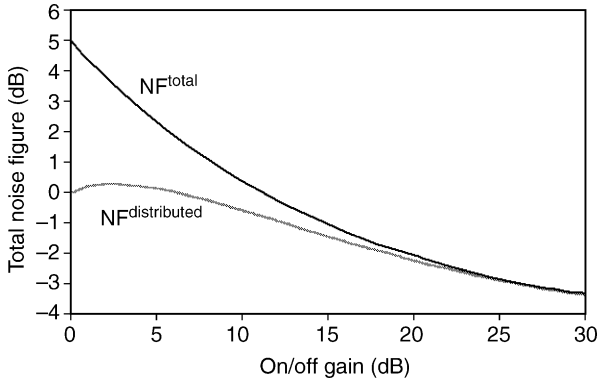


Figure 13 Total noise figure of the hybrid amplifier and noise figure of the distributed amplifier versus on/off gain (noise figure of the discrete amplifier equals to 5 dB; standard single-mode fiber of 100 km; fiber loss of 20 dB at signal wavelength; input signal power of 0 dBm at 1550 nm; Raman backward pumping ranging from 0 mW to 600 mW at 1450 nm).

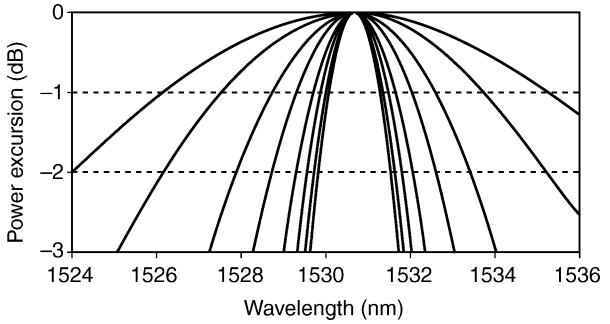


Figure 14 Dependence of the power excursion on the number of fiber spans. Number of spans equals, respectively, to 1, 2, 5, 10, 20, 30, 40, and 50 (amplifier with 15 m long alumino-silicate EDF; peak absorption of 1 dB m^{-1}; $D = 0.6$; fiber loss of 8.5 dB; flat fiber loss spectrum).

Table 1 Correspondence between parameters used for the EDFA and Raman amplifier models with the single-channel and the multichannel approaches (the multichannel approach is also used to describe the power evolution of pump beam when WDM pump scheme is used)

	Single-channel model	*Multichannel model*
Power	$P_k^{+,-}(z)$	$P_k^{+,-}(\lambda, z)$
Sum of power	$P_k^{+}(z)$ $+P_k^{-}(z)$	$\sum_\lambda P_k^{+}(\lambda, z)$ $+P_k^{-}(\lambda, z)$
Absorption coefficient	α_k	$\alpha(\lambda)$
Emission coefficient	g_k	$g(\lambda)$
Raman gain coefficient	g_R	$g_R(\Delta\nu_{p,s})$

With EDFAs, the main factors in the gain flatness are the path-average medium inversion, the EDF codopant composition, and fiber host material. Depending on the bandwidth and the wavelength of the WDM signal, the way the EDF is optimized may vary.

In the following example, we focus on the path-average medium inversion for the case of C-band applications. Figure 15 shows the output spectrum of a WDM signal amplified by an alumino-silicate EDFA having, respectively, a medium inversion of $D = +0.4$ and $D = 0$. The peak-to-peak gain excursion is, respectively, 2 dB and 14 dB for $D = +0.4$ and $D = 0$.

Application with C-band WDM signals requires amplifiers operating with high-medium inversion. As can be extrapolated from Figure 15, if the WDM signal is centered around 1560 nm, a lower medium inversion is more suitable. Indeed, as shown in Figure 3, when the medium inversion is close to zero the peak gain wavelength changes from 1530 nm to 1560 nm and higher; the optical window for L-band applications.

Regarding the doping elements, we show in Figure 16 the gain spectrum for a germano-silicate and an alumino-silicate EDF, respectively. The peak-to-peak gain excursion is 8 dB with the germano-silicate EDF compared to 2 dB with the alumino-silicate EDF.

With Raman amplifiers, the gain flatness can be adjusted by using wavelength-multiplexed pumps. In Figure 17 we show the power spectrum when a Raman amplifier is pumped with three wavelength-multiplexed pumps located, respectively, at 1430 nm, 1455 nm, and 1480 nm. The maximum power excursion between the channels is 2 dB.

For comparison, we also show in Figure 18 the power spectrum of the Raman amplifiers when only

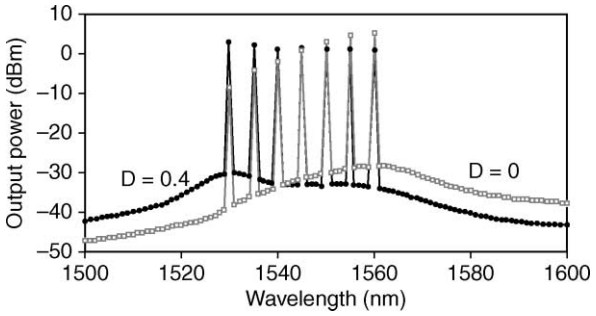


Figure 15 Gain spectrum of an alumino-silicate EDF for $D = 0.4$ and $D = 0$ (C-band WDM signal made of 7 channels equally spaced ranging from 1530 nm to 1560 nm; flat input power of −10 dBm per channel; gain of channel located at 1545 nm of 10 dB for $D = 0.4$ and $D = 0$; EDF with peak absorption of 1 dB m^{-1}; EDF length of 30 m for $D = 0.4$; EDF length of 100 m for $D = 0$).

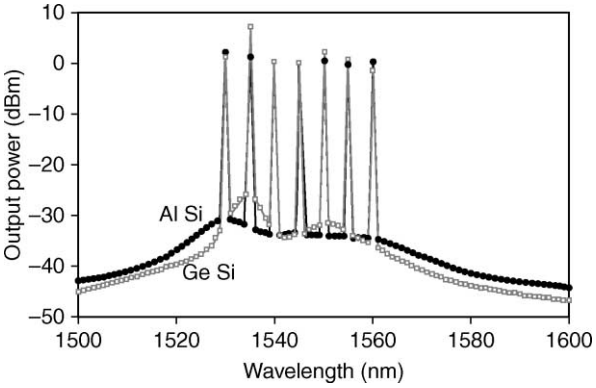


Figure 16 Gain spectrum of an alumino-silicate EDF and of a germano-silicate EDF for $D = 0.4$ (C-band WDM signal made of 7 channels equally spaced ranging from 1530 nm to 1560 nm; flat input power of −10 dBm per channel; gain of channel located at 1545 nm of 10 dB for $D = 0.4$ and $D = 0$; alumino-silicate EDF of 30 m with peak absorption of 1 dB m^{-1}; germano-silicate EDF of 40 m with peak absorption of 1 dB m^{-1}).

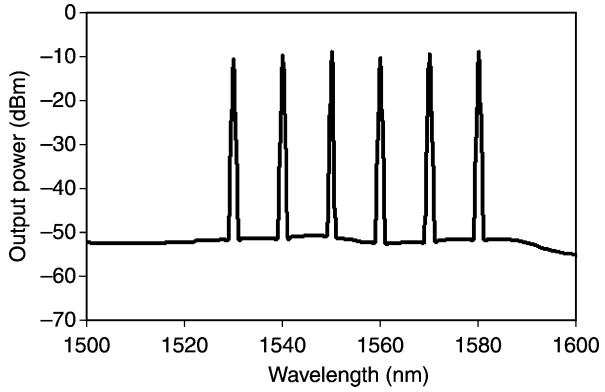


Figure 17 Output power spectrum of a Raman amplifier pumped with three wavelength-multiplexed pumps (standard single-mode fiber of 100 km; WDM signal made of 6 channels equally spaced ranging from 1530 nm to 1580 nm; pumping in forward direction with 100 mW per pump located at 1430 nm, 1455 nm, and 1480 nm).

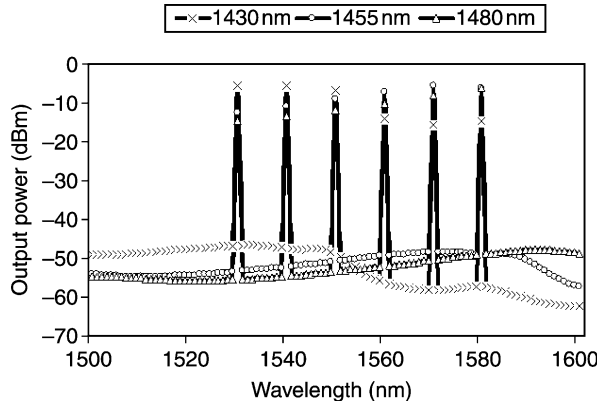


Figure 18 Output power spectrum of Raman amplifiers pumped at three different wavelengths (standard single-mode fiber of 100 km; WDM signal made of 6 channels equally spaced ranging from 1530 nm to 1580 nm; pumping in forward direction with 300 mW at, respectively, 1430 nm, 1455 nm, and 1480 nm).

one of the three multiplexed pumps is used, with the same total pump power. The power excursion becomes 10 dB for the 1430 nm pump, 7 dB for the 1455 nm pump, and 8.5 dB for the 1480 nm pump, which are much higher values than are obtained with the multiplexed pump scheme.

To further reduce the gain excursion of the amplifiers an external gain equalizing filter can be used. The filter, which is located next to the amplifier, is manufactured so that its attenuation spectrum matches the gain spectrum of amplifier, as illustrated in Figure 19. Gain equalizing filters are generally made with thin film filters or fiber Bragg gratings (FBG).

In order to obtain amplification over ultra-wide bandwidths, amplifiers of smaller bandwidths, e.g., C-band and L-band EDFAs, can be operated in parallel by means of Mux and Demux. Another option is to use a Raman amplifier with multiple pumps having their wavelengths distributed over a very large bandwidth.

Long-Haul Transmission System Types

Long-haul optical transmission systems are appropriate for different types of applications. We distinguish here three types of transmission system: the submarine systems; the terrestrial systems; and the unrepeatered systems. Each of these applications requires a careful design of the system, with particular attention to the optical amplifiers. In this section we focus on the key points of their design.

Submarine Systems

Submarine systems are used to transmit information across the oceans, i.e., over distances of up to several thousand kilometers. The transmission fiber and the in-line optical amplifiers, which are packaged in the repeater modules, are deployed undersea at the same time. The drastic constraints imposed on the reliability of the repeaters lead to their high cost, and consequently minimizing their number in a link is essential. The optimum span length, and therefore the number of in-line amplifiers, depends on the total distance of the transmission system. To illustrate this point, we derive from eqn [6] which number of spans of a particular length can be used for a given received OSNR, launched signal power, amplifier noise figure, and span loss per km. Based on the number of spans and span length, we then calculate the maximum distance that can be reached. The result is represented in Figure 20.

With the assumption made here regarding the amplifier characteristics, in order to reach distances of more than 5000 km, span lengths close to 50 km are required.

Terrestrial Systems

Terrestrial systems are used to connect traffic from cities to cities and from cities to submarine-system landing sites. Their total distances vary from a few hundred kilometers to several thousand kilometers. Generally with terrestrial systems, the deployments of the transmission fibers and the transmission systems are done at different times. During its lifetime, the same fiber plant might be used with a different generation of transmission system. Furthermore, because of the geographical constraints that can be imposed on the

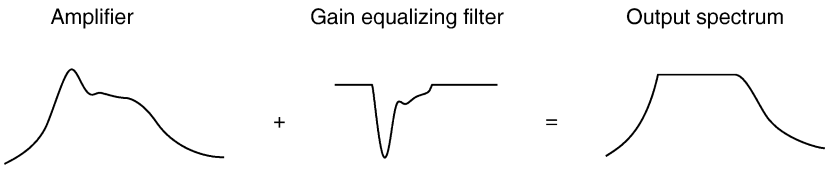


Figure 19 Combination of optical amplifier and gain equalizing filter to provide a flat gain amplifier.

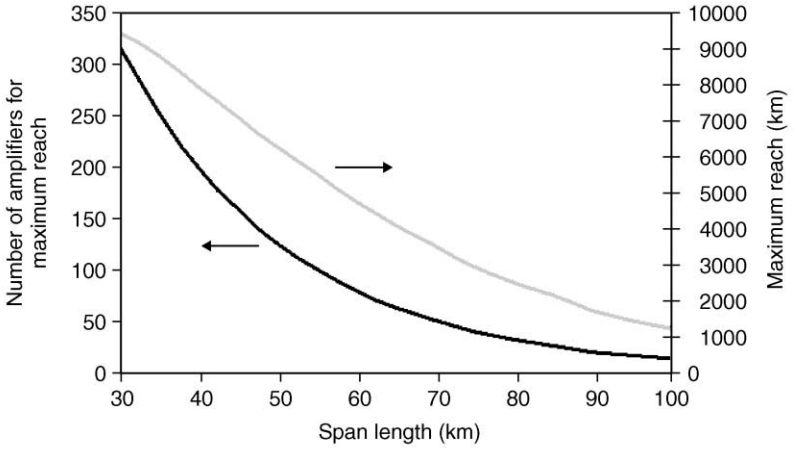


Figure 20 Dependence of the maximum distance and span number of a system on its span length (signal power at the output of the amplifier of −2 dBm; amplifier noise figure of 5 dB; fiber loss of 0.2 dB/km; required OSNR of 20 dB/0.1 nm).

locations of the terminals and the in-line optical amplifiers, the different span lengths may greatly vary in a given link. Consequently, an important characteristic of the terrestrial systems is their capability to operate with different span parameters.

For terrestrial systems using large numbers of WDM channels, maintaining the gain excursion of the optical amplifiers as little as possible under different operating conditions is crucial. Approaches to limit the gain excursion under these variations can be taken. With EDFAs of a given length, the medium inversion depends on its operating gain. The EDFA can be forced to operate at constant gain, independent of the span loss, by adding an adjustable attenuator to the amplifier, e.g., in the middle of a dual-stage amplifier, such that the sum of the attenuator loss and the span loss remains constant. With Raman amplifiers, the control of the gain excursion can be realized by adjusting the power of the pumps when a wavelength-multiplexed pumping scheme is used. Another possibility is to insert dynamically adjustable gain flattening devices along the link. The automatic adjustment of the gain flatness requires the need of a feedback generally provided by an optical channel power monitor.

Unrepeatered Systems

Unrepeatered systems are used to connect mainland sites separated by a water area, such as islands or cities located on the coast. Their main characteristic is that they do not use any electronic component along the transmission fiber; therefore optical amplifiers cannot be placed under water. The length of such systems is typically of a few hundred kilometers, meaning that, even for those systems where very low loss fiber is used, the loss of the transmission fiber is in the order of several tens of dB. To allow sufficiently high OSNR at the receiver, several options can be used. Among them are the use of very high-power boosters, distributed Raman amplification, and remotely pumped discrete amplification. Remote amplification is obtained by inserting, along the submerged fiber cable, a piece of EDF. A 1.48 μm pump is launched into the EDF from the mainland via the transmission fiber or via a dedicated fiber running along the transmission fiber. Figure 21 shows a schematic of a possible configuration for unrepeatered systems when a very high-power booster, distributed Raman amplification, and remote discrete amplification are used simultaneously.

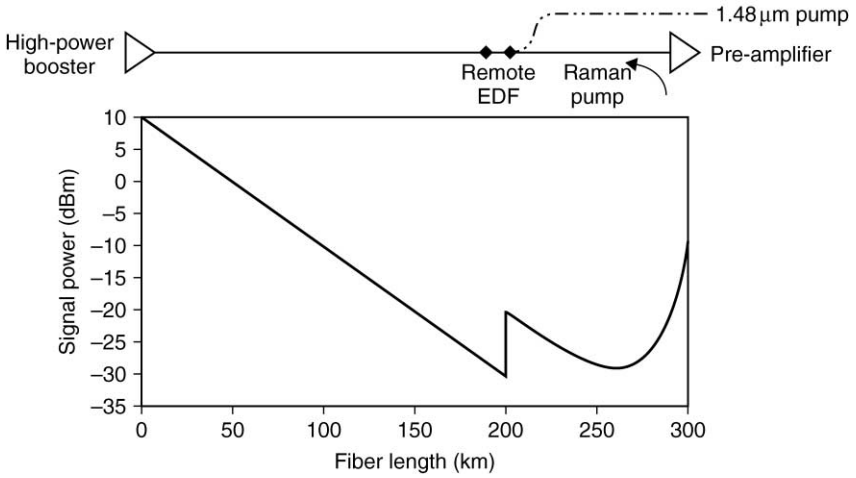


Figure 21 Schematic of unrepeatered systems using high-power booster, remote EDFA, and distributed Raman amplification.

The total equivalent noise figure, NF_{Equ}^{Total}, of an unrepeatered transmission system, using remote discrete amplification and distributed amplification, is given by

$$NF_{Equ}^{Total} = T NF^{Remote} + \frac{1}{G_{Remote}} \times \left(NF_{Equ}^{Distributed} + \frac{NF^{Discrete} - 1}{G_{On/Off}} - T \right) \quad [26]$$

where NF^{Remote} and G_{Remote} are, respectively, the noise figure and the gain of the remote amplifier; T is the loss of the transmission fiber between the remote amplifier and the pre-amplifier ($T < 1$); $NF_{Equ}^{Distributed}$ and $G_{On/Off}$ are, respectively, the equivalent noise figure and on/off gain of the distributed amplifier; and $NF^{Discrete}$ is the noise figure of the pre-amplifier.

Assuming that the gain and noise figures of the remote EDFA are, respectively, 10 dB and 7 dB, that the remote EDFA is located at 100 km from the pre-amplifier, that the fiber loss per km is 0.2 dB/km, that the Raman distributed amplifier provides 30 dB of on/off gain with an equivalent noise figure of −3 dB, and that the pre-amplifier noise figure is 5 dB, the equivalent total noise figure of the unrepeatered system is −10 dB.

List of Units and Nomenclature

A_{eff}	effective area
ASE	amplified spontaneous emission
B_{eff}	effective erbium doping area
D	normalized medium inversion
dB	ratio in decibel
dBm	power in decibel
DBR	distributed Bragg reflector
DCM	dispersion compensating modules
Demux	optical de-multiplexers
DFB	distributed feedback
DSF	dispersion-shifted fiber
EDF	erbium-doped fiber
EDFA	erbium-doped fiber amplifiers
FBG	fiber Bragg gratings
g_R	Raman gain coefficient
G	gain
$G_{On/Off}$	on/off gain
h	Planck's constant
Length	meter (m)
MPI	multi-path interference
Mux	optical multiplexers
n_{eq}	equivalent input noise factor
NF	noise figure of the amplifier
NRZ	non-return to zero
NZDSF	non-zero dispersion shifter fiber
N_2/N_1	medium inversion
OSNR	optical signal-to-noise ratio
Power	watt (W)
SSMF	standard single-mode fiber
TDM	time-division multiplexing
Time	second (s)
WDM	wavelength-division multiplexing
τ	lifetime

See also

Coherent Lightwave Systems. **Dispersion Management**. **Fiber and Guided Wave Optics:** Nonlinear Effects (Basics). **Optical Amplifiers:** Basic Concepts; Erbrium Doped Fiber Amplifiers for Lightwave Systems; Raman, Brillouin and Parametric Amplifiers.

Optical Communication Systems: Architectures of Optical Fiber Communication Systems; Basic Concepts; Historical Development; Lightwave Transmitters; Optical Time Division Multiplexing; Wavelength Division Multiplexing. **Solitons:** Soliton Communication Systems.

Further Reading

Agrawal G (2001) *Nonlinear Fiber Optics*. London: Academic Press.

Chesnoy J (2002) *Undersea Fiber Communication Systems*. London: Academic Press.

Desurvire E (1994) *Erbium-Doped Fiber Amplifiers, Principles and Applications*. New York: John Wiley & Sons.

Desurvire E, Bayart D, Desthieux B and Bigo S (2002) *Erbium-Doped Fiber Amplifiers, Device and System Developments*. New York: John Wiley & Sons.

Emori Y, Kado S and Namiki S (2002) Broadband flat-gain and low-noise Raman amplifiers pumped by wavelength-multiplexed high-power laser diodes. *Optical Fiber Technology* 8: 107.

Hansen PB and Eskildsen L (1997) Remote amplification in repeaterless transmission systems. *Optical Fiber Technology* 3: 221.

Islam MN (2004) *Raman Amplifiers for Telecommunications*. New York: Springer-Verlag.

Kaminow I and Li T (2002) *Optical Fiber Telecommunications IV*. London: Academic Press.

Keiser G (2000) *Optical Fiber Communications*. New York: McGraw-Hill.

Raman, Brillouin and Parametric Amplifiers

M F S Ferreira, University of Aveiro, Aveiro, Portugal

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Glass fibers for optical communications are made of fused silica, an amorphous material, to which dopant materials of various kinds can be added to produce changes in refractive index. A number of third-order nonlinear processes can occur; these can grow to appreciable magnitudes over the long lengths available in fibers, even though the nonlinear coefficients in the materials are relatively small. The effects are particularly important in single-mode fibers, in which the small mode field dimensions result in substantially high light intensities with relatively modest input powers.

Nonlinear optical effects, such as stimulated Raman scattering (SRS), stimulated Brillouin scattering (SBS), and four-wave mixing (FWM), can cause some problems such as nonlinear loss, pulse distortion, and cross-talk between signals in wavelength division multiplexed (WDM) transmission systems. However, these effects are also useful for efficient amplification and frequency conversion of injected signals, as well as for generation of new frequencies. In this chapter the main properties of fiber Raman amplifiers (FRAs), fiber Brillouin amplifiers (FBAs), and fiber parametric amplifiers (FPAs) will be reviewed, regarding their applications in the field of fiber optic communications.

Fiber Raman Amplifiers

Stimulated Raman scattering is likely to occur in fibers, producing appreciable amplification for down-shifted signals when moderate pump power levels are used. This is largely because the process is phase-matched, so that the long interaction lengths available in fibers will enhance the process. Additionally, the small mode field cross-section yields high intensities for relatively modest power levels.

In the SRS process one has simultaneously the absorption of a photon from the pump beam at frequency ω_p and the emission of a photon at the signal (Stokes) frequency, stimulated by the incident signal beam at frequency ω_s (Figure 1a). The difference in energy is taken up by a high energy

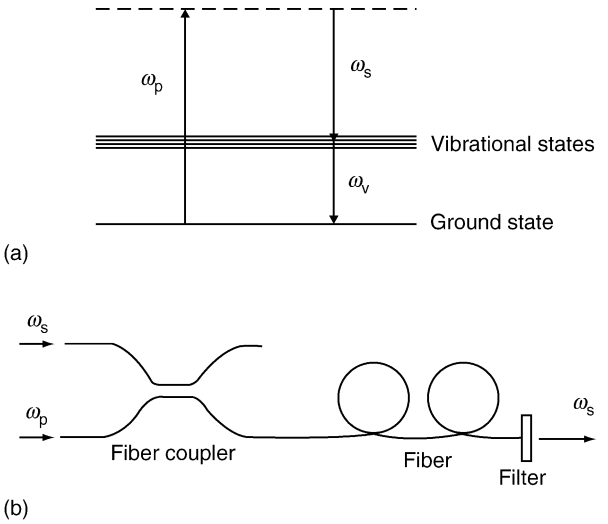


Figure 1 (a) Energy-level illustration of the process of stimulated Raman scattering and (b) schematic diagram of a fiber Raman amplifier.

phonon (molecular vibration) at frequency ω_v. Thus, SRS provides for energy gain at the signal frequency at the expense of the pump.

Figure 1b shows schematically a fiber Raman amplifier. The pump and the signal beams are injected into the fiber through a WDM fiber coupler. The case illustrated in Figure 1a shows the two beams co-propagating inside the fiber, but the counter-propagating configuration is also possible. In fact, the SRS can occur in both directions, forwards and backwards. It was confirmed experimentally that the Raman gain is almost the same in both cases.

Raman Gain and Bandwidth

Figure 2 shows the Raman gain coefficient for fused silica as derived from the spontaneous Raman spectrum. At the peak, the Raman gain coefficient for silica fibers is 9.4×10^{-14} m W^{-1} for a pumping wavelength $\lambda_p = 1.0\,\mu$m and varies as λ_p^{-1}. At this wavelength, a shift of 15 THz corresponds to 50 nm. The most significant feature of the Raman gain in silica fibers is that $\bar{g}_R$ extends over a large frequency range (up to 40 THz) with a broad dominant peak near 13 THz. This behavior is due to the amorphous nature of silica glass, whose molecular vibrational energy levels merge together to form a band.

The use of oxide glasses as dopants for fiber fabrication is suitable to enhance the Raman gain coefficient. For example, in the case of pure GeO_2, the Raman gain coefficient is about 9.2 times that for pure SiO_2 glass. In fact, doping the core of an SiO_2 Raman fiber amplifier with GeO_2 has two beneficial effects. First, the index difference Δ is raised, which increases the effective waveguiding and in turn reduces the effective cross-sectional area A_e. As a consequence, the rate of SRS is increased. Secondly, the Raman gain coefficient increases above that of pure SiO_2 in proportion to the concentration GeO_2, and is given by

$$g_R = (1 + 80\Delta)\bar{g}_R \qquad [1]$$

where $\bar{g}_R$ is the Raman coefficient for pure SiO_2.

In spite of the positive aspects mentioned above, doping with GeO_2 also increases the fiber losses, due to the concomitant rise in Rayleigh scatter and in the number of dopant-dependent loss centers. Thus, greatly increasing the GeO_2 concentration will not be worthwhile if the increased fiber losses outweigh the improvement in amplifier gain.

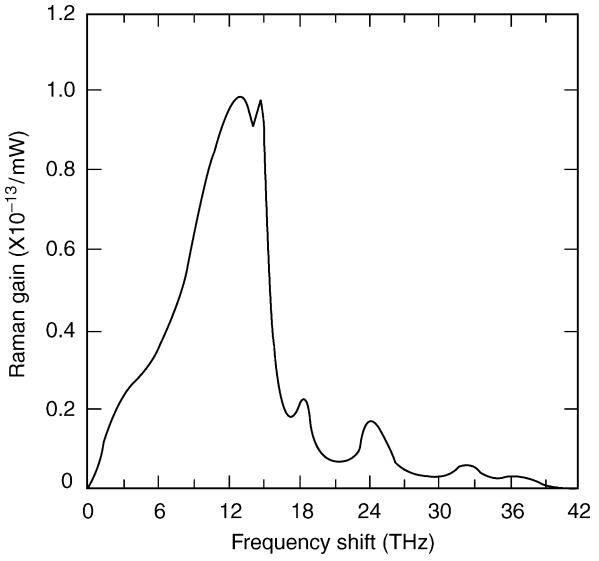


Figure 2 Measured Raman gain spectrum for fused silica at a pump wavelength $\lambda_p = 1.0\,\mu$m. The Raman gain scales inversely with λ_p. Reproduced with permission from Stolen RH and Ippen EP (1973) *Applied Physics Letters* 22: 276. © 1973 American Institute of Physics.

Gain Saturation

The interaction between the pump and Stokes waves is governed by the following coupled equations:

$$\frac{dP_s}{dz} = -\alpha_s P_s + \frac{g_R}{A_e} P_p P_s \qquad [2]$$

$$\frac{dP_p}{dz} = -\alpha_p P_p - \frac{\omega_p}{\omega_s}\frac{g_R}{A_e} P_s P_p \qquad [3]$$

where A_e is the effective cross-sectional area of the fiber mode, α_s and α_p are absorption coefficients which account for the fiber loss at the signal and pump frequencies, respectively, and the signal wave is considered to be co-propagating with the pump wave. Assuming that $\alpha_s = \alpha_p = \alpha$ (which is a reasonable approximation around the 1.5 μm wavelength region in a low loss fiber), the following approximate solutions of eqns [2] and [3] can be written when $P_p(0) \gg P_s(0)$:

$$P_p(z) = \frac{P_p(0)\exp(-\alpha z)}{1 + F(z)} \qquad [4]$$

$$P_s(z) = \frac{P_s(0)\exp\left(\Gamma(1 - e^{-\alpha z}) - \alpha z\right)}{1 + F(z)} \qquad [5]$$

where

$$\Gamma = \frac{g_R P_p(0)}{\alpha A_e} \qquad [6]$$

and

$$F(z) = \frac{\omega_p P_s(0)}{\omega_s P_p(0)} \exp\left(\Gamma(1 - e^{-\alpha z})\right) \qquad [7]$$

The dimensionless parameter $F(z)$ accounts for the effects of pump depletion and the attendant saturation of the gain seen by the signal.

Figure 3 shows the output signal power $P_s(L)$ versus the input signal power $P_s(0)$ for a Raman amplifier with length $L = 8$ km and several values of the input pump power. Typical values $g_R = 6.7 \times 10^{-14}$ m W^{-1}, $\alpha = 0.2$ dB km^{-1}, $A_e = 30 \times 10^{-12}$ m^2, $\lambda_p = 1.46$ μm and $\lambda_s = 1.55$ μm were assumed. The transfer characteristics are linear for $P_p(0) = 0.1$, 0.2, and 0.4 W, but the effects of pump depletion become discernible for input signal levels $P_s(0) > -15$ dB m when $P_p(0) = 0.6$ W. For a pump power $P_p(0) = 0.8$ W the linear behavior of the Raman amplifier cannot be observed, even for signal power levels as low as -25 dB m.

Since, in the absence of Raman amplification, the signal power at the amplifier output would be $P_s(L) = P_s(0) \exp(-\alpha L)$, the amplifier gain is given by

$$G_R = \frac{P_s(L)}{P_s(0) \exp(-\alpha L)} = \frac{\exp(g_R P_p(0) L_e / A_e)}{1 + F(L)} \qquad [8]$$

where

$$L_e = \frac{1}{\alpha}[1 - \exp(-\alpha L)] \qquad [9]$$

is the effective interaction length and L is the actual fiber length. The gain given by eqn [8] is seen to be a function of the input signal power (i.e., a saturation nonlinearity) through the term $F(L)$.

When $F(L) \gg 1$ we have from eqn [5], that

$$P_s(L) = \frac{\omega_s}{\omega_p} P_p(0) \exp(-\alpha L) \qquad [10]$$

In this case, the amplifier output reaches the pump level irrespective of the input signal level. This implies that any spontaneous Raman scatter in the fiber will be amplified up to power levels comparable to that of the pump. This is obviously inadmissible in an amplifier application, thus this operating regime must be avoided.

On the other hand, when $F(L) \ll 1$ we have, from eqn [5], that

$$P_s(L) = P_s(0) \exp(g_R L_e P_p(0)/A_e - \alpha L) \qquad [11]$$

In this case, the amplifier gain is

$$G_R = \frac{P_s(L)}{P_s(0) \exp(-\alpha L)} = \exp(g_R L_e P_0 / A_e) \qquad [12]$$

From eqn [12], the Raman gain in decibels is expected to increase linearly with the pump input power. Raman gains were experimentally confirmed to be almost the same for forwards and backwards pumping.

Figure 4 illustrates the variation of the amplifier gain (full curves, left scale) and the saturation parameter $F(L)$ (dashed curves, right scale) with the pump power $P_p(0)$ for a fixed input signal power $P_s(0) = 0$ dBm and amplifier lengths $L = 2$, 4, and 8 km. The full curve corresponding to $L = 2$ km shows a linear dependence of the gain on the pump power. However, the other two full curves show how the pump depletion limits the achievable gain. The black point on these curves marks the value of the pump power at which the depletion parameter F

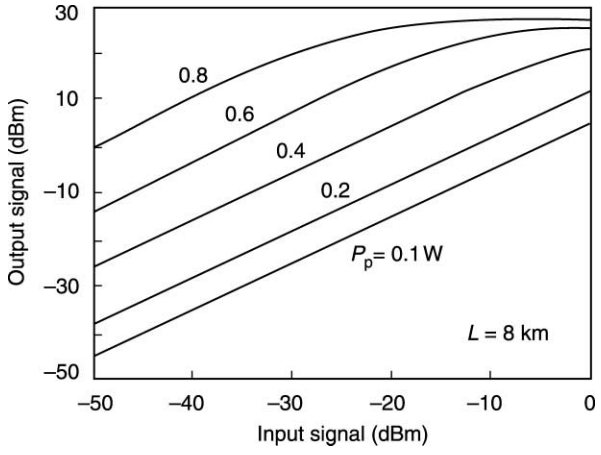


Figure 3 Transfer characteristics for a Raman amplifier with a length $L = 8$ km and several values of the input pump power.

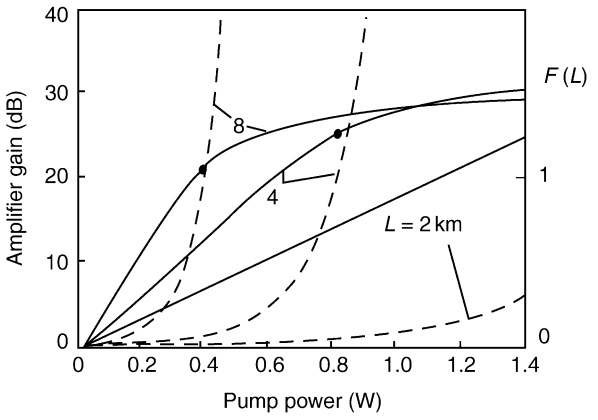


Figure 4 Raman amplifier gain (full curves, left scale) and parameter $F(L)$ (dashed curves, right scale) against input pump power for three values of the amplifier length.

becomes equal to unity. Only operating at pump powers sufficiently below each point ensures linear operation of the amplifier, whilst choosing to operate the amplifier significantly above that point will mean $F \gg 1$ and spontaneous scatter will be amplified to intolerable levels.

Noise Light Power

The dominant noise light in fiber Raman amplification is due to the amplified spontaneously Raman scattered light. In fact, a part of the pump energy is spontaneously converted into Stokes radiation extending over the entire bandwidth of the Raman gain spectrum and is amplified together with the signal. The output thus consists not only of the desired signal but also of background noise extending over a wide frequency range (~10 THz or more).

It has been shown that, for both forwards and backwards pumpings, the noise light power is equivalent to a hypothetical injection of a single-photon per unit frequency at the fiber input end for forwards pumping, and at some distance away from the fiber output for backwards pumping. Assuming that $G_R \gg 1$ and $\alpha_L \gg 1$, the noise light powers for forwards pumping, $P_{fasp}(L)$, and backwards pumping, $P_{basp}(L)$, are approximately given by

$$P_{fasp}(L) \approx h\nu_s \Delta\nu_R (G_R - 1) \exp(-\alpha L) \qquad [13]$$

$$P_{basp}(L) \approx h\nu_s \Delta\nu_R (G_R - 1)/\ln G_R \qquad [14]$$

where $\Delta\nu_R$ is the Raman gain bandwidth, h is Planck's constant, and ν_s is the signal frequency. We can observe from eqn [13] that the noise light power in the case of forward pumping decreases as the fiber length is increased. However, in the case of backwards pumping, it depends mainly on the Raman gain, being nearly independent of fiber length and loss.

FRA Performance and Applications

One main limitation affecting the operation of fiber Raman amplifiers is the requirement of a relatively high-power laser as a pump source. The experiments near 1.55 μm are usually carried out using color-center lasers for this purpose. However, such lasers are too bulky for communication applications and the use of compact high-power laser diodes (LDs) becomes desirable. This objective has recently been achieved with the advent of high-power multimode LDs. In some cases, the output lights from various LDs are combined to attain the necessary pump power.

In a Raman amplifier, SRS has to compete frequently with other nonlinear effects. For example, since the intrinsic gain coefficient for SBS is two orders of magnitude larger than that for SRS, in some cases the SBS may occur at lower pump powers and may affect significantly the SRS process. In particular, the Raman gain may not only be reduced but also become unstable. However, the SBS can be suppressed taking into account its narrow linewidth, which is typically less than 100 MHz. Therefore, if a pump laser with linewidth broader than the Brillouin bandwidth is employed, the SBS effect would not be significant. When a semiconductor laser is used as pump source, its spectral linewidth can be artificially increased by direct frequency modulation, resulting in effective SBS suppression.

Considering its broad bandwidth (>5 THz), fiber Raman amplifiers can be used to amplify several channels simultaneously in a multichannel communication system. The same characteristic also makes such amplifiers suitable for amplification of short optical pulses.

Considering a repeaterless transmission system, the FRA can be used in three different configurations: forwards pumping, backwards pumping, and bi-directional pumping. In the first case, the signal amplification occurs near the fiber input, where the signal power is relatively high. This configuration for the Raman amplification does not generally affect the transmission quality and can compensate for the transmission loss by approximately the value of the Raman gain. For example, 10 dB Raman gain provides an increase in transmission distance of 50 km when the fiber attenuation is 0.2 dB km^{-1}. In the case of backwards pumping, the signal amplification occurs near the fiber output end before photodetection, where the signal power may be very weak, comparable with the Raman noise power. This scheme for the Raman amplification can considerably reduce the minimum detectable signal power, which results in a corresponding increase of the transmission distance. However, the transmission quality is affected due to backwards noise light. The bi-directional pumping configuration uses both forwards and backwards pumping simultaneously and its performance can be estimated from those for the respective unidirectional schemes considered above.

The FRAs can also be used in nonregenerative all-optical multirepeater systems. In this case, the transmission loss is compensated for by both forwards and backwards Raman gain providing the periodically injected pump light. This scheme can be used to overcome the fiber loss in soliton-based communication systems. In this case, the pulse broadening due to chromatic dispersion of the

fiber is also eliminated, which provides the possibility of achieving ultralong distance distortionless transmission.

Fiber Brillouin Amplifiers

The process of SBS can be described as a classical three-wave interaction involving the incident (pump) wave of frequency ω_p, the Stokes wave of frequency ω_s, and an acoustic wave of frequency ω_a. The pump creates a pressure wave in the medium owing to electrostriction, which in turn causes a periodic modulation of the refractive index. This process is illustrated schematically in Figure 5. Since the acoustic wavefronts are moving away from the incident pump wave, the scattered light is shifted downward in frequency to the Stokes frequency:

$$\omega_s = \omega_p - \omega_a \qquad [15]$$

The response of the material to the interference of the pump and Stokes fields tends to increase the amplitude of the acoustic wave. Therefore, the beating of the pump wave with the acoustic wave tends to reinforce the Stokes wave, whereas the beating of the pump wave and the Stokes waves tends to reinforce the acoustic wave. This explains the appearance of the stimulated Brillouin scattering process.

In spite of the apparent similarity between SBS and SRS, these processes differ in three important aspects: i) Brillouin amplification occurs only when the pump and signal beams counter-propagate inside the fiber; ii) the Stokes shift for SBS is smaller by three orders of magnitude compared with that of SRS; and iii) the Brillouin gain spectrum is extremely narrow, with a bandwidth <100 MHz. These characteristics significantly affect the performance and the fiber Brillouin amplifiers.

Brillouin Gain and Bandwidth

Physically, each pump photon in the SBS process gives up its energy to create simultaneously a Stokes photon and an acoustic phonon. Since both the energy and the momentum must be conserved in this process, the frequencies of the three waves must satisfy eqn [15], whereas the wave vectors are related by

$$\vec{k}_p = \vec{k}_s + \vec{k}_a \qquad [16]$$

where $\vec{k}_p$ and $\vec{k}_s$ are the wave vectors of the pump and signal beams, respectively. The frequency ω_a and the wave vector $\vec{k}_a$ of the acoustic wave satisfy the dispersion relation:

$$\omega_a = |\vec{k}_a| v_a = 2v_a |\vec{k}_p| \sin(\theta/2) \qquad [17]$$

where v_a is the acoustic velocity, θ is the angle between the pump and signal waves, and $|\vec{k}_p| \approx |\vec{k}_s|$ was used. Equation [17] shows that ω_a vanishes in the forward direction ($\theta = 0$), whereas it is maximum in the backward direction ($\theta = \pi$). The frequency shift ν_B in the backward direction is given by

$$\nu_B = \frac{\omega_a}{2\pi} = \frac{2nv_a}{\lambda_p} \qquad [18]$$

where eqn [17] was used with $|\vec{k}_p| = 2\pi n/\lambda_p$, λ_p is the pump wavelength, and n is the refractive index. Considering $n = 1.45$ and $v_a = 5.96$ km s^{-1} as typical values for silica fibers, we obtain $\nu_B = 11.1$ GHz at $\lambda_p = 1.55$ μm.

Assuming a monochromatic pump, the spectrum of the scattered light depends on the uniformity of the fiber parameters. For a uniform medium the spectrum is Lorentzian with a width $\Delta\nu_B$ determined by the acoustic attenuation. For 1.5 μm light the width is about 17 MHz. However, both the frequency shift ν_B and the gain bandwidth $\Delta\nu_B$ can vary from fiber to fiber because of the guided nature of light and the presence of dopants in the fiber core. In particular, the inhomogeneities in the core section along the length determine an increase of the amplifier bandwidth, which can exceed 100 MHz, although typical values are 50–60 MHz for $\lambda_p \approx 1.55$ μm.

The peak value of the Brillouin gain coefficient is given by

$$g_B(\nu_B) = \frac{2\pi n^7 p_{12}^2 \gamma}{c\lambda_p^2 \rho_0 v_a \Delta\nu_B} \qquad [19]$$

where p_{12} is the longitudinal elasto-optic coefficient, ρ_0 the average material density, n the refractive index, c the vacuum speed of light, and γ is a numerical factor that depends on the polarization properties of the optical fiber, which assumes the value $\frac{1}{2}$ for long lengths of nonpolarization-preserving optical fiber. By using the parameter values typical of fused silica, g_B is estimated to be about 2.5×10^{-11} m W^{-1}, which is two orders of magnitude larger than the Raman gain coefficient at $\lambda_p = 1.55$ μm.

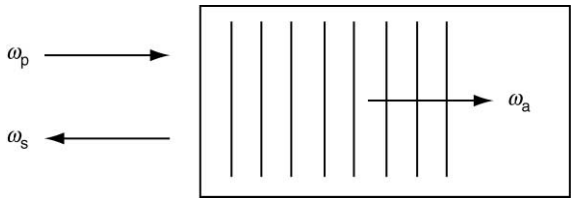


Figure 5 Schematic illustration of the stimulated Brillouin scattering process.

The expression for the Brillouin gain in eqn [19] assumes that the spectral width of the pump beam ($\Delta\nu_p$) is much narrower that the Brillouin linewidth ($\Delta\nu_B$). When such condition is not verified, the Brillouin gain is reduced according with the relation:

$$\tilde{g}_B(\nu_B) = \frac{\Delta\nu_B}{\Delta\nu_B + \Delta\nu_p} g_B(\nu_B) \qquad [20]$$

For this reason, fiber Brillouin amplifiers need to be pumped by narrow-linewidth semiconductor lasers ($\Delta\nu_p < 10$ MHz).

Gain Saturation

The slowly varying steady-state Fourier amplitudes of the Stokes (E_s) and pump (E_p) electric fields are related by the following coupled-wave equations:

$$\frac{\partial E_s}{\partial z} = \left[j\frac{K_1 K_2 |E_p|^2}{\delta - j\Gamma} + \frac{\alpha}{2} \right] E_s \qquad [21]$$

$$\frac{\partial E_p}{\partial z} = -\left[j\frac{K_1 K_2 |E_s|^2}{\delta + j\Gamma} + \frac{\alpha}{2} \right] E_p \qquad [22]$$

where Γ^{-1} is the acoustic phonon lifetime resulting in a spontaneous Brillouin scattering linewidth $\Delta\nu_B = \Gamma/\pi$, $\alpha = \alpha_s \approx \alpha_p$ is the absorption coefficient, and the coupling constants are

$$K_1 = \frac{K_2 \rho_0 n^2 \varepsilon_0}{2\nu_a}, \qquad K_2 = \frac{n^3 p_{12} \omega_s (\hat{e}_p \times \hat{e}_s)}{2c\rho_0} \qquad [23]$$

ε_0 being the free-space permittivity and $\hat{e}_p$ and $\hat{e}_s$ the unit vectors in the directions of the pump and Stokes optical fields, respectively. In eqns [21] and [22] it is assumed that the pump wave is launched at $z = 0$ and propagates in the $+z$ direction, whereas the signal wave is launched at $z = L$ and propagates in the $-z$ direction. The parameter $\delta \equiv \omega_p - \omega_s - \omega_a$ takes into account a possible detuning of the signal from the SBS gain line center.

Equations [21] and [22] can be written in terms of optical power as follows:

$$\frac{dP_s}{dz} = \alpha P_s - \frac{g_B(\delta)}{A_e} P_p P_s \qquad [24]$$

$$\frac{dP_p}{dz} = -\alpha P_p - \frac{g_B(\delta)}{A_e} P_s P_p \qquad [25]$$

where A_e is the effective core area and $g_B(\delta)$ is the Brillouin gain coefficient, given by

$$g_B(\delta) = \frac{\Gamma^2}{\delta^2 + \Gamma^2} g_B(\nu_B) \qquad [26]$$

where $g_B(\nu_B)$ is given by eqn [19].

Considering that $a \ll 1$, the following approximate solutions of eqns [24] and [25] can be written:

$$P_s(z) = P_s(0) D(z) e^{\alpha z} \qquad [27]$$

$$P_p(z) = P_p(0) D(z) H(z) e^{-\alpha z} \qquad [28]$$

where

$$D(z) = \frac{[P_p(0) - P_s(0)]}{[P_p(0) H(z) - P_s(0)]} \qquad [29]$$

and

$$H(z) = \exp\left\{ g_B [P_p(0) - P_s(0)][1 - \exp(-\alpha z)]/(A_e \alpha) \right\} \qquad [30]$$

The Brillouin amplifier gain is given from eqn [27] as

$$G_B = \frac{P_s(0)}{P_s(L) e^{-\alpha L}} = \frac{1}{D(L)} \qquad [31]$$

Figure 6 shows the output signal power $P_s(0)$ versus the input signal power $P_s(L)$ for a Brillouin amplifier with length $L = 8$ km and several values of the input pump power. Typical values $g_B(0) = 2.5 \times 10^{-11}$ m W^{-1}, $\alpha = 0.2$ dB km^{-1}, $A_e = 30 \times 10^{-12}$ m^2 were assumed. The transfer characteristics are linear for $P_p(0) = 0.2$ mW, but the effects of pump depletion become discernible

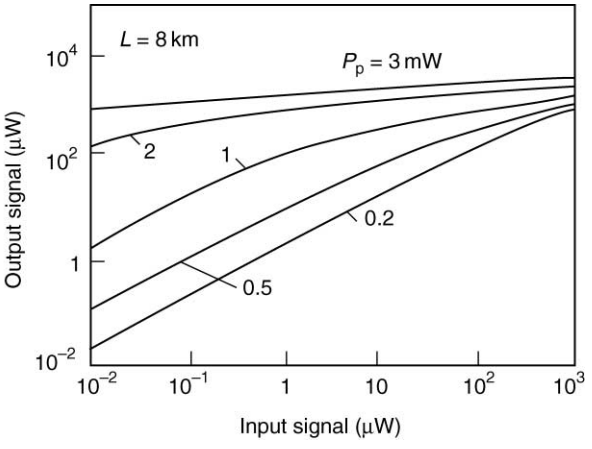


Figure 6 Transfer characteristics for a Brillouin amplifier with a length $L = 8$ km and several values of the input pump power.

for input signal levels $P_s(0) > 2\ \mu W$ when $P_p(0) = 1$ mW. For a pump power $P_p(0) = 2$ mW, linear behavior of the Brillouin amplifier cannot be observed even for signal powers as low as 10 nW.

The dependence of the Brillouin amplifier gain on the input pump power is illustrated in Figure 7 for several amounts of the signal detuning and two values of the signal power. The fiber length is $L = 5$ km and the gain bandwidth is $\Delta\nu_B = 30$ MHz, whereas the other parameter values are equal to those assumed in Figure 6. The saturated gain shows a significant dependence on the input signal power. An input signal of 1 nW, which is comparable to the original value of spontaneous emission, can be amplified by about 60 dB for a pump power $P_p(0) \approx 5$ mW. However, the amplifier gain can be considerably reduced when the signal is detuned from the gain peak. For example, for an input signal power $P_s(L) = 1$ nW and a pump power $P_p(0) \approx 3$ mW, a detuning of 15 MHz (which corresponds to a reduction of the gain coefficient to half of its maximum value) determines a reduction of about 25 dB on the amplifier gain.

Substituting eqn [28] for $P_p(z)$ in eqn [21], the following result is obtained for the Stokes field at the amplifier output:

$$E_s(0) = E_s(L)\sqrt{G_B}\,\exp\{-\alpha L/2 - j\phi\} \qquad [32]$$

where

$$\phi = \frac{\delta}{2\Gamma}\ln(G_B) \qquad [33]$$

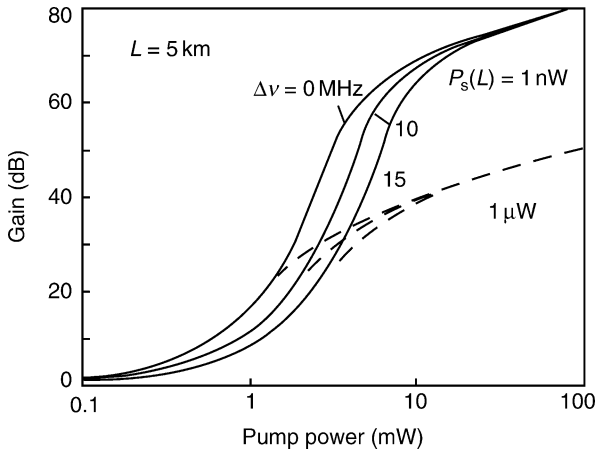


Figure 7 Brillouin amplifier gain against the input pump power for an amplifier length $L = 5$ km and different amounts of detuning from the gain center. The full curves and the dashed curves correspond to input signal powers $P_s(L) = 1$ nW and $P_s(L) = 1\ \mu$W, respectively. Reproduced with permission from Ferreira MFS, *et al.* (1994) *Optical and Quantum Electronics* 26: 35, with permission from Kluwer Academic Publisher.

represents a nonlinear phase change, which depends on the SBS gain, input signal power, and frequency detuning. This nonlinear phase shift can impose a stringent limit on the pump- and signal-frequency stability when FBAs are used in some phase-sensitive detection schemes.

Noise Light Power

To describe the spontaneous emission noise we must add a spontaneous emission term to the equations describing the evolution of the Stokes and pump waves. The amplified spontaneous scattered power per unit frequency is then obtained in the form:

$$P_{asp}(\nu, z) = (h\nu/A_e)g_B(\delta)(N+1)G(\nu, z) \times \int_z^L P_p(z')[G(\nu, z')]^{-1}dz' \qquad [34]$$

where

$$G(\nu, z) = \exp\left\{\int_z^L [P_p(z)g_B(\delta)/A_e - \alpha]dz\right\} \qquad [35]$$

is the gain function, h is Planck's constant, and N is the thermal equilibrium number of acoustic phonons, given by

$$N = \frac{1}{\exp(h\nu_B/kT)^{-1}} \approx kT/h\nu_B \approx 500 \qquad [36]$$

where k is the Boltzmann constant and T is the absolute temperature. The total amplified signal and spontaneous powers are given by integration of $P_s(\nu, z)$ over the gain profile.

Similarly to the behavior of the gain, the noise power depends significantly on the signal magnitude in the saturation regime and becomes particularly high for low signal powers. This high noise level, which is about 20 dB above that of an ideal amplifier, imposes some limitations on the use of FBA as a receiver preamplifier.

FBA Performance and Applications

Fiber Brillouin amplifiers can provide high gains at low pump powers. However, the gain bandwidth is small and the amplified spontaneous emission noise is substantially larger than that observed with other amplifiers, leading to a noise figure that is quite large (>15 dB). This fact limits their usefulness as a preamplifier. Use as in-line amplifiers at high signal levels is possible in some configurations. Some problems in this application are concerned with the narrow bandwidth and the small saturation power of

the FBAs. In particular, the Brillouin linewidth strictly limits the bandwidth of data signals that can be amplified in a FBA.

The Brillouin gain bandwidth can be intentionally extended by more than one order of magnitude by applying frequency modulation to the pump laser. Of course, the bandwidth enhancement is accompanied by a reduction of the peak gain. Consequently, a higher pump power will be necessary to achieve the same gain.

The narrow bandwidth of the FBAs can be used to advantage in coherent lightwave systems to amplify selectively the carrier of a phase- or amplitude-modulated wave, thus allowing homodyne detection, using the amplified carrier as local oscillator. The scheme should work well at bit rates $>100\ \text{Mb s}^{-1}$ because the modulation sidebands then fall outside the amplifier bandwidth, and the optical carrier can be amplified selectively. However, an inevitable accompaniment of this narrow-band amplification is the nonlinear phase shift induced by the pump on the signal, given by eqn [33]. This phase shift imposes the most stringent limit on pump- and signal-frequency stability when FBAs are used for self-homodyne detection. In the case of amplitude-shift keyed (ASK) signals, for example, a phase stability of about 0.1 rad for the amplified signal carrier may be required, which corresponds to a pump-signal frequency offset <100 kHz. The nonlinear phase shift can be used with advantage in some self-homodyne schemes which require some specific adjustment of the carrier phase. In the case of a self-homodyne coherent receiver for phase-shift keyed (PSK) signals, for example, a quadrature phase correction is required. Typically, this requires a detuning of only a few MHz from the SBS gain peak, well within the Brillouin gain bandwidth.

Another application of the narrow linewidth associated with the Brillouin gain, is as a tunable narrow-band filter for channel selection in a densely packed multichannel communication system. A channel can be selectively amplified through Brillouin amplification by launching a pump beam at the receiver end so that it propagates inside the fiber in a direction opposite to that of the multichannel signal. Different channels can be selectively amplified by tuning the pump laser. The adjustable bandwidth and high out-of-band rejection can be used to advantage in this case. In another approach, the fiber Brillouin amplifier can be used to selectively amplify multiple subcarrier-multiplexed (SCM) channels carried by a single optical carrier within a WDM signal, or to amplify a backward-propagating supervisory signal.

Fiber Parametric Amplifiers

In both stimulated Raman scattering and stimulated Brillouin scattering, the optical fiber plays an active role through the participation of molecular vibrations or acoustic phonons. However, in many other nonlinear phenomena, the fiber plays a passive role except for mediating the interaction among several optical waves. Since these phenomena involve the light-induced modulation of some medium parameter, they are known as parametric processes.

The parametric processes can be classified as second-order or third-order processes, depending on whether the second-order susceptibility or the third-order susceptibility is responsible for them. Concerning the second-order parametric processes, they are not significant in silica fibers, due to the small value of the second-order susceptibility. In contrast, some third-order parametric processes, namely third-harmonic generation, four-wave mixing, and parametric amplification, can exhibit relatively high conversion efficiencies. The parametric gain in optical fibers can be used to make fiber parametric amplifiers, which find some interesting applications for squeezing and optical fiber communications.

Phase-Matching in Parametric Processes

The efficiency of the four-wave mixing process depends on the relative phase among the interacting optical waves. A high efficiency requires matching of the frequencies as well as of the wave vectors. The latter requirement is often referred to as phase-matching. In contrast with SRS and SBS, where phase-matching is automatically fulfilled as a result of the active participation of the nonlinear medium, the phase-matching condition in parametric processes requires a specific choice of the frequencies and the refractive indices.

Considering the case in which two photons at energies ω_1 and ω_2 are annihilated with simultaneous creation of two photons at frequencies ω_3 and ω_4, the matching of the frequencies imposes the condition:

$$\omega_1 + \omega_2 = \omega_3 + \omega_4 \qquad [37]$$

On the other hand, the phase-matching is given by the condition $\Delta k = 0$, where

$$\begin{aligned}\Delta k &= k_3 + k_4 - k_1 - k_2 \\ &= (\tilde{n}_3\omega_3 + \tilde{n}_4\omega_4 - \tilde{n}_1\omega_1 - \tilde{n}_2\omega_2)/c \end{aligned} \qquad [38]$$

where the $\tilde{n}_i$ are the effective refractive indices of the fiber modes. It is relatively easy to satisfy the phase-matching condition when $\omega_1 = \omega_2$. In this case, a strong pump at $\omega_1 = \omega_2 \equiv \omega_p$ generates a

low-frequency sideband at ω_3 and a high-frequency sideband at ω_4, when we assume $\omega_4 > \omega_3$. These sidebands are referred as the Stokes and anti-Stokes bands, respectively, which are also often called the signal and idler bands. The frequency shift of the two sidebands is given by

$$\Omega_s = \omega_p - \omega_3 = \omega_4 - \omega_p \qquad [39]$$

The effective indices $\tilde{n}_i$ in eqn [38] can be written as

$$\tilde{n}_i = n_i + \Delta n_i \qquad [40]$$

where Δn_i is the change of index refraction due to waveguiding. As a consequence, considering the eqn [40] and the case $\omega_1 = \omega_2$, eqn [38] can be written in the form:

$$\Delta k = \Delta k_m + \Delta k_w \qquad [41]$$

where

$$\Delta k_m = (n_3\omega_3 + n_4\omega_4 - 2n_1\omega_1)/c \qquad [42]$$

results from the material dispersion and

$$\Delta k_w = (\Delta n_3\omega_3 + \Delta n_4\omega_4 - (\Delta n_1 + \Delta n_2)\omega_1)/c \qquad [43]$$

results from the waveguide dispersion. We must note that although the two pump waves have the same frequency, Δn_1 and Δn_2 are, in general, different if these waves propagate in different modes.

The material contribution Δk_m can be expressed in terms of the frequency shift Ω_s given by eqn [39], if we use an expansion of Δk_m about the pump frequency $\omega_p = \omega_1 = \omega_2$. By retaining up to terms quadratic in Ω_s in this expansion, one obtains

$$\Delta k_m = \beta_2\Omega_s^2 \qquad [44]$$

where $\beta_2 = d^2k/d\omega^2$ is the group-velocity dispersion (GVD) coefficient at the pump frequency. In the description of the evolution of the Stokes and anti-Stokes waves, an effective propagation constant appears, given by

$$\kappa = \Delta k + \Delta k_{NL} = \Delta k_m + \Delta k_w + \Delta k_{NL} \qquad [45]$$

In eqn [45], Δk_{NL} represents a nonlinear contribution, given by

$$\Delta k_{NL} = 2\gamma P_p \qquad [46]$$

where $P_p = P_1 = P_2$ is the power in each pump wave and

$$\gamma = \frac{\bar{n}_2\omega}{cA_e} \qquad [47]$$

is the nonlinearity coefficient, $\bar{n}_2$ being the nonlinear-index coefficient.

Parametric Gain

The evolution of the Stokes and anti-Stokes waves are governed by the following coupled equations:

$$\frac{dE_3}{dz} = 2j\gamma P_p \exp(-j\kappa z)E_4^* \qquad [48]$$

$$\frac{dE_4^*}{dz} = -2j\gamma P_p \exp(j\kappa z)E_3 \qquad [49]$$

The general solution of eqns [48] and [49] is given by

$$E_3(z) = \left(c_1\exp(g_P z) + c_2\exp(-g_P z)\right)\exp\left(-j\kappa z/2\right) \qquad [50]$$

$$E_4^*(z) = \left(c_3\exp(g_P z) + c_4\exp(-g_P z)\right)\exp\left(j\kappa z/2\right) \qquad [51]$$

where the coefficients c_i are determined from the boundary conditions and the parametric gain g_P is given by

$$g_P = \sqrt{(2\gamma P_p)^2 - \left(\frac{\kappa}{2}\right)^2} \qquad [52]$$

In the range

$$-6\gamma P_p \le \Delta k \le 2\gamma P_p \qquad [53]$$

the radicand in eqn [52] is positive, so a gain occurs. The maximum gain is achieved when $\kappa = 0$ and it is given by

$$g_{Pmax} = \Delta k_{NL} = 2\gamma P_p = \bar{g}_P\frac{2P_p}{A_e} \qquad [54]$$

where

$$\bar{g}_P = \frac{\omega_p\bar{n}_2}{c} \qquad [55]$$

Considering that $\bar{n}_2 = 3.2\times10^{-20}\ m^2\ W^{-1}$ and assuming a pump wavelength $\lambda_p = 1\ \mu m$, we find that the gain coefficient $\bar{g}_P$ is larger by a factor of 2 than the peak of the Raman gain coefficient g_R. Consequently, the threshold for the parametric process is lower than with the Raman effect. However, the Raman effect dominates in general for long fibers, since it is difficult to maintain phase matching because of fluctuations of the core diameter. One can define a coherence length, given by

$$L_{coh} = 2\pi/\kappa_{max} \qquad [56]$$

where κ_{max} is the maximum tolerable phase mismatch. Significant four-wave mixing occurs only if $L < L_{coh}$.

In the case of a monomode fiber, Δk_w is very small and can be ignored. For λ_p below the zero of dispersion $\lambda_D \approx 1.3\ \mu m$ ($\beta_2 > 0$), one has to keep both Δk_m and Δk_{NL} small by using small frequency shifts Ω_s and low pump powers. When Δk_m dominates, we have

$$L_{coh} \approx \frac{2\pi}{|\Delta k_m|} = \frac{2\pi}{\beta_2 \Omega_s^2} \qquad [57]$$

In the visible range, one has $\beta_2 \approx 50 - 60\ ps^2\ km^{-1}$, and $L_{coh} \approx 1\ km$ for $f_s = \Omega_s/2\pi \approx 50\ GHz$. For a pump wavelength above the zero dispersion ($\beta_2 < 0$), the negative value of Δk_m must be compensated by Δk_{NL}. For phase matching one has

$$\kappa = 0 \approx \Delta k_m + \Delta k_{NL} = \beta_2 \Omega_s^2 + \Delta k_{NL} \qquad [58]$$

From this, we obtain

$$\Omega_s = \sqrt{2\gamma P_p/|\beta_2|} \qquad [59]$$

which coincides with the frequency shift characteristic of the modulation instability.

Amplifier Gain and Bandwidth

Four-wave mixing can be used, for example, in a fiber-optical parametric amplifier. In fact, the main difference between the four-wave mixing experiments and the parametric-amplification experiments is whether a signal at the phase-matched frequency is co-propagated with the pump or not. In the absence of a signal, the signal and idler waves are generated from amplifications of noise.

Let us consider the situation in which a signal wave with a power P_3 is launched at $z = 0$, and simultaneously an idler wave appears, with $P_4(z = 0) = 0$. In this case, considering the solution given by eqs [50] and [51] and taking into account eqn [52], we can write the following results for the signal and idler powers:

$$P_3(z) = P_3(0)\left[1 + \left(1 + \frac{\kappa^2}{4g_P^2}\right)\sinh^2(g_P z)\right] \qquad [60]$$

$$P_4(z) = P_3(0)\left(1 + \frac{\kappa^2}{4g_P^2}\right)\sinh^2(g_P z) = P_3(z) - P_3(0) \qquad [61]$$

The single-pass gain of the parametric amplifier becomes

$$G_P = \frac{P_3(L)}{P_3(0)} = 1 + \frac{P_4(L)}{P_3(0)} = 1 + \left(1 + \frac{\kappa^2}{4g_P^2}\right)\sinh^2(g_P L)$$
$$= 1 + \frac{(2\gamma P_p)^2}{g_P^2}\sinh^2(g_P L) \approx \left(1 + \frac{\kappa^2}{4g_P^2}\right)\sinh^2(g_P L) \qquad [62]$$

The approximation holds for $g_P L \gg 1$. According to eqn [52], amplification ($g_P > 0$) occurs only in the case:

$$|\kappa| < 4\gamma P_p \qquad [63]$$

Since $|\kappa|$ must be small, this also means good phase-matching. If one has complete phase-matching ($\kappa = 0$) and still $g_P L \gg 1$,one obtains exponential behavior for the amplifier gain:

$$G_P = \sinh^2(g_P L) \approx \frac{1}{4}\exp(2g_P L) = \frac{1}{4}\exp(4\gamma P_p L) \qquad [64]$$

From Eqs [44] and [45], we have that the phase mismatch κ and the frequency shift Ω_s are related by

$$\kappa = \beta_2 \Omega_s^2 + \Delta k_w + \Delta k_{NL} \qquad [65]$$

From eqn [65] we have

$$\frac{\Delta\kappa}{\Delta\Omega_s} = 2|\beta_2|\Omega_s \qquad [66]$$

In order to estimate the amplification bandwidth $\Delta\Omega_s$, we set

$$\Delta\kappa = \kappa_{max} = 4\gamma P_p \qquad [67]$$

From Eqs [66] and [67] we obtain

$$\Delta\Omega_s = \frac{2\gamma P_p}{|\beta_2|\Omega_s} = \frac{2}{|\beta_2|\Omega_s L_{NL}} \qquad [68]$$

where $L_{NL} = (\gamma P_p)^{-1}$ is the nonlinear length. Considering typical values $|\beta_2| = 20\text{–}60\ ps^2\ km^{-1}$, $\Omega_s/2\pi = 10\text{–}100\ THz$, and $L_{NL} \sim 1\ m$, we obtain a bandwidth in the range ~10–100 GHz. This bandwidth lies between that of a fiber Raman amplifier (~5 THz) and a fiber Brillouin amplifier (~100 MHz).

For $|\kappa| > 4\gamma P_p$ the parametric gain becomes imaginary and there is no longer amplification, but rather a periodical power variation of the signal and idler waves. In particular, if $|\kappa| \gg 4\gamma P_p$, then

$g_P \approx j\kappa/2$ and the parametric amplifier gain is

$$G_P = (2\gamma P_p L)^2 \frac{\sin^2(\kappa L/2)}{(\kappa L/2)^2} \quad [69]$$

A convenient measure for the width in this case corresponds to a wavevector mismatch $\Delta\kappa = 2\pi/L$, where L is the fiber length. Using eqn [66], we then have

$$\Delta\Omega_s = \frac{\pi}{L|\beta_2|\Omega_s} \quad [70]$$

Amplifier Performance and Applications

Parametric amplification can induce considerable crosstalk in WDM communication systems. Such crosstalk affects the system performance, particularly for channel spacings less than 10 GHz. Considering a channel spacing of 5 GHz and -20 dB allowed crosstalk, calculations show that the input power per channel should be smaller than 1 mW for a 15 km-long monomode fiber, and 0.1 mW for 100 km-long fiber. However, this effect can be avoided by an appropriate frequency separation of the channels.

Besides the limitation mentioned above, parametric amplification may be useful in the field of optical fiber communication systems. In fact, parametric amplifiers have a bandwidth of ~100 GHz, which is smaller that the bandwidth of fiber Raman amplifiers (~5 THz) but still sufficiently large for many applications. On the other hand, the frequency shift in FPAs can be as large as ~100 THz, which must be compared with the pump-signal frequency shift of ~13 THz of FRAs. These characteristics of FPAs allow a considerable flexibility in the choice of the pump source.

Parametric amplifier gains above 40 dB, with a 30-m-long fiber for input signal powers less than 1 mW and a pump power $P_p \approx 70$ W, were already measured. Such large gain values show clearly the potential of fiber parametric amplifiers provided phase-matching can be achieved. In fact, this condition puts stringent limits on the control of frequency shift Ω_s between the pump and signal waves. One method to satisfy the phase-matching requirement consists in using large birefringence of polarization-preserving fibers. Several parametric amplifiers using this kind of fiber have been demonstrated. Moreover, the possibility to realize a parametric laser was also demonstrated, by placing the fiber inside a Fabry–Perot cavity.

Fiber-optic parametric amplifiers have also attracted considerable attention in the context of a phenomenon known as squeezing. In contrast to coherent light containing equal minimum noise fluctuations in cophasal and quadrature components, in squeezed states one of these components acquires a noise fluctuation below the quantum-noise level at the expense of the other component. This phenomenon can occur in a parametric amplifier for certain values of the relative phase between the signal and idler waves. In fact, spontaneous emission at the signal and idler frequencies generates photons with random phases. Four-wave mixing increases or decreases the number of specific signal-idler photon pairs, depending on their relative phases. The noise reduction below the quantum-noise level can be verified using a phase-sensitive detection scheme in which the phase of the local oscillator is conveniently adjusted. The phase-sensitive nature of parametric amplifiers can also be usefully exploited to reduce the timing jitter in soliton communication systems.

List of Units and Nomenclature

A_e	effective fiber core area, m^2
c	vacuum speed of light, m s^{-1}
E_p	amplitude of the pump wave electric field, V m^{-1}
E_s	amplitude of the signal wave electric field, V m^{-1}
g_B	Brillouin gain coefficient, m W^{-1}
G_B	Brillouin amplifier gain
g_P	parametric gain, m^{-1}
G_P	parametric amplifier gain
g_R	Raman gain coefficient, m W^{-1}
G_R	Raman amplifier gain
$\vec{k}_a$	wave vector of the acoustic wave, rad m^{-1}
$\vec{k}_p$	wave vector of the pump wave, rad m^{-1}
$\vec{k}_s$	wave vector of the signal wave, rad m^{-1}
L	amplifier length, m
L_{coh}	coherence length, m
L_e	effective amplifier length, m
L_{NL}	nonlinear length, m
n	refractive index
$\bar{n}_2$	nonlinear-index coefficient, m^2 W^{-1}
p_{12}	longitudinal elasto-optic coefficient
P_{asp}	amplified spontaneous scattered power, W
P_p	pump power, W
P_s	signal (Stokes) power, W
ν_a	acoustic velocity, m s^{-1}
ν_B	Brillouin frequency shift, s^{-1}
α_p	absorption coefficient at the pump frequency, m^{-1}

α_s	absorption coefficient at the signal frequency, m^{-1}
β_2	group-velocity dispersion coefficient, $s^2\ m^{-1}$
γ	nonlinearity coefficient, $W^{-1}\ m^{-1}$
Γ	inverse acoustic phonon lifetime, s^{-1}
δ	frequency detuning from Brillouin gain peak, s^{-1}
$\Delta\nu_B$	Brillouin gain bandwidth, s^{-1}
$\Delta\nu_p$	pump linewidth, s^{-1}
$\Delta\nu_R$	Raman gain bandwidth, s^{-1}
ε_0	free-space permittivity, $F\ m^{-1}$
ρ_0	average material density, $kg\ m^{-3}$
ϕ	nonlinear phase shift, rad
ω_a	acoustic wave frequency, rad s^{-1}
ω_p	pump wave frequency, rad s^{-1}
ω_s	signal (Stokes) wave frequency, rad s^{-1}
Ω_s	sideband frequency shift, rad s^{-1}

See also

Optical Communication Systems: Wavelength Division Multiplexing. **Nonlinear Optics, Basics:** Four-Wave Mixing. **Scattering:** Raman Scattering.

Further Reading

Agrawal GP (1992) *Fiber-Optic Communication Systems*, Ch. 8. New York: John Wiley & Sons.

Agrawal GP (1995) *Nonlinear Fiber Optics*, 2nd edn, Ch. 8–10. San Diego, CA: Academic Press.

Boyd RW (1992) *Nonlinear Optics*, Ch. 8–10. San Diego, CA: Academic Press.

Buck JA (1995) *Fundamentals of Optical Fibers*, Ch. 7 and 8. New York: John Wiley & Sons.

Cotter D (1987) Fiber nonlinearities in optical communications. *Optical Quantum Electronics* 19: 1–17.

Cotter D (1993) Nonlinearity in optical fiber communication. In: Eason RW and Miller A (eds) *Nonlinear Optics in Signal Processing*, Ch. 9. London: Chapman & Hall.

Forghieri F, Thach RW and Chraplyvy AR (1997) Fiber nonlinearities and their impact on transmission systems. In: Kaminow IP and Koch TL (eds) *Optical Fiber Telecommunications – IIIA*, Ch. 8. San Diego, CA: Academic Press.

Sauter EG (1996) *Nonlinear Optics*, Ch. 5–11. New York: John Wiley & Sons.

Schubert M and Wilhelmi B (1986) *Nonlinear Optics and Quantum Electronics*, Ch. 11 and 12. New York: John Wiley & Sons.

Shen YR (1984) *The Principles of Nonlinear Optics*, Ch. 9–11, 14, 26. New York: John Wiley & Sons.

Semiconductor Optical Amplifiers

M J Connelly, University of Limerick, Limerick, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

There has been rapid growth in the deployment and capacity of optical fiber communication networks over the past 25 years. This growth has been made possible by the development of new optoelectronic technologies that can be utilized to exploit the enormous bandwidth of optical fiber. Today, systems are operational which operate at bit rates in excess of 100 Gb/s. Optical technology is the dominant carrier of global information. It is also central to the realization of future networks that will have the capabilities demanded by society. These capabilities include virtually unlimited bandwidth to carry communication services of almost any kind, and full transparency that allows terminal upgrades in capacity and flexible routing of channels. Many of the advances in optical networks have been made possible by the optical amplifier.

Optical amplifiers can be divided into two classes: the optical fiber amplifier and semiconductor optical amplifier (SOA). The former dominates conventional system applications such as in-line amplification to compensate for optical link losses. However, due to advances in optical semiconductor fabrication techniques and device design, the SOA is showing great promise for use in evolving optical communication networks. It can be utilized as a general gain element but also has many functional applications, including optical switching and wavelength conversion. These functions, where there is no conversion of optical signals into the electrical domain, are required in transparent optical networks.

We will review SOA basics, technology (materials and structures), signal transmission performance (pattern effects, crosstalk, and ultrashort pulse amplification), and some important functional applications (optical switching and wavelength conversion).

Basic Principles

The SOA is based on similar technology as a laser diode. Optical gain is achieved by electrically pumping a suitable semiconductor material, such that a

population inversion occurs between the material conduction and valence bands. An incoming photon can then be amplified when the resulting stimulated emission exceeds losses due to stimulated absorption and internal material losses. SOAs can be designed to operate in either the 1300 nm or 1550 nm optical communication windows. The principle of operation of an SOA with low residual reflectivities ($R \approx 0$) is shown in Figure 1. An input optical power experiences a single-pass gain $G = \exp(gL)$ after traveling through the SOA active waveguide of length L. The net gain coefficient $g = \Gamma g_m - \alpha_{int}$, where Γ is the optical confinement factor (the fraction of the propagating signal power confined to the SOA waveguide), g_m the material gain, and α_{int} the optical loss coefficient. g_m is a function of the injected carrier (electron) density and wavelength. The single-mode active waveguide can support two orthogonal polarization modes: transverse electric (TE) and transverse magnetic (TM).

The amplification process also adds broadband noise, amplified spontaneous emission (ASE) to the output signal. The amplifier noise figure (NF) is a measure of the signal-to-noise ratio (SNR) degradation of the signal after amplification. It is defined as the ratio of the SNR of the input to that of the output when the input noise is shot noise limited. Expressions for the output SNR are usually calculated assuming that the SOA is followed by a narrowband optical filter and an ideal photodetector, such that the main source of detector noise is the beat noise between the signal and the ASE falling within the filter bandwidth. In this case the SOA noise figure is given by

$$\mathrm{NF} = 2n_{sp}K \qquad [1]$$

with the excess noise factor K given by

$$K = \frac{(1 + RG)(G - 1)\Gamma g_m}{(1 - R)G(\Gamma g_m - \alpha_{int})} \qquad [2]$$

where n_{sp} is the effective population inversion parameter. To achieve a low NF, the internal losses must be small and a value of n_{sp} close to one is required. A favorable value of n_{sp} is achieved by operating the SOA at high gain.

A low R is required to prevent the SOA from oscillating at high gains. The residual reflectivity is manifest as ripples in the amplifier gain spectrum and ASE spectrum.

SOA Structures

The key parameters required for a practical SOA are:

- low reflectivity ($<10^{-4}$) to ensure low gain ripple (<0.5 dB);
- low optical loss coefficient to achieve a high gain;
- high material gain to allow low drive current operation;
- low polarization sensitivity (<0.5 dB) because the polarization state of the optical signal coming from a link fiber is usually random;
- high saturation output power (P_{sat}), defined as the output power at which the gain is reduced by 3 dB; and
- low fiber-to-chip coupling losses (<3 dB per facet).

A schematic diagram of a commercial SOA chip is shown in Figure 2. The active waveguide consists of a 0.2 μm thick InGaAsP bulk active layer sandwiched between 0.1 μm thick InGaAsP separate confinement heterostructure (SCH) layers. The central section of the active waveguide is 600 μm long with a constant width of 1.4 μm. The mode-expanding active waveguide tapers are 150 μm long with a width that changes linearly from 1.4 μm to 0.4 μm at the tip. The tapers allow optical coupling to an underlying passive waveguide enabling efficient coupling to the input and output optical fibers. The structure provides a high confinement factor because of the refractive index mismatch between the layers in the gain section. The p-n junction formed by the p-InP

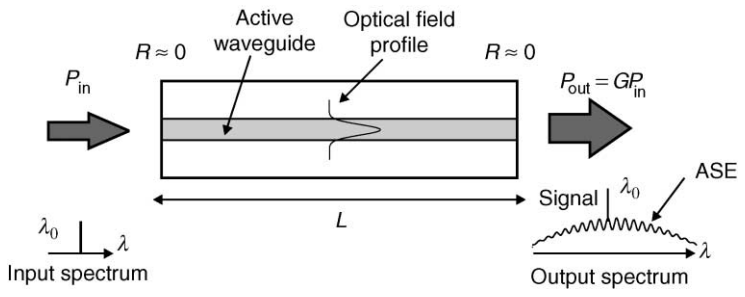
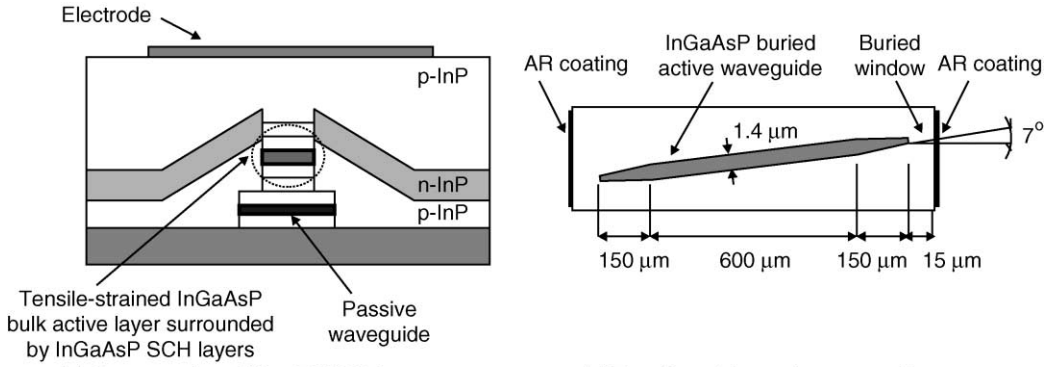


Figure 1 SOA basic structure.



(a) Cross-section of the SCH SOA structure.

(b) Top view of the active waveguide.

Figure 2 SCH SOA with tensile-strained bulk active layer.

Table 1 Typical specifications of a commercial SOA

Wavelength of peak gain	1550 nm
Peak fiber-to-fiber gain	22 dB
Noise figure	6.5 dB
Saturation output power	10 dBm
Polarization sensitivity	0.2 dB
Gain ripple	0.1 dB
3 dB bandwidth	40 nm
Drive current	200 mA

and n-InP layers acts as a current block, thereby providing good confinement of the injected carriers from the drive current to the active layer. Very low-reflectivity ($<10^{-6}$) is obtained by combining buried windows with antireflection-coated tilted facets (7° tilt angle).

Because of the active waveguide asymmetry, the TE confinement factor is larger than the TM confinement factor. g_m is isotropic in bulk material. The introduction of tensile strain, due to the lattice mismatch between the active layer and SCH layers, causes the bulk material band structure to change in such a way that the TM material gain is greater than the TE material gain. The introduction of the correct amount of tensile strain compensates for the difference in the TE and TM confinement factors leading to low polarization dependence.

Other SOA structures that can achieve low polarization dependency are based on active waveguides with near square cross-section having almost the same TE and TM confinement factors or the combination of compressively strained quantum wells (higher TE gain) and tensile strained quantum wells (higher TM gain). Typical specifications and characteristics of a commercial SOA are listed in Table 1 and shown in Figure 3.

As Figure 3 shows, gain saturation in conventional SOAs is manifest at output signal powers well below P_{sat}. Because the gain recovery time (carrier lifetime) in SOAs is fast (typically 0.1–1 ns), this can lead to pattern effects in single-channel systems and severe crosstalk in wavelength division multiplexed (WDM) systems. This problem can be greatly reduced by the use of a gain-clamped SOA (GC-SOA). In a GC-SOA, lasing action is produced, at a wavelength remote from the operating wavelength range, by the introduction of wavelength specific feedback. Once lasing begins, the carrier density in the SOA active layer is clamped at a fixed value. Changes in the input signal power lead to opposing changes in the lasing mode power. This has the effect of keeping the carrier density fixed (i.e., clamped), thereby making the signal gain relatively insensitive to changes in the total input power. A common method of providing this feedback is the use of distributed Bragg reflectors (DBRs) as shown in Figure 4. A typical GC-SOA gain versus output power characteristic is shown in Figure 5.

Basic Applications of SOAs in Optical Communication Systems

The three basic applications of SOAs in optical communication systems are: booster amplifier, in-line amplifier, and preamplifier. The main requirements of SOAs for such applications are listed in Table 2.

A booster amplifier is used to increase the power of a high power signal. Boosting laser power in an optical transmitter can be used to overcome external modulator losses, compensate for splitting and tap losses in optical distribution networks and to increase the power budget of optical links. The most critical requirement of a booster amplifier is a high P_{sat}, to obtain a high output signal power and minimize pattern effects. SOAs with P_{sat} in excess of 10 dBm are now available commercially.

The sensitivity of a conventional optical receiver is limited by thermal noise. The sensitivity is the

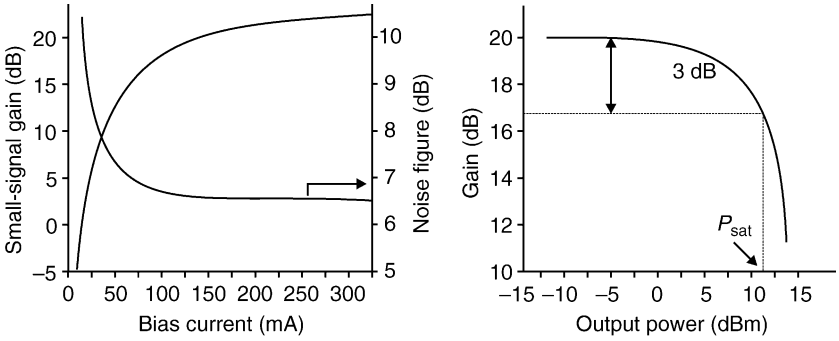


Figure 3 Typical commercial SOA small-signal gain and noise figure versus bias current and gain versus output power characteristics.

minimum signal power required at the receiver input to achieve a desired bit-error-rate (typically 10^{-9}). An optical preamplifier can be used to increase the power level of an optical data signal prior to detection and demodulation, leading to an increase in sensitivity. The performance of a preamplified optical receiver is dependent on the detector signal-to-noise ratio, $\text{SNR} = i_{\text{sig}}^2 / \bar{i}_{\text{noise}}^2$. The signal photocurrent i_{sig} is proportional to the amplified signal power. The mean square noise current $\bar{i}_{\text{noise}}^2$ includes circuit noise (thermal noise and detector dark current noise), signal shot noise, spontaneous emission shot noise, signal-spontaneous beat noise, and beat noise between the spectral components of the spontaneous emission. The best improvement in SNR occurs when the SOA is operated in the signal-spontaneous beat noise limit. In this regime, the signal power is sufficiently large such that the dominant receiver noise is the signal-spontaneous beat noise that falls within the signal bandwidth. This usually requires that the spontaneous emission from the SOA be reduced using a narrowband optical filter. The noise figure is a critical parameter in this application and should be as low as possible.

In long-haul optical transmission systems, in-line optical amplifiers can be used to compensate for link losses thereby increasing the spacing between optical regenerators. The main advantages of in-line SOAs are transparency to data rate and modulation format (unsaturated operation), bidirectionality, WDM capability, simple mode of operation, low power consumption, and compactness. The latter two advantages are important for remotely located optical components. To avoid noise accumulation in the link, it may be necessary to follow each SOA by a narrowband optical filter, but this can prevent capacity enhancements using WDM.

An example of a single channel optical transmission system utilizing booster, in-line, and

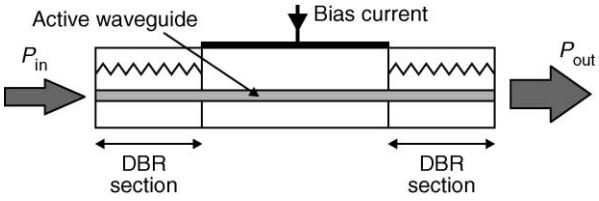


Figure 4 Schematic structure of a GC-SOA.

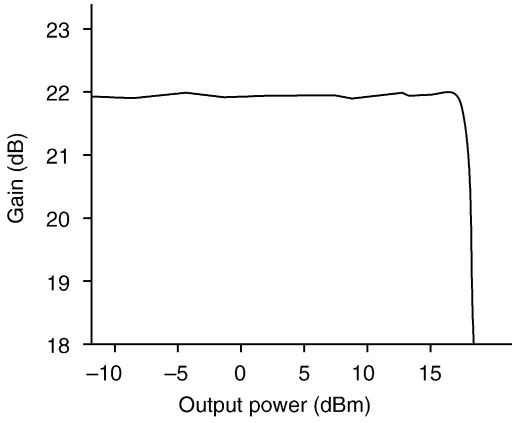


Figure 5 Typical GC-SOA gain versus output power characteristic.

preamplifier SOAs is shown in Figure 6. The transmitter comprised a 1309 nm gain-switched laser diode directly modulated with a 10 GHz sinusoid to produce a train of 40 ps wide pulses at a repetition rate of 10 GHz. At 1309 nm fiber dispersion is small and the maximum transmission distance is mainly limited by the link losses. The laser output was connected to an external modulator, driven by a 2^{31}-1 pseudo-random bit sequence (PRBS) to produce an optical data stream with an extinction ratio of 13 dB. A booster SOA was used to increase

Table 2 Requirements of basic SOA applications

Requirement	*Booster amplifier*	*In-line amplifier*	*Preamplifier*
High gain	Yes	Yes	Yes
High P_{sat}	Yes	Yes	Not critical
Low noise figure	Not critical	Yes	Yes
Low polarization sensitivity	Not critical	Yes	Yes
Optical filter	Not necessary	Not critical	Yes

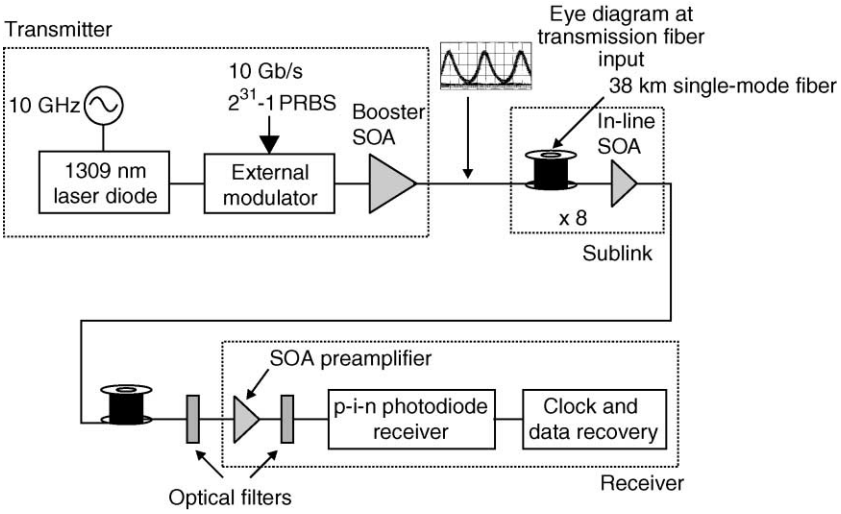


Figure 6 10 Gbit/s transmission experiment utilizing booster, in-line and preamplifier SOAs. (Adapted from Kuindersma *et al.* (1996) 10 Gbit/s RZ transmission at 1309 nm over 420 km using a chain of multiple quantum-well semiconductor optical amplifier modules at 38 km intervals. European Conference on Optical Communications 1996.)

the average transmitted power to between 0 and 2 dBm (7–9 dBm peak power). The transmission fiber length was 420 km with 12 in-line SOAs, used to compensate for fiber loss, spaced at 38 km intervals. At the receiver the signal was passed through a 1 nm bandpass filter, to reduce the accumulated spontaneous emission, amplified by an SOA preamplifier and filtered by a 1 nm bandpass filter. The signal was then detected by a p-i-n photodiode followed by an electronic clock and data recovery circuit. The SOA preamplifier and filter increased the receiver sensitivity from −14 dBm to −31 dBm for a BER of 10^{-10}. The receiver penalty after 420 km was 5 dB. In this experiment, the main limitation on transmission distance was the accumulation of spontaneous emission within the optical filter bandwidth.

Pattern Effects and Crosstalk

When an SOA is operated in the unsaturated regime, the amplifier gain is independent of the number of input signals and signal data rate. Outside this regime, the SOA will cause distortion because at high input powers, the gain saturates and compresses. Dynamic gain saturation occurs on the same time scale as the gain recovery time. This leads to pattern effects where the power of an incoming bit affects the gain experienced by subsequent bits. This is particularly important when the bit rate is of the same order as the inverse of the gain recovery time as shown in Figure 7. In WDM systems cross-gain modulation (XGM) between the channels can lead to severe interchannel crosstalk.

A further complication in WDM systems using SOAs is interchannel crosstalk caused by four-wave mixing (FWM). FWM is a coherent nonlinear process between two optical fields within the SOA, resulting in gain modulation at the beat frequency between the fields, and in the process generating new sidebands. In WDM systems with equally spaced wavelengths, the net effect of FWM is the generation of crosstalk terms interfering with the desired signals. Although the crosstalk can be relatively low in power terms, it can produce significant power penalties due to the coherent beat noise phenomena. The level of FWM crosstalk increases as the channel spacing decreases and the channel output power increases.

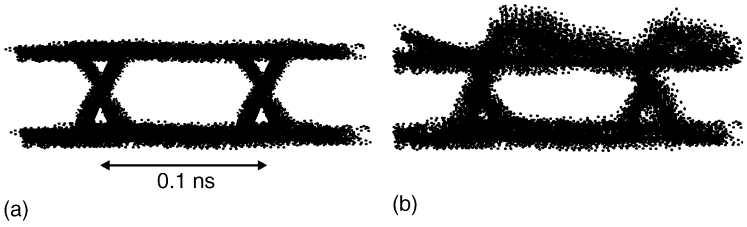


Figure 7 Typical eye diagrams of 10 Gb/s nonreturn to zero data (a) before and (b) after amplification by an SOA. ASE noise is present in the output eye diagram.

FWM crosstalk is of particular importance in dense WDM systems where the inter channel spacing is less than 100 GHz. Figure 8 shows a typical SOA output spectrum for an 8-channel multiplex with FWM-generated crosstalk signals.

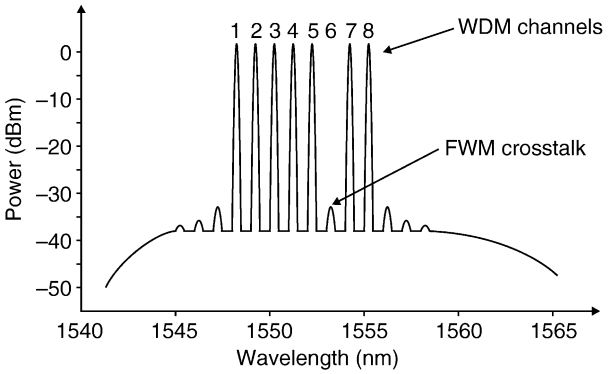


Figure 8 Typical SOA output spectrum for an 8-channel WDM multiplex with 100 GHz spacing. Channel 6 has been dropped to show the FWM crosstalk signal caused by channels 5 and 7.

Ultrashort Pulse Amplification

Optical time division multiplexing is an efficient method for increasing the bit rate of optical transmission systems. This requires the time interleaving of short optical pulses. The transmission distance is limited by the fiber group velocity dispersion, which is proportional to the pulse spectral width. Soliton transmission is also possible where the pulse shape is preserved as it propagates in the fiber.

Because an SOA has a very large bandwidth (typically 5 THz) it is capable of amplifying pulses as short as 100 fs. An input pulse can be amplified without significant distortion if the pulse energy is much less than the saturation energy E_{sat} of the SOA. Typical SOA saturation energies are of the order of a few pico Joules. As the pulse energy approaches E_{sat}, considerable spectral broadening and distortion can result. For an input pulse width τ_p (full width at half maximum) of the order of 10–100 ps spectral broadening is primarily due to self phase modulation (SPM). SPM is caused by gain saturation, which leads to intensity-dependent changes in the SOA active layer refractive index in response to carrier density variations. The degree of spectral broadening and distortion also depends on the input pulse shape. If τ_p is much less than the carrier lifetime, the output pulse power $P_{out}(t)$ and phase $\phi_{out}(t)$ are given approximately by

$$P_{out}(t) = P_{in}(t)\exp[h(t)] \qquad [3]$$

$$\phi_{out}(t) = \phi_{in}(t) - \frac{1}{2}\alpha h(t) \qquad [4]$$

with

$$h(t) = \left[1 - \left(1 - \frac{1}{G_0}\right)\exp\left(\frac{-U_{in}(t)}{E_{sat}}\right)\right]^{-1} \qquad [5]$$

$$U_{in}(t) = \int_{-\infty}^{t} P_{in}(\tau)\mathrm{d}\tau \qquad [6]$$

where $P_{in}(t)$ and $\phi_{in}(t)$ are the input pulse power and phase, respectively. G_0 and α are the amplifier unsaturated gain and linewidth enhancement factor, respectively. Typical values for α are in the range 2–10 and depend on the amplifier active region material and operating conditions. The output pulse spectrum can be obtained from

$$S(\omega) = \left|\int_{-\infty}^{\infty} [P_{out}(t)]^{1/2}\exp[i\phi_{out}(t) + i(\omega - \omega_0)t]\mathrm{d}t\right|^2 \qquad [7]$$

where $\nu_0 = \omega_0/2\pi$ is the pulse optical frequency. The output pulse chirp (frequency variation) is given by

$$\Delta\nu_{out}(t) = -\frac{1}{2\pi}\frac{\partial\phi_{out}}{\partial t} = \Delta\nu_{in}(t) + \frac{\alpha}{4\pi}\frac{\partial h}{\partial t} \qquad [8]$$

where $\Delta\nu_{in}(t)$ is the input pulse chirp. Using the above theory the shape, chirp and spectrum of an amplified zero-chirp (transform limited) Gaussian

pulse of input power

$$P_{\text{in}}(t)=\frac{E_{\text{in}}}{\tau_0\sqrt{\pi}}\exp\left(-\frac{t^2}{\tau_0^2}\right)$$

is shown in Figure 9. E_{in} is the pulse energy and $\tau_{\text{p}}\approx 1.665\tau_0$. The amplified pulse is asymmetric because the leading edge of the pulse experiences a larger gain than the trailing edge. The amplified pulse spectrum is broader than the input pulse and also develops a multipeak structure. This is due to the SPM-induced frequency chirp imposed on the pulse as it propagates through the amplifier. In this case the chirp is linear around the peak of the pulse power spectrum. It can be compensated by transmitting the pulse through an optical fiber with anomalous group-velocity dispersion. In practice optical pulses can be far from Gaussian and can also have an initial chirp. In this case the induced chirp and resulting pulse spectrum can be more complex and more difficult to compensate.

When the input pulse width is less than ~10 ps, the above theory is no longer adequate and other nonlinear effects within the SOA such as carrier heating and spectral hole burning must be taken into account. The resulting pulse power and spectral distortions induced by an SOA on such pulses can be very complex.

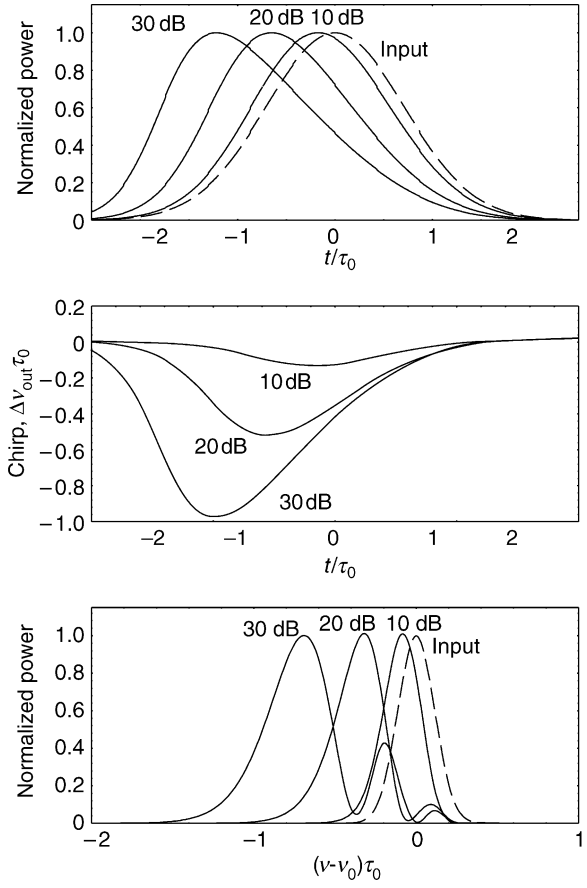


Figure 9 Amplified pulse shape, frequency chirp and spectrum for a transform-limited Gaussian input pulse with $E_{\text{in}}/E_{\text{sat}} = 0.1$ and $\alpha = 5$. ν is optical frequency. The parameter is the amplifier unsaturated gain.

Functional Applications

SOAs can be used to perform functions that are particularly useful in optically transparent networks. Recent advances in photonic integrated circuit and optoelectronic device packaging technology have made the deployment of SOA functional elements feasible. Two of the most important applications of SOAs are optical switching and wavelength conversion.

An example of a 2 × 2 SOA switch module used for optical routing is shown in Figure 10. Larger switch matrices can be constructed using this basic element. The module is based on a hybrid structure consisting of four integrated SOAs mounted on a silicon submount. The SOA-array is aligned to input and output polymer waveguides through V-grooves and alignment indentations. An incoming data packet can be routed to any output port by switching on the appropriate SOA. Switching times of the order of 1 ns are possible.

Ultrafast switching (<100 fs) can be achieved by incorporating SOAs in non-linear loop mirror structures, such as the terahertz optical asymmetric demultiplexer (TOAD) shown in Figure 11. Switching is achieved by placing an SOA offset from the center of an optical fiber loop mirror and injecting data into the loop via a 50:50 coupler. The two

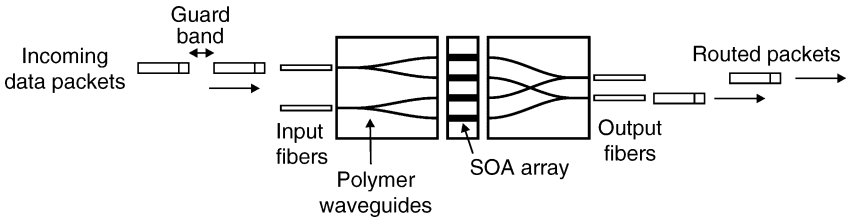


Figure 10 2 × 2 hybrid SOA switch module.

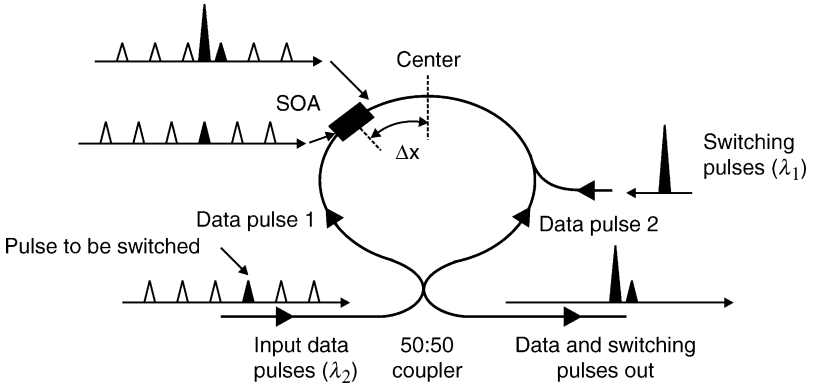


Figure 11 Ultrahigh speed optical switching using a TOAD.

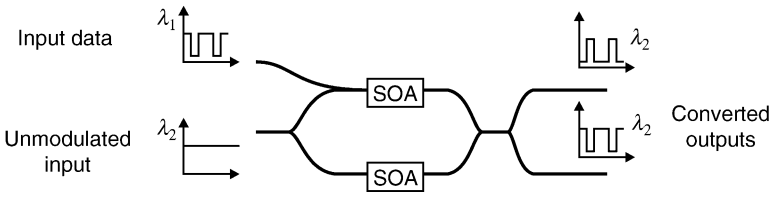


Figure 12 SOA Mach–Zehnder wavelength converter.

counter-propagating data pulse streams arrive asynchronously at the SOA. A switching pulse is timed to arrive after one data pulse but just before its replica. The switching pulse power is adjusted to impart a phase change of π radians onto the replica, so the data pulse is switched out when the two counter-propagating components interfere on their return to the coupler. The TOAD can also be used to demultiplex high-speed time division multiplexed pulse streams.

All-optical wavelength converters will play an important role in broadband optical networks. Their most important function will be to avoid wavelength blocking in optical cross-connects in WDM networks. Wavelength converters increase the flexibility and capacity of a network using a fixed set of wavelengths and can be used to centralize network management. In packet switching networks, tunable wavelength converters can be used to resolve packet contention and reduce optical buffering requirements.

Wavelength conversion in SOAs can be achieved using XGM, cross-phase modulation (XPM), or FWM. The design of an SOA wavelength converter based on XPM requires that one or more SOAs be incorporated into an interferometer, an example of which is the Mach–Zehnder interferometer (MZI) shown in Figure 12. An input data signal at wavelength λ_1 is used to modulate the upper SOA refractive index, which controls the phase shift experienced by a second unmodulated input signal at λ_2 in one of the interferometer arms. When the input data signal is low (logic 0), the arms are in phase and the λ_2 signal appears at the top output. When the input pump signal is high (logic 1) it produces an extra π phase shift between the two arms, causing them to be totally out of phase and the λ_2 signal to appear at the bottom output. Both outputs contain a copy of the original data signal imposed onto the new wavelength. Up and down conversion is possible. The lower SOA is used to equalize the gain experienced by the λ_2 signal in each of the arms. An optical filter is required at the output to remove λ_1. Such wavelength converters can operate at very high bit rates (>10 Gb/s). An important advantage of the MZI structure is that it also gives 2R regeneration (re-amplification and reshaping) of the input data signal with wavelength conversion. This is because the interferometer non-linear response increases the converted signal extinction ratio compared to the input data. All-optical 3R (2R + retiming) regenerators are of great interest in future optical communication networks to restore degraded transmission signals. SOA-based interferometric structures can be used to achieve 3R (2R + retiming) regeneration at data rates >80 Gb/s.

See also

Coherent Lightwave Systems. **Lasers:** Semiconductor Lasers. **Optical Amplifiers:** Basic Concepts; Erbrium Doped Fiber Amplifiers for Lightwave Systems; Optical

Amplifiers in Long-Haul Transmission Systems. **Optical Communication Systems:** Optical Time Division Multiplexing; Wavelength Division Multiplexing. **Semiconductor Physics:** Band Structure and Optical Properties.

Further Reading

Agrawal GP and Olsson NA (1989) Self phase modulation and spectral broadening of optical pulses in semiconductor laser amplifiers. *IEEE Journal of Quantum Electronics* 25: 2297–2306.

Connelly MJ (2002) *Semiconductor Optical Amplifiers*. Boston, MA: Kluwer Academic.

Desurvire E (1994) *Erbium-Doped Fiber Amplifiers: Principles and Applications*. New York: Wiley.

Durhuus T, Mikkelsen B, Joergensen C, Lykke Danielsen S and Stubkjaer KE (1996) All-optical wavelength conversion by semiconductor optical amplifiers. *Journal of Lightwave Technology* 14: 942–954.

Ghafouri-Shiraz H (1995) *Fundamentals of Laser Diode Amplifiers*. New York: Wiley.

Shimada S and Ishio H (1992) *Optical Amplifiers and their Applications*. New York: Wiley.

OPTICAL COATINGS

Contents

Anti-Counterfeiting and Decorative Coatings
Diamond Optical Devices and Coatings
Laser Damage in Thin Film Coatings
Optical Black Surfaces
Thin Film Optical Coatings
X-Ray Coatings

Anti-Counterfeiting and Decorative Coatings

R W Phillips and R L Bonkowski, Flex Products, Inc., Santa Rosa, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Scientists are well known for the diversity of their interests. One topic that captures the interest of virtually all scientists, and people in general, is money. Not only are people engaged in obtaining money, but they are also fascinated about the manufacture of money and the protection of its value.

The relationship between manufacturing techniques and the protection of value documents (currency, stock shares, bonds, etc.) is not accidental. Technology has always been employed to counter the contemporary threats of counterfeiters. These threats vary, depending on the skill set and resources available to the criminal, as well as the counterfeiting strategy adopted. While we refer to banknotes repeatedly in this chapter, the information applies, as well, to passports, stamps, tax banderoles, identification cards, visas, licenses, checks, and bankcards. Banknotes can be considered the most challenging application for optical anticounterfeiting technologies. Stamps, banderoles, and checks are single-use items, and do not have to be as durable as currency. Identification cards, passports, and the like are used many times but are generally in protective covers and are examined by trained personnel. Currency must endure the rigors of commerce and travel, while still being easily authenticated by the average citizen.

The ever-increasing availability of low-cost computer hardware and software, along with color photocopiers and printers, has stimulated the use of optical and electro-optical technologies that were previously applied in a limited fashion or not at all. These optical techniques include: transmitted light devices, such as watermarks and threads; luminescent materials; reflective, interference, and diffraction schemes; modulation techniques (Moiré, for example); and light polarization effects.

Each of these concepts either has been or is likely to be applied to value documents to prevent counterfeiting. Each has its own set of potential benefits and

application challenges. Also, some of these effects have been successfully applied to commercial decorative products. It is important to understand the difference between commercial and security uses. Effects aimed at commercial uses are mainly decorative or aesthetic, whereas security devices are functional and integrated into a security strategy. To ensure economic success, decorative products must be easily manufactured and widely distributed. In contrast, security products are carefully controlled in both their availability and the diffusion of manufacturing expertise.

Historical Perspective

Paper currencies (banknotes) have been printed for hundreds of years. Their history has been marked by a succession of technical changes in an effort to prevent counterfeiting. Sophisticated manufacturing techniques, such as intaglio printing and watermarks, along with special papers, made it difficult for counterfeiters to produce acceptable imitations. Intaglio is the centuries old method of printing, using incised or engraved metal plates. The tremendous pressure of the intaglio press forces the paper into an ink-coated plate engraving, forming patterns of ink on the paper.

In the eighteenth and nineteenth centuries, banknote printers often used unique engravings of guilloche patterns (intricate, overlapping curves), vignettes, and portraits as anticounterfeiting devices. These sophisticated patterns effectively discouraged early, relatively unskilled, counterfeiters. The development of photography and lithographic printing in the late nineteenth century created new opportunities for counterfeiting. Photochemical technology made it possible to produce suitable reproductions using print shop facilities throughout the world.

Photochemical techniques had significant limitations, however. The printed notes were passable but not high-quality imitations of genuine intaglio notes. Making lithographic plates required specific skill, considerable equipment, and left behind meaningful evidence (films, plates, press blankets, etc.) for police to find.

Until the 1980s, counterfeiting detection methods by governmental officials were relatively uncomplicated, as three general perpetrators could be classified in fairly predictable ways (Table 1).

Until the advent of home computers, optical scanners, color printers, and copiers, the challenge for law enforcement was difficult but focused. Amateur counterfeiting was virtually nonexistent. The effort required to make a few notes was daunting, the risk high, and potential profit meager. This meant that anticounterfeiting police efforts could be concentrated on criminal counterfeiters. Government-sponsored counterfeiting was certainly known (both the Germans and the British attempted to de-stabilize their opponents' national economies during World War II), but this was more an issue of foreign policy, diplomacy, and espionage and not the main province of national police.

The situation changed quickly and radically in the late 1980s. Color photocopiers had improved in cost, quality, and size to the point that commercial establishments were buying hundreds of machines. The 1990s introduced numerous new technologies to seduce the casual counterfeiter. These included home computers, graphics editing software, color printers, and low-cost scanners.

The amateur counterfeiter now had an array of sophisticated tools that could copy, edit, and reproduce virtually any kind of document. Banknotes made by the techniques of 'digifeiting' (a term coined by Dr. Sara Church of the US Treasury Department) were not of high quality, but passable in small quantities in low risk situations. It was now possible, economical, and relatively safe to make one or a dozen bills, especially in low denominations.

Police efforts were now severely challenged as the proliferation of amateur counterfeiters made investigation and prosecution costly. Especially in the United States, average citizens were lured by the prospect of augmenting their incomes by printing one or two notes a week. When apprehended, effective prosecution and punishment of such small-stakes criminals was expensive and unpopular.

Large-production criminal counterfeiting was augmented by new digital and electro-optical

Table 1

Class of threat	*Typical volume of notes*	*Resources and skills (1980)*
Amateur	Small (<100)	Retail paper, ink. Household chemistry
Criminal	Medium to large (1000 to 1 000 000)	Litho press, darkroom, plate making, cutting equipment. Skilled printing personnel
State-sponsored	Large (millions)	Intaglio press, currency paper. Skilled engravers and printers, Major distribution networks

breakthroughs. Lithographic plates could now be made without cameras or chemistry. Laser ablated plates made directly on the printing press (such as Heidelberg/Presstek) accurately duplicated features as small as 10 or 15 microns. Further, computer-to-plate technology with erasable plates presented an unusual opportunity for counterfeit production in small press shops.

The use of new electronic and optical tools by amateur and criminal counterfeiters coincided with a rise in state-sponsored counterfeiting. The United States, in particular, was victimized in the early 1990s by high-quality, intaglio-printed notes. One such note, dubbed the 'super $100' was allegedly produced in huge volumes in the Middle East and circulated widely in Europe.

Various means of detecting counterfeits were promoted but remain largely ineffective today. The so-called 'detector pen' applies a small amount of iodine solution onto a banknote. The iodine reacts with starch (often used to stiffen paper to give it a banknote 'feel'), turning the pen mark from light brown to black. Unfortunately, the presence or absence of starch in the paper is not a foolproof criterion. US currency paper contains no starch, while the currencies of other countries, such as Canada, have a starch component.

Concurrent with the availability of new high-tech counterfeiting tools, optical science has been hard at work to create protection schemes, utilizing the latest optical effects and manufacturing technologies. These new technologies have been largely outside the skill set and resources of amateur, criminal, and even state-sponsored counterfeiters.

An analysis of all the types of optical materials used for anticounterfeiting purposes as well as for decorative purposes shows that all are based on the control of electromagnetic radiation in the ultraviolet (UV), visible, and infrared. Exploiting various degrees of light transmission, reflectance, absorption, or scatter controls this radiation. In particular, all anticounterfeit devices fall into one or more of five basic categories: light patterning, light re-emission, light diffraction, polarization, or light interference. Characteristics of an effective anticounterfeiting solution include:

- quick verification by the average person;
- ability to function in a variety of lighting situations;
- deterrence of scanning and electronic photocopying;
- compatibility with existing banknote production schemes and equipment;
- ability to be verified by machines;
- durable and reliable despite strenuous handling conditions;
- tight control of materials and manufacturing processes.

Devices Based on Light Patterning

Watermarks

Images can be created in the paper substrate during the manufacturing process. Watermarks have been in use for at least 500 years and are produced by the arrangement of the paper fibers in the cylinder-mold paper machine or by impression in wet paper in the Fourdrinier paper machine. Because they are not on the surface of a document, watermarks are not readily visible in reflected light, but become apparent when light is transmitted through the paper substrate. Credible counterfeits, however, have been produced using images created by imprinting paper with waxy or oily substances. The fatty 'ink' is not readily detected on the surface of the document and serves to change the light transmission characteristics of the paper.

Watermarks serve chiefly to authenticate the paper substrate and to deter the practice of 'raising' notes, that is, bleaching a genuine low-denomination note and re-printing it as a higher value. The most effective watermarks are produced by the cylinder-mold process and use an image that is a duplicate of that image printed in ink (Figure 1) elsewhere on the document. This gives the examiner a reference point to compare the watermark as, for example, on all high-denomination US currency.

Watermarks are an effective counterfeiting deterrent when the general population is educated about their presence and their function. The limited opportunity to expose banknotes to a source of transmitted light (outdoors or in a dimly lit restaurant, for example) somewhat reduces their effectiveness.



Figure 1 Watermark portrait of Clara Schumann, famous pianist and wife of composer Schumann: far left – portrait in transmitted light; center – portrait in reflected and transmitted light; and right – banknote in reflection. Courtesy of Papierfabrik Louisenthal.

Threads

Thin strips of polymer or metallized polymer can be imbedded in the paper substrate during the paper-making process. Such strips are approximately 12 microns thick and 1–5 millimeters wide. These strips are known as threads because Crane & Company developed the original concept in the 1860s, using colored silk thread. Typically, threads are installed border-to-border in the narrow direction of the banknote. Because they are inside the paper, they are not seen in reflected light, but are readily visible when light is transmitted through the substrate. Such threads are effective deterrents against counterfeiting achieved by scanning or photocopying, as the scanner uses only reflected light and cannot duplicate the buried thread. Threads have evolved into several varieties:



Figure 2 Windowed thread. Reproduced with permission from De La Rue International Limited.



Figure 3 Fluorescent thread embedded into US $100 note. Reproduced with permission from Crane and Co., Inc.

- De-metallized threads: Thin polymer coated with an opaque metal. The metal is selectively removed to create text or images.
- 'Window' threads: Metallized polymer that is 'woven' in and out of the paper to permit observation in both reflected and transmitted light (Figure 2).
- Fluorescent threads: Polymer impregnated with ultraviolet inks that emit a specific color under UV irradiation (Figure 3). See the section 'Devices Based on Light Re-emission' later in this article.
- Diffractive threads: Metallized polymer strips that have been embossed with diffractive or holographic patterns and attached by hot-stamping or adhesive to the surface of the note. See the section 'Devices Based on Light Diffraction.'

Potential drawbacks to the use of threads are durability and the temptation of the general public to try to extract them from the banknote. Also, some threads can be simulated by printing or by attaching a polymer film to the surface of the banknote.

Laser Perforation

A modern 'watermark' concept is laser perforation of the banknote substrate. Precise patterns of microscopic holes (ca. 100 microns diameter) can be drilled using a high-power laser apparatus. The 200-Franc Swiss banknote features denomination numbers that are not readily visible in reflected light. However, when a light source is held behind the note, the numerals fabricated by laser perforation are clearly visible (Figure 4). In addition to being easily authenticated by non-expert observers with no special viewing apparatus, laser perforation has forensic advantages. The penetrations made by laser beams have a particular shape, spacing, and dimension, which are not easily mimicked. By clever adaptation of laser scanning software, holes of different shapes can be made and inter-mixed for additional security complexity. As a result, manual

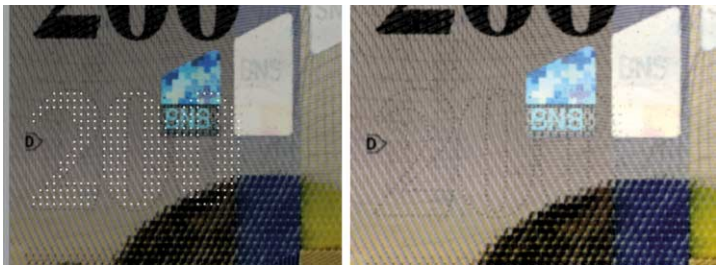


Figure 4 Microperf® laser holes in banknote. Left image in transmitted light; right image in reflected light. Reproduced with permission from Orell Füssli Security Printing Ltd.

pin pricks and mechanically pierced holes can be discriminated as forgeries.

Reflective Foils

One of the earliest defenses against photocopying and digital scanning was the use of reflective metallic foils affixed to the surface of the document. Such foils yield >80% reflection during scanning and the resultant scanned image is usually a black field.

Attachment of foils to currency paper or documents is carried out by applying a thermoset adhesive under heat and pressure and releasing the optical film from its carrier web by means of a separate polymeric release layer. This process is known as hot-stamping. Alternatively, an ultraviolet curing adhesive can be used to affix the foil to the paper substrate using a cold roll-on process.

Foiling may impede the amateur counterfeiter; since decorative foiling of stationery and packaging is well known, it presents little challenge to the professional perpetrator.

Planchettes

Planchettes are small flakes or fragments a few millimeters in diameter and 10–25 microns thick. These are often colorful or reflective materials and can feature a variety of effects including iridescence, diffraction, fluorescence, magnetic properties, or thermochromic properties. These fragments are attached to the banknote substrate by an adhesive (Figure 5). Planchettes discourage scanning and photocopying, since the copier cannot duplicate their special properties. Unfortunately, planchettes have been found to be non-durable and easy to simulate. The generally random placement makes it difficult for an examiner to authenticate a genuine note.

Thermochromic Inks

Inks that react to temperature changes have been employed on checks and identity cards. Thermochromic inks fluctuate from one color to another when the temperature of the printed image is raised. While it is possible for body temperature to activate this change, the effect is more certain if infrared heat is used. Since the process requires several seconds to activate (depending on the heat source), thermochromic materials have not been widely used on banknotes where immediate authentication is necessary.



Figure 5 Peruvian 100 Nuevo Soles note showing iridescent planchettes. Reproduced with permission from Flex Products, Inc.

Colored Fibers

Extremely fine fibers (ca. 15 microns diameter) have been incorporated into currency papers for decades. The United States, for example, uses a few hundred grams of red and blue fibers per ton of paper in making currency. While of some forensic interest, the amount and size of the fibers makes them nearly undetectable by the average citizen.

Alternatively, fibers may be impregnated with fluorescent materials, or with absorbers that selectively absorb the wavelengths of light typically used in copy machines, producing obvious forgeries when copied.

Moiré Effects

Images can be created with very fine details that produce aliasing errors when digitized. Aliasing is the creation of new and spurious patterns through the superposition of two patterns whose periodicities are unequal. It represents interaction between the spatial frequency of the genuine document and the sampling frequency of the copying device. Aliasing is commonly known as moiré in the printing industry. Typical moiré generating patterns are constructed as closely spaced concentric circles, often found around the eyes in portraits on banknotes. When scanned digitally, the resulting copy will contain spurious images (scan traps) that are easily identified as fraudulent.

For example, if a genuine document is designed in a 1000 dpi (dots per inch) printing process it can later be scanned on a 1000 dpi scanner and, theoretically, all the data can be reproduced accurately. If, however, the scanner and the document are not in synchrony, the scanner will miss pieces of data, depending on the registration error, resulting in a digitized image that does not resemble the original.

In the scanned section of a US $5 note (Figure 6), one can see patterns of curved lines intersecting in the background, whereas the authentic image consists of fine parallel lines that do not intersect. The curvature and intersections are artifacts of the moiré created by digital scanning.

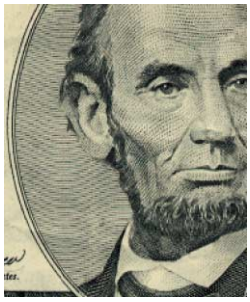


Figure 6 Scanned section of US $5 note showing moiré effect. Authentic image has no intersecting lines. Courtesy of Flex Products, Inc.

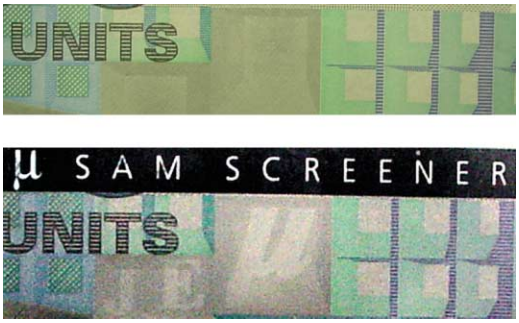


Figure 7 Top image – hidden image as normally seen; Bottom image – Image decoded with visible line screen. Courtesy of Joh Enschedé Security Printing Ltd.



Figure 8 Hidden image in 32¢ US postage stamp decoded by invisible line screen (left image – normal view; right image – with decoder in place). Courtesy of Flex Products, Inc.

Modulation Techniques

Images produced by offset lithography are based on patterns (screens) of fine dots. By varying dot density, an array of shadings and tones can be created. More progressive modulation techniques, such as amplitude modulation (varying dot sizes) and frequency modulation (varying dot spacing) are often used to produce high-resolution images. To deter counterfeiting, banknotes can be printed using pixels, which are lines of equal length but assorted angular orientation. This technique is known as screen angle modulation (SAM).

SAM techniques can be used to defeat digital scanners by preventing the scanning of all available information due to under-sampling, or by concealing hidden patterns which are not demodulated by the human eye, but can be detected with suitable filters (Figure 7). A variant of this technology, called Scrambled Indicia®, has been used to protect US postage stamps (Figure 8).

Devices Based on Light Re-emission

Ultraviolet Fluorescent Inks

Ultraviolet (UV) inks are widely used on banknotes as semicovert deterrents. Images printed in UV inks are challenging to detect in normal lighting situations but are readily visible under UV (black light) illumination. Such printed images require a special light source for authentication and are subject to degradation over time from the handling and chemical exposure of banknotes. Typically, these products are sensitive to either short UV at 254 nanometers or long UV at 366 nanometers.

UV inks provide good protection against photocopying but can be readily simulated by more advanced counterfeiters.

UV Fibers

Several countries mix minute quantities of UV luminescent fibers into the paper substrate. In most cases, these are randomly placed and invisible to the naked eye.

Devices Based on Light Diffraction

Diffractive structures range from simple periodic surface relief structures (textured surfaces) to multilevel holographic structures that form a continuum of products ranging from decorative coatings to anti-counterfeit devices. Figure 9 shows a generalized structure for all known types of reflective-based relief structures.

Most devices are based on reflection; however, in a few instances where clear polymer is used, devices may also be based on diffractive light transmission. The sizes, shapes, and orientation of the grooves vary according to the exact device. The opaque aluminum layer insures that the design works in reflected light. Most of these structures exhibit a rainbow effect where the color moves from short to long wavelength as the viewing angle increases. This rainbow effect can be found for all types of structures with grooves that have sizes near that of light wavelengths. The color shifts in a direction opposite to that of a thin film interference structure. Further, the rainbow effect can only be seen in a direction perpendicular to the groove direction. In the special case where the groove dimensions are less than the wavelength of light, the

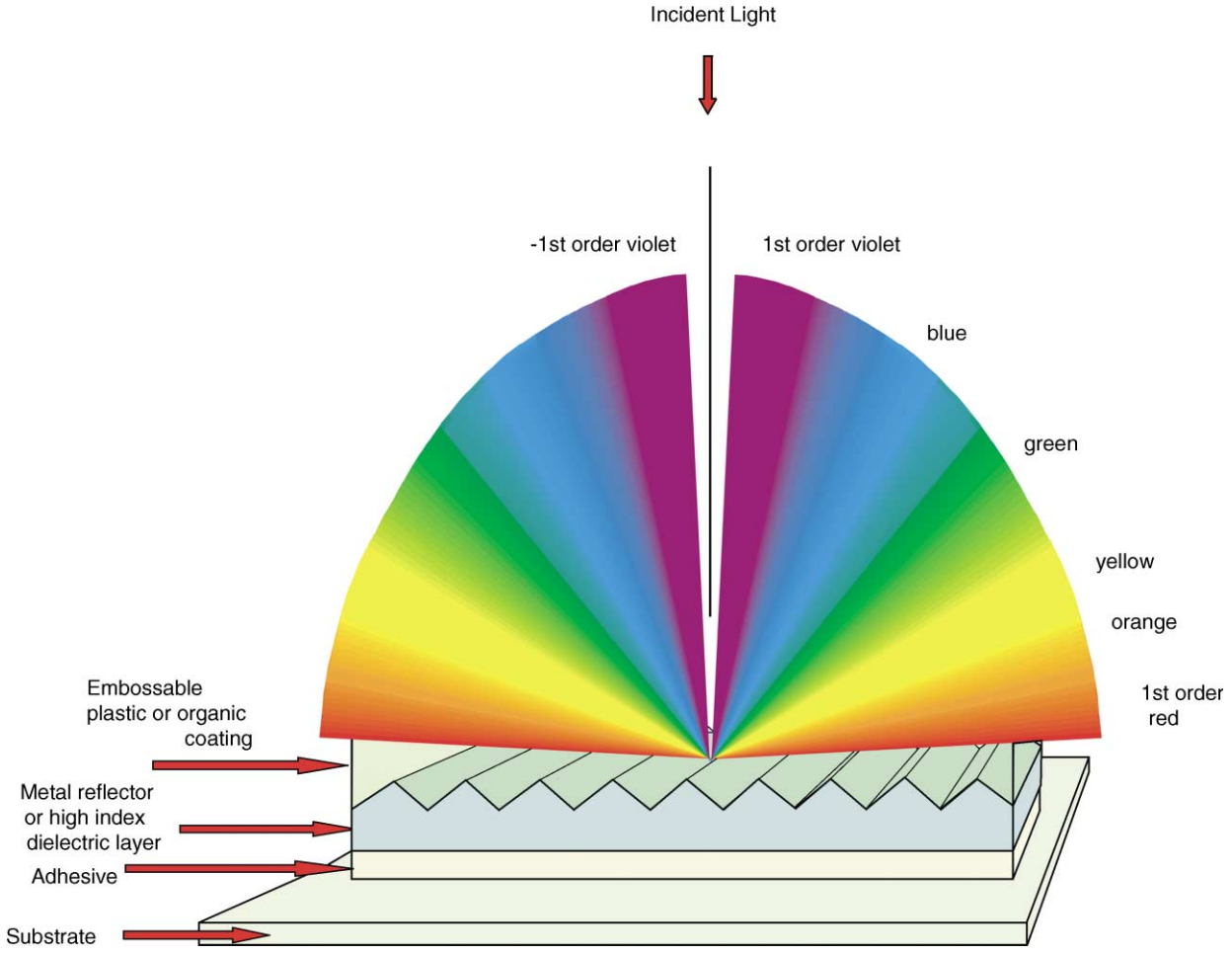


Figure 9 Cross-section of diffraction device. Courtesy of Flex Products, Inc.

rainbow effects disappear and are replaced by another type of diffraction (zero-order diffraction), where only specific colors are seen in reflection. Groove height and width for all diffraction products range from 0.1 to 10 microns.

For a unique set of discrete angles and for a given spacing d, the diffracted light from each facet is in phase with the light diffracted from any other facet, so they combine constructively:

$$Gm\lambda = \sin\alpha + \sin\beta \qquad [1]$$

where $G = 1/d$ is the groove density or pitch, α is the angle between the incident light and the normal to the grating, β is the angle between the diffracted beam and the normal to the grating and m is an integer called the diffraction order.

For $m = 0$, $\beta = -\alpha$ for all λ, the grating acts as a mirror and the wavelengths are no longer separated. This is called specular or zero-order reflection.

In contrast to grating structures (linear grooves) that only modify light intensities, holographic structures are complex surface contours that modify both the phase and intensity of the reflected light. Light reflecting off a metallized holographic structure is capable of producing an image in either two dimensions (2D) or three dimensions (3D). Heat and pressure may emboss the relief structure into a metallized polymeric film by hot-stamping. Alternatively, an embossing metal roller presses the relief image into a thermoplastic film or thin ultraviolet curable coating. The embossed structure is subsequently metallized with opaque aluminum or with a high-index dielectric material such as zinc sulfide or zirconium dioxide. The latter layers allow printed information on the substrate to show, while at the same time provide a holographic image. However, such high-index holograms are not as bright, since some of the reflective light is lost in transmission. Polymethylmethacralate (PMMA), polycarbonate (PC), polyvinyl chloride (PVC), and polyester (type G) are typical materials that can accept such an embossing.

Thick metallized embossed plastic films of ca. 50 microns are used for decorative labels. Thin metallized embossed hardcoats of ca. 7 microns, which can be released from carrier sheets or embossed thin plastic films, are used for the security field.

Holograms used for high-security applications have one characteristic that distinguishes them from ordinary holograms or other diffractive devices used for decorative purposes. This characteristic finds itself in the complication of images, multilevel images at different focal depths, different images that appear at various viewing angles, images of 3D objects superimposed with synthetic 2D images, micro-images, or images that appear to move as the device is tilted back and forth. Specific types of diffractive structures are described below. The list starts with simple holograms used in decorative markets and progresses to high-security devices.

Diffractive grating: A structure comprised of square or sinusoidal linear grooves that breaks light into its component colors. An assemblage of abutting islands, each randomly oriented and with its own parallel grooves, gives the viewer a twinkling color effect (Figure 10).

2D hologram: A relief structure that, upon viewing with a point source of light, shows a 2D image of objects, text, logos, or lines that emerge over a background. Such holograms are often referred to as 'cartoon images' since they are images of original drawings (Figure 11). Because of their simplicity and frequent use as decorative packaging, these holograms are generally not used in the security market.

Multilevel 2D hologram: A relief structure that shows a simulated 3D effect. This is accomplished by creating an image with two or more layers of flat artwork, with each layer at a slightly different focal depth (Figure 12). These images are well suited for labels, cards, and other decorative products. They have also been used on coupons, gift certificates, and tickets.

3D hologram: A relief structure that reconstructs a 3D image from a real or computer-synthesized 3D object (Figure 13). A 3D dove has been used on the VISA credit card as a security device for many years, though many now consider it more of a marketing logo than a security feature.

2D/3D hologram: A relief structure that reconstructs an image showing both a 3D and a 2D object, created from a 3D sculpture and flat art drawings (Figure 14). The artwork is created in such a way that the images appear at different depths of view.

Stereo-hologram: A relief structure that reconstructs a 3D image such that upon tilting from side to side one can see around the object, and additional features can then be viewed. At any given angle, it appears to be a 3D hologram. The image is created through multiple exposures of a 3D object during the holographic mastering process.

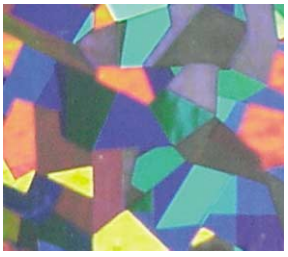


Figure 10 Picture of a typical diffractive assemblage of gratings. Reproduced with permission from Amagic Holographics Inc./K-Laser Technology, Inc.



Figure 12 Multilevel, 2D hologram. Courtesy of American Bank Note Holographics, Inc.



Figure 11 2D hologram. Courtesy of American Bank Note Holographics, Inc.



Figure 13 3D hologram. Courtesy of American Bank Note Holographics, Inc.

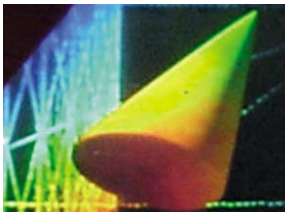


Figure 14 2D/3D hologram. Courtesy of American Bank Note Holographics, Inc.



Figure 15 True color hologram. Reproduced with permission from Light Impressions International, Inc.



Figure 16 High-definition security hologram. Number 50 appears in various colors in first angle of view. Colored shells appear at another angle of view. Reproduced with permission from DeLaRue International Limited.

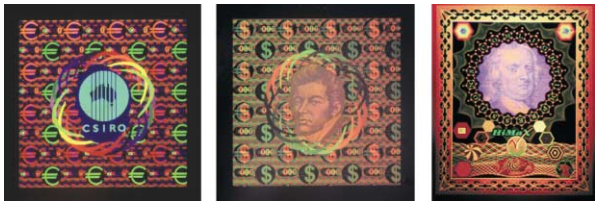


Figure 17 Left and center images – Exelgram™ at two different viewing angles. Reproduced with permission from CSIRO. Right image – HiMax® OVD high security device showing portrait of Sir Isaac Newton. Reproduced with permission from Light Impressions International, Inc.

True color hologram: In contrast to the holograms described above, a true color hologram is created by diffraction angle control of the colors red, green, and blue from a 3D object or color transparency. The relief structure, when viewed in white light, reveals a picture in true color, as opposed to the rainbow effect seen in regular 2D and 3D holograms (Figure 15).

High-definition security hologram: A relief structure that produces an image composed of multiple elements, such as fine line text and complex patterns that simulate movement as the device is tilted back and forth. Different images also appear at different angles (Figure 16). These images occur because different images are recorded in the holographic mastering process at different angles. The complicated imagery and the use of multiple images is the major holographic technique in use today for anticounterfeiting.

High-definition diffraction gratings: The relief structure in Figure 17 shows a multitude of elements, each with colors that change upon tilting. It consists of a very fine line grating structure, 1000 lines or greater per millimeter. These gratings are sometimes combined with holographic images that change depictions upon tilting, thus providing a high level of security.

Kinegrams®: This grating structure, based on computer-generated graphic images, is composed of light diffracting microstructures of less than one micron, with blaze (saw-tooth) contours that when viewed display optokinematic movements. This security device was first used in 1987 on Austrian 5000 Schillings. When the orientation of the blaze grating changes across the device, two distinct images are possible, depending on the view with respect to the blaze angles. The images do, however, require an aluminum reflective layer deposited onto the relief structure in order to generate the highly reflective colored images.

A 180° rotation of the device can lead to a marked change in the diffraction properties and, consequently, the image properties. This stands in contrast to holographic symmetrically based structures that yield identical images. An example of a Kinegram® security device is shown in Figure 18. The letter 'K' on the front facet of the cube is filled with an asymmetric structure, whereas the background is realized by using the same structure rotated by 180°. When viewed from one side, the letter 'K' appears bright on a dark background. Upon rotation in its plane by 180°, the contrast between the letter 'K' and the background is reversed.

Zero-order diffraction grating: This relief structure is quite different from holographic or normal grating structures, as it is based on an abrupt change in the



Figure 18 The Kinegram® security device as seen in two views differing from each other by a rotation of 180° in its plane, while illumination and viewing conditions remain unchanged. Reproduced with permission from OVD Kinegram Corporation.

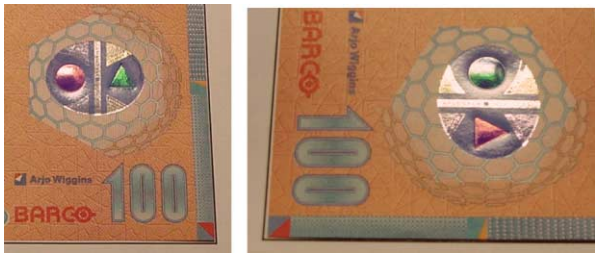


Figure 19 Zero-order diffraction images in a DID® security device: two views 180° apart in planar rotation. Reproduced with permission from Hologram Industries.

refractive index from approximately 2.3 to 1.5. Rather than employing an aluminized layer for reflection, the structure uses a thin, high-index layer ($n = 2.3$, for example) embedded in an organic polymer with a refractive index of about 1.5. The high-index layer appears as a parallel series of lines, approximately 0.2 microns wide, with another parallel series of lines beneath the first set, but oriented under the spacing of the first set of lines at a depth of about 0.2 microns.

This structure not only changes color as the device is tilted back and forth but also as it is rotated. In rotation, the color changes from red to green (Figure 19). Further, the light is polarized as it reflects. When a linear polarizer is inserted between the eye and the device, the color changes to black. The colors seen are based on specular zero-order reflection, unlike standard diffraction gratings where the reflected color is based on first-order reflection. The preeminent effect is when the device is placed over a black background. Over a white background the colors are subtler. Viewing reflection in zero order allows one to see the reflected colors in both point and diffuse lighting. The combination of polarization effects and color in rotation is ideally suited to high-security applications.

Devices Based on Light Interference

Thin-Film Foils

There are two types of thin-film interference devices. The first is based on an all-dielectric filter, and the second is based on a Fabry–Perot resonance cavity structure of the type absorber/dielectric/reflector. The all-dielectric filter uses pairs of high and low refractive index materials at a quarter-wave thickness. A typical filter is the one used initially on Canadian currency that employed the design: ZrO_2 1QW/SiO_2 3QW/ZrO_2 1QW/SiO_2 3QW/ZrO_2 1QW, where the QW (quarter-wave optical thickness) is centered at 585 nm. This design exhibits a gold-to-green color shift with viewing angle.

A Fabry–Perot filter also has a gold-to-green color shift and typically uses the design Cr 5 nm/MgF_2 4QW @ 604 nm/Al (opaque). This filter has slightly higher color saturation than the all-dielectric filter. The design is opaque and, therefore, the color observed is independent of the substrate color onto which it is affixed. This stands in contrast to the all-dielectric filter that must be laid over a black background for optical color saturation. If the all-dielectric filter is placed on a white background, the transmitted light passing through the optical filter will reflect off the white background and combine with the filter's reflection to give white light. However, with a varied colored background, the reflected color will vary depending on the combination of reflected color from the substrate and the reflected color from the filter.

The gamut of reflected colors from a metal–dielectric foil covers all four-color quadrants in the CIE Lab color space. Brightness of the foil depends on the materials used and on the design. Figure 20 shows the theoretical gamut of colors as plotted on the a^* b^* plane for the design, using the materials Cr, MgF_2 and Al.

Foils based on thin film interference layers contain a reflectance peak property moving from long wavelength to short wavelength as the viewing angle is increased. This is true across the UV, through the visible and into the near-infrared. To the human eye, the color moves from red to blue. The anti-counterfeit foil shown in Figure 21 exemplifies this effect.

Shifts from long to short wavelength occur because the path light difference (see the bold line in Figure 22) between the rays of light that reflect from the surface and the reflector are a function of the cosine of the incidence angle.

The path length difference is shown by the term $2d \cos \theta$, where θ is the angle of incidence within the

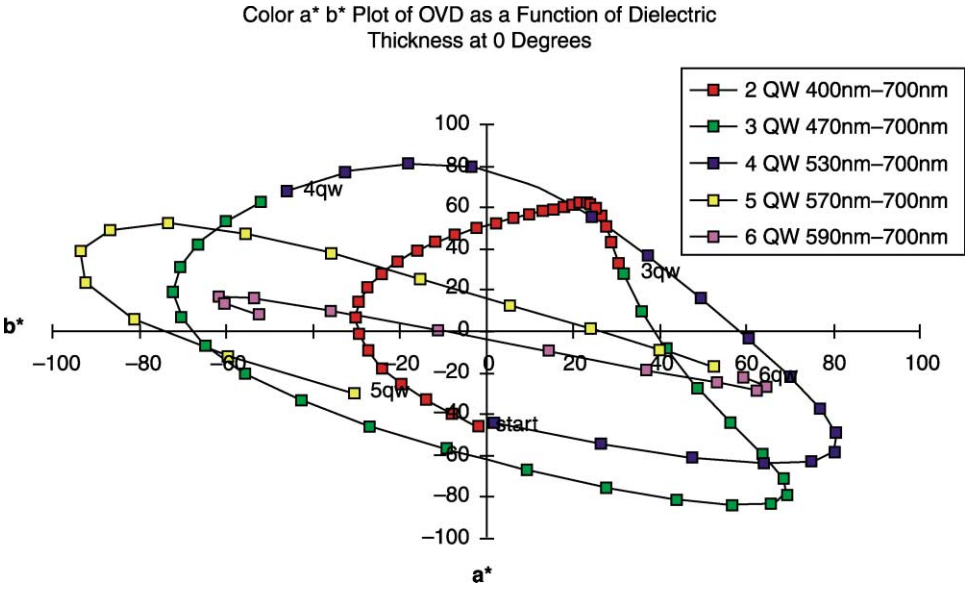


Figure 20 Gamut of colors at 0° incidence viewing under Illuminant C produced from Fabry–Perot designs as a function of quarter-wave optical thickness. Courtesy of Flex Products, Inc.



Figure 21 Anticounterfeit interference foil on brake shoes. Reproduced with permission from November AG.

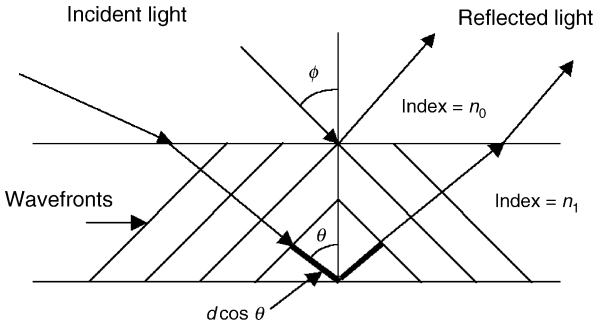


Figure 22 Diagram showing the principle of optical shift with viewing angle. Courtesy of Flex Products, Inc.

filter and d is the distance between the surface of the filter and the reflector. The angle θ and the angle ϕ are related by Snell's law: $n_0 \sin \phi = n_1 \sin \theta$.

Light interference occurs where the wavefronts superimpose on one another. The fractional number of wavelengths that can fit into this path length difference is given by $\Psi = 2n_1 d \cos(\theta)/\lambda$, where λ is the wavelength of light. Waves reflected from these two surfaces will interfere constructively when Ψ is an integral number of wavelengths and interfere destructively when Ψ is at multiples of half wavelengths. As the viewing angle (ϕ) increases, the angle θ also increases. Thus, constructive interference translates into a reflectance peak that moves toward shorter wavelengths as the angle of viewing is increased due to its cosine dependence.

The Fabry–Perot filter can use any transparent dielectric (colored or clear), a reflective metal, although optical metals are preferable since they have the highest reflectance (e.g., Ag, Al, Ni, Cu), and an absorber that is partially transmissive and partially reflective. Thin metals, absorbing dielectrics, and metal nitrides are typical materials used as an absorbing layer.

Light Interference Inks

Inks based on interference pigments have been developed over the last thirty years, starting with pigments based on high-index thin film coated micas (typically TiO_2 coated micas) and culminating with

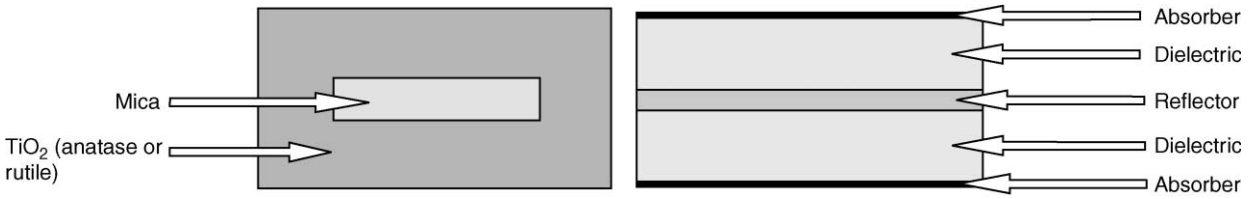


Figure 23 Left image – cross-section of interference mica; right image – cross-section of optically variable pigment (OVP®). Courtesy of Flex Products, Inc.

OVP® pigments based on the Fabry–Perot structure (Figure 23).

Inks created from micas are known as pastel or pearlescent inks, whereas inks formulated from OVP® are known as optically variable inks (OVI®). The latter inks are used as a primary defense against counterfeiting, defeating color copying and color printing from computer images. The difficulty with using interference micas for security features is that they can be readily purchased in the commercial world. In contrast, OVP® is a tightly controlled material and cannot be purchased by the general public.

OVI® security ink has been formulated into intaglio, silkscreen, flexographic and gravure applications. Its use on security documents encompasses over 75 banknotes worldwide, including the newly issued Euro notes, post-1996 $10, $20, $50, and $100 US banknotes, and banknotes for China. With the use of such inks, postage stamps, tax stamps, identification cards, and other security documents have also been effectively protected against counterfeiting. Color control for OVI® is assured because OVP® can be blended according to additive color theory to produce a desired color and color shift. Two examples of OVI® – the 100 Pula Botswana note and the new 100 Euro note – can be seen respectively in Figures 24 and 25.

OVI® inks have been extremely successful for banknote protection and might even be a *de facto* standard, because they meet the seven characteristics of an effective anticounterfeit solution as described earlier.



Figure 24 100 Pula Botswana note showing green-to-blue color shift OVI® with latent images. Reproduced with permission from SICPA.



Figure 25 100 Euro note showing magenta-to-green OVI®. Courtesy of Flex Products, Inc.

Devices Based on Light Polarization

The use of light polarization as an anticounterfeit feature has taken several avenues, one as coatings or pigments, the other as invisible images that are detected with a linear polarizer. With the exception of the zero-order diffraction device described earlier, polarization devices are based on cholesteric liquid crystals, using either nematic (planar oriented) layers or helical twisted layers.

Liquid crystal pigments are constructed from cholesteric polysilozanes that have been oriented and cross-linked in place with photopolymerization. With this cross-linking, pigments are relatively stable to temperature changes. These materials can form helical structures in either right or left twist orientation. Such materials reflect 50% of a band of wavelengths, centered at wavelength, λ_0, of the incoming light as circularly polarized light and 50% of the band centered at λ_0 as transmitted, circular polarized light, but with the opposite twist. The remaining wavelengths pass through the structure with no interaction. The amount and type of the

chiral phase (components that inherently orient into a helical structure) in the polymer determine the pitch of the helical structure. This, in turn, determines the wavelength of reflection, i.e. $\lambda_0 = nP$ where n is the average refractive index and P is the pitch length of the material (pitch length determined by 360° rotation of the helix). At tilted viewing angles, the color changes according to the formula $\lambda = \lambda_0 \cos\theta$ where θ is the viewing angle relative to the vertical, and λ_0 is the reflected wavelength seen perpendicular to the surface. By the judicious combination of left- and right-handed helical based layers, or by inserting a half-wave retarder layer between two layers having the same helical structure, the resulting pigments reflect nearly 100% at λ_0.

When these materials are properly processed and ground (particles 2–50 microns thick, with an aspect ratio of 3:1 in lateral/thickness dimensions), particles with the helical structure perpendicular to the planar surfaces can be realized. Such pigments, or liquid crystal polymeric coatings, when printed, have color-shifting properties with viewing angles that defeat color copiers and the like, since toners have no color-shifting properties (Figure 26). These pigments also have an additional security feature, namely, that when viewed through a circular polarizer having the opposite twist, the color-shifting pigments appear black.



Figure 26 Anticounterfeit label showing liquid crystal coating viewed straight-on and at tilt angles. Courtesy of Flex Products Inc.

A special technology known as linear photo-polymerization (LPP) is a process to produce photo-patterned liquid crystal polymer (LCP) $\lambda/4$ phase retarders, which, when combined with a series of layers involving a reflector, and a linear polarizer, shows positive and negative images. Without the diagnostic linear polarizer, however, the image is invisible (Figure 27).

Adding a nematic liquid crystal layer, which exhibits a helical twist configuration, extends this technology. The latter technology can produce colored images when viewed with a linear polarizer. Such devices can be coated or hot-stamped onto a substrate containing viewable information.

Combination Devices

Beginning in 2000, combinations of holographic structures with liquid crystal or interference thin films were developed, which could make counterfeiting even more difficult. Such structures emerge as a hologram floating on a color shifting background. Magnetic layers, integrated into part of the optical interference stack, allow spatial orientation and 3D effects during printing.

Decorative Coatings

Coatings Based on Diffractive Pigments

When pigments based on diffractive effects are created, individual flakes have surface contours of groove-like structures of 2000–3500 lines mm^{-1}. In spray-out coatings using such pigments, diverse interplay of subtle rainbow colors results (Figure 28). The effect is particularly striking in full sunlight as, for example, on curved surfaces found on cars and motorcycles. Some manufacturers have successfully used finely shredded diffraction gratings

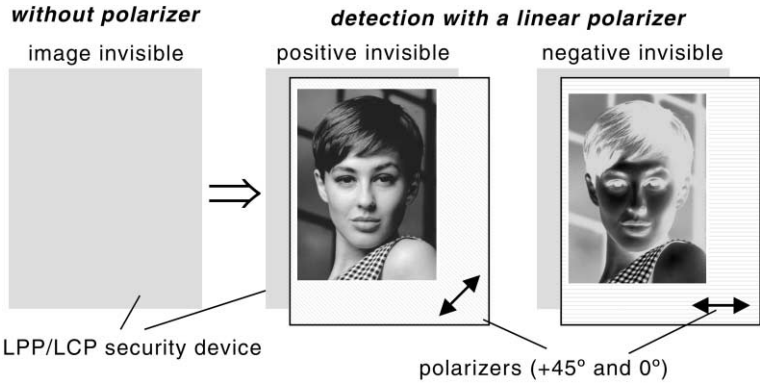


Figure 27 Reflective optical LPP/LCP security device showing a complex LPP/LCP retarder image visualized with one polarizer; rotation of the polarizer changes the image from positive to negative. Reproduced with permission from Rolic Research, Ltd.



Figure 28 Motorcycle painted with diffraction-based pigments. Courtesy of Flex Products, Inc.

(ca. 25 microns thick) as flakes mixed in paints for decorative purposes. The use of such thick flakes seriously limits the applicability and uniformity of the coating.

Coatings Based on Interference Pigments

Pigment-based interference effects are widely known and have been used for decades in the paint, plastics, and cosmetic industries. These structures are based on TiO_2 coated mica, where the mica thickness is between 300 and 600 nm with planar diameters ranging from 1 to 200 microns, depending on the application.

Depending on the thickness of the TiO_2, different reflective colors of silver, gold, red, blue, and green are achievable. For maximum color reflection, a black substrate or viewing at a particular angle is required so as to prevent dilution of the reflected color from light reflected from the substrate surface. Colored mica pigments are platelets of mica coated with colored oxides. Metallic-like colors are also formed from iron oxide coated micas. In the first instance, two colors may be seen: the platelet color (due to absorption) and the interference color at one angle. Only the interference color may be seen at another viewing angle. If the substrate is colored, the color switches from the interference color to the substrate color as the viewing angle changes. More recently, laminar pigments with an all-dielectric optical stack have been produced with a low-index central core, encapsulated with high-index layers that cause interference.

Interference pigments based on absorber/dielectric/metal/dielectric/absorber have also been introduced into the decorative marketplace. These structures are based on aluminum as the central layer, a dielectric spacer layer, and either a wavelength selective absorber layer such as iron oxide (Fe_2O_3) or a nonwavelength selective absorber such as chromium (Cr). In Fe_2O_3 based pigments, the face-on colors are mainly found in the yellows and reds whereas for the Cr-based pigments, a full color gamut is possible.

Diffractive Foils

These foils are based on diffractive and grating surfaces where the contours are generally less than 2000 lines mm^{-1}. These materials show a glittering effect, i.e., multiple points of color, as the viewing angle is changed.

Interference Polymeric Films

These films are made from multilayer extrusions of pairs of polymers of 50 or more layers. Each pair is composed of two different polymers with a slightly different refractive index. Each layer is approximately one-quarter wave in thickness. Due to the small refractive index differences between layers, the resultant film is approximately 25 microns thick. Uniform color at a given viewing angle is generally not possible due to the lack of thickness control. Hence, they tend to look like oil slicks, with each color changing as the viewing angle changes (Figure 29). Interference polymeric films have found use as decorative gift wrapping material.

Holographic Foils

As indicated earlier, holograms designed for non-security markets may range from simple 2D holograms to more complex ones that function for either decorative or brand protection purposes. Another application of holographic foils is in holographic fabrics, whereby small holograms approximating several millimeters in diameter are affixed in a pattern to the cloth (Figure 30).

Aluminum Flake Based Inks and Coatings

These inks are composed of leafing or nonleafing aluminum powders and a polymeric vehicle. The use of leafing pigments produces images and coatings of high brilliance, whereas nonleafing pigments produce a matte or satin-like finish.



Figure 29 Visual effect from multilayer extruded polymeric film. Reproduced with permission from Englehard Corporation.

Summary and Challenges for the Future

Optical scientists have shown tremendous creativity in discovering and employing clever technical devices in the fight against counterfeiting of valuable documents. Optical solutions have been created that meet the functional as well as the economic needs of the document security industry. A cadre of research institutions and manufacturing companies has accepted the challenge of protecting the instruments of our economic and physical security.

This vigilance will have to continue in the future. Nearly as fast as new defenses are developed, criminals are illicitly using new tools to defeat the protection schemes. Financing for sophisticated counterfeiting efforts is likely to increase, whether sourced by international terrorists or drug dealers. Likewise, the more we move toward an information society, the more we depend on machines to count, sort, and verify documents and money. This machine-driven trend will mandate new features, especially in banknotes and cards.



Figure 30 Fabrics showing holographic effects. Reproduced with permission from Amagic Holographics, Inc./K-Laser Technology, Inc.

It is likely that the future of optical anticounterfeiting technology will be strongly influenced by:

- Active, rather than passive, devices. That is, devices that can be applied to a document and later stimulated or interrogated by a machine to elicit a specific effect or signal.
- Schemes that may be merged with new biometric technologies, especially for individualized documents, such as passports and identification cards. In such applications, it will be necessary to confirm that a document is genuine, and that it is in the possession of its rightful owner.
- The need to clearly differentiate security applications from decorative or aesthetic uses. Effective security products must always be carefully controlled and their availability limited.
- The willingness of the public to accept regular changes in the appearance of value documents. Currency will have to undergo redesign on a three-to-six year cycle, rather than the long cycles typical in the past.

This chapter was intended only as a survey of the issues involved in anticounterfeiting protection. To learn more about the optical methods discussed here, or to learn more about the science of anticounterfeiting, please consult the "Further Reading" section and the Internet sites listed in Table 2.

Table 2 Internet sites: Banknote Design & Protection

Sponsor	*Web Address*
Amagic Holographics, Inc.	http://www.amagic-usa.com/
American Bank Note Holographics, Inc.	http://www.abholographics.com/
Bank of England	http://www.bankofengland.co.uk/banknotes/index.htm
Bundesdruckerei GmbH	http://www.bundesdruckerei.de/en/products/index.html
Crane & Co.	http://www.crane.com/specialty_papers/default.asp
Engelhard Corporation	http://www.engelhard.com/
European Central Bank	http://www.ecb.int/
Flex Products, Inc.	http://www.ocli.com/markets/flex_products.html
Giesecke & Devrient	http://www.gdm.de/eng/main/home/index.php4
Graphic Security Systems Corp.	http://www.graphicsecurity.com
Hologram Industries	http://www.hologram-industries.com/pages/index.htm
Joh Enschedé	http://www.joh-enschede.nl/divisies/jess-engels.html
K-Laser Technology, Inc.	http://www.klaser.com.tw
Light Impressions International, Inc.	http://www.lightimpressions.com/
November AG	http://www.november-ag.de
OVD Kinegram Corporation	http://www.kinegram.com/
Papierfabrik Louisenthal GmbH	http://www.louisenthal.de/
Rolic Research, Ltd.	http://www.rolic.com/
SICPA Security Inks	http://www.sicpa.com/default.asp?lang=e
Thomas De La Rue	http://www.delarue.com/
United States Bureau of Engraving	http://www.bep.treas.gov/

The authors express their gratitude to Ms. Sue Spadini for the technical editing of this work, and to the many individuals and companies who provided exemplars and illustrations. Please see the figure captions and the reference lists for their names.

List of Units and Nomenclature

2D	two-dimensional
3D	three-dimensional
CIE	Commission Internationale de l'Eclairage; International Commission on Illumination
CIELAB color space	(L^* = Lightness; $a^*.b^*$ = chroma coordinates)
CMYK	cyan-magenta-yellow-black
QW	quarter wave

See also

Holography, Techniques: Color Holography; Holographic Interferometry; Holographic Recording Materials and their Processing; Sandwich Holography and Light in Flight. **Microscopy:** Interference Microscopy. **Optical Coatings:** Thin-film Optical Coatings. **Polarization:** Introduction; Matrix Analysis.

Further Reading

Committee on Next-Generation Currency Design, National Materials Advisory Board (1993) *Counterfeit Deterrent Features for the Next Generation Currency Design*. Washington, DC: National Academic Press.

Demus D, Goodby J, Gray GW, Spiess HW and Vill V (eds) (1998) *Handbook of Liquid Crystals*. Weinheim, Germany: Wiley-VCH.

Fagan WF (1990) *Optical Security and Anticounterfeiting Systems*. Volume 1210. SPIE Conference – The International Society for Optical Engineering. Bellingham, WA.

Faris SM (1994) US Patent 5,364,557: *Aligned Cholesteric Liquid Crystal Inks*.

Glausch R, Kieser M, Maisch R, *et al.* (eds) (1998) *Special Effect Pigments*. Hanover, Germany: Vincentz Verlag.

Hutley MC (1982) *Diffraction Gratings*. London: Academic Press.

Phillips RW and Bonkowski RL (2000) *Security Enhancement of Holograms with Interference Coatings*. Volume 3973. SPIE – The International Society for Optical Engineering. Bellingham, WA.

Phillips RW and Nofi M (1999) *Colors by Chemistry or by Physics?* Society of Vacuum Coaters 42nd Annual Technical Conference Proceedings, pp. 494–499, Chicago, IL.

Phillips RW, Nofi M and Slusser R (1994) *Color Effects From Thin Films Designs*. Eighth International Conference on Vacuum Web Coating, pp. 270–284. Las Vegas, NV.

Proceedings of the International Conference on Security Documents for the 21st Century (1987) San Diego, CA. Prepared by Oak Ridge National Laboratories, CONF-87040175, Contract DE-A-C05-840R2140, Oak Ridge, TN 37831.

Rancourt JD (1987) *Optical Thin Films, Users' Handbook*. New York: Macmillan.

US House of Representatives (1984) *The Currency Design Act*. Hearing Before the Sub-Committee on Banking, Finance, and Urban Affairs, Second Session on H.R. 6005, Serial No. 98-98. Washington, DC: US Government Printing Office.

Van den Assem B, Brongers D, Rath J, *et al.* (1994) *Practical Guide for the Security Documents Printing Sector*. The Hague: SDU Uitgevererij Koninginnegracht.

Van Renesse RL (1994) *Optical Document Security*. London: Artech House.

Van Renesse RL (1996) *Optical Security and Counterfeit Deterrence Techniques*. Volume 2659. SPIE Conference – The International Society for Optical Engineering. Bellingham, WA.

Van Renesse RL (1998) *Optical Security and Counterfeit Deterrence Techniques II*. Volume 3314. SPIE Conference – The International Society for Optical Engineering. Bellingham, WA.

Van Renesse RL and Vliegenthart WA (2000) *Optical Security and Counterfeit Deterrence Techniques III*. Volume 3973. SPIE Conference – The International Society for Optical Engineering. Bellingham, WA.

Diamond Optical Devices and Coatings

D M Aslam, Michigan State University, East Lansing, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

What do a human, a plant, the photoresist used in microfabrication, and a polymer have in common? Why has carbon become a crucial part of all known living systems? The answer to these questions may be embedded in carbon's hybrid bonding structure (sp, sp^2, and sp^3 C–C bonds). Carbon has 4 electrons in its outermost shell (the valence shell), which is the second shell. At first sight, one would expect that these 4 electrons are distributed among its one 2s and three 2p orbitals. However, it is well known that, in the diamond lattice, each carbon atom bonds with its four neighbors. Are the electrons in 2s and 2p orbitals

responsible for the four bonds? One possible answer is that the 2s, $2p_x$, $2p_y$, and $2p_z$ orbitals of the carbon atom form bonds with their neighbors. There are two problems with this answer. First, these bonds will have different strengths (s and p orbitals), which is in contradiction with the experimental fact indicating an equal strength for all the four C–C bonds in the diamond. Second, the computed strength of such bonds will be lower than that found experimentally. What is then the nature of C–C bonds found in organic compounds?

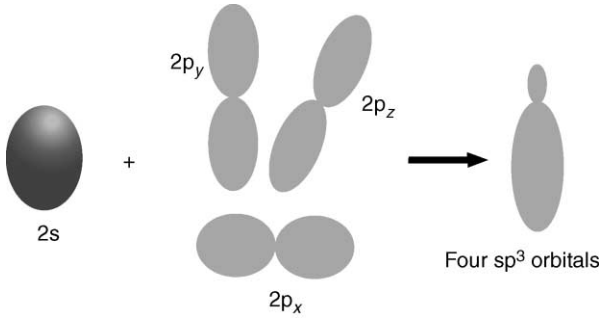


Figure 1 Hybridization of s and p orbitals in carbon to form sp^3 orbitals, which form the so-called sigma bonds responsible for the strong diamond lattice.

The answer to this difficult question was provided in 1931 by Linus Pauling (1901–1994), the winner of two Nobel prizes. With the help of quantum mechanical calculations, Pauling showed how one s orbital and three p orbitals can mix, or hybridize, to form four equivalent atomic orbitals with tetrahedral orientation. The resulting hybrid orbitals are called sp^3 orbitals as shown in Figure 1. The sp^2 or sp^1 orbital results if two or one p orbital combine with the s orbital, respectively, to form the hybrid. While sp^3 C–C bonds lead to a diamond lattice, the sp^2 C–C bonds lead to graphite, and a material which contains both sp^2 and sp^3 C–C bonds is called diamond-like carbon (DLC). The optical properties, such as the bandgap, of DLC depend on the sp^3/sp^2 ratio. If a sheet consisting of a monolayer of graphite, which consists of sp^2 bonds, is wrapped into a pipe, the resulting structure is called a single-wall nanotube (SWNT), as shown in Figure 2a. A multiwall nanotube (MWNT) has many such layers (Figure 2b–c). Other elements such as oxygen (in H_2O) and nitrogen (in NH_3) also form sp^3 orbitals but carbon seems to be only one which is capable of forming sp, sp^2, or sp^3 bonds, depending upon the bonding environment (Figure 3). This may explain why nature has selected carbon

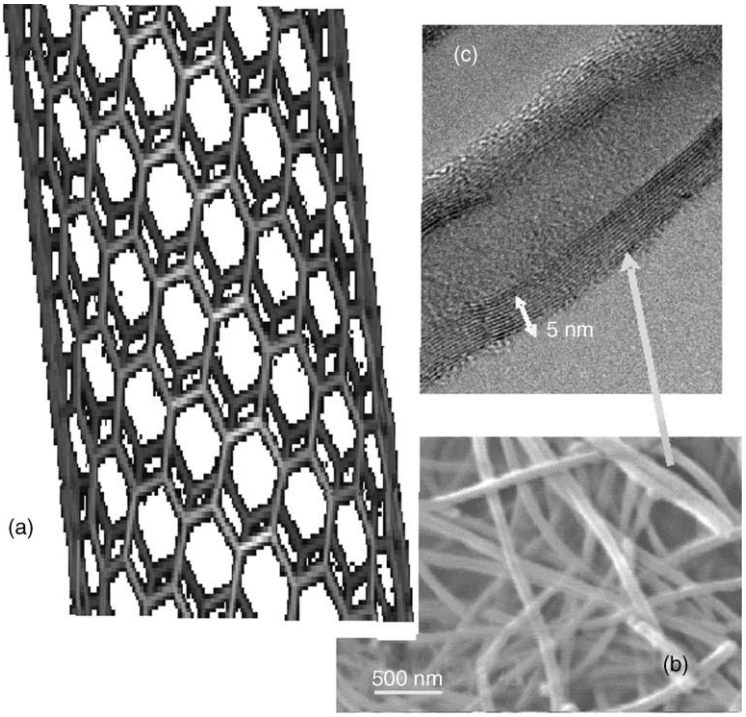


Figure 2 CNT: (a) drawing of a single wall CNT showing sp^2 bonded layer; (b) a mat of multi-wall CNTs; and (c) a TEM of MWNT showing 13 sp^2 bonded layers.

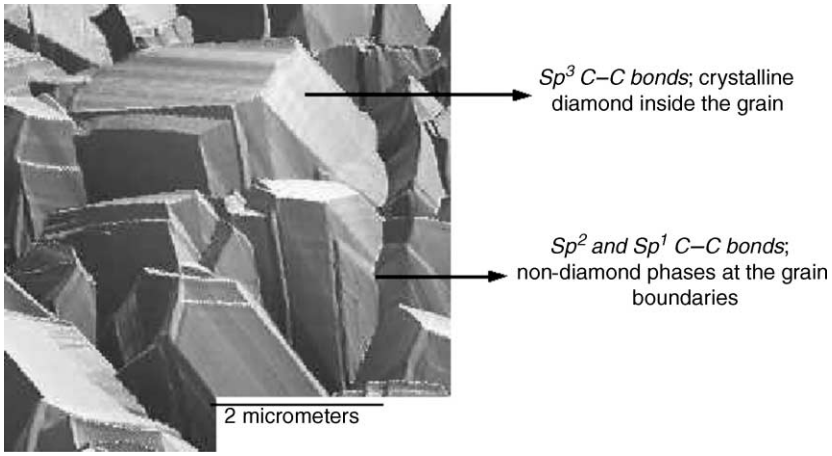


Figure 3 Polycrystalline diamond film; crystalline regions separated by grain boundaries containing non-diamond phases.

Table 1 Comparison of semiconductor properties

Property	*Diamond*		*Si*	*GaA*	*β*-SiC
	Poly-C	*Crystalline*			
Fig. of merit					
Johnson, 10^{23} W Ω s^{-2}		73 856	9	62.5	10 240
Keyes, 10^2 W cm^{-1} s^{-2} °C		444	13.8	6.3	90.3
Sat. electron vel., 10^7 cm s^{-1}		2.7	1	1	2.5
Carrier mobilities, (cm^2 V^{-1} s^{-1})					
Holes	1–165	1600	600	400	50
Electrons		2200	1500	8500	400
Resistivity, Ω cm	10^{-2}–10^8	10^{-3}–10^{13}	10^{-4}–10^8	?–10^8	150
Breakdown field, MV cm^{-1}	0.1–1	10	0.3	6	4
Band gap, eV		5.45	1.12	1.42	3
Dielectric constant	6.7	5.5	11.7	12.5	9.7
Thermal expan. coeff., 10^{-6} $°C^{-1}$	2.6	1.1	2.6	5.9	4.7
Thermal cond., W cm^{-1} K^{-1}	20	4–22	1.5	0.5	5
Lattice constant, Å		3.57	5.43	5.65	4.36
Density, g/cm^3		3.52	2.32	5.31	3.215
Melting point, °C		4000[a]	1412	1240	2540
Hardness, kg mm^{-2}		10 000	1000	600	3500
Poisson ratio	0.11	0.15	0.23	0.315	
Young's modulus, 10^{12} Pa	0.8–1.2	1.1–1.2	0.155	0.085	0.7

[a]Diamond may convert to graphite well below 4000 °C depending upon the ambient and it oxidizes at 650 °C to form CO_2/CO in O_2.

(perhaps due to its bonding flexibility, to form long C–C chains and rings) as the material of choice for building living systems.

The sp^3 bonding in a diamond lattice makes it a very unique material with a combination of properties not matched by any other known materials (Table 1). In fact, the C–C sp^3 bond in diamond is responsible for its remarkable properties leading to current and potential applications as shown in Figure 4. Although diamond (for centuries) and its bonding structure (for over 70 years) has been known for a long time, the growth of diamond films using chemical vapor deposition (CVD) took a long time to mature. Again, this delay relates to the difficulty with which the diamond lattice with sp^3 C–C bonds is fabricated.

Diamond Growth by CVD

First evidence of diamond growth by CVD, by Eversole in 1952–1953, led to the use of H_2 and CH_4 in the hot filament CVD (HFCVD) to grow diamond on diamond substrates (homoepitaxial growth) by Angus in 1971. The inexpensive CVD polycrystalline diamond (poly-C) was grown on non-diamond substrates by Deryagin in 1976, Spitsyn in 1981, and by Matsumoto *et al*. in 1983. Currently,

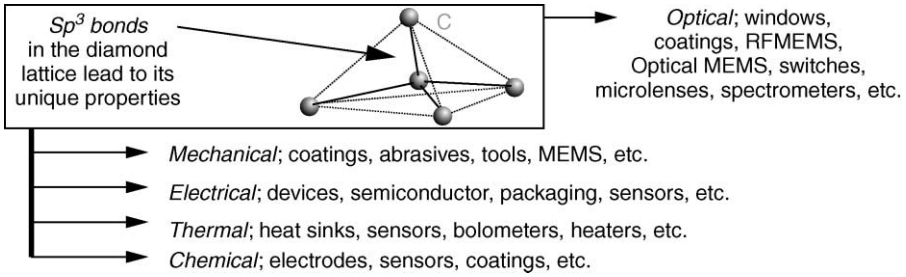


Figure 4 Sp^3 bonds in diamond make it a unique material.

Table 2 Different diamond deposition methods

Methods	*HFCVD*	*MPCVD*	*DC-arc Jet CVD*	*Combustion synthesis*	*Multiple pulsed laser*
Deposition rate (μm/hr)	0.1–10	0.1–10	30–150	4–40	3600
Substrate temperature (°C)	300–1000	300–1200	800–1100	600–1400	50
Deposition area (cm^2)	5–900	5–100	<2	<3	N/A
Advantages	Simple, large area	Quality, stability	High rate, good quality	Simple, high rate	Ultra-high rate
Disadvantages	Contamination, fragile filament	Rate	Contamination, small area	Small area	Expensive

HFCVD, hot filament CVD.

it is believed that, during the CVD of diamond, the CH_3 is responsible for deposition of C as diamond and non-diamond phases. The atomic hydrogen, present in the growth environment, removes the non-diamond phases leaving behind the diamond phase. Near the end of the 1990s, basic science of CVD diamond was well understood, and today diverse plasma and thermal techniques have been developed to produce poly-C films several mm thick and over 12 inches in diameter. Optically smooth 300-micrometer thick undoped poly-C wafers are available on the market. Although there are some reports of n-type poly-C and crystalline diamond growth, the well-established techniques exist only for *in situ* doping of p-type diamond.

A number of diamond growth techniques are currently available for the growth of diamond and carbon nanotubes. Microwave plasma CVD (MPCVD) can be used for the growth of both diamond and carbon nanotubes. As shown in Table 2, the deposition rate of diamond shows a large variation (0.1–3600 micrometers/hour) depending on the growth technique. For applications of diamond in optical coatings, optoelectronic devices and optical systems, the study of transmission, reflection and absorption of light for diamond is very important. These properties of diamond are responsible for its beauty and glamor. For applications in optoelectronic materials, devices and systems, the micro-fabrication and -machining technologies are crucial, particularly in the case of inexpensive poly-C films.

Absorption, Reflection, and Transmission in Diamond

The optical properties, which have been studied for hundreds of years, of naturally occurring single crystal diamond have led to the gem commerce. The fundamental concept involved in the tremendous historical fame of diamond is its high refractive index, n_d (in the range of 2.38–2.64 depending upon the wavelength of light). The refractive index of a material is a measure of decrease of velocity of light (which is also true for all electromagnetic waves) when the wave enters the material. Mathematically, n_d is defined by

$$n_d = V_{vac}/V_d \qquad [1]$$

where V_{vac} and V_d are velocities of light in vacuum and diamond, respectively.

When a beam of light of intensity I_0 in air is incident normally on the surface of diamond, the reflected part of the beam, characterized by the reflection coefficient R, depends on n_d, and R is given by

$$R = \left(\frac{n_a - n_d}{n_a + n_d}\right)^2 \qquad [2]$$

where n_a is refractive index of air and n_d is given by Sellmeier dispersion equation:

$$n_d = \sqrt{1 + \frac{4.3\lambda^2}{\lambda^2 - 0.0121} + \frac{0.33\lambda^2}{\lambda^2 - 0.031}} \quad [3]$$

where λ is the wavelength of light expressed in micrometers. When the beam enters diamond, part of it is absorbed in diamond. The absorption, characterized by the absorption coefficient α, is caused by: (i) impurities (such as nitrogen, boron, hydrogen, etc.); and/or (ii) intrinsic defects (such as dangling bonds, sp^2 bonds, and sp bonds). However, if the energy of light is close to or above the bandgap of diamond (5.5 eV), the absorption can be caused by electronic transitions from valence to conduction bands. The intensity $I(x)$ at a distance x below the diamond surface is given by

$$I(x) = I_0(1 - R)\exp(-\alpha x) \quad [4]$$

The fundamental expressions for α are different for indirect and direct type semiconductors. However, the modeling of α, for energies close to the bandgap of diamond, can be accomplished empirically using Urbach's rule according to the following equation:

$$\alpha = \alpha_0 \exp\left[\frac{\sigma(T)(E - E_0)}{kT}\right] \quad [5]$$

where α_0 and σ are Ubach tail parameters, and E_0 (6.5 eV) is direct gap energy of diamond. K is Boltzmann constant and T is the temperature. As at $T = 300$ K, $\alpha_0 = 4.23 \times 10^{11}$ cm^{-1} and $\sigma = 0.585$, eqn [5] can be simplified:

$$\alpha = 4.23 \times 10^{11}\, e^{22.6(E-6.5)} \quad [6]$$

where the units for α and E are in cm^{-1} and eV, respectively. Using the expression $E = 1.24/\lambda$, where E is in eV and λ is in micrometers, eqn [6] can be written as

$$\alpha = 4.23 \times 10^{11}\, e^{27.9(1/\lambda - 5.24)} \quad [7]$$

where λ is in micrometers. The variations of α and n_d with wavelength and energy of light are shown in Figure 5, which are plotted using eqns [3] and [7]. Impurities and intrinsic defects in diamond strongly affect α, n_d, and its other properties.

Impurities in Diamond

A pure diamond with perfect lattice structure is transparent to ultraviolet (UV), visible, infrared (IR), millimeter and microwave regions of the electromagnetic spectrum. However, even an impurity-free diamond has some absorption in the infrared due to phonon excitations and an ultraviolet cutoff due to band-to-band absorption. The defects, impurity-related and intrinsic, in single crystal diamond have been extensively studied because they strongly affect its properties. In fact, diamonds are classified according to the type of defects that are dominant.

Impurities in diamond give rise to additional energy levels in its bandgap. If the density of these impurities is sufficiently high, they can cause absorption of electromagnetic radiation (including light) on one hand and break the lattice symmetry on the other. The breaking of the lattice symmetry can lead to one-phonon IR absorption, which is forbidden in pure diamond. The fact that impurities affect the optical properties of diamond led to its classification into types Ia, Ib, IIa, and IIb which correspond to pale yellow, golden yellow, no color, and blue, respectively. The transmission properties of these various types of diamond are depicted in Figure 6. The dominant impurities are nitrogen

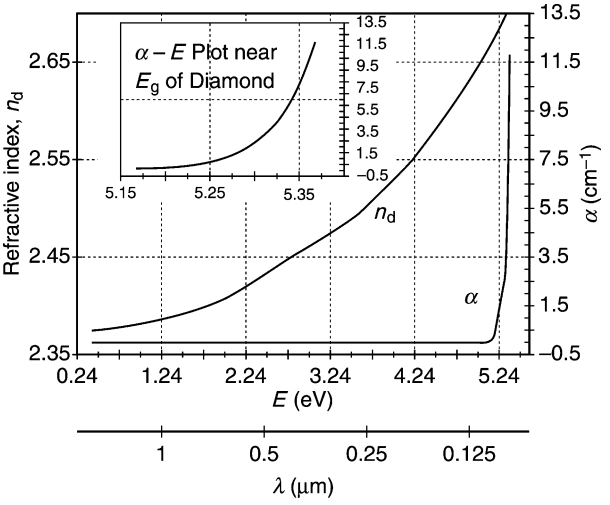


Figure 5 Wavelength and energy dependence of absorption coefficient and refractive index of diamond.

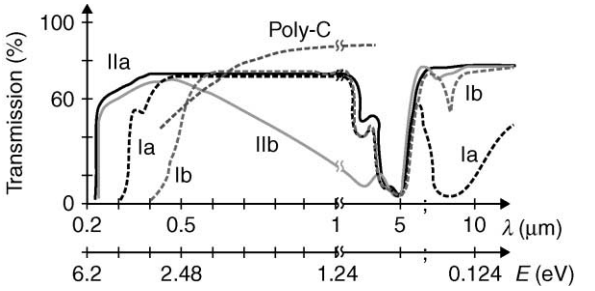


Figure 6 Transmission spectra of natural diamond and CVD polycrystalline diamond (poly-C).

(5–2500 ppm atomic) in type I and boron (10 ppm atomic) in type II. It may be pointed out that boron-doped diamond is a semiconductor with p-type conductivity measured at room temperature. Nitrogen-doped diamond does not show n-type conductivity at room temperature as the donor levels are too deep to be ionized.

Polycrystalline CVD Diamond

Due to the lack of a reliable hetero-epitaxial diamond growth technique, the crystalline diamond films are very expensive (require single crystal diamond substrates for their growth) and, consequently, their widespread commercial use is severely limited. Polycrystalline diamond (poly-C) grown by CVD, which can be as inexpensive as any other material if mass produced, is very attractive for optical and other applications (Figure 2). The properties of high-quality (with high density of sp^3 C–C bonds) poly-C films are comparable to those of crystalline films.

As shown in Figure 6, the transmission properties of poly-C are comparable to or better than those of type I and II diamonds for part of the energy range. Free-standing commercial poly-C films are typically polished and have thicknesses in the range of 50–300 micrometers (the least expensive are 300 μm thick). White poly-C films transparent to visible light are also available. Such films, with optically smooth surface, can be considered free of scattering effects. These films can be considered as a transparent plate in air and, for normal incidence of light with energy less than the bandgap, the transmission coefficient is given by

$$T = \frac{(1 - r_1^2)(1 - r_2^2)}{1 + r_1^2 r_2^2 + 2r_1 r_2 \cos(2\delta)} \qquad [8]$$

where the Fresnel coefficients are given by

$$r_1 = \frac{n_a - n_d}{n_a + n_d}, \qquad r_2 = \frac{n_d - n_a}{n_d + n_a} \qquad [9]$$

where n_a and n_d are refractive indexes of air and poly-C, respectively, and δ is phase change for a poly-C film with a thickness of t:

$$\delta = \frac{2\pi}{\lambda} n_d t \qquad [10]$$

Using $n_a = 1.0003$, $n_d = 2.4$, and $t = 100$ μm, eqn [8] can be written as

$$T = \frac{1}{1.49 - 0.493 \cos\left(\frac{3015}{\lambda}\right)} \qquad [11]$$

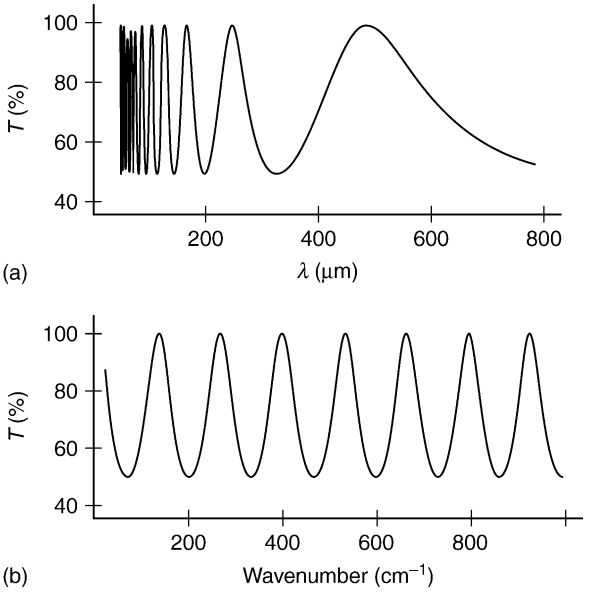


Figure 7 Transmission coefficient of a polished poly-C film with a thickness of 100 μm as a function of (a) wavelength and (b) wavenumber.

or

$$T = \frac{1}{1.49 - 0.493 \cos(0.048\, k)} \qquad [12]$$

where λ is in μm and k, the wavenumber, is in cm^{-1}. Variation of T as a function of λ and k, is shown in Figure 7. A similar behavior is observed for non-polished poly-C films if the wavelength is larger than the surface roughness of the films. For a sample with a very rough surface, scattering effects need to be included in eqn [8].

Electroluminescence in Poly-C

In a luminescence process, the electron must first be excited to higher energy, which can be accomplished by light (photoluminescence), electron beam (cathodoluminescence) or electrical energy (electroluminescence). A subsequent electron transition to lower energies leads to emission of light. A common example of electroluminescence is a forward biased p–n junction made out of direct bandgap type semiconductor such as GaAlAs. The light emitting diodes (LED) and LASERS are based on electroluminescence. Red, green, and blue LEDs are made from GaAlAs, GaP, and SiC/GaN, respectively. Indirect bandgap type semiconductors, such as diamond, typically do not exhibit electroluminescence. The fact that electroluminescence has experimentally been observed in crystalline diamond and poly-C seems to be, at first sight, a contradiction of its

indirect bandgap. In fact, defects in diamond are believed to be responsible for their luminescence properties. All types of diamond are found to have a number of defects related to vacancies and impurities. Due to a unique combination of properties of diamond (Table 1), its electroluminescence properties offer a very unique application potential in optical microelectromechanical systems (MEMS), displays and communication devices in IR, visible and UV regions.

Electroluminescence (EL) in optoelectronic devices is produced by excitation of electrons by electronic current. Yellow-green EL has been observed in Schottky diodes made of boron-doped poly-C. Blue-green and blue-violet EL has been observed in free-standing poly-C films, with thicknesses in the range of 15–400 micrometers, when the samples were subjected to applied fields in the range of 2 V/μm. The EL peaks are observed at energies in the range of 3–3.85 eV, with the main peak occurring at 3 eV.

Recently, a very interesting type of EL has been observed in poly-C and carbon nanotubes, which seems to have an application potential in displays and optical MEMS. Under applied electric fields in the range of 20–40 V/μm, electrons are known to tunnel out of a poly-C film into vacuum and the process is called field emission. EL has been observed during the field emission process, and this new effect is called field emission EL (FEEL). A typical FEEL spectra is shown in Figure 8. It is believed that FEEL results from direct type defects in poly-C, which are populated during the field emission process. The electrons in these defects, instead of being field emitted, make transitions to lower energy levels leading to light emission. The main FEEL peak is observed at 2.6 eV which is in the visible range.

Poly-C Applications

The current applications of diamond, which include optical windows, temperature sensors, heat sinks, cutting tools, surgical instruments, and optical coatings, are based on natural, synthetic, and CVD diamond. The cost of producing diamond has always been prohibitive in widespread diamond applications. Fortunately, over the past 20 years, the development of CVD diamond technologies has eradicated the fears of diamond being too expensive for its widespread applications. As the CVD poly-C can be as inexpensive as any other electronic material, widespread applications of diamond are expected for poly-C. In the area of micro- and nanostructures, the poly-C can provide a unique opportunity for both polycrystalline and crystalline microdevices. Using the present fabrication techniques, the grain size in poly-C can be varied in the range from a few tens of nanometers (nanocrystalline diamond) to hundreds of micrometers (commercial free-standing poly-C films). This leads to a whole range of device structures ranging from inter- to intra-grain devices. An intra-grain device made in poly-C will be an inexpensive single crystal device with properties better that those found in natural/synthetic diamonds, because an intra-grain poly-C device will have a higher degree of purity (because it is grown by CVD).

The applications of poly-C in microsystems (current, in-progress, and future) include optical MEMS, biological/protective/hard coatings, sensors,

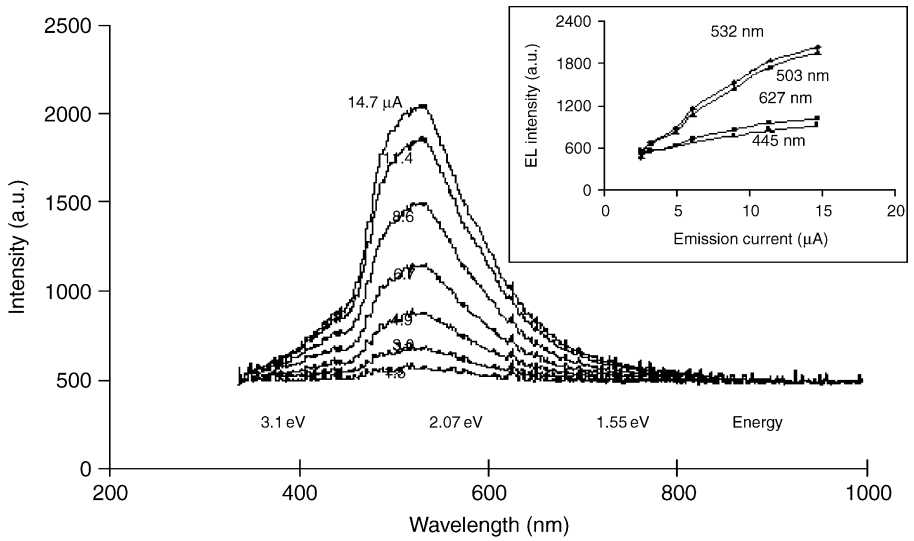


Figure 8 FEEL spectra in poly-C as function of wavelength and emission current (inset).

field emission displays, and WIMS (wireless integrated microsystems). Intra-grain microlenses, RFMEMS, heat pipes, sensors, microabrasives, and spectrometers may lead the way for unique poly-C applications.

Optical MEMS

A basic MEMS structure, with electrostatic actuation mechanism, is shown in Figure 9. Currently, most of the Optical MEMS and RFMEMS (radio frequency MEMS) devices are made from poly-Si (polycrystalline Si) using surface or bulk micro-machining fabrication techniques. The ability to integrate micro-optical elements with movable structures and micro-actuators has opened up many new opportunities for optical and optoelectronics systems, as MEMS allow a more effective manipulation of optical structures than the conventional methods. Optical MEMS have applications in displays, sensing, data storage, and communication systems. Some of the examples are optical switches, optical cross-connects, wavelength division add/drop mixers, tunable filters/lasers/detectors, dispersion components, polarization dispersion components, and spectrometers.

Material-related limitations put some constraints on the current optical MEMS technologies and applications. Diamond is an excellent material for optical MEMS and can extend the application range and reliability. The use of poly-C in optical MEMS can also remove the traditional cost constraints associated with the use of single crystal diamond. Although the poly-C optical MEMS are in their infancy, recent developments are very exciting. Figure 10 shows a poly-C resonator structure fabricated using a poly-C fabrication technology compatible with Si processing techniques. Such structures, if used as filters and switches, can lead to a superior performance as compared to any other currently used material. Poly-C on-chip spectrometers are also expected to operate in a large wavelength range.

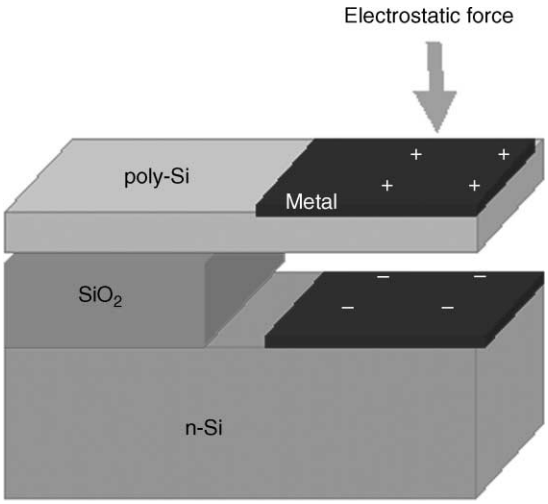


Figure 9 A simple MEMS structure with electrostatic moving of the cantilever beam.

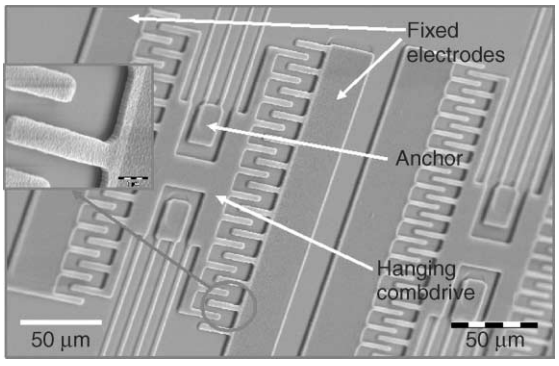


Figure 10 A poly-C MEMS combdrive for high-quality factor resonators, inertial sensors, optical MEMS, etc. Poly-C is superior to poly-Si as its Young's modulus is 4 times higher than that of poly-Si.

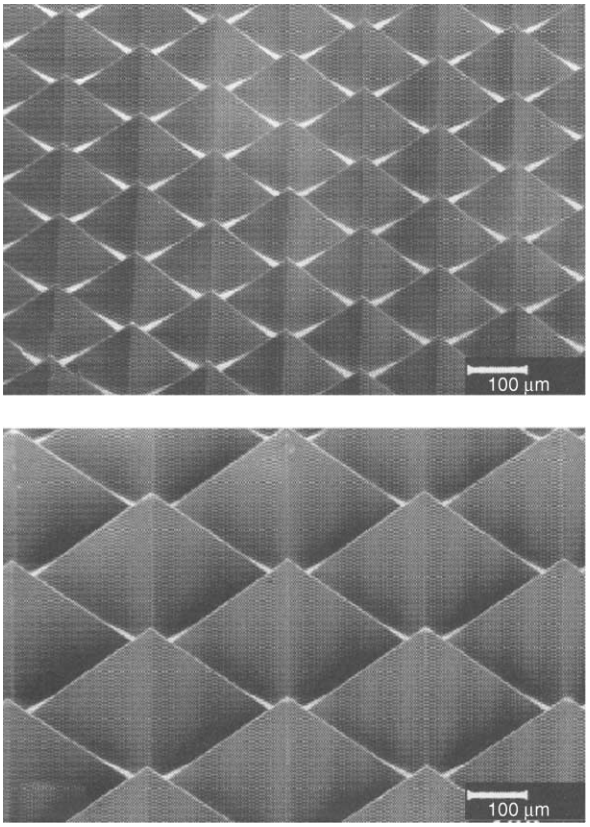


Figure 11 Poly-C microabrasive structures two different densities of pyramids. (Courtesy of Dr. H. Bjorkman of Upsala University.)

Microabrasives

Poly-C is also a very promising material for microabrasives. Micromachined poly-C MEMS structures have recently been fabricated for possible use for microabrasives. Figure 11 shows micromachined MEMS structures which are expected to enhance the life of a tool. Such abrasives by design are inherently cost-effective because they are batch fabricated.

See also

Light Emitting Diodes.

Further Reading

Asmussen J and Reinhard D (2002) *Diamond Handbook*. New York: Marcel Dekker.

Kim U and Aslam DM (2003) Field emission electroluminescence on diamond and carbon nanotube films. *Journal of Vaccum Scientific Technology B* 21(4): 1291–1296.

McMurry J (1996) *Organic Chemistry*. New York: Brooks/Cole.

Pan LS and Kania DR (1996) *Diamond: Electronic Properties and Applications*. Boston, MA: Kluwer.

Spear KE and Dismukes JP (1994) *Synthetic Diamond: Emerging CVD Science and Technology*. New York: Wiley.

Laser Damage in Thin Film Coatings

D Ristau, Laser Zentrum Hannover e.V., Hannover, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since the very beginning of laser technology, Laser-induced Damage Thresholds (LIDT) of optical components were always an obstacle for the development of laser systems operating at high power levels. In such systems, the surface or the bulk of the optical components can be damaged by laser radiation of sufficiently high power or energy density resulting in an uneconomical standstill of the laser and its application. In the early days of high power laser development, mainly inclusions in laser rod materials were discussed as a major complication for an augmentation of the output power in solid state laser systems. Nowadays, in the course of the development of optical materials with excellent quality and power handling capability, the problem of laser-induced damage has shifted from the bulk to the surface of the optical component. The optical surface is subjected to various production steps and environmental influences, which modify its structure and composition. Especially, the thin film coating, which is deposited on the optical surface to adapt its reflectance and transmittance to the application, contributes predominantly to the reduction of the LIDT values. As a consequence, the measurement and optimization of the power handling capability of thin films is considered as one of the primary research areas in modern optics technology and is supported by an extensive scientific community.

In the following, a brief review will be given on selected fundamental damage mechanisms in thin films, considering also the scaling of LIDT values for different operation conditions. Also, current standards for the measurement of LIDT will be described, and examples illustrating some practical aspects of high power optical coatings will be presented. Finally, a summary of the present state of the art will be given and recent trends in laser technology will be discussed in respect to research in laser-induced damage.

Fundamental Damage Mechanisms in Thin Films

During thirty years of intense research in laser damage of thin films, a broad spectrum of different damage and degradation mechanisms had been discovered and studied for a variety of laser operation conditions and wavelengths. For prominent high-power laser systems often applied in industrial production fields or fundamental research, specific models for laser-induced breakdown have been developed, which are also of practical relevance. In many experiments, damage sites with the morphology illustrated in Figure 1 are observed. The coating is delaminated from the surface of the optical component in the center of the laser beam area. Obviously, mechanical stress has built up in the coating by heating, which is induced by the absorbed laser power. At a certain stress level, exceeding the adhesion strength of the coating to substrate surface, the thin film cracks and may even delaminate from the component. In other cases of this absorption-induced damage effect, the film reaches its melting point prior to the mechanical threshold and

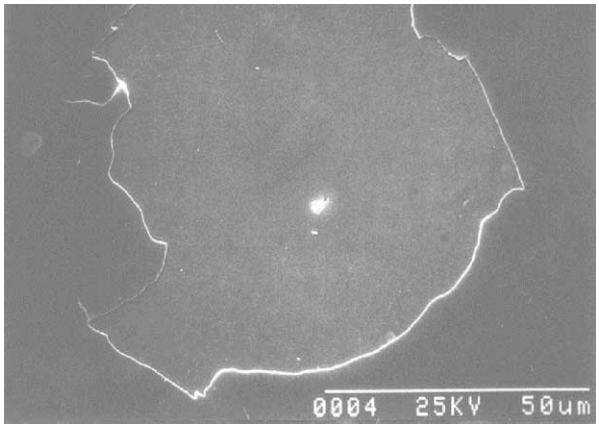


Figure 1 SEM-picture of a damage site on an alternating coating of SiO_2/HfO_2. The system was irradiated by a Nd:YAG-laser with an energy density of 45 J/cm^2. As a dominant damage mechanism, a delamination of the coating by extreme mechanical stress is suggested. Reproduced with permission from Ristau D, Ebert J and Welling H (1989) Optische Beschichtungen für Hochleistungslaser. *Laser und Optoelektronik* 21(4): S.49.

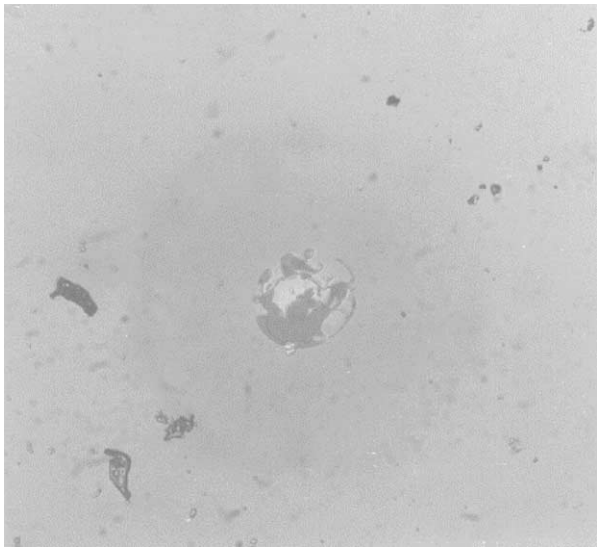


Figure 2 Nomarski micrograph (magnification 250×) of a damage site on an antireflecting coating of SiO_2/Ta_2O_5. The system was irradiated with an energy density of 51 J/cm^2 at a wavelength of 1.064 μm. A delamination of the coating in the center of the beam is surrounded by a recrystallization zone.

evaporates or changes its crystalline structure. The morphology of such damage sites exhibits a discoloring or an increased surface roughness in the center of the laser beam (see Figure 2). Absorption-induced damage is dominated by the instantaneous heating of the coating material in the area of interaction with the laser beam and can be described on the basis of the heat diffusion equation. In this approach, the source term in the heat diffusion equation is given by the heat coupled into the system by the absorbed laser power, and the boundary conditions are determined by the geometry and the structure of the layer system as well as the spatial laser beam profile. Apparently, absorption-induced damage has to be considered only for components with significant absorptance at their operation wavelength. Typical examples for coating materials and wavelength regimes, where absorption-induced damage is frequently observed, are the wavelength 10.6 μm of the CO_2-laser with ZnSe, ZnS, and fluorides as coating materials, or the DUV/VUV-wavelengths dominated by excimer lasers and fluoride coating materials. The effect of absorption-induced damage in optical materials is also of practical relevance for cw- and long-pulse operation conditions of the laser system corresponding to long thermal diffusion lengths. For cw-irradiation, the temperature rise in the component can be calculated by numerical methods involving finite elements or differences. For boundary conditions with circular symmetry and a Gaussian beam profile (beam diameter w, power P), an analytical expression can be derived for the temperature rise ΔT in the center of an irradiated circular component

$$\Delta T = \frac{2\beta_s P}{\pi^{3/2} k w} \tan^{-1}\left(\frac{16\kappa t_I}{w^2}\right)^{1/2} \qquad [1]$$

In this model the temperature rise is dependent on the thermal properties (k: thermal conductivity, κ: thermal diffusivity), on the surface absorption of the component (β_S), and on the beam diameter. For irradiation times t_I, long compared to the typical heat diffusion time w^2/κ, eqn [1] reduces to the asymptotic dependence

$$\Delta T \to \frac{\beta_s P}{\pi^{1/2} k w} \qquad t_I \to \infty \qquad [2]$$

Obviously, the temperature rise ΔT scales with the P/w for these long-time irradiation conditions compared to the short pulse regime, where the scaling of the temperature rise is proportional to P/w^2. For the application of optical components in high power cw- and long-pulse lasers, this P/w scaling law has to be considered carefully, because the power handling capability depends critically on the size of the irradiated area. Also, the onset of damage may be delayed in respect to the start of irradiation by time constants in the range of several 10 seconds.

In the course of the development of improved coating processes, optical absorptance could be reduced to very low levels in the near infrared and visible spectral region. For example, high reflecting mirrors for the wavelength of the Nd:YAG-laser with total losses below 1 ppm can be produced with

advanced ion beam sputtering processes nowadays. Therefore, another principal damage mechanism, which is based on inclusions or defects in the coatings, is often found in the near infrared spectral range. A detail of a damage site representing inclusion dominated damage induced by a pulsed Nd:YAG-laser in a coating of HfO_2/SiO_2 is shown in Figure 3. In the laser beam area numerous craters are located, which extend from circular voids with different diameters in a range below 1 μm. At the high magnification of the employed electron microscope, the multilayer structure of the coating system can be resolved indicating an origin of damage approximately in the same depth of the layer structure at the substrate surface. The underlying damage mechanism is initiated by inclusions or other defects with an absorption, which is significantly higher than for the surrounding thin film material. As a consequence, the inclusion is rapidly heated during the interaction with the laser beam and explodes removing the covering layer structure. This inclusion breakdown can be modeled by calculating the temperature behavior $T(r)$ at the radial position r of a sphere (radius g, thermal conductivity k_E, diffusivity κ_E) which is embedded in a medium with defined properties (see eqn [3]). The interaction of the sphere, which has a size in the range of the wavelength λ and a refractive index n', with the laser irradiation, is described by the Mie absorption coefficient Q. For the instant of damage, a critical temperature at the outer surface of the sphere ($r = g$) is assumed, which is often related to the melting point of the layer material

$$T(r) = \frac{3QI}{4\pi k_E g}\left[\frac{k_E}{3k} + \frac{1}{6}\left(1 - \frac{r^2}{g^2}\right) - \frac{2gb}{r\pi}\int_0^\infty \mathrm{d}y \exp\left(-\frac{y^2 t_p}{\tau_1}\right) \times \frac{(\sin y - y\cos y)(\sin ry/g)}{y^2\left[(c\sin y - y\cos y)^2 + b^2 y^2 \sin^2 y\right]}\right] \quad [3]$$

$$Q = \pi g^2\left[1 + \frac{2\exp(-8\pi n' g/\lambda)}{8\pi n' g/\lambda} + \frac{\exp(-8\pi n' g/\lambda) - 1}{(8\pi n' g/\lambda)^2}\right]$$

$$\tau_1 = \frac{g^2}{\kappa_E} \quad b = \frac{k}{k_E}\left(\frac{\sqrt{\kappa_E}}{\sqrt{\kappa}}\right) \quad c = 1 - \frac{k}{k_E}$$

Besides the thermal properties of the host material (thermal conductivity k, diffusivity κ), the properties and the statistic distribution of the inclusions have to be known for a quantitative description of inclusion breakdown.

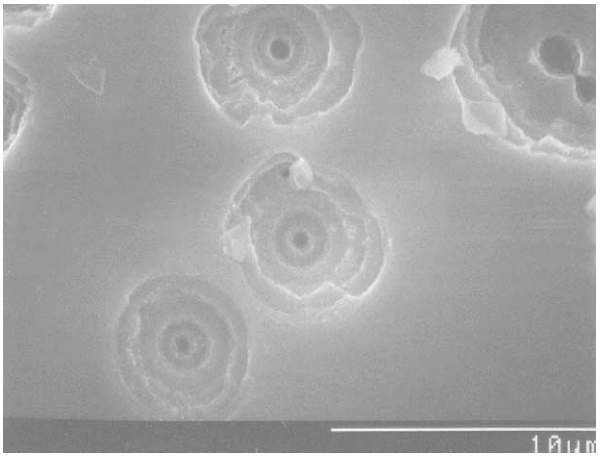


Figure 3 SEM-picture of damage site on a high reflecting coating of SiO_2/HfO_2. This inclusion-dominated breakdown mechanism was initiated at an energy density of approximately 130 J/cm^2 with a Nd:YAG-laser.

In contrast to the nearly deterministic behavior of absorption induced damage, defect dominated damage mechanisms often exhibit a statistical nature. Since the inclusions are not homogeneously distributed in the layer structure, a variation of the threshold value is observed for different damage sites on a single sample. The assessment of the inclusion parameters is extremely difficult in practice. Therefore, the inclusion model is mainly employed for the prediction of tendencies for the LIDT as a function of the beam parameters and the properties of the layer materials.

Different aspects related to the substrate polishing, cleaning, and deposition process have to be considered for an investigation in the various origins of inclusions or defects in a layer structure. Besides contamination of the substrate surface prior to the deposition process, particles may be generated during the deposition process by mechanical abrasion of material from moving parts in the plant or by delamination of material coated on the inner walls. Also, the deposition source may contribute to the formation of defects by sputtering of particles or droplets from the melt or the target. Especially in low energetic thermal deposition processes, defects are often the origin for the formation of nodules. These conical faults in the layer structure reduce the LIDT because they are less resistant against intensive laser radiation than the bulk of the coating system.

Besides models, which are based on the transformation of laser radiation into heat, direct electronic excitation has to be considered as a breakdown mechanism for short-pulse lasers. In these intrinsic breakdown models for dielectric materials, the

generation of a high electron density in the conduction band is assumed. This carrier generation can be explained by high field ionization of electrons in the valence band of the dielectric material. Many theoretical concepts introduce the avalanche effect as an important mechanism contributing to the generation of free electrons. Electrons excited by strong field ionization can interact with the electric field and accumulate excitation energy. By collisions with the lattice, this energy can be also transferred to the electrons in the valence band. Thus, when the energy of the free electrons exceed a critical value, other electrons can be excited to the conduction band, and a carrier generation like an avalanche arises. Another theoretical approach is the multiphoton model which is formulated on the basis of multiphoton absorption for the generation of free electrons. In most intrinsic models it is assumed, that laser-induced damage breakdown occurs when the electron density in the conduction band reaches a critical value of approximately 10^{21} 1/cm^3. At this high electron density, a plasma state is reached and the impinging radiation is coupled efficiently into the layer structure resulting in its destruction.

Intrinsic damage is often characterized by a deterministic damage behavior with a well defined threshold value which is also characteristic for damage in bulk materials. As a prominent example, damage induced by ultra-short pulses with a duration below 1 ps can be often attributed to intrinsic mechanisms, because the diffusion length of thermal effects can be neglected in comparison to the intensive interaction of the laser radiation with the electrons of the coating materials. In this case, the generation rate dn/dt of electrons in the conduction band can be described on the basis of the impact ionization rate and multiphoton excitation

$$\frac{dn}{dt} = \alpha n(t) q I(t) + \beta_m \left[q I(t) \right]^m \qquad [4]$$

In this model $I(t)$ is the power density of the laser radiation, and q is a correction factor representing interference effects in the coatings, which may result in a local power exceeding the power density of the radiation impinging on to the layer system. The electron excitation mechanisms are described by the impact ionization coefficient α and the m-photon absorption coefficient β_m. The relative contributions of both excitation mechanisms to laser-induced breakdown are depending on the pulse duration and the bandgap energy of the layer material.

In a rough approximation, the damage threshold increases with the bandgap energy of the materials as a consequence of the photon absorption coefficient β_m decreasing with its order m. For short pulse durations below a few 10 fs, multiphoton processes contribute significantly in the range of some 10% to the free-carrier generation. In most materials, avalanche ionization develops to the dominant generation mechanism for pulse durations above 100 fs. These tendencies could be demonstrated for single layers of selected oxide material deposited by ion beam sputtering on quartz substrates, which were tested in a pulse duration regime from 20 fs to 1 ps. According to the theoretical model, the expected increase of the damage also influencing values with the pulse duration was observed in the damage experiments.

In practice, the power handling capability of optical components is also often limited by imperfections like scratches, digs, and areas with high roughness on the optical surface. In most deposition techniques applied in optical technology, the layer system tends to replicate or even enhance surface imperfections of the substrate. In the application, additional weak points may be introduced by improper handling or contamination of the optical surfaces. If the contaminants are not removed by the impinging high-power laser radiation, they increase the surface absorption and act as initiation points for laser-induced damage. Voids, grooves, pores, or scratches reduce the power capability of the optical element, because they act as concentrators for the electric field.

Units and Scaling of Laser-Induced Damage Threshold

According to the theoretical models of laser-induced damage, the appropriate units of measurement for LIDT values are mainly given by the dominant damage mechanism and the irradiation time. For pulsed laser radiation and dielectric breakdown effects, the laser-induced damage threshold is usually reported in W/cm^2. In the case of absorption-induced damage or inclusion dominated breakdown and a pulse duration short compared to the thermal diffusion time in the layer structure, LIDT values are often scaled in J/cm^2. The unit of linear power density (W/cm) is indicated for the LIDT of long-pulse, cw-lasers or other sources, which induce a temperature increase in the optical component with rise times in the regime of seconds.

Laser-induced damage thresholds of optical thin films are critically dependent on the operation conditions of the applied laser system and on the design of the layer structure. For example, even the damage mechanism can change if the wavelength or the pulse duration is varied. Therefore, LIDT

values should be scaled with the irradiation parameters only for small intervals, in cases where the fundamental damage mechanism is clarified. In all other situations, the extrapolation of threshold values is extremely difficult and may lead to inaccurate results. Especially, an overestimation of the LIDT value holds severe dangers in practical applications, where the replacement of a damaged component and the repair of other parts impaired by the damage event causes high expenses. If the damaged components consist of toxic materials (e.g., ZnSe, GaAs, CdTe, ThF_4, chalcogenides, Be, Cr, etc.) severe health hazards may occur, and an expensive decontamination of the environment may be necessary. Therefore, the determination of the laser-induced damage threshold using the actual laser parameters under controlled environmental conditions is always recommended for unclear cases.

In the present state of research in laser-induced damage mechanisms, only few tendencies are generally accepted as confident for scaling of LIDT values. As a function of the pulse duration, laser-induced damage thresholds increase for longer pulses. For inclusion and absorption dominated breakdown, a $t_p^{1/2}$-law is often used in the pulse duration regime between 10^{-10} to 10^{-8}s. This dependency can be extended to other pulse regimes up to pulse lengths in the ms-range, if the exponent is replaced by a value between 0.5 to 2. In respect to the laser wavelength, a decrease of the LIDT value with decreasing wavelength is observed for most materials and operational conditions. Investigations in the influence of the beam diameter have been performed by many research groups, indicating a decrease of the LIDT values for increasing beam diameters. Especially for inclusion dominated breakdown, the event of damage for a certain laser irradiated site will be dependent on the distribution of inclusions at that position. If the beam size is small, the probability for interrogating a defect vulnerable to damage is low. By increasing the spot diameter, this probability will asymptotically reach unity, because at a certain beam diameter, at least one defect will always be covered by the beam. Therefore, the onset of laser-induced damage is not dependent on the beam diameter for inclusion-dominated breakdown, which is often encountered in conventional optical coating systems. Another special case is the scaling of the cw-damage threshold with the beam diameter. Since the cw-LIDT is expressed in linear power density, the power handling capability increases more slowly with the beam diameter for the cw-lasers than expected from the normal pulsed operation. For example, extrapolating the LIDT value of 100 W at a beam diameter of 1 mm for a laser mirror to a beam diameter of 10 mm, results in a maximum power load for the component of 1 kW according to the correct P/w-law. The threshold power would be extremely overestimated to approximately 3.2 kW, if the general P/w^2-dependence is applied.

Measurement of Laser-Induced Damage Thresholds

As a consequence of the complicated relation between the laser damage mechanism and a broad spectrum of thin film properties and laser parameters, laser-induced damage threshold measurements have to be performed under well-defined conditions. To investigate the comparability of LIDT measurements, an extended international round-robin experiment has been conducted on coated optics for the wavelength of 1.064 μm at the beginning of the 1980s. This experiment indicated the need for a clearly specified LIDT-measurement procedure, and conceptual work was initiated to develop a corresponding ISO-Standard series. During recent years, an International Standard (ISO 11254) has been adopted covering testing conditions relevant for most typical laser applications. In the first part of ISO 11254 1 on 1-testing of optical surfaces, in respect to laser damage, is described. The fundamental approach of the standard measurement procedure is illustrated in Figure 4. A laser source operating in transversal and longitudinal single mode is employed for the irradiation of the sample surface. The beam

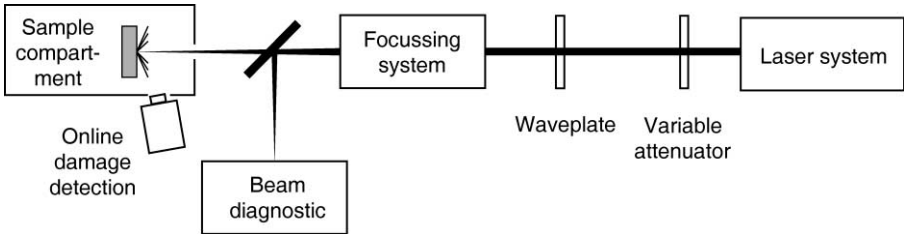


Figure 4 Fundamental setup for the measurement of laser-induced damage thresholds.

parameters of the laser are assessed by a beam diagnostic system which monitors the spatial and temporal profile as well as the energy of the laser radiation in the target plane. In order to achieve the high energy density levels necessary to destroy the surface of the specimen, a well characterized focusing system is installed. For adjustment of the laser energy in the target plane, an attenuating system is employed. In the 1 on 1 measurement protocol, each site on the sample surface is subjected to a single laser pulse once only. In a test sequence, the test surface is examined by pulses of different energies, covering low values without damage and high values causing damage. After the test, the specimen will be inspected with a Nomarski interference contrast or a darkfield microscope at a magnification of 150× or higher to identify the damaged sites. For the evaluation of the damage test, the damage probability method is recommended. In this evaluation scheme, the ratio of the number of damaged sites to the total number of sites objected to a certain energy or power level is interpreted as the damage probability. The plot of these damage probability values as a function of energy or power, which is called the survival curve of the optical component, provides an insight into the damage mechanisms involved. The damage threshold is given by the highest quantity of laser radiation for which the extrapolated probability of damage is zero. A typical example for a 1 on 1 damage test at the wavelength 1.064 μm and the extrapolated LIDT value is depicted in Figure 5.

In most catalogs of optics manufacturers, the 1 on 1 LIDT values are used to illustrate the power handling

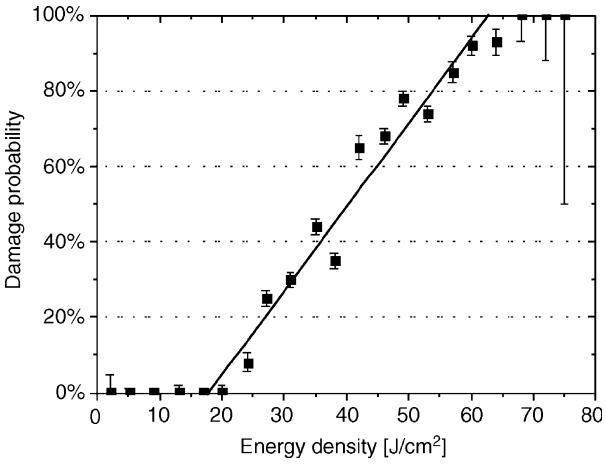


Figure 5 Survival curve of an anti-reflective layer system (two layer V-coating, SiO_2/Ta_2O_5) on fused silica substrate for the Nd:YAG-laser wavelength. The measurement was performed at a wavelength of 1.064 μm, a beam diameter of 420 μm, and a pulse duration of approximately 15 ns, respectively.

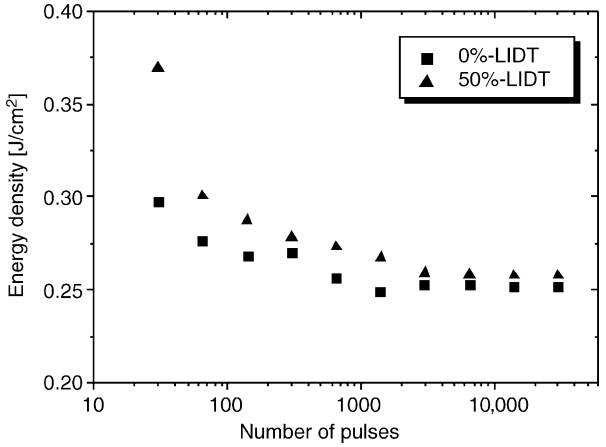


Figure 6 Characteristic damage curve of a high reflecting mirror of SiO_2/TiO_2 for fs-laser systems. The layer structure was tested with a high repetition (1 kHz repetition rate) fs laser system operating at a pulse duration of 150 fs.

capability of their products, even though 1 on 1 data are of limited importance for practical applications, where an optical component is always subjected more than one time to a laser beam. This customary situation is covered by Part 2 of the ISO-LIDT-Standard which describes a damage test procedure for a series of pulses (S on 1-tests). For an assessment of the reliability, the concept of the characteristic damage curve is introduced by this standard. This curve is deduced directly from the S on 1 test data by plotting the energy density for a selected damage probability as a function of the number of pulses (see Figure 6). By an extrapolation of the characteristic damage curve to high pulse numbers in the order of 10^9 to 10^{12} shots, the lifetime of the optical component can be roughly estimated. For a certification of optical components in respect to their power handling capability, a third part of ISO 11254 is under development, which is concentrated on different testing protocols of a defined surface fraction at power levels expected in the application. The fundamental approach of these tests is a simulation of conditions at the upper limits of operation parameters encountered in practice.

Optical Coatings for High Power Lasers

For the development and application of coating systems with high LIDT values, several major aspects have to be considered. In the first approach, appropriate materials and processes have to be selected which deliver coatings with sufficient power resistance. In many studies, a correlation of the LIDT values of the constituent single layers to the power

handling capability of a layer system was observed. Therefore, single layers of potential deposition materials were investigated at the wavelengths of prominent laser systems from the VUV- to the FIR-spectral region. For specific deposition processes, often a clear ranking of the coating materials in respect to their power handling capability could be demonstrated. An example for a Nd:YAG-laser with a pulse duration of 15 ns is illustrated by Table 1. With the exception of the ZrO_2-layer, which exhibited severe inhomogeneity and a high inclusion density, a correlation can be assumed between the LIDT value and the melting point of the bulk materials under these operation conditions. This dependency can be attributed to absorption-induced damage as well as to the inclusion model, and therefore, the damage morphology has to be studied in each case to identify the dominating damage mechanism. In the fs-regime or the UV/VUV-spectral range, often a dependence of the power handling capability on the bandgap of the materials is found. According to the relationships of the avalanche and other dielectric breakdown models, materials with highest bandgaps are frequently encountered at the top of an LIDT ranking.

For the influence of the coating process on the laser-induced breakdown of optical coatings, no clear tendencies can be detected. For example, Ion Beam Sputtering (IBS), which is considered as the deposition process for coatings with extremely low losses and contamination, cannot always surpass conventional thermal deposition processes in laser stability. In general, ion or plasma assisted deposition techniques produce coatings with lower LIDT values than conventional coating processes for most operation conditions. Sol-gel processes, which were developed for the deposition of removable coatings on large optics in laser fusion, can reach superior power handling capabilities. In general, as a consequence of the complicated relation between the production parameters and laser-induced breakdown, production processes for optical coatings have to be optimized separately for different wavelength regimes and irradiation conditions.

Table 1 Laser induced damage thresholds of selected single-layer coatings in relation to the melting points of the corresponding bulk materials. In the last column, LIDT values of anti-reflective coatings (AR-Coating) composed of the high index material and SiO_2 are compiled to illustrate the correlation between single LIDT values and the damage threshold of layer systems

Material	*Melting point (°C)*	*LIDT value (J/cm²) single layer*	*LIDT value (J/cm²) AR-coating*
TiO_2	1775	13 ± 1	14 ± 1
Ta_2O_5	1918	28 ± 2	32 ± 5
HfO_2	2758	41 ± 3	46 ± 6
ZrO_2	2700	34 ± 4	28 ± 1
Al_2O_3	2072	39 ± 1	
SiO_2	1723	34 ± 7	

Besides the fundamental production parameters, the design of the coating system is of similar importance for the achievement of high LIDT values, because damage is directly driven by the electric field strength in the layer structure. A standing wave field pattern is depicted in Figure 7 for a high reflecting stack of Ta_2O_5 and SiO_2. The design consisting of layers with an optical thickness of 1 QWOT is typical for most standard laser mirrors. The power density reaches extreme values always at the interfaces between the layers and even exceeds the incoming irradiation power density (100%). Interfaces between the layers can be considered as weak points in a coating system, because additional contamination may occur during the switching of the material in the production process. Also, the adhesion between the adjacent layers may be reduced, and mechanical stress may be built up by the different materials. To improve damage thresholds of laser mirrors, the thickness of the first few layer pairs can be adjusted to shift the points of maximum field strength into the bulk of the layer with higher damage resistivity (see Figure 8). Also, a thick layer can be attached to the system in order to stabilize the outer layer pair in respect to thermal or mechanical stress. Another technique, which can be applied to enhance the stability of the interface is the codeposition of materials, resulting in a gradual interface with a mixing zone of materials between the layers. This region of codeposited material exhibits a higher resistance against mechanical stress resulting in an improvement of LIDT values of up to 20%.

The effect of the internal electric field strength distribution on the power handling capability of a layer system should be most apparent for intrinsic damage mechanisms, which are dominant in the ultra short pulse regime below 1 ps. According to the fundamental model (see eqn [4]), a direct relationship of the damage threshold to the maximum field strength value within the layer system is expected. As an example, an investigation in the threshold behavior of ion beam sputtered coating systems with different field strength values is illustrated in Figure 9. In this experiment, the maximum field strength in the last low index layer of SiO_2 has been adjusted to factors between 0.4 and 1.6 of the impinging field strength (see upper diagram in Figure 9) by depositing different designs on a basic 1-QWOT layer stack of TiO_2/SiO_2. In the lower part of Figure 9, the S on 1 LIDT values, measured with a fs-laser, are depicted for selected pulse numbers N. For all layer systems,

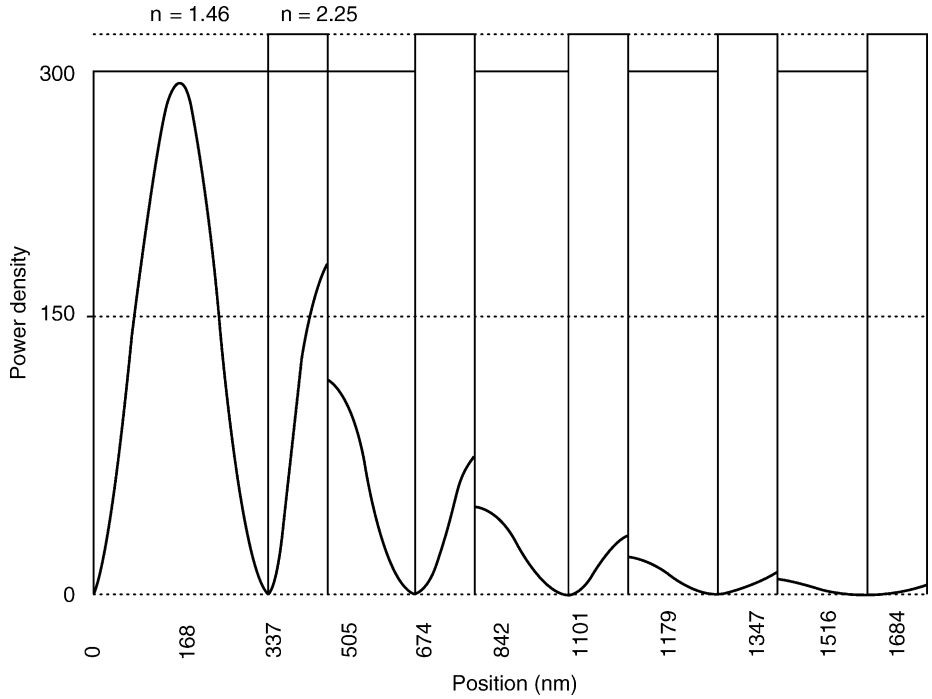


Figure 7 Distribution of the power density in a high reflecting layer stack of SiO_2/Ta_2O_5. The first nine layers of the system and a final thick stabilizing layer of SiO_2 next to the air interface are depicted. The incoming energy density is calibrated to 100%. Reproduced with permission from Ristan D, Ebert J and Welling H (1989) Optische Beschichtungen für Hochleistungslaser. *Laser und Optoelektronik* 21(4): S.53.

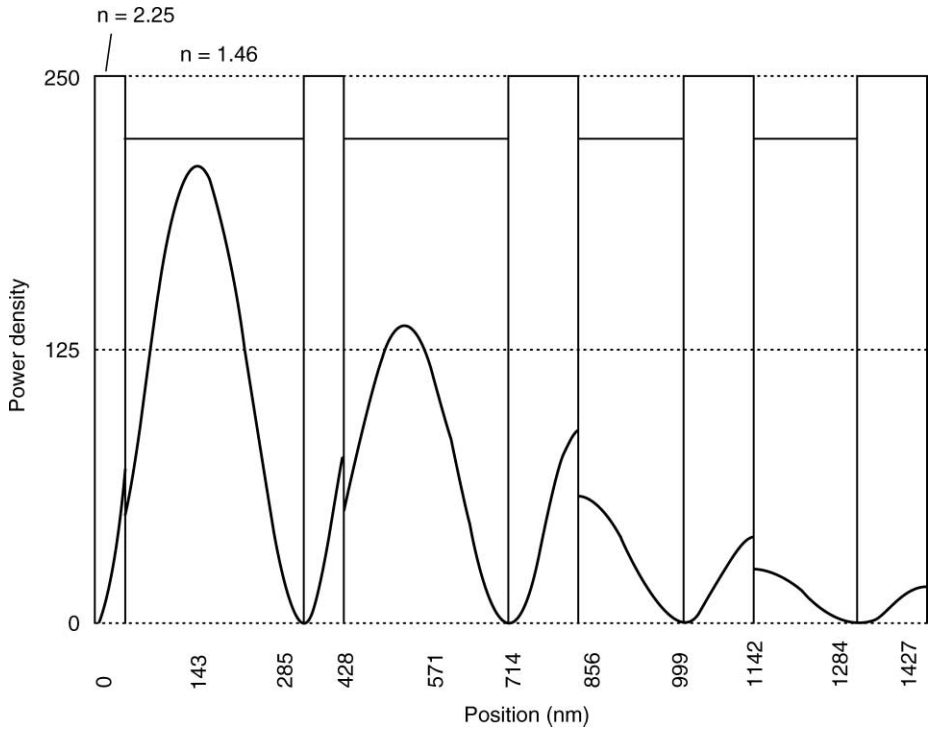


Figure 8 Distribution of power density in a high reflecting layer stack of SiO_2/Ta_2O_5. The first nine layers of the system, which is designed for reduced power density at the interface in the first layer pair are depicted. Reproduced with permission from Ristan D, Ebert J and Welling H (1989) Optische Beschichtungen für Hochleistungslaser. *Laser und Optoelektronik* 21(4): S.53.

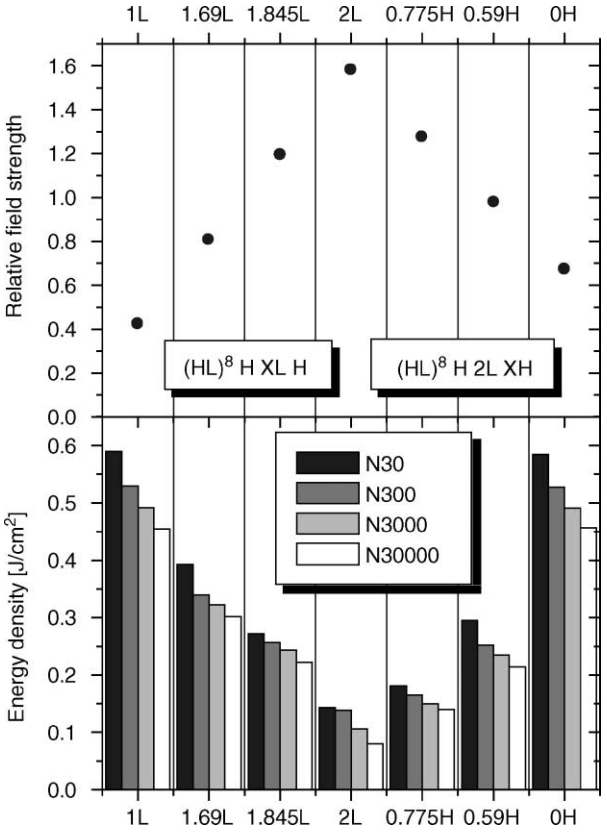


Figure 9 Damage thresholds (lower diagram) and internal field strength relative to the field strength of the impinging wave (upper diagram). The damage thresholds were measured with an ultrashort pulse laser operating at a repetition rate of 1 kHz, a pulse duration of 150 fs, and a beam diameter of 100 μm on the sample surface. Damage thresholds for selected numbers between $N = 30$ to $N = 30\,000$ of pulses are indicated by columns. The plotted maximum field strength in the last SiO_2-layer is adjusted on the basis of design variation.

which were deposited in separate deposition runs, a strong correlation between the maximum field strength and the laser threshold is observed. This experiment clearly demonstrates the role of the electric field strength distribution in high-power coatings and the potential of advanced thin film design strategies.

Besides the properties of the coatings, the quality of the substrate has to be considered in respect to surface imperfections and to the polishing procedure. For example, substrates for the production of high-quality optics with enhanced power handling capability in the VIS/NIR-spectral range should be polished to a surface roughness of less than 1 nm_{rms} with a surface imperfection value $5/1 \times 0,010$ according to ISO 10110. For an illustration of the effect of the polishing grade on the LIDT values, selected results are compiled in Table 2 for substrates of BK7-glass and high reflecting mirrors.

Table 2 Laser-induced damage threshold values of samples polished using compounds of different grain size. Besides the LIDT values of the uncoated surfaces, also data for anti-reflective coatings of selected materials on these surfaces are summarized

Coating type	*LIDT value (J/cm^2) polishing type I*	*LIDT value (J/cm^2) polishing type II*
Bare surface	76 ± 1	68 ± 12
Ta_2O_5/SiO_2	44 ± 2	38 ± 1
HfO_2/SiO_2	47 ± 16	35 ± 5
Nd_2O_3/MgF_2	34 ± 6	24 ± 8

Substrates with the polishing Type I and Type II were processed with powder of different grain diameter (2 μm and 3 μm) in conventional pitch polishing. For the LIDT values of the bare substrate and the coatings, a clear relation to the surface roughness can be observed. Especially for optical coatings with significant transmittance, subsurface damage in the substrate has to be taken into account. As a consequence of the chemical and mechanical interaction of the surface with the polishing compounds, impurities and dislocation are introduced in the surface structure resulting in a reduction of the damage threshold.

Summary

In the course of the rapid development of laser technology, a large background in high power optical coatings had been built up during the last three decades. Nowadays, the corresponding experience in the production of high power coatings is mainly located at industrial companies and a few research institutes, which are also involved in the characterization of optical coatings. For an illustration of the present state in thin film technology, damage threshold values of advanced optical components are summarized in Table 3 for laser systems and operation conditions often applied in industrial production environments.

Besides the spectacular experiments in laser fusion, isotope separation, and fundamental physics, industrial applications of lasers in material processing, medicine, information technology, and semiconductor lithography are considered as major pacemakers for the progress of high power optics. Therefore, trends in these technology fields will dominantly govern the future development of thin film technology. For example, in semiconductor lithography at a wavelength of 157 nm, which would open the way to feature sizes well below 100 nm, optical coatings with extended lifetimes are still on rank 6 of the list of challenges to achieve an effective lithographic production facility. In next generation lithography, optics

Table 3 Laser-induced damage threshold values of selected optical coating systems for laser applications (Types: HR: high reflecting mirror, AR: antireflective coating, th: thermal evaporation, IBS: ion beam sputtering)

Laser system wavelength	*Type*	*Laser-induced damage threshold ISO 11254*
193 nm, ArF-excimer	AR/th	1–2 J/cm^2 (1 on 1, 20 ns)
	HR/th	2–4 J/cm^2 (1 on 1, 20 ns)
248 nm, KrF-excimer	AR/th	10 J/cm^2 (1 on 1, 30 ns)
	HR/th	>20 J/cm^2 (1 on 1, 30 ns)
	HR/IBS	>3 J/cm^2 (1 on 1, 30 ns)
1,064 μm, Nd:YAG	AR/th	>60 J/cm^2 (12 ns, 0,25 mm)
	HR/th	>100 J/cm^2 (12 ns, 0,25 mm)
	HR/IBS	>80 J/cm^2 (12 ns, 0,25 mm)
10,6 μm, CO_2-Laser	AR/th	>20 J/cm^2 (100 ns, 1,4 mm) >2 kJ/cm^2 (1,2 ms, 250 μm) >3 kW/mm (cw, 100 μm)
	HR/th	>25 J/cm^2 (100 ns, 1,4 mm) >2 kJ/cm^2 (1,2 ms, 250 μm)

of even smaller wavelengths, around 13 nm, have to be optimized to achieve the 60 nm node. Ultra-short pulse lasers gain importance as innovative tools for material processing, laser medicine, and biology, as well as the analysis and control of chemical reactions. For the development of fs lasers needed in these applications, special high-power broadband coating systems are required. In addition to high LIDT values, these coatings have to fulfil demands with respect to their group delay dispersion for the compensation of dispersion effects in the laser systems. Other challenges have to be expected from applications, where a combination of the high-power handling capability with additional properties has to be achieved. As a typical example, high-power laser coatings with improved mechanical or chemical stability for applications in laser medicine can be considered.

List of Units and Nomenclature

dn/dt	generation rate of electrons in the conduction band
g	radius of a spherical inclusion; (cm)
$I(t)$	power density of laser radiation,
k	thermal conductivity; (W/(cm °C))
k_E	thermal conductivity of a spherical inclusion; (W/(cm °C))
LIDT	laser induced damage threshold (J/cm^2; W/cm^2; W/cm)
n'	refractive index of a spherical inclusion; (1)
P	output power of a laser; (W)
q	correction factor representing interference effects in coatings
Q	Mie absorption coefficient of a spherical inclusion;
QWOT	quarter wave optical thickness: unit for the thickness of the layer
t_I	irradiation time for a cw-laser; (s)
$T(r)$	temperature in a spherical inclusion at the radial position r; (°C)
w	diameter of a laser beam with Gaussian profile; (cm)
α	impact ionization coefficient; (1)
β_m	m-photon absorption coefficient; (1)
β_S	surface absorption of an optical component; (1)
ΔT	temperature rise in the center of an irradiated circular spot; (°C)
κ	thermal diffusivity; (cm^2/s)
κ_E	thermal diffusivity of a spherical inclusion; (cm^2/s)
λ	wavelength of a laser; (cm)

See also

Optical Coatings: Thin-Film Optical Coatings.

Further Reading

Bloembergen N (1973) Role of cracks, pores and absorbing inclusions on laser damage threshold of transparent dielectrics. *Applied Optics* 12(4): 661–664.

Genin FY and Stolz CJ (1996) *Morphologies of laser induced damage in hafnia-silica multilayer mirror and polarizer coatings*, Third International Workshop on Laser Beam and Optics Characterization. In: Morin M and Giesen A (eds). *Proc. SPIE 2870*, 439–448.

Guenther AH, *et al. Proceedings of the Symposium on Laser Induced Damage in Optical Materials, NBS (NIST) Special Publication 1969–1989*, SPIE vols. 1989–2003. CD-Rom version of collected papers from 1969 to 1998 available: Washington: SPIE.

Guenther KH, Ebert J, Kiesel E, *et al.* (1984) 1.06-m laser damage of thin film optical coatings: a round-robin experiment involving various pulse lengths and beam diameters. *Applied Optics* 23: 3743.

ISO 11254 (2001) *Test methods for laser induced damage threshold of optical surfaces*. Part 1: 1 on 1-test, 2000, Part 2: S on 1 test.

ISO 11254-3 (2002) *Laser and laser-related equipment – Determination of laser-induced damage threshold of optical surfaces – Part 3: Assurance of laser power handling capabilities*. Draft Standard. International Organization of Standardisation.

Jasapara J, Nampoothiri AVV, Rudolph W, Ristau D and Welsch E (2001) Femtosecond laser pulse induced breakdown in dielectric thin films. *Physical Review B* 63(4): 045117.

Puttick K, Holm R, Ristau D, *et al.* (1997) *Continuous wave CO_2 laser induced damage thresholds in optical*

components, Proceedings of the Symposium on Laser Induced Damage in Optical Materials. SPIE vol. 3244, 188–197.

Rudolph W, Mero M, Liu J, *et al.* (2003) *Femtosecond pulse damage and pre-damage behavior of dielectric thin films*, Proceedings of the 34th Annual Symposium on Optical Materials for High Power Lasers. SPIE vol. 4932, 202–215.

Walker TW, Vaidyanathan A, Guenther AH and Nielsen P (1979) *Impurity breakdown in thin films*, Proceedings of the Symposium on Laser Induced Damage in Optical Materials. NBS Special Publication, No.568, 479–495.

Wood R (1990) *Laser Damage in Optical Materials*. Bristol and New York: Adam Hilger. Reprint with corrections.

Optical Black Surfaces

S M Pompea, National Optical Astronomy Observatory, Tucson, AZ, USA

S H McCall, Stellar Optics Research International Corporation (SORIC), Thornhill, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Black surfaces play an important role in many, if not most, optical systems. Optical instruments and systems of all types often rely on black surfaces to help minimize the effect that stray or off-axis light can have in degrading optical system performance. The strategic selection and placement of appropriate black surfaces can often limit significant detrimental stray light effects in an optical systems, thereby dramatically improving system performance. The high emissivity of black surfaces also gives them an important role in the design and construction of black bodies, calibration surfaces, and radiator surfaces.

The characterization and selection of black surfaces is an important field of optics and optical system design. However, the selection of black surfaces is a specialized undertaking requiring careful study to ensure the proper selection of particular surfaces for each optical system. A black surface that works well in one application may not be at all appropriate for a different application, where different systems performance goals are desired. The wavelength of operation of the system is an important consideration since a surface which has low reflectance at one wavelength band may have different reflective properties at another wavelength or range of wavelengths. Specialized measuring devices called goniophotometers are often used to characterize the reflectance and scatter of black surfaces for a given wavelength or set of wavelengths. Goniophotometric measurements can be used to create a specialized functional description of reflectance and scatter off a surface, called the Bidirectional Reflectance Distribution Function (BRDF) or Bidirectional Scatter Distribution Function (BSDF). In practice, having the BRDF or BSDF of a surface is extremely valuable for surface selection and is more useful than other reflectance and scatter measurement measurements or descriptions. Comparing the measured BRDF or BSDF of several surfaces helps to distinguish the general optical properties of these surfaces and allows a direct comparison of the value of these different surfaces in an optical system as they contribute to overall system performance.

The BRDF distinguishes, at a particular wavelength, how a surface specularly reflects or how it scatters light in different directions when incident light from a particular direction relative to the surface (at a given angle of incidence) interacts with the surface. By understanding the BRDF of a surface, the optical designer can understand how the surface behaves in a given system location when illuminated from a specific direction. The BRDF of a black surface is also important because it can be used as a mathematical function in stray light analysis programs to predict system performance. The BRDF and other properties of surfaces can be input into stray light codes so that the propagation of light from black surfaces can be modeled through Monte Carlo-based propagation models. Using these models, the stray light can be directed through the placement of black surfaces into directions less likely to degrade system performance. Thus a measure of stray light control can be gained through absorption by black surfaces or by redirection of the stray light into less critical directions.

Black surfaces may be considered a subset of a class of surfaces known as spectrally selective surfaces. The term ‘spectrally selective surface’ indicates that the spectral properties of many surfaces are different when examined in various spectral regions. For applications such as the design of blackbodies, space radiators, and baffles, surfaces with unique spectral emissive and reflection properties are needed. Large databases of spectrally selective surfaces are available to allow the optical designer flexibility in choosing surfaces for applications where specific optical reflectance and optical scatter properties

are desired. Spectrally selective surfaces are important for applications that range from solar energy to telescope design. In some applications, it is desirable for a surface to be highly absorbing at one wavelength and more reflective at another, rather than to be highly absorbing of light over a wide range of wavelengths. An analysis of the spectral properties of a large number of surfaces allows the optical designer to choose surfaces optimized for a particular application.

Black surfaces come in many types, but may be generally characterized as being more shiny (specular) or more diffuse (Lambertian) in their reflective and scattering properties, depending on the form of their BRDF curve. Surfaces are often described in computer models as a combination of their respective specular and Lambertian properties for a given wavelength. Many black surfaces have significant reflectivity at some region of the electromagnetic spectrum, even if they are highly absorbing at one wavelength. Some black surfaces are constructed to be remarkably nonreflecting over a range of wavelengths; these surfaces are often used as baffle surfaces for telescopes and instruments that must operate over a wide range of wavelengths. No black surface can work well in all optical systems. Optical designers must understand the specific optical properties of a black surface to determine if the surface can work in their particular system. To characterize a surface, it is helpful to know at least the BRDF of the surface at the wavelength of interest as well as basic materials properties of the surface. The specific BRDF at the most appropriate wavelength for the system can be used with the optical system stray light model to determine if the surface can fulfill the designer's, expectations of it for the system performance. The placement and choice of black surfaces in an optical system are often critical decisions with highly significant effects on overall system performance.

For this reason, optical designers rely greatly on databases of the reflection and scatter properties of black surfaces in order to make the right choice of surface. These databases have been created to provide the optical engineer with the information needed to understand a wide variety of candidate surfaces and select the most appropriate ones. These databases not only describe the optical properties of surfaces but also elucidate material properties so that the selection process can examine the heat resistance, outgassing rates, fade resistance, durability, and a myriad of other surface or material properties that may be relevant.

The reflectance and scatter data on black surfaces are often taken using a variety of reflectance-measuring instruments and the data format often varies greatly. When published and unpublished data sets are compiled, they are often confusing and sometimes even contradictory. This has made the comparison of surfaces a difficult proposition. In recent years, there have been successful attempts to gather existing large amounts of existing data so that comparisons among black surfaces can be made in a data-rich environment so that anomalous data can be ignored. A significant amount of reflectance data on black surfaces has been created by private or governmental organizations such as NASA and NIST – the US National Institute for Standards and Technology. These data now appear in databases of BRDF and reflectance measurements. Designers are relying increasingly on databases of BRDF surfaces rather than the individual measurements that can be found in the optical physics and applied optics published literature.

Uses of Black Surfaces

Table 1 lists some of the major areas in which black surfaces or black coatings may be used in instruments and optical systems. The optical properties of desirable surfaces may be different for the specific categories in the table. Plausible surfaces for these different applications can be found in materials databases such as the SOLEXIS™ database which contains a useful compilation of data from private

Table 1 Possible uses for black surfaces and black coatings in optical systems

Apertures	Baffles	Barrels	Calibration screens	Choppers
Cold shields	Detectors	Dewars	Diffusion screens	Domes for telescopes and instruments
Enclosures and curtains for rooms and instruments	Integrating sphere enclosures	Lens edges	Light traps	Mounts
Radiators	Radiometers	Solar collector absorbers and concentrators	Standards and references of reflectance	Targets

and governmental data compilations from the US, Canada, and the former Soviet Union.

Selection of Optical Black Surfaces

The selection of appropriate black surfaces for a given application becomes progressively harder as the number of optical requirements (e.g., reflectivity, scatter, emissivity, etc.) and materials requirements (e.g., stiffness, coefficient of thermal expansion, etc.) increase. Because of this, the selection process for specialized surfaces for space applications can be quite involved, since there are a large number of criteria, each of which may be regarded as an essential requirement.

The optical system designer must consider at least the following aspects of the system in selecting a surface:

- the wavelength(s) and waveband(s) of interest;
- the position of the surface in optical system;
- the general robustness required of the surface;
- the environment of the optical system;
- the cost including inspection, handling, and assembly; and
- the production and delivery schedule, including spares.

The selection process must carefully address the system needs as well. Poor decisions are often made when only a cursory study of potential black surfaces is made or when databases and the literature on the subject are neglected. Therefore, an important part of the selection process is the management of the timing of the selection, in order to give the designer time to access available information and to influence the system design in a timely manner.

If black surfaces are used for stray light control, then the use of these surfaces must be planned at the beginning of the optical system design process. If the placement of black surfaces and the array of potential surface treatments are examined early in the program, the optimal choices can often be determined in a careful manner that enhances the system performance. If the selection is saved until the very end of the program, the choice of surfaces will be more limited and improvements in system performance may be minimal, even with the most excellent black surfaces. This is because the location of the surfaces in the system often proves to be as critical as the choice of surface. If the surface is in the wrong location, the best optical black may not be able to improve the stray light performance. For best results, the layout of the system and the placement of baffles must be considered first, before the baffle surfaces are selected.

An optimal flowchart of good system design for black surface selection is given in Figure 1. Figure 2 describes a work flowchart where black surface selection is neglected until the end of the program, often with dire consequences for system performance, cost, and schedule.

For space-based systems, the effect of atomic oxygen in a low-earth orbit environment is an important consideration. For space systems, the designer must also examine how black surfaces may be affected by or contribute to contamination and outgassing. Other considerations are the generation of particles during launch, the effect of solar

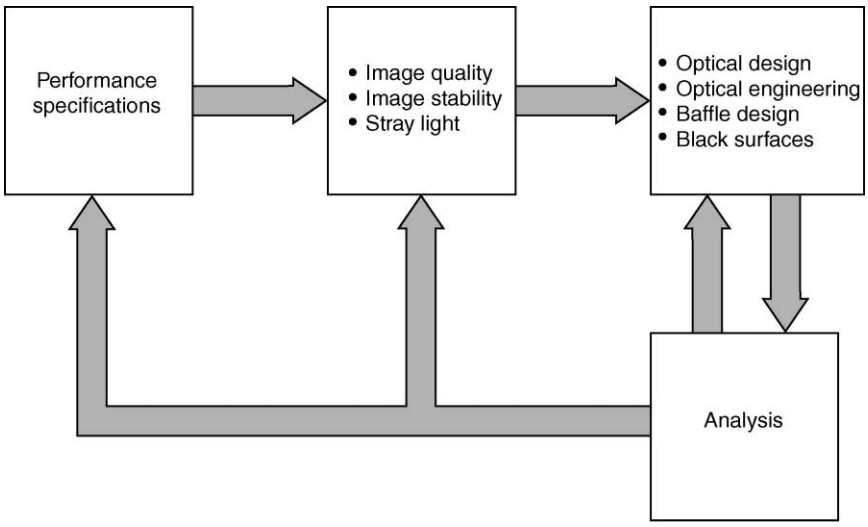


Figure 1 Diagram of the system design process which treats black surface selection as an important part of basic system design, to be done early in the design process. This gives the designer an opportunity to conduct system tradeoffs to optimize system performance.

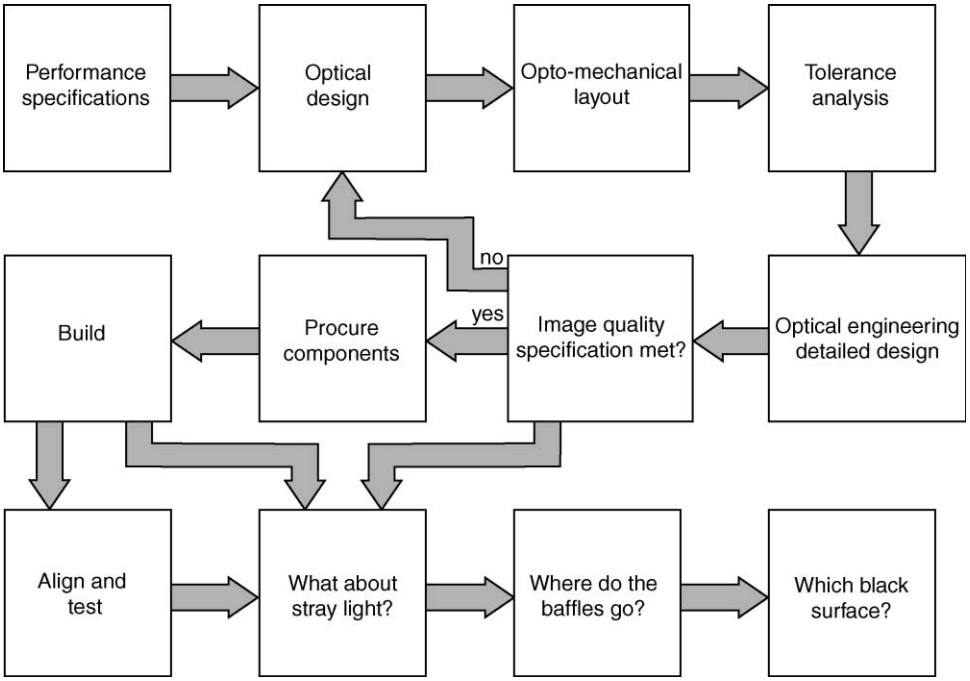


Figure 2 Diagram of a poor system design process. This process addresses the selection of black surfaces late in the system design process, where selection is significantly more difficult, expensive, and risky. Even with a late selection, there may be severe impacts on schedule.

ultraviolet radiation on black surfaces, and how spacecraft charging may affect conductive, dielectric, or insulating black surfaces. Since black surfaces may have a large surface area in an optical system, the designer must ensure that the black surfaces do not contribute significantly to the risk of system failure through some degradation process that may contaminate other optical surfaces. For space-based systems, data on surface outgassing and the past performance of a surface in a specific space environment is essential.

There are a variety of resources available to assist in the selection of black surfaces. The best databases allow the user to create 2-dimensional and 3-dimensional BRDF plots for extensive lists of potential black surfaces. For most surfaces, with no preferred directionality in the structure of the surface, a 2-dimensional BRDF plot is adequate. Most of these databases provide information on a wide variety of spectrally selective materials including black, white, and reflective surfaces for thermo-optical applications for ground and space-based instrumentation. The BRDF plots can be examined by the user or fed into existing, commercially available optical design and analysis software packages. The results of these models indicate whether the chosen surfaces give adequate performance or whether more specialized coatings are necessary to achieve the desired level of performance.

A significant problem for many designers is that the formulations of many paints and other surface treatments often change. Sometimes the name of the coating remains the same even as the optical properties change, due to a change in the compounds or methods used to prepare the coating. Sometimes the name changes with or without a change in the formulation. These often unpublished changes increase the risk of using older plots from the published literature as definitive optical modeling data. Experts in the optical blacks area continuously investigate these changes and their implications and can advise optical designers of the implications of recent changes in the paint and coating arena. These experts are often aware of situations where coating failures have occurred or where mission performance was degraded by unanticipated changes in coatings, even if reports of this incidents were never published.

In summary, the selection process, by its very nature, must be a deliberate and careful effort to examine the best surface options. Some of the better surfaces which were considered laboratory standards (e.g., 3M Black) are no longer easily available and replacements must be found. For the most systematic decision-making process, access is needed to the best optical black surface databases, often commercial databases with hundreds to thousands of BRDF curves on several hundred materials. The use of

these optical surface databases can substantially reduce the costs, time, and scope of a measurement program, as well as streamline the selection process. Additionally it allows the designer to consider materials used successfully in other programs which may not be familiar to the designer. Finally, the designer may avoid using materials that have been rendered unreliable through changes in the manufacturing process or in the coating chemistry.

Characterization of Black and Spectrally Selective Surfaces

Optical Characterization

An understanding of black surfaces relies heavily upon an understanding of the meaning and use of the BRDF. While the optical properties of surfaces can be characterized through specular and diffuse reflectance measurements, the use of BRDF measurements has become extremely important. BRDF measurements are equally adept at characterizing highly specular and highly diffuse surfaces and the myriad of surfaces with both characteristics. The BRDF is often used in general to describe the angle-dependent optical scatter from these surfaces. This is important as the angular distribution of scatter can be used to calculate if the scatter from a black surface will be a limiting factor in the resolution or noise level of the system. BRDF measurements can even be used in computer graphics visualization programs that illustrate what a surface may look like under various illumination conditions. For example, the BRDF is used in computer graphics models to render a certain color for an automobile or a countertop in a kitchen. For some black surfaces, the angular distribution of scatter can be used to calculate a variety of surface parameters such as the surface roughness, given certain assumptions about the surface. For opaque black surfaces, BRDF measurements have proved to be one of the best overall descriptors of the optical properties of the surface.

BRDF is defined as the ratio of scattered radiance (watts/(cm^2 sr)) to surface irradiance (watts/cm^2); its units are inverse steradians. Radiance is used in the definition, in order to make the BRDF independent of the parameters of the measuring instrument, such as the detector aperture and distance to the detector. BRDF measurements are made by a variety of instruments for a number of wavelengths. For example, BRDF measurements of surfaces have been made in wavelengths ranging from the vacuum ultraviolet to the far infrared. The most common BRDF measurements made on black surfaces are made at laser wavelengths of 0.6328 and 10.6 micrometers. However, these measurements may be of limited value if the system is operating at substantially different wavelengths, where the surface's BRDF is significantly different.

To make a BRDF measurement requires a light source, a sample mounted and illuminated by that source, a receiver (detector) to measure the scatter from the sample, and the computer/electronics package to accurately record the detection of the scattered light. Additionally the sample and/or light source must be able to be moved so that the sample can be illuminated from nearly every angle, and scattered light from the surface can be detected in a hemisphere around the sample. The geometry of the BRDF measurement is illustrated in Figure 3.

To make a BRDF measurement on black surfaces, appropriate sources, detectors, and low-scatter optics must be used. This becomes more difficult for wavelengths less than about 0.25 micrometers. At longer, infrared wavelengths, diffraction effects become important. Despite the instrumental difficulties, it is often better to measure the amount of scatter at a specific wavelength λ and angle θ_s, rather than try to predict it from theoretical models. This is because the extrapolation to new angles or wavelengths using scaling theories is not always accurate or even possible. Methods for wavelength scaling (extrapolating from BRDF data for one wavelength to data for any other wavelength) or angle scaling (where data taken at particular incident and scatter angles are extrapolated to other angles) will not be discussed as BRDF curves are now widely available for many surfaces at a variety of wavelengths and angles.

The definitions of BRDF are given in Table 2.

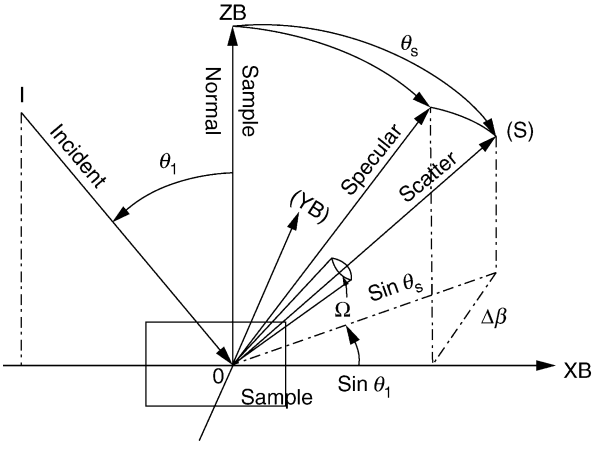


Figure 3 Geometry of BRDF measurement and the angle conventions accepted by the American Society for Testing Materials (ASTM). The plane of incidence is the I-0-ZB plane and the scatter plane is the S-0-ZB plane.

Table 2 BRDF term definitions

Bidirectional reflectance distribution function BRDF	The sample radiance divided by the sample irradiance. The procedures given in this practice are correct only if the field of view (FOV) determined by the receiver field stop is sufficiently large to include the entire illuminated area
Scatter	The radiant flux that has been redirected (by interaction with the sample) over a range of angles
Sample coordinate system	A coordinate system is fixed to the sample and used to a specify position on the sample surface for the measurements. The sample coordinate system may be application and sample specific. A Cartesian coordinate system is recommended for flat samples, with the origin at the geometric center of the sample face with the Z axis normal to the sample
Plane of incidence, PLIN	The plane defined as containing the central ray of the incident flux and the sample normal
Beam coordinate system, XB, YB, ZB	A Cartesian coordinate system defined with the origin on the central ray of the incident flux at the sample surface, the XB axis in the plane of incidence (PLIN) and the ZB axis normal to the surface. The angle of incidence, scatter angle, and incident and scatter azimuth angles are defined with respect to the beam coordinate system
Incident direction	The central ray of the incident flux specified by θ_i and φi in the beam coordinate system
Angle of incidence, θ_i	Angle between the central ray of the incident flux and the ZB axis
Incident azimuth angle, φ_i	The fixed 180° angle from the XB axis to the projection of the incident direction onto the XB–YB plane. It is convenient to use a beam coordinate system in which $\varphi_i = 180°$ since this makes φ_s the correct angle to use directly in the familiar form of the grating equation. Conversion to a sample coordinate system is straight forward, provided the sample location and rotation are known
Scatter direction	The central ray of the scattered flux specified by θ_s and φ_s in the beam coordinate system
Scatter angle, θ_s	Angle between the central ray of the scattered flux and the ZB axis
Scatter azimuth angle, φ_s	Angle from XB axis to the projection of the scatter direction onto the XB–YB plane
Specular direction	The central ray of the reflected flux that lies in the PLIN with $\theta_i = \theta_s$ and $\varphi_s = 0$
Delta theta, $\Delta\theta$	The angle between the specular direction and the scatter direction
Delta beta, $\Delta\beta$	The projection of $\Delta\theta$ onto the XB–YB plane, that is, the $\Delta\theta$ angle measured in direction cosine space. For scatter in the PLIN, $\Delta\beta = \sin\theta_s - \sin\theta_i$. For scatter out of the PLIN, the calculation of $\Delta\beta$ becomes more complicated
Receiver solid angle, Ω	The solid angle subtended by the receiver aperture stop from the sample origin
Sample radiance, L_e	A differential quantity that is the reflected radiant flux per unit projected receiver solid angle Ω per unit sample area. L_e is often an average calculated from scattered power, P_s collected by the projected receiver solid angle $\Omega\cos\theta_s$, from the illuminated area, A. The receiver aperture and distance from the sample determines Ω and the angular resolution of the instrument
Sample irradiance, E_e	The radiant flux incident on the sample surface per unit area. In practice, E_e is an average calculated from the incident power, P_i, divided by the illuminated area, A. The incident flux should arrive from a single direction; however, the acceptable degree of collimation or amount of divergence is application specific and should be reported

Table adapted from ASTM E1392-90 "Standard Practice for Angle Resolved Optical Scatter Measurements on Specular or Diffuse Surfaces," American Society for Testing Materials, 1990.

Given these definitions, BRDF, is mathematically described by

$$\text{BRDF} = \frac{L_e}{E_e} = \frac{(P_s/\Omega A\cos\theta_s)}{(P_i/A)} = \frac{P_s}{P_i\Omega\cos_s}\ [\text{sr}^{-1}]$$

The BRDF is a function of many measurable quantities. It depends on the angle of incidence of the incident flux and the wavelength of this radiation. It also depends on the direction in which the scattered light is measured as well as the wavelength and polarization of the incident and scattered flux. The BRDF can be calculated for a given angular configuration of incident source and detector from the average radiance divided by the average irradiance. The BRDF is therefore referenced by the incident power. The $\cos\theta_s$ term in the formula accounts for the foreshortening of A as the angle θ_s increases, so that true radiance is being used in the calculation. Because of this, the BRDF is constant for a surface with Lambertian (uniformly diffuse) properties. The detector must have a field of view so that all of A can be viewed and such that P_i includes the entire beam. In practice, P_i is measured by removing the sample from the measurement position. Then the detector is moved into the straight-through position. Often a reference detector is used to monitor the incident beam power so that any changes after the initial V_i, measurement can be recorded.

For BRDF measurements to be meaningful, the instrument signature must be examined and understood. This is particularly important in the measurement of low-scatter surfaces. The instrument signature is the BRDF when the sample is removed and there is no sample scatter. In this case the stray light and scatter signature of the device will be apparent. Angular positioning errors and electronic noise can also be important. For example, as θ_s approaches 90° the accuracy of measuring θ_s becomes critical due to the 1/cos θ_s term in the BRDF equation. Ideally the instrument signature is a small fraction of the BRDF being recorded. An understanding of the instrument signature provides the needed information to decide where the BRDF data loses its meaning.

The BRDF is most often displayed in a plot, where the BRDF value (vertical axis) and the angle of scatter (horizontal axis) are plotted on logarithmic scales. In these 'log–log' plots it is easy to examine the scatter distribution close to the specular direction which, for shiny surfaces, represents the bulk of the scattered light. For surfaces that are diffuse in nature, a log–log plot of the BRDF generally yields a straight line. The slope of this plot is characteristic of the surface of the process which is used to produce the surface. This straight-line plot is useful in comparing samples made with the same process. Deviations from the slope of the line or the value of the points on the line indicate that there may be problems with a sample. Thus the simplest scattering instrument can be used for quality control of black samples. For a fixed angle of incidence and a fixed detection angle, variations in the BRDF value may indicate changes in the materials or the processes which created their surfaces. The quality control process can be performed by looking for variations in the BRDF among the different samples.

The BRDF can be used as a measure of surface roughness – an important attribute for a diffuse black surface. As random surface roughness features increase on a surface, the scatter from the surface also increases. Because of this, the BRDF can be used to characterize surface roughness parameters. If the surface roughness is all in the form of small, micro-rough surface features, the BRDF value then is very much proportional to the surface power spectral density function (PSD), which is a measure of the power plotted against spatial frequency. The characterization of surface roughness by BRDF scatter measurements only works if the scattered energy is due to surface micro-roughness. If the scatter is from larger surface features or from surface contamination, then it will be difficult to accurately characterize the surface roughness. Thus only smooth, mirror-like black surfaces can be used for this simple leap from BRDF values to surface roughness parameters. Some sample BRDFs of black surfaces are found in Figures 4–7.

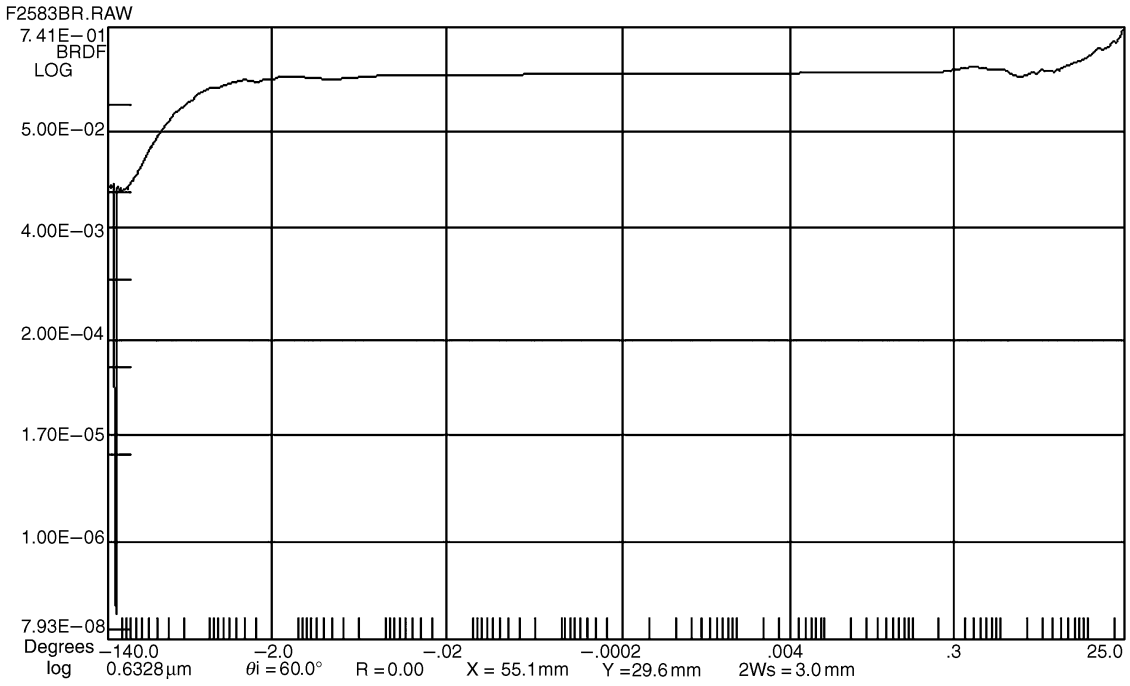


Figure 4 BRDF at wavelength of 0.633 micrometers and 60° incidence angle of black anodize sample. Plot generated by SOLEXIS using CASI software from Schmitt Measurement Systems, Portland, OR, USA.

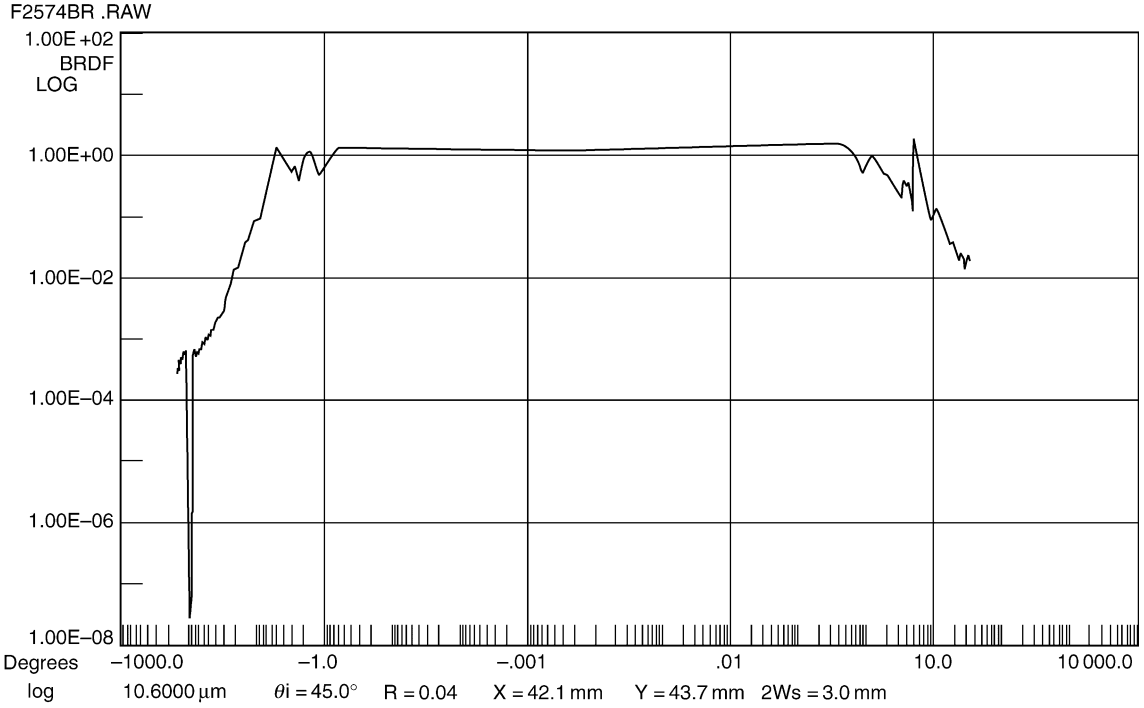


Figure 5 BRDF at wavelength of 10.6 micrometers of black anodize at 45° incidence. Plot generated by SOLEXIS using CASI software from Schmitt Measurement Systems, Portland, OR, USA.

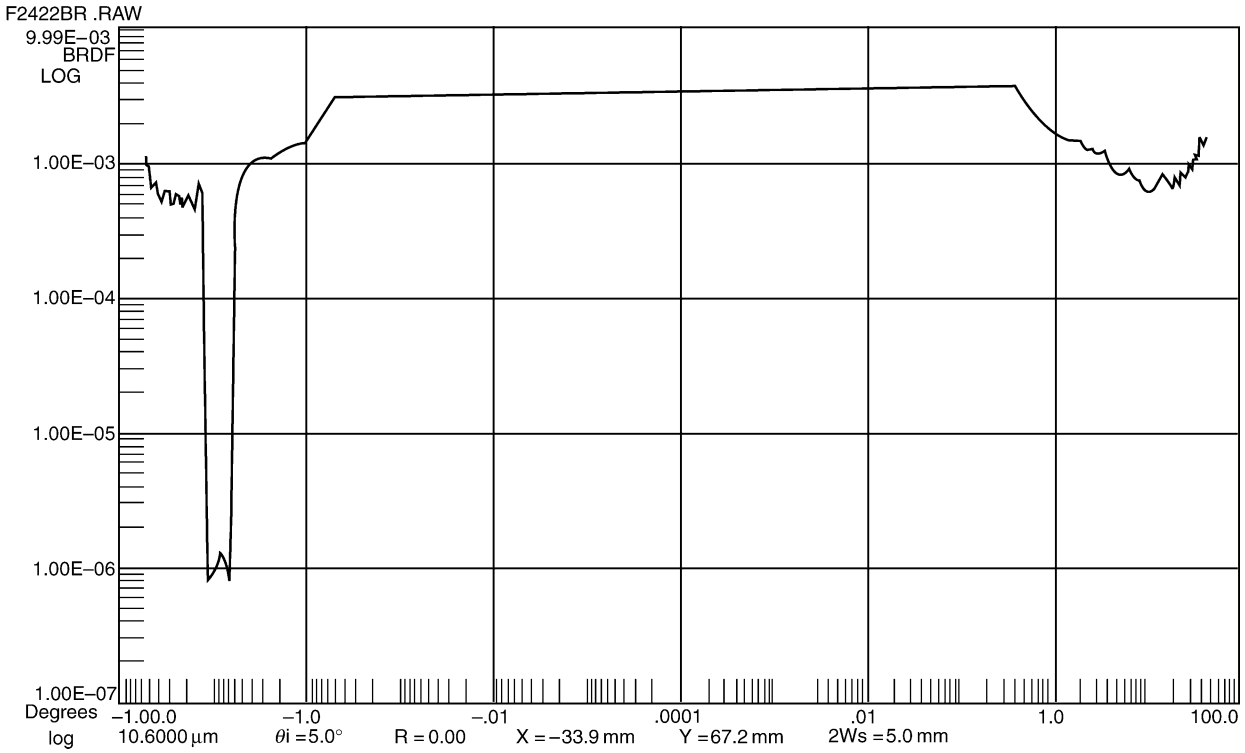


Figure 6 BRDF at a wavelength of 10.6 micrometer of Martin Optical Black at 45° incidence. Plot generated by SOLEXIS using CASI software from Schmitt Measurement Systems, Portland, OR, USA.

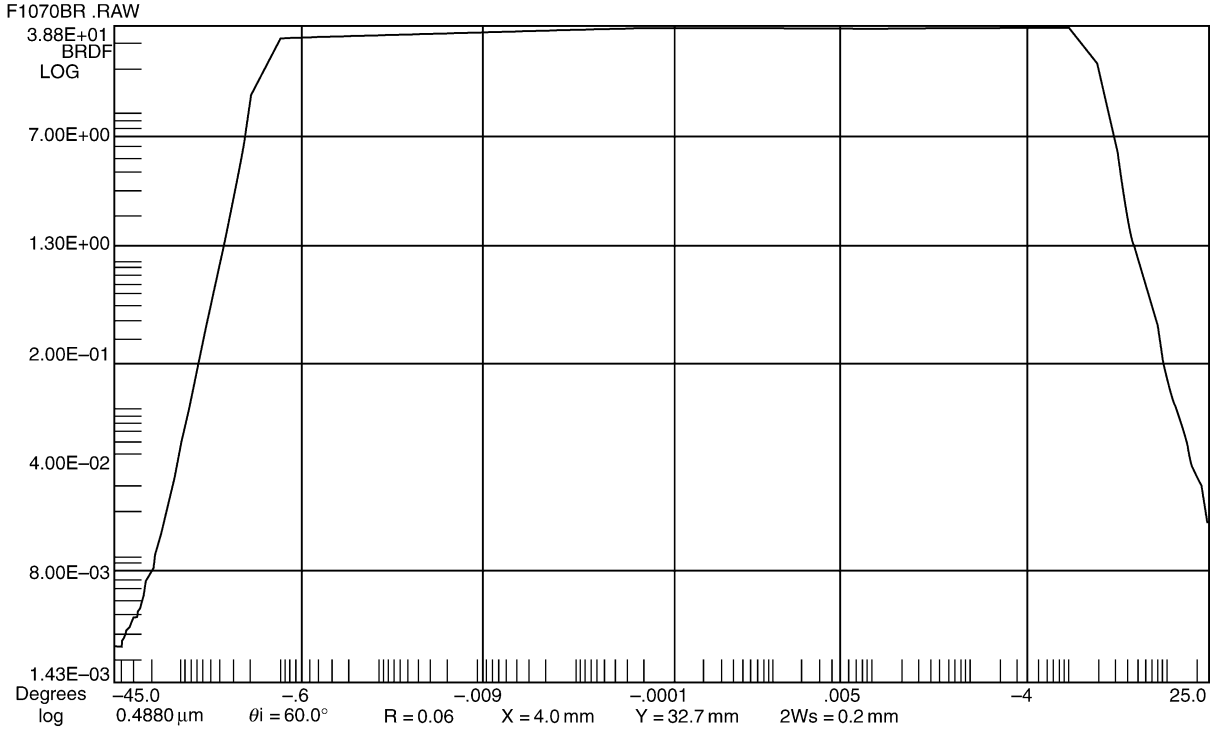


Figure 7 BRDF at a wavelength of 0.488 micrometers of a glossy black surface at 60° incidence. Plot generated by SOLEXIS using CASI software from Schmitt Measurement Systems, Portland, OR, USA.

Examples of Black Surfaces

Black surfaces appear in many forms. They may be either diffuse or specular in nature, or both. Black surfaces may be created by painting a surface, by altering a surface (such as through an anodization process), or by any number of other surface treatment effects. For example, absorbing compounds may be added to paint or to a surface to decrease reflection. Adding large (compared to the wavelength) craters or holes to a surface may result in multiple reflections within the surface leading to decreased reflection. Surfaces may be modified to increase their light-scattering ability, with the hope that reflection in a specific direction will be reduced. Sometimes surfaces may be created in such a way as to enhance thin-film interference effects, leading to reduced reflection. These and other techniques may be combined together in an attempt to reduce reflection and make the material appear blacker in a particular wavelength range of interest or in a particular direction.

Some sample black surfaces are given in Table 3, although databases are much more current on current names and manufacturers.

Many black surfaces use surface texture to enhance their light-absorbing ability. Scanning electron micrographs of surfaces reveal the surface morphology and how multiple reflections can be used for light trapping. Surface chemistry, changes in the index of refraction, and the use of particles of different sizes are also used to create better absorbers. A variety of other proprietary methods are used as well (Figures 8–11).

Future Developments

There are many new kinds of black surfaces being developed for specialized applications. Black surfaces can be created using a variety of new techniques and materials. For example, new surfaces can be created using laser ablation or chemical microetching of surfaces. Surfaces can also be created using new materials such as carbon nanotubes or by using the technology used to create flocked materials. These new kinds of surfaces generally have good absorption at a variety of wavelengths. Even with the creation of these new surfaces, the greatest gain in an optical system will probably be attained through a concerted effort to choose the appropriate black surfaces early in the program and to place them properly in the optical system. An early selection of surfaces and an analysis of their performance will lead to a higher level of system performance than a search for an exotic, ultrablack surface at a late stage

Table 3 Selected optical black surfaces

Name	*Manufacturer*	*Notes*	*Type*
Aeroglaze L300	Lord Corporation Erie, PA	Formerly called Chemglaze L300	Paint
Aeroglaze Z004	Lord Corporation Erie, PA	Formerly called Chemglaze Z004	Paint
Aeroglaze Z302	Lord Corporation Erie, PA	Formerly called Chemglaze Z302	Paint
Aeroglaze Z306	Lord Corporation Erie, PA	Formerly called Chemglaze Z306	Paint
Aeroglaze Z306 with microspheres	Lord Corporation Erie, PA	Formerly called Chemglaze Z306 with microspheres	Paint
Aeroglaze Z307	Lord Corporation Erie, PA	Formerly called Chemglaze Z307	Paint
Aeroglaze Z313	Lord Corporation Erie, PA	Formerly called Chemglaze Z313	Paint
Ames 24E (Ames 24E2)	NASA Ames Research Center Moffett Field, CA		Paint
Cardinal 6450	Cardinal Industrial Finishes South Elmonte, CA	Formerly 'Cardinal 6550'	Paint
Cornell Black	Cornell University		Paint
DeSoto Flat Black	Courtaulds		Paint
Electrically Conductive Black Optical Paint	Jet Propulsion Laboratory Pasadena, CA	Has no trade name	Paint
IITRI Bone Black D-111 (IITRI D111)	Illinois Institute of Technology Chicago, IL		Paint
LMSC Black	Lockheed Martin Palo Alto		Painted multilayer coating
MH21-I	Illinois Institute of Technology Chicago, IL		Paint
MH55	Illinois Institute of Technology Chicago, IL		Paint
MH2200	Illinois Institute of Technology Chicago, IL	Formerly 3M's ECP 2200 paint, but sold to IIT	Paint
Solarchem	Eastern Chem Lac Corporation Malden, MA		Paint
463-3-8	Akzo Coatings, Inc. Orange, CA	Formerly called 'Cat-a-lac 463-3-8' diffuse black paint	Paint
443-3-8	Akzo Coatings, Inc. Orange, CA	Formerly called 'Cat-a-lac 443-3-8' glossy black	Paint
443-3-17	Akzo Coatings, Inc. Orange, CA	Formerly called 'Sikkens 443-3-17'	Paint
Infrablack	Lockheed Martin Denver, CO	For Al substrates only	Anodization process
Martin Black	Lockheed Martin Denver, CO	For Al substrates only	Anodization process
Martin Black, Enhanced	Lockheed Martin Denver, CO	For Al substrates only	Anodization process
Infrablack	Lockheed Martin Denver, CO	For Al substrates only	Anodization process
Black Chrome Type Surfaces	Lockheed Martin Denver, CO	For many kinds of metal substrates	Electrodeposition process
Black Cobalt Type Surfaces (Cobalt Black)	Lockheed Martin Denver, CO and many more companies	For many kinds of metal substrates. References to Black Copper, Black Steel etc. are covered by Black Cobalt	Electrodeposition process, and can be followed by chemical or thermal oxidation
Black Nickel	−NIST	For many kinds of metal substrates	Deposition and etching process
(NBS Black)	−Ball Aerospace		
(Ball Black)	−Lockheed Martin and others	Ball modified the process developed by NBS	
Black Kapton™ Foil			Foil
Black Tedlar™ Foil			Foil
Boron Black	Lockheed Martin Denver, CO	For many kinds of metals	Plasma spray deposition process
Boron Carbide	Lockheed Martin Denver, CO	For Ti substrates only	Proprietary process
Silicon Carbide	Lockheed Martin Denver, CO and other companies	For many kinds of metals	Chemical vapor deposition
Textured Surfaces	−NASA Ames Research Center Moffat Field, CA −SPIRE Corporation Bedford, MA −Optics MODIL Oak Ridge National Laboratory	For many kinds of metal substrates	Either: −sputter coated −ion beam etched −sputter coated then etched

Note that finish names and manufacturers change regularly. Table adapted from SOLEXIS™ database, Stellar Optics Research International Corporation (SORIC), 78 Normark Dr., Thornhill, Ontario, Canada L3T 3R1.

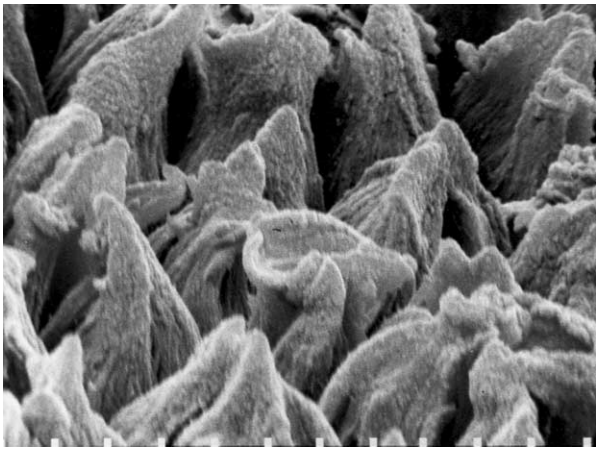


Figure 8 Scanning electron micrograph of Martin Optical Black surface on 6061T6 aluminum. This surface is a good absorber at ultraviolet, visible, and near infrared wavelengths. (Courtesy of Stephen Pompea and Donald Shepard.)

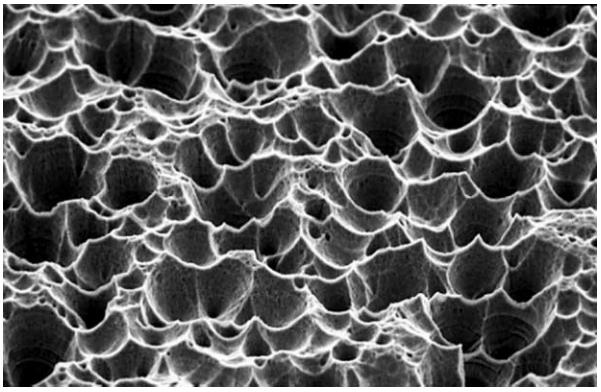


Figure 9 Scanning electron micrograph of phosphoric acid etched electroless nickel surface. This surface is a good absorber in the visible wavelength region. (Courtesy of Stephen Pompea and Donald Shepard.)

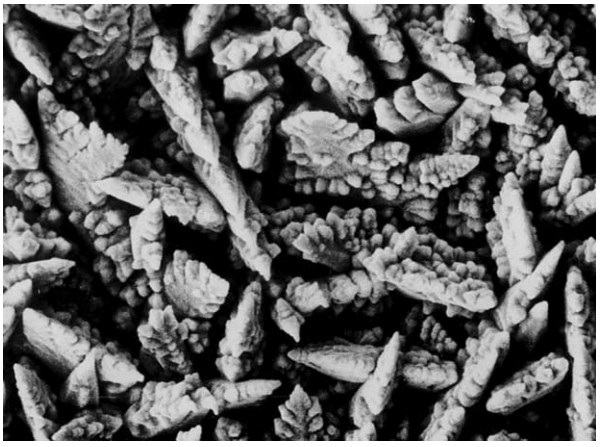


Figure 11 Scanning electron micrograph of tungsten hexafluoride surface. This surface is a good absorber at visible and near infrared wavelengths. (Courtesy of Stephen Pompea and Donald Shepard.)

in the program. Although better black surfaces are constantly being developed, the most important future development in the area of black surfaces may be a wider realization of the importance of an early and comprehensive surface selection process. The value of starting the selection process early using black surface databases has been confirmed repeatedly by optical designers.

See also

Instrumentation: Photometry; Telescopes. **Scattering:** Scattering from Surfaces and Thin Films.

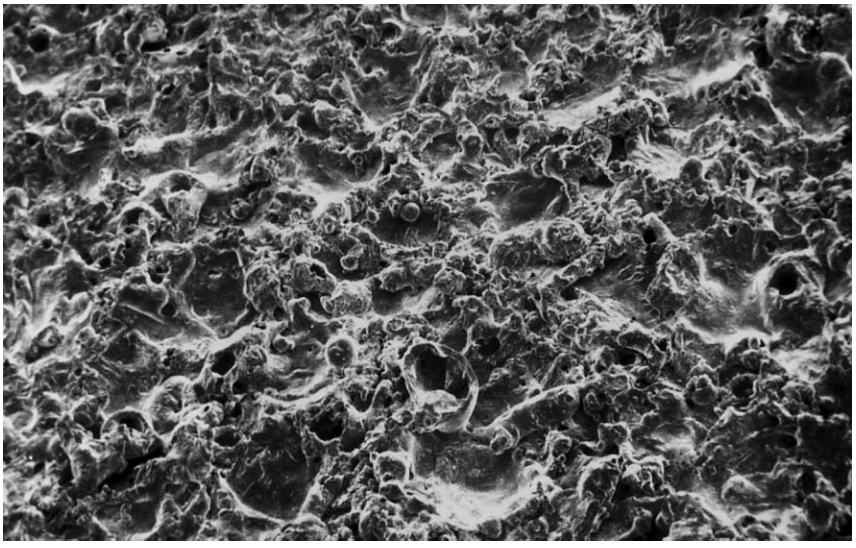


Figure 10 Scanning electron micrograph of Aeroglaze Z-306 painted surface with embedded microspheres. This surface is used primarily at visible wavelength. (Courtesy of Stephen Pompea and Donald Shepard.)

Further Reading

American Society for Testing Materials (ASTM) (1990) ASTM E1392-90. *Standard Practice for Angle Resolved Optical Scatter Measurements on Specular or Diffuse Surfaces*. West Conshohocken, PA: ASTM.

Bennett JM and Mattsson L (1989) *Introduction to Surface Roughness and Scattering*. Washington, DC: Optical Society of America.

Bohren CF and Huffman DF (1983) *Absorption and Scattering of Light by Small Particles*. New York: John Wiley and Sons.

Breault RP (1995) Control of stray light. In: *Handbook of Optics*, 2nd edn. New York: Optical Society of America.

Church EL and Takacs PZ (1995) Surface scattering. In: Bass M (chief ed.) *Handbook of Optics*, 2nd edn. New York: Optical Society of America.

Granqvist CG (1989). *Spectrally Selective Surfaces for Heating and Cooling Applications*. Volume TT 1. Bellingham, WA: SPIE Optical Engineering Text.

Gunderson JA (1977) *Goniometric Reflection Scattering Measurements and Techniques at 10.6 Micrometers*. Thesis, University of Arizona.

Henninger JH (1984) Solar absorptance and thermal emittance of some common spacecraft thermal control coatings. *NASA Reference Publication 1121*.

Jelinsky P and Jelinsky S (1987) Low reflectance EUV materials: a comparative study. *Applied Optics* 26(4): 613–615.

Persky MJ (1999) Review of black surfaces for space-borne infrared systems. *Rev. Sci. Instruments* 70(5): 2193–2217.

McCall SHCP, Pompea SM, Breault RP and Regens NL (1992) Reviews of properties of black surfaces for ground and space-based optical systems. In: Breault RP (ed.) *Proceedings of the SPIE: Stray Radiation II*, vol. 1753.

Pompea SM and Breault RP (1995) Optical black surfaces. In: *Handbook of Optics*, 2nd edn. New York: Optical society of America.

Pompea SM, McCall SHCP (1992) Outline of selection process for black buffle surfaces in optical systems. In: Breault RP (ed.) *Proceedings of the SPIE: Stray Radiation in Optical Systems* II, vol. 1753.

Pompea SM, Shepard DF and Anderson S (1988) BRDF measurements at 6328 Angstroms and 10.6 micrometers of optical black surfaces for space telescopes. In: Breault RP (ed.) *Proceedings of the SPIE: Stray Light and Contamination in Optical Systems*, vol. 967, pp. 236–248.

Smith SM (1980) Far infrared (FIR) optical black bidirectional reflectance distribution function (BRDF). *Proceedings of the SPIE: Radiation Scattering in Optical Systems* 362: 161–168.

SOLEXIS™ Database of BRDF Measurements (2003) Stellar Optics Research International Corporation (SORIC), Ontario, Canada.

Stover J (1995) *Optical Scattering: Measurement and Analysis*, 2nd edn. Bellingham, WA: SPIE Optical Engineering Press.

Wolfe WL (1978) Optical materials. In: Wolfe WL and Zissis GJ (eds) *The Infrared Handbook*. Ann Arbor, Michigan: ERIM.

Thin Film Optical Coatings

D Ristau, Laser Zentrum Hannover e.V., Hannover, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Thin-film optical coatings are employed to adapt the properties of an optical surface to its specific application. Commonly, optical coatings comprise a sequence of single layers of different transparent materials with a thickness in the range of the operation wavelength. Depending on this coating design, the spectral characteristics of an optical surface can be tuned to nearly any required transfer function including a suppression of reflection for a broad spectral range by an antireflective coating or even the achievement of highest reflectance values up to 99.999% using a high reflecting layer stack. Besides these basic functions, optical coatings can be employed to realize a variety of transfer spectra specified for multiple wavelengths or wavelength ranges to manipulate radiation according to even extremely sophisticated demands in optical systems for laser technology, communication, or fundamental research. Moreover, the desired optical properties can be combined with additional specifications, for example, concerning the mechanical or chemical stability of an optical surface by selecting adapted coating materials. As a consequence of this enormous versatility, thin-film optical coatings can be found in nearly every product and development of modern optical technology today. Aside from the large market of consumer optics including ophthalmology, camera objectives, binoculars, and optical systems for data storage, progress in optical coating technology is predominantly driven by demanding applications in many future technology fields. As one prominent pace-setter, semiconductor lithography imposes new demands on optical coatings extending the operation wavelength of stepper systems in chip production to the extreme ultraviolet (EUV)-spectral range. Other examples can be found in communication technology, where thin-film filters with a bandwidth below

0.5 nm are required, or also in laser material processing, where coatings with high power handling capability and lifetime are indispensable for the development of cost-efficient laser tools.

In this article a concentrated review will be given on the design, production, and on specific properties of optical coatings. After an introduction to the fundamental principles of operation of coating systems, basic expressions for the optical transfer function of a single layer will be derived and extended to layer structures on the basis of the matrix method. Selected conventional and ion processes for the production of coatings will be discussed with respect to the achievable optical and structural properties of the coatings. Also, the thickness control of the coatings during production will be considered introducing the crystal monitoring technique as well as optical measurement strategies at a single wavelength or in a broad spectral range. The article will conclude with a review of the present state of optical coatings and an analysis of future trends.

Design of Optical Coatings

The colorful appearance of optical coatings and microstructures, which can be observed on some butterflies and other insects, was present in nature long before mankind invented them. The beginning of the technical evolution of optical surfaces may be dated back to Greek culture, where polished metal surfaces were used as mirrors. Considering the application of transparent coatings, the investigations of Joseph Fraunhofer in the reflectivity of aged glass surfaces during the year 1819 should be mentioned as a first approach towards an antireflective coating. On the basis of controlled etching, this technique was further developed for different glass materials by Dennis Taylor in 1891 to realize the first antireflective coatings on lenses for applications in optical devices. In 1899, the Fabry–Perot theory was developed as a milestone in the theoretical description of multilayer structures. This theory describes the interference of partial waves reflected from two parallel optical surfaces, which can be considered as the fundamental arrangement of a multilayer structure (see Figure 1). In an optical thin-film system, layers of different transparent materials are composed on a stack deposited on the surface of the optical component (substrate refractive index n_T). In general, to adjust the spectral characteristics of the layer system to the demanded specifications, at least two layer materials with a high refractive index n_H and a low index n_L have to be selected. The choice of the thin-film materials is dependent on the application wavelength and the desired stability of the coating system as well as on their optical properties essential for the specific design and the targeted optical quality. Similar to the Fabry–Perot principle, the optical function of the layer structure is based on the interference of partial waves reflected and transmitted at the interfaces between the layers. The trace of the incoming wave and the induced partial waves are depicted in Figure 2 for a single layer, which can be considered as the basic unit for coating systems even with highest complexity. For a theoretical description of the transfer function of this single layer, the contributions of the individual partial waves have to be calculated and accumulated to form the total transmitted and reflected wave. In this classic approach, the incoming wave can be described by the function of a plane wave with amplitude E_0, wavenumber $k = 2\pi/\lambda$ incorporating the wavelength λ, and frequency ω:

$$E(x,t) = E_0 \exp[-\mathrm{i}(kx - \omega t)] \qquad [1]$$

At the first interface, a part of this plane wave is directly reflected with an amplitude of $A_0 = E_0 r_1$

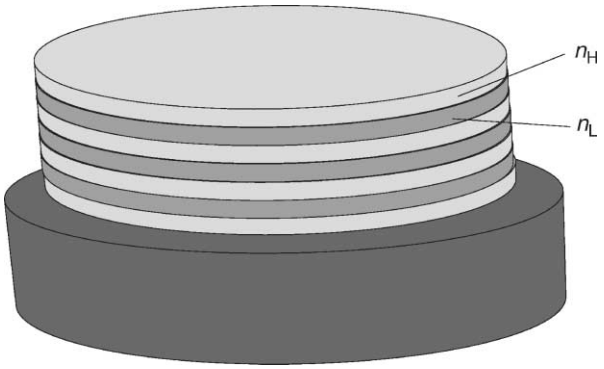


Figure 1 Basic structure of a thin-film system. Transparent layers of at least two different materials (n_H, n_L) are deposited on a substrate with refractive index n_T.

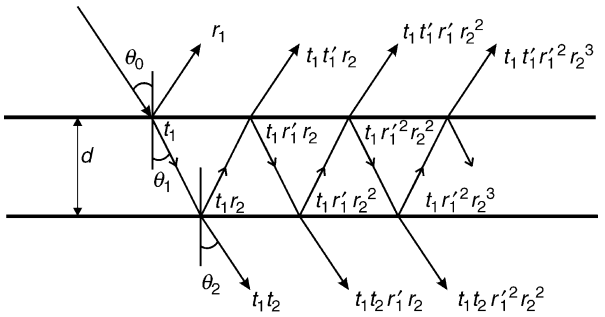


Figure 2 Schematic path of a plane wave incident onto a single layer with refractive index n_1. The reflection and transmission coefficients for the amplitudes at the interfaces $m = 1, 2$ are denoted by t_m and r_m, respectively. A prime added to the coefficients indicates a transfer of the wave opposite to the direction of incidence.

(order zero), which can be calculated using the reflection coefficient r_1 given by the Fresnel formulas. After passing the first interface, a reflection and transmission of the incoming wave at the second interface has to be considered. Also a phase shift has to be taken into account, which is introduced by the path of the wave from the front to the rear interface of the single layer. This phase shift δ_i is dependent on the thickness d_i of the layer and the refractive index n_i:

$$\delta_i = \frac{2\pi n_i d_i \cos\theta_i}{\lambda} \qquad [2]$$

For oblique incidence also the angle of incidence θ_i in the layer, which has to be calculated by Snell's law, has to be taken into account. The index i has been chosen in this consideration for convenience to account for the layer number in coating systems, which will be discussed later. Following the wave reflected the first time inside the layer, a further phase shift for the path from the rear side to the front side has to be taken into account. This partial wave is then transmitted by the front surface resulting in an amplitude A_1 of the first-order wave leaving the layer:

$$A_1 = t_1 t'_1 r_2 E_0 \exp(-i2\delta) \qquad [3]$$

where t'_1 is the transmission coefficient for the amplitude of the wave passing the layer at the interface to the ambient. By following the wave further through the layer, the amplitudes A_k of each partial wave of order k can be determined:

$$A_k = t_1 t'_1 r_2 E_0 \exp(-i2\delta) * (r'_1 r_2 \exp(-i2\delta))^{k-1} \qquad [4]$$

Obviously the contributions of the partial waves follow the rule of a geometric expansion, and therefore the total amplitude of the wave reflected by the layer, which is the sum of the amplitudes of the partial waves, can be expressed in closed form:

$$A_{\text{tot}} = \left[r_1 + \frac{t_1 t'_1 r_2 \exp(-i2\delta)}{1 - r'_1 r_2 \exp(-i2\delta)} \right] E_0 \qquad [5]$$

The reflection coefficient r_S of the single layer is given by the ratio between the total amplitude of the reflected wave and the amplitude of the incoming wave:

$$r_S = \frac{r_1 + r_2 \exp(-i2\delta)}{1 + r_1 r_2 \exp(-i2\delta)} \qquad [6]$$

In this equation, the coefficients r'_1 and t'_1 have been replaced by using the relations

$$t_1 t'_1 = (1 - r'_1)(1 - r_1) = (1 - r_1)(1 + r_1) = 1 - r_1^2 \qquad [7]$$

In principle, this method of compiling the contributions of the partial waves reflected at the layer interfaces can be extended also to the calculation of systems with more than one layer. But, considering the enormously increasing number of partial beams, the resulting equations become extremely complicated for multilayer structures. Therefore, the matrix formalism, which can be deduced from the boundary conditions of the electric and magnetic field at the interfaces, is usually employed for the calculation of thin-film systems. The major advantage of this approach, which was developed in the 1940s, is the representation of each layer by a single matrix M_i containing all specific parameters of the film. The matrix M_i of a thin film at position i within the layer system relates the electric (E_{i-1}) and magnetic (H_{i-1}) field strength at the front face to the field strength values (E_i and H_i) at the rear face of the layer:

$$\begin{pmatrix} E_{i-1} \\ H_{i-1} \end{pmatrix} = \mathbf{M}_i \begin{pmatrix} E_i \\ H_i \end{pmatrix} = \begin{pmatrix} \cos\delta_i & \frac{i}{n_i}\sin\delta_i \\ i n_i \sin\delta_i & \cos\delta_i \end{pmatrix} \begin{pmatrix} E_i \\ H_i \end{pmatrix} \qquad [8]$$

By considering the boundary conditions for the field strengths at the interfaces, the matrix components given in eqn [8] can be determined. On the basis of the matrix formalism, the transfer function of a layer system can be simply accomplished by multiplication of the matrices M_i representing the constituent single layers. Thus, the transfer matrix M_S of a stack, which is formed by a number K of single layers with layer 1 located at the first interface with respect to the incoming wave, is given by

$$\mathbf{M}_S = \mathbf{M}_1 \mathbf{M}_2 \mathbf{M}_3 \ldots \mathbf{M}_i \mathbf{M}_{i+1} \ldots \mathbf{M}_K \qquad [9]$$

For calculating the reflection coefficient r_{SK} of the entire arrangement including the substrate (index of refraction n_T) and the ambient medium (index of refraction n_0), again the ratios of the amplitudes have to considered:

$$r_{SK} = \frac{n_0 \mathbf{M}_{11} + i n_0 n_T \mathbf{M}_{12} - i\mathbf{M}_{21} - n_T \mathbf{M}_{22}}{n_0 \mathbf{M}_{11} + i n_0 n_T \mathbf{M}_{12} + i\mathbf{M}_{21} + n_T \mathbf{M}_{22}} \qquad [10]$$

As a consequence of its perceptible structure, the matrix method can be readily transferred into computer codes, which cover the calculation of all important properties of even extremely complicated layer systems. In particular, considering the fundamental relations of the matrix method, also the field strength distribution within the layer structure can be determined. In software packages, which are commercially available as standard tools for optical thin-film technology, often the film thickness D_i of the layers is expressed in the units of QWOT (quarter wave optical

thickness) at the design wavelength λ_Z:

$$D_i = \frac{4n_i d_i}{\lambda_Z} \qquad [11]$$

In the case of normal incidence, the matrices M_i reduce to simple expressions for layers of integer QWOT values, because the corresponding phase shift terms are multiples of $\pi/2$. Clearly, an integer QWOT value of optical thickness also represents an extreme value in the spectral behavior at the design wavelength. This fact is often exploited for a precise thickness control by optically monitoring the reflectance or transmittance of the growing layer at the design wavelength λ_Z: by recording the transfer behavior as a function of time during production, extreme values and accordingly thickness values of multiple QWOT can be accurately determined. Thus, if a layer structure is composed only of multiple QWOT layers, switching between the layer materials can always be precisely performed at a minimum or maximum read-out of the optical monitor tuned to the design wavelength. In respect of this advantage in production, most coating designs are based on QWOT layers indicating the practical importance of the unit QWOT. For an efficient description of the designs, often a notation in capital letters is used, which indicate a 1 QWOT layer of a certain coating material. For example an alternating QWOT-stack with 11 layers of a high-index material (H, refractive index n_H) and a low-index material (L, refractive index n_L) is described by the sequence HLHLHLHLHLH or even more condensed by $(HL)^5H$. The QWOT stack is the fundamental design for dielectric high reflecting mirrors and represents the basis for a variety of functional coatings in optical thin-film technology. By applying the matrix method, an approximate expression for the reflectance R_S at the design wavelength can be derived for such a stack:

$$R_S \approx 1 - 4n_0 n_T \frac{n_L^{2N}}{n_H^{2(N+1)}} \qquad [12]$$

This expression is only valid for a large number N of QWOT layers and underlines the role of the materials for the practical production of optical coatings. Highest reflectivity with lowest number of layers can be achieved only by selecting materials with a high contrast of their refractive indices. This contrast also influences the spectral broadness of the high-reflecting region of the stack (Figure 3).

Besides highly reflecting mirrors, antireflective coatings are another important class of functional coatings for modern optics and laser technology. Antireflective coatings often involve non-QWOT layer structures of more than two materials in order to accomplish the demands of lowest residual reflectivity for a broad spectral range. Eye glasses with antireflective coatings covering the entire visible spectral range can be considered as a prominent example for consumer optics, which impose requirements on the coatings also in respect to their stability against cleaning and other environmental influences during normal use. In this context, the introduction of optical plastic materials is an additional challenge in consumer optics and ophthalmology, which has to be mastered by adapted designs and production processes. In contrast to the typical demands of consumer optics, antireflective coatings on laser windows are

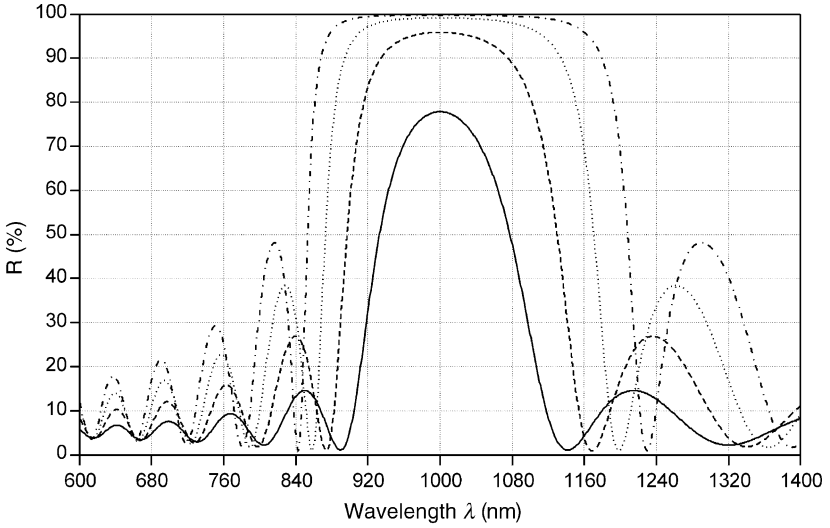


Figure 3 Spectra calculated for high-reflecting QWOT stacks (15 layers, substrate index $n_T = 1.52$) composed of a low-index material ($n_L = 1.46$) and different materials with high indices of refraction varying between $n_H = 1.7$ (solid), 1.9 (dashed), 2.1 (dotted) and 2.3 (dash-dotted). The spectral broadness and the reflectivity increase with the ratio n_H/n_L.

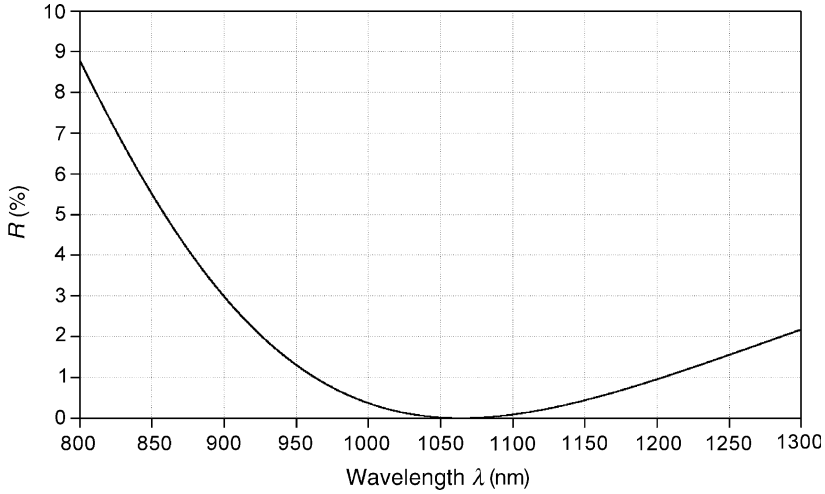


Figure 4 Spectrum of an antireflective coating for the wavelength 1.064 μm, which is designed as a V-coating with two layers of Ta_2O_5 ($n_H = 2.03$; $d_H = 0.383$ QWOT) and SiO_2 ($n_L = 1.46$; $d_L = 1.308$ QWOT) for a quartz substrate (substrate index $n_T = 1.46$).

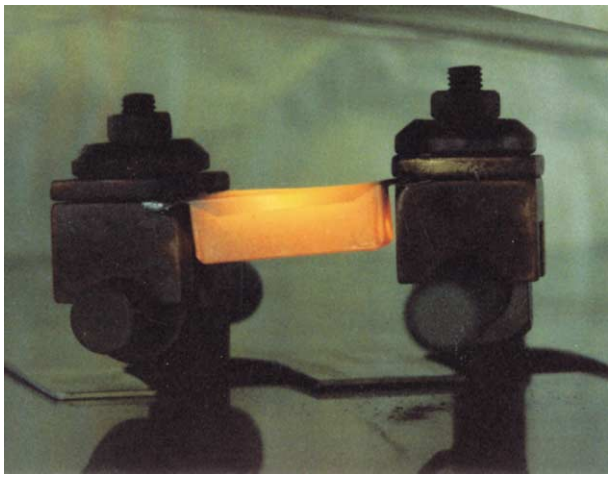


Figure 5 Boat evaporation source in operation. The deposition material is put into the boat which is made of a metal with high melting point. A high electric current is conducted through the boat for heating to temperatures up to 2000 °C.

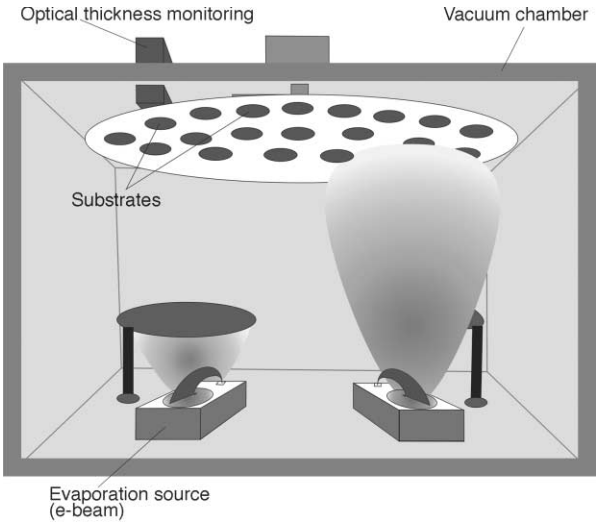


Figure 6 Outline of a modern box coater used for the production of optical coatings with e-beam deposition processes.

often restricted to one wavelength, but have to fulfill extreme requirements with respect to optical losses and power handling capability. Therefore, most antireflective coatings in laser technology are composed only of two thin layers to minimize the total thickness, which is often a direct measure of the expected losses of the layer system. With such a coating, which is often called V-coating according to its spectral characteristic (Figure 4), the reflectivity of the window surface can be suppressed ideally for one wavelength.

Production of Optical Coatings

The industrial production of optical coatings started in the 1940s on the basis of glass bell jar systems equipped with diffusion pumps and boat evaporators for the deposition of single layers of metals for mirrors and fluoride materials for antireflective coatings. Metal containers are used as boat evaporation sources. These containers are filled with the deposition material and heated by electric currents up to several 100 A (Figure 5). In the course of the rapid improvement of vacuum systems and especially the development of the electron beam evaporator over the next three decades, new materials with improved optical and mechanical quality could be combined to form more complicated dielectric coating systems. The layout of an e-beam deposition plant of these times, which can still be considered as fundamental for the advanced deposition systems of today, is illustrated in Figure 6. In such a box coater, the

substrates are installed in a calotte which is rotated to achieve a uniform distribution of the layer thickness. The deposition materials are put in water-cooled crucibles and are subjected to an electron beam operated at currents in the range between 0.1 A to more than 10 A and voltage levels of several kilovolts. As a consequence of the interaction with the energetic electrons, the deposition material can be heated to high temperatures above 5000 K allowing the evaporation of most of the desired deposition materials. These achievable high temperatures in conjunction with the good controllability are the major advantages of e-beam sources (Figure 7) in comparison with the classical boat evaporation process. In particular, e-beam evaporation opened the way for the deposition of oxide materials and other optical materials with high evaporation temperatures, which exhibit optical properties as well as a high mechanical and chemical stability superior to the material classes accessible with boat evaporation. However, the cost efficient boat evaporation technique is still applied today for the deposition of metals or temperature-sensitive materials for coating systems in the vacuum ultraviolet (VUV)/UV and mid-infrared (MIR) spectral ranges. If the shutter above the deposition source is opened, the evaporated material reaches the substrates and forms the layer by condensation. This condensation process is sensitively dependent on the environmental conditions of the substrates and the deposition sources. Sufficient microstructural properties of the growing layer can generally be achieved only under fine vacuum below 100 Pa and heating of the substrates to temperatures in the range of some 100 °C. Even under well-optimized conditions, the typical microstructure of such thermally evaporated coatings exhibits specific deficiencies (Figure 8) which influence the optical and mechanical properties of the layer systems. Often, microcrystallites with diameters in the range of a few nm are observed, which are arranged in columns of a typical size growing from the substrate interface to the outer surface of the layer. For a simulation of the growing mechanisms for this low energetic deposition process, uncomplicated models with a few input parameters already lead to satisfactory results. Assuming a kinetic energy of 0.1–0.3 eV, which is typical for thermal evaporation according to the Boltzmann distribution, the impinging atoms can move only a small distance or even adhere directly at their position of incidence on the growing layer. As a consequence of this small mobility, the atoms are not able to reach their energetically optimum positions in the layer, and they even obstruct the path of the following atoms by shadowing. On the basis of these assumptions, computer simulations of the growing process result in structures which are comparable to the columnar growth observed in practice. The packing density of such a columnar structure is significantly below the bulk values of the corresponding deposition materials, and it contains also a high number of defects and voids, which reduce the optical quality of the coating. In particular, the coating tends to

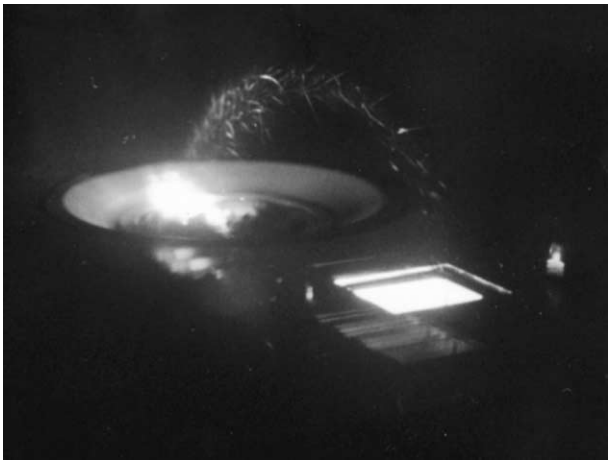


Figure 7 Electron beam source in operation. The deposition material is put in a water-cooled crucible and subjected to an e-beam. Since the zone heated by the e-beam is entirely embedded in the surrounding deposition material, very low contamination levels and temperatures above 5000 °C can be achieved. For the generation of the e-beam an arrangement similar to the system of a cathode ray tube is employed: the electrons emitted by a hot filament are focused by an aperture and accelerated by a high voltage.

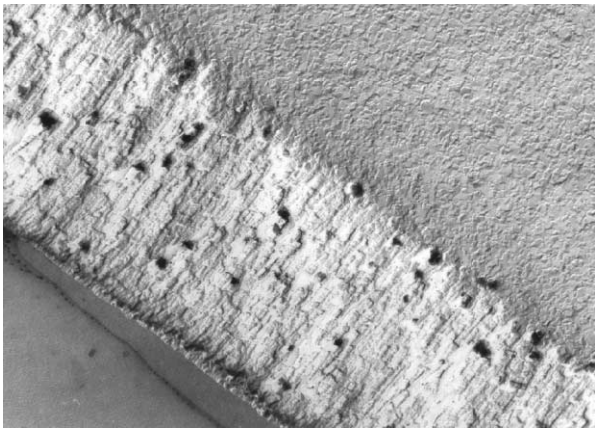


Figure 8 Microstructure of a MgF_2 single layer (geometric thickness approximately 780 nm) deposited by conventional boat evaporation. For the preparation of the TEM picture, the coating has been fractured and replicated.

adsorb water from the normal atmospheric environment into its internal structure leading to a shift of the spectral characteristics as a function of the ambient conditions. Besides this loss of spectral stability, the adsorbed water also induces an increased optical absorption of the coatings in the VUV and the MIR spectral ranges.

The increasing demand in laser technology and other optical applications for coatings far exceeds the quality level of conventionally deposited coatings and this was the driving force for the development of advanced concepts for the production of optical coatings during the last three decades. Considering the fundamental growth mechanism of conventional optical coatings, two basic approaches for an improvement of the microstructure are available. On one hand, the mobility of the atoms on the surface of the growing layer can be increased by the insertion of additional energy from other sources. This concept leads to the development of ion-assisted deposition processes, where the additional energy is transferred to the adatoms by collisions with ions generated directly in the plant by an ion source at energy levels of approximately 30–300 eV (Figure 9). Especially, by the deposition of oxide materials, which can be assisted directly with oxygen ions for balancing their stoichiometry, coating systems with high environmental stability and low optical losses can be achieved for the VIS up to the MIR spectral range. On the other hand, the energy of the adatoms can be directly increased in sputter processes which deliver deposition energies orders of magnitude higher than the thermal energy range. Besides adapted activated reactive magnetron sputter processes, particularly ion beam sputtering (IBS) was developed for a coating process delivering high-end coatings mainly for laser technology with an unsurpassed optical quality (Figure 10). For example, mirrors with total losses below one ppm can be produced for the Nd:YAG laser (wavelength 1.064 μm) with adapted IBS processes.

In addition to stable deposition processes, precise control of the layer thickness with an accuracy below 1 nm is a further important step towards the realization of complex optical coatings. Most deposition plants are equipped with quartz crystal monitors, which measure the optical thickness on the basis of the resonance frequency shift of an oscillating quartz crystal induced by the increase in mass of the deposited coating. The relative accuracy of these monitors ranges from 2% to 10% and is well suited for the regulation of the deposition rate. In order to achieve the required precision for high-quality coatings, optical monitors are often employed, which record the transmittance of the growing layer system at a single wavelength or even in a broad spectral range. Since these systems account directly for the optical properties of the deposited layer system, an efficient and accurate control of the film thickness is possible allowing the production of extremely sensitive filters with bandwidths below 1 nm for telecommunications or mirrors with defined group delay dispersion for fs-laser systems (Figure 11).

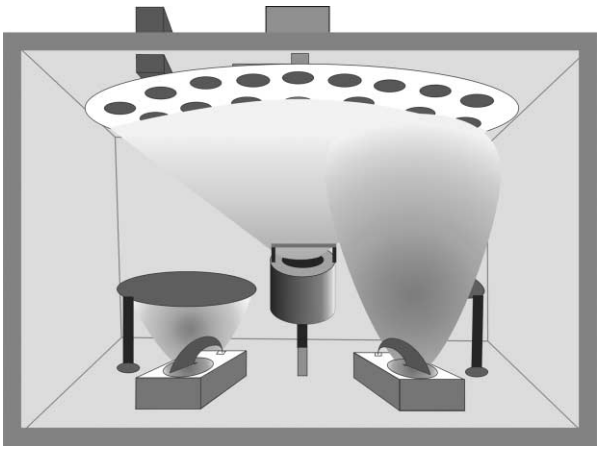


Figure 9 Ion assisted deposition. An additional ion source is installed, which delivers inert or reactive ions impinging on the growing layer. The energy transfer from the ions to the adatoms results in improved microstructure of the produced coatings with higher packing density.

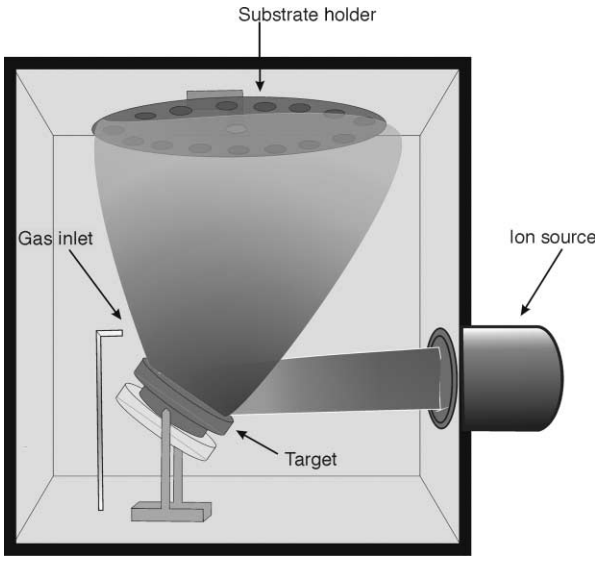


Figure 10 Principle of the ion beam sputtering (IBS) process. Inert ions with an energy in the range 1–2 KeV are generated in an ion source and directed onto the target which consists of the desired deposition material. Material is sputtered from the target and reaches the substrates. As a consequence of the high energy of the adatoms (5–30 eV) the formed layer exhibits a superior microstructure.

Quality Parameters of Optical Coatings

Besides the spectral characteristics, a variety of specific parameters has to be considered for each application of optical coating systems (Table 1). For laser systems the laser-induced damage threshold and the optical losses have to be evaluated in the context of technical and also economic requirements. Absorption is a loss channel, which transforms a fraction of the impinging radiation energy into heat, inducing a temperature increase in the coating. This generated temperature profile influences the imaging properties of the optical component via thermal expansion and temperature-induced alterations of the refractive indices of the coating as well as the substrate materials. Despite numerous technical problems, absorption always implies the loss of laser energy which is produced with high economic expense. Absorption losses are related to defects and impurities, which are introduced during the production process into the layer by a variety of contamination mechanisms. Furthermore, an unbalanced stoichiometry with a higher metal content is often observed as a cause for increased absorption in dielectric optical coatings, which are mainly composed of binary metal compounds. Absorptance values of optical components are determined by laser calorimetric techniques described in ISO 11551. In this measurement scheme, the temperature increase in the sample is monitored during irradiation with a well-characterized laser beam. After this heating phase, the laser beam is blocked and the temperature curve during the cool-down cycle is also recorded. On the basis of the measured temperature development, the absorptance can be calculated taking into account the laser power and the heat capacity of the specimen.

All mechanisms, which deflect radiation from its specular direction, contribute to optical scattering representing the second major loss channel in optical coatings. Since scatter losses are mainly attributed to microstructural imperfections and the surface roughness of the layers, a scaling with wavelength can be deduced from the corresponding theoretical models. Especially in the VUV/UV spectral range, high scatter values are apparent and may dominate the losses of the optical components. For longer wavelengths, the contribution of scattering is decreased to a point in the MIR spectral range where it can be neglected in comparison to the absorption losses. Scatter effects

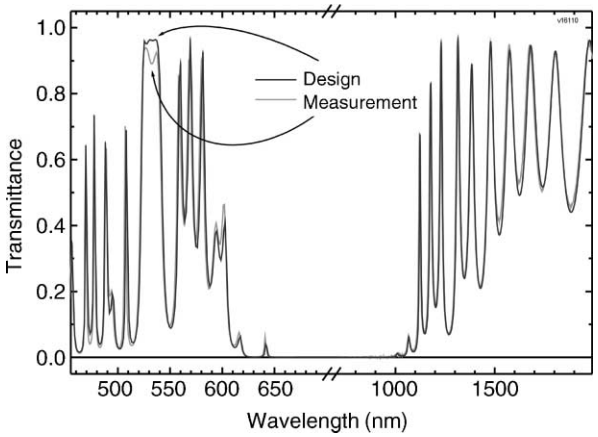


Figure 11 Example of a complicated layer system deposited under the control of an advanced optical broadband monitoring device: double chirped mirror (56 non-QWOT layers of SiO_2/TiO_2). Only small deviations between the calculated spectrum and the spectrum of the produced coating can be observed.

Table 1 Selected quality parameters of optical coatings and surfaces in conjunction with corresponding ISO standards and measurement principles

Specification	*Parameter/Unit*		*Standard/Measurement principle*
Laser induced damage threshold (LIDT)	cw-LIDT	W/cm	ISO 11254-1: cw-laser irradiation
	1 on 1-LIDT	J/cm^2	ISO 11254-1: irradiation with single pulses
	S on 1-LIDT	J/cm^2	ISO 11254-2: repetitive irradiation with pulses
	Assurance	J/cm^2	DIS 11254-3: irradiation sequence
Optical losses	Absorptance	ppm	ISO 11551: laser calorimetry
	Total scattering	ppm	ISO 13696: integration of scattered radiation
Transfer function	Reflectance	%	CD 13697: precise laser ratiometric method
	Transmittance	%	DIS 15368: spectrophotometry
Surface quality	Scratch/Digs		ISO 10110: 13 parts containing different types of imperfections
	Roughness		
Stability	Abrasion		ISO 9211: different test methods
	Environmental Stability		ISO 9022: 21 parts containing a variety of conditioning methods

may impair the imaging properties or the resolution, and they reduce the efficiency of an optical system. Besides the economic aspect of each loss effect, scattering can also be considered as relevant for safety in high-power laser systems, where some 100 ppm of scattered radiation carry enough power to injure or blind the operator. In the ISO 13696 standard, measurement procedures based on the collection or integration of the total radiation scattered by the sample are described.

Depending on their application, coatings have to withstand also manifold mechanical, chemical, and environmental influences. Severe mechanical and chemical requirements are imposed onto coatings for ophthalmology or consumer optics, which are often cleaned with abrasive and aggressive cleaning solvents and tissues. To qualify coatings for these applications, various testing procedures have been developed and are routinely employed in industrial quality management. Coatings for laser systems and precision optics are predominantly certified according to conditioning methods described in the ISO standard series 9211. Test procedures with respect to environmental influences are extensively outlined in the more than 20 parts of ISO 9022, which is dedicated to a variety of climatic, mechanical, and chemical conditioning cycles.

Summary and Outlook

Optical coatings are the major tool for engineering the properties of functional optical surfaces for an extensive spectrum of applications. On the basis of the matrix algorithm, a comprehensive theoretical model for the spectral transfer characteristics of optical coatings is available, which has been transferred to commercial computer programs. Now, most design problems can be solved with these computer codes, whereas the reproducible and precise production of optical coatings with high quality could only be partially achieved until today (Table 2). Besides conventional thermal processes, new ion-assisted deposition processes and sputtering concepts are gaining importance in optical thin-film technology. For the quality management of optical coatings, a variety of often standardized testing procedures has been developed during the last two decades including the determination of the power handling capability, the measurement of the optical losses and a broad spectrum of conditioning methods for optical coatings.

Demanding future challenges will be imposed on optical coatings by the development of semiconductor lithography in the EUV spectral range to achieve nodes well below 100 nm in chip production. Another important trend is directed by new laser applications in material processing, medicine and communication technology, operating at extremely high repetition rates and power levels, or with ultrashort pulses in the fs-regime. Other requirements will be defined by the rapid development of optical technology towards a broad range of special products with short innovation cycles obligating significant progresses in thin-film technology in the direction towards flexible and reproducible process concepts

Table 2 Selected quality parameters of optical coating systems for laser applications

Laser, wavelength	*Type*	*Absorption ISO 11551*	*Total scattering ISO 13696*	*Laser induced damage threshold, ISO 11254*
157 nm, F_2 excimer	HR/th		1–4%	
193 nm, ArF excimer	AR/th	0.7–2.5%	0.2–0.5%	1–2 J/cm^2 (1on1, 20 ns)
	HR/th	0.4–2.0%	0.2–2.5%	2–4 J/cm^2 (1on1, 20 ns)
248 nm, KrF excimer	AR/th		<0.025%	10 J/cm^2 (1on1, 30 ns)
	HR/th	<500 ppm	<0.2%	>20 J/cm^2 (1on1, 30 ns)
	HR/IB		<0.1%	>3 J/cm^2 (1on1, 30 ns)
633 nm, HeNe laser	HR/th	<30 ppm	<30 ppm	–
	HR/IB	<5 ppm	<5 ppm	–
1.064 μm, Nd:YAG	AR/th	<20 ppm	<100 ppm	>60 J/cm^2 (12 ns, 0.25 mm)
	HR/th	<50 ppm	<100 ppm	>100 J/cm^2 (12 ns, 0.25 mm)
	HR/IB	<1 ppm	<1 ppm	>80 J/cm^2 (12 ns, 0.25 mm)
10.6 μm, CO_2 laser	AR/th	<0.16%	–	>20 J/cm^2 (100 ns, 1.4 mm) >2 kJ/cm^2 (1.2 ms, 250 μm) >3 kW/mm (cw, 100 μm)
	HR/th	<0.10%	–	>25 J/cm^2 (100 ns, 1.4 mm) >2 kJ/cm^2 (1.2 ms, 250 μm)

Types: HR: highly reflecting mirror; AR: antireflective coating; th: thermal evaporation; IB: ion beam sputtering.

for coatings with increasing complexity. In all fields of modern optics, optical thin films will represent an enabling technology which will play a key role for many future applications and products.

List of Units and Nomenclature

A_0	amplitude of partial wave of order zero
A_1	amplitude of first-order partial wave
A_k	amplitude of the partial wave of order k
d_i	thickness of layer number i
D_i	film thickness expressed in units of QWOT
E_0	amplitude of a plane wave
E_i, H_i	electric field strength and magnetic field strength at the rear front of layer number i
IAD	ion-assisted deposition
IBS	ion beam sputtering
k	wavenumber of a plane wave $k = 2\pi/\lambda$
M_i	matrix of layer number i
M_S	composite matrix of a layer stack of K single layers
n_0	index of refraction of the ambient medium
n_H	refractive index of a high-index material
n_i	refractive index of layer number i
n_L	refractive index of a low-index material
n_T	refractive index of the substrate
QWOT	quarter wave optical thickness: unit for the thickness of the layer
r_1, r_1'	reflection coefficients of the ambient–layer interface
r_2	reflection coefficient at the layer–substrate interface
r_S	reflection coefficient of a single layer
r_{SK}	reflection coefficient of a layer stack of K single layers
R_S	reflectance of a QWOT stack
t_1, t_1'	transmission coefficients of the ambient–layer interface
δ_i	phase shift of a plane wave in layer number i
λ	wavelength of a plane wave
λ_Z	design wavelength
θ_i	angle of incidence in layer number i
ω	frequency of a plane wave

See also

Optical Coatings: Laser Damage in Thin Film Coatings.

Further Reading

Guenther AH (ed.) (2002) *International Trends in Applied Optics*. Bellingham, Washington: SPIE Press.

ISO 11551 (1997) *Test Method for Absorptance of Optical Laser Components*. International Organization for Standardization.

ISO 13696 (2002) *Test Method for Radiation Scattered by Optical Components*. International Organization for Standardization.

ISO 9022 (1996, 1997) *Optics and Optical Instruments – Environmental Test Methods – Parts 1–21*. International Organization for Standardization.

ISO 9211 (1994–2001) *Optical Coatings, Parts 1–4*. International Organization for Standardization.

ISO/FDIS 15368 (2003) *Measurement of Reflectance of Plane Surfaces and Transmittance of Plane Parallel Elements*. International Organization for Standardization.

Jacobson MR (1989) *Deposition of Optical Coatings*. SPIE Milestone Series, Volume MS 6. Bellingham, Washington: SPIE Press.

Kaiser N and Pulker HK (2003) *Optical Interference Coatings*. Berlin: Springer.

Liddell HM (1980) *Computer-Aided Techniques for the Design of Multilayer Filters*. Bristol, UK: Adam Hilger.

Macleod HA (2001) *Thin Film Optical Filters*. Bristol, UK: Institute of Physics Publishing.

Rancourt JD (1996) *Optical Thin Films*. Bellingham, Washington: SPIE Press.

Thelen A (1988) *Design of Optical Interference Coatings*. New York: McGraw-Hill.

X-Ray Coatings

P Dhez, Université Paris Sud, Orsay, France

© 2005, Elsevier Ltd. All Rights Reserved.

In all spectral ranges, there exist two principal options for coating an optical element. This may be a simple layer, as for example a metallic layer is frequently used in the visible spectral region. The second option is to use a multilayer coating, making use of the optical properties of two or more materials to achieve enhanced reflectivity over a limited spectral range. In both cases, a knowledge of the refractive

index of the materials (real or complex) is required, and of course one must choose materials that can be coated in a practical deposition process. A frequent side-advantage of coating is to make it possible to choose the material of the substrate for its mechanical or cooling properties without regard to the optical properties of the coatings.

Originally (at the beginning of the twentieth century) the progress of X-ray optics was hampered by a poor understanding of the nature of these rays, and in particular by the observation that they are very weakly reflected and refracted. After the electromagnetic nature of these rays had been confirmed, the problem of producing polished surfaces that were sufficiently smooth on an atomic scale to reflect X-rays still remained. Optical coatings gave little hope of improvement due to the growth of roughness during the coating process. Thus early attempts to make X-ray mirrors produced poor results and little progress was made until the 1960s. However, in the past 30 years, much technical progress has been made, and now many optical principles used in the visible spectral region have been extended to very short wavelengths. Even refractive and diffractive focusing lenses (not treated here) have been implemented in the X-ray region.

In the first section we outline the history of the discovery of the properties of X-rays, which were achieved mainly in experiments using crystals. This allows us to understand how their electromagnetic nature was discovered, and how they interact with materials. We then describe the phenomenon of the 'total reflection' of X-rays, which occurs only at very high angles of incidence ('grazing incidence') and its main applications in X-ray optics. Finally, we outline the principles of mirrors based on interference, their applications in the X-ray region, their advantages and their limitations.

X-rays Compared to the Other Electromagnetic Spectral Regions

Infrared radiation was discovered by W. Herschel in 1800, and ultraviolet rays by J. Ritter in 1801. These were the only 'invisible rays' known before Maxwell's contribution to the understanding of the electromagnetic spectra through the development of his equations in 1869. The discovery of radio waves by H. Hertz in 1887 confirmed that our eyes are blind to most of the electromagnetic spectrum.

In 1895, W. Roentgen discovered some previously unknown 'rays' that had the surprising property of being able to penetrate low-density material opaque to visible light. This discovery was a side-effect of his studies of low-pressure electrical discharges, fashionable at the time. The nature of these unknown or 'X'-rays as Roentgen named them, remained unknown for 12 years or so. Were they electromagnetic waves or particles? Roentgen used lenses, prisms and mirrors to try to detect a reflective, refractive or focusing effect, with inconclusive results. Although the electromagnetic nature of the X-rays was favored, the magnitude of the frequency or the refractive index was unknown. We note that natural radioactivity and gamma rays were discovered a year after X-rays, and the same questions concerning their nature were raised. However, in neither case did the ignorance of the nature of the radiation impede their use in medical radiography or in cancer therapy.

In 1906, the experiments of C. Barkla provided the first evidence of the electromagnetic nature of X-rays. In scattering X-rays from blocks of a light element (in this case, carbon), he was able to show that the scattered radiation was polarized, and therefore had the nature of a transverse wave. This kind of experiment had been performed almost a century previously to demonstrate the transverse wave nature of visible light. In 1916, following a suggestion by M. von Laue that X-rays might be diffracted from a grating of period comparable to their wavelength, Friedrich and Knipping succeeded in diffracting X-rays from the atomic planes of a natural crystal. This experiment established definitively the electromagnetic nature of X-rays and also the scale of their wavelength.

In 1915 W. L. Bragg summarized the diffraction rules for crystals through his equation [1], now known as the simplified Bragg law:

$$2d_i \sin\theta_i = N\lambda_i \qquad [1]$$

This equation relates the diffracted X-rays wavelength λ_i to a set of parallel atomic planes separated by a distance d_i and to the incidence angle θ_i. N is an integer indicating the order of diffraction, as in the grating equation. In using this equation, care is needed because of two peculiarities compared to the usual mirror or grating equations in the visible range. In the Bragg equation, the angle of incidence θ is measured from a given atomic plane to the incident rays. However, inside a crystal several different atomic planes exists, not only the planes parallel to the surface as indicated in **Figure 1**. In addition, θ is not measured from the normal as is usual in optics; hence the angle θ is referred to as the glancing or grazing angle of incidence.

We note that the Bragg equation is not restricted to X-rays, but is a coherent scattering phenomenon which is common to all cases where waves, such as

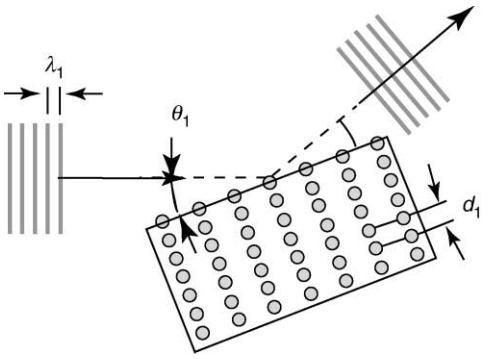


Figure 1 The geometry for the conditions leading to selective reflection on a periodic medium.

sound waves or ripples on water, are diffracted by regular arrays of scattering centers.

In order to visualize the phenomena, we refer to Figure 1 which is a schematic representation of a plane wave incident upon a two-dimensional array of scattering centers (electrons around nucleus, in the case of X-rays), shown here as dots. As the sine function has a maximum value of 1, the maximum wavelength λ that can be reflected from a spacing d is given by $\lambda = 2d_{max}$. As the largest value of $2d$ that can be found in natural crystals is about 2 nm (in natural muscovite mica, for example), X-rays of 2 nm, with an equivalent energy of about 620 eV, are the 'softest' (least energetic) that can be reflected in this manner.

The model shown in Figure 1, similar to that proposed by M. von Laue, does not take into account the detailed interaction of the atoms with the X-rays, and therefore ignores effects such as absorption and refraction, and discrete atomic transitions. A.H. Compton attempted the first treatment of the interaction of matter with X-rays from an atomic viewpoint. In 1917, he developed a classical theory in which the electrons bound to the nucleus are treated as harmonic oscillators. In this way, he was able to derive an expression for the atomic scattering coefficients and to describe resonance effects at frequencies corresponding to specific atomic transitions. A few years later, he was able to establish a link between these 'atomic scattering factors' and the macroscopic concept of refractive index. This new formulation allowed the use of the Fresnel equations, widely used in other spectral regions, to calculate the intensity of X-rays reflected from smooth, polished surfaces. These surfaces may be amorphous as well as crystalline.

The refractive index can be written in several ways; since it is close to unity over the X-ray range, the most common form is complex:

$$n = 1 - \delta - i\beta \qquad [2]$$

where δ is the refractive index decrement, and the imaginary part β the absorption index. For photon energies corresponding to X-rays, both have small negative values, typically 10^{-3}–10^{-6}, and are correcting factors to the vacuum index 1. As in other photon energy ranges, the absorption and the refraction coefficient variations are linked through the Kramers–Krönig relations. In the visible range the index values are usually larger than 1, being below 1 only around specific optical thresholds. In the UV region, the index values become smaller than unity at energies above about 30 eV, which corresponds to the energy range for which atoms can be ionized.

At an early stage, Roentgen demonstrated X-ray absorption by sheets of materials. The demonstration of refraction took much longer and was more difficult. The first direct X-ray refraction index measurement was achieved in 1924, using a prism and an X-ray tube. Figure 2 is the scheme of the experiment, which is easy to perform in the visible range to measure the refraction of a transparent material. The footprint of the direct beam on the photographic plate is made before inserting the prism. In the visible range for which $n > 1$, the beam is refracted at larger angle than the glancing incidence angle, and so below the direct beam direction as shown in Figure 2a. At the contrary, as indicated in Figure 2b, the X-ray beam is refracted toward the prism surface and such an experiment effectively demonstrated that the index was slightly inferior to unity, as suggested by A.H. Compton. We will return

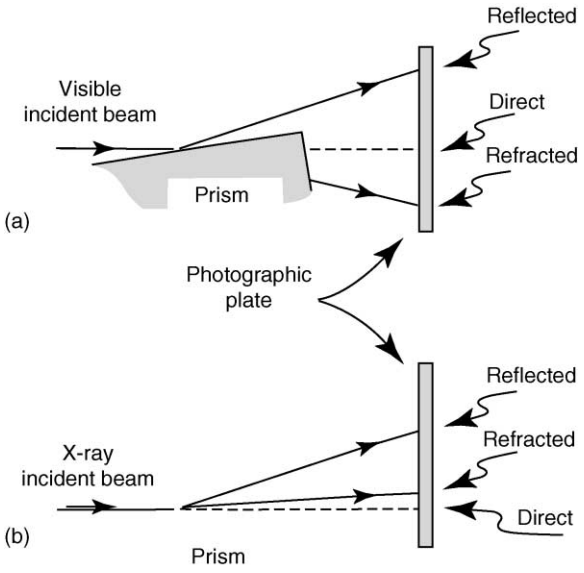


Figure 2 The geometry of the experiment used to evaluate the magnitude and the sign of the refractive index in the X-ray and the visible ranges.

to this peculiarity in the following section, when the total reflection phenomenon is described.

More details about the successive steps in our understanding and the pioneering works on X-rays can be found in standard textbooks.

Total Reflection X-Ray Mirrors

Using Fresnel's equations and tabulated values of the refractive index one can verify that, on any material, grazing incidence is necessary to reflect X-rays. Reflectivity calculations over the range from EUV to X-rays (30 eV to 10 keV), with approximations to allow calculations for rough surfaces, can be carried out using the web site of the Center for X-ray Optics of Lawrence Berkeley Laboratory http://www-cxro.lbl.gov, and curves similar to those presented below can be reproduced.

In the case of normal incidence, Fresnel's equation indicates that for any material the X-ray reflectivity $R^2 = |(n-1)/(n+1)|^2$ is very low, typically below 10^{-4}. This is a direct consequence of the X-ray index values, which are close to unity, as mentioned above.

In the visible range, normal incidence reflectivity from the surface of a transparent medium is low but can be increased just by going to glancing incidence. This is also the case in the X-ray region but in addition as the real part of the X-ray refractive index is lower than that of the vacuum, we have an 'external' total reflection phenomenon. From Figure 2 one can see how, for a sufficiently small incidence angle, the refracted wave becomes parallel to the interface. This is well known in the visible range (in which case it is the 'internal' reflection) and the phenomenon is used, for example, in fiber optics.

Figure 3 shows examples of calculated reflectivity variations versus grazing incidence angles, for perfectly smooth glass (SiO_2), nickel, and platinum surfaces. Curves are drawn for the 8 keV energy photons, corresponding to the Cu Kα line produced by conventional X-ray tubes. One observes that for this energy the reflectivity is low beyond a grazing angle around 1°, then rises quickly at low angles. This is typical of the total reflection phenomenon. The grazing angle corresponding to a 50% reflectivity is called the critical angle Θ_c, its value being roughly proportional to the refractive index decrement at the relevant photon energy. Notice that weaker absorption materials, lower Z, lead to more abrupt reflectivity variation but with smaller critical angles.

Such reflectivity variations at grazing incidence were systematically studied with X-ray tubes as early as the 1930s. The penetration depth of X-rays at grazing incidence was also studied. The evanescent wave inside the material, under total reflection conditions, is confined to a surface layer of a few tens of wavelengths. Hence, in the total reflection of X-rays, a coating having a thickness of a few tenths of a micron, on any substrate, is all that is required.

Figure 4 shows examples of the variation in nickel reflectivity calculated over the energy range 0.3–10 keV, for several grazing incidence angles. Nickel is one of the coatings commonly used for X-ray mirrors. The abrupt reflectivity variations and the dips correspond to the energies of the nickel atomic thresholds L (0.85 keV) and K (7.5 keV) for which the index tables predict a rapid change in absorption and refraction coefficients. Choosing an incidence angle provides a practical way of limiting the reflected spectral range. The energy cut-off of the

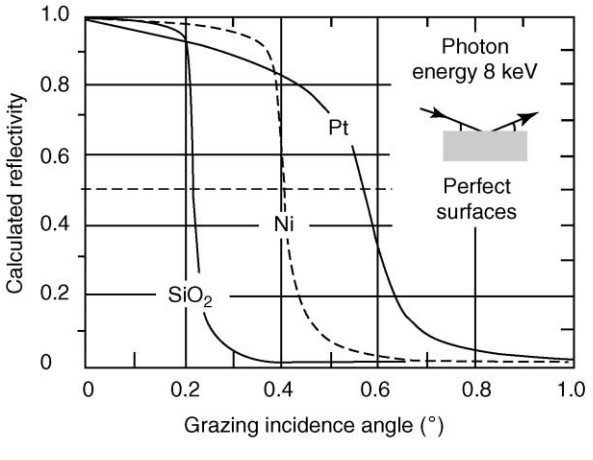


Figure 3 Examples of calculated reflectivity variations versus the incidence angle in the case of 8 keV X-rays on perfect platinum, nickel and glass surfaces.

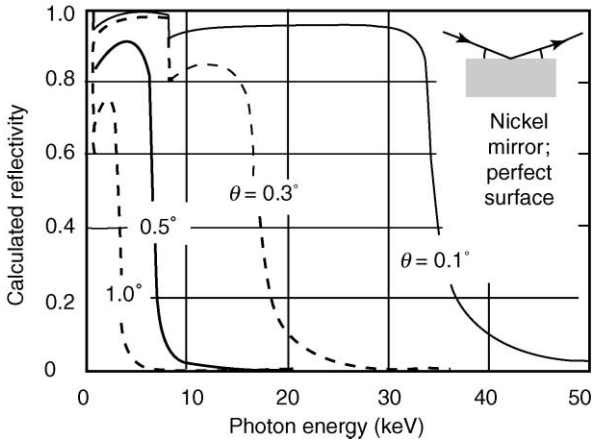


Figure 4 Calculated reflectivity variations versus photon energy, on a perfect nickel surface, for different incidence angles.

X-ray mirror acts as a filter for the higher-energy photons. That is effectively the method of suppressing the short wavelengths from an incident spectrum if they are not desired in an experiment.

When using highly divergent sources such as X-ray tubes, the restriction to operation at grazing incidence leads to a strong limitation of the collected solid angle, and thus of the beam size and angular aperture. These constraints are less stringent on highly collimated beams emitted by synchrotron sources. Thus, most synchrotron beam lines use total reflection optics in the form of curved mirrors, to refocus the beams and remove the undesired high X-ray energies. Recently, X-ray imaging optics, consisting of a series of parabolic and hyperbolic nested grazing incidence mirrors to increase the collected aperture, have been launched on satellites.

The reflectivity curves in Figures 3 and 4 are calculated for perfect surfaces. Real surfaces, even polished surfaces, have roughness created by defects over a large and continuous range of spatial frequencies, ranging from several centimeters down to the atomic scale. Coating a surface replicates the large period defects. Eventually new defects are added at near-atomic scale by nucleation of the coating material. Interferometric methods have been developed to study the range of defects from centimeter to micron spatial periods, whereas the atomic force microscope (AFM) is able to measure roughness from the micron down to the 0.1 nm scale. As waves are scattered most efficiently by asperities of a size comparable to their wavelength, scattering measurements are used in the X-ray region to probe defects of atomic scale. The prevalence of scattering explains why mirrors that reflect well in the visible range were disappointing when used in early X-ray experiments. The recent development of adaptive polishing methods has almost solved this problem.

The main effect of scattering by an optical system is not to decrease the number of photons coming out, but to modify their angular distribution. Scattering measurements are experimentally studied to test and evaluate the spatial defect distribution. In X-rays, one first characterizes the incident beam angular distribution in measuring its angular variation intensity after a scanning slit. Performing the same measurement after the optic, across the reflected beam in the case of mirrors, provides the modified angular distribution. From such a measurement, a few theoretical models adapted to the X-ray range can be used to deduce the mirror spatial defect distribution and so to check progress in polishing methods. The simplest

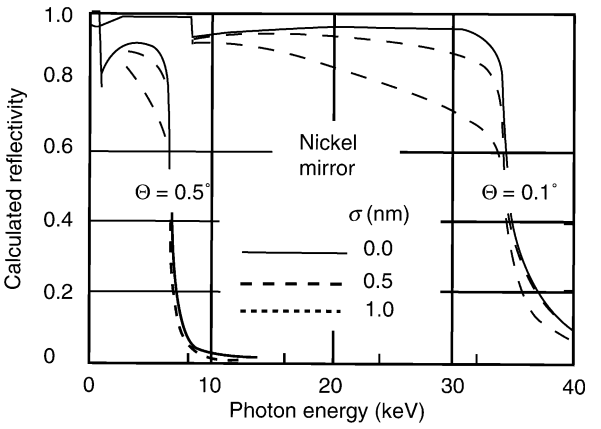


Figure 5 Calculated reflectivity variations versus photon energy, on nickel surfaces having different roughness, for different incidence angles.

model, coming from early crystal studies, is describing defects by a single parameter named roughness and denoted σ. Figure 5 shows an example of calculated reflectivity; on nickel surfaces versus photon energy and for increasing roughness; it shows how many photons have deviated from the symmetrical direction. Corresponding angular distributions can also be calculated.

The coating systems used for X-ray mirrors are material dependent. Deposition methods are similar to those used for visible optics coatings, but specific effort has been devoted to avoid nucleation growth to achieve uniform smooth layers. In general, one tries to give the evaporated particles, ions or clusters a kinetic energy lying within a specific range, which promotes mobility on the surface but is below the energy needed for implantation. In the case of electron beam deposition the energy distribution is thermal, but an auxiliary ionic bombardment can be used during deposition to provide the necessary energy to the previously deposited atomic layer (ion assisted deposition). For each material, the specific sputtering deposition and ionic bombardment conditions need to be carefully studied.

Until the 1970s, Bragg reflection by crystals and the total reflection phenomenon was the only way to reflect X-rays.

Multilayer X-Ray Mirrors

In the visible range, optics are frequently coated with stacks of dielectric multilayers obtained by periodic deposition of layers of two amorphous transparent materials. Such stacks are highly selective reflectors, or bandpass filters. They operate on the principle of

constructive interference between all the waves partially reflected by the successive interfaces. They can be regarded as analogous to two-dimensional Bragg reflectors, in which the interfaces play the same role as atomic planes in crystals. Hence the Bragg equation [1] can be applied, the relevant period d being the thickness of an individual bilayer, if the stack is truly periodic.

The concept of using artificially deposited stacks of thin films, of alternating high and low electron density, to reflect and diffract X-rays dates back at least to the 1920s. It was not until 1944, however, that DuMond and Youtz were able to demonstrate significant X-ray reflection from such an artificially layered structure. Their work used alternating layers of copper and gold, laid down by evaporation. Unfortunately these materials diffuse into one another, reducing the electron density contrast and hence the X-ray reflectivity. In 1967 Dinklage and Frerichs succeeded in making a structure of alternating layers of iron and magnesium, which do not interdiffuse quickly, and thus produced the first stable X-ray reflecting multilayer.

Since then, enormous progress has been made in the fabrication and application of such multilayer reflectors, principally because of the efforts of two pioneers. Troy Barbee, while at Stanford University, developed open loop sputtering methods for controlling the thickness of the many thin (~nm) layers required to make up such a stack, while Eberhard Spiller at IBM developed an evaporation method using *in situ* X-ray monitoring to achieve similar thickness control. Almost all of the many multilayer deposition systems in use around the world are based on one of these two methods.

As for the total reflection mirror, only recent progress in methods for deposition and controls has made this new kind of X-ray optics practical. Present applications range from the extreme ultra violet (30 eV) region to hard X-rays having energies of a few tens of keV.

Prediction of the reflectivity variations can be simulated by calculations on the same web site mentioned in the preceding section. Multilayer mirrors are based on interference, so the reflectivity coefficient and the width of the reflected bandpass are determined by the amount of absorption inside the stack and the partial reflectivity at each interface. The more efficient pairs of materials have the lowest X-ray absorption and the highest refractive index difference at the required energy. The selection of real materials is constrained by the coating conditions and the need to achieve smooth interfaces. In particular, a low interdiffusion coefficient between the successive material pairs is a requirement for temporal stability. Epitaxial deposition methods have been considered but no useful pair of materials has yet been found.

An example of a flourishing application is multilayer-coated normal incidence X-ray optics. Optical schemes used in the visible region for telescopes and microscopes have excellent optical performance and high numerical aperture but a very low reflectivity in the X-ray region if coated with a single layer. However, if coated with multilayers they can be used, providing that extremely low-roughness surfaces can be produced. The interface qualities of the deposited thin layers presently limit normal incidence optics to the soft X-ray energy range, typically below 300 eV so around the carbon edge. This corresponds to periods of about 2 nm, or a few atomic layers, and the resulting roughness seriously decreases the reflectivity.

Figure 6 is a reflectivity curve versus energy for a quasinormal incidence Mo/Si multilayer mirror optimized to get the first-order peak at 93 nm. On Mo/Si mirrors, precise control of the layer period and the quality of the interfaces allow the manufacture of mirrors with reflectivity only a few percent below the theoretical limit shown here. Considerable effort has been invested in such mirrors, with a view to applications to EUV projection lithography.

It is possible to build polarizer mirrors over the soft X-ray range. An incidence angle of 45° corresponds to the Brewster angle, because the X-ray index is close to unity. X-ray multilayer and semitransparent selective beam splitters now permit the development of Michelson interferometers for newly developed soft X-ray lasers.

Figure 7 is an example of a tungsten/carbon (W/C) reflectivity curve for a multilayer adapted to the 8 keV energy of Cu Kα, mentioned before. The curve

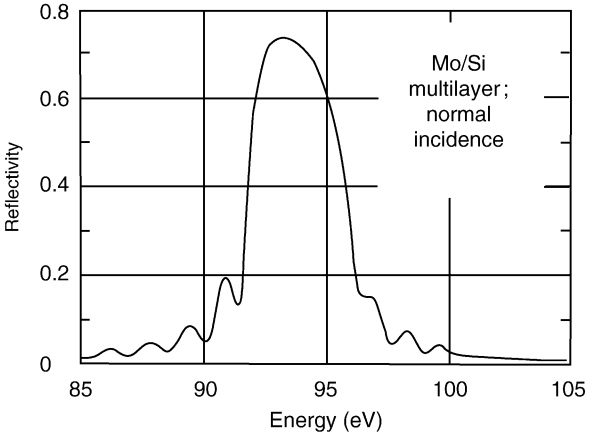


Figure 6 Calculated reflectivity versus incidence angle for a typical Mo/Si multilayer, optimized around 93 eV, under quasi-normal incidence.

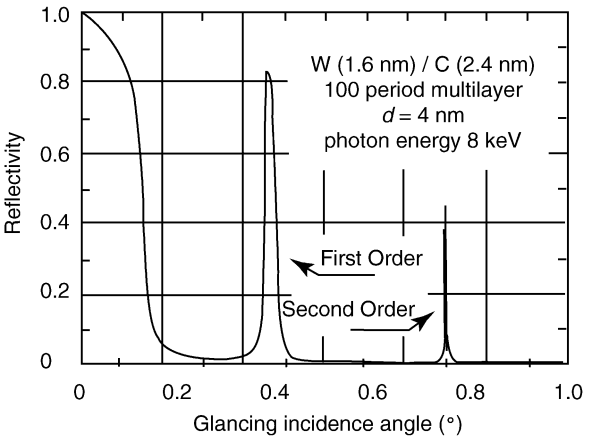


Figure 7 Calculated reflectivity versus incidence angle for a typical W/C multilayer adapted to reflect 8 keV photons.

of reflectivity versus glancing angle has been calculated from the total reflection range to the second-order peak. This kind of multilayer is being used in X-ray tubes to premonochromatize the beam and make it parallel. In such an application, the period of the coating is continuously graded laterally to satisfy locally the Bragg law at each point of the curved mirror.

As a final example of the new possibilities offered by X-ray multilayers, we mention the possibility of widening the reflected bandpass by varying continuously the period from the bottom to the top of the deposited stack. This is the equivalent of multilayer neutron 'supermirrors'. Such interference mirrors are usable at quite high X-ray energies, and give high reflectivity and a larger incidence angle than total reflection mirrors working over the same energy range.

The continuous fast development of coating technologies and X-ray multilayer applications can be followed through the proceedings of dedicated meetings at, for example, on a web site like http://PXRM.

Summary and Perspectives

After the discovery of X-rays in 1895, almost 25 years were necessary to understand their characteristics compared to the visible range. Only in 1923 did F. Holweck and R.A. Millikan demonstrate the continuity of electromagnetic radiation from X-rays to ultraviolet. The selective Bragg reflection by natural crystals is limited to energy commonly larger than about 2 keV photons. Below this energy the total reflection mirror was the only optical device to reflect and focus wavelengths shorter than the far UV until the development of adapted interferential mirrors.

During the 1960s, the EUV and X-UV ranges gained interest due to the spread of fusion research and in the astrophysics community studying the Sun and stars. For both communities, the continuum and the lines emitted at short wavelengths by ionized matter were a convenient tool to evaluate high plasma temperatures. At about the same time, the development of synchrotron radiation sources increased the need for X-ray optics. In addition, this last source offered the possibility to attempt new optical solutions with convenient intense and parallel beams from UV to X-rays. Finally, the concomitant progress in various microtechnologies afforded the necessary tools to produce new adapted optical devices for short wavelengths. At the present time, adapted optics are under development to take advantage of various applications from microscopy to interferometry over the very large electromagnetic energy X-ray range.

As a matter of fact, it is now possible to tailor a choice of mirror coatings adapted to each X-ray photon energy and application. Such coatings are specific because they must be optimized to take into account the inevitable absorption energy dependence.

Progress achieved in blank polishing and evaporation methods has led to efficient and high-quality X-ray mirrors even able to keep the spatial coherence of synchrotron sources. Such technological advances are also steps to master surfaces and interfaces at the atomic scale. They are now becoming applicable to the other spectral ranges and to various micro- and nanotechnologies.

List of Units and Nomenclature

Crystal period	d	[nm]	$1\ \text{nm} = 10^{-9}\ \text{m}$
Incidence Angle	Θ	[degree]	
Photon Energy	E	[eV]	$(h/2\pi)\omega\lambda = hc = 1239.842\ \text{eV.nm}$
Wavelength	λ	[nm]	$1\ \text{nm} = 10^{-9}\ \text{m}$

See also

Diffraction: Fresnel Diffraction. **Geometrical Optics:** Lenses and Mirrors. **Incoherent Sources:** Synchrotrons. **Optical Coatings:** Thin-film Optical Coatings.

Further Reading

Attwood D (1999) *Soft X-Rays and Extreme Ultraviolet Radiation*. Cambridge: Cambridge University Press.

Ditchburn RW (1963) *Light*. Appendix XIX, 2nd edn. London: Blackie.
Hecht E (1998) *Optics*, 3rd edn. Reading, MA: Addison-Wesley.
Henke BL, Gullikson EM and Davies JC (1993) *Atomic Data and Nuclear Data Tables* 54: 181.
James RW (1982) *The Optical Principles of the Diffraction of X-Rays*. Woodbridge, CT: Ox Bow Press.
Michette A and Pfauntsch S (1997) *X-Rays: The First Hundred Years*. New York: John Wiley.
Mould RF (1993) *A Century of X-Rays and Radioactivity in Medicine*. Bristol, UK: Institute of Physics Publishing.
Paratt LG (1954) Surfaces studies of solids by total reflection of x-rays. *Physical Review* 95: 359.
Spiller E (1972) Low-loss reflection coatings using absorbing materials. *Applied Physics Letters* 20(9): 365–367.
Underwood JH and Attwood DT (1984) The renaissance of x-rays optics. *Physics Today, April*: 44–51.

OPTICAL COMMUNICATION SYSTEMS

Contents

Basic Concepts
Historical Development
Architectures of Optical Fiber Communication Systems
Free Space Optical Communications
Lightwave Transmitters
Local Area Networks
Optical Time Division Multiplexing
Wavelength Division Multiplexing

Basic Concepts

S Lee and A E Willner, University of Southern California, Los Angeles, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since modern lightwave communication started in the 1970s, extensive efforts have been made to increase data rate and transmission distance. Recently, the bit rates for single-channel and multichannel systems have exceeded 40 and 2 Tb/s on single fibers respectively, for transoceanic distances. One revolutionary development in lightwave systems has been the deployment of erbium doped fiber amplifiers (EDFAs). These amplifiers facilitate transmission of an optical signal over long distances, by providing periodic analog-like amplification rather than digital-like regeneration. The wide bandwidth provided by EDFAs has made possible the increased use of wavelength division multiplexing (WDM) in which multiple lightwave signals, each having a different wavelength, are co-propagated on a single fiber. EDFAs and WDM techniques can enhance the lightwave system capacity, both in terms of obtainable transmission distance, and total number of data rates (see Figure 1).

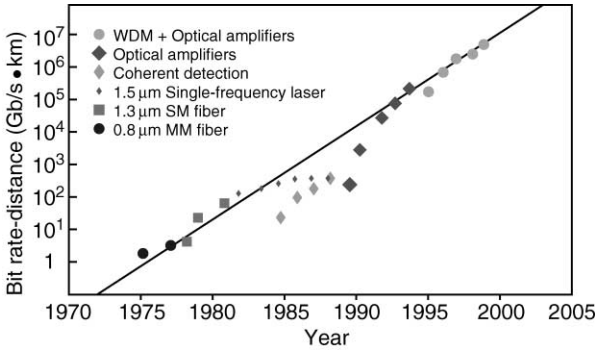


Figure 1 Bit rate product. Data from T. Li (private communication).

Optical Fiber

Most telecommunication fibers are made of silica, therefore following the attenuation of silica material. Figure 2 shows the silica fiber attenuation as a function of wavelength. The downward slope towards 1.6 μm and upward slope away from 1.6 μm, are the theoretical limits due to Rayleigh scattering and absorption of silica material, respectively. The absorption peak near 1.4 μm is due to water molecule absorption. There are two low attenuation regions: ~1.3 μm and 1.55 μm, both of which are used for communications in general, and some other wavelengths are also used for shorter distance communications. Figure 2 also shows the comparison between gain bandwidth of EDFA (~3 THz) and the low attenuation window around 1.55 μm (~25 THz). The typical values of attenuation at 1.3 μm and 1.55 μm for single-mode fiber (SMF) are ~0.35 dB/km, and 0.2 dB/km, respectively.

When an electromagnetic wave interacts with the bound electrons of a silica fiber, the medium response depends upon optical frequency (ω). This property, referred to as material dispersion, manifests itself through the frequency dependence of the refractive index $n(\omega)$. Fiber material dispersion plays a critical role in propagation of short optical pulses since different spectral components associated with the pulse travel at different speeds, given by $c/n(\omega)$. Consequently, the optical pulse at the output of the fiber is broadened in time. The commonly used system parameter is called the dispersion parameter D (the negative D value is called normal dispersion, and the positive D value is called anomalous dispersion), and is measured in ps/(nm·km); D for SMF is ~+17 ps/(nm · km) at 1.5 μm. However, because of dielectric waveguiding, the effective mode index is slightly lower than the material index $n(\omega)$, with the reduction itself being ω-dependent (waveguide dispersion). This results in a waveguide contribution that must be added to the material contribution to obtain the total dispersion (see Figure 3).

Generally, the waveguide contribution to dispersion D is to shift zero dispersion wavelength (λ_0) to longer wavelengths. Furthermore, the waveguiding can be tailored to shift the λ_0 from 1.3 μm to 1.5 μm, the wavelength at which the fiber has the minimum attenuation loss. The fiber having λ_0 in the neighborhood of 1.5 μm, is called dispersion-shifted fiber (DSF); D for DSF is usually between −2.5 and +2.5 ps/nm·km, at a wavelength of 1.5 μm. The typical dispersion parameter, D, as a function of a wavelength for both SMF and DSF, is shown in Figure 4.

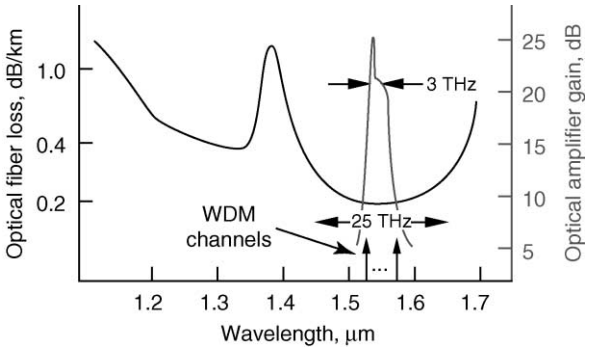


Figure 2 Silica fiber attenuation versus wavelength.

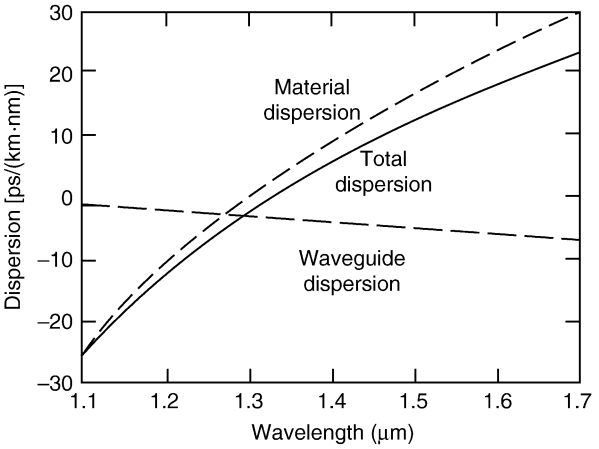


Figure 3 Total dispersion (D) and contribution of material dispersion and waveguide dispersion in conventional SMF.

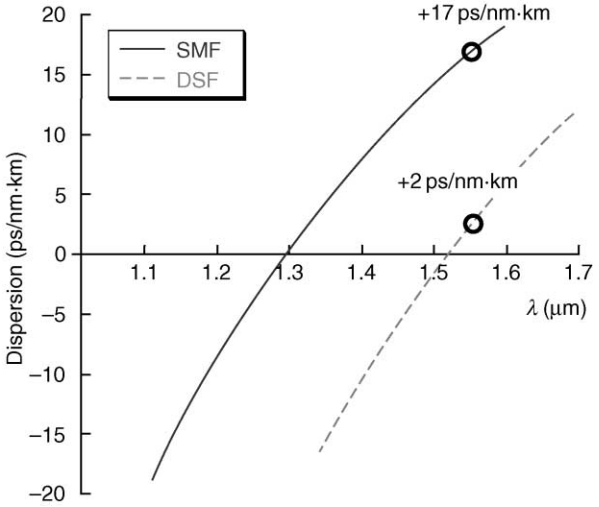


Figure 4 Dispersion versus wavelength for conventional SMF and DSF.

In addition to chromatic dispersion, optical fiber has another dispersion properties related to polarization. An optical wave of arbitrary polarization can be represented as the superposition of two orthogonally polarized modes. In an ideal fiber, these two modes are indistinguishable, and have the same propagation constants owing to the cylindrical symmetry of the waveguide. However, in real fibers there is some residual anisotropy due to unintentional circular asymmetry, usually caused by noncircular waveguide geometry or asymmetrical stress around the core, as shown in Figure 5.

In either case, the loss of circular symmetry gives rise to two distinct orthogonally-polarized modes with different propagation constants (differential phase velocity) responsible for polarization mode dispersion (PMD) in the fiber, and can be related to the difference in refractive indices (birefringence) between the two orthogonal polarization axes. The differential phase velocity indicated is accompanied by a difference in the group velocities for the two polarization modes, therefore the pulse at the end of the transmission fiber is broadened by this differential group delay (DGD). DGD is usually expressed in units of ps/km for a short length (1 m to 1 km) of birefringent fiber. However, DGD does not accumulate along a long fiber link in a linear fashion. Instead, because of random variations in the perturbations along a fiber span, the DGD in one section may either add to or subtract from another section of the fiber. As a result, average DGD in long fiber spans accumulates in a random-walk-like process that leads to a square root of transmission-length dependence. Therefore, average DGD is expressed in $ps/km^{1/2}$ in long fiber spans, referred to as the PMD of the fiber, and the typical PMD parameter has a value of 0.01 to 10 $ps/km^{1/2}$. Because of the many perturbations that act on a real-world fiber (e.g., temperature, vibration, etc.), transmission properties typically vary with time. Therefore, the PMD of the fiber link fluctuates randomly, thus causing random fluctuations in system performance.

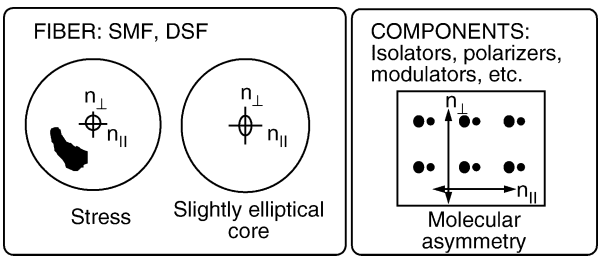


Figure 5 Origin of PMD.

It is well known that optical fibers show nonlinear behavior under conditions of high power and long interaction length. The power-times-distance products for amplified transmission systems can be large enough to make fiber nonlinear effects the dominant factor in the design of long-distance systems.

There are two categories of fundamental optical nonlinear effects: stimulated scattering effects and refractive index effects. Stimulated scattering effects arise from parametric interactions between light and acoustic or optical phonons in the fiber. Two nonlinear effects fall into this category: stimulated Brillouin scattering (SBS) and stimulated Raman scattering (SRS). The main difference between the two is that optical phonons participate in SRS, while acoustic phonons participate in SBS. In a simple quantum-mechanical picture, a photon of the incident field is annihilated to create a photon at a longer wavelength. The new photon is co-propagated and counter-propagated along the original signal in SRS, while counter-propagated in SBS. Refractive index effects are caused by modulation due to changes in light intensity. There are three types of refractive index effects: self-phase modulation (SPM), cross-phase modulation (XPM), and four-wave mixing (FWM). SPM introduces the change of optical phase by its own intensity, and, this leads to frequency chirp of the pulse, depending on the relative position from the peak. When this nonlinear frequency chirp interacts with the fiber dispersion, the pulse either broadens or compresses, depending on the sign and the amount of dispersion along the fiber. In XPM, the phase of the signal in one wavelength channel is modulated by the intensity fluctuation of the other wavelength channels. In a WDM system, XPM imposes far more damaging effects than SPM because XPM is stronger by a factor of two than SPM when the channel power is the same, and large number of WDM channels can contribute to XPM. FWM is the generation of modulation sideband at new frequencies, due to the phase modulation of channels between lights at different frequencies in multi-channel system. FWM causes penalties in a WDM system if the newly generated frequency is either equal to or close to the frequency of existing WDM channels.

Fiber nonlinearities impose significant degradation in optical transmission system, and limit the system on allowable number of channels, channel power, and channel spacing (see Figure 6). But, the effects can be well suppressed by carefully designing the fiber dispersion profile (i.e., dispersion map) in the transmission system, and controlling the optical launch power for each fiber span.

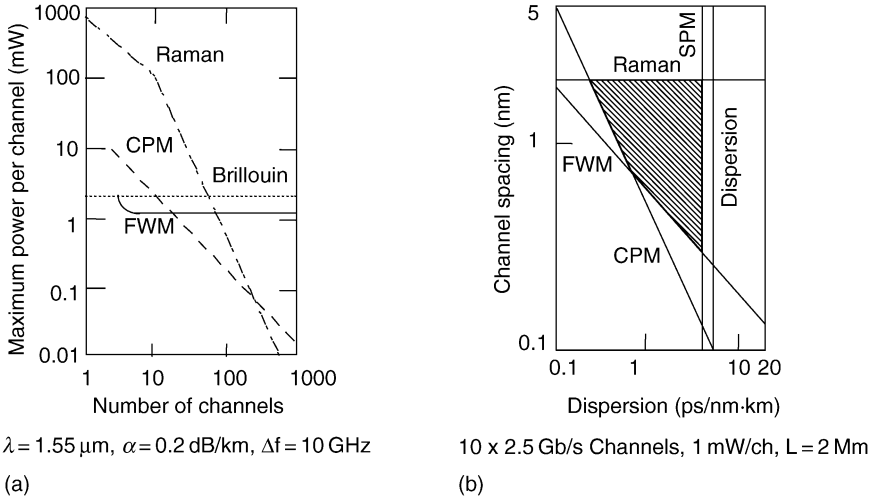


Figure 6 Limitation due to fiber nonlinearities in (a) maximum channel power and number of channels, and (b) channel spacing and dispersion. Reproduced with permission from Chraplyvy AR (1990) Limitation on lightwave communications imposed by optical-fiber nonlinearities. *IEEE Journal of Lightwave Technology* 8:1548–1557. Copyright © 1990 IEEE.

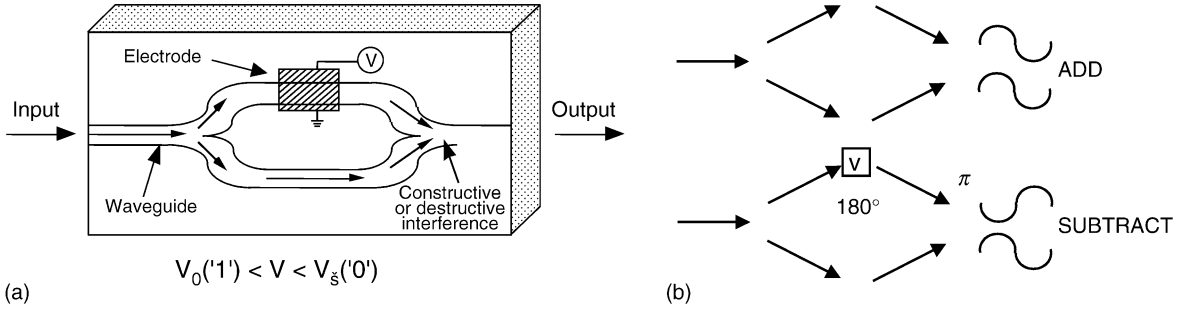


Figure 7 (a) EOM: MZ interferometer on LiNbO2, (b) operation of EOM; constructive and destructive interference resulting in intensity modulation.

Communication Components

Distributed feedback laser (DFB) is widely used as a light source for metro, long-haul, and undersea applications, due to its narrow spectral width, and wavelength stability. Fabry–Perot (FP) lasers, and vertical cavity surface emitting lasers (VCSEL) are used for local area networks (LAN), and for access applications, because these are cheaper than DFB lasers. For an application, where the cost is the most important parameter with low data modulation speed, light-emitting diode (LED) is used as a light source.

In optical data modulation, the simplest and easiest technique is turning ON or OFF the laser, depending on the binary logic '1' or '0' (direct modulation). This type of modulation is simple and cheap compared to external modulation technique. But, its application is limited by the bandwidth of modulation, and the induced frequency chirp (i.e., difference in optical frequency at the turn ON state and just before the turn OFF state of laser) that limits the transmission distance. Two different types of external modulators are widely used for digital optical data transmission. Electro absorption modulator (EAM) uses the Franz–Keldysh effect, where it is observed that the wavelength of the optical absorption edge in a semiconductor is lengthened by applying an electric field. Electro-optic modulator (EOM) utilizes the linear electro-optic effect, called the Pokels effect, which changes the refractive index of the material caused by and proportional to an applied electric field, so therefore changes the phase of the optical signal. This phase change is used to modulate the intensity of a lightwave through a Mach–Zehnder (MZ) interferometer (see Figure 7). Because the Pockels effect exists only in crystal, lithium niobate crystal ($LiNbO_2$) or electro-optic polymers are used for EOM.

EDFA is a wideband optical amplifier that has merits in that: (i) erbium ions (Er^{3+}) emit light in the 1.55 μm loss-minimum band of optical fiber, (ii) a circular fiber-based amplifier is inherently compatible with a fiber optics system; (iii) it provides

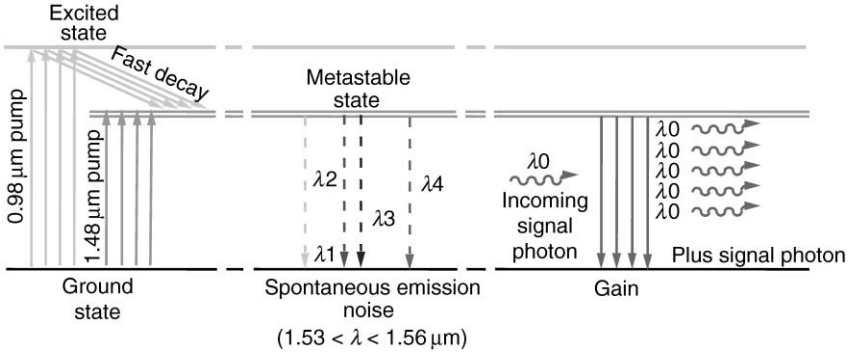


Figure 8 Energy level and transition diagram of EDFA. Reproduced with permission from Willner AE (1997) Mining the optical bandwidth for a terabit per second. *IEEE Spectrum* Apr.: 32–41. Copyright © 1997 IEEE.

amplification independent of bit-rate, modulation format, power, and wavelength; and (iv) it has low distortion and low noise during amplification. EDFA contains a gain medium (i.e., erbium-doped fiber) that must be inverted by a pump source. A signal initiates stimulated emission, resulting in gain, and spontaneous emission occurs naturally, which results in noise.

Figure 8 shows the energy level and transition diagram of EDFA. A 0.98 μm pump photon is absorbed and excites a carrier (Er^{3+} ions) into higher excited states, and the excited carrier decays rapidly to the first excited state that has a very long lifetime of ~10 ms (metastable state). In contrast, a 1.48 μm pump photon is absorbed and excites a carrier into a metastable state. This long metastable state is one of the key advantages of EDFA, and intersymbol distortion and interchannel crosstalk are negligible as a result of these slow dynamics. Depending on the external optical excitation signal, this carrier will decay in a stimulated or spontaneous fashion to the ground state and emit a photon. Both absorption and emission spectra have an associated bandwidth depending on the spread in wavelengths that can be absorbed or emitted from a given energy level. This is highly desirable because (i) a pump laser does not need to be an exact wavelength; and (ii) the signal may be at one of several wavelengths, especially in a WDM system. Figure 9 shows the bandwidth in the 1.48 μm absorption and 1.55 μm fluorescence spectrum of a typical erbium-doped amplifier.

When the noise characteristics of EDFA are considered, noise figure (NF) is an important parameter, which is defined by $NF \equiv SNR_{in}/SNR_{out}$, where SNR_{in} (SNR_{out}) is the electrical equivalent signal-to-noise ratio (SNR) of the optical wave going into (coming out) of the amplifier, if it were be detected. The typical value of NF in commercially-available EDFA is 4.5–6 dB. The importance of NF in optical communication systems is presented later in this chapter, when the required optical SNR (OSNR) for error-free optical data transmission is considered. Figure 10 shows an example of EDFA diagram.

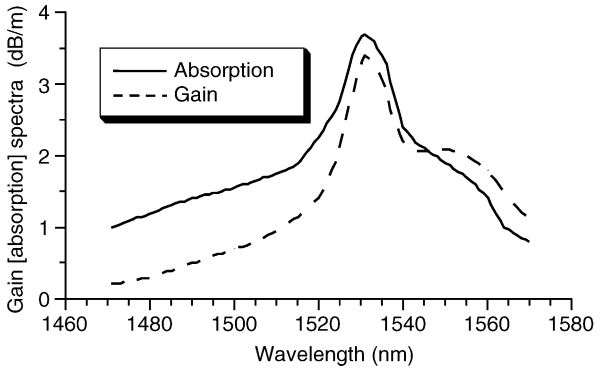


Figure 9 The absorption and fluorescence spectra for erbium near 1.5 μm. Reproduced with permission from Miniscalco WJ (1991) Erbium-doped glasses for fiber amplifiers at 1550 nm. *IEEE Journal of Lightwave Technology* 9: 234–250. Copyright © 1991 IEEE.

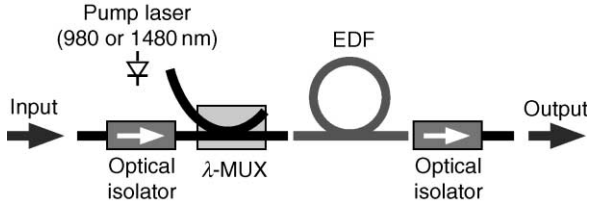


Figure 10 Typical EDFA diagram.

Recent development of EDFA in longer wavelength regimes or L-band (1570–1620 nm) has doubled the useful transmission bandwidth over the conventional band or C-band (1525–1265 nm). L-band gain of EDFA is generally achieved by proper arrangement of pump lasers through longer lengths of erbium-doped fiber than conventional C-band EDFA. Figure 11a shows the concept of wideband EDFA. The incoming

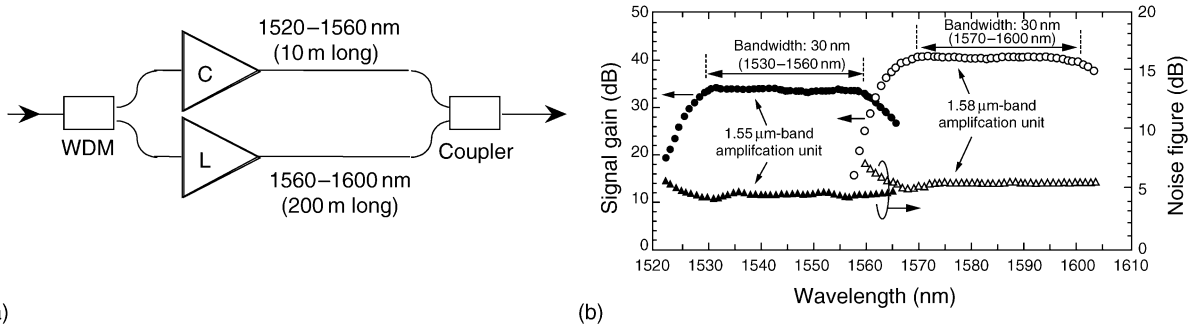


Figure 11 Wide band EDFA. Reproduced with permission from Yanada M (1997) *Electronics Letters*. Copyright © 1997 IEEE.

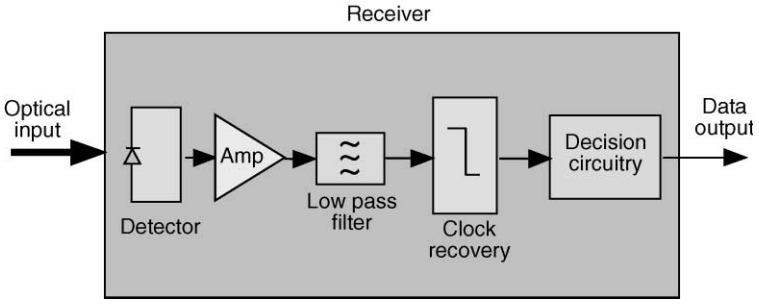


Figure 12 Block diagram of functions performed in advanced receiver packages.

WDM signal is split into two bands (C- and L-bands), amplified separately, and combined at the output. Gain and NF characteristics of this wideband EDFA is shown at Figure 11b.

Because SRS transfers the energy of a lower wavelength channel into the higher wavelength channel in the optical fiber, the transmission fiber itself can be used as a gain medium on specific channel (or channels) when the proper pump signal (or signals) are provided in the fiber. Even though Raman gain is much lower than the gain of EDFA (i.e., <0.1 dB gain/mW pump in Raman amplifier (RA) compared to a few dB gain/mW pump in EDFA), RA has several advantages over EDFA (i) RA has the capability to provide gain at any signal wavelengths, (ii) the amplification window is expandable by combining multipump wavelengths; and (iii) RA offers improved noise performance because the signal is amplified over transmission fiber (i.e., distributed amplifier). Especially improved noise performance increases system OSNR, which can be used to extend system reach or create more wavelength channels on the system. The Raman pumps are typically located ~100 nm shorter than the wavelength of the signal to be amplified, and a few 100s mW pump lasers are used to get a 10–15 dB gain.

In general, telecom receiver (RCVR) consists of photo diode (PD) with transimpedence amplifier (TIA), followed by limiting amplifier (LA) or automatic gain control amplifier (AGC), low pass filter, and clock data recovery circuit (CDR) (see Figure 12).

Vertically illuminated positive-intrinsic-negative (PIN) photodiode and avalanche photo diode (APD) are the most common PD that is used in RCVR. In PIN, incoming light is absorbed in the n^- (low doped or intrinsic) region and generates an electron-hole pair, which drifts to a depletion region and collected n^+ and p^+ regions, therefore generating electric current. APD is used for high-sensitivity detection up to 10 Gb/s. It is structured as absorption, grading, and multiplication layers. The multiplication layer of APD provides additional transit time delay compared to PIN, therefore the bandwidth of APD is smaller than PIN in general. TIA is a broadband low noise amplifier that is used to convert and amplify the weak current from PD to the desired output voltage. A low pass filter is used to suppress the noise and data distortion from high frequency components with a typical cut-off frequency ~0.7 bitrate. LA and AGC are amplifiers that limit the output voltage as a constant. LA is a nonlinear amplifier that makes a decision either to make logic '1' or '0' based on input voltage level, and limits output voltage as fixed values. AGC is a linear amplifier that limits the output voltage by adjusting the gain of the amplifier based on input voltage level. CDR consists of a phase lock loop (PLL) with a voltage controlled oscillator (VCO)

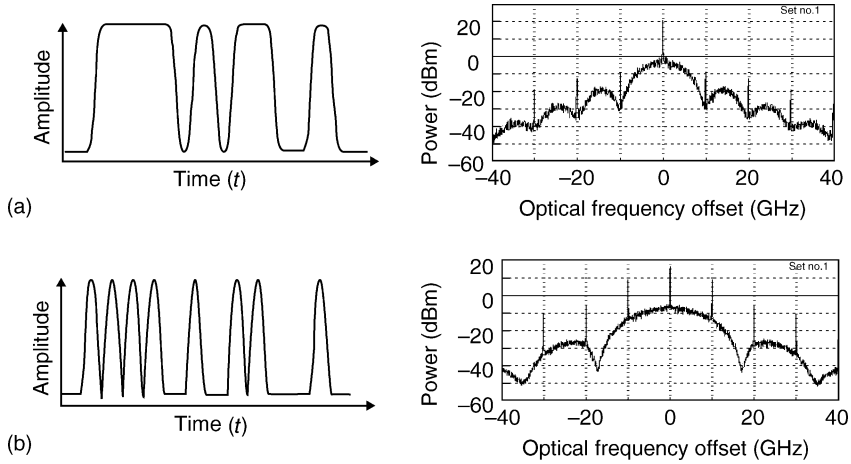


Figure 13 Amplitude modulation (a) nonreturn-to-zero (NRZ), and (b) return-to-zero (RZ).

and data flip-flop circuit. The PLL is used to synchronize the VCO to the incoming high-speed bit stream, therefore recovering the system clock. The recovered clock is used in D-FF to retime and regenerate the data.

Data Modulation Formats

In digital optical communications, numerous modulation formats have been developed and utilized, in order to achieve higher spectral efficiency (i.e., ratio of individual channel bit rate to dense WDM (DWDM) separation), and to improve system performance over fiber impairments. A simple amplitude modulation of lightwave is still the most widely used method for optical communication. The easiest amplitude modulation technique is to using a nonreturn-to-zero (NRZ) format. Figure 13a shows an example of NRZ binary sequence in time domain, and the corresponding frequency domain spectrum. In NRZ, data '1' and '0' correspond to the ON and OFF state of a lightwave transmitter, respectively, occupying entire bit time (i.e., binary symbol time). When the ON or OFF state of a lightwave transmitter does not occupy the entire bit time (T_B) (i.e., when the data pulse duration (T_d) is less then (T_B), it is called the return-to-zero (RZ). Figure 13b shows a binary sequence of RZ, and its frequency spectrum when $T_d \approx 0.5\, T_B$. The Fourier transform of the square wave is the Sinc function. Therefore, the frequency spectrum of an ideal NRZ and RZ follows the pattern of a Sinc square function. Note that the first null of frequency spectrum of RZ is about twice as high as the null of NRZ.

Even though NRZ is better in spectral efficiency than RZ, RZ has an advantage over NRZ in tolerance of fiber impairments. Figure 14 shows an example on eye closure penalty comparison between NRZ and RZ data formats at a 16-channel WDM system at 40 Gb/s data rate. It is clearly seen that the penalty of RZ is much less severe than the penalty of NRZ, as the transmission distance increases. (Note, nonlinearities and dispersion increases as the transmission distance increases.)

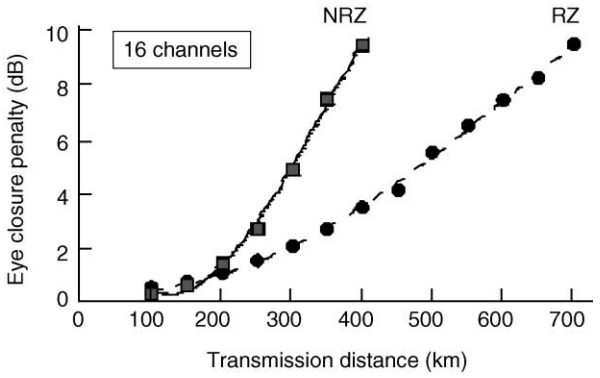


Figure 14 System performance of different modulation format (NRZ vs. RZ).

There has been much effort to increase the spectral efficiency, to improve the system performance over fiber impairments, or to achieve both simultaneously. Figure 15 shows some of examples of these. Chirped RZ (CRZ) is the modulation format inducing frequency chirp on conventional RZ. CRZ is robust on fiber impairment at the cost of losing spectral efficiency. In carrier-suppressed RZ (CS-RZ), a strong un-modulated optical carrier is removed from conventional RZ, therefore suppressing the nonlinear effects. Duo-binary is a three-level signal modulation format generated by a series of delay and added

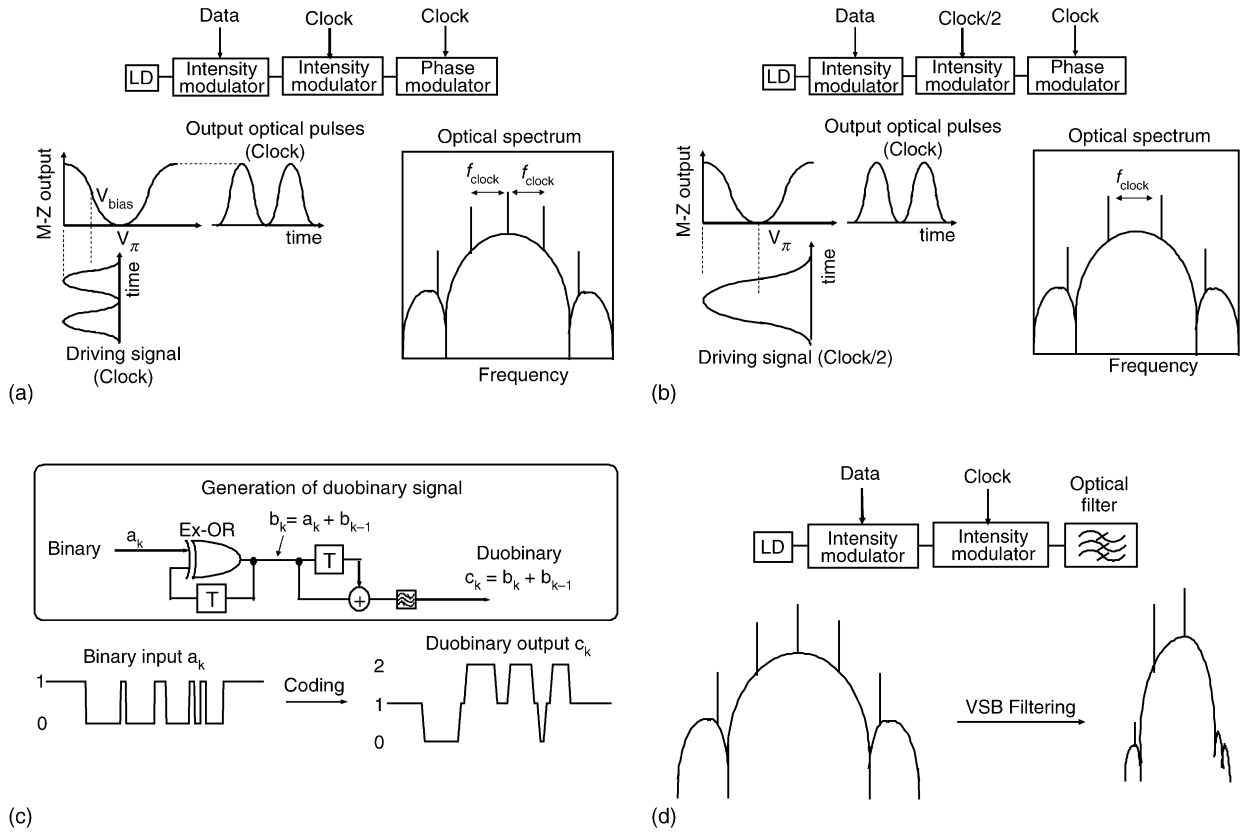


Figure 15 Examples of bandwidth efficient modulation formats. (a) Chirped RZ (CRZ) generation, (b) carrier suppressed-RZ (CS-RZ) generation, (c) duo-binary generation, and (d) vestigial single side band (VSB) RZ generation.

circuits in the transmitter. The advantage of duo-binary is that its bandwidth is reduced to one half of conventional NRZ, therefore having enhanced spectral efficiency, and has a high tolerance in chromatic dispersion, and suppression of fiber nonlinearites.

Single-sideband (SSB) or vestigial singleside band (VSB) can be generated into either NRZ or RZ, by either optically utilizing optical filtering with a sharp cut-off filter, or electrically applying tapped-delay-line filter approximation on both amplitude and phase modulators. SSB (or VSB) has about a two times higher spectral efficiency and higher tolerance on chromatic dispersion than conventional modulation formats.

Data Multiplexing

At their most basic, optical networks can imitate electrical networks in which time division multiplexing (TDM) is overwhelmingly used for digital data transmission. A fiber can carry many time-multiplexed channels, in which each channel can transmit its data in an assigned time slot. A typical TDM link is shown in Figure 16, in which N

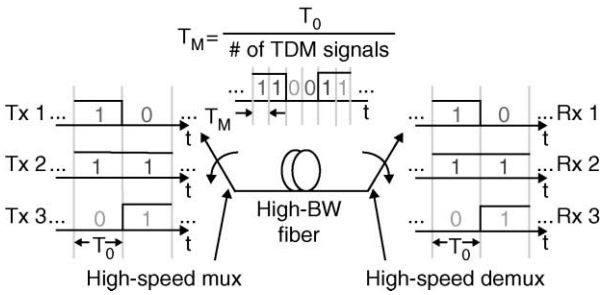


Figure 16 Concept of bit-interleaving TDM.

transmitters are sequentially polled by a fast multiplexer to transmit their data. The time-multiplexed data are sequentially and rapidly demultiplexed at the receiving node.

Time multiplexing and demultiplexing functions can be performed either electrically with an electrical time (DE)MUX switch and an optical transmitter/receiver, or optically with multiple optical transmitters/receivers and an optical time (DE)MUX switch. The major advantage of TDM is that there is no output–port contention problem (each data bit occupies its own time slot and there is only a single high-speed signal present at any given

instant). The disadvantage of TDM is that the scheme requires a ultra-high-speed switching component if the individual signals are themselves high-speed and if there are many users.

In WDM, multiple wavelength channels are transmitted through a single fiber, therefore enabling the fiber to carry more throughput. By using wavelength-selective devices for the ON and OFF ramps, independent signal routing also can be accomplished. Figure 17a shows a simple point-to-point WDM system in which several channels are multiplexed at one node, the combined signals transmitted across a distance of fiber, and the channels demultiplexed at a destination node. As shown in Figure 17b, the wavelength becomes the signature address for either the transmitter or the receivers, and the wavelength will determine the routing path through an optical network.

Many interesting challenges face the eventual implementation of WDM systems and networks. Several of these challenges involve the control and management of the data through this novel high-speed network. Figure 18 shows a small subset of critical component technologies typically required for a WDM system, including multiple-wavelength transmitters, multiport star couplers, passive and active wavelength routers, EDFAs, and tunable optical filters.

Some generic goals that a WDM-device technologist aims to achieve include; large wavelength tuning range; multi-user capability; wavelength selectivity and repeatability; low cross-talk; high extinction ratio; minimum excess losses; fast wavelength tunability; high-speed modulation bandwidth; low residual chirp; high finesse; low noise; robustness; high yield; potential low cost. Depending on system requirements, device availability and cost, WDM technologies divide into two; course-WDM (CWDM) and dense-WDM DWDM. Figure 19 shows the wavelength allocation for CWDM and

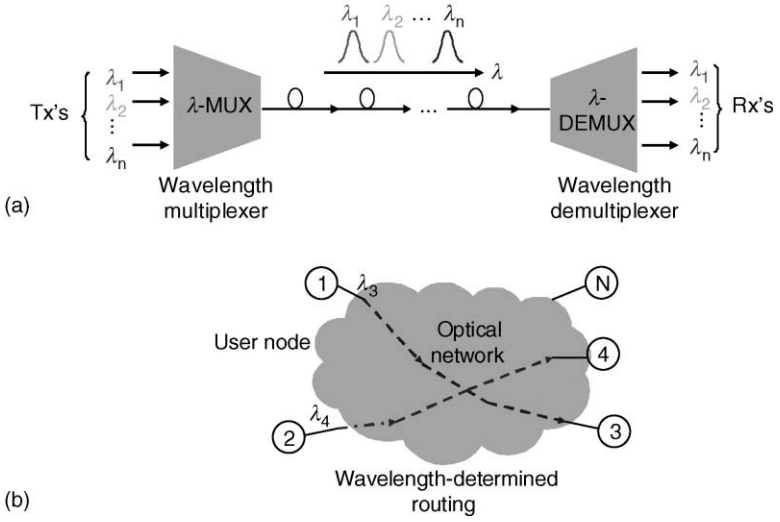


Figure 17 (a) A simple point-to-point WDM transmission system; (b) a generic multiuser network in which the communication links and routing paths are determined by the wavelengths used within the optical switching fabric.

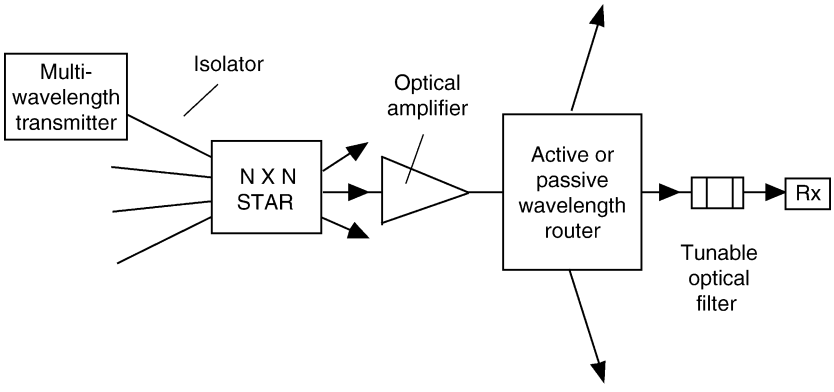


Figure 18 Schematic of a small subset of enabling device technologies for a WDM system.

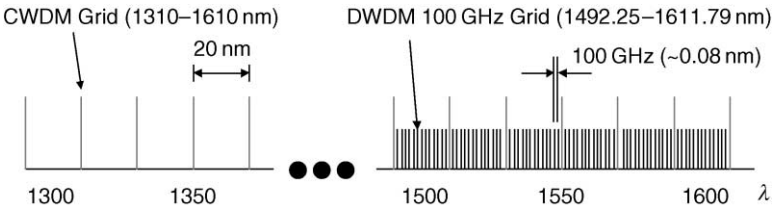


Figure 19 WDM wavelength allocation.

DWDM. CWDM technology uses an International Telecommunication Union (ITU) standard 20 nm spacing between the wavelengths from 1310 nm to 1610 nm. Also DWDM technology uses an ITU standard 100 GHz or 200 GHz between wavelengths, arranged in several bands at around 1500–1600 nm.

Dispersion Management and Compensation

As mentioned previously, chromatic dispersion is one of the most basic characteristics of fiber, although it is possible to manufacture fiber that induces zero chromatic dispersion. But such fiber is incompatible with the deployment of a WDM system since harmful nonlinear effects are generated, therefore chromatic dispersion must exist, and must be compensated for. The effect of chromatic dispersion is cumulative and increases quadratically with the data rate (see Figure 20).

The quadratic dependence of dispersion with the data rate is a result of two effects. First, a doubling of the data rate will double the Fourier-transformed frequency spectrum of the signal, thereby doubling the effect of dispersion. Second, the same doubling of the data rate makes the data pulse only half as long in time and therefore twice as sensitive to temporal spreading due to dispersion. The conventional wisdom for the maximum distance over which data can be transmitted is to consider a broadening of the pulse equal to the bit time period. For a bit period B, a dispersion value D and a spectral width $\Delta\lambda$, the dispersion-limited distance is given by $L_D = 1/(D{\cdot}B{\cdot}\Delta\lambda)$ (see Figure 21).

In theory, compensation of chromatic dispersion for high-speed or long-distance systems can be fixed in value if each link's dispersion value is known. A simple yet elegant solution is to create a dispersion 'map', in which the designer of a transmission link alternates elements that produce positive and then negative dispersion (see Figure 22). This is a very powerful concept: at each point along the fiber the dispersion has some nonzero value, effectively

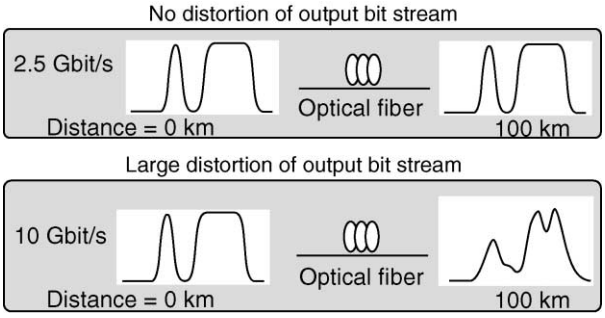


Figure 20 Pulse broadening at two different data rates (2.5 Gbit/s, and 10 Gbit/s) as a result of quadratic nature of chromatic dispersion.

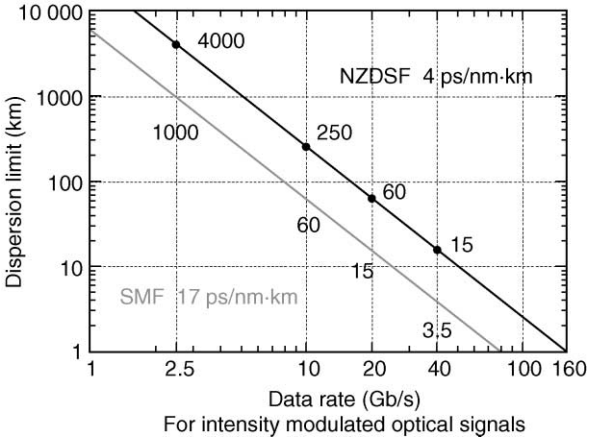


Figure 21 Dispersion-limited (uncompensated) transmission distance in single-mode fiber (SMF) as a function of data rate for intensity-modulated optical signals. Courtesy of L.D. Garrett.

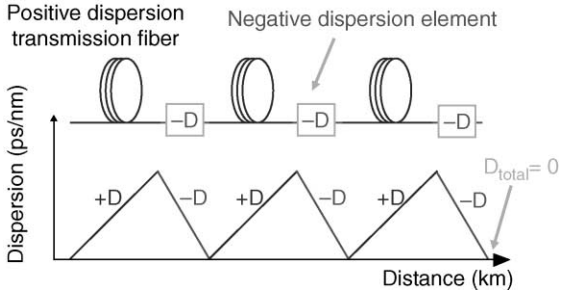


Figure 22 Typical static management of chromatic dispersion.

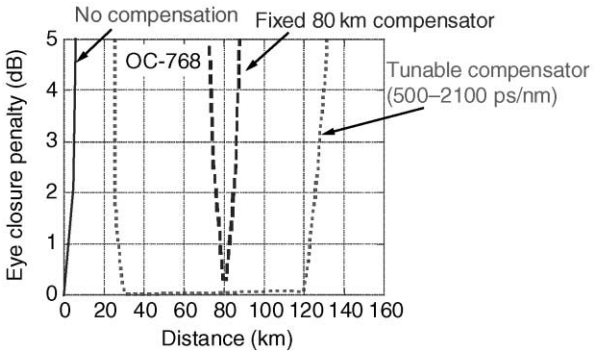


Figure 23 Effect of tunable compensation at 40 Gb/s (OC-768); it allows a wide range of transmission distance. On the contrary, the 80 km fixed compensator allows only small range of transmission around 80 km.

eliminating FWM and XPM, but the total accumulated dispersion at the end of the fiber link is zero, so that minimal pulse broadening is induced. The specific system design, as to the periodicity of management, depends on several variables, but a typical number for SMF as the embedded base is a compensation at every 80 km in a 10 Gb/s system. Dispersion-compensating fiber (DCF) has been generally used as a dispersion compensating element.

However, there are several important aspects of optical systems and networks that make tunable dispersion compensation solutions attractive, including: (i) it significantly reduces the inventory of different required types of compensation modules; (ii) it tunes to adapt to routing path changes in a reconfigurable network, (iii) it tracks dynamic changes in dispersion due to environment, and (iv) it achieves a high degree of accuracy necessary for 40 Gb/s channels (see Figure 23).

One brute-force method to achieve tunable dispersion compensation is to build a module with optical switches used to add or remove sections of fixed DCF to achieve a discrete set of dispersion compensation. Many other elegant, yet viable, approaches have been developed for tunable dispersion compensation, and these approaches include linearly chirped FBG with nonuniform heating, nonlinearly chirped Fiber Bragg grating (FBG) with a simple mechanical stretcher, virtually imaged phase array, electronic tap delay filter with weights, multiple stage all pass filter, etc.

System Performance Parameters

An eye diagram provides a simple and useful tool to visualize intersymbol interference between data bits. Figure 24a shows a perfect eye diagram. A square bit stream (i.e., series of symbol '1's and '0's) is sliced into sub-bit stream with predetermined eye intervals (i.e., several bit periods), and displayed through bit analyzing equipment (e.g., digital channel analyzer), overlapping the sliced sub-bit stream in order to obtain the eye diagram. When a perfect transmitter and receiver (i.e., infinite receiver bandwidth with zero noise characteristics and jitter), and a perfect transmission media (i.e., no dispersion and nonlinearites) are used, the received eye diagram is shaped as a perfect rectangular. In reality, the transmitter and receiver have a limited bandwidth with noise and jitter, and the transmission media (i.e., optical fiber) has dispersion and nonlinearites. Therefore, the eye diagram deviates from the perfect rectangular shape. Figure 24b shows the eye diagram close to a real situation. The shape of the eye is generally broadened and distorted (i.e., eye is closed) due to limited bandwidth and fiber impairments, and noise and timing jitter are added onto this broadened and distorted eye shape.

The Q-factor (q) is also an important system parameter widely used in long-distance optical transmission system design. It is defined as the electrical signal-to-noise ratio before the decision circuit at receiver. The Q-factor is directly related to bit-error-rate (BER: the percentage of bits that has an error relative to total number of bits received in a specific time) by: $\mathrm{BER} = 0.5 \cdot \mathrm{erfc}(q/2^{0.5})$, where $\mathrm{erfc}(x)$ is the complementary error function. The Q-factor is an unitless linear ratio, and is express in dB by $20 \cdot \log(q)$. As a conventional relationship, Q-factor of 15.6 dB (i.e., linear ratio 6) is required to achieve $\mathrm{BER} = 10^{-9}$. Figure 25 shows the relationship of eye diagram, Q-factor, and BER.

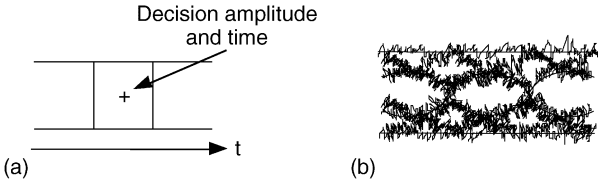


Figure 24 (a) A perfect square eye diagram, (b) a closed eye diagram due to bandwidth, fiber impairments, noise, and timing jitter.

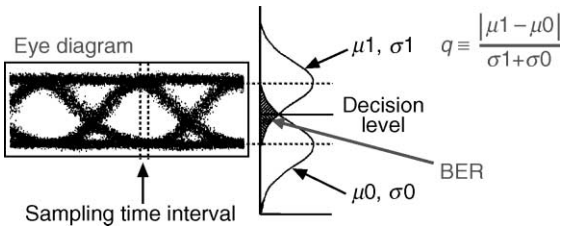


Figure 25 The relationship of eye diagram, Q-factor, and BER ($\mu 1$, $\mu 2$: mean levels of logic '1' and '0', $\sigma 1$, $\sigma 2$: standard deviation of logic '1' and '0').

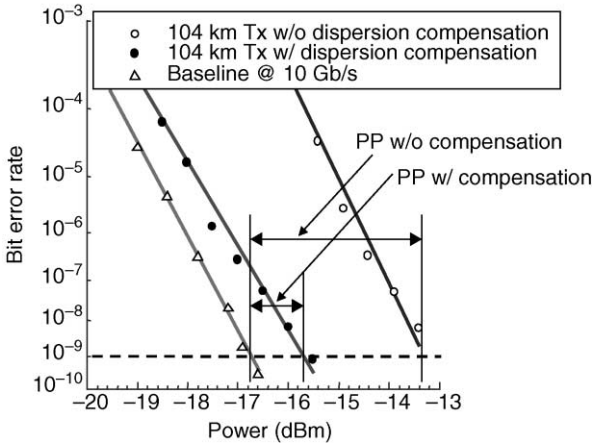


Figure 26 BER curve and power penalty.

Power penalty (PP) is one of the most important system parameters, and is defined as the received optical power difference in dB with and without signal impairments at a specified BER (conventionally 10^{-9}), from measured BER versus optical power curve. Therefore, 1 dB PP means that a system with signal impairments requires 1 dB more optical power at the receiver in order to achieve the same BER performance compared to the system without signal impairments. Figure 26 shows the typical BER curve as a function of received optical power and PP. Note that PP reduces from ~3.5 dB to ~1 dB with dispersion compensation.

In long-haul and undersea transmission system with many EDFA chains, OSNR is an important system parameter to design and characterize the optical transport system. OSNR is defined as the ratio of the power of signal channel to the power of ASE in a specified optical bandwidth (conventionally 0.1 nm). The OSNR (in dB) of a signal channel at the end of the system is approximated by: $\text{OSNR (dB)} = 58 + P_{\text{out}} - L_{\text{span}} - \text{NF} - 10\log(N_{\text{amp}})$, when the system has N_{amp} fiber spans, span loss L_{span} (in dB) followed by an optical amplifier with output power P_{out} (in dBm) per channel launched into the span and noise figure (NF) (in dB). As an example, a typical OSNR requirement for $\text{BER} = 10^{-9}$ is about 17 dB at 10 Gb/s data rate. The required OSNR should be increased by 6 dB when the data rate is increased by a factor of 4. Improving the amplifier's NF can increase OSNR, and the improved OSNR can be used to increase the system reach, reducing the channel power in order to suppress the nonlinearities, etc.

See also

Fiber and Guided Wave Optics: Light Propagation. **Imaging:** Multiplex Imaging. **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Leonid K, Sergio B and Alan W (1996) *Optical Fiber Communication Systems*. Norwood, MA: Artech House Inc.

Govind PA (1995) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Emmanuel D (1994) *Erbium-Doped Fiber Amplifiers*. New York: Wiley.

Ivan K and Tingye L (2002) *Optical Fiber Telecommunications IV A Components*. San Diego, CA: Academic Press.

Ivan K and Tingye L (2002) *Optical Fiber Telecommunications IV B Systems and Impairments*. San Diego, CA: Academic Press.

Historical Development

G Keiser, PhotonicsComm Solutions, Inc., Newton Center, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article describes the development of optical fiber communication systems. After describing some of the motivations for using optical fiber communications and the advantages of this technology, the key milestones and the principal people involved in developing optical fibers and compatible light sources are presented. Following this, the article looks at the evolution of fielded systems and the use of optical fiber links in undersea applications.

One of the principal needs of people since antiquity has been to communicate. This need created interests in devising communication systems for sending messages from one distant place to another. Many forms of such systems have appeared over the years. The basic motivations behind each new one were either to improve the transmission fidelity, to increase the data rate so that more information could

be sent, or to increase the transmission distance between relay stations. Prior to the nineteenth century, all communication systems operated at a very low information rate and involved only optical or acoustical means, such as signal lamps or horns. One of the earliest known optical transmission links, for example, was the use of a fire signal by the Greeks in the eighth century BC for sending alarms, calls for help, or announcements of special events.

The invention of the telegraph by Samuel FB Morse in 1838 ushered in the era of electrical communications. In the ensuing years an increasingly larger portion of the electromagnetic spectrum, shown in Figure 1, was utilized for the conveying of information from one place to another. The reason for this trend is that, in electrical systems, the data usually are transferred over the communication channel by superimposing the information onto a sinusoidally varying electromagnetic wave, which is known as the carrier. When reaching its destination the information is removed from the carrier wave and processed as desired. Since the amount of information that can be transmitted is related directly to the frequency range over which the carrier operates, increasing the carrier frequency theoretically increases the available transmission bandwidth and, consequently, provides a larger information capacity. Thus the trend in electrical communication system developments was to employ progressively higher frequencies (shorter wavelengths), which offer corresponding increases in bandwidth or information capacity. This activity led to the birth of communication mechanisms such as radio, television, microwave, and satellite links.

Another important portion of the electromagnetic spectrum encompasses the optical region shown in Figure 1. In contrast to electrical communications, transmission of information in an optical format is carried out, not by frequency modulation of the carrier but by varying the intensity of the optical power. Similar to the radio-frequency spectrum, two classes of transmission medium can be used: an atmospheric channel or a guided-wave channel. Whereas transmission of optical signals through the atmosphere was done thousands of years ago, the use of a guided-wave optical channel such as an optical fiber is a fairly recent application.

Fiber optic communication systems have a number of inherent advantages over their copper-based and radio-transmission counterparts. Fiber optic cable can transmit at a higher capacity, thereby reducing the number of physical lines and the amount of equipment needed for a given transmission span. The lower weight and smaller size of optical fibers offer a distinct advantage over heavy, bulky copper cables in crowded underground cable ducts, in ceiling-mounted cable trays, and in mobile platforms such as aircraft and ships. Their dielectric composition is an especially important feature of optical fibers, since this ensures freedom from electromagnetic interference between adjacent fibers, eliminates ground loops, and results in extremely low fiber-to-fiber crosstalk. In addition, an optical fiber affords a high degree of data security since the optical signal is well confined within the optical waveguide.

Optical Fiber Development

Since an atmospheric channel requires a line-of-sight link and because it can be adversely affected by weather conditions, a guided-wave channel is the preferred approach in most cases for communication system applications. A challenge in using an optical fiber channel is to have a flexible, low-loss medium that transfers the optical signal over long distances

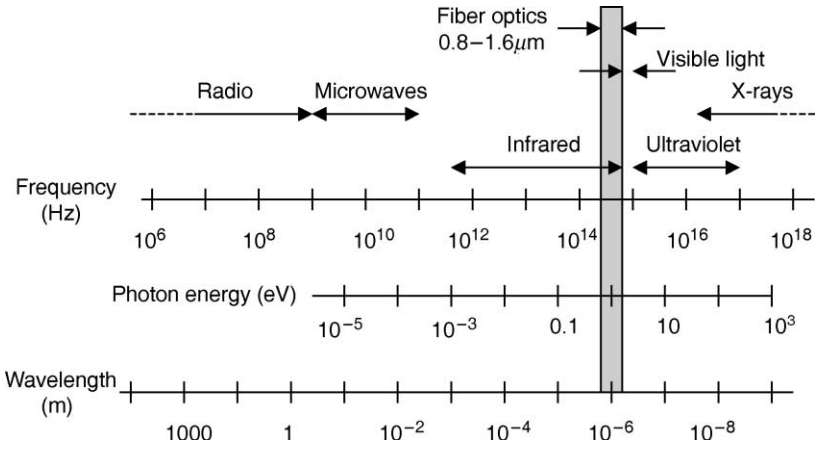


Figure 1 Electromagnetic spectrum for electrical and optical communications.

without significant attenuation and distortion. Glass is an obvious material for such applications. The earliest known glass was made around 2500 BC and glass already was drawn into fibers during the time of the Roman Empire. However, such glasses have very high attenuations and thus are not suitable for communication applications. One of the first known attempts of using optical fibers for communication purposes was a demonstration in 1930 by Heinrich Lamm of image transmission through a short bundle of optical fibers for potential medical imaging. However, no further work was done beyond the demonstration phase since the technology for producing reasonably low-loss fibers with good light confinement was not yet mature.

Further work and experiments continued on using optical fibers for image transmission and by 1960 glass-clad fibers had attenuations of about one decibel per meter. This attenuation allowed fibers to be used for medical imaging, but it was still much too high for communications. Optical fibers had attracted the attention of researchers at that time, because they were analogous in theory to plastic dielectric waveguides used in certain microwave applications. In 1961, Elias Snitzer demonstrated this similarity by drawing fibers with cores so small they carried light in only one waveguide mode. He published a classic theoretical description of single-mode fibers in May 1961. However, to be useful for communication systems, optical fibers would need to have a loss of no more than 10 or 20 decibels per kilometer.

As a reminder, decibels measure the ratio of the output power to the input power on a logarithmic scale. The abbreviation for a decibel is 'dB.' The power ratio in decibels is given by the expression '10 log (power out/power in).' As an example, a power loss of 20 dB over a 1-km distance in an optical fiber means that only 1% of the light injected into the fiber comes out of the other end. Table 1 gives some typical examples of various power ratios and their decibel equivalents.

In the early 1960s Charles Kao pursued the idea of using a clad glass fiber for an optical waveguide, building on optical waveguide research being done at the Standard Telecommunication Laboratories in England. After he and George Hockman painstakingly examined the transparency properties of various types of glass, Kao made a prediction, in 1966, that losses of no more than 20 dB/km were possible in optical fibers. In July 1966, Kao and Hockman presented a detailed analysis for achieving such a loss level. Kao then went on to actively advocate and promote the prospects of fiber communications, which generated interest in laboratories around the world to reduce fiber loss. It took four years to reach Kao's predicted goal of 20 dB/km, and the final solution was different from what many had expected.

To understand the process of making a fiber, consider the schematic of a typical fiber structure shown in Figure 2. A fiber consists of a solid dielectric (glass or plastic) cylinder of refractive index n_1 called the core. This is surrounded by a dielectric cladding which has a refractive index n_2, that is less than n_1, in order to achieve light guiding in the fiber. This structure is then encapsulated by a buffer coating to protect the fiber from abrasion and outside contaminants. The first step in making a fiber is to form the clear glass rod or tube called a preform. This normally is done by a vapor-phase oxidation process. The preform has two distinct regions that correspond to the core and cladding of the eventual fiber. Fibers are made from the preform by precision feeding it into a circular furnace that softens the end of the preform to the point where it can be drawn into a very thin filament which becomes the optical fiber.

Most researchers had tried to purify the compound glasses used for standard optics, which are easy to melt and draw into fibers. A different approach was taken at the Corning Glass Works where Robert Maurer, Donald Keck, and Peter Schultz started with fused silica. This material can be made extremely pure, but has a high melting point and a low refractive index. Silica has the approximate attenuation versus wavelength characteristic shown in Figure 3. Note that for early silica fibers there are regions of low attenuation around 850, 1310, and 1550 nm, which the literature refers to as the first, second, and third windows, respectively. The large attenuation spikes in the 1000- and 1400-nm spectral

Table 1 Examples of some optical power ratios and their decibel equivalents

Power ratio (P_{out}/P_{in})	*Decibel equivalent (dB)*
0.01	−20
0.10	−10
0.50	−3
1	0
2	3
10	10
100	20

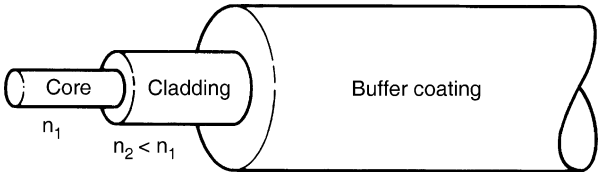


Figure 2 Typical optical fiber structure.

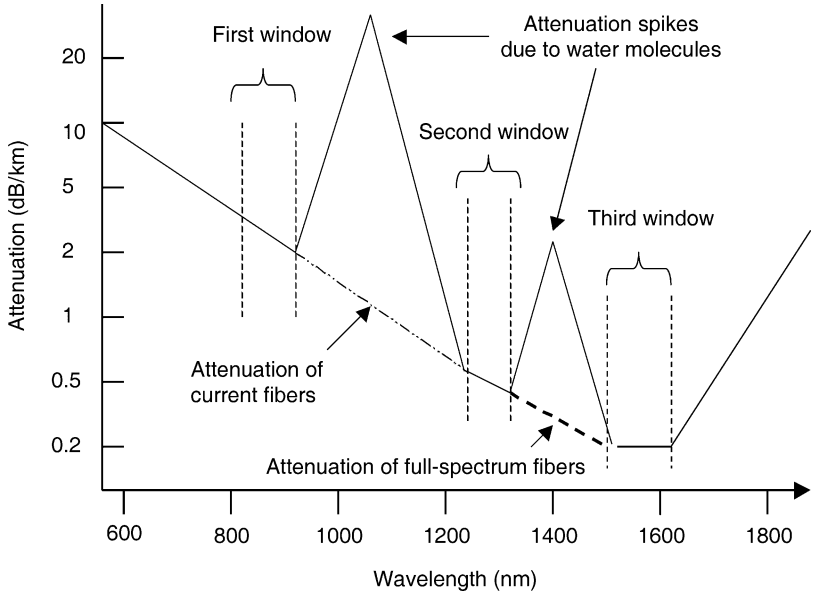


Figure 3 Attenuation versus wavelength approximation for silica fibers.

regions are due to absorption by residual water molecules in the glass. Although ultra-pure material processing techniques can eliminate these spikes to produce what are known as full-spectrum or low-water-peak fibers, most installed fibers still have a relatively large attenuation between 1350 and 1500 nm.

Note that since the attenuation spikes no longer are present in various types of fibers, the idea of operating windows has been replaced by the concept of spectral bands. The International Telecommunications Union (ITU) has designated six spectral bands for use in intermediate-range and long-distance optical fiber communications within the 1260 to 1675-nm region. The regions, which are known by the letters O, E, S, C, L, and U, are defined as follows:

- Original Band (O-Band): 1260 to 1360 nm
- Extended Band (E-Band): 1360 to 1460 nm
- Short Band (S-Band): 1460 to 1530 nm
- Conventional Band (C-Band): 1530 to 1565 nm
- Long Band (L-Band): 1565 to 1625 nm
- Ultra-long Band (U-Band): 1625 to 1675 nm

The Corning team made cylindrical preforms by depositing purified materials from the vapor phase. To produce a fiber that has light guiding properties they carefully added controlled trace levels of titanium to the core to make its refractive index slightly higher than that of the cladding without raising the attenuation significantly. In September 1970, they announced the fabrication of single-mode fibers with an attenuation of 17 dB/km at the 633-nm helium-neon line. The fibers were fragile, but independent tests at the British Post Office Research Laboratories facility in Martlesham Heath, England confirmed the low loss.

This dramatic breakthrough was the first among the many developments that opened the door to fiber optic communications. The ensuing years saw further reductions in optical fiber attenuations. By mid-1972, Maurer, Keck, and Schultz at Corning had made multimode germania-doped fibers with a 4-dB/km loss and much greater strength than the earlier brittle titania-doped fibers. In order to couple a sufficient amount of optical power into a fiber, early applications used multimode fibers with a refractive-index gradient between core and cladding of around 2% and core diameters of 50 or 62.5 micrometers.

Single-mode fibers have much smaller core diameters on the order of 9 micrometers in order to allow only one propagation mode. This type of fiber has a much higher transmission capacity since the effect of modal dispersion is eliminated. The first single-mode fibers were optimized to have a zero dispersion value at 1310 nm, since the silica material used at that time exhibited a low loss within a spectral band around this wavelength. Figure 4 shows the dispersion characteristic of this type of fiber as a function of wavelength in the S-, C-, and L-bands. As a result of its widespread use in early single-mode transmission systems, this fiber design has been standardized by the International Telecommunication Union (ITU-T) under the designation Recommendation G.652.

Standard G.652 silica fibers provide the lowest attenuation at 1550 nm, but have a much larger signal

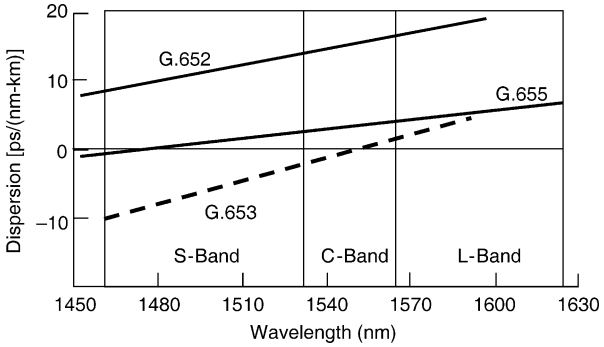


Figure 4 Dispersion characteristics of three types of standard optical fibers.

dispersion at this wavelength than at 1310 nm. Since system designers wanted to use fibers at a 1550-nm wavelength, in order to transmit high-speed data over longer distances, fiber manufacturers overcame the larger signal dispersion limitation by creating the so-called dispersion-shifted fibers. This was done through a clever manipulation of the core and cladding designs that allowed the zero-dispersion point to shift to longer wavelengths. In particular, the ITU-T created a specification for operation at 1550 nm which is designated as recommendation G.653. Figure 4 also shows the dispersion characteristic of this type of fiber as a function of wavelength.

Although the G.652 and G.653 fibers work well for single-wavelength operation, a different type of fiber, having non-zero dispersion within a broad spectral range, is needed when implementing systems that use many independent light sources simultaneously within a particular wavelength band. This led to the specification of the non-zero dispersion shifted fiber (NZDSF) in the ITU-T Recommendation G.655 that are designated to operate around 1550 nm. The main purpose of having a positive dispersion value over the entire operating spectrum is to mitigate a nonlinear optical effect called four-wave mixing (FWM), which is analogous to intermodulation distortion in electrical systems. Figure 4 shows the dispersion characteristic of the G.655 fiber in the S-, C-, and L-bands.

Light Sources

The fiber by itself is not practical unless there is a compatible optical source for launching light signals into it. The most suitable device for this is a semiconductor light-emitting diode (LED) or a laser diode. In the 1960s, a great deal of effort took place to achieve laser action in pn-junction diodes. The early devices were GaAs and GaAsP lasers that operated at a temperature of 77 K, emitted at a wavelength around 850 nm, and had high lasing threshold current densities. To make devices that were more application-friendly by operating at room temperature, structures consisting of sandwiched layers of AlGaAs and GaAs were investigated. Finally, in 1970, researchers at Bell Laboratories and a team at the Ioffe Physical Institute in Leningrad made the first semiconductor diode lasers based on a layered AlGaAs/GaAs/AlGaAs structure that were able to emit continuous-wave light in the 850-nm region at room temperature.

Major improvements in laser diode performance and reliability followed this achievement during the next decade. In addition, around 1976, room-temperature laser diodes operating at longer wavelengths, in the 1100 to 1600-nm region, started to appear. Of particular interest were GaInAsP/InP-based laser diodes emitting in the 1310-nm and 1550-nm low-loss windows of optical fibers. The progressive development of ever-improving devices during the 1980s and 1990s, included single-frequency emission with narrow linewidths under continuous operation, low levels of chirp under direct modulation, high output power, and the ability to tune specially constructed laser diodes over a wavelength range of up to 30 nm.

Fielded Systems

The bit rate-distance product $B \times L$, where B is the transmission bit rate and L is the repeater spacing, measures the transmission capacity of optical fiber links. Since the inception of optical fiber communications in the mid-1970s, the link transmission capacity has experienced a tenfold increase every four years. Several major technology advances spurred this growth. Among the technology developments are laser diodes emitting over an extremely narrow spectral band, optical amplifiers, fibers with low losses and low dispersions, and the concepts of wavelength division multiplexing.

Some of the initial telephone-system field trials in the USA were carried out in 1977 by GTE in Los Angeles and by AT&T in Chicago. These transmission links were largely for the trunking of telephone lines, which are digital links consisting of time-division-multiplexed 64-kb/s voice channels. Similar demonstrations were carried out in Europe and Japan. Applications ranged from 45 to 140 Mb/s with repeater spacings around 10 km.

With the advent of high-capacity fiber optic transmission lines in the 1980s, service providers established a standard signal format called synchronous optical network (SONET) in North America and synchronous digital hierarchy (SDH) in other

parts of the world. These standards define a synchronous frame structure for sending multiplexed digital traffic over optical fiber trunk lines. The basic building block and first level of the SONET signal hierarchy is called the synchronous transport signal–level 1 (STS-1), which has a bit rate of 51.84 Mb/s. Higher-rate SONET signals are obtained by byte-interleaving N STS-1 frames, which are then scrambled and converted to an optical carrier–level N (OC-N) signal. Thus the OC-N signal will have a line rate exactly N times that of an OC-1 signal. For SDH systems, the fundamental building block is the 155.52-Mb/s synchronous transport module–level 1 (STM-1). Again, higher-rate information streams are generated by synchronously multiplexing N different STM-1 signals to form the STM-N signal. Table 2 shows commonly used SDH and SONET signal levels.

The development of optical sources and photodetectors capable of operating at 1310 nm allowed a shift in the transmission wavelength from 850 nm to 1310 nm. This resulted in a substantial increase in the repeaterless transmission distance for long-haul telephone trunks, since optical fibers exhibit lower power loss and less signal dispersion at 1310 nm. Intercity applications first used multimode fibers, but in 1984 switched exclusively to single-mode fibers, which have a significantly larger bandwidth. Bit rates for long-haul links typically range between 155 and 622 Mb/s (OC-3 and OC-12), and in some cases up to 2.5 Gb/s (OC-48), over repeater spacings of 40 km. Both multimode and single-mode 1310-nm fibers are used in local area networks, where bit rates range from 10 Mb/s to 100 Mb/s over distances ranging from 500 m to tens of kilometers.

In the next step of system evolution, links operating in the low-loss window around 1550 nm attracted much attention for high-capacity, long-span terrestrial and undersea transmission links. These links routinely carry traffic at 2.5 Gb/s over 90-km repeaterless distances. By 1996, advances in high-quality lasers and receivers allowed single-wavelength transmission rates of 10 Gb/s (OC-192).

In 1989, the introduction of optical amplifiers gave a major boost to fiber transmission capacity. Although there are GaAlAs-based solid-state optical amplifiers for the first window and InGaAsP amplifiers for the second window, the most successful and widely used devices are erbium-doped fiber amplifiers (commonly called EDFAs) operating in the 1530 to 1560-nm range and Raman fiber amplifiers that are used for operation in the 1560 to 1600-nm region.

During the same time period, impressive demonstrations of long-distance high-capacity systems were made using optical soliton signals. A soliton is a nondispersive pulse that makes use of nonlinear dispersion properties in a fiber to cancel out chromatic dispersion effects. As an example, solitons at rates of 10 Gb/s have been sent over a 12 200-km experimental link using optical amplifiers and special modulation techniques.

Entrance of Wavelength Division Multiplexing

The use of wavelength division multiplexing (WDM) offers a further boost in fiber transmission capacity. The basis of WDM is to use multiple sources operating at slightly different wavelengths to transmit several independent information streams over the same fiber. Although researchers started looking at WDM in the 1970s, during the ensuing years it generally turned out to be easier to implement higher-speed electronic and optical devices than to invoke the greater system complexity called for in WDM. However, a dramatic surge in WDM popularity started in the early 1990s, as electronic devices neared their modulation limit and high-speed equipment became increasingly complex.

Figure 5 shows the concept of implementing many closely spaced wavelengths within a spectral band centered around 1552.524 nm. This scheme is referred to as dense WDM or DWDM. Conceptually, the DWDM scheme is the same as frequency

Table 2 Commonly used SONET and SDH transmission rates

SONET level	*Electrical level*	*Line rate (Mb/s)*	*SDH equivalent*	*Common rate name*
OC-1	STS-1	51.84	–	
OC-3	STS-3	155.52	STM-1	155 Mb/s
OC-12	STS-12	622.08	STM-4	622 Mb/s
OC-24	STS-24	1244.16	STM-8	
OC-48	STS-48	2488.32	STM-16	2.5 Gb/s
OC-96	STS-96	4976.64	STM-32	
OC-192	STS-192	9953.28	STM-64	10 Gb/s
OC-768	STS-768	39,813.12	STM-256	40 Gb/s

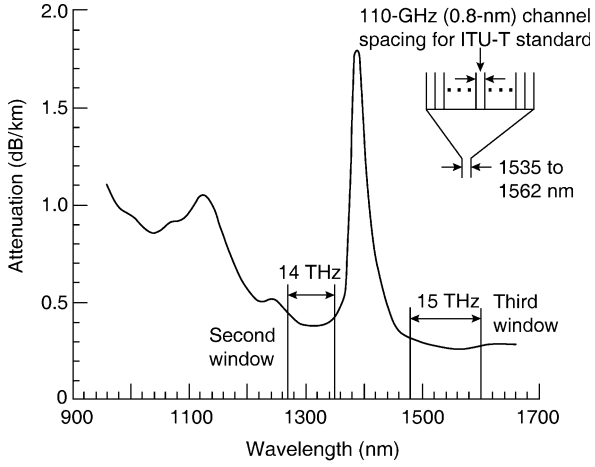


Figure 5 Wavelength division multiplexing (WDM) concept.

division multiplexing (FDM) used in microwave radio and satellite systems. Just as in FDM, the wavelengths (or optical frequencies) in a DWDM link must be properly spaced to avoid interference between channels. In an optical system this interference may arise from the fact that the center wavelength of laser diode sources and the spectral operating characteristics of other optical components in the link may drift with temperature and time, thereby giving rise to the need for a guard band between wavelength channels.

Since WDM is essentially frequency division multiplexing at optical carrier frequencies, the ITU developed DWDM standards that specify channel spacings in terms of frequency. The ITU-T Recommendation G.694.1, which is entitled 'Dense Wavelength Division Multiplexing (DWDM),' specifies WDM operation in the S-, C-, and L-bands for high-quality, high-rate metro area network (MAN) and wide area network (WAN) services. It calls out for narrow frequency spacings of 100 to 12.5 GHz (or, equivalently, 0.8 to 0.1 nm at 1550 nm). This implementation requires the use of stable, high-quality, temperature-controlled and wavelength-controlled (frequency-locked) laser diode light sources.

With the production of full-spectrum (low-water-content) fibers, the development of relatively inexpensive optical sources, and the desire to have low-cost optical links operating in metro and local area networks, came the concept of coarse WDM (CWDM). In 2002, the ITU-T released a standard aimed specifically at CWDM. This is Recommendation G.694.2, which is entitled 'Coarse Wavelength Division Multiplexing (CWDM).' The CWDM grid is made up of 18 wavelengths defined within the range 1270 nm to 1610 nm (O-through L-bands) spaced by 20 nm with wavelength-drift tolerances of ± 2 nm. This can be achieved with inexpensive light sources that are not temperature-controlled. The targeted transmission distance for CWDM is 50 km on single-mode fibers, such as those specified in ITU-T Recommendations G.652, G.653, and G.655.

Wavelength tunability of a source is an important property of WDM systems. Obviously it is not desirable or practical to maintain an inventory of dozens of lasers that emit at different wavelengths for WDM applications. The ideal tunable laser should be adjustable to emit at a specific wavelength across a broad spectral range. One such device is a distributed Bragg reflector (DBR) laser diode that can be tuned over a 10 to 20 nm spectral range. Work on perfecting such devices are still underway.

Starting in the mid-1990s, a combination of EDFAs and WDM was used to boost fiber information capacity to even higher levels and to increase the transmission distance. A major system consideration in these super-high capacity links is to ensure that there is appropriate link and equipment redundancy, so that alternate paths are available in case of disruptions in communications resulting from cable ruptures (for example, caused by errant digging from a backhoe) or equipment failures at an intermediate node. Such disruptions otherwise could have a devastating effect on a large group of users.

Undersea Optical Cable Systems

The first transoceanic fiber optic cable systems were installed in the Atlantic and Pacific Oceans in 1988 and 1989. Initially these systems operated at 280 Mb/s per fiber pair using 1310-nm lasers and single-mode fibers. The links consisted of a series of point-to-point optical fiber segments between electronic-based undersea regeneration points that were located nominally 60 km apart. Later the transmission capacity of these links was upgraded to 2.5 Gb/s and the regenerator spacing was increased to 100 km by converting the 1310-nm multiple-frequency light sources to 1550-nm single-frequency laser diodes. Later, the regenerator spacing was increased to 140 km.

Although these cable systems significantly improved the quality of the international telephone service, the optical-to-electrical conversion process at each regeneration point remained a capacity bottleneck. The introduction of erbium-doped optical fiber amplifiers (EDFA) eliminated this bottleneck from undersea lightwave systems by amplifying signals

directly in the optical domain. Since EDFAs operate over a 30-nm wavelength band, they are well suited for use with undersea WDM links, which have provided a further capacity increase. These undersea optical amplifiers are typically spaced about 45 km apart.

One example of the many worldwide installations of optically amplified WDM networks is the SEA-ME-WE-3 Cable System. This undersea network runs from Germany to Singapore, connecting more than a dozen countries in between. Hence the name SEA-ME-WE, which refers to Southeast Asia (SEA), the Middle East (ME), and Western Europe (WE). The network has two pairs of undersea fibers with a capacity of eight STM-16 wavelengths per fiber.

See also

Lasers: Optical Fiber Lasers. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Wavelength Division Multiplexing.

Further Reading

Bergano NS (1997) Undersea amplified lightwave system design. In: Kaminow IP and Koch TL (eds) *Optical Fiber Telecommunications*, vol. IIIA, pp. 302–335. New York: Academic.

Goralski WJ (2002) SONET/SDH, 3rd edn. New York: McGraw-Hill.

Hecht J (1999) *City of Light*. New York: Oxford University Press.

Joyner CH (1997) Semiconductor laser growth and fabrication technology. In: Kaminow IP and Koch TL (eds) *Optical Fiber Telecommunications*, vol. IIIB, pp. 163–199. New York: Academic.

Kaiser P and Keck DB (1988) Fiber types and their status. In: Miller SE and Kaminow IP (eds) *Optical Fiber Telecommunications II*, pp. 29–54. New York: Academic.

Kao KC and Hockman GA (July 1966) Dielectric-fibre surface waveguides for optical frequencies. *Proceedings IEE*, vol. 113, 1151–1158.

Keiser G (2003) *Optical Communications Essentials*. New York: McGraw-Hill.

Keiser G (2000) *Optical Fiber Communications*, 3rd edn. Burr Ridge, IL: McGraw-Hill.

Mollenauer LF, Gordon JP and Mamyshev PV (1997) Solitons in high bit rate long-distance transmission. In: Kaminow IP and Koch TL (eds) *Optical Fiber Telecommunications*, vol. IIIA, pp. 373–460. New York: Academic.

Nyman B, Farries M and Si C (Oct 2001) Technology trends in dense WDM demultiplexers. *Optical Fiber Technology* 7(4): 255–274.

Ramaswami R and Sivarajan KN (2002) *Optical Networks*, 2nd edn. San Francisco, CA: Morgan Kaufmann.

Snitzer E (May 1961) Cylindrical dielectric waveguide modes. *Journal Optical Society of America* 51: 491–498.

Trischitta PR and Marra WC (1998) Applying WDM technology to undersea cable networks. *IEEE Communication Magazine* 36(2): 62–66.

Architectures of Optical Fiber Communication Systems

G Keiser, PhotonicsComm Solutions, Inc., Newton Center, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Simple Optical Fiber Links

The most basic optical fiber communication system architecture is a point-to-point link consisting of the elements shown in Figure 1. The key sections are a transmitter consisting of a light source and its associated drive circuitry, a cable offering mechanical and environmental protection to the optical fibers contained inside, and a receiver consisting of a photodetector plus amplification and signal-restoring circuitry. Very long links may include periodic optical amplifiers for boosting the level of the optical signal. In networks that use only simple point-to-point links, the optical fiber is used principally as a transmission medium and all the switching and processing of information bits is done by electronics.

At the transmitting end of an optical fiber link, information bits consisting of electronic pulses are used to modulate the light output of the source. The resulting light pulses are coupled into an optical fiber. As they travel along the fiber these signal pulses become progressively weakened and distorted due to attenuation and dispersion mechanism that are characteristic of an optical fiber. At the end of the fiber link a photodetector converts the optical pulses back to an electrical format for switching to another link or for processing at the end station.

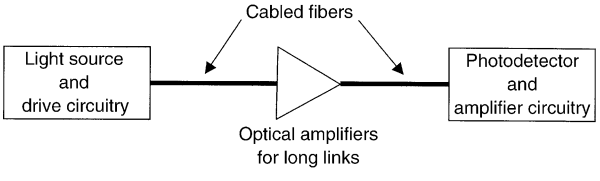


Figure 1 Basic elements of a point-to-point fiber optic link.

The cabled fiber is one of the most important elements in an optical fiber link, since it determines operational parameters such as transmission distance and achievable data rates. In addition to protecting the glass fibers during installation and service, the cable may contain copper wires for powering optical amplifiers or signal regenerators, which are needed periodically in long-distance links for amplifying and reshaping the signal.

Analogous to copper cables, the installation of optical fiber cables can be either aerial, in ducts, undersea, or buried directly in the ground. As a result of installation and/or manufacturing limitations, individual cable lengths will range from several hundred meters to several kilometers. Practical considerations such as reel size and cable weight determine the actual length of a single cable section. The shorter segments tend to be used when the cables are pulled through ducts. Longer lengths are used in aerial, direct-burial, or undersea applications. Splicing together individual cable sections forms continuous transmission lines for these long-distance links. For undersea installations, the splicing and repeater-installation functions are carried out on board a specially designed cable-laying ship.

One of the principal characteristics of an optical fiber is its attenuation as a function of wavelength, as shown in Figure 2. Early technology used the 800 to 900 nm wavelength band, since in this region the fibers made at that time exhibited a local minimum in the attenuation curve, and optical sources and photodetectors operating at these wavelengths were available. This region originally was referred to as the 'first window'. By reducing the concentration of hydroxyl ions and metallic impurities in the fiber material, in the 1980s manufacturers were able to fabricate optical fibers with very low loss in the 1100 to 1600 nm region. Two windows are defined here, these being the 'second window' centered around 1310 nm and the 'third window' centered around 1550 nm.

In 1998 a new ultra-high purifying process patented by Lucent Technologies eliminated virtually all water molecules from the glass material. The resulting fiber is designated as either a low-water-peak or a full spectrum fiber. By dramatically reducing the water-attenuation peak around 1400 nm, this process opens the transmission region between the second and third windows to provide around 100 nm more bandwidth than in conventional single-mode fibers. This particular fiber allows wider spectral tolerances on optical components, which reduces costs for moderate-distance metropolitan network applications.

Since the attenuation of low-water-peak fibers makes the designation of transmission windows obsolete, the concept of operational spectral bands arose for the 1260–1675 nm region. As shown in

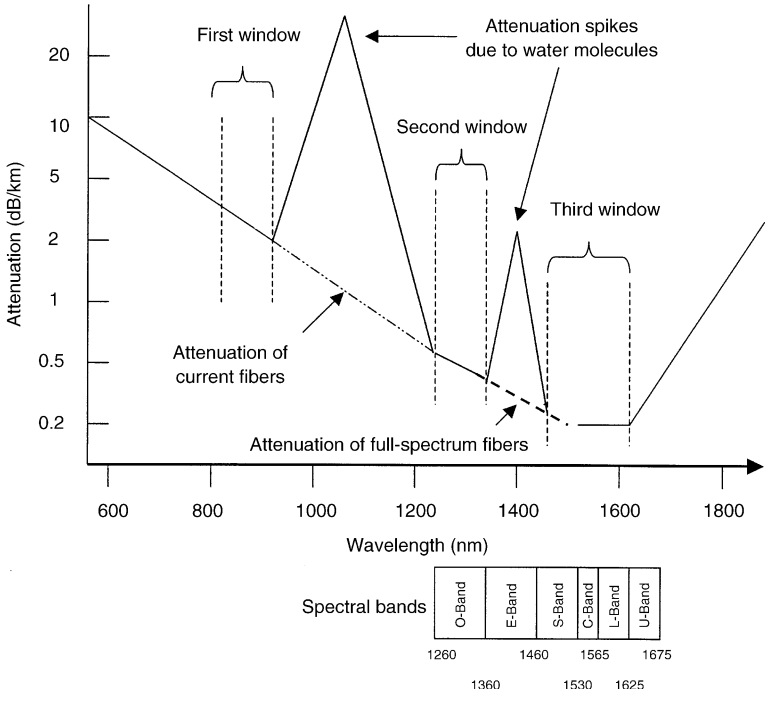


Figure 2 Attenuation versus wavelength for silica fibers and the concept of spectral bands.

Figure 2, the regions are known by the letters O, E, S, C, L, and U, which are defined as follows:

- Original band (O-Band): 1260–1360 nm;
- Extended band (E-Band): 1360–1460 nm;
- Short band (S-Band): 1460–1530 nm;
- Conventional band (C-Band): 1530–1565 nm;
- Long band (L-Band) 1565–1625 nm;
- Ultra-long band (U-Band): 1625–1675 nm.

Once the cable is installed, a light source that is dimensionally compatible with the fiber core is used to launch optical power into the fiber. Semiconductor light-emitting diodes (LEDs) and laser diodes are suitable for this purpose, since their light output can be modulated rapidly by simply varying the bias current at the desired transmission rate, thereby producing an optical signal. The electric input signals to the transmitter circuitry for the optical source can be either of an analog or digital form. For high-rate systems (nominally greater than 1 Gb/s), direct modulation of the source can lead to unacceptable signal distortion. In this case, an external modulator is used to vary the amplitude of a continuous light output from a laser diode source.

After an optical signal is launched into a fiber, it will become progressively attenuated and distorted with increasing distance because of scattering, absorption, and dispersion mechanisms in the glass material. At the receiver a photodiode will detect the weakened optical signal emerging from the fiber end and convert it to an electric current. The design of an optical receiver is inherently more complex than that of the transmitter, since it has to interpret the content of the weakened and degraded signal received by the photodetector. The principal figure of merit for a receiver is the maximum optical power necessary at the desired data rate to attain either a given error probability for digital systems or a specified signal-to-noise ratio for an analog system. The ability of a receiver to achieve a certain performance level depends on the photodetector type, the effects of noise in the system, and the characteristics of the successive amplification stages in the receiver.

Networks of Links

With the advent of fiber optic transmission lines, the next step in the evolution of the digital multiplexing scheme used by telecommunication providers was a standard signal format called synchronous optical network (SONET) in North America and synchronous digital hierarchy (SDH) in other parts of the world. In the mid-1980s, efforts started on developing a standard that would allow network engineers to interconnect fiber optic transmission equipment from various vendors through multiple-owner trunk networks. This resulted in a series of standards for SONET by the American National Standards Institute (ANSI) and a series of recommendations for SDH from the International Telecommunications Union (ITU). Examples of particular interest are the ANSI T1.105.06 standard and the ITU-T G.957 recommendation. Although there are some implementation differences between SONET and SDH, all SONET specifications conform to the SDH recommendations. A key characteristic of SONET and SDH is that they are usually configured as a ring architecture. This is done to create loop diversity for uninterrupted service protection purposes in case of link or equipment failures. The SONET/SDH rings commonly are called self-healing rings, since the traffic flowing along a certain path can be switched automatically to an alternate or standby path following failure or degradation of the link segment.

The basic building block and first level of the SONET signal hierarchy is called the Synchronous Transport Signal – Level 1 (STS-1), which has a bit rate of 51.84 Mb/s. Higher-rate SONET signals are obtained by byte-interleaving N STS-1 frames, which are then scrambled and converted to an Optical Carrier – Level N (OC-N) signal. Thus the OC-N signal will have a line rate exactly N times that of an OC-1 signal. For SDH systems the fundamental building block is the 155.52 Mb/s Synchronous Transport Module – Level 1 (STM-1). Again, higher-rate information streams are generated by synchronously multiplexing N different STM-1 signals to form the STM-N signal.

Three main features, each with two alternatives, classify all SONET/SDH rings, thus yielding eight possible combinations of ring types. First, there can be either two or four fibers running between the nodes on a ring. Second, the operating signals can travel either clockwise only (which is termed a unidirectional ring) or in both directions around the ring (which is called a bidirectional ring) Third, protection switching can be performed either via a line-switching or a path-switching scheme. Upon link failure or degradation, line switching moves all signal channels of an entire OC-N channel to a protection fiber. Conversely, path switching can move individual payload channels within an OC-N channel (e.g., an STS-1 channel within an OC-12 channel) to another path.

Of the eight possible combinations of ring types, the following two architectures have become popular for SONET and SDH networks:

- two-fiber unidirectional path-switched ring (two-fiber UPSR);
- two-fiber or four-fiber bidirectional line-switched ring (two-fiber or four-fiber BLSR).

The common abbreviations of these configurations are given in parentheses. They also are referred to as unidirectional or bidirectional self-healing rings (USHR or BSHR).

Figure 3 shows a two-fiber unidirectional path-switched ring network. By convention, in a unidirectional ring the normal working traffic travels clockwise around the ring, as indicated by the heavy arrows. For example, the connection from node 1 to node 3 uses links 1 and 2, whereas the traffic from node 3 to node 1 traverses links 3 and 4. Thus, two communicating nodes use a specific bandwidth capacity around the entire perimeter of the ring. If nodes 1 and 3 exchange information at an OC-3 rate in an OC-12 ring, then they use one-fourth of the capacity around the ring on all the primary links. In a unidirectional ring the counter-clockwise path is used as an alternate path for protection against link or node failures. To achieve this, the signal from a transmitting node is dual-fed into both the primary and protection fibers. This establishes a designated protection path on which traffic flows counter-clockwise, i.e., from node 1 to node 3 via links 5 and 6 (in that order), as shown in Figure 3b. A heavy line and a dashed line indicate the primary and protection paths, respectively.

Consequently, two identical signals from a particular node arrive at their destination from opposite directions, usually with different delays, as denoted in Figure 3b. The receiver normally selects the signal from the primary path. However, it continuously compares the fidelity of each signal and chooses the alternate signal in case of severe degradation or loss of the primary signal. Thus, each path is individually switched based on the quality of the received signal. For example, if path 2 breaks or equipment in node 2 fails, then node 3 will switch to the protection channel to receive signals from node 1.

Figure 4 illustrates the architecture of a four-fiber bidirectional line-switched ring. Here two primary fiber loops (with fiber segments labeled 1p through 8p) are used for normal bidirectional communication, and the other two secondary fiber loops are standby links for protection purposes (with fiber segments labeled 1s through 8s). In contrast to the two-fiber UPSR, the four-fiber BLSR has a capacity advantage

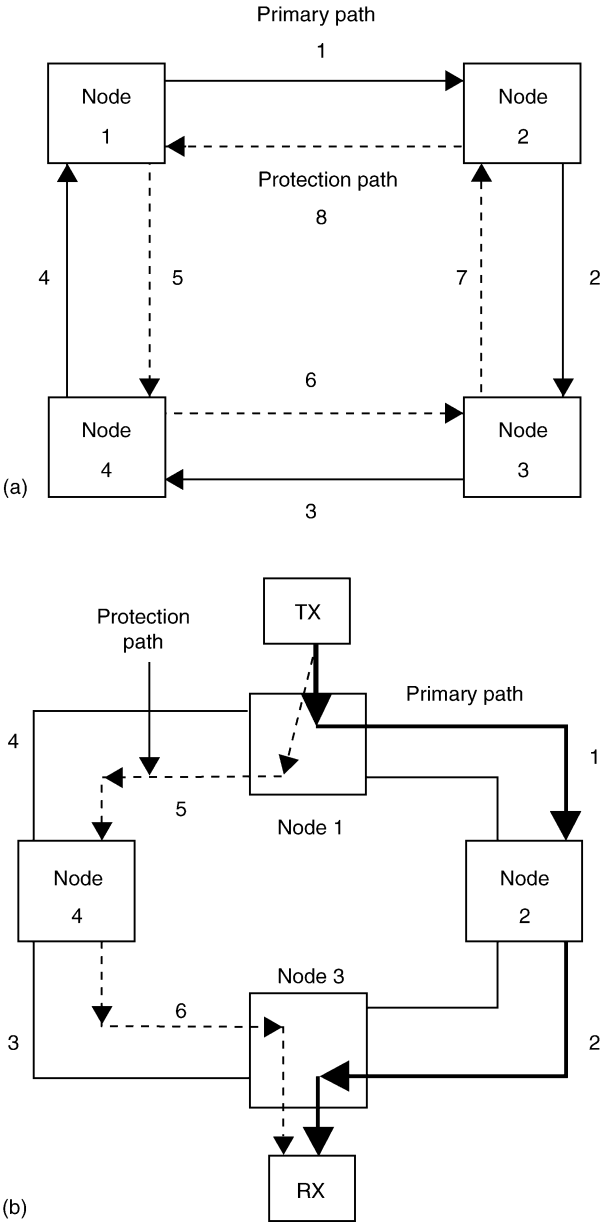


Figure 3 Schematic of a two-fiber unidirectional path-switched ring network.

since it uses twice as much fiber cabling and because traffic between two nodes is only sent partially around the ring.

To see the function and versatility of the standby links in the four-fiber BLSR, consider first the case where a transmitter or receiver circuit card used on the primary ring fails in either node 3 or 4. In this situation the affected nodes detect a loss-of-signal condition and switch both primary fibers connecting them to the secondary protection pair, as shown in Figure 5. The protection segment between nodes 3 and 4 now becomes part of the primary bidirectional loop. The exact same reconfiguration scenario will

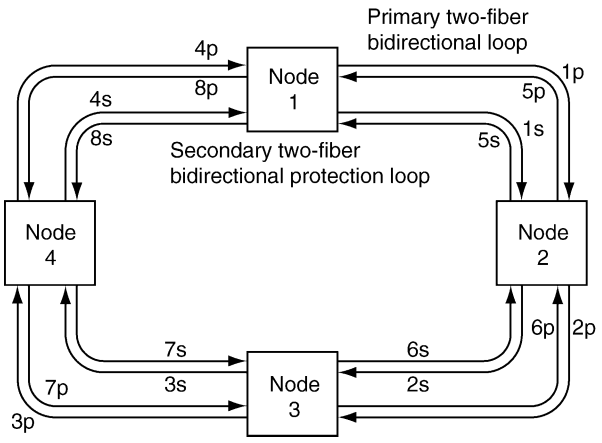


Figure 4 The architecture of a four-fiber bidirectional, line-switched ring.

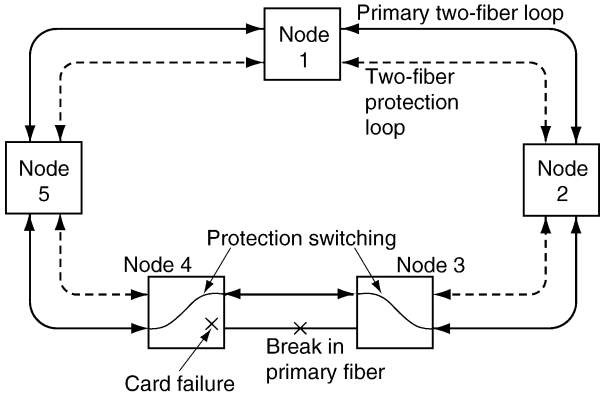


Figure 5 Reconfiguration of a four-fiber BLSR under transceiver failure.

occur when the primary fiber connecting nodes 3 and 4 breaks. Note that in either case the other links remain unaffected.

Now suppose an entire node fails, or both the primary and the protection fibers in a given span are severed, which could happen if they are in the same cable duct between two nodes. In this case the nodes on either side of the failed span internally switch the primary-path connections from their receivers and transmitters to the protection fibers, in order to loop traffic back to the previous node. This process again forms a closed ring, but now with all of the primary and protection fibers in use around the entire ring, as shown in Figure 6.

Single-Span WDM Links

An interesting and powerful aspect of an optical communication link is that many different wavelengths can be sent along a fiber simultaneously in the 1300–1600 nm spectrum. The technology of combining a number of wavelengths onto the same fiber is known as wavelength division multiplexing or WDM. Figure 7 shows the basic WDM concept. Here N independent optically formatted information streams, each transmitted at a different wavelength, are combined with an optical multiplexer and sent over the same fiber. Note that each of these streams could be at a different data rate. Each information stream maintains its individual data rate after being multiplexed with the other streams, and still operates at its unique wavelength. Conceptually, the WDM scheme is the same as frequency division multiplexing (FDM) used in microwave radio and satellite systems.

To see the potential of WDM, consider the characteristics of a high-quality optical source such as a distributed feedback (DFB) laser, which has a very narrow frequency spectrum on the order of 1 MHz, which is equivalent to a spectral linewidth of 10^{-5} nm. When using such a source, a guard band of 0.4–1.6 nm is typically employed. This is done to take into account possible drifts of the peak wavelength due to aging or temperature effects, and to give both the manufacturer and the user some leeway in specifying and choosing the precise peak emission wavelength. With such spectral band widths, simplex systems make use of only a very

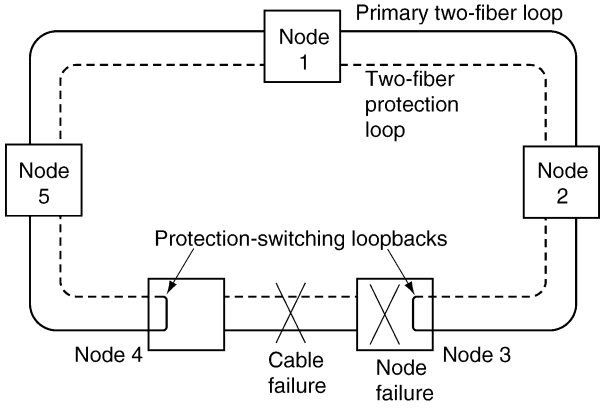


Figure 6 Reconfiguration of a four-fiber BLSR under node or fiber-cable failure.

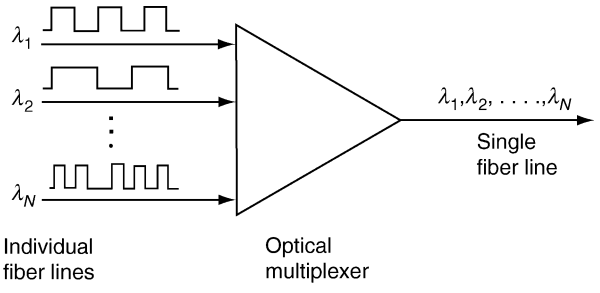


Figure 7 The basic WDM concept.

small portion of the transmission bandwidth capability of a fiber. This can be seen from Figure 2, which depicts the attenuation of light in a silica fiber as a function of wavelength. The curve shows that the two low-loss regions of a single-mode fiber extend over the wavelengths ranging from about 1260 to 1360 nm (the O-band) and from 1460 to 1625 nm (the S-, C-, and L-bands).

These regions can be viewed either in terms of spectral width (the wavelength band occupied by the light signal and its guard band) or by means of optical bandwidth (the frequency band occupied by the light signal). To find the optical bandwidth corresponding to a particular spectral width in these regions, consider the fundamental relationship $c = \lambda\nu$, which relates the wavelength λ to the carrier frequency ν, where c is the speed of light. Differentiating this yields $\Delta\nu = (c/\lambda^2)\Delta\lambda$, where the deviation in frequency $\Delta\nu$ corresponds to the wavelength deviation $\Delta\lambda$ around λ. Thus the optical bandwidth is $\Delta\nu = 14$ THz for a usable spectral band $\Delta\lambda = 80$ nm in the O-band. Similarly, $\Delta\nu = 15$ THz for a usable spectral band $\Delta\lambda = 120$ nm in the region spanning the S-, C-, and L-bands.

Since the spectral width of a high-quality source occupies only a narrow optical bandwidth, the two low-loss windows provide many additional operating regions. By using a number of light sources, each emitting at a different peak wavelength that is sufficiently spaced from its neighbor so as not to create interference, the fidelities of the independent messages from each source are maintained for subsequent conversion to electrical signals at the receiving end.

Since WDM is essentially frequency division multiplexing at optical carrier frequencies, the WDM standards developed by the ITU specify channel spacings in terms of frequency. A key reason for selecting a fixed frequency spacing, rather than a constant wavelength spacing, is that when locking a laser to a particular operating mode it is the frequency of the laser that is fixed. The ITU-T Recommendation G.692 specifies selecting the channels from a grid of frequencies referenced to 193.100 THz (1552.524 nm) and spacing them 100 GHz (0.8 nm at 1552 nm) apart. Suggested alternative spacings include 50 GHz and 200 GHz.

Historically the term dense WDM (DWDM) was used somewhat loosely and generally referred to the spacings denoted by Recommendation G.692. In 2002 the ITU-T released an updated standard aimed specifically at DWDM. This is Recommendation G.694.1, which specifies WDM operation in the S-, C-, and L-bands for high-quality, high-rate metro area network (MAN) and wide area network (WAN) services. It calls out for narrow frequency spacings of 100 to 12.5 GHz (or, equivalently, 0.8 to 0.1 nm at 1550 nm).

Alternatively, by using the larger spectral window available in low-water-content fibers, one can relax the spectral tolerances on the optical components and space the individual sources much further apart. This is the basis of coarse WDM (CWDM), which is useful for metropolitan area networks where variations in attenuation between different wavelength channels is not critical over short transmission distances. In 2002 the ITU-T released Recommendation G.694.2, which is aimed specifically at CWDM. The CWDM grid is made up of 18 wavelengths defined within the range 1270–1610 nm (O- through L-bands) spaced by 20 nm with wavelength-drift tolerances of ±2 nm.

A key feature of WDM is that the discrete wavelengths form an orthogonal set of carriers which can be separated, routed, and switched without interfering with each other. This holds as long as the total optical power intensity is kept sufficiently low to prevent nonlinear effects such as stimulated Brillouin scattering and four-wave mixing processes from degrading the link performance.

The implementation of WDM networks requires a variety of passive and/or active devices to combine, distribute, isolate, and amplify optical power at different wavelengths. Passive devices require no external control for their operation, so they are somewhat limited in their application in WDM networks. These components are mainly used to split and combine or tap off optical signals. The performance of active devices can be controlled electronically, thereby providing a large degree of network flexibility. Active WDM components include tunable optical filters, tunable sources, and optical amplifiers.

Passive Optical Networks

Another step towards realizing the full potential of optical fiber transmission capacity is the concept of all-optical WDM networks to extend the versatility of communication networks beyond architectures such as those provided by SONET. These networks can be classified as either broadcast-and-select or wavelength-routing networks. In general, broadcast-and-select techniques employing passive optical stars, buses, or wavelength routers are used for local network applications, whereas active optical components form the basis for constructing wide-area wavelength-routing networks. Broadcast-and-select networks can be categorized as single-hop or multihop networks. Single-hop refers to networks where

information transmitted in the form of light reaches its destination without being converted to an electrical form at any intermediate point. On the other hand, intermediate conversion can occur in a multihop network.

Figure 8 shows two alternate physical architectures for a WDM-based local network. Here N sets of transmitters and receivers are attached either to a star coupler or a passive bus. Each transmitter sends its information at a different fixed wavelength. All the transmissions from the various nodes are combined in a passive star coupler or coupled onto a bus and the result is sent out to all receivers.

Each receiver sees all wavelengths and uses a tunable filter to select the one wavelength addressed to it. In addition to point-to-point links, this configuration can also support multicast or broadcast services, where one transmitter sends the same information to several nodes. An interesting point is that these passive WDM networks are protocol transparent. This means that different sets of communicating nodes can use different information-exchange rules (protocols) without affecting the other nodes in the network. This is analogous to current time-division-multiplexed telephone lines in which voice, data, or facsimile services are sent in different time slots without interfering with each other.

Although the architectures of single-hop broadcast-and-select networks are fairly simple, there needs to be careful dynamic coordination between the nodes. For example, if the network is set up so that a transmitter sends its information at a unique fixed wavelength, then the destined receiver needs to be informed when a message is being sent to it, so that it can tune its selective filter to that wavelength. Also conflicts need to be resolved in cases where two stations transmitting at different wavelengths want to send information to the same recipient simultaneously.

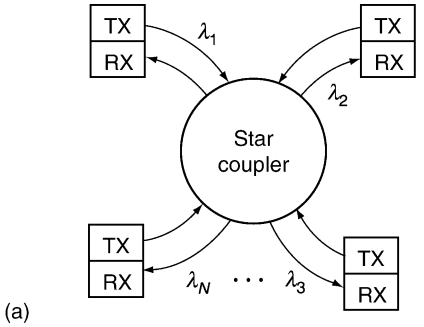


Figure 8 Two alternate physical architectures for a WDM-based local network.

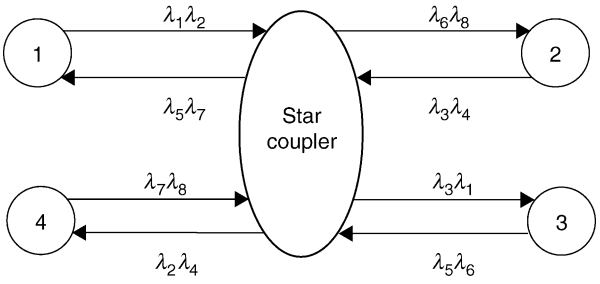


Figure 9 Example of a four-node broadcast-and-select multihop network.

A drawback of single-hop networks is the need for rapidly tunable lasers or receiver optical filters. The designs of multihop networks avoid this need. Multihop networks generally do not have direct paths between each node pair. Each node has a small number of fixed-tuned optical transmitters and receivers. Figure 9 shows an example of a four-node broadcast-and-select multihop network where each node transmits on one set of two fixed wavelengths and receives on another set of two wavelengths. Stations can send information directly only to those nodes that have a receiver tuned to one of the two transmit wavelengths. Information destined for other nodes will have to be routed through intermediate stations.

The flow of traffic can be seen from Figure 9. If node 1 wants to send a message to node 2, it first transmits the message to node 3 using λ_1. Then node 3 forwards the message to node 2 using λ_6. In contrast to single-hop networks, with this scheme there are no destination conflicts or packet collisions in the network, since each wavelength channel is dedicated to a particular source–destination link. However, for H hops between nodes, there is a network throughput penalty of at least $1/H$.

Wavelength-Routed Optical Networks

Wavelength-routed networks overcome the limitations of passive optical networks through wavelength reuse, wavelength conversion, and optical switching. The physical topology of a wavelength-routed network consists of optical wavelength routers interconnected by pairs of point-to-point fiber links

in an arbitrary mesh configuration, as illustrated in Figure 10. Each link can carry a certain number of wavelengths, which can be directed independently to different output paths at a node. Each node may have logical connections with several other nodes in the network, where each connection uses a particular wavelength. Provided the paths taken by any two connections do not overlap, they can use the same wavelength. Thereby the number of wavelengths is greatly reduced. For example, the connection from node 1 to node 2 and from node 2 to node 3 can both be on λ_1, whereas the connection between nodes 4 and 5 requires a different wavelength (λ_2).

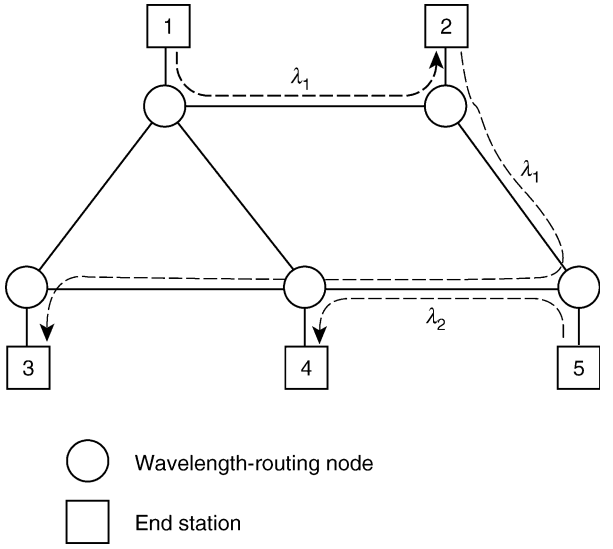


Figure 10 The physical topology of a wavelength-routed network.

A high degree of path modularity, capacity scaling, and flexibility in adding or dropping channels at a user site can be achieved by introducing the concept of an optical cross-connect architecture in the physical path structure (the so-called path layer) of an optical network. These optical cross-connects (OXCs) operate directly in the optical domain and can route very high-capacity WDM data streams over a network of interconnected optical paths.

Contentions could arise when channels having the same wavelength but traveling on different input fibers enter the OXC and need to be switched simultaneously to the same output fiber. This could be resolved by assigning a fixed wavelength to each optical path throughout the network, or by dropping one of the channels and retransmitting it at another wavelength. However, in the first case, wavelength reuse and network scalability (expandability) are reduced, and in the second case the add/drop flexibility of the OXC is lost. These blocking characteristics can be eliminated by using wavelength conversion at any output of the OXC.

As an example consider the 4×4 OXC shown in Figure 11. Here two input fibers are each carrying two wavelengths. Either wavelength can be switched to any of the four output ports. The OXC consists of three 2×2 switch elements. Suppose that λ_2 on input fiber 1 needs to be switched to output fiber 2 and that λ_1 on input fiber 2 needs to be switched to output fiber 1. This is achieved by having the first two switch elements set in the bar state (the straight-through configuration) and the third element set in the cross state, as indicated in Figure 11. Obviously without wavelength conversion there would be wavelength contention at both output ports.

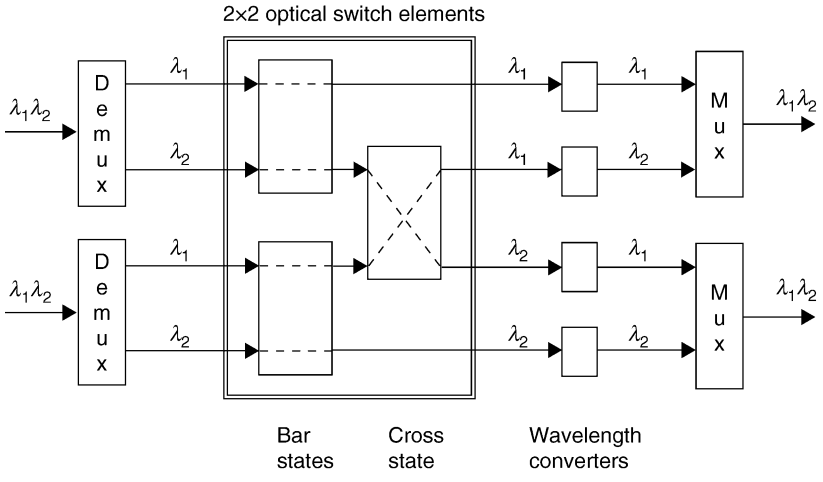


Figure 11 Example of a 4×4 optical cross-connect (OXC).

By using wavelength converters, the cross-connected channels can be converted to noncontending wavelengths.

See also

Lasers: Optical Fiber Lasers. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Blumenthal DJ, Prucnal PR and Sauer JR (1994) Photonic packet switches: architectures and experimental implementations. *Proceedings of the Institute of Electrical and Electronics Engineering* 82(11): 1650–1667.

Forouzan BA (2001) *Introduction to Data Communications and Networking*, 2nd edn. Burr Ridge, IL: McGraw-Hill.

Goralski WJ (2002) *SONET/SDH*, 3rd edn. New York: McGraw-Hill.

Gumaste A and Antony T (2003) *DWDM Network Designs and Engineering Solutions*. Upper Saddle River, NJ: Prentice Hall.

Hecht J (2002) *Understanding Fiber Optics*, 4th edn. Upper Saddle River, NJ: Prentice Hall.

Iannone E and Sabella R (1996) Optical path technologies: a comparison among different cross-connect architectures. *Journal of Lightwave Technology* 14(10): 2184–2196.

Keiser G (1999) A review of WDM technology and applications. *Optical Fiber Technology* 5(1): 3–39.

Keiser G (2000) *Optical Fiber Communications*, 3rd edn. Burr Ridge, IL: McGraw-Hill.

Keiser G (2002) *Local Area Networks*, 2nd edn. Burr Ridge, IL: McGraw-Hill.

Keiser G (2003) *Optical Communications Essentials*, New York: McGraw-Hill.

Ramaswami R and Sivarajan KN (2002) *Optical Networks*, 2nd edn. San Francisco: Morgan Kaufmann.

Special Issue (1995) *Undersea Communications Technology. AT&T Technical Journal* 74(1): 4–92.

Trischitta PR and Marra WC (1998) Applying WDM technology to undersea cable networks. *IEEE Communications Magazine* 36(2): 62–66.

Zang H, Jue JP, Sahasrabuddhe L, *et al.* (2001) Dynamic lightpath establishment in wavelength-routed WDM networks. *IEEE Communications Magazine* 39: 100–108.

Free Space Optical Communications

R Martini, Stevens Institute of Technology, Hoboken, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Free space optical (FSO) communication is the wireless transmission of data via a modulated optical beam directed through free space, without fiber optics or other optical systems guiding the light. The fundamental idea goes back to ancient times, as light (or smoke) signals were used to transmit information. From a more modern point of view, Graham Bell's patent on the photophone may mark the onset of modern FSO techniques, as it transmitted audio signals (i.e., voice) via the modulation of sunlight. A renaissance of FSO systems started with the availability of lasers, light sources with high output power and high coherence, which allowed the accurate direction of the light beams over long distances. During the 1970s and 1980s the main proposed application of FSO systems was for secure and long distance (50–1000 km) communication, mainly targeted for ground–satellite or satellite–satellite communication. This focus changed drastically over the last decade as a new market for FSO grew in the establishment of high bandwidth data link and their integration over a locally restricted area.

The main competitors in this market are the fiber-based optical network, the RF communication system, as well as the low bandwidth copper cable-based system. In comparison to the closely related wireless radio frequency transmission, the higher frequency of the optical carrier ($\sim 10^{14}$–10^{15} Hz) thereby allows for much higher transmission rates, comparable to those of typical fiber optic networks. On the other hand, the use of an optical carrier also results in much more directed beam propagation, which requires an undisturbed line-of-sight between emitter and detector. This restricts the application of most FSO systems to a range between a few hundred meters up to several kilometers, which are still favorable for distribution of high bandwidth networks over a locally restricted area. This makes FSO a highly attractive candidate for the 'last-mile' distribution of high-bandwidth Ethernet to the individual homes.

The simplicity of setup of FSO links, as well as their modularity, is thereby their biggest advantage, compared to fiber-based networks. It makes them not only highly cost efficient, but it eases the maintenance or

allows for fast and easy upgrades, as it does not require any extensive and time-consuming installation – in contrast to the installation of fiber optic cables. This advantage makes FSO highly attractive for temporary installations (emergency or short-time high broadcast situations, such as the Olympic Games), as well as to overcome geometrical restraints (river, seas, etc.). As most detector and emitter systems are typically based on the same electro-optical components used for fiber optic networks, they seamlessly integrate and expand an existing network without complications in data handling.

Standard FSO System

Figure 1 shows schematically the setup for a typical FSO system connecting two separate local networks. It consists of: an electronic data input; a small but powerful light source which can be modulated; emitter optics which shape the emitted beam to a highly directed beam; the atmosphere as transmitting device; detector optics which receive the transmitted light and focus it onto a photodetector; and an electronic amplifier which serves as data output.

Compared to a fiber optical communication system, the main difference is the use of the atmosphere as the transmission medium, in contrast to guided optical propagation. The transmission properties are thereby determined by the atmospheric conditions, which can lead to degradation and/or redirection of the beam. The main influences which have to be taken into account are atmospheric scattering, molecular absorption, turbulence effects (including beam wander and fading), as well as the impact of low-visibility weather situations like rain, snow, or fog. These depend strongly on wavelength and linewidth of the used emitter system, distance between emitter and receiver, as well as the environmental conditions. As a consequence, the characteristics of the atmospheric channel determine the layout of the other FSO components – and have to be considered first.

Atmospheric Losses

The overall losses of the atmospheric transmission are determined by the Beer–Lambert absorption law:

$$I = I_0 \exp[-\alpha L] \qquad [1]$$

where I is the intensity at the receiver, I_0 the intensity of the transmission beam after the emitter optics, and L is the range (i.e., distance between emitter and detector optics). The total extinction coefficient $\alpha = \alpha_{\text{Rayleigh}} + \alpha_{\text{Mie}} + \alpha_{\text{Absorption}}$ is based on three different contributions: Rayleigh scattering, Mie scattering, and absorption.

In contrast to absorption, the losses due to scattering originate from redirection of light and will be discussed first. Depending on the size of the scattering partner, two separate regimes have to be taken into account. For scatter smaller than the wavelength, the effect can be described based on Rayleigh's classical formula. The scattering cross-section of an element with mass m, charge q, natural oscillation wavelength λ_0, is then:

$$\sigma_s = \frac{fq^4\lambda_0^4}{6\pi\varepsilon_0^2 m^2 c^4}\frac{1}{\lambda^4} \qquad [2]$$

where ε_0 is the dielectric constant, c the speed of light, and λ the wavelength of the incident light, leading to an extinction coefficient of $\alpha_{\text{Rayleigh}} = \sigma_s N_s$, where N_s is the concentration of the scattering elements. The λ^{-4} dependence in the scattering predicts stronger scattering for shorter wavelengths and Rayleigh scattering is therefore dominant for short and visible wavelengths (which gives the sky its blue color). For typical FSO

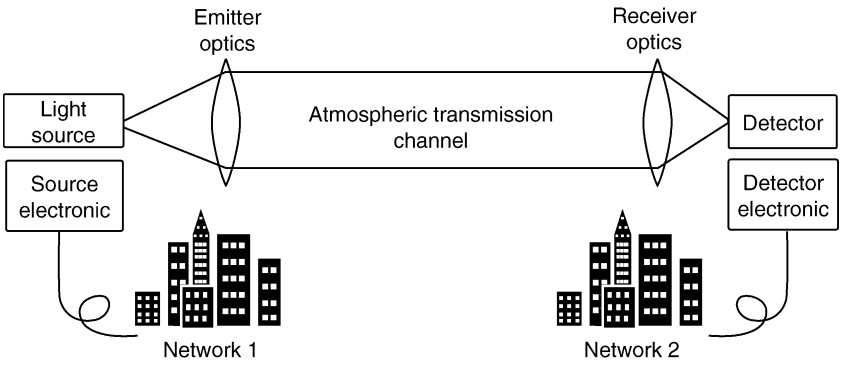


Figure 1 Typical components of a FSO system.

wavelengths in the region of near-infrared radiation (NIR, 800–2000 nm) or mid-infrared radiation (MIR, 2000–20 000 nm), the contribution of Rayleigh scattering is typically relatively low and can be neglected.

On the other hand, Mie scattering is, in most cases, the strongest contribution for FSO systems – but also the hardest to predict. It is due to resonant scattering effects, where the particle size is of the wavelength of the scattered light. For the typical FSO wavelengths these particles include water and oil droplets, dust, smog, fly ash, and aerosols (which can originate from pollution). A correct calculation of the Mie scattering coefficient is possible, but requires knowledge of the exact particle concentration as well as their size distribution. As these numbers are not only difficult to measure, but also fluctuate continuously depending on weather situation, time of day, season, temperature, climate, pollution, and other factors, a numerically correct evaluation is not possible.

Historically, most systems rely on an empirical formula, which estimates the attenuation coefficient to

$$\sigma_{\text{Mie}} = \frac{3.91}{V}\left(\frac{\lambda}{550}\right)^{-p} \qquad [3]$$

where V denotes the visibility and the value of p ranges from 1.6 for high visibility ($V > 50$ km) and 1.3 for average visibility (50 km $> V >$ 6 km) down to 0.585 $V_{1/3}$, respectively for low visibility. The formula is based on data collected in the 1950s and 1960s. More accurate measurements show that the actual scattering cross-sections differ quite strongly, especially for low-visibility cases. Especially, the predicted dependence on wavelength is not correct. The failure of such a general model is not surprising, as the actual scattering properties are strongly dependent on the weather situation and the environmental conditions. It is the focus of ongoing research to find more accurate models describing the scattering condition in multiple environments and conditions to provide the background for more accurate empirical models.

Apart from the light redirection due to scattering, absorption either by molecules or aerosols is nearly as important for the transmission calculation of a FSO system. The strongest molecular absorptions are typically due to water, carbon dioxide, as well as ozone. A typical absorption spectrum can be easily calculated using standard software together with the publicly available HITRAN database, both available from ONTAR. The upper part of Figure 2 shows a low-resolution transmission spectrum for the interesting wavelength region for a path length of 1 km.

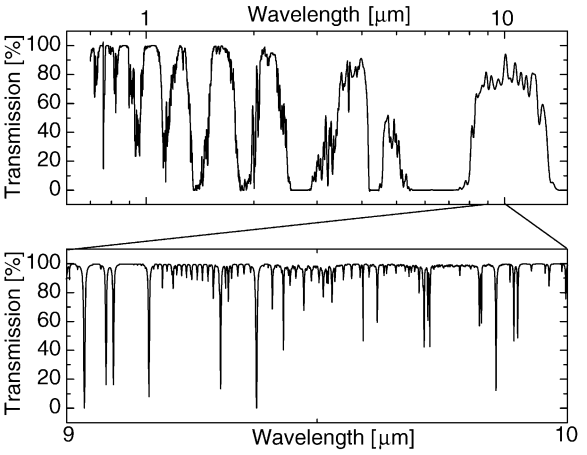


Figure 2 Upper: Low-resolution transmission spectrum of the atmosphere showing molecular absorption for a path length of 1 km. Lower: high-resolution spectrum of the atmospheric windows around 9.5 μm.

Within this spectrum, several well-known atmospheric windows exist, which are naturally preferred for FSO applications. These include the region around 830 nm, 1550 nm, and from 8 μm to 12 μm. For the exact selection of a wavelength, a closer look is needed. As an example, a high-resolution analysis of the atmospheric windows around 9.5 μm is shown in the lower part of Figure 2, which identifies additional absorption lines as well as highly transparent wavelength ranges within the window. Such substructure has to be taken carefully into account for the FSO wavelength selection, especially if lasers are employed, which have a linewidth comparable to the width of the absorption line.

Besides the molecules, aerosol particles also contribute to the absorption spectrum. Calculation of their contribution can also be performed using available software package, such as Air Force's MODTRAN and LOWTRAN program as well as OPAC from the LMU University in Munich. Figure 3 shows a typical low-resolution transmission spectrum for 50% humidity as well as 99% humidity, where by scattering as well as absorption effects have both been included in the calculation. The occurrence of sharp absorption features, as well as the increasing losses for wavelengths above 10 μm, is typically due to the aerosol absorption, whereas the scattering dominates the function for wavelengths below 3 μm. Losses due to the aforementioned Rayleigh scattering are also shown in the same diagram, demonstrating its low influence at longer wavelengths.

MODTRAN and LOWTRAN also allow the calculation of transmission spectra under rain and snow conditions, where rain and snow consist of

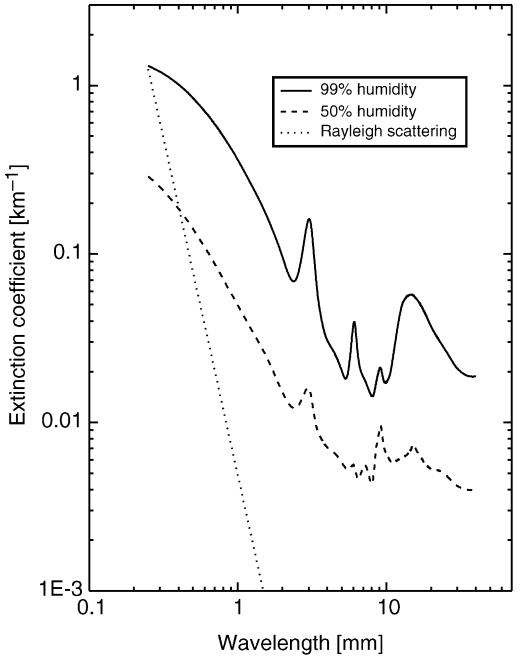


Figure 3 Typical low-resolution transmission spectrum calculated based on absorption and scattering losses from aerosols for 50% humidity (dashed line) and 99% humidity (solid line). The dotted line shows additionally the losses due to Rayleigh scattering.

clusters of water and aerosol with diameters from 200 μm to 2 mm and above. Depending on the severity of the precipitation, the losses can range from ~0.5 dB/km for light rain up to 13 dB/km for cloudbursts, requiring an additional power margin in the preparation of the FSO link. Significantly stronger is the impact of fog, as typically particles with much lower sizes are present. The losses inflicted by a moderate fog (visibility 500 m) range above 20 dB/km, surpassing the effect of a cloudburst; for thick fog the attenuation can even grow beyond 200 dB/km. However, depending on the size distribution within the fog itself, the losses can be highly wavelength dependent. Recent experiments show that long wavelength FSO systems operating in the MIR wavelength region have a much stronger stability under the influence of thick fog than one operating at NIR wavelengths.

Turbulence and Scintillation

In contrast to a static medium, the atmosphere is a dynamic mixture which undergoes continuous changes due to wind, temperature, and other influences. This creates a constant change in the parts of the atmosphere which are contributing to the FSO channel and therefore in the optical properties of the beam path. The index of refraction can differ, not only in its total value, but also within subsections of it, creating a random fluctuation in the beam paths. Such localized changes in the refractive index of the air are schematically shown in Figure 4, where turbulences of different sizes modify the beam path and disturb the beam front. The turbulence in the atmosphere can be quantified by the refractive-turbulence structure coefficient C_n^2, which can range from $10^{-13}\,\mathrm{m}^{-2/3}$ in the middle of the afternoon (highest value) to $10^{-17}\,\mathrm{m}^{-2/3}$ one hour after sunrise or sunset. As the turbulence effect is typically due to strong temperature difference, C_n^2 is strongest near hot surfaces (e.g., tar roofs or asphalt in the summer) and decreases with altitude. The effects of such turbulences on a directed light beam (such as used in FSO systems) include mainly beam wander and scintillation effects.

Beam wander is thereby based on large size turbulences which act as a lens on the propagating beam. Consequently, a narrow focused beam could be randomly steered away from the detector area, creating a fluctuation in the detected intensity.

Stronger impact for FSO links has scintillation, which is better known as the twinkling of a star or the heat shimmer of the horizon on a hot day. It is based on the continuous variation of the beam front which can lead to intensity fluctuations (like the twinkling) and loss in the image resolution (heat shimmer). Strong intensity fluctuations can thereby cause signal loss as well as receiver saturation, both resulting in downtime in the link. Such local intensity disturbances can occur up to sizes of $\sqrt{\lambda L}$, reaching, for example, 4 cm for a 1.55 μm link after 1 km. For small fluctuations, the intensity follows a classical stochastic behavior and can be described by the irradiance variation (normalized to its mean value) as

$$\sigma^2 = \exp\left[0.5\left(\frac{2\pi}{\lambda}\right)^{6/7} L^{11/6} C_n^2\right] - 1 \qquad [4]$$

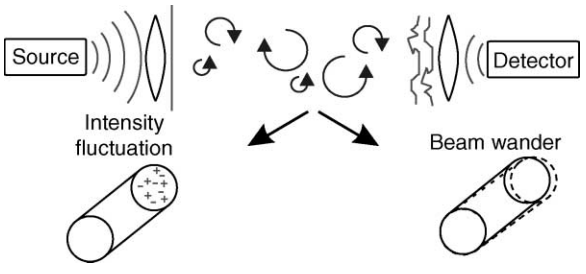


Figure 4 Schematic depiction of atmospheric turbulences with different sizes in FSO systems. Shown below are illustrations of its main effects on a directed beam causing intensity fluctuations and beam wander.

with usual values in the range from 5 to 70%. Nonetheless, it should be pointed out that actual turbulence effects can differ quite strongly from these predictions, due to local inhomogeneities in temperature distribution. Consequently, test measurements are necessary for the correct evaluation of a specific installation site.

What makes scintillation effects difficult to handle for communications systems is the relatively slow frequency of fluctuations in the order of a few hundred Hz or less. The corresponding downtimes for a link can range from 1 ms to 10 ms, or even up to seconds. This is magnitudes higher than typical downtimes in standard communications links and cannot be compensated by standard forward error correction systems. One possibility to reduce the influence of scintillation is the method of aperture averaging, whereby multiple beams originating from slightly different positions are deployed and targeted at the same receiver. Assuming that the variation in each of the separate beam paths is random and uncorrelated, the signal-to-noise ratio will decrease by the square root of the number of beams. Recent experiments have impressively demonstrated this effect and allow reduction of intensity fluctuation to an acceptable level.

Another alternative to decreasing the scintillation effects is in the application of longer wavelength FSO systems. Whereas beam wander is relatively independent from the wavelength, scintillations effects scale with $\lambda^{7/6}$, as turbulences smaller than the wavelength cannot obscure its beam front. Using a 2 times longer wavelength, for example, should have the same impact as signal averaging with 4 different beams.

Beam Spreading and Pointing

Based on optical diffraction theory, the beam leaving the emitter optics will be divergent and its beam diameter will consistently increase during transmission through the optical channel. Assuming an optimal alignment and using standard geometrical optics, the losses due to beam spreading can estimate the ratio of emitted power (P_E) to received power (P_R):

$$R_{BS} = \frac{P_R}{P_E} = \frac{D_R}{(D_E + L\theta)} \qquad [5]$$

where D_R is the diameter of the receiving optics, D_E the diameter of the emitting optics, and θ the divergence of the beam. The minimal divergence of the beam is typically determined by the light source and is given by $\theta_L = \lambda / D_L M^2$, where D_L is the initial size of the laser beam and M^2 the laser-mode structure parameter, for a typical semiconductor laser θ_L below 0.1 mrad. In many cases, the divergence of the emitted beam is commonly enlarged by slight defocusing of the emitting optics – if the power budget allows for the additional losses. The beam size at the receiver is typically several times larger than the receiving aperture, which not only eases the alignment of the system, but also reduces pointing losses. This includes the effect of beam wander due to turbulence as well as swaying of the buildings or the mount. If those effects are solely compensated for by an increased beam spread, divergence can easily reach values of 2–20 mrad for short-distance links. Obviously this is not an option for longer-distance FSO links, as the losses will increase dramatically. In these cases, active tracking systems are incorporated in the emitter optics allowing for continuously adjustment to achieve and keep an optimal alignment. Typical tracking devices are thereby based on servo-driven Gimbal mounts, steering mirrors, or recently, MEMS devices.

Laser Eye Safety

As the atmospheric transmission channel of most FSO systems is not completely isolated, there exists the possibility of a human person intercepting the beam. To allow for this, most FSO systems are designed to be eye-safe following the voluntary ANSI standard, which limits the output irradiance to a certain value depending on the deployed wavelength. As radiation below 1400 nm is still being focused on the retina, it causes a higher irradiance, which results in a lower damage threshold. For 800 nm laser systems, the minimum permissible exposure (MPE) level for a 10 s long exposure is therefore recommended not to exceed 1 mW cm^{-2}, whereas for 1.55 μm or longer wavelength systems, an MPE of 100 mW cm^{-2} is suggested. The recommendations for incoherent light sources such as LEDs are thereby less stringent as they cannot be focused quite as strongly as laser light.

The Light Source

Following the discussion above, the optimal light source for FSO system should be emitting at a wavelength away from any absorption line (or should be much broader then the absorption line), have a relatively long wavelength as well as a relative low divergence, be eye-safe, and allow for high modulation speed. In reality, most FSO systems use either GaAsAl laser diodes or LEDs emitting at 0.8 μm, or the typically more expensive 1.5 μm laser diodes

based on the InGaAsP semiconductor system. The output power of both laser systems is typically in the range of 1 mW to 100 mW, which for the longer wavelength case can be boosted to 1–10 W using erbium-doped fiber amplifiers (or even higher for specific applications). The longer wavelength system has the additional advantage of higher ANSI standards as discussed above, allowing stronger eye-safe emission power, but is typically far more expensive.

Both laser types were developed and optimized for standard fiber-based communication systems allowing direct modulation up to 10 Gbit/s. To achieve higher bandwidth in FSO systems, standard techniques known from fiber-based networks can easily be deployed, like wavelength division multiplexing which increases the bandwidth up to state-of-the-art levels of fiber-based networks.

Contrary to semiconductor lasers, LEDs provide a much less expensive alternative, but allow only for modulation up to 10–100 MHz. Their broad emission spectrum is thereby an advantage, as it is less sensitive to sharp absorption lines which – as discussed earlier – can have a drastic impact on the FSO link performance. Consequently, LEDs have been successfully implemented in several commercial FSO systems targeted at the short-range market.

Recently developed sources, which have also been applied for FSO systems, include vertical-cavity surface-emitting lasers as well as quantum cascade lasers (QCL) which emit in the mid-infrared regime. The latter devices have been predicted to have a strong impact on the FSO market, as they combine high-output power (up to 1 W) together with fundamentally high modulation bandwidth (2.5 Gbit/s recently demonstrated and higher rates predicted) at wavelengths ranging from 3.5 μm to 12 μm and beyond. They can access the long-wavelength atmospheric window atmospheric for the FSO systems (typically reserved for CO_2 lasers) with the convenience of direct semiconductor lasers like direct electrical modulation, small size, and low power consumption. Recently, first FSO experiments, employing experimental QCLs, demonstrated the predicted higher stability of the link in low-visibility situations, including strong fog as well as high dust concentration. Nonetheless, to date the high cooling requirement for QCLs restricts more general use of them, but their very fast development, recent reports of cw-operation of room-temperature QCLs as well as FSO links with TE-cooled QCLs in pulse operation, give hope for a solution to this problem.

The Detector System

Once the light reaches the receiver, it is typically collimated using a short focal length telescope or a single lens system. To optimize efficiency without sacrificing ease of alignment a compromise has to be made between a wide aperture for greater light collection and a short focal length. The concentrated light then hits the detector, which, depending on the wavelength, is either Si or InGaAs-photodiodes. Si-based devices can detect up to wavelengths of 1100 nm and InGaAs systems are typically used for the 1.5 μm region, respectively. To enhance the sensitivity of the detectors, avalanche photodiodes are preferred which internally enhance the detected signal by a factor ranging from 20 to 100. To ensure high bandwidth operation (10 MHz–10 GHz), the size of the detecting device has to be kept relatively small (20 to 100 μm) avoiding parasitical capacitance effects – but raising the requirement on the detecting optics.

The sensitivity of these NIR detectors is typically limited by the shot-noise, which is due to statistical fluctuations of the received phonons and scales with the square root of its number. For practical purposes the internal noise of the detector is quantified using the corresponding amount of incident optical power leading to a comparable signal. The values for this noise equivalent power (NEP) can range from microwatts to tens of nanowatts. Other sources for detector noise can include a background current from the detector–amplifier combination, typically referred to as Johnson noise.

Mid-infrared detectors are additionally objected to strong background radiation due to thermal emission, which peaks around the 10 μm at temperature of 300 K. The use of additional cold filters and high-resolution spectral filters can help to overcome this restriction, but still the sensitivity of those detectors is lacking compared to NIR systems.

For communication system the noise is more commonly quantified as bit error rate (BER), quantifying the average percentage of wrongly received bits (submitted '1', detected '0' or vice versa). A BER greater than 10^{-9} is usually referred to as an error-free system, as the residual errors can be corrected employing forward error correcting algorithms. In contrast, the performance of FSO systems can more easily be measured in its signal-to-noise ratio (SNR). Following standard communication theory, a peak SNR of 12 and an average SNR of 6 is required to achieve a BER of 10^{-9}, guaranteeing error-free communication. Consequently range limits for FSO links are typically based on SNR = 6.

Link Budget and FSO Range Equation

Classical link budget estimation is a commonly used tool to access the performance of FSO systems. To take into account the losses due to atmospheric absorption, scattering, turbulence, and beam spreading, as well as optical losses in the used receiver and emitter optics, they are deducted from the starting power emitted from the FSO light source. The estimated intensity at the detector is then compared with the intensity needed to establish error-free communication (for example SNR = 6), the calculated difference is the link margin of the specific system, i.e., the amount of additional losses under which the FSO link will still be working.

Comparing this number to losses from nonpredictable fluctuations (like heavy rain or fog) gives an insight under which implications the link will fail. Using weather station data, the frequency of such extremely low-visibility occurrences can be obtained and a number for the reliability of the FSO link be estimated. Typical well-installed FSO systems estimate very high reliabilities, between 99.9% and 99.999%, which have been obtained in installation and verified in long-term measurements.

Another way to estimate the performance for a given system is based on the FSO range equation, where the received power on the detector (P_R) is calculated to

$$P_R = P_E \frac{D_R}{(D_E + L\theta)^2} TK \exp[-\alpha L] \qquad [6]$$

T denotes the optical losses of the transmission and receiver system, and K denotes additional losses for incoherent lights sources and is given by the ratio of LED emitting area to detector area, and is $K = 1$ for laser light sources. This equation can also be used to estimate losses for different weather situations and obtain the maximum range L of a specific FSO system.

Recent Research and Future Considerations

Existing FSO systems have impressively demonstrated their versatility and stability, as well as their seamless integration into existing networks, on multiple occasions. A prime example is the re-establishment of communication links between Merrill–Lynch Brokerage and Wall Street in New York, in the aftermath of the terrorist attack on September 11th, which destroyed one of the central fiber optic hubs in downtown Manhattan. The installation of a commercial FSO system by Terabeam ensured a fast replacement and a stable continuation of their work.

The demonstrated performance of such links compares to standard fiber-based links. The inclusion of wavelength division multiplexing as well as other standard or nonstandard techniques in FSO systems proves the strengths of a system, which is based on identical hardware for emitter and detectors as fiber-based networks. Nonetheless, FSO allows a multitude of applications not possible in fibers, which current research projects are targeted at. This includes, for example, long wavelength FSO communication systems, continuously ground-satellite links and free-space quantum communication for extreme secure data transmission – to list only a few of the possibilities. One very obvious example is extreme wavelength division multiplexing, which uses one link operating in all atmospheric windows at the same time. Such a FSO system would not only allow very high data rates and extreme stability against weather effects, but could also employ new kinds of wavelength-based modulation and encoding schemes increasing the security of the data link. Concluding the actual state of the FSO system as well as ongoing research, free space optical communication has all the ingredients needed to be one of the mayor key ingredients of future networks – and based on the increasing demand of bandwidths they will contribute strongly.

List of Units and Nomenclature

c	speed of light
C_n	refractive-turbulence structure coefficient
D_E	diameter of the emitting optics
D_R	diameter of the receiving optics
I	intensity
L	range (i.e., distance between emitter and detector optics)
m	mass
M^2	laser-mode structure parameter
N_s	concentration
P_E	emitted power
P_R	received power
q	charge
T	optical losses of the transmission and receiver system
V	visibility
α	extinction coefficient
ε_0	dielectric constant
θ	divergence of the light beam
λ	wavelength
λ_0	natural oscillation wavelength
σ_s	scattering cross-section

See also

Environmental Measurements: Doppler Lidar; Optical Transmission and Scatter of the Atmosphere. **Imaging:** Infrared Imaging; Lidar.

Further Reading

Andrews LC and Phillips RL (1998) *Laser Beam Propagation through Random Media*. Bellingham, CA: SPIE Optical Engineering Press.

Begley DL (ed.) (1991) *Selected Papers on Free-Space Laser Communications*. SPIE Milestone Series. Bellingham, CA: SPIE Optical Engineering Press.

IEE Proceedings Optoelectronics, *Special Issue on Free Space Optical Communications*. December 1996.

Jha AR (2000) *Infrared Technologies*. New York: John Wiley & Sons.

Killinger D (2002) *Free Space Optics for Laser Communications Through the Air*. Washington: Optics and Photonics News, Optical Society of America, p. 36.

Measures R (1984) *Laser Remote Sensing*. New York: John Wiley & Sons.

Mecherle GS (ed.) (2003) *Free-Space Laser Communication Technologies XV*. Bellingham, CA: SPIE Optical Engineering Press.

Weichl H (1990) *Laser Beam Propagation in the Atmosphere*. Bellingham, CA: SPIE Optical Engineering Press.

Willebrand H and Ghuman BS (2001) *Free-Space Optics: Enabling Optical Connectivity in Today's Networks*. Indianapolis, IN: SAMS.

Wolfe WL and Zissis GJ (1978) *The Infrared Handbook*. Ann Arbor: Enviromental Research Institute of Michigan.

Zuev VE (1988) *Laser Beams in the Atmosphere*. New York: Consultant Bureau.

Lightwave Transmitters

J G McInerney, National University of Ireland-Cork, Cork, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical fiber telecommunications has changed human society forever, providing the capacity for affordable and ubiquitous communication. It provides data rates and transport economies far in excess of those available using purely electronic means. It is doubtful whether the Internet and worldwide web, or the vast infrastructure of wireless cellular telephony, would be viable without optical fiber technology. Its rapid development and acceptance has been driven largely by contemporaneous successes in manufacturable low-loss optical fiber cables, sensitive pin-FET photoreceivers, and reliable high-performance laser diode-based transmitters. The earliest developments in silica glass fibers at Standard Telecommunication Laboratories, UK, (then part of the US-based ITT Corporation) and Corning Glass Works (USA) in the 1960s and 1970s, were pivotal in determining the directions of modern fiber technology, as were later key inventions of rare-earth doped fiber amplifiers at the University of Southampton (UK) and AT&T Bell Laboratories (USA) in the 1980s and early 1990s. However, perhaps the greatest contributors to the success of fiber optics have been semiconductor injection lasers, or laser diodes, which in various forms have made up the vast majority of all fiber optic transmitters, and the entirety of long-haul transmitters, as well as pump sources for the doped-fiber amplifiers which enable modern networks. The fascinating story of the development of fiber optics is to a large extent driven by the dramatic progress in the development of high-performance, reliable semiconductor lasers. These emitted at wavelengths near 1300 and 1550 nm, which respectively correspond to the dispersion and attenuation minima of single-mode silica optical fibers.

Semiconductor Laser Principles

Semiconductor lasers were first demonstrated in research laboratories at General Electric, IBM Corporation and the MIT Lincoln Laboratory (all USA) as early as 1962, although it took almost two decades for the basic science and engineering of materials, fabrication, reliability, and performance design issues to be developed sufficiently for their use in practical communication systems. Early successes were obtained using the GaAs/AlGaAs lattice-matched material system, in which a lightly doped GaAs or ternary compound AlGaAs active layer is sandwiched between two lattice-matched heterojunctions with n- and p-doped AlGaAs layers with higher Al fractions than the active layer, so that their refractive indices are lower and their bandgaps higher, providing optical and charge-carrier confinement. In all cases the material is epitaxially grown, meaning that the entire laser chip forms a single crystal. The gain, and hence the laser emission, occurs at the direct bandgap of the active layer, at wavelengths in the range 750–870 nm depending on the Al fraction.

Later evolutions used the quaternary compound GaInAsP clad by InP guiding layers, where the In and P fractions are chosen to match the lattice constant to the InP substrate.

In general, optical gain in a laser diode is generated by creating an electron-hole plasma in the vicinity of the forward-biased junction, a situation corresponding to population inversion in a conventional laser, and with sufficient forward current the material becomes transparent, in that the gain exactly equals the absorption and scattering losses, at a particular wavelength. When this arrangement is enclosed within a suitable optical cavity, and the forward current is increased further, the system begins to lase, that is to oscillate continuously to produce coherent optical radiation at a wavelength which satisfies two basic conditions: it must be within the set of wavelengths at which the material produces sufficient gain, and it must be close to one of the electromagnetic modes at which the optical cavity is resonant. These conditions, which define lasing threshold, are characterized by the equality of the optical gain and the total losses, that is the material losses (absorption, scattering, and free-carrier plasma) and the cavity losses which include the light output.

Early laser diodes in both the GaAs/AlGaAs and GaInAsP/InP systems were of the edge-emitting type, in which the light is emitted from the edge of the laser chip, perpendicular to the growth direction in the plane of the wafer. Although this geometry complicates laser production flow, particularly the testing function which requires that the wafer be scribed into bars to define the laser output facets, edge-emitting lasers (EELs) make up the vast majority of optical fiber transmission sources. An alternative laser diode structure is the vertical cavity surface-emitting laser (VCSEL), in which emission occurs in the growth direction perpendicular to the plane of the wafer. VCSELs offer major advantages in spectral stability, beam quality, manufacturability and cost, but their output powers are low (a few mW) and they are not yet available at the key telecommunication windows around 1300 nm and 1550 nm, due to difficulty in forming the necessary high-reflectivity Bragg mirrors in the GaInAsP/InP material system. Major initiatives are currently underway to produce long-wave VCSELs in the GaAs/AlGaAs system using InAs quantum dots, or dilute nitrides such as InGaAsN, or hybrid approaches such as InP-based gain media fused to high-reflectivity mirrors using GaAs/AlGaAs, dielectric coatings, air gaps, and others. At present, however, all long-haul optical fiber transmission is based on edge-emitting laser diodes fabricated in GaInAsP/InP.

General Structure and Requirements of Optical Fiber Communication Systems

The first optical fiber communication links were simple point-to-point affairs, consisting essentially of series-connected transmitters, fibers, and receivers. The transmitters were simple Fabry–Perot laser diodes, onto which data were encoded by direct digital modulation of the injection current at rates ~100 Mbit/s. The first commercial optical fiber link was built in 1976, a single fiber cable linking two switching centers of the Illinois Bell Telephone Co just 2.5 km apart in the Chicago metropolitan area, using 850 nm GaAs laser technology. In 1988, the first transatlantic fiber link, AT&T's TAT-8, was completed linking endpoints in New Jersey, England, and France with three fiber pairs carrying signals generated by 1300 nm GaInAsp single mode lasers. In these early systems, data were regenerated frequently by repeaters each consisting of a receiver, signal conditioner and transmitter. Each fiber in TAT-8 carried a single optical frequency modulated at 280 Mbit/s and the construction cost of the system was ~$50 k/km, resulting in an economic figure of merit of over $600 k/Mbit/s. Twenty-five years later, current systems cost the same per kilometer to build in absolute dollars despite inflation, a major effective cost reduction enabled largely by replacement of the expensive repeaters by doped-fiber optical amplifiers. Moreover, each fiber is now highly multiplexed, with the potential for hundreds of wavelengths each carrying 10 Gbit/s signals generated by 1550 nm laser transmitters, so that the unit cost has plummeted to ~$200/Mbit/s, a reduction of three and a half orders of magnitude.

In the remainder of this article, we will describe transmitters for both short- and long-haul systems where the channel spacing is only tens of GHz in the optical carrier frequency. Modern short-haul systems either feed client networks as part of bi-directional transceivers, or drive metro and local access networks connected directly by photonic switches to terminals of the long-haul systems.

DWDM Transceivers

Modern practice uses modular units containing transmitters and receivers allowing bi-directional data flow. In the send mode, data are applied to a long-haul transmitter at the local (client) source and coupled through the external fiber network to the remote long-haul receiver. In the receive mode, data arrive from a remote long-haul transmitter and are coupled (via a regenerator including signal

conditioning, clock recovery, and error correcting steps) to a local short-haul transmitter leading to the client's internal network. Each transceiver card therefore contains a long-haul transmitter and receiver set (known collectively as dense wavelength division multiplexing (DWDM) transport devices) and a short-haul transmitter and receiver set (known as client interfaces or local transport devices).

Transmitter Requirements

Short-haul transmitters may be simple Fabry–Perot edge-emitting laser diodes, or vertical cavity laser diodes, or even superluminescent LEDs for sub-Gbit/s data rates. They are generally not required to be at one of the minimum-loss telecom windows near 1550 nm, as the transmission distances are short. However, as client networks grow into large metropolitan networks, it is likely that 1300 nm sources will be used to minimize dispersion. Power requirements are modest, a few mW peak, at data rates ~1–10 Gbit/s. Single mode laser emission and polarization control are not required by present day systems, although polarization-mode dispersion may limit transmission at higher data rates. Both return-to-zero (RZ) and nonreturn-to-zero (NRZ) coding may be used, and direct modulation of the laser amplitude is the norm.

For long-haul transmitters the situation is dramatically different. Operation in one of the standard erbium-doped fiber amplifier (EDFA) bands near 1550 nm is mandatory: the C (conventional) band extends from 1530–1565 nm, the S (short-wave) band from 1460–1530 nm) and the L (long-wave) band from 1565–1625 nm. In the C-band, output powers ~10 mW are generally sufficient, and powers above ~30 mW are limited by nonlinear effects such as self-phase modulation and four-wave mixing in the fiber. In the S- and L-bands, where the available gain from EDFAs is less, lasers may need to operate close to the nonlinear limit, several tens of mW, although various schemes (differential pumping, gain-flattening filters, etc.) are being evaluated to flatten the global EDFA gain spectrum.

In current practice, a single fiber utilizing DWDM, using only the C-band, can carry up to 1 Tbit/s comprising 100 channels at 10 Gbit/s on each channel, the current standard. For such performance a channel spacing of 50 GHz in optical frequency is required. Using all three bands with similar density would enable ~3 Tbit/s, which could be doubled by using 25 GHz channel spacings. Ultimately, even higher channel densities and hence overall system data rates are possible, with a practical limit ~10 Tbit/s. The channel spacing is a key parameter in that it determines the laser design, specifically its spectral stability and linewidth, and the overall transmitter design in that laser chirp (modulation induced dynamic spectral shift) must be less than half the channel spacing. In practice, only single-wavelength and coarse wavelength division multiplexing (CWDM) systems may be modulated directly. All modern DWDM systems require external modulation, for example using electro-absorption or Mach–Zehnder interferometric devices outside the laser cavity. For the latter cases, stabilization of the laser wavelength is required for example by on-chip gratings to form distributed feedback (DFB) or distributed Bragg reflector (DBR) lasers, or by external fiber gratings.

Directly Modulated Lasers

Direct modulation of semiconductor lasers is convenient and effective: by modulating the laser injection current, one modulates the carrier density and hence the optical gain, resulting in modulation of the laser output up to ~10 Gbit/s. The actual modulation bandwidth is determined by interactions between photons and carriers, whose respective decay lifetimes, τ_s and τ_p, are determined by cavity losses and total recombination rates. The simplest form of such interactions is described by the photon and carrier rate equations for the injected carrier density N and photon density S in a single lasing mode:

$$dN/dt = J/ed - gNS - N/\tau_s \quad [1]$$

$$dP/dt = gNS + \beta N/\tau_s - S/\tau_p \quad [2]$$

where J is the injected current density, e the electronic charge, d is the thickness of the active region (hence J/ed is the rate of injection of carriers per unit volume), g is the gain coefficient per unit length and β is the fraction of the spontaneous emission coupled into the lasing mode, typically $\sim 10^{-2}$–10^{-5} for edge-emitting lasers (product of geometric and spectral overlap factors). These equations simply state that the rate of change of the carriers or photons is given by the rate of generation less the loss rate. Additional terms are required in the presence of optical feedback or coupling (in which case the photon density is replaced by the complex amplitude and phase of the optical electric field) or for extremely rapid modulation where the traveling wave nature of the disturbances in the photon fields is significant. These so-called traveling-wave rate equations for the photons are of the form:

$$dS_+/dt + cdS_+/dz = gNS_+ + \beta N/\tau_s \quad [3]$$

$$dS_-/dt - cdS_-/dz = gNS_- + \beta N/\tau_s \qquad [4]$$

with the same carrier rate equation as [1] using the total photon number S made up of the forward-traveling component S_+ and the backward-traveling component S_- : $S = S_+ + S_-$.

Both sets of rate equations are based on approximations such as homogeneous gain broadening, neglect of transverse field effects, and diffusion. These are valid in most situations but care is required when analyzing lasers with large transverse apertures, or when ultrashort (picosecond) pulses are involved.

Small-Signal Modulation

Small-signal modulation may be analyzed by a linear perturbation analysis of the rate equations, leading to a conjugate pair of poles in the response function, with damped relaxation oscillations at frequency:

$$f_R = (1/2\pi)\sqrt{(gS_0/\tau_p)} \qquad [5]$$

or, in terms of the injected current density J (assuming $S \sim G(J - J_{th})/ed$ with Γ the electromagnetic confinement factor of the mode):

$$f_R = (1/2\pi)\sqrt{[\Gamma g(J - J_{th})/ed]}$$

f_R can be regarded as the resonant frequency for the interaction between the carriers and photons. As a practical matter, lasers can be modulated up to $\sim 2f_R$ but the modulation response rolls off rapidly with increasing frequency above f_R. Using explicit expressions for the threshold gain, we can write f_R as:

$$f_R = (1/2\pi)\sqrt{[(GN_T g\tau_p + 1)(J/J_{th} - 1)/\tau_s\tau_p]} \qquad [6]$$

where N_T is the transparency carrier density $\sim 10^{18}\,\mathrm{cm}^{-3}$. f_R can take values in excess of 10 GHz for a well-designed laser. It should be noted, however, that high modulation bandwidth almost always requires high differential gain dg/dN and short τ_p. Increasing dg/dN requires p-doping (which increases optical loss) and/or some special quantum-confining structure such as wells, wires, or dots. Decreasing τ_p inevitably leads to greater optical loss and thus a higher threshold.

The damping rate of the relaxation oscillations is also important in limiting modulation bandwidth, and also (when increased) in reducing the sensitivity of the laser to optical feedback. For bulk or quantum well lasers, the condition for critical damping is $A^2 = 4B$ with $A = (S/g\tau_s) - \sigma[N - N_T]/g\tau_p$ and $B = \sigma[1 + (S_0/g\tau_s) - (1 - \beta)(N - N_T)]$. For quantum dot lasers there are additional damping terms due to carrier transport and thermalization which may restrict small-signal modulation bandwidths to a few GHz.

Experimentally, small-signal modulation properties may be determined using a simple sampling oscilloscope, low-noise tunable signal generator and spectrum analyzer. A combination of these elements with automated frequency sweeping may be found in a scalar network analyzer acting as an *s*-parameter test set. The modulated laser is connected to port 1 and a high speed photodetector connected to port 2, then a swept-frequency measurement of the transfer characteristic s_{21} is performed. In addition to the limitations on modulation response due to the carrier-photon resonance, practical lasers are limited by RC parasitics in the laser chip (junction impedance) and its package. For the highest modulation speeds, packages need to be designed as microwave transmission line components with effective impedance matching and low back reflections, as characterized by the voltage standing wave ratio (VSWR).

Large-Signal Modulation

Large-signal direct modulation can be analyzed by numerical integration of the rate equations. In general, its results are beyond the scope of this review, but for effective high-speed response, digital modulation of laser diodes should be carried out with a pre-bias close to threshold. Modulation bandwidth also increases with laser power, which is limited for systems considerations by fiber nonlinearities, and for reasons of laser reliability. In pulse code modulation, care must be taken that the laser is near critical damping, to minimize thermal patterning effects due to long strings of ones (high power) or zeros (low power). Experimentally, large signal modulation is analyzed by constructing eye diagrams, in which pulse traces on a fast oscilloscope are continuously overlaid when the laser is modulated using pseudorandom binary pulses. When the signal quality is high, the high and low digital levels are easily distinguishable, resulting in an open 'eye' in the accumulated traces.

Requirements for Externally Modulated Lasers

The requirement for external modulation occurs when the laser chirp exceeds half the desired channel spacing in the communication system. Chirp occurs due to changes in the refractive index, and hence the optical phase, during modulation. This dynamic phase shift then results in an instantaneous frequency shift.

For direct modulation we have instantaneous wavelength

$$\lambda(t) = \lambda(0)(\mu(t)/\mu(0) \quad [7]$$

Where $\mu(t)$ is the refractive index at time t. μ shifts with carrier density N due to combinations of several factors (free carrier plasma dispersion, thermal bandgap shrinkage, bandgap renormalization, dynamic Burstein–Moss effects) which give $d\mu/dN \sim -3 \times 10^{-20}\,\text{cm}^3$, which for large signal digital modulation can result in tens of GHz. The actual degree of chirp depends on the so-called antiguiding or linewidth enhancement factor a given by

$$\alpha = -2k_0(d\mu/dN)/(dg/dN) \quad [8]$$

where $k_0 = 2\pi/\lambda_0$ is the free-space wavenumber. α as defined is positive, typically in the range 2–7 depending on the material, operating carrier density, and wavelength relative to the gain peak. Distributed feedback (DFB) or distributed Bragg reflector (DBR) lasers, which use integrated gratings to control the lasing wavelength, can have lower chirp by detuning from the gain peak but therefore are likely to require careful temperature stabilization. Quantum dot lasers, based on 3D nanoclustered active regions, have the potential for very low chirp, narrow linewidths, and reduced wavelength shifts, so that directly modulated DWDM transmitters may be possible.

Integration of laser diodes with electro-absorption or Mach–Zehnder type modulators is achieved in the transmitter optical subassembly and may in some cases be accomplished by monolithic integration on the same chip. Using separate modulators allows each device to be optimized independently but requires optical assembly, alignment, and retention. Mach–Zehnder modulators in X-cut lithium niobate, for example, have essentially zero chirp and can operate to 40 Gbit/s and above, with dynamic extinction ratios (ratio between fully on and fully off) of ~20 dB. Integrated semiconductor M-Z modulators have chirp an order of magnitude less than those of directly modulated lasers, so that channel spacings ~25 GHz are possible in DWDM systems with suitable temperature and wavelength controls. Although the details of such modulators are beyond the scope of this article, the laser requirements are to generate ~10 mW of continuous power with stable center wavelength and narrow linewidth. Such lasers are usually mounted in hermetic packages (e.g., the current 14-pin butterfly standard) with integrated thermo-electric cooler, power monitor photodiode, an optical isolator to suppress optical backreflections which cause instabilities and self-pulsing, and optional wavelength locking optics.

Reliability of Lasers in Fiber Optic Systems

From the early days, when laser diode lifetimes were measured in seconds even at cryogenic temperature, enormous progress has been made in achieving materials purity and reducing crystalline defects. Today's fiber optic laser transmitters have projected lifetimes of decades. In common with electronic devices, diode lasers fail at a rate given by a 'bathtub' curve, that is the failure rate $r(t)$ defined as the probability of failure per unit time at time t, has relatively high values at low t (early failures) and high t (wearout failures) and very low values in between. In terms of the population of lasers $n(t)$ we have:

$$r(t) = (-1/n(t))\, dn(t)/dt \quad [9]$$

If Δn is the number of samples which fail in time Δt, then assuming Δt begins at $t = 0$, the effective or average failure rate r_{eff} over the interval is

$$r_{\text{eff}}(t) = \Delta t \Delta n/n(0) \quad [10]$$

In reliability science it is customary to define r_{eff} in FITs (failures integrated in time) with the time-span chosen to yield statistically significant numbers. For electronic and photonic devices, it is customary to select the time interval $\Delta t = 10^9$ h (1 billion operating hours), so that if 1% of the devices fail in 10 years ($\sim 10^5$ hours) we have $r_{\text{eff}} \sim 100$ FITs, which is the order of magnitude required by modern fiber optic laser sources.

In terms of actual statistical models, failure rates can be estimated using failure probability density functions $f(t)$. This is related to $r(t)$ by $f(t) = r(t)S(t)$ where $S(t)$ is the cumulative probability of surviving until t. When $r(t)$ is nearly constant, as in early failures due to material defects or process errors, the statistics are approximately exponential:

$$f(t) + (1/\tau) \exp(-t/\tau) \quad [11]$$

or Weibull-distributed (an exponential is a Weibull distribution with $\beta = 1$):

$$f(t) = (\beta/t^{\beta})t^{b-1} \exp[-(t/\tau)^b] \quad [12]$$

but this description is not suited to wearout failures for which $f(t)$ shows a significant increase as the population ages. For such cases, the lognormal distribution is often applicable:

$$f(t) = (1/\sigma t\sqrt{(2\pi)}) \exp[-(\ln t - \ln \tau)^2/2\sigma^2] \quad [13]$$

where in each case τ is approximately the mean time to failure (MTTF).

Laser diodes undergo standardized qualification tests to demonstrate sufficient reliability prior to being used in commercial systems. Operationally, failure statistics must be tabulated and the best fit obtained. Burn-in procedures are used to screen early failures. Chip and module failures should be distinguished clearly, as the latter includes many additional factors such as thermal and power management, light coupling, mechanical or chemical integrity.

Given that effective lifetimes of decades are required, it is clearly necessary to accelerate aging to produce statistically meaningful failure rates in reasonable times $\sim 10^3$ h. For the most common modes of laser chip degradation, due to recombination-induced aggregation of defects in the crystalline epitaxial material, these are thermally activated and current driven, so that it is customary to use a modified Arrhenius law at temperature T:

$$\tau \sim (1/J^n) \exp(E_a/k_B T) \qquad [14]$$

where J is the operating current density, n is an index ~ 2, E_a is the activation energy, and k_B is Boltzmann's constant. In terms of thermal acceleration, we have a factor $F = \tau(T_A)/\tau(T_B)$ with:

$$F = \exp[(E_a/k_B)(1/T_A - 1/T_B)] \qquad [15]$$

Typical values of E_a for laser diodes are ~ 1 eV so that acceleration factors $\sim 10^3$ are possible.

Finally, it should be noted that laser diodes are generally subject to catastrophic failure in the event of overdriving or static discharge, the failure mode being thermal facet damage at $\sim 1-10$ MW/cm^2 for continuous-wave operation, the exact value depending on the material, surface preparation and specifically surface state absorption and its temperature coefficient.

Conclusions

Laser diodes are ideal sources for optical fiber communication systems and have propelled the development of fiber optics from its origins in the 1960s to the present day. They are compact, rugged, efficient, and reliable sources of light at key wavelength ranges such as 1300–1310 nm (short haul high speed links) and 1500–1600 nm (long haul amplified systems). They are capable of direct modulation at gigabit rates for simple systems but require external modulation and careful wavelength control for dense wavelength division multiplexed terabit systems. Transmitter lasers for real systems must satisfy stringent reliability conditions.

See also

Lasers: Edge Emitters; Semiconductor Lasers. **Semiconductor Physics:** Quantum Wells and GaAs-based Structures.

Further Reading

Telcordia Technologies (formerly Bellcore): document family FR-796 on Reliability and Quality Generic Requirements for telecommunications equipment. Latest issue is 003 dated Jul 2003. The specific document related to laser reliability is GR-468 – Generic Reliability Assurance Requirements for Optoelectronic Devices used in Telecommunications Equipment. Standards for other network photonic, electronic and mechanical components may be found in the document family FR-2063 Network Equipment-Building System Family of Requirements, issue 003 dated Feb 2003 and specifically FR-357 – Network Equipment-Building System Family of Requirements: Component Design for Reliability.

Adams MJ and Thomas B (1977) Detailed calculations of transient effects in semiconductor injection lasers. *IEEE Journal of Quantum Electronics* QE-13: 580.

Agrawal GP (2002) *Fiber Optic Communication Systems*, 3rd edn. New York: Wiley.

Alferov ZhI, Andreev VM, Portnoi EL and Trukan MK (1970) AlAs-GaAs heterojunction injection lasers with a low room-temperature threshold. *Soviet Physics – Semiconductors* 3: 1107.

Basov NG, Krokhin ON and Popov YuM (1961) Production of negative-temperature states in p-n junctions of degenerate semiconductors. *Soviet Physics – JETP* 13: 1320.

Bernard MGA and Duraffourg G (1961) Laser conditions in semiconductors. *Physica Status Solidi* 1: 699.

Bhattacharya P (1996) *Semiconductor Optoelectronic Devices*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.

Cheng J and Dutta NK (eds) (2000) *Vertical Cavity Surface-Emitting Semiconductor Lasers: Technology and Applications*. London: Taylor & Francis.

Chuang S-L (1995) *Physics of Optoelectronic Devices*. New York: Wiley.

Coldren LA and Corzine SW (1995) *Laser Diodes and Photonic Integrated Circuits*. New York: Wiley.

Dupuis RD (1987) An introduction to the development of the semiconductor laser. *IEEE Journal of Quantum Electronics* QE-23: 651.

Fukuda M (1991) *Reliability and Degradation of Semiconductor Lasers and LEDs*. Norwood, MA: Artech House.

Fukuda M (1998) *Optical Semiconductor Devices*. Wiley.

Guekos G (ed.) (1998) *Photonic Devices for Telecommunications*. New York: Springer.

Hall RN, Fenner GE, Kingsley JD, Soltys TJ and Carlson RO (1962) Coherent light emission from GaAs junctions. *Physical Review Letters* 9: 366.

Hayashi I, Panish MB, Foy PW and Sumski S (1970) Junction lasers which operate continuously at room temperature. *Applied Physics Letters* 17: 109.
Holonyak N Jr and Bevacqua SF (1962) Coherent (visible) light emission from Ga(As1-xPx) junctions. *Applied Physics Letters* 1: 82.
Ito R, Nakashima H, Kishino S and Nakada O (1975) Degradation sources in GaAs-AlGaAs double-heterostructure lasers. *IEEE Journal of Quantum Electronics* QE-11: 551.
Kaminow IP and Li T (eds) (2002) *Optical Fiber Telecommunications IV (two parts)*. Amsterdam: Elsevier.
Kaminow IP and Koch TL (eds) (1997) *Optical Fiber Telecommunications III (two parts)*. London: Academic Press.
Kapon E (1998) *Semiconductor Lasers, (two parts)*. London: Academic.
Kartalopoulos SV (1999) *Introduction to DWDM Technology*. New York: Wiley.
Kartalopoulos SV (1999) *Understanding SONET/SDH and ATM*. IEEE Press.
Kasap SO (2001) *Optoelectronics and Photonics: Principles and Practices*. Englewood Cliffs, NJ: Prentice-Hall.
Kazovsky L, Benedetto S and Willner A (1996) *Optical Fiber Communication Systems*. Norwood, MA: Artech House.
Keiser G (1991) *Optical Fiber Communications*. New York: McGraw-Hill.
Kressel H and Lockwood HF (1974) A review of gradual degradation phenomena in electroluminescent diodes. *Journal de Physique* 35: C3–223.
Lasher G and Stern F (1964) Spontaneous and stimulated recombination radiation in semiconductors. *Physical Review* 133: A553.
Milonni PW and Eberly JH (1988) *Lasers*. New York: Wiley.
Morthier G and Vankwikelberge P (1997) *Handbook of Distributed Feedback Laser Diodes*. Norwood, MA: Artech House.
Nathan MI, Dumke WP, Burns G, Dill FH Jr and Lasher G (1962) Stimulated emission of radiation from GaAs p-n junctions. *Applied Physics Letters* 1: 62.
Numai T (2004) *Fundamentals of Semiconductor Lasers*. New York: Springer.
Paoli TL (1977) Changes in the optical properties of cw (AlGa) as junction lasers during accelerated aging. *IEEE Journal of Quantum Electronics* QE-13: 351.
Paoli TL and Ripper TE (1970) Direct modulation of semiconductor lasers. *Proceedings of IEEE* 58: 1457.
Pearsall TP (2003) *Photonics Essentials: An Introduction with Experiments*. New York: McGraw-Hill.
Piprek J (2003) *Semiconductor Optoelectronic Devices: Introduction to Physics and Simulation*. London: Academic.
Quist TM, Rediker RH, Keyes RJ, Krag WE, Lax B, McWhorter AL and Zeiger HJ (1962) Semiconductor maser of GaAs. *Applied Physics Letters* 1: 91.
Reinhart FK, Hayashi I and Panish MB (1971) Mode reflectivity and waveguide properties of double-heterostructure injection. *Lasers* 42: 4466.
Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: Wiley.
Senior JM (1992) *Optical Fiber Communications*, 2nd edn. Engelwood cliffs, NJ: Prentice-Hall.
Siegman AE (1986) *Lasers*. University Science Books.
Verdeyen JT (1994) *Laser Electronics*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.
Yariv A (1989) *Quantum Electronics*, 3rd edn. New York: Wiley.
Zachos TH and Ripper JE (1969) Resonant modes of GaAs junction lasers. *IEEE Journal of Quantum Electronics* QE-5: 29.

Local Area Networks

E Wong, University of Melbourne, Melbourne, VIC, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Local area networks (LANs) are privately owned networks that are used to carry internal traffic loads within an organization. A LAN may exist on many levels categorized by its application and speed. This is illustrated in **Figure 1** whereby, at the simplest requirement, a low-speed LAN may be used to interconnect a cluster of personal computers and workstations. Such a scenario may exist in a department within a university campus, or a building within a technology park. If need be, these individual departmental or building LANs may be interconnected using a moderately high-speed backbone LAN. The backbone LAN may also provide interconnection to servers which support mass storage devices that are shared by the individual lower speed LANs. In larger organizations, a much higher speed LAN is linked to the backbone LAN to support bulk data transfer between mainframes, supercomputers, and other office equipment.

The common transmission mediums for traditional low-speed LANs are coaxial cable and twisted pairs. High-speed backbone LANs advocate the use of the optical fiber. The continuing progress in fiber-optic technology and devices, and the availability of appropriate standards have given rise to LANs that can span distances of hundreds of kilometers or more, and that operate at data rates in the order of hundreds to thousands of megabits per second (Mb/s).

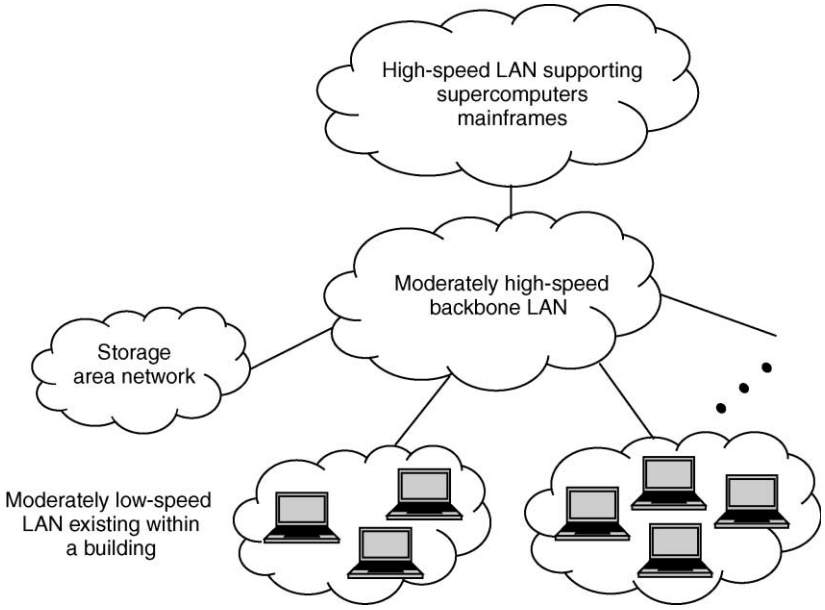


Figure 1 Interconnection of local area networks (LANs).

The increase in demand, to extend LAN capabilities towards higher data rates and across larger geographical areas, is fueled primarily by the need to provide new bandwidth intensive services such as multimedia teleconferencing, image processing, and streaming video, to a significantly higher number of end-users.

This article focuses on three existing fiber-optic LAN standards, namely Ethernet, Fiber Distributed Data Interface (FDDI), and Distributed Queue Dual Bus (DQDB). While the current uses of FDDI and DQDB are not as widespread as Ethernet, both standards have significantly contributed to the initial deployment of fiber-optics for LAN applications, with standard-based products still commercially available. In general, standards are required to ensure interoperability between components and equipment from different vendors and suppliers. Two key features are addressed by the standards: the physical topology, and the medium access control (MAC) protocol. A network's physical topology defines the way in which the users, or nodes, are interconnected to each other and to the transmission medium. Common optical LAN topologies are the star (e.g., Ethernet), bus (e.g., DQDB and Ethernet), and ring (e.g., FDDI). In comparison, the MAC protocol controls network access such that its capacity is efficiently and fairly shared amongst all nodes.

Access control can be either centralized or decentralized. In the former, a node wishing to transmit must wait until it receives permission from a controller, whereas in the latter, nodes collectively determine the transmission order. Further, capacity allocation to each node can be either synchronous or asynchronous. The synchronous approach guarantees a specific capacity to each node, e.g., using time division multiple access (TDMA), whereby nodes are allowed to transmit only during a designated time slot. In contrast, the asynchronous approach allows capacity to be dynamically allocated to each node. The asynchronous approach can be further categorized into round robin, reservation, and contention schemes. With round robin, nodes take it in turn to transmit in a cyclic manner and a node may transmit up to a maximum amount of data during its turn. With the reservation scheme, time on the medium is divided into slots, and a node wishing to transmit must reserve future slots by transmitting additional control information into the network. With the contention scheme, nodes randomly access the network. The Ethernet standard, discussed in the following section, is based on the contention scheme.

Ethernet

Overview

The IEEE (Institute of Electrical and Electronics Engineers) 802.3 standards committee first standardized 10 Mb/s Ethernet in 1985. It is now the most widely used technology for high-speed LANs, with standards extended to cater for operating speeds of 100 Mb/s, 1 Gb/s and 10 Gb/s, and encompassing the use of the optical-fiber transmission medium. Low-speed 10 Mb/s and 100 Mb/s Ethernet LANs use the

carrier-sense multiple access with collision detection (CSMA/CD) MAC protocol, whereas a switched technique is used for the 1 Gb/s and 10 Gb/s options.

In CSMA/CD, a node wishing to transmit an Ethernet frame first listens to the transmission medium by detecting the presence of signals on the channel (i.e., carrier-sense). This process allows the node to determine if another transmission is in progress. If the medium is in use, the node continues to listen until the medium is idle, and then transmits immediately. While transmitting, the node continues to listen to the medium. It may happen that two or more nodes may attempt to transmit at the same time. If this occurs, there will be a collision, and data from both transmissions will become garbled.

A collision can be indicated by a combined signal that exceeds a predefined CD threshold voltage at the carrier-sense receiver, taking into account the collision between frames from nodes that are located furthest apart. If a collision is detected during transmission (i.e., collision detection), the transmitting node ceases transmission and proceeds to transmit a brief jamming signal to notify all nodes of the collision. After transmitting the jamming signal, the node must back off before attempting to transmit the same data frame again. The back-off period is random, and is determined by a binary exponential back-off algorithm. Every time a node attempts to retransmit, the mean value of the random back-off period is doubled until the tenth attempt. After that the node is allowed to retransmit a further six times with the same mean random delay before discarding the frame and reporting an error to higher control layers.

The transmission medium remains unusable when frames collide but only until the collision is detected. As such, the minimum transmission time of a frame must be longer than the maximum propagation delay of the network so that collision is detected prior to the end of transmission. If shorter frames are used, then collision detection does not occur, and CSMA/CD thus exhibits the same performance as its less efficient predecessor protocol, CSMA. Accordingly, the CSMA/CD standards specify a minimum frame length and maximum network size, which place fundamental limitations on the practical application of a pure CSMA/CD LAN.

IEEE 802.3 10 Mb/s Ethernet

In 1993, the IEEE 802.3 committee extended the 10 Mb/s Ethernet standard to include the optical-fiber transmission medium. The standard notation of this option is 10BASE-F. Each transmission link comprises a pair of graded-index 62.5 μm core multimode optical fibers (MMF), one for each direction of transmission. The transmission wavelength is 850 nm. 10BASE-F can be further categorized into three specifications. The first, denoted 10BASE-FP (fiber passive star), specifies a broadcast star topology that can interconnect up to 33 nodes and repeaters. Each of these devices is attached to a passive star coupler via two optical-fiber links with distances up to 500 m. The passive star coupler splits the optical power of any incoming frame equally amongst its output links, allowing all nodes or repeaters in the network to receive the frame, albeit at a lower optical power level.

The second and third specifications, denoted 10BASE-FL (fiber link) and 10BASE-FB (fiber backbone) respectively, define point-to-point links over distances of up to 2 km. 10BASE-FL supports both nodes and repeaters, whereas 10BASE-FB supports only repeaters. A repeater is a device that detects the optical signals on its incoming link, performs electrical signal regeneration, and then optically retransmits the signals on its output link. Repeaters used in 10BASE-FB have an additional function in which the detected signals are retimed with a local clock before being retransmitted. This prevents the propagation of timing distortions, allowing distances of up to 15 cascaded repeater links to be implemented. The characteristics of the three specifications for 10BASE-F, along with the 100 Mb/s and 1 Gb/s specifications to be discussed in the following sections, are summarized in Table 1.

IEEE 802.3 100 Mb/s Ethernet (Fast Ethernet)

Fast Ethernet refers to the standard developed by the IEEE 802.3 committee to provide low-cost, Ethernet-compatible LANs operating at 100 Mb/s. The optical-fiber transmission medium option is denoted 100BASE-FX. Originally defined for FDDI (covered in the section on FDDI below), 100BASE-FX specifies the use of two 62.5 μm MMF links, each operating at a unidirectional rate of 100 Mb/s. While 10 Mb/s Ethernet can only operate in half-duplex mode, whereby a station can either transmit or receive a frame but cannot do both simultaneously, Fast Ethernet can operate in either half- or full-duplex modes. A node operating in full-duplex mode can simultaneously transmit and receive data frames.

In order to deploy full-duplex Fast Ethernet, all repeaters must be replaced with switching hubs. Figure 2a illustrates a repeater located in the central point in an active star topology. Each node is attached to the repeater via two optical links. The repeater retransmits frames received from any one of its input

Table 1 Summary of IEEE 802.3 physical and MAC layer specifications for 10 Mb/s, 100 Mb/s and 1 Gb/s Ethernet

	10BASE-FP	*10BASE-FL*	*10BASE-FB*	*100BASE-FX*	*1000BASE-SX*	*1000BASE-LX*
Data rate	10 Mb/s	10 Mb/s	10 Mb/s	100 Mb/s	1 Gb/s	1 Gb/s
Transmission wavelength	850 nm	850 nm	850 nm	850 nm	850 nm	1310 nm
Transmission medium	62.5 μm MMF pair	62.5 μm MMF pair	62.5 μm MMF pair	62.5 μm MMF pair	50 μm or 62.5 μm MMF pair	10 μm SMF, 50 μm or 62.5 μm MMF pair
MAC protocol	CSMA/CD (half duplex)	CSMA/CD (half duplex)	CSMA/CD (half duplex)	CSMA/CD (half duplex) Switched (full duplex)	CSMA/CD (half duplex) Switched (full duplex)	CSMA/CD (half duplex) Switched (full duplex)

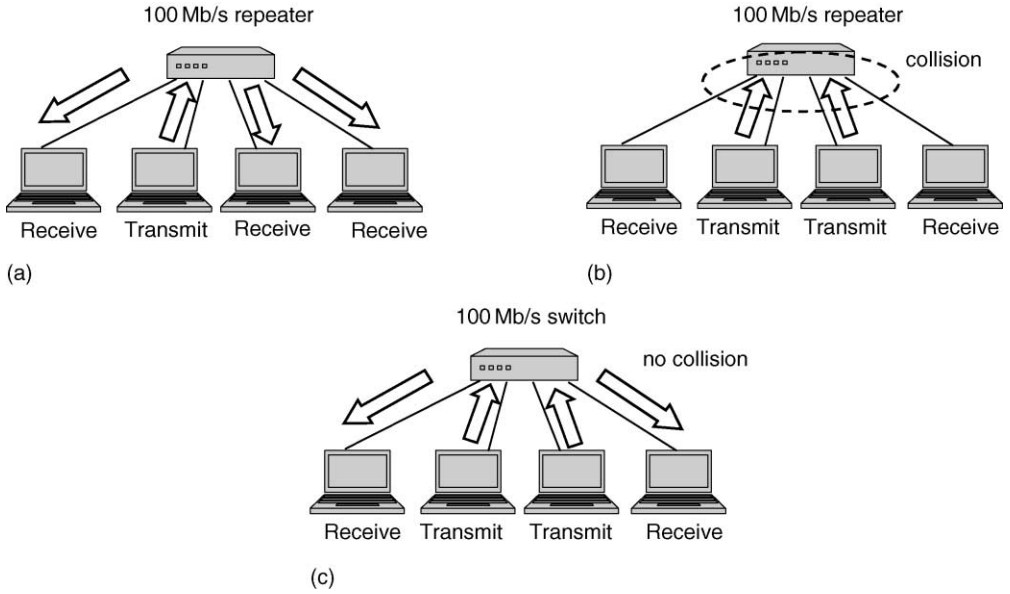


Figure 2 (a) Shared LAN repeater. (b) Collision at a shared LAN repeater due to simultaneous transmission of frames. (c) LAN switch (adapted from William S (2004) *Data and Computer Communications*, 7th edn, Ch. 15–16. Upper Saddle River, NJ: Pearson Education).

links to all of its output links. If more than one node transmits at the same time, then collisions will occur at the repeater, as illustrated in Figure 2b. A switching hub, on the other hand, switches an incoming frame from a particular input to an appropriate output line. The same switching function can be performed simultaneously on other input–output links without interference to the frames involved, as illustrated in Figure 2c. As such, each node has a dedicated capacity equal to that of the transmission rate of the LAN, and the switching hub can support mixed capacity links, e.g., 10 Mb/s and 100 Mb/s. With dedicated access to the medium and appropriate buffering at the switch to avoid contention of the same output link, there is no need to implement the CSMA/CD access algorithm. Nonetheless, the conventional Ethernet frame format is used to ensure inter-operability between existing 10 Mb/s and 100 Mb/s LANs.

IEEE 802.3 1 Gb/s Ethernet (Gigabit Ethernet)

The IEEE 802.3 1 Gb/s Ethernet standard, commonly referred to as Gigabit Ethernet, was finalized in 1998. An example of a Gigabit Ethernet backbone LAN is shown in Figure 3, whereby a Gigabit Ethernet backbone switch connects central servers to high-speed workgroup switches. In turn, each workgroup switch supports gigabit high-performance workgroup servers, as well as 100 Mb/s workstations and hubs. As in the 100 Mb/s standard, Gigabit Ethernet uses the conventional Ethernet frame format and allows both half-duplex and full-duplex operation. For half-duplex operation, the CSMA/CD access protocol is

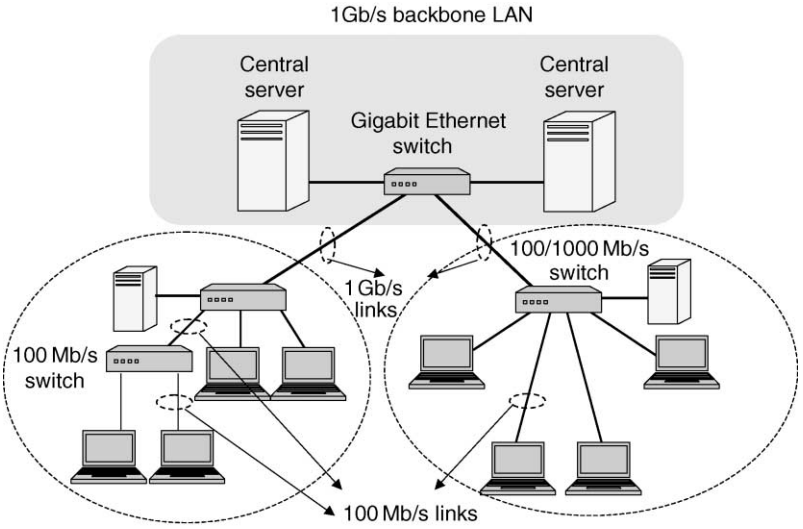


Figure 3 Gigabit Ethernet backbone network (adapted from William S (2004) *Data and Computer Communications*, 7th edn, Ch. 15–16. Upper Saddle River, NJ: Pearson Education).

equipped with two additional features. The first feature, known as carrier extension, extends the propagation time of short frames by adding nondata symbols to the end of frames. The minimum length of a frame for Gigabit Ethernet is 512 bytes as opposed to the specified minimum of 64 bytes in the preceding two standards. The second feature, known as frame bursting, allows a node to transmit multiple consecutive frames, up to a specified maximum, without having to undergo the contention process for each transmission. Both features improve the efficiency of the protocol.

There are two optical-fiber transmission medium specifications for Gigabit Ethernet. The short wavelength option, denoted 1000BASE-SX, is for 850 nm transmission. Specified maximum link distances are 275 m using 62.5 μm MMF, and 550 μm using 50 μm MMF. The long wavelength option, denoted 1000BASE-LX, is for 1310 nm transmission. Achievable duplex link distances are 550 m using 62.5 μm or 50 μm MMF, and 5 km using 10 μm SMF.

IEEE 802.3 10 Gb/s Ethernet (10GbE)

The Ethernet standard was extended to incorporate 10 Gb/s (10GbE) in 2002, fueled by ever-increasing Internet traffic and bandwidth intensive applications. 10GbE fulfills not only LAN specifications, but also metropolitan and wide-area network specifications. It is envisioned that with the emergence of 10GbE metropolitan and wide-area networks, the deployment of 10GbE LANs will enable a seamless transport of Ethernet frames across different network boundaries, thus simplifying network management and optimizing operational costs.

Unlike its predecessor standards, the physical layer options for 10GbE include only the optical-fiber transmission medium and allow only for full-duplex operation. Four specifications are defined for 10GbE. 10GBASE-S is designed for 850 nm transmission on 63.5 μm MMF, while 10BASE-L and 10BASE-LE use 10 μm single mode fiber (SMF) for 1310 nm transmission and 1550 nm transmission, respectively. The corresponding maximum achievable distances are 300 m, 10 km, and 40 km, respectively. An alternative specification, 10GBASE-LX4, is designed for wavelength division multiplexed (WDM) transmission links. The specification caters for four 2.5 Gb/s transmission channels within the 1270–1355 wavelength region on either 10 μm SMF, 50 μm MMF, or 62.5 μm MMF. Single-mode transmissions can reach distances of up to 10 km, while multimode transmissions can reach distances up to 300 m. The characteristics of these physical layer specifications are summarized in Table 2.

Fiber Distributed Data Interface (FDDI)

Overview

The Fiber Distributed Data Interface (FDDI) standard was developed by the Accredited Standards Committee X3T9.5, and approved by both ANSI (American National Standards Institute) and ISO (International Standards Organization) in 1989. While not as

Table 2 Summary of IEEE 802.3 physical and MAC layer specifications for 10 Gb/s Ethernet

	10GBASE-S	*10GBASE-L*	*10GBASE-E*	*10GBASE-LX4*
Data rate	10 Gb/s	10 Gb/s	10 Gb/s	4 × 10 Gb/s
Transmission wavelength	850 nm	1310 nm	1550 nm	1270 nm to 1355 nm
Transmission medium	62.5 μm MMF pair	10 μm SMF pair	10 μm SMF pair	10 μm SMF or 50 μm, 62.5 μm MMF pair
MAC protocol	Switched (full duplex)	Switched (full duplex)	Switched (full duplex)	Switched (full duplex)

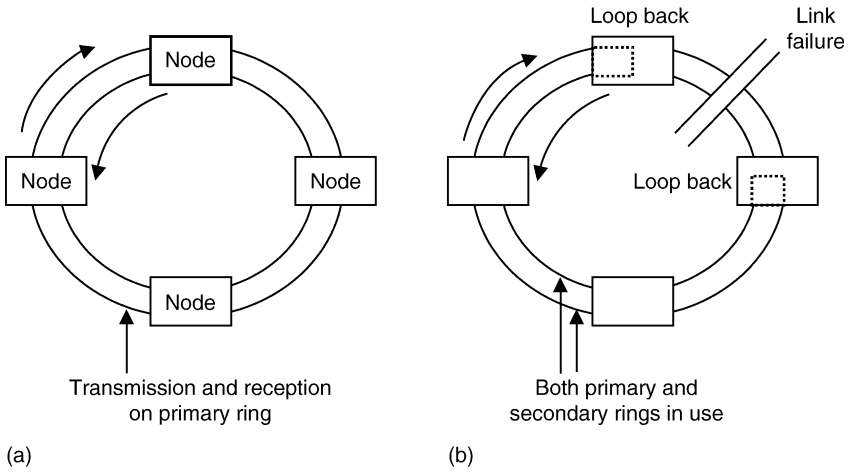


Figure 4 (a) FDDI ring network. (b) Use of secondary ring for protection against link failures.

widely deployed as the Ethernet family of LANs, due to the high cost of equipment and devices, it remains as the first to be standardized for high-speed fiber-optic LAN applications and is now mainly used to interconnect existing FDDI-compatible LANs. Extensions have been made to the FDDI standard to incorporate single-mode fiber operation for increased signal reach.

FDDI specifies a dual counter-propagating ring network as shown in Figure 4a. Each ring operates at the speed of 100 Mb/s, and can reach a maximum length of 100 km. The dual ring topology is self-healing against fiber-cuts and node failures. Under normal operation, frame transmission and reception are performed on a primary ring. The secondary ring is utilized to form a closed loop when a node or link failure occurs, as illustrated in Figure 4b. In FDDI, the interface between each node and the optical-fiber transmission medium is a repeater. Data frames, each of which carries its source and a destination address in a defined field in the frame header, are transmitted sequentially around the ring from one repeater to another. Each frame can vary in length by up to 4550 bytes. Each repeater copies the destination address field of all incoming frames and, upon recognizing the destination address as its own, additionally copies the remainder of the frame. The repeater attached to the source node also facilitates the removal of frames it has transmitted after a round-trip propagation around the ring.

FDDI employs a distributed timer-controlled, token-passing mechanism, referred to as timed token rotation protocol, to control access to the shared optical fiber ring. While Ethernet-based LANs are highly suited to support throughput sensitive services which can sustain reasonable delay in return for high throughput, FDDI provides an additional capability to support delay sensitive services such as packetised voice and video. For such services, frames must be received within a certain time interval for quality of service requirements. A node connected to an FDDI network can therefore transmit both classes of services: the former referred to as burst traffic and the latter as stream traffic. In addition, the timed token rotation protocol can limit the amount of frames that can be transmitted by a particular node for each class of service, ensuring efficient and fair network access amongst all nodes and services. FDDI is thus well-suited to carry the mixture of burst and stream traffic expected in a backbone network which supports a number of smaller LANs.

MAC Protocol

Token-passing mechanism

In FDDI, access to the shared transmission medium is governed by a timed token rotation protocol. A token is a special frame created during network (re)-initialization by a pre-assigned node. The token circulates the ring network and gives the node that holds it the right to transmit. Figure 5 gives an example of a single frame transmission from Node A to Node C, and from Node B to Node A. Node A has a frame to transmit to Node C and thus waits for a token (Figure 5a). Upon recognizing that a token frame is passing through, it aborts the token transmission without regenerating the token bits. After Node A has completely received the token, the node begins transmitting frame FA addressed to Node C, and appending a new token upon completion (Figure 5b). Meanwhile, Node B regenerates and retransmits FA as it circulates past. When FA reaches Node C, the node copies FA in addition to regenerating and retransmitting FA. Node B captures the token issued by A and transmits FB to Node A (Figure 5c). Eventually FA recirculates back to its originating node, Node A, whereby it is removed from the ring (Figure 5d). At the same time, Node B finishes transmitting FB and releases the token. Nodes C and D, with no frames to transmit or receive, regenerate and retransmit both FB and the token. When FB reaches Node A, it is copied by the node. Node A also regenerates and retransmits both FB and the token. FB is finally removed from the ring by Node B (Figure 5e), and the token remains in the ring to be received by the next node with a frame to transmit (Figure 5f).

Aside from source and destination address fields, each dataframe also contains a frame status field which comprises error detected (E), address recognized (A), and frame copied (C) indicator bits. Each node can check passing frames for errors and can set the E indicator if an error is detected using the standard cyclic redundancy check procedure. If a node detects its own address, it sets the A indicator, and upon copying the frame, the station may also set the C indicator. When a dataframe recirculates back to its source node, the indicator bits are examined. If both A and C are not set, then the destination node does not exist or is inactive. On the other hand, if the frame has not been copied by an active destination node or an error has been detected, higher control layers are notified for future retransmission.

Time-controlled mechanism

FDDI uses a time-controlled mechanism to prevent nodes from dominating the network capacity. During initialization, a target token rotation time (TTRT) is established based on the shortest allowable time between token arrivals at each node. The same TTRT value is then stored at all nodes. In addition, each node has a maximum capacity guaranteed for the transmission of its stream traffic. The total transmission time of this stream traffic and the latency of the ring (i.e., total round-trip propagation time, the transmission time of a maximum length frame and the token frame) must never exceed the TTRT. During network operation, each node measures the

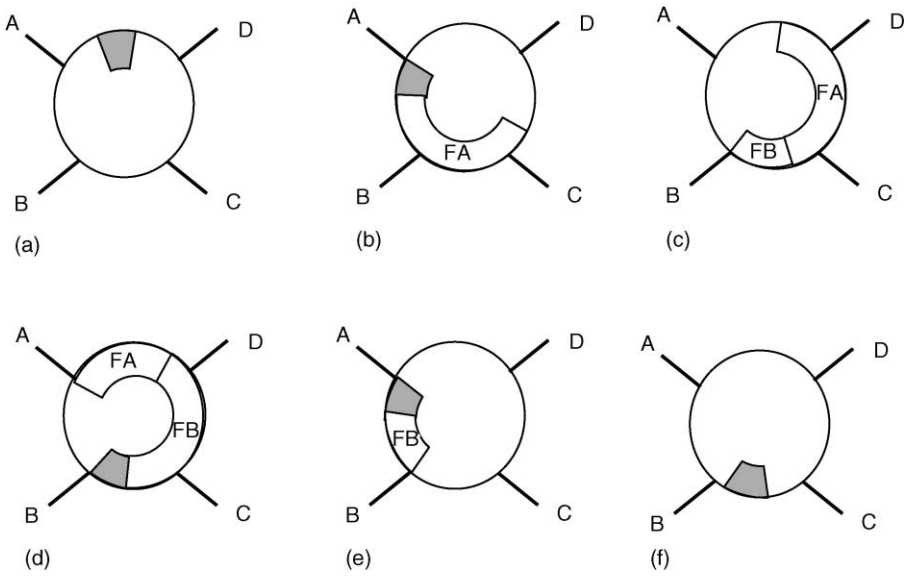


Figure 5 Fiber Distributed Data Interface (FDDI) token passing mechanism.

time interval between successive token arrivals, using a token-rotation timer (TRT). The TRT is enabled only upon a token arrival, and expires if the counter reaches the TTRT value. In turn, a late counter (LC) is incremented and TRT is reset to zero. The number of TRT expirations since a token was last received is, therefore, counted by LC. Any token that arrives at a node with a nonzero LC value is referred to as a late token. Likewise, an early token is one which arrives at a node with a zero LC value.

The action of a node when it receives a token depends on whether the token is early or late. When a late token is received, LC is reset to zero but TRT continues to run. The node then transmits for the period of time which corresponds to the predetermined bandwidth for steam traffic transmission. The node is not allowed to transmit any bursty traffic frames. In the event of an early token reception, the node transfers the current value of TRT to a second timer, the token-holding timer (THT) but does not enable it. TRT is reset to zero and is enabled to track the next rotation of the counter. The node then transmits stream traffic frames for the allocated period of time. After that, THT is enabled and bursty traffic frames are transmitted until THT reaches the TTRT value. The THT is then reset and the node releases the token into the ring. The transmission of bursty frames may be further controlled by two additional mechanisms. The restricted-token mechanism allows only two nodes in the network to use all the available bursty bandwidth, and restricts all other nodes to their stream bandwidth. This is facilitated by the use of a restricted token which is issued by one of the two nodes. The priority mechanism provides each station with the capability to transmit bursty frames according to eight priority levels. Each priority level corresponds to a different time threshold in the THT.

Note that the time-controlled mechanism of the MAC protocol guarantees a maximum response time for the ring because, in the worst case, the time between the arrival of two successive tokens will never exceed twice the value of TTRT. Most multimedia applications, such as video conferencing, real-time video, and entertainment video can be supported on FDDI with stream service and a small TTRT. However, since the TTRT cannot be less than the ring latency, applications requiring time bounds less than twice the ring latency cannot be supported by FDDI. Long and continuous isochronous services which must adhere to strict periodic access require FDDI-II.

FDDI-II

FDDI-II provides the additional circuit-switching capability to support constant bit-rate telecommunication applications. Like FDDI, the transmission bit-rate of FDDI-II is 100 Mb/s. FDDI-II can operate in either complete isochronous mode, complete packet-switching mode, or in hybrid mode in which both isochronous as well as bursty and stream services are supported. Strict periodic access required by isochronous services are accomplished with transmission opportunities repeated every 125 μs. At 100 Mb/s, time on the ring is divided into several frames, each 1562.5 bytes long. Out of the 1562.5 bytes, 1560 bytes are used for data transmission while the remaining 2.5 bytes are used as interframe gap. In turn, the 1560 bytes are divided into 16 wideband channels of 96 bytes each. Each wideband channel thus provides a bandwidth of 6.144 Mb/s, sufficiently adequate for television broadcast and multiple voice channels. For operation in the hybrid mode, some of the wideband channels are allocated for packet mode transmission. Likewise, if complete packet-switching operation is required, all wideband channels are allocated for packet switching transmission.

IEEE 802.6 Distributed Queue Dual Bus (DQDB)

Overview

The IEEE 802.6 Distributed Queue Dual Bus (DQDB) standard, finalized in 1990, specifies a dual bus network with an operating speed of 150 Mb/s. DQDB is now mainly used for MAN applications, although it can be readily deployed for LAN backbone interconnection. Figure 6 illustrates the basic architecture of a DQDB network with access nodes, Node 1 to Node N, distributed along two counter propagating buses, BUS A and BUS B. Each access node is considered to be connected to a lower speed LAN. At the head end of each bus, a slot generator continuously generates fixed-size time slots which propagate along its respective bus. The contents of each time slot are regenerated at each access node it passes, and upon reaching the end of the bus, a sink terminal receives and removes the time slots from the network. Data are transported across the network from one access node to another in the form of fixed-size segments. A node transmits a segment by changing the contents of a passing time slot. Each segment needs only be transmitted on one of the two buses, the choice depending on the locations of the transmitting and receiving nodes. For example, referring to Figure 6, a downstream transmission from Node i to Node $i+1$ requires the use of BUS A. On the contrary, an upstream transmission from Node i to Node $i-1$ uses BUS B.

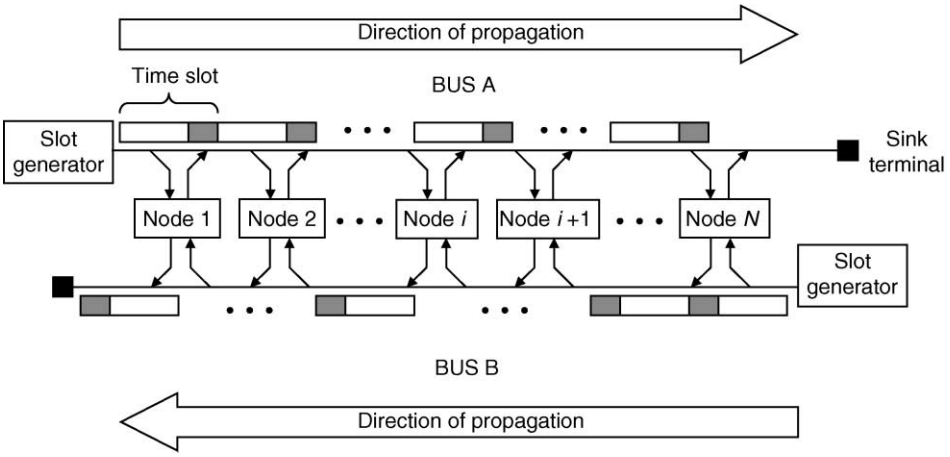


Figure 6 Distributed Queue Dual Bus (DQDB) architecture.

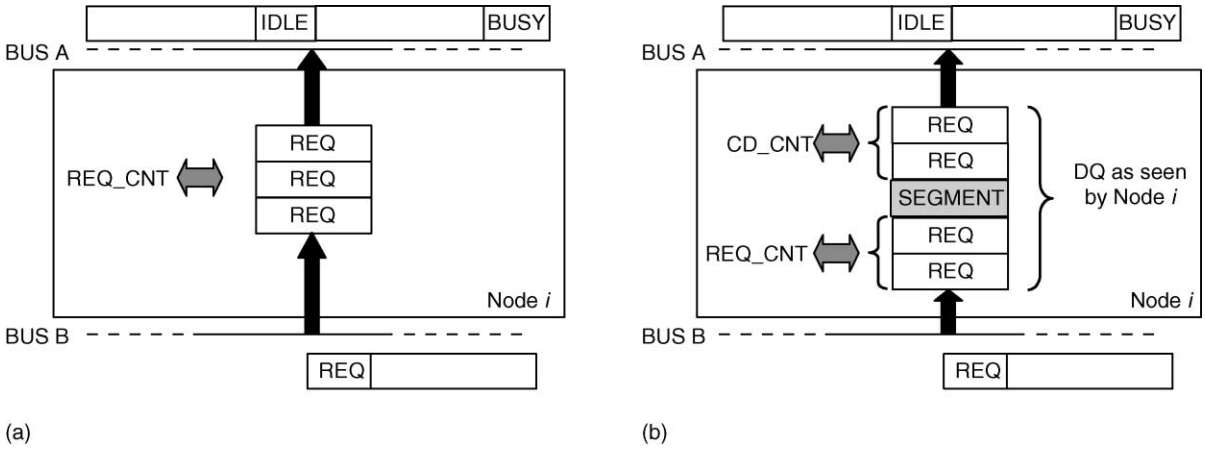


Figure 7 Behavior of Node *i* during (a) idle state; and during (b) countdown state.

MAC Protocol

In DQDB, permission to transmit segments into the network is determined by a global distributed queueing algorithm. The algorithm allows each node to continuously monitor both buses, counting the requests and available time slots that pass by. In addition to maintaining a global first-in-first-out queue (DQ) for each bus, the algorithm relies on information in the header of time slots, namely the BUSY and REQ fields, and counters at the access nodes to coordinate transmissions. A 'BUSY = 0' field represents the availability of a time slot for the transport of a segment. For example, time slots generated at the head end of each bus have default 'BUSY = 0' and 'REQ = 0' fields. The BUSY field is set to 1 when a segment is transmitted into an available time slot.

The REQ field is used in conjunction with the counters to inform nodes of an impending transmission from downstream. To simplify the explanation, consider only packets propagating on BUS A. A node wishing to transmit a segment downstream needs to queue the segment in the appropriate DQ. The node with an impending transmission must then notify all upstream nodes that it has inserted a segment in the DQ so that an available time slot can be reserved for the transmission. This is accomplished by setting the first time slot propagating in the opposite bus, BUS B, with REQ = 0 to REQ = 1. As such, the node maintains the outstanding segments in the DQ belonging to downstream nodes via the value in the REQ field.

To illustrate, Figure 7a shows an arbitrary node, Node *i*, in an idle state, whereby the node has no segments to transmit downstream on BUS A. With every 'REQ = 1' observed on BUS B, a request counter, REQ_CNT is incremented by one. However, with every empty slot, i.e., 'BUSY = 0', observed on

Table 3 Comparison of standardized local area networks

	Ethernet	*Fast Ethernet*	*Gigabit Ethernet*	*10GbE*	*FDDI*	*DQDB*
Data rate	10 Mb/s	100 Mb/s	1 Gb/s	10 Gb/s	100 Mb/s	150 Mb/s
MAC protocol	CSMA/CD	CSMA/CD or Switched technique	CSMA/CD or Switched technique	Switched technique	Timed token rotation	Distributed queue
Standard	IEEE 802.3	IEEE 802.3	IEEE 802.3	IEEE 802.3	ANSI-FDDI	IEEE 802.6

the BUS A, REQ_CNT is decremented by one. When Node *i* has a segment to transmit, it enters the countdown state as illustrated in Figure 7b. The value in REQ_CNT is transferred into an independent countdown counter, CD_CNT, and REQ_CNT is reset to zero. Node *i* also sets 'REQ = 1' in the opposite bus, BUS B. As in the idle state, each 'REQ = 1' observed on BUS B increments REQ_CNT while each 'BUSY = 1' observed on BUS A decrements it. When CD_CNT reaches zero, Node *i* transmits its segment in the first time slot with 'BUSY = 0' on BUS A and sets 'BUSY = 1'. The node returns to the idle state if there are no other segments to transmit. The algorithm allows the node to determine the number of time slots reserved by downstream nodes and the time to transmit a segment into an available time slot. Note that both the idle and countdown states are replicated for transmissions on BUS B.

The DQDB standard also specifies a limited support of isochronous services. Nodes which provide and receive such services are reserved a certain amount of time slots which are indicated with additional header information. The remaining time slots are used by all other access nodes for the transport of packet-switched traffic, and access to these time slots is managed by the global distributed queueing algorithm. A priority mechanism based on using multiple DQs, each with a different priority level, may also be implemented to guarantee access to certain segments in the network.

Summary

Due to the explosive demand for high-bandwidth applications, fiber-optics has become an essential technology in LANs. Table 3 compares the various fiber-optic LAN standards discussed in this article. The Ethernet family remains the most prevalent of LAN technologies. Its widespread use is contributed by the availability of low-cost chipsets, mature and familiar management and analysis tools. The future of fiber-optic LANs operating beyond 10 Gb/s will depend upon WDM solutions. The current usage of one transmission channel per network is impeded primarily by the limitation of the speed of electro-optic conversion, and thus does not fully exploit the massive transmission bandwidth of the optical fiber. With WDM, multiple transmission channels are concurrently transmitted on a single optical fiber, thereby greatly increasing the capacity of an optical LAN. Ongoing research efforts have been directly towards advancing WDM component technology and addressing prevailing issues such as polarization control, chromatic dispersion, and four-wave mixing among others.

List of Units and Nomenclature

CD_CNT	countdown counter
CSMA/CD	carrier-sense multiple access with collision detection
DQ	global first-in-first-out queue
DQDB	distributed queue dual bus
FDDI	fiber distributed data interface
LAN	local area network
LC	late counter
MAC	medium access control
MAN	metropolitan area network
MMF	multimode fiber
REQ_CNT	request counter
SMF	single mode fiber
TDMA	time division multiple access
THT	token-holding timer
TRT	token rotation timer
TTRT	target token rotation timer
WDM	wavelength division multiplexing

See also

Information Processing: Optical Neural Networks.

Further Reading

IEEE Std 802.6 (1990) IEEE standards for local and metropolitan area networks: distributed queue dual bus (DQDB) subnetwork of a metropolitan area network (MAN).

IEEE Std 802.3 (2002, Revision of IEEE Std 802.3, 2000 ed.) IEEE Standard for Information Technology – Telecommunications and Information Exchange Between Systems – Local and Metropolitan Area

Networks – Specific Requirements Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications.
ISO 9314-1,2 (1989) Information processing systems – Fibre Distributed Data Interface (FDDI) – Part 1: Token Ring Physical Layer Protocol (PHY) and Part 2: Token Ring Media Access Control (MAC).
Marco C, Enrico G and Luciano L (1997) *Metropolitan Area Networks*. New York: Springer-Verlag.
Mills A (1995) *Understanding FDDI*. Englewood Cliffs, NJ: Prentice Hall.
William S (2004) *Data and Computer Communications*, 7th edn, Ch. 15–16. Upper Saddle River, NJ: Pearson Education.

Optical Time Division Multiplexing

L P Barry, Dublin City University, Dublin, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

As the demand for bandwidth continues to grow, driven by the massive increase in Internet usage, so will the necessity to have communications networks that can handle very high data rates. The use of optical fiber networks is the clear choice for such systems given the huge available bandwidth of the fiber transmission medium. However, in a basic optical communication system comprising a laser transmitter, an optical fiber transmission medium, and a receiver, the capacity is essentially limited by the speed at which light can be modulated at the transmitter. To overcome this limitation, which is basically due to the speed of available electronics, it is necessary to use optical multiplexing techniques, and one such technique, known as optical time division multiplexing (OTDM), will be the subject of this article.

Principles of Time Division Multiplexing

Time division multiplexing (TDM) has long been the traditional method for electrically combining information channels. If we take the most fundamental data rate to be that of a simple voice call at 64 kbit/s, then the transmission of data at higher bit rates is achieved by electrically multiplexing a large number of 64 kbit/s channels in the time domain. With the evolution of standards, a number of different data rates have been specified as standard transmission rates. In Europe the synchronous digital hierarchy (SDH) has a basic data rate of 155.52 Mbit/s, which is known as the synchronous transport module – Level 1 (STM-1). This particular data rate is essentially obtained by electrically multiplexing over 2000 voice calls in the time domain (with some of the capacity required for overhead information). By subsequently multiplexing a number of STM-1 channels together we can obtain transmission at the higher standard data rates of STM-4 (622 Mbit/s), STM-16 (2.48 Gbit/s), and STM-64 (9.88 Gbit/s), as outlined in Table 1.

Figure 1 illustrates how basic electrical TDM is used in standard optical communication systems. Clearly as we approach the higher data rates of STM-64, a serious level of electrical multiplexing is required, and as the data rates increase so does the cost and complexity of the electrical equipment at the transmitter and receiver. Indeed the present state of the art in electronics seems to suggest 40 Gbit/s as the limit for electrically multiplexed communication systems. However, as we stated earlier, communications traffic has been growing explosively over the last decade and will continue to do so. In order to meet this demand for capacity, and better exploit the massive available bandwidth of optical fiber, it is necessary to use optical multiplexing techniques for communications systems. The two main optical multiplexing techniques available are wavelength division multiplexing (WDM) and optical time division multiplexing (OTDM). WDM essentially involves transmitting data at a number of different wavelengths on the same fiber. Although electrical multiplexing may be limited to data transmission rates of around 40 Gbit/s using one laser, by multiplexing together N different wavelength channels each carrying 40 Gbit/s, we can achieve overall data rates up to and beyond a terabit/s.

The second of these optical multiplexing techniques, OTDM, is the subject of this article. Whereas WDM multiplexes optical data channels in the

Table 1 Standard data rates for SDH transmission systems

SDH standard	*Data rate (Mbit/s)*
STM-1	155.52
STM-4	622.08
STM-16	2488
STM-64	9953
STM-256	39 813

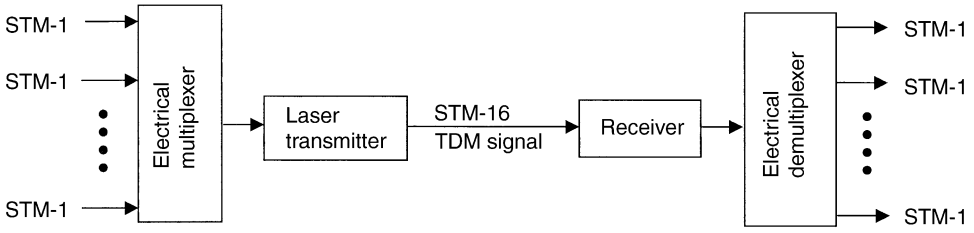


Figure 1 Electrical time division multiplexing (ETDM) in a basic optical communication system.

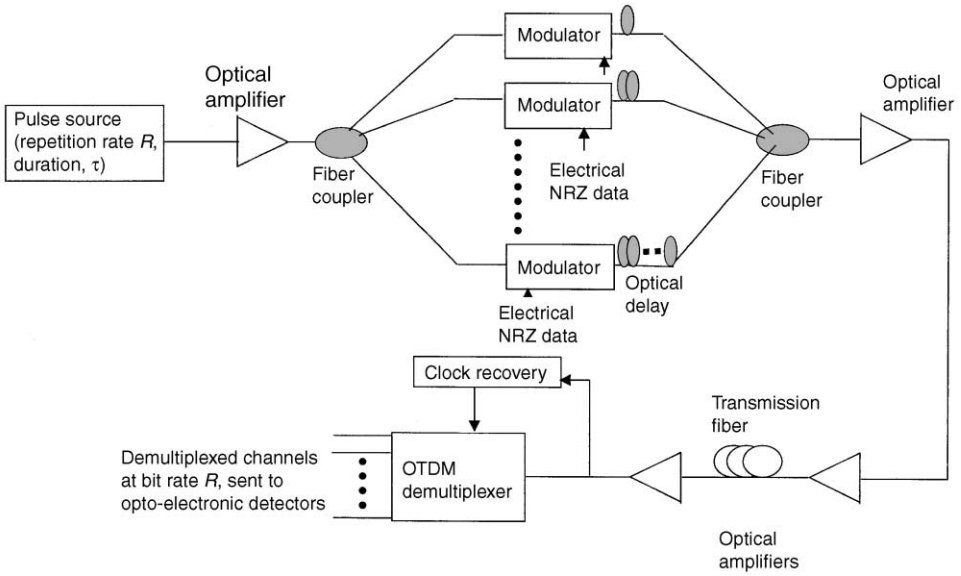


Figure 2 Basic configuration for an OTDM transmission system comprising transmitter, transmission fiber, and receiver.

wavelength domain, the basic principle of OTDM communications is to increase the system capacity by multiplexing optical data channels in the time domain. This multiplexing is usually achieved by a process known as bit-interleaving. This process can be explained by considering Figure 2, which shows a basic schematic of an OTDM based transmission system. The main component of the overall system is a source of ultrashort optical pulses (pulse duration τ) at a certain repetition rate, R. The optical pulse source is initially split into N channels using a passive fiber coupler, and each pulse train is subsequently modulated by electrical data which is at a data rate of R. The resulting output from each modulator is essentially an optical data channel where the data are represented using ultrashort optical pulses (return-to-zero data format). The data from each modulator then passes through a fixed fiber delay line which delays each channel by a time equal to $1/RN$ relative to its adjacent channel (as shown in Figure 2). The N modulated and delayed optical data channels are then recombined in another passive fiber coupler to form the OTDM data signal. We can consider that the delay lines essentially assign each data channel to a specific bit slot (of width $1/RN$) in the overall multiplexed signal. The multiplexed data signal may then be transmitted over optical fiber before arriving at the receiver which is responsible for demultiplexing the optical signal into its discrete channels.

The duration of the optical pulse source is extremely important in determining the maximum overall data rate which can be achieved. The overall data rate is basically defined by the temporal separation between channels in the multiplexor, but in order to avoid cross-channel interference, this separation must be significantly greater than the pulse duration. Thus to increase the overall data rate we must use shorter optical pulses. However, as we reduce the optical pulse width to raise the data rate we need to take into account the problems that may be encountered as this high-speed data signal propagates over optical fiber, and also the difficulty in demultiplexing a high-speed OTDM data signal. The following section will look in greater detail at

these key elements of the OTDM communication system, namely the optical pulse source, the transmission of the ultrahigh capacity signal, and the demultiplexing at the receiver.

Key Elements of an OTDM Communication System

Ultra-Short Optical Pulse Sources

As stated earlier, the optical pulse source is a key element in any OTDM-based communication system. The important characteristics of the pulse source that will affect its usefulness in an OTDM system are the pulsewidth, the spectral width, and the temporal jitter. The pulse duration clearly has to be short enough to support the desired overall transmission rate. For example, if we wish to design an OTDM system with an aggregate data rate of 100 Gbit/s (which will have a delay of 10 ps between each channel in the multiplexed signal), then the pulsewidth should normally be less the 30% of the channel spacing to avoid cross-channel interference (i.e. around 3 ps). For terabit/s OTDM systems we would require pulsewidths less than 0.5 ps. The spectral width of the pulse source is also important, as it will have a major impact on how the pulse will evolve during propagation in the fiber. A standard figure of merit which is employed is the time–bandwidth product, and ideally we require the pulse source to be transform limited, which implies that the spectral width is as small as possible for the associated pulsewidth. The impact of temporal jitter on the pulses can be easily understood by considering the above example of a 100 Gbit/s system, employing 3 ps pulses, with the multiplexed channels spaced by 10 ps. Obviously if the jitter on the pulses becomes significant in relation to the channel spacing, then this can also lead to interference between adjacent channels in the overall OTDM system. A large number of techniques have been employed to develop ultrashort pulse sources suitable for use in OTDM systems, but the three main methods which will be described below are active mode-locking, gain-switching, and external modulation of a CW light signal.

Active mode-locking of laser diodes normally involves modulating the amplitude of the optical field inside the laser cavity at a frequency which is equal to the mode spacing of the laser. This can be achieved by applying an electrical sinusoidal signal at the correct frequency, and results in the generation of optical pulses at the repetition rate of the applied signal. This technique has been successfully employed in the generation of subpicosecond pulses at repetition rates up to and beyond 40 GHz, with excellent spectral and temporal jitter characteristics. However, an inherent problem in all mode-locked pulse sources is the difficulty in synchronizing the mode-locked frequency to a specific SDH standard data rate.

Gain switching of semiconductor laser diodes is probably the simplest and most reliable technique to generate optical pulses. The technique, which is presented in Figure 3, involves applying a high-power electrical pulse (or electrical sinusoidal signal) to the laser in conjunction with a certain bias current. By ensuring that the electrical pulse signal and the bias signal have the correct level, the relaxation oscillation phenomenon of the laser results in the production of optical pulses with durations between 10 and 30 ps, at the repetition rate of the applied electrical signal. The frequency of the electrical signal applied to the laser is essentially arbitrary (provided it is not larger than the modulation bandwidth of the diode), thus making it straightforward to synchronize the optical pulse train to a SDH line-rate. The main problem with this technique is that the spectral width of the pulses generated is such that the pulses are far from transform-limited, which would affect their subsequent propagation in the fiber. In addition, temporal jitter on gain-switched pulses can also be a problem. However, by using novel arrangements such as external injection into the gain-switched laser, this difficulty can be overcome.

The third pulse generation technique mentioned above involves external modulation of a cw light signal with an electro-absorption modulator. The experimental configuration for this pulse generation technique is shown in Figure 4. By biasing the modulator around its null point, and applying an electrical sinusoidal signal to it, the cw light passing through the modulator becomes shaped into optical pulses. The optical pulse train which is generated due

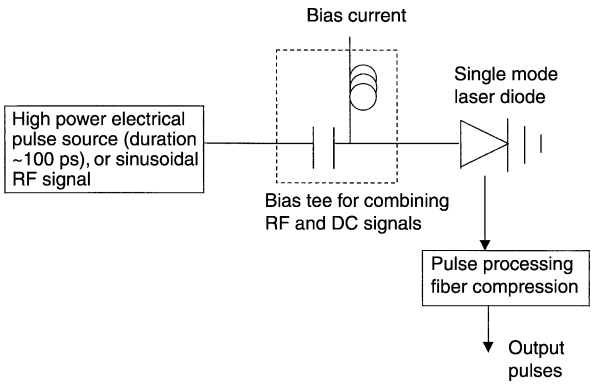


Figure 3 Pulse generation using the gain-switching technique followed by pulse compression.

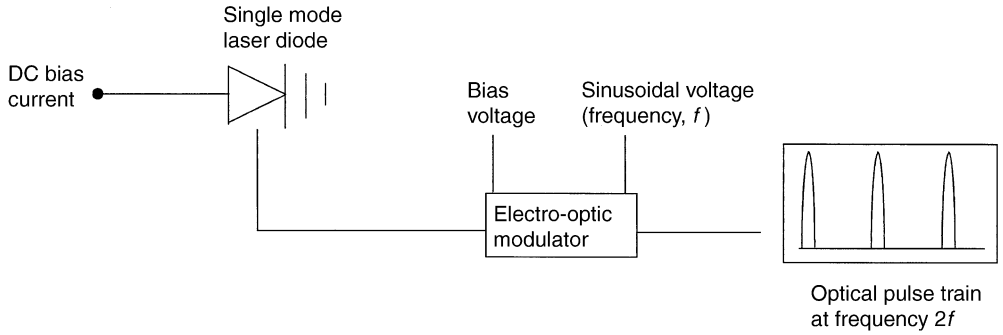


Figure 4 Pulse generation based on shaping of cw light using the nonlinear response of an external modulator biased about its null point.

the nonlinear response of the electro-absorption modulator is at a repetition rate of twice the applied electrical signal. The pulses generated are normally transform limited with extremely low temporal jitter, and the repetition rate is arbitrary as in the gain-switching technique (with the limit being ultimately determined by the modulation bandwidth of the modulator). This method may be readily used to generate pulses at repetition rates up to 40 GHz with pulsewidth around 5 ps, and it has the advantage that the optical source and the modulator can be integrated in a single device to form a compact pulse source suitable for OTDM communication systems.

It should also be noted that in addition to the various techniques that have been outlined, it is possible to use pulse compression in order to reduce the pulse width. The main issues concerned with pulse compression are the shape and spectral width of the optical pulses after compression. One of the most attractive methods of achieving pulse compression involves using nonlinear compression in dispersion-decreasing fiber, with the main advantage of this technique being the ability to maintain a transform limited pulse after compression. By employing this compression scheme, optical pulse sources at 10 GHz with pulsewidths below 200 fs have been developed, and such pulsewidths would be suitable for use in Tbit/s OTDM systems.

Transmission of an OTDM Signal over Fiber

The transmission performance of an OTDM data signal is vital in determining the distance over which the data can be transmitted successfully. The main fiber parameters, which will affect the signal transmission, are attenuation and dispersion. If we consider OTDM systems operating at a wavelength of around 1550 nm (minimum loss wavelength), to overcome the fiber attenuation and maintain a suitable optical power budget for the system, optical amplifiers are normally employed. If we thus assume that the amplifiers overcome the fiber loss problem, then the maximum transmission distance will be limited by the fiber dispersion. In a very basic OTDM system operating at 1550 nm, with transmitter and receiver linked using standard fiber (dispersion parameter of about 16 ps km^{-1} nm^{-1}), the maximum transmission distance will be limited by the overall data rate, and the pulsewidth and spectral width of the optical pulse source. For example, consider a 40 Gbit/s OTDM data signal which is formed using 8 ps optical pulses with a spectral width of 40 GHz (0.32 nm). From the spectral width we can calculate the signal broadening due to dispersion to be around 5 ps km^{-1}, and as the pulses broaden and spread into adjacent channels of the OTDM signal then this interference will make it increasingly difficult to correctly detect the signal at the receiver. In this case after propagation through 5 km of fiber, the signal pulses will have already dispersed to around 25 ps duration, the same value as the temporal bit slot into which each channel is placed. This will clearly result in serious loss of signal integrity. A straightforward possibility to increase the transmission distance of OTDM systems is to employ dispersion shifted fiber; the dispersion parameter is now around 1–2 ps km^{-1} nm^{-1}, at an operating wavelength of 1550 nm. This reduction in dispersion will obviously increase the allowed transmission distance by about an order of magnitude for the example described above. However, to develop ultrahigh-speed, long-haul OTDM communications we require more complex transmission schemes. Two possibilities for this include dispersion compensation, and soliton transmission techniques.

Dispersion compensation basically involves compensating for the dispersion that has been encountered during transmission by using some fiber with a total dispersion of opposite sign but equal magnitude to the transmission fiber. Dispersion compensation

may also be achieved using a suitably designed fiber grating. For a long-haul OTDM system, a dispersion compensator may be used every 50 or 60 km, in conjunction with the optical amplifier, thus allowing us to compensate both fiber loss and attenuation periodically along the link. The main limitation, however, with the dispersion compensation technique is caused by the dispersion slope of the fiber, as for ideal compensation it is necessary to compensate completely for the dispersion slope in addition to the overall fiber dispersion. The second technique that may be employed to greatly extend the transmission distance of high-speed OTDM communication systems is the use of soliton transmission. The basic principle of soliton transmission is to use optical data pulses with a particular shape, pulsewidth, and peak power, such that as the pulse propagates, the effects of fiber dispersion and nonlinearity counterbalance to allow the signal to propagate undistorted. By using optical amplifiers to ensure that the optical pulse peak power does not vary too much along the transmission link, OTDM transmission at data rates greater than 100 Gbit/s, over distances approaching 500 km have been demonstrated.

Demultiplexing of OTDM Signal at Receiver

For OTDM communication systems with data rates above 40 Gbit/s, it is not feasible to use electrical switching. Indeed for ultrahigh-speed systems operating at 100 Gbit/s and beyond, the only solution for demultiplexing is to use all-optical switching in which optical control signals are used to switch the OTDM data signal. All-optical switches normally use nonlinear optical effects either in optical fiber or semiconductors. A typical scenario (as presented in Figure 5) involves the injection of both the OTDM data signal and an optical control signal into the nonlinear device. The control signal consists of high-power ultrashort pulses at the repetition rate of the individual channels within the temporally multiplexed signal, and by synchronizing it with one of the OTDM channels it is possible to demultiplex this channel from the high-speed OTDM signal.

Two of the most popular all-optical demultiplexers are the nonlinear optical loop mirror (NOLM), and the terahertz optical asymmetric demultiplexer (TOAD). The NOLM is based on the nonlinear refractive index of optical fiber. It essentially consists of a 2×2 fiber coupler with its two outputs joined using a certain length of fiber. When an OTDM signal is injected into an input port of the coupler, it splits into two counterpropagating components in the fiber loop, and when these components recombine and interfere at the coupler, the overall signal is output through its initial input port. If we now inject a high-power control signal directly into the loop, such that it propagates unidirectionally, and is

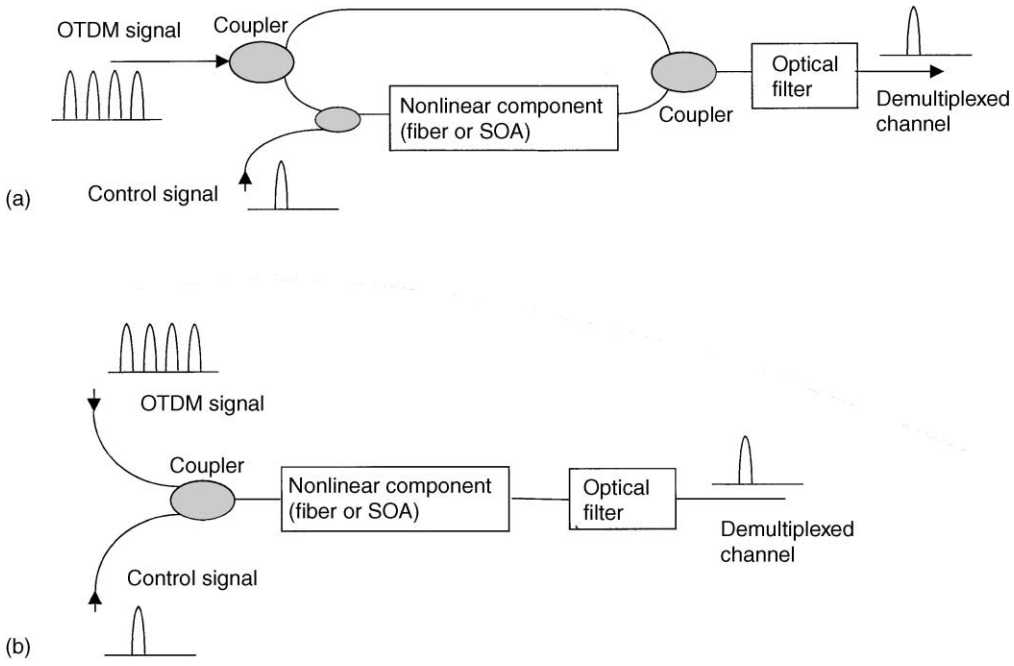


Figure 5 Configuration of OTDM demultiplexers using (a) an interferometer based on nonlinear phase shift in fiber or semiconductor optical amplifiers, and (b) four-wave mixing in fiber or semiconductor optical amplifiers.

synchronized with one of the OTDM data channels, then the phase shift induced by the control on the copropagating signal channel (via the nonlinear refractive index of the fiber) results in that particular channel being switched out to the second input port of the coupler. The remaining data channels of the OTDM signal are once again output through the same port that they entered the coupler. As this particular nonlinear effect in fiber (known as the Kerr effect) has a femtosecond response time, it should be possible to realize extremely high-speed switching with the NOLM. Demultiplexing at data rates in excess of 100 Gbit/s has already been achieved using this technique. One disadvantage of this method is that the nonlinear index coefficient of standard fiber is very small, so in order to achieve the required phase shift from the control pulse, it is necessary to use a fiber loop of around 1 km in length (depending on the power of the control pulse used). This length may affect the overall stability of the demultiplexer and its ability to be readily integrated into high-speed OTDM systems. However, it should be noted that recent developments in the design of high-nonlinearity fibers may reduce this problem.

The second all-optical switch mentioned above is known as a TOAD. Whereas the NOLM is based on the nonlinear refractive index of the fiber, the TOAD is based on a nonlinear optical effect in semiconductor optical amplifiers (SOA). The overall setup for a TOAD is very similar to that for a NOLM, except that a semiconductor amplifier, as shown in Figure 5, replaces the length of fiber in the loop. As for the NOLM, the OTDM data signal is injected into the TOAD using one input of the coupler, such that it splits into two counterpropagating signals, while the control signal propagates unidirectionally in the loop containing the SOA. By synchronizing the control with one channel in the OTDM signal, this particular channel is switched out. Although the response time of a TOAD-based switch may not be as fast as the NOLM, it does have the major advantage that it can be developed into a very compact demultiplexer, suitable for deployment in OTDM-based communication systems.

In addition to the NOLM and the TOAD structures, which are both interferometric-based switches, it is also possible to use four-wavemixing (FWM) in optical fiber or semiconductor optical amplifiers to carry out the demultiplexing of OTDM signals. In this case, the high-power control pulse is synchronized with one of the OTDM channels such that the wavelength of this channel is shifted as it passes through the nonlinear device. Optical filtering may then be used to select out the demultiplexed channel.

As we discussed in the section on demultiplexing, whatever specific all-optical switching technique is employed, we need to generate an optical clock signal for use as the control pulses. This implies that it is necessary to extract a clock at the base rate of the individual data channels from the high-speed OTDM signal. Depending on the base data rate, which may be anywhere from 2.5 to 40 Gbit/s for typical OTDM systems, different clock recovery techniques have been demonstrated. These range from basic electronic clock recovery schemes to more advanced techniques based on self-pulsating laser diodes and optical phase-locked loops.

OTDM Networking Issues

The previous section has predominantly dealt with the basic elements required to implement a high-speed OTDM transmission link. However, if we are to use OTDM technology for high-speed networking, then additional elements will be required. One of the most important of these is an add/drop multiplexer (ADM). The ADM allows us to switch out one channel from the OTDM signal, and then insert another data channel for transmission in the vacant bit slot. If we consider employing OTDM in a ring configuration network, then a typical architecture for the network will be as shown in Figure 6. The high-speed OTDM data signal propagates around the closed fiber loop, and each node in the network is responsible for switching out the data it requires, in addition to inserting the data it wishes to transmit onto the fiber ring. Add/drop multiplexing is thus imperative in developing such an OTDM ring network.

The configuration for each node in an OTDM ring network is shown in Figure 7. At each node it is initially necessary to recover the base rate clock signal. The recovered clock is then used in conjunction with the OTDM signal to switch out the required data channel from the multiplexed signal. When an add/drop multiplexer is used for this purpose we obtain two outputs, one being the demultiplexed data, and the other being the OTDM signal less the switched-out data channel. A simple fiber coupler can subsequently be used to allow this particular node to insert the data it wishes to send into the free bit slot of the overall OTDM signal. This obviously requires control of the relative delay between the OTDM signal and the local data channel to be inserted onto the fiber ring, to ensure that the local data are inserted into the vacant bit slot. A number of techniques have been used for implementing add/drop multiplexers in

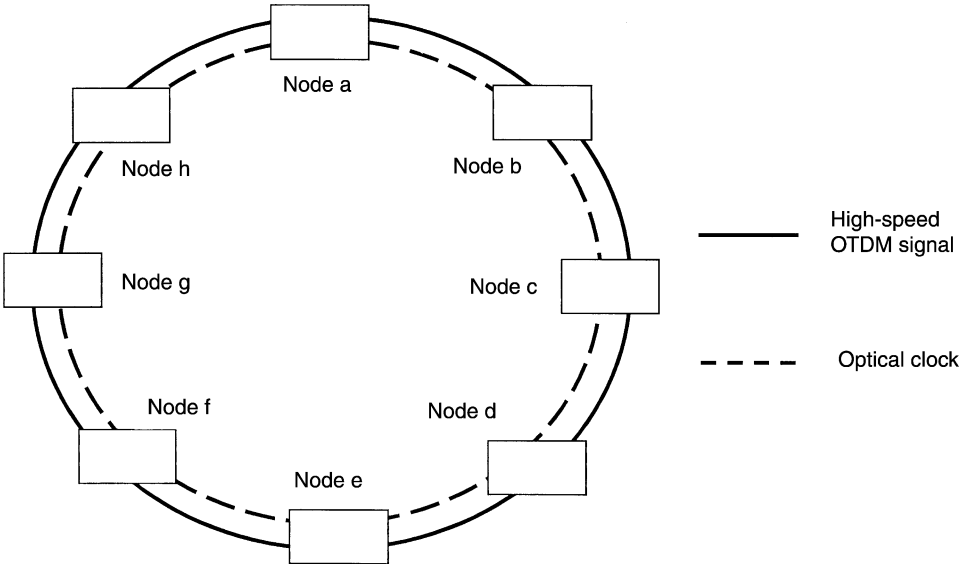


Figure 6 Schematic representation of a ring network employing OTDM with a separate fiber for clock distribution.

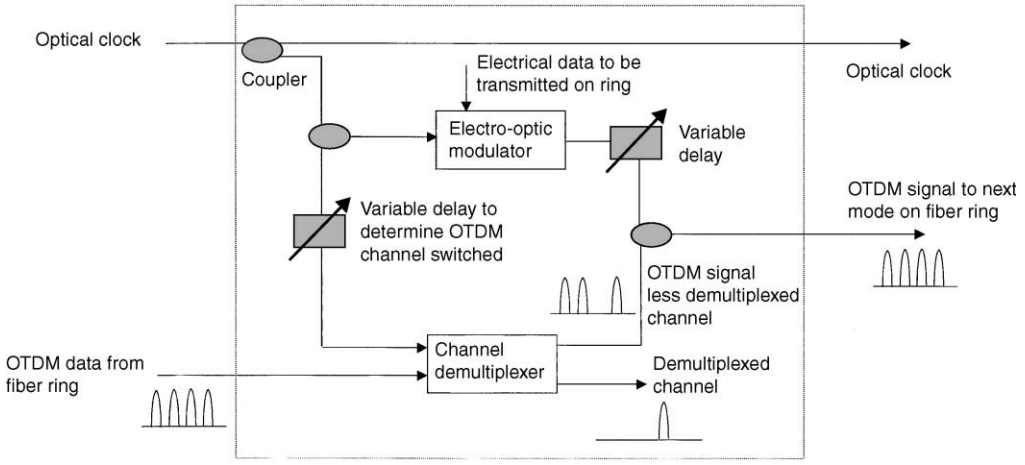


Figure 7 Configuration of an add/drop multiplexing node for use in OTDM-based communication systems.

OTDM-based networks. These have used either fiber interferometers, or interferometers using semiconductor optical amplifiers, in which we obtain both the demultiplexed data signal and the OTDM signal less the switched-out channel (to which the local data channel is then added). The continuing development and enhancement of ADM devices for OTDM systems will be vital for the future implementation of OTDM communication systems.

Conclusion

In this article we have explained the main ideas involved in building OTDM communication systems, and also introduced the principal technologies which are required to do so. OTDM is an extremely wavelength-efficient technique for delivering high-capacity data signals, and it may be used both for long-haul transmission and networking around metropolitan areas. In addition, unlike WDM, it does not require a different wavelength source for each channel in the multiplexed signal. Despite these and other advantages, OTDM is still way behind WDM when it comes to commercial maturity, with no OTDM systems currently available in the telecommunications market. However, the technologies required to implement high-speed OTDM systems (outlined in Table 2), including ultrashort pulse sources, clock extraction circuitry, and demultiplexing devices, are developing rapidly. It therefore seems

Table 2 Main elements of an OTDM communications system

Device	*Principal function*	*Technical implementation*
Optical pulse source	Generate ultrashort optical pulses suitable for transmission in high-speed OTDM systems	Gain-switching, mode-locking, or external modulation techniques
Multiplexer	Multiplex individual pulse data channels in the temporal domain using appropriate delays to achieve an OTDM signal	Passive couplers with fixed fiber delay lines, or planar waveguide circuits (for accurate control of delays)
Optical amplifier	Amplify the OTDM signal during fiber transmission to overcome fiber loss	Fiber doped with rare earth materials (e.g. erbium) in conjunction with pump laser
Transmission fiber	Transmit the OTDM data signal from transmission site to required receiver	Standard single-mode fiber, dispersion shifted fiber, or other specialty fiber may be used
Clock recovery	Extract clock (at base rate of individual channels) from OTDM data signal at receiver, for use in demultiplexing	Electronic or optical phase-locked loops, self-pulsating lasers, or mode-locked ring laser techniques
Demultiplexer	Demultiplex one of the channels from the overall OTDM signal	Interferometric or four-wave mixing techniques based on fiber or semiconductor nonlinearity
Add/drop demultiplexer	Extract one channel from OTDM data signal, and insert a local channel into the empty time slot for transmission	Similar technical implementations to demultiplexer described above

likely that in the future, OTDM technology may be used for enhancing the overall capacity of optical communication systems.

One possibility for employing OTDM systems would be for upgrading installed WDM networks. The current method for improving WDM networks involves using more wavelength channels, but it may also be feasible to increase the overall capacity by developing hybrid WDM/OTDM communications systems. In such a system, each wavelength channel may carry aggregate data rates in excess of 100 Gbit/s, achieved using OTDM. The design of these hybrid networks would permit switching of the optical data signals to be carried out at two different levels in the overall network. The WDM signals could be switched coarsely using passive filtering devices, and the OTDM data channels could be switched with fine granularity using one of the all-optical switching techniques available.

List of Units and Nomenclature

Dispersion	[ps km^{-1} nm^{-1}]
Transmission data rate	[bit s^{-1}]
ADM	Add drop multiplexer
CW	Continuous wave
FWM	Four-wave mixing
NOLM	Nonlinear optical loop mirror
OTDM	Optical time division multiplexing
SDH	Synchronous digital hierarchy
SOA	Semiconductor optical amplifier
STM	Synchronous transport module
TOAD	Terahertz optical asymmetric demultiplexer
WDM	Wavelength division multiplexing

See also

All-Optical Signal Regeneration. **Optical Communication Systems:** Basic Concepts; Wavelength Division Multiplexing.

Further Reading

Agrawal GP (1997) *Fibre Optic Communications*, 2nd edn., pp. 329–335. New York: Wiley.

Cole M (2000) *Introduction to Telecommunications: Voice, Data, and the Internet*, chapter 5. Englewood Cliffs, NJ: Prentice Hall.

Islam M (1992) *Ultrafast Fiber Switching Devices and Systems*. Cambridge, UK: Cambridge University Press.

Kawanishi S (1998) Ultrahigh-speed optical time-division-multiplexed transmission technology based on optical signal processing. *IEEE Journal of Quantum Electronics* 34: 2064–2079.

Midwinter JE and Guo YL (1992) *Optoelectronics and Lightwave Technology*, pp. 148–178. Chichester, UK: Wiley.

Mukherjee B (1997) *Optical Communications Networks*, pp. 493–506. New York: McGraw-Hill.

Nakazawa M, Kubota H, Suzuki K, Yamada E and Sahara A (2000) Ultrahigh-speed long distance TDM and WDM soliton transmission technologies. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 363–396.

Prucnal PR, Santoro MA and Sehgal SK (1986) Ultrafast all-optical synchronous multiple access fiber networks.

IEEE Journal of Selected Areas of Communication 4: 1484–1493.

Spirit DM, Ellis AD and Barnsley PE (1994) Optical time division multiplexing: systems and networks. *IEEE Communications Magazine* 32(12): 56–62.

Spirit DM, Blank LC and Ellis AD (1995) Optical time division multiplexing for future high capacity network applications. In: Smith DW (ed.) *Optical Networking Technology*, pp. 54–77. London: Chapman and Hall.

Tucker RS, Einstein G and Korotky SK (1988) Optical time-division multiplexing for very high bit rate transmission. *IEEE Journal of Lightwave Technology* 6: 1737–1749.

Wavelength Division Multiplexing

J Bowers and H F Chou, University of California, Santa Barbara, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Wavelength division multiplexing (WDM) has enabled a revolution in communications technology. This article describes the technology, critical components of WDM systems, and transmission impairment and limits to transmission distance and capacity.

Multiplexing

There are three techniques used in fiber optics to multiplex the information that is transmitted from one point to another. The first is space division multiplexing (SDM), which simply consists of having multiple fibers in a cable or conduit. In a typical installation, there may be between 8 and 1000 fibers in parallel. The second approach is time division multiplexing (TDM) (*see* **Optical Communication Systems:** Optical Time Division Multiplexing), which consist of interleaving multiple data streams into one higher bit rate stream. For example, 672 voice channels (DS0) at 64 kbit/s are multiplexed into one T1 data stream at 1.544 Mbit/s. Eighty-four of these T1 channels are then multiplexed into a SONET (Synchronous Optical NETwork) OC-3 signal at 155 Mbit/s. The third multiplexing approach is wavelength division multiplexing (WDM), which consists of sending multiple signals at different wavelengths. These signals can be used to fill up the entire low loss transmission band of an optical fiber (Figure 1). The low loss transmission band extends from 1200 nm to 1600 nm, which corresponds to over 40 THz. One approach is to pack the signals densely together, as in dense wavelength division multiplexing (DWDM). This is a relatively expensive, but very high capacity approach. The other approach is to space the signals widely apart, (coarse WDM or CWDM) resulting in very low cost components, but with fairly limited transmission capacity.

Dense Wavelength Division Multiplexing (DWDM)

Most DWDM systems use multiple beams spaced at 100 GHz spacing centered at 193.1 GHz as defined by an International Telecommunications Union (ITU) standard (Figure 2a). Some higher capacity systems use finer spacing, 50, 25, or even 12.5 GHz spacing to pack more information into the fiber. The 100 GHz spacing used in DWDM systems corresponds to 0.8 nm spacing at 1.55 mm. In principle, 40 THz of information can be sent down an optical fiber. If high spectral efficiency (~1 bit/Hz) is maintained across the band, then this would correspond to 40 Tbit/s. Practical DWDM systems are limited by the bandwidth of erbium-doped fiber amplifiers (EDFAs) and by signal impairments discussed below. Consequently, most systems utilize a single band, typically the C-band extending from 1530 nm to 1565 nm. Additionally some transmission systems use multiplexers to combine other bands, typically C and L bands for additional transmission capability. The wavelength definitions are shown in Figure 1. Subbands are often used within a band to facilitate adding and dropping a group of wavelengths within a band. Typically 40 10 Gbit/s signals spaced at 100 GHz are used to fill the C-band. Higher capacity

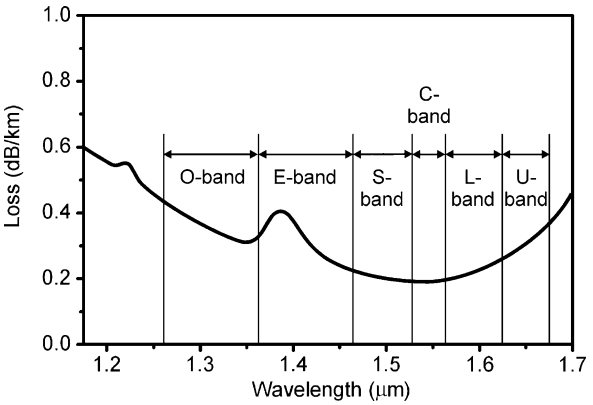


Figure 1 Transmission spectrum of optical fiber. Also shown are the standard transmission bands.

systems with 80 to 200 wavelengths have also been deployed commercially.

The historical evolution of optical transmission systems can be seen in Figure 3. The lower line shows the increasing bit rate demonstrated in experimental TDM systems, which began in 1978 with 45 Mbit/s systems and has now reached 1.28 Tbit/s. The upper line shows what has been demonstrated with DWDM with over 10 Tbit/s transmission demonstrated.

Coarse Wavelength Division Multiplexing (CWDM)

The fine frequency requirements of DWDM systems result in significant cost for transmitters, multiplexers and other components. A low-cost approach is to space the colors widely apart. The ITU standard is 20 nm spacing from 1271 to 1611 nm (Figure 2b). Note that the spacing is uniform in wavelength, but varies in frequency because the frequency spacing is related to the wavelength spacing by

$$\Delta f = -\frac{c}{\lambda^2}\Delta\lambda \qquad [1]$$

The wide spacing of signal wavelengths means that temperature stabilization of elements is not needed and the yield of components is high, resulting in a lower-cost system.

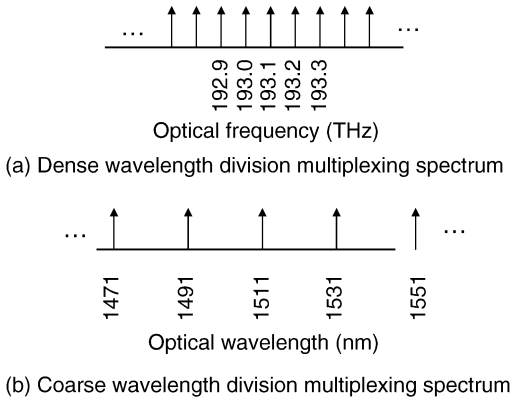


Figure 2 (a) DWDM spectrum (b) CWDM spectrum.

Point-to-Point DWDM Links

Point-to-point DWDM links were widely deployed during the 1990s. The essential elements are shown in Figure 4, namely single frequency lasers, optical multiplexers and demultiplexers, optical amplifiers, optical add drop multiplexers (OADM), and optical receivers. These will be described in more detail in subsequent sections. Essential parts of the design of these links are managing the signal-to-noise ratio and managing the pulse broadening. With periodic amplification in EDFAs and dispersion compensation, links of 1000 km are common. The widespread use of forward error correction (FEC) circuits allows error-free transmission over a link with limited signal-to-noise ratio. With careful design, ultra-long-haul

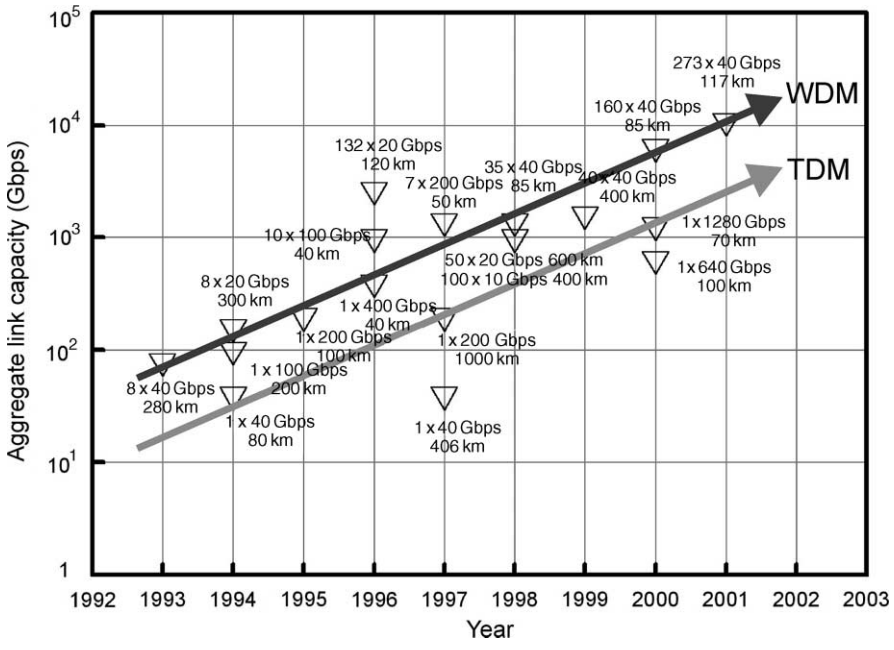


Figure 3 Historical evolution of transmission systems.

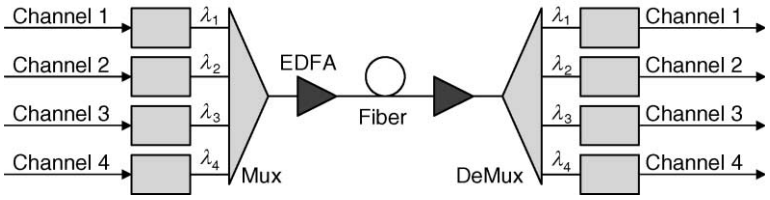


Figure 4 Point-to-point WDM link.

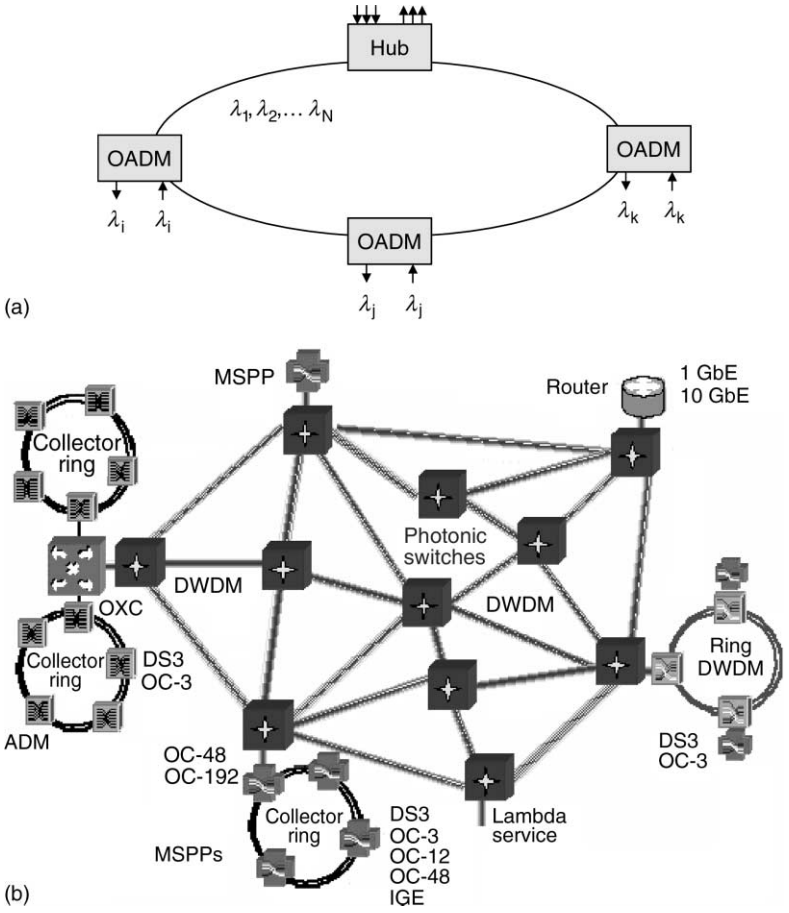


Figure 5 (a) WDM ring network. (b) DWDM mesh networks (after Basch B, Gringeri S, Goudreault R and Ravinkumar S (2002) Evolution of the photonic core within metropolitan and regional networks. *NFOEC*. Technical Proceedings).

transmission over 5000 km is possible, even with many 10 Gbit/s channels.

DWDM Networks

Point-to-point DWDM links are widely deployed in opaque networks with electronic regeneration and switching at nodes. The next stage in optical networks is the deployment of DWDM transparent ring and mesh networks. An example of a simple ring network is shown in Figure 5a where most of the light signals transit nodes transparently. The OADM elements drop particular wavelengths at each node. As more wavelengths are deployed in these ring networks, it becomes necessary to dynamically reallocate wavelengths and resources. Reconfigurable optical add/drop multiplexers (ROADMs) and optical switches have been developed to allow for dynamic transparent mesh networks (Figure 5b). This allows service providers to provision their networks remotely and quickly adapt to changing demand and restore vital connections when service interruptions occur.

Critical WDM Components

The previous sections described WDM networks. This section describes in more detail the critical components used in these networks.

Multiplexers/Demultiplexers

Multiplexers (MUXs) are used to combine multiple wavelength channels into a single fiber. They are usually wavelength selective and utilize wave phenomena such as diffraction and interference. A simple cascade of 3 dB couplers is not a practical multiplexer since the output power decreases exponentially with channel count. Due to the reciprocity of electromagnetic waves, most MUXs can also serve as a DEMUX. For high channel count systems, a multistage design may be necessary to overcome the limitations of a single MUX/DEMUX. This approach also provides a more flexible upgrading plan. In this case, wavelength channels can be grouped into bands or interleaved. Diffraction gratings and arrayed waveguide gratings (AWGs) are among the available technologies for MUX/DEMUX.

In a diffraction grating, narrow slits or micro structures are spaced evenly on the grating plane and serve as a secondary light source for the input light. The transmitted or reflected light (depending on the design) forms a diffraction pattern on the image plane. The locations of peak intensity on the image plane depend on the wavelength and the grating pitch. Therefore, wavelength channels can be separated in space and fibers can be positioned properly to couple individual channels. An AWG is a planar integrated and digitized form of diffraction grating as shown in Figure 6. It is usually fabricated on a silicon substrate with silica waveguides (silica-on-silicon). The input light is coupled into an array of waveguides with incremental difference in length. These tributaries are then combined in a free-propagating output coupler to several output waveguides. Due to the interference of these tributaries from the arrayed waveguides, different wavelength channels are focused to different output waveguides.

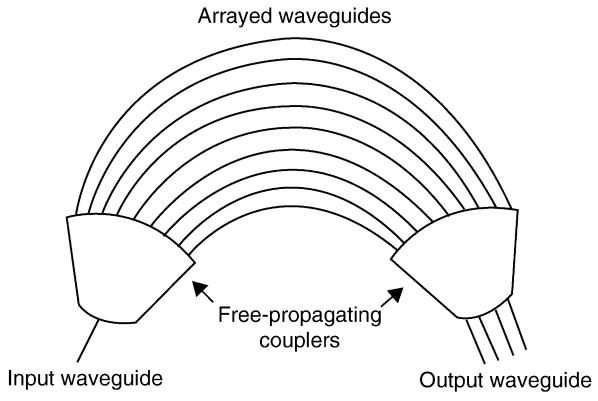


Figure 6 Schematic diagram of an AWG.

Optical Add/Drop Multiplexer

An optical add/drop multiplexer (OADM) provides critical networking capability for WDM systems. In an OADM, some of the wavelength channels are dropped from the input and are either terminated or routed. New channels at these wavelengths are then added with the through channels. This function can be done with the combination of a MUX and a DEMUX of any kind as shown in Figure 7a. However, if only one wavelength channel needs to be processed while others pass through, a simpler configuration can be used where a circulator drops the reflected wavelength from the fiber Bragg grating and a new channel is added with a coupler (Figure 7b).

The two approaches described above are termed static since the wavelength processed is fixed. To enable a more flexible network, reconfigurable OADM (ROADM) are required, where the wavelengths to be added or dropped can be dynamically specified. Figure 7c shows a configuration of ROADM based on a high port-count optical switch and a set of MUX/DEMUX. A tunable laser is highly desirable in this scenario since the wavelength of the added channel can be reconfigured to resolve wavelength contention in the network.

Optical Amplifiers

Erbium-doped fiber amplifiers (EDFA) (*see* **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems) provide simultaneous amplification of wavelength channels in the entire C-band (1530 to 1565 nm) and part of the L-band (1565 to 1605 nm). EDFAs were a key enabling technology for the deployment of WDM systems, especially for long-haul transmission where costly regenerators were eliminated. Both 980 nm and 1480 nm pumps can be utilized to achieve gain in erbium-doped fibers. The 980 nm pumps provide low-noise and high-gain operation while the 1480 nm pumps are more favorable for high-output power. One of the most significant advantages of an EDFA over a semiconductor optical amplifier (SOA) is that crosstalk among wavelength channels is much smaller in EDFA due to a millisecond-scale spontaneous emission lifetime. One critical issue for EDFAs is that the gain is not flat over the operation wavelength range and peaks around 1532 nm. However, this can be equalized by specially designed filters.

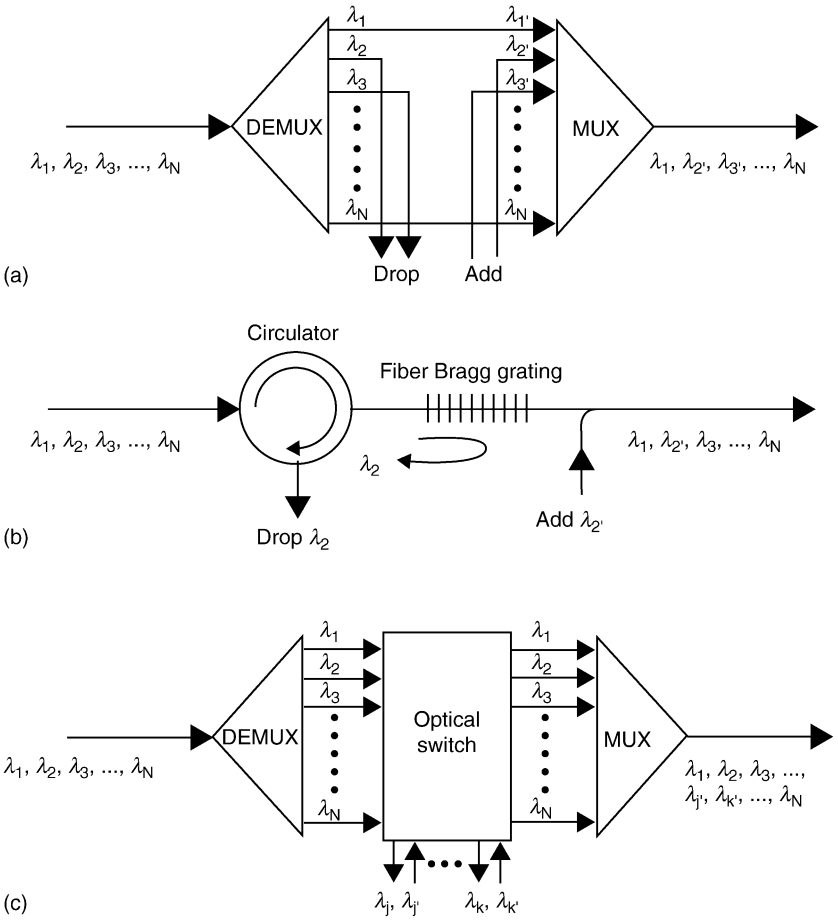


Figure 7 (a) A parallel OADM architecture which allows multiple simultaneous add's and drop's; (b) a linear OADM architecture for processing a single channel; (c) an ROADM which enables reconfigurable processing of multiple wavelength channels.

Outside the gain spectrum of EDFAs, Raman amplifiers can be used, where the gain peak is about 13 THz (100 nm) below the frequency of the pump signal. Also, Raman amplification occurs in the same fiber used for transmission so that it can provide a distributed gain over the span of the fiber, which results in a better signal-to-noise ratio. This is particularly important for ultra-long-haul transmission. A counter-propagating pump scheme is preferable to reduce the influence of pump noise and depletion. The biggest challenge for Raman amplifiers is the availability of high-power pump lasers at the required wavelengths.

Optical Filters

Filters can find many applications in WDM systems, such as in multiplexers, demultiplexers, OADM, receivers and optical amplifiers. Fiber Bragg grating, Fabry–Perot filter, and multilayer thin-film filters are among the most widely used filters and utilize the interference property of optical waves. There are some general requirements for filters in WDM systems. First, the top of the pass-band must be flat enough so that cascading of filters in the network does not lead to a significant decrease of bandwidth. In addition, the roll-off of the pass-band must be sharp so that cross-talk from other channels is minimized.

A fiber Bragg grating has a periodic perturbation of refractive index made by exposing a photosensitive fiber to interfering ultraviolet lights. These perturbations reflect the input light as it propagates. At a specific wavelength, these reflected beams add in phase, which results in a strong overall reflection. The reflections at other wavelengths do not add up and these wavelengths will eventually pass through the grating. In short, the fiber Bragg grating acts as a wavelength-selective reflector and is usually used with a circulator or isolator to manage the reflected light. A Fabry–Perot filter is composed of two partially transmitting interfaces and a cavity in between. The underlying mechanism is similar to the fiber Bragg grating but the transmission is periodic in optical frequency. The spacing is called

the free spectral range, which depends on the optical length of the cavity. The bandwidth of the passband is a strong function of the reflectivity of the partially transmitting interfaces. A multilayer thin-film filter is an extended form of Fabry–Perot filter where several layers of dielectric reflectors are formed with dielectric cavities in between. This allows for more tuning dimensions in the design of the passband shape in order to meet the requirements of WDM systems and the temperature tolerance is better than other kinds of filter.

Laser Sources

Semiconductor lasers (*see* **Lasers:** Semiconductor Lasers) are widely used in optical communication because they are compact, low cost and efficient. As a light source for transmission, single frequency operation is necessary for high-capacity WDM systems. Multilongitudinal mode lasers such as Fabry–Perot lasers have excess bandwidth which causes excessive penalty due to dispersion unless the transmission distance is short. They also result in crosstalk in DWDM systems. The most widely used single-frequency laser in WDM systems is the distributed-feedback laser (DFB), where a corrugated grating structure is fabricated on top of the gain region and only the wavelength which satisfies the Bragg condition can oscillate in the laser cavity and all other modes are suppressed.

As WDM networks advance, tunable lasers are strongly desired to reduce the excessive inventory of fixed wavelength lasers for DWDM and also to provide reconfigurability in the network. One possible structure is the distributed Bragg reflector (DBR) laser, in which the Bragg grating is decoupled from the gain region and the Bragg wavelength can be tuned independently by isolated current injection. More complicated variations of this concept such as the three-section DBR laser and the sampled grating DBR laser can provide more continuous and wider tuning in wavelength.

Photonic Switches

Photonic switches are necessary in transparent networks to allow reconfigurability in networks, and allow rapid provisioning and fault recovery. They provide connectivity between rings and between point-to-point links without requiring the DWDM signal to be broken into individual wavelengths and to be electronically regenerated. Photonic switches can be used to switch fibers, wavebands, or individual wavelengths. They are inherently transparent to bit-rate and data format. A number of different technologies have been developed.

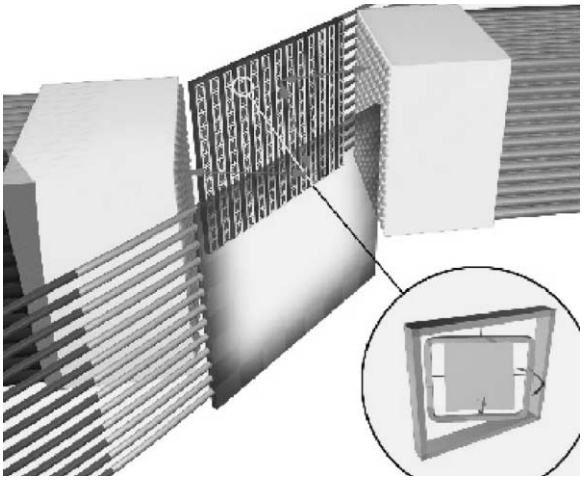


Figure 8 3D MEMS optical switch.

Microelectromechanical system (MEMS) switches are widely used for 2×2 protection switches, wavelength selective switches and for nonblocking crossbar switches with sizes from 8×8 to 1024×1024. The technology usually used for large switches is called 3D MEMS switches (Figure 8), in contrast to 2D technologies where the optical beams are confined to a two-dimensional plane.

Transmission Issues

Chromatic Dispersion

Chromatic dispersion (*see* **Fiber and Guided Wave Optics:** Dispersion) means that the group velocity, which is the propagation speed of an optical signal, changes with wavelength. Since every real pulse has a finite bandwidth, each slice of wavelength within this bandwidth propagates with a different speed and results in the distortion of the pulse as it propagates along the fiber. Depending on the sign of the dispersion and the initial phase distribution, the pulse can either be broadened or compressed. However, if the fiber is long and nonlinear effects are small, the pulses will eventually be broadened and result in intersymbol interference. Dispersion compensating fibers (DCFs), which have an opposite sign of dispersion, can be used to reverse this process and restore the original pulse shape. In general, it is advantageous to minimize the length of DCF by keeping the dispersion low. Dispersion-shifted fibers (DSF) have a zero-crossing of dispersion around 1.5 μm and hence a reduced dispersion for the C-band.

For a WDM system, a very low dispersion will result in more severe nonlinear impairments since the distance pulses can overlap with each other along the fiber is long. Nonzero-dispersion-shifted fibers

(NZ-DSF), which introduce a suitable amount of dispersion, can be used instead to solve this issue. Chirped fiber Bragg grating can also be used, where different wavelengths are reflected at different positions in the grating so as to compensate the dispersion caused by transmission. It is capable of compensating a given amount of dispersion with a much shorter length when compared to DCF. For high channel count WDM systems, the dispersion experienced by each channel may be different and dispersion compensation needs to be implemented on a per-channel basis after a demultiplexer.

Polarization Mode Dispersion

Polarization mode dispersion (PMD) means that the group velocity dispersion of the two orthogonal polarizations in the fiber are different and cannot be simultaneously compensated by a single DCF. It is usually manifested as the splitting of a pulse at the receiver even after chromatic dispersion compensation and causes intersymbol interference. PMD can originate from the imperfections in the fiber or environmental factors like temperature and mechanical pressure. Therefore, the amount of PMD can vary with time for a given link, which means that dynamic compensation is necessary. Traditional electrical equalization after the optical signal is detected can be used for compensation but the highest bit-rate achievable is limited by the speed of electronics. Optical compensation is another choice. The optical signal is split into fast and slow components and the fast component is delayed properly before being recombined with the slow component.

Nonlinear Impairments

The most important nonlinear impairments for WDM systems are the cross-phase modulation (CPM) and the four-wave mixing (FWM). They are both third-order nonlinear effects, where the nonlinear dielectric polarization is proportional to the third power of the electric field of the signal. In CPM, the phase of a wavelength channel can be modified by other channels, which increases the chirp and results in more dispersion penalty. On the other hand, FWM can generate new waves at other wavelengths, causing unwanted crosstalk if the generated wave happens to coincide with another channel. The existence of dispersion in fiber can help to reduce the effects of both impairments since the pulses at different wavelengths propagate at different speeds and walk away from each other so that nonlinear interactions are reduced. This is why NZ-DSF is preferred over DSF. Self-phase modulation (SPM) is another third-order nonlinearity, where the phase distribution of a pulse is altered by its intensity profile. It will lead to the distortion of the pulse. Other nonlinear effects include stimulated Brillouin scattering and stimulated Raman scattering. Both originate from the interactions of phonons and photons in the fiber and cause power fluctuations at different wavelengths.

Forward Error Correction

To reduce the errors of a transmission link, extra information relating to the data can be sent by the transmitter and used by the receiver to detect and correct errors in the received data. This technique is called forward error correction (FEC) and imposes redundancy to the transmission which means that the bit-rate transmitted will be higher than the actual data rate. A measure of performance for FEC is the coding gain, which is defined as the decrease in receiver sensitivity at a given bit-error-rate. Reed–Solomon codes are one of these codes. With a 7% redundancy, up to 6 dB of coding gain can be achieved. This could correspond to more than 9 orders of magnitude improvement in bit-error-rate. Therefore, the use of FEC can increase the spacing between regenerators in ultra-long-haul transmission and reduces the cost. It is also useful to overcome error floor problems due to crosstalk in a WDM system, where a further increase in received power cannot reduce the bit-error-rate. Crosstalk among wavelength channels can happen due to nonlinear impairments during fiber transmission or due to imperfect channel isolation in components like multiplexers, demultiplexers, filters, and switches. The use of FEC can greatly increase the system margin.

Conclusions

WDM technology has enabled a huge increase in the transmission capacity of optical fibers and a huge reduction in the cost of fiber optic transmission since the cost of the installed fiber is shared between the signals traveling on different wavelengths. As new components have been developed, WDM technology has expanded simple point-to-point links to transparent networks. Great progress has been made over the past 20 years to solve transmission impairments and extend the reach and bandwidth of these systems. This is likely to continue to meet the increasing communication needs of our society.

List of Units and Nomenclature

Distance (km, m, μm)
Bit rate (bit/s)
Wavelength (nm)

See also

Color and the World. **Information Processing:** All-optical Multiplexing/Demultiplexing. **Fiber and Guided Wave Optics:** Optical Fiber Cables; Passive Optical Components. **Lasers:** Edge Emitters; Semiconductor Lasers. **Modulators:** Modulation and Demodulation of Optical Signals. **Optical Amplifiers:** Basic Concepts; Optical Amplifiers in Long-Haul Transmission Systems; Semiconductor Optical Amplifiers; **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Basic Concepts; Historical Development; Lightwave Transmitters; Optical Time Division Multiplexing. **Solitons:** Soliton Communication Systems.

Further Reading

Agrawal GP (2002) *Fiber Optic Communication Systems*, 3rd edn. New York: John Wiley & Sons Inc.

Basch B, Gringeri S, Goudreault R and Ravinkumar S (2002) Evolution of the photonic core within metropolitan and regional networks. *NFOEC*. Technical Proceedings.

Green PE (1993) *Fiber Optic Networks*. Upper Saddle River, NJ: Prentice Hall.

Ramaswami R and Sivarajan KN (2002) *Optical Networks: A Practical Perspective*, 2nd edn. San Francisco, CA: Morgan Kaufmann Publishers.

OPTICAL MATERIALS

Contents

Color Filter and Absorption Glasses
Heterogeneous Materials
Lightweight Mirrors
Measurement of Optical Properties of Solids
Optical Glasses
Plastics
Sculptured Thin Films
Smart Optical Materials
Sol-Gel Materials

Color Filter and Absorption Glasses

J E Shelby, Alfred University, Alfred, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical filter and absorbing glasses can be categorized by either the optical behavior of the filter, for example, bandpass, shortpass or longpass, neutral density, or interference filters – or by the method used to filter the light, i.e., glasses which absorb light themselves and those which serve as high quality, transparent substrates for coatings which either absorb or reflect light. The term 'filter' usually refers to glasses used in optical systems, to either transmit or prevent transmission of specific wavelengths. The term 'absorber' is usually reserved for glasses which are used for controlling light to provide protection from damage to humans, for example, for eye protection, as in sunglasses, or welder and laser goggles, or to materials exposed to light, for example, plastics and other components in the passenger compartment of automobiles. Manufacturing technology is normally based on the method used to produce the device, such as control of absorbing species in the glass or method of application of optical coatings. Terminology for optical applications is almost always based on optical behavior and not on the method used to produce that behavior.

Bandpass Filters

The first glasses used to filter light were colored by inclusion of ions from the first row of transition metals. These colored glasses still provide most of

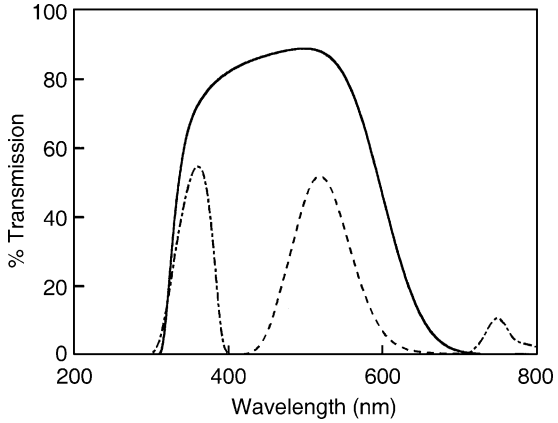


Figure 1 Transmission spectra for bandpass filters with a large bandwidth (solid line), medium bandwidth (dashed line), and narrow bandwidth (dot dash line).

the glasses used to control the wavelength of light. They are widely used as absorbers in windows in buildings and automobiles, to control the entrance of ultraviolet and infrared light into these environments. These filters allow the passage of light over a significant range of wavelengths, while absorbing wavelengths outside that range. They are designated by the peak transmission wavelength, or central wavelength, and the bandwidth, or wavelength range, over which light is transmitted with at least some designated transmission value. Filters based on absorption by transition metal or rare earth ions usually have larger bandwidths and somewhat diffuse cutoffs at the edges of the transmitted spectral region (Figure 1).

The most common bandpass filters absorb radiation in a desired wavelength region. Historically, the first bandpass filters, which are still the most common, relied on absorption of light as a result of ligand field (also known as crystal field) effects and/or charge transfer processes within a glass, to create absorption bands in the spectrum.

Ligand field effects will produce optical absorption bands in glasses which contain transition metal (usually row 3d) or rare earth (lanthanide, or 4f) ions. The differences in the electronic structures of the transition metal and rare earth ions result in very different absorption spectra. Absorption spectra, due to transition metal ions, are characterized by broad bands and are frequently accompanied by charge transfer bands, due to the same ions, while the bands due to absorption of light by rare earth ions are very sharp and are not accompanied by charge transfer absorption bands.

Iron oxide was probably the first, and is still the most widely used additive for production of glasses which primarily transmit light in the visible portion of the spectrum. A large, broad absorption band due to ferrous ions is centered at $\approx$1050 nm in common soda-lime-silica glasses. The tails of this band extend into the visible portion of the spectrum, allowing passage of the blue-green portion of the spectrum, but absorbing much of the red light, and to wavelengths in excess of 2000 nm, absorbing much of the near infrared spectrum.

Ferric iron produces relatively small ligand-field absorption bands in the region from 380 to 440 nm. These bands remove most of the blue part of the spectrum, allowing passage of light with longer wavelengths. Both ferrous and ferric ions also display charge transfer bands in the ultraviolet, effectively eliminating transmission of light with a wavelength $<$300 nm. Glasses containing enough iron oxide, to effectively eliminate both ultraviolet and near infrared transmission, exhibit various shades of green, depending on the ratio of ferrous to ferric ions present. Melting under strongly reducing conditions yields blue-green glasses, whereas melting under strongly oxidizing conditions yields yellow-green glasses. Glasses containing $\geqq$0.5 wt% iron oxide are used in windows of automobiles to minimize the amount of ultraviolet and infrared radiation entering the passenger compartment.

Copper oxide is also used to remove a portion of the near infrared spectrum from sunlight and to produce glasses which transmit in the green and blue region. The primary ligand-field absorption band of copper is due to cupric ions and is located near 800 nm. This band in similar in shape to that due to ferrous ions, with tails extending into the visible and near infrared regions of the spectrum. The shift in band position toward the visible, however, results in less absorption in the infrared and more in the visible than for ferrous iron. As a result, glasses containing copper are more strongly colored than those containing ferrous iron and are less effective filters for infrared light. Glasses colored with cupric ions can be used to produce a transmission window from about 350 to 600 nm in soda-lime-silica glasses. Cuprous ions do not create any significant absorption bands.

Glasses containing hexavalent chromium, which must be produced in a strongly oxidizing environment, effectively remove ultraviolet light between 300 and 400 nm, but allow transmission of ultraviolet light in a narrow range centered around 300 nm. A small tail of the charge transfer band due to Cr^{6+}extends into the visible, producing a yellow-green glass. The presence of a set of much weaker ligand-field bands, centered around 650 nm due to trivalent chromium, also contributes to the green color of these glasses. Combining chromium with copper ions in a glass yields a filter with a small transmission window centered at about 520 nm.

Combination of chromium with iron results in suppression of both hexavalent chromium and ferrous iron, yielding a green glass without either the strong ultraviolet absorption of hexavalent chromium or the strong near infrared absorption of ferrous iron.

Manganese oxide can be used to produce bandpass filters which block most of the visible, but allow transmission of ultraviolet and infrared light. Addition of manganese ions to a glass results in a set of absorption bands with a maximum intensity at about 500 nm and a tail extending to around 900 nm. Since these ions do not absorb in the ultraviolet, the transmission spectra of these glasses has a window between about 300 and 400 nm, which overlaps the transmission window due hexavalent chromium. The absorption bands due to manganese ions are relatively weak, so that concentrations of manganese oxide in excess of 1 wt% are needed to effectively block visible transmission.

Cobalt oxide is the strongest colorant of the oxides of 3d transition metals. A small addition of cobalt oxide to silicate glasses creates strong absorption bands between about 450 and 700 nm. The optical window of these glasses in the ultraviolet is centered near 380 nm and is somewhat broader than that of glasses containing manganese oxide, which have similar spectra. Since cobalt oxide is such a strong absorber, black glasses are easily produced by use of cobalt oxide in combination with other oxides. If cobalt oxide is combined with nickel oxide, the resulting glass retains a bandpass in the ultraviolet, while combining cobalt oxide with iron oxide, which has a complementary absorption spectrum, yields a black glass which also effectively eliminates transmission in the ultraviolet and near infrared.

The optical spectra of glasses containing most rare earth ions are characterized by sharp absorption bands. Some rare earth ions create spectra with many absorption bands, while others display only a few bands. The band structures which result from rare earth ions are not particularly useful for creation of bandpass filters, although Yb^{3+} ions do effectively remove light around 1100 nm, without absorbing light in the rest of the spectrum.

While not commonly used for filters, glasses containing rare earth ions have been used for wavelength calibration standards for uv-vis spectrophotometers. A common wavelength standard called a 'didymium' glass contains a combination of neodymium and prasedymium oxides. Since the optical spectrum from each of these ions contains many sharp bands, these glasses provide known wavelengths for absorption bands over a wide region of the spectrum. Neodymium ions, which have absorption bands near the D line of sodium, are used in goggles for eye protection for glassblowers, to remove the intense yellow light resulting from the sodium in the glass.

Interference Filters

Interference filters are a special case of bandpass filters which are created by application of dielectric films to glass substrates. The dielectric films are deposited on a glass substrate in alternating layers of high and low refractive index materials. The combination of refractive indices and thickness of the dielectric films determines the wavelength regions where constructive and destructive interference occur, i.e., those wavelength regions which will transmit or reflect the incident light. Since the choice of materials will control the indices, production of filters for different wavelengths usually relies on deposition of layers of differing thicknesses.

These filters can be produced with very narrow bandwidths ($\pm$10 nm) or with broader bandwidths (commonly $\pm$50 or $\pm$80 nm). The peak transmission frequently does not occur at the midpoint of the bandwidth. Since the wavelengths for constructive and destructive interference depend upon the path length through the dielectric layers, these filters are extremely dependent upon the incident light angle, with a shift toward lower wavelengths for incident angles other than normal to the filter surface. Interference filters typically reduce throughput of light as compared to absorption-based bandpass filters, but provide for much narrower transmission windows. If the dielectric layers are deposited in such a manner that their thickness varies along the length of the substrate, interference filters can be produced with a continually varying central transmission wavelength along the length of the substrate. While convenient for applications, which might otherwise require several filters of different central wavelengths, these linearly variable interference filters are much more expensive than those designed to block specific wavelengths.

Dispersion Filters

Dispersion filters are based on the passage of light through a medium consisting of a matrix material and dispersed, very small particles of a material which has a very different dispersion curve. The original version of these filters, called a Christiansen filter, consists of small glass particles with a controlled particle size suspended in a liquid. The dispersion curve for the glass is much less wavelength-dependent than that of the liquid. Transmission of light occurs at the wavelength where the refractive index of the glass and the liquid are equal, while light of other

wavelengths is scattered. Since the refractive index of a typical liquid is highly temperature sensitive, the transmitted, or peak, wavelength can be shifted over a wide region by controlling the temperature of the mixture. The bandwidth of dispersion filters can be very narrow ($<\pm 10$ nm).

The large temperature dependence of the traditional dispersion filter can be useful in that it provides an optical filter with an adjustable peak wavelength. On the other hand, this sensitivity to temperature also results in a serious control problem, since a temperature change of only a few degrees can shift the transmission window by up to 100 nm. This problem can be decreased or even eliminated by making a solid dispersion filter using two glasses of different dispersions, with one glass distributed in the other, or a glass dispersed in some other transparent solid such as a polymer with a similar thermal expansion coefficient. Since the refractive index of solids tends to vary with thermal expansion coefficient, these solid state dispersion filters can be almost temperature independent.

Neutral Density Filters

Neutral density filters are used to reduce the throughput of light to some desired value over a wide spectral region. Gray glasses are used to reduce the intensity of light from a bright source such as the sun. Black glasses are used to completely block light in the visible region. Neutral density filters can be produced using a combination of transition metal ions or by use of optical coatings. They are characterized by their percent transmission, range of wavelength where the transmission is constant, and the 'ripple', or spectral flatness, within the designated spectral region (Figure 2).

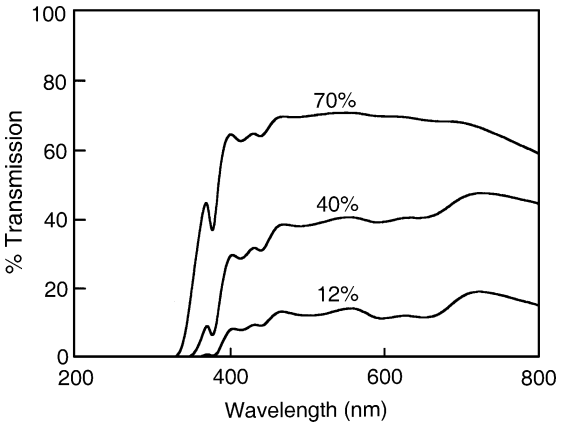


Figure 2 Transmission spectra for neutral density filters with different % transmission.

An approximation to a neutral density filter for visible light is often created by combining the absorptions in the ultraviolet and near infrared from ferrous and ferric ions, with the multiple absorption bands centered in the region of 500 to 600 nm due to divalent cobalt ions. While the resulting absorption spectrum displays small ripples throughout the region from 350 nm to beyond 1200 nm, the transmission of the glass can be controlled to within a few percent by the proper balance of ferric and ferrous ions combined with the desired concentration of CoO. These glasses are generally gray, or, if enough of iron and cobalt oxides are present, black. Combination of CoO with NiO can produce similar filters covering a smaller wavelength range which maintain transmission in the ultraviolet. Filters with different transmission values can be produced by either varying the composition of the glass or by use of different glass thicknesses. A continuously graded neutral filter can be produced from a single piece of glass by etching or polishing to form a wedge. With more effort, a step-graded filter can be produced by etching or polishing to form sharply delineated thickness steps along the length of a glass plate.

Neutral density filters can also be produced by coating a glass substrate with a metal film. Coated neutral density filters have less ripple than absorptive filters based on iron and cobalt oxides dissolved in the glass. In order to produce a constant transmission throughout the desired wavelength region, the metal used (commonly inconel) should have a constant reflectance across the desired wavelength region. Use of a metal with a varying wavelength dependence in its reflectivity, such as silver, will result in a wavelength dependent transmission. Continuously graded neutral filters can be produced by varying the thickness of the film along the length of a strip of glass.

Longpass/Shortpass Filters

Longpass filters are designed to pass light with wavelengths longer than the specified cutoff value. They are frequently used to block ultraviolet light while transmitting visible light or to block visible light while transmitting infrared light. Shortpass filters provide the opposite effect, i.e. they pass light with wavelengths shorter than the specified cutoff value, while blocking light with longer wavelengths. Shortpass filters are often used to block passage of heat (infrared light), while allowing transmission of visible light (Figure 3).

Ultraviolet blocking longpass filters can be produced by varying the basic composition of a glass or

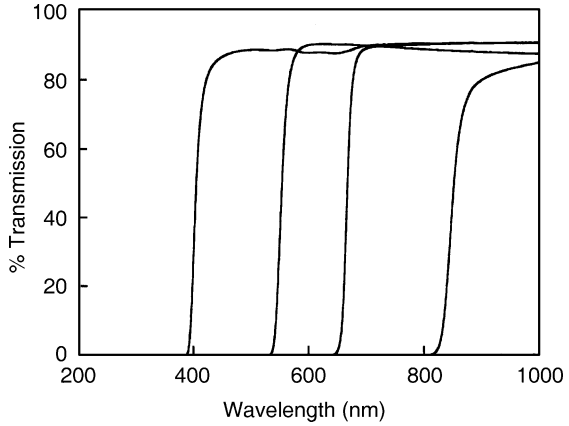


Figure 3 Transmission spectra for longpass filters with different cutoff wavelengths.

by addition of specific ions which absorb strongly in the ultraviolet. Varying the primary glassforming oxide can alter the ultraviolet cutoff frequency of a glass. Vitreous silica, for example, has an ultraviolet cutoff beyond 200 nm, so that it is a very effective transmitter of ultraviolet light. Vitreous germania, on the other hand, has an ultraviolet cutoff near 400 nm, so that it acts as an effective longpass filter for most ultraviolet light. While vitreous silica is widely used as an optical material, the poor chemical durability of vitreous germania prevents its use in most optical applications.

The ultraviolet cutoff wavelength of silicate glasses is shifted toward the visible by addition of alkali and alkaline earth oxides to silica. The addition of these oxides creates nonbridging oxygens in the vitreous network, which shift the ultraviolet cutoff toward the visible. These glasses are effectively opaque beyond their cutoff wavelength, creating longpass filter glasses with cutoff wavelengths which vary with the nonbridging oxygen concentration. Impurities such as iron or copper can also reduce the ultraviolet transmission of silicate glasses due to the creation of charge transfer bands, producing longpass filters with cutoffs in the range of 250 to 300 nm.

Longpass filters with tightly controlled ultraviolet cutoff wavelengths are often made by adding either cerium or a combination of cerium and titanium oxides to silicate glasses. Additions of cerium oxide result in a strong absorption band near 320 nm, which provides protection from ultraviolet light. Cerium also is routinely used to reduce radiation-induced coloration of glasses exposed to high energy radiation. An absorption band due to trivalent titanium also serves to cut off ultraviolet light, while producing a yellow glass. The use of a combination of cerium and titanium produces a very effective ultraviolet filter which eliminates transmission of almost all ultraviolet light.

Long bandpass absorption filters can also be based on precipitation of colloids of metals or semi-conductors in an otherwise colorless glass. The glasses are initially prepared in a colorless state, with the materials which will form the colloids dissolved as ions. The glasses are then reheated into the glass transformation region, where the desired ions are reduced to the atomic state and precipitated as colloids. Since this process is known as 'striking', these filters are known as striking filters. Colloids can be produced by striking from several elements, including silver, gold, copper, arsenic, antimony, lead, and bismuth, and from a number of compounds, including the cadmium chalcogenides and ZnS.

Formation of silver colloids in common soda-lime-silicate glasses results in a very strong optical absorption band centered near 400 nm. Glasses containing silver colloids are yellow, with a bandpass window in the ultraviolet. Since these glasses absorb so strongly at a wavelength commonly used to irradiate photosensitive polymers in the production of electrical circuit boards, they are frequently used for the production of beam masks. A pattern of silver colloids is easily produced in the surface of common float glasses by depositing metallic silver through a mask onto the surface of the glass and then heating the glass to 500 °C for a short time.

Copper can also be used to form longpass filters with a cutoff in the red portion of the spectrum by a similar process to that used to form the silver colloid filters. In this case, it is necessary to use a copper salt in place of the metallic silver, but the remainder of the process is essentially the same (Figure 4).

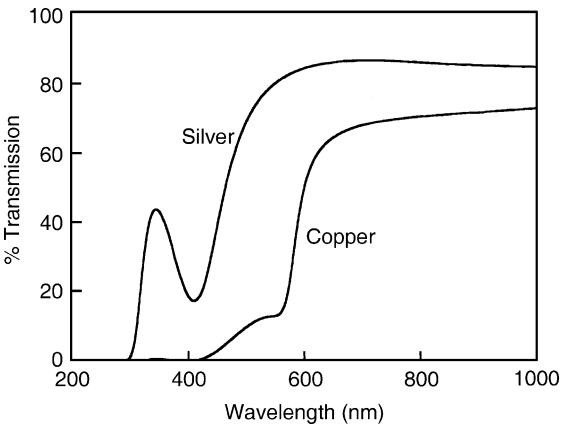


Figure 4 Transmission spectra for longpass filters containing silver or copper colloids.

Precipitation of semi-conducting colloids containing CdS, CdSe, and CdTe in silicate glasses results in sharp cutoffs in the range from the near ultraviolet to the near infrared. The cutoff is the direct result of the bandgap of these crystallites. Since these materials form a broad series of solid solutions, the desired cutoff wavelength can be obtained by use of an appropriate mixture of two of these compounds in the glass batch. Glasses containing only CdS, for example, have cutoffs in the range of 400 to 500 nm. Combining CdS with CdSe will shift the cutoff toward the infrared, with cutoffs in the range of 500 to 700 nm. Combining CdS with CdTe produces glasses with cutoffs near the edge of the visible region. Use of CdTe alone results in filters with cutoffs around 800 nm.

While the most common explanation for the shift in the cutoff in glasses containing the cadmium chalcogenides is based on solid solution arguments, it has been shown that the size of the particles also strongly affects the position of the cutoff for glasses containing CdSe. In this case, the energy gap, which determines the cutoff wavelength, varies inversely with the square of the particle size.

Glasses containing CdS and ZnS can be heat-treated to form filters with cutoffs in the range from 340 to 518 nm. Heat treatment at lower temperatures yields CdS colloids, with a cutoff around 500 nm. At higher heat treatment temperatures, the absorption edge shifts toward the ultraviolet, approaching the value of 340 nm representative of the lattice absorption of ZnS. This behavior has been interpreted as indicating that CdS crystals precipitate at lower temperatures, while at higher temperatures, solid solution crystals of ZnS and CdS are formed.

Colloids of arsenic, antimony, bismuth, and lead form less interesting filter glasses. The optical properties of these metals are such that there is no sharp cutoff wavelength. Growth of colloids produces wavelength-dependent scattering, with a long tail extending into the visible or infrared. These glasses can be produced in colors ranging from light tan to black by progressive heat treatments of the originally colorless glasses in hydrogen. The hydrogen serves to reduce the ions to the atomic state. The atoms then diffuse to form colloids. The growth process creates a sharp interface between the colored and colorless region of the glass, which can be used to produce windows for visible light of highly transmitting glass surrounded by essentially opaque glass.

Absorber Glasses

Glass filters are used to remove unwanted wavelengths from light. When the application is less demanding than those of complex optical devices or systems, the glasses are commonly called 'absorbers' rather than filters. These applications usually involve protection of humans or materials from damage by light, for example, sunglasses and goggles for welders and glassworkers, or to enhance the environment by controlling the entrance or exit of ultraviolet and infrared light. Most of these applications are variations on the filters discussed above.

The most common use of glass as an absorber for both protection from damage and enhancement of the environment is found in the soda-lime-silicate glasses used as windows in buildings and automobiles. The presence of a relatively large concentration of iron oxide provides absorption in both the ultraviolet and infrared regions of the spectrum. Since glasses containing iron do absorb some light in the visible, a compromise between total absorption of unwanted, nonvisible light and the desired visible light must be made. As a result, these glasses usually contain between 0.5 and 2 wt% iron oxide. Changes in melting conditions which alter the ratio of ferrous to ferric ions in the glass are used to tune the glass for specific applications. While these glasses are technically broad bandpass filters, their common applications are based on the favorable location and bandwidth of their transmission window, which roughly corresponds to the visible region.

Iron is also used for protection from ultraviolet and infrared radiation in welder goggles. These glasses contain much more iron (up to 8 wt% or more) than those used for windows, since the demand for protection from the intense light created in welding is much greater. The high absorption coefficient for infrared light results in considerable heating of the glass during lengthy use. A thin film of gold is often applied to the glass to reflect infrared radiation and reduce heating of the glass.

Other absorber applications are more demanding in terms of absorption wavelength. As mentioned earlier, glasses containing neodymium have a strong absorption band at the wavelength of the yellow sodium emission line. These glasses transmit visible light at other wavelengths, allowing their use by glassblowers, who need to be able to see the unheated part of the glass and their immediate surrounding area, while simultaneously reducing the intense yellow light given off by the sodium in the glasses normally used for glassworking.

Protection of the eye from laser radiation is even more demanding due to the very high intensity over very narrow wavelength ranges. Common ultraviolet longpass filters are usually satisfactory for absorption of light from ultraviolet lasers. Specific ions must be added to glasses to provide protection from many

visible and near infrared lasers, since the glass must simultaneously permit transmission of visible light of other wavelengths and total absorption at the laser wavelengths. Light from lasers which operate in the far infrared is readily absorbed by most glasses, since the infrared cutoff of oxide glasses lies around 5 micrometers. Great caution should always be used in selecting glasses for eye protection from lasers. No assumptions regarding the degree of protection should ever be made without insuring that the glass is designed for the conditions under which it will be used.

See also

Optical Coatings: Thin-Film Optical Coatings. **Optical Materials:** Optical Glasses.

Further Reading

Bamford CR (1977) *Colour Generation and Control in Glass*. Amsterdam: Elsevier Scientific Publishing Company.

Deeg EW (1986) Optical glasses. In: Boyd DC and MacDowell JF (eds). *Commercial Glasses*, pp. 1–34. Columbus, OH: The American Ceramic Society, Inc.

Edmund Industrial Optics Catalog N017 (2001) pp. 68–87. Barrington, NJ: Edmund Industrial Optics.

Hillmann E and Neuroth N (1998) Glasses for eye protection. In: Bach H and Neuroth N (eds). *The Properties of Optical Glass*, pp. 370–379. Heidelberg, Germany: Springer-Verlag.

Kolberg U (1998) Coloured glasses. In: Bach H and Neuroth N (eds). *The Properties of Optical Glass*, pp. 350–370. Heidelberg, Germany: Springer-Verlag.

Neuroth N (1998) Transmission and reflection. In: Bach H and Neuroth N (eds). *The Properties of Optical Glass*, pp. 82–95. Heidelberg, Germany: Springer-Verlag.

Paul A (1982) *Chemistry of Glasses*, pp. 204–276. London: Chapman and Hall.

Vogel W (1991) Optical properties of oxide glasses. In: Ullmann DR and Kreidl NJ (eds). *Optical Properties of Glass*, pp. 1–36. Westerville, OH: The American Ceramic Society, Inc.

Weyl WA (1951) *Coloured Glasses*. Society of Glass Technology: Sheffield, England.

Heterogeneous Materials

U Kreibig and M Quinten, Aachen University of Technology, Aachen, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The term heterogeneous materials – being a quite diffuse classification of matter with discrete building units of different kinds, irrespective of their sizes – is defined here by scaling these inhomogeneities in relation to optical light wavelengths. Matter consisting of atoms in a regular structural order (or, alternatively, in fully statistical disorder) is classified as homogeneous since the inhomogeneities are too small to effect optical fields. This applies to crystalline, liquid, and glassy or amorphous matter likewise.

All geometrical/topological and chemical heterogeneities on a larger length-scale than the atomic scale, i.e., single crystallites, organic macromolecules, domains, extended lattice defects, clusters, and nanoparticles of chemical compositions differing from their host material, etc., result in a macroscopic material that is heterogeneous, in optical properties and optical experiments. Examples are multilayer-structures, heterogeneous in one dimension, three-dimensional heterogeneous composites, i.e., isometric building units distributed in some homogeneous host (photographic systems, drugs, cosmetics, color pigments in binders, etc.), and three-dimensional heterogeneous densely packed systems (poly- or nanocrystalline materials, ceramics, or nanoceramics, etc.).

The materials most intensively investigated, in the field of nanoscience, represent a subclass of heterogeneous materials.

Most of the common materials in our everyday environment have heterogeneous structures, often on larger characteristic size-scales. More generally, heterogeneous materials are so widespread that they are much more frequent than homogeneous materials. Some more examples, relevant for their optical properties, may illustrate this: pigment colors, photovoltaic composites, aerosol-systems, hydrosol-systems, many minerals, photonic crystals, island film devices, nanosensors, special biosystems, and polymer-systems. Most interesting optical features arise when these inhomogeneities are below or of the order of light wavelengths. Recent research has concentrated on a narrow size region on the lower nanometer scale where many-atoms-systems change from molecular to solid-state behavior with increasing sizes. This 'transition' is accompanied by drastic changes of almost all physical and chemical

properties, hence these properties develop strong dependences on particle size in that size region, and materials with uncommon properties can be developed.

In contrast to the widespread applications and research perspectives, the basic theoretical understanding of heterogeneous material is generally still poor. One reason is that, since the beginning of the development of optics as a science, the homogeneous materials like glass, water, etc. were preferentially treated since they were easier to describe theoretically. Another reason is the extremely wide variability of existing heterogeneous structures usually based upon low symmetries. This variability is too large, and the resulting optical properties too complex and multifarious to enable a general and uniform description of heterogeneous matter on a quantitative level, like the one successfully developed for the much simpler topology of homogeneous crystals.

Qualitative general features, however, can be obtained from investigations of model systems, with especially simplified topological structures. In the following we describe the model system of nanoparticle/host-matter (or, synonymously, cluster-matter), which is based upon nanoparticles/clusters packed closely in a host material, thus forming macroscopic or microscopic heterogeneous many-particle systems. Even this model system exhibits broad variability of topologies and of chemical compositions, as shown in Figure 1.

The restriction in the present contribution to this topologically simple model system appears to be justified by the fact that the optical properties discussed below for this system are also qualitatively characteristic of other arbitrary heterogeneous materials. We also admit that the nanos may consist of a nucleus and a surrounding shell of different structure and/or different material.

Heterogeneous matter can be defined by its heterogeneous building units and by the borders of these units and by the host material and the borders between building units and host, i.e., the internal interfaces.

It is a general feature that structures are visible because of property changes at their terminating borders: borders create structures, and structures require borders. The more refined the structures are, the more increases the topological extent of the borders. This means that when sizes of building units, for example, in cluster matter decrease, the relative amount and the effects of surface/interface regions increase. Such interfaces in heterogeneous matter are not two-dimensional (as suggested by the term 'interface') but consist of three-dimensional

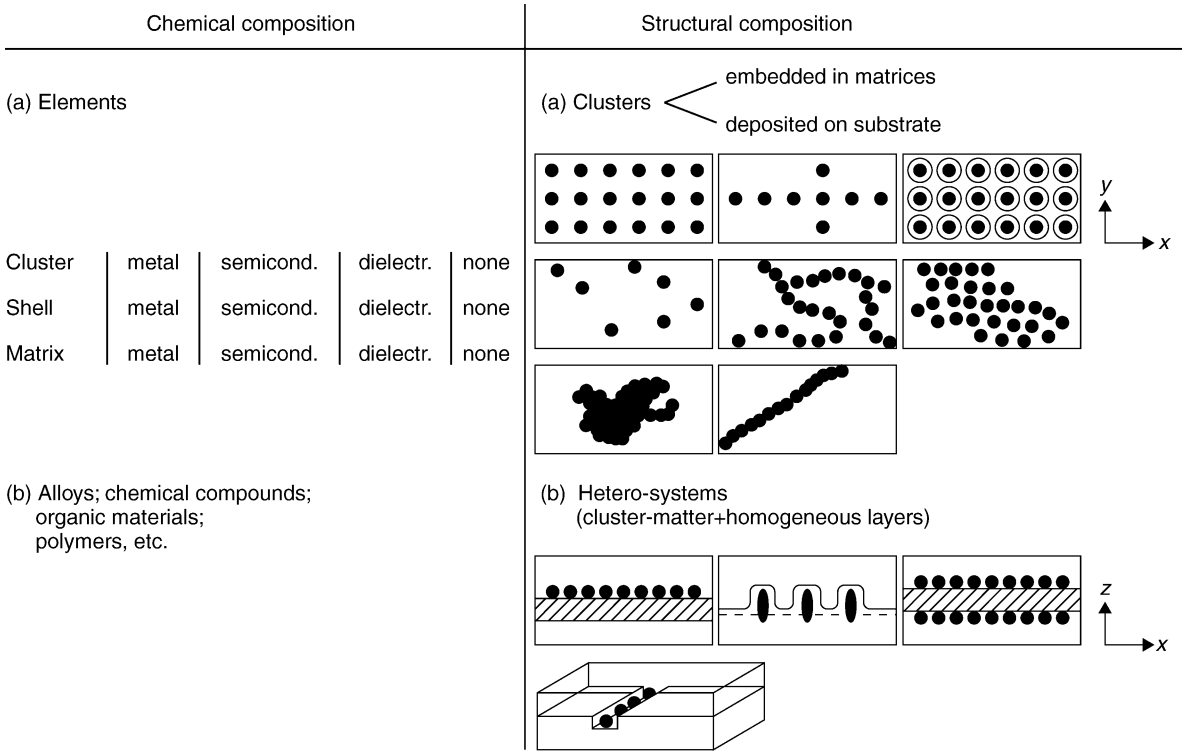


Figure 1 Examples of heterogeneous nanoparticle/host materials. Left: possible material combinations of the three building units: cluster, shell, matrix. Right: some examples of sample topologies.

interlayers, the atomic and electronic structure and the chemical composition of which are usually in complex disorder. In the case of very small particles the interface volume can easily exceed the volume of the building units themselves. There are, however, also examples where the atomic arrangement forms regular interfaces, such as at coherent grain boundaries.

Hence, heterogeneous materials are characterized by various building units, homogeneous matrices and the interface regions. To describe them, we will focus on our introduced model system of nanoparticle/host matter, where we further specialize to spherical shapes of the particles.

The investigation and description of optical properties of heterogeneous materials have to deal with the interior of the building units, e.g., characteristic optical size effects, and with the interfaces, e.g., optical interface effects. The importance of both increase with decreasing characterizing structural lengths. The role of the host materials, when present, can be different, depending on their concentration: in dilute particle systems they may act as bulk-like homogeneous matter (e.g., the electrolyte in colloidal liquid systems); at low amounts, the host may also develop size effects (e.g., the water layers in the opal).

General Optical Properties

Among the experimental characterization methods most frequently applied to heterogeneous matter, the optical ones are essential, ranging from spectroscopy in the frequency domain (from FIR to UV), to femtosecond analysis in the time domain.

Optics can mean physical optics, i.e., the science of optical material properties, including light emission (the 'active' properties) and light absorption (the 'passive' properties), and instrumental optics, i.e., special experimental techniques based upon propagation of light to analyze such samples.

The most usual experiments for the investigation of linear optical response are refraction, absorption, reflection, and elastic scattering. Also, there are the emission responses and the broad field of nonlinear optical responses.

Refraction, reflection, and absorption are typical of homogeneous matter. Elastic light scattering, on the other hand, can be used to define optical heterogeneous matter, since elastic scattering is a direct consequence of the existence of interior interfaces forming deviations from the homogeneous structure. While the grainy structure on the atomic level in homogeneous matter does not give rise to disturbances of the shape of a plane interacting optical wave, this does occur with heterogeneous matter. Each inhomogeneity gives rise to a wavefront deformation and, hence, to light scattering. Scattering spectra and intensities provide important information about the heterogeneous material, so such optical methods are frequently used for its identification.

Besides scattering, the electrical charging of the interior interfaces induced by the light field, can lead to spectacular multipolar optical absorption features, unseen in related macroscopic homogeneous materials, such as Mie resonance, discussed below.

The optical properties of our model system are composed of various contributions which all act together in a complex way, which can be described as:

1. the optical properties of the single building units. In the case of solid state units, they are characterized by the optical material properties, frequency dependent and, generally, complex dielectric function $\varepsilon(\omega)$ or, alternatively, the (generally complex) refraction index $n(\omega)$, which, in the nanometer-size region, both develop strong variations with size.
2. the optical interactions among all building units in the sample. There are different effects:
   - interference effects among the scattering radiation of the units;
   - electromagnetic coupling effects;
   - the contributions of host materials, if present; and
   - the interactions between heterogeneous building units and host.

The Optical Nanoparticle Resonances (Mie Resonances)

Metallic and semiconductor nanoparticles can exhibit special, spectrally selective absorption and scattering bands which do not have counterparts in the molecular state, in thin film structures, or in the extended bulk. They represent the most spectacular optical properties of matter with nanosized heterogeneities. Therefore they will now be treated in more detail in the frame of our model material of spherical inclusions in a homogeneous embedding medium. We begin with a comparison to thin film structures.

The interband transitions (in Ag below 325 nm) are almost equal in both spectra shown in Figure 2, of a thin Ag film and spherical Ag nanoparticles, respectively, while above 325 nm the flat relaxator spectrum of the conduction electrons in the metal film differs strongly from the oscillator spectrum of the same electrons in the clusters. The inset demonstrates the excitation of the dipolar surface oscillator mode by

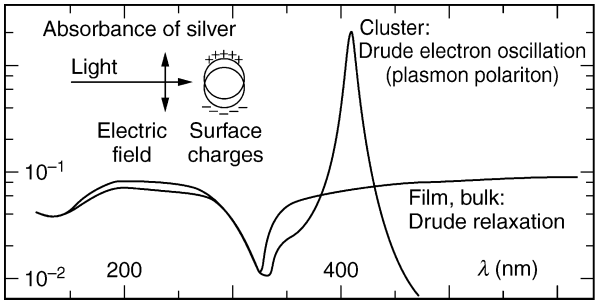


Figure 2 Comparison of absorbances of a thin Ag film and spherical small Ag nanoparticles.

incident polarized light. This figure exemplifies the above-mentioned effect that charging of interfaces by the incident electromagnetic lightwave can give rise to electromagnetic multipole resonances, the Mie-resonances.

The monotonic conduction electron absorption spectrum of the thin film, ranging from UV to zero frequency (the dc current), is thereby compressed in the nanoparticles into the narrow Mie-absorption band with peak position in the blue, containing almost the whole oscillator strength of the conduction electrons. In particular, the Ag-spheres are therefore transparent in the low frequency region, and if we assume the conductivity according to the Drude model as being definitive for a metal, we must assume that these particles are no longer metallic. The conduction electrons no longer follow the Drude relaxator behavior; instead the narrow peak indicates the properties of an oscillator.

The origin of this oscillator and its excitation can be described, being a quantum phenomenon; due to the incident electrical field, the electron plasma, the 'electronic particle', is shifted against the fixed ionic background, the 'ionic particle', causing surface charges modulated with the light frequency. The surface charges produce restoring forces to the plasma, following the periodic field strength, which is the condition required for oscillator behavior. As every free oscillator the particle develops resonances at special eigenfrequencies, and in the case of our metal clusters these are close to the peak frequency of the Mie-absorption, the oscillations of which not being free but forced by the incident field. Analogous to the mechanical case, the resulting resonance frequencies of free and forced oscillations differ slightly. To distinguish them, Mie resonances in metallic particles, coupled to the external electromagnetic field, are called plasmon polaritons.

In the frame of classical electrodynamics each single particle in the heterogeneous material acts as a spherical nano-antenna. This acts as receiving antenna by absorption and, as with every other kinds of antenna, simultaneously becomes a transmitting antenna by re-radiating elastically scattered light.

In the case of Cu, Ag, and Au, and many other metals, these absorption frequencies are in the visible, hence beautiful colors appear. However, there are a lot of metals where this oscillator is overdamped due to low electron mobility. Under particular conditions, such resonances also occur in semiconductor nanoparticles, but due to the lower charge densities, in the IR. At weak fields, the optical response of the oscillator is linear. At higher fields it becomes nonlinear, and thus second and higher harmonics can be excited.

Such unique collective excitations are not restricted to electrons but also spherical surface phonon polaritons can be excited if the particles are nonmetallic (in this case, free electron excitations do not cover the spectral region of phonon excitations). Only in the most simple case of particles which are extremely small compared to wavelength ($R/\lambda \leq 1/20$), the dipolar mode of excitation described so far (i.e., homogeneous polarization of the whole particle) comprises the whole response. For larger particle sizes, a complicated, nonhomogeneous polarization of the whole particle occurs, which can be described by an orthogonal multipole expansion, i.e., multipolar oscillations. The according size (R-) dependent multipolar contributions, which are excited simultaneously and sum up to the total optical response. Further complications arise in larger particles by retardation-induced phase shifts of the lightwaves along the particles.

A complete description of this complex multipole excitation was given by Gustav Mie on the basis of classical electrodynamics. Mie's theory is the exact solution within the frame of Maxwell theory for spherical symmetry, as are the Fresnel-formulae for the planar geometry. However Mie's theory does not describe exactly the conditions which we observe in real heterogeneous systems, hence several extensions of Mie's theory will be presented in the section entitled *Beyond Mie's Theory,* below.

Mie's theory also gives some qualitative insight into the optical behaviour of other kinds of heterogeneous matter. A highly abridged summary of this theory will be presented below.

The Theory of Mie

In the following, an outline of the Mie theory for one spherical particle is given by listing the mathematical steps:

1. Introduction of spherical coordinates r, θ, ϕ. Particle radius R;

2. Incidence of a plane, monochromatic, harmonic electromagnetic wave (circular frequency ω, wavelength λ);
3. Solution of Maxwell's equations for the incident and scattered fields in the surrounding dielectric host medium with (real) dielectric constant $\varepsilon_{\text{matrix}}$, and in the interior of an arbitrary spherical particle of (complex) dielectric function $\varepsilon(\omega)$.

   There are two solutions for transverse fields: one with the radial component of the electric field $E_r = 0$, called 'transverse electric (TE) solution', and one with the radial component of the magnetic field $H_r = 0$, called 'transverse magnetic (TM) solution'. These solutions can be obtained in terms of vector spherical harmonics which can be derived from corresponding scalar potentials Π_E and Π_H. These are solutions of the Helmholtz-equation:

$$\nabla^2\Pi_{E,H} + K^2_{\text{transverse}} \times \Pi_{E,H} = 0, \qquad K^2_{\text{transverse}} = \varepsilon_{\text{transverse}}\omega^2/c^2 \qquad [1a]$$

   In spherical coordinates, the Helmholtz equation is solved by a separation of variables ansatz and a multipole expansion, yielding an infinite number of independent partial solutions with multipolar order L, the TM partial waves or 'electric partial waves', and the TE partial waves or 'magnetic partial waves'.

   A third solution, the potential Γ, that only affects the electric field, can be found as the solution of

$$\nabla^2\Gamma = 0, \qquad \varepsilon_{\text{longitudinal}} = 0 \qquad [1b]$$

   and represents longitudinal excitations in the particle.
4. Boundary conditions at $r = R$:

   for the transverse fields:

$$E_\theta^{\text{incident}} + E_\theta^{\text{scattered}} = E_\theta^{\text{interior}}$$
$$E_\phi^{\text{incident}} + E_\phi^{\text{scattered}} = E_\phi^{\text{interior}} \qquad [2a]$$

$$H_\theta^{\text{incident}} + H_\theta^{\text{scattered}} = H_\theta^{\text{interior}}$$
$$H_\phi^{\text{incident}} + H_\phi^{\text{scattered}} = H_\phi^{\text{interior}} \qquad [2b]$$

   for the longitudinal fields:

$$E_r^{\text{incident}} + E_r^{\text{scattered}} = E_r^{\text{interior}} \qquad [3]$$

   (Sauter–Forstmann condition)

   The linear response function of the particle material used here, is ε, where $\varepsilon = \varepsilon_1 + i\varepsilon_2$ is averaged over the particle volume, including the surface–interface region.

   The magnetic permeability μ is set 1 for the investigated high frequency regions.
5. Computation of cross-sections for extinction, scattering, and absorption from Poynting's law:

   - Extinction, i.e., the sum of absorption and scattering losses:

$$\sigma_{\text{ext}} = \frac{2\pi}{k^2}\sum_{L=1}^{\infty}(2L+1)\text{Re}\{a_L + b_L\} \qquad [4]$$

   - Elastic scattering:

$$\sigma_{\text{sca}} = \frac{2\pi}{k^2}\sum_{L=1}^{\infty}(2L+1)(|a_L|^2 + |b_L|^2) \qquad [5]$$

   - Absorption, i.e., the energy dissipation (production of heat and thermal radiation):

$$\sigma_{\text{abs}} = \sigma_{\text{ext}} - \sigma_{\text{sca}} \qquad [6]$$

   L indicates the multipolar order of the Mie resonances.

The coefficients a_L and b_L, following from the Maxwell boundary conditions for transverse fields, are

$$a_L = \frac{m\psi_L(mx)\psi'_L(x) - \psi'_L(mx)\psi_L(x)}{m\psi_L(mx)\eta'_L(x) - \psi'_L(mx)\eta_L(x)} \qquad [7a]$$

$$b_L = \frac{\psi_L(mx)\psi'_L(x) - m\psi'_L(mx)\psi_L(x)}{\psi_L(mx)\eta'_L(x) - m\psi'_L(mx)\eta_L(x)} \qquad [7b]$$

with $m = 2x/n_{\text{Matrix}}$, where m denotes the complex-valued index of refraction of the particle $\varepsilon = n'^2$ and n_{matrix} the real-valued index of refraction of surrounding medium. k is the wavevector and $x = k{\cdot}R$, the size parameter. $\Psi_L(z)$ and $\eta_L(z)$ are Riccati–Bessel functions of the first and third kind. The prime indicates differentiation with respect to the argument in parentheses. Including longitudinal fields, eqn [7] will be slightly modified.

Applying Lambert–Beer's law, these cross-sections are related to the absorption-, scattering-, and extinction-constant $A(\omega)$, $S(\omega)$, and $E(\omega)$, respectively, and the resulting intensity loss ΔI of a many-particle system of number-density Z of the (noninteracting) particles, is

$$\Delta I = I_0(1 - \exp(-Z{\cdot}\sigma_{\text{ext}}{\cdot}d)), \qquad d = \text{sample thickness} \qquad [8]$$

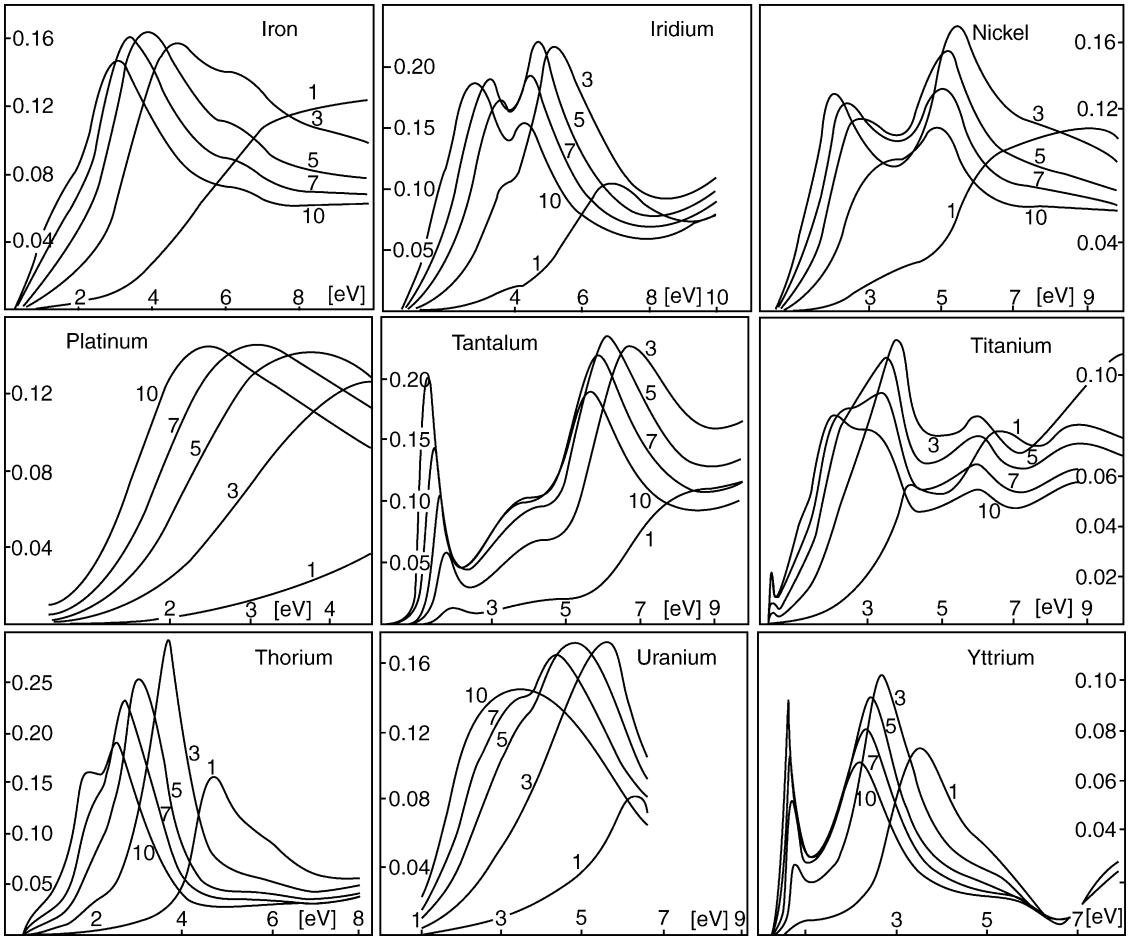


Figure 3 Mie extinction spectra of various particle materials. $2R = 10$ nm. The clusters are embedded in matrix materials with dielectric constants between 1 and 10, which are indicated for each spectrum. Obviously, the peak structure of the spectra strongly depends on the matrix.

Figure 3 gives a compilation of some spectra of general interest determined by evaluating Mie's theory.

Beyond Mie's Theory

Application of Mie's theory to experimental results is disappointing because, in most cases, quantitative coincidence is not obtained. One reason is that the simple assumptions incorporated either in Maxwell's theory or in Mie's derivation are not met by reality.

Mie's theory does not apply, if:

1. particle sizes in a diluted cluster matter sample are not uniform;
2. particle shapes differ from the sphere;
3. $\varepsilon_{\text{matrix}}$ varies with frequency, $\varepsilon_{\text{matrix}}(\omega)$
4. cluster-dependent dielectric functions differ from the according functions of the bulk;
5. the (step-function like) Maxwellian boundary conditions at the particle-matrix interface are not applicable;
6. dielectric functions of particles and/or matrix are nonlinear;
7. surrounding/embedding matrix material is absorbing;
8. particle structures are heterogeneous (core-shell structures, multi-grain clusters, etc); and
9. particles form close-packed aggregates instead of being well-separated, and hence, electromagnetic coupling among particles is important.

Of course, this list may be continued. Regarding realistic samples, it is obvious already from these deviations that quantitative correspondence between experiment and Mie theory usually cannot be expected. But, over time, each of these extra effects has been treated and included into extensions of Mie's

theory, which should therefore be applied, instead of the original theory. In the following several of these effects will be treated in some detail.

Particle Shapes Deviating From the Sphere

Even metals of high structure-symmetry, like the fcc-Au, can yield nanoparticles with low symmetry habits. Figure 4 shows Au nanoparticles with exotic shapes ranging from triangle and hexagonal platelets over fivefold symmetry multitwin-particles to rodlike shapes. The majority of particles are multigrain structures of irregular isometric shapes.

Multishell Particles

In experiments, particle surfaces often react chemically with the surrounding media, thereby forming compound shells. Core/shell structures can be clearly visible in the optical absorption spectra. Figure 5 shows, as an example, the absorption spectra of nanoparticles of Ag core/Au shell and of Au core/Ag shell, respectively with varied shell thicknesses.

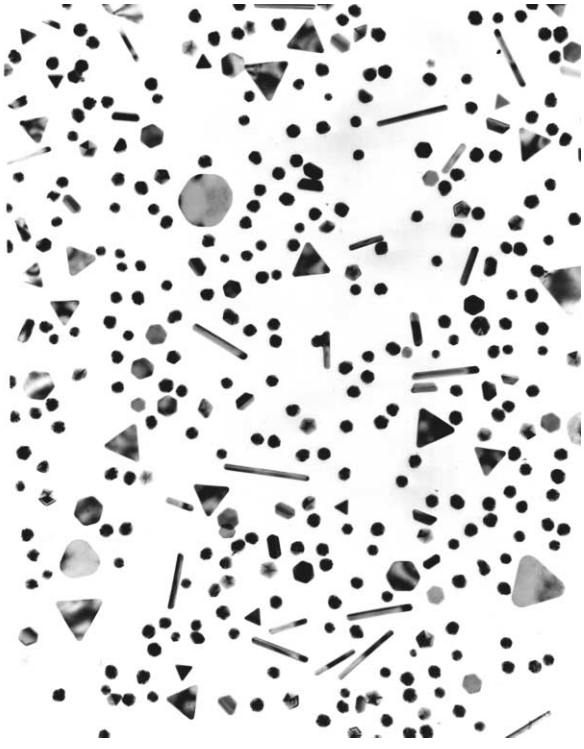


Figure 4 Chemically (Esiguondy method) produced Au clusters with extraordinary variety of different shapes. (Of special interest are the small particles of five-fold symmetry.) Mean sizes of the irregular, almost spherical particles: 36 nm.

Resonances will also occur, e.g., in metal-coated dielectric particles. An example is shown in Figure 6, with extinction spectra of silver-coated titanium-dioxide particles, computed by an extension of Mie's theory.

Aggregates of Nanoparticles with Electrodynamic Particle–Particle Coupling

The optical extinction and scattering spectra of many-particle systems with dense packing forming aggregates (see Figure 4) are strongly influenced by electromagnetic interactions: each particle feels the

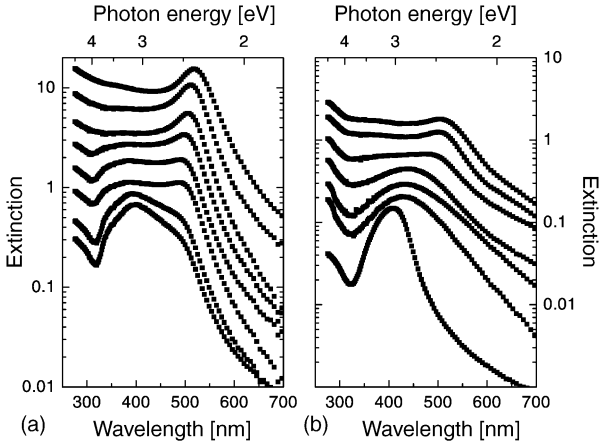


Figure 5 Measured extinction spectra of (a) silver-coated gold particles (gold core $2R = 17$ nm, shell thickness varying from $d = 0$ to $d = 3.6$ nm from top to bottom), and (b) gold-coated silver particles (silver core $2R = 18$ nm, shell thickness varying from $d = 0$ to $d = 5$ nm from bottom to top).

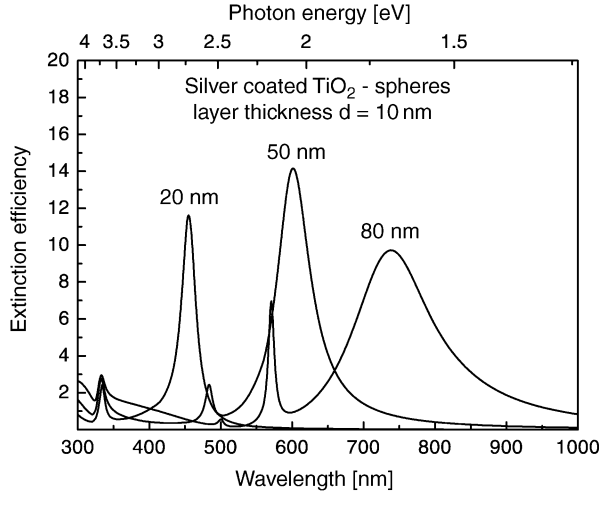


Figure 6 Calculated extinction spectra of spherical silver-coated titanium dioxide particles. The shell thickness is kept constant at $d = 10$ nm, the core sizes amount to $2R = 30$ nm, 50 nm, and 80 nm, respectively, as indicated in the graph.

scattering near fields of all its neighbors, which give rise to splitting, broadening, and shifts of resonances. This is clearly demonstrated with colloidal Ag and Au particles in Figure 7.

The splitting, broadening, and shifts can be assigned to the shape and size of the resulting many-particle aggregates. An example for the influence of the topology of the many-particle system is given in Figure 8 with extinction and scattering spectra of silver particle aggregates of $N = 5$ identical silver particles with $2R = 40$ nm, which were computed by the Generalized Mie Theory (GMT). As a rough rule-of-thumb, such electrodynamic coupling effects of the oscillators only can be ignored if the smallest next neighbor center-to-center distance d_{cc} exceeds $d_{cc} \approx 5 \cdot R$.

Size Effects

The extraordinary practical success of Mie's theory is based upon the fact that this example of classical electrodynamics enables us to compute numerically the optical response for arbitrary realistic particle materials. In contrast quantum theoretical calculations are restricted, beyond the molecule-solid state transition, to jellium and jellium-related hypothetical material.

Classical electrodynamics, being a phenomenological theory to describe light propagation, does not yield information about optical material properties. These enter via the dielectric functions inserted into the Maxwell boundary conditions. They must be taken from elsewhere, such as from experiments or from quantum solid state theory model calculations. This twofold basis of Mie's theory is illustrated in Figure 9.

There are two different sources for particle size dependences of the optical extinction spectra:

1. The electrodynamics: the larger the particles, the larger are the number of contributing multipoles and the influences of retardation effects, both influencing strongly the optical spectra;
2. The optical material properties describing the particle material which, in the nanoscale size region, exhibit size dependences. These size dependences can be treated by corrections entering the bulk material function.

In general, the dielectric function $\varepsilon(\omega)$ of a metal or semiconductor is the sum of susceptibility contributions of the lattice, of the conduction electrons, and of the interband (electron–hole) excitations:

$$\begin{aligned}\varepsilon(\omega) &= \varepsilon_1(\omega) + i\varepsilon_2(\omega) \\ &= 1 + \chi^{\text{lattice}}(\omega) + \chi^{\text{free electrons}}(\omega) \\ &\quad + \chi^{\text{interband}}(\omega)\end{aligned} \qquad [9]$$

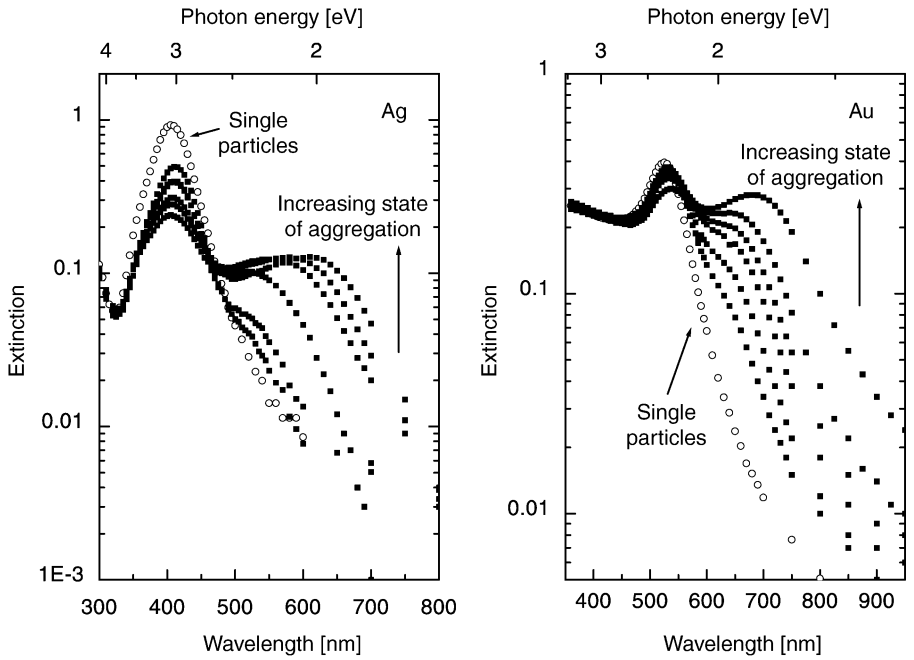


Figure 7 Measured extinction spectra of aggregated samples of (a) silver particles with $2R = 28$ nm, and (b) gold particles with $2R = 38$ nm. The samples contain random distributions of aggregates of various topologies, but with increasing state of aggregation from bottom to top.

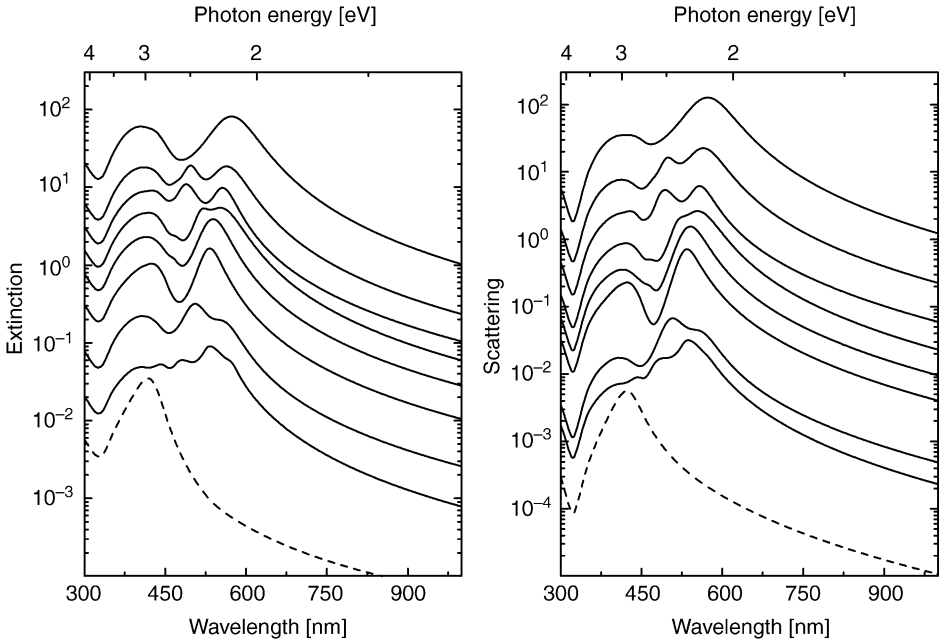


Figure 8 Calculated extinction and scattering spectra of aggregates with $N = 5$ almost touching identical silver spheres of $2R =$ 40 nm. From top to bottom, the spectra belong to the following particle arrangements in the aggregate: ▪▪▪▪▪ ▪▪▪▪ ▪▪▪ ▪▪ ▪ ▪▪▪ ▪▪▪. The dashed line represents the single particle spectra.

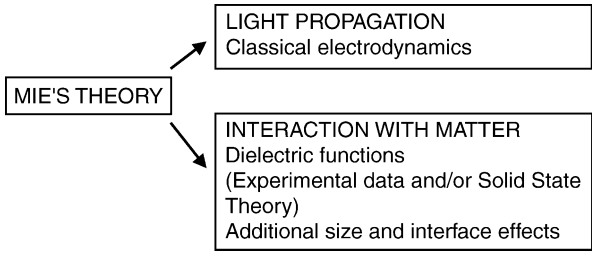


Figure 9 Scheme of the basics of Mie's theory.

A detailed expression was given by Bassani (here, the lattice contributions, mainly important in the IR, were disregarded):

$$\varepsilon(\omega) = 1 - \underbrace{\frac{(ne_0^2/\varepsilon_0 m_{\text{eff}})}{\omega^2 + i\gamma\omega}}_{\text{conduction electrons}} + \underbrace{\chi^{\text{interband}}(\omega)}_{\text{interband transitions}} \qquad [10]$$

with

$$\chi^{\text{interband}} = \frac{2\hbar^2 e^2}{\varepsilon_0 m_{\text{eff}}^2} \sum_{i,f} \int_{\text{BZ}} \frac{2d^3k}{(2\pi)^3} |\mathbf{e} \times M_{\text{if}}(\mathbf{k})|^2 \times \left\{ \frac{1}{[E_f(\mathbf{k}) - E_i(\mathbf{k})][(E_f(\mathbf{k}) - E_i(\mathbf{k}))^2 - \hbar^2\omega^2]} + i\frac{\pi}{2\hbar^2\omega^2} \delta(E_f(\mathbf{k}) - E_i(\mathbf{k}) - \hbar\omega) \right\} \qquad [11]$$

In this equation, the quantity n is the conduction electron density, $\mathbf{k}$ the wavevector of the electrons with norm k, and BZ denotes the Brillouin Zone. The vector $\mathbf{e}$ is the polarization vector of the incident light. M_{if} is the transition matrix element with i = initial and f = final states.

To which degree of approximation a single dielectric function can describe the optical response of a nanoparticle consisting of inner core and surface with differing local optical properties (i.e., polarizabilities), and how this function has to be modelled as an average, is still a problem. An easy, but only approximate answer is to identify terms in the eqns [10] and [11], which may change if the cluster size is reduced down to the nanometer scale, and to introduce correction terms.

In the case of the conduction electrons of eqn [10], these are the electron density n, the effective electron mass m_{eff}, and the relaxation frequency γ. In the terms of the interband transitions (index inter), these are the transition matrix elements M_{if} and the band structure energies of the initial and final states E_i and E_f. All of these quantities change, when the particle size is reduced, in different size regions and to different amounts, and detailed quantum theoretical investigations are required to derive proper correction terms.

The application of the continuum-based dielectric functions and of Mie's theory fail when the molecule-solid state transition region is reached.

The effects of size-dependent dielectric functions upon cluster absorption spectra are demonstrated in the series of absorption spectra of Figure 10, measured on highly diluted heterogeneous systems containing spherical Ag, Au, and Cu particles which were embedded in a glass matrix. If $\varepsilon(\omega)$ were independent of size, then the spectra should be identical, in the investigated particle size region. So, their differences reflect clearly that $\varepsilon = \varepsilon(R)$.

To describe the size effects of the conduction electrons, the relaxation frequency γ of eqn [10] has been modified introducing a parameter that is called A-parameter which also contains size effects and also particle–host interface effects. To explain the electronic size effects in detail, quantum effects have to be considered taking the discretization of the bandstructure into account, but they can also be treated by a simple classical approach: this is the mean free path effect (FPE), assuming the mean flight time of the effective conduction electrons (i.e., those with Fermi energy E_F) between subsequent surface collisions amount to

$$\tau\gamma_{FPE} = A_{FPE}\frac{v_{Fermi}}{R} \qquad [12]$$

The A-parameter is a measure of the effectiveness of the single collision in view of relaxation. Complete 'memory loss' of the drift momentum in each collision would lead to $A_{FPE} = 1$. Taking electron correlation effects and energy relaxation into account, A-values larger than 1 may also be realistic. Values larger than unity also occur in multigrain particles, due to grain boundary 'collisions'.

Various quantum mechanical models of electronic particle size effects result in relations, formally coinciding with the $1/R$-dependence in eqn [12]. Most relevant is the value of A for particles *in vacuo* with clean surface. This value has been determined theoretically and experimentally to $A \approx 0.25$ for Ag particles.

Properties of Internal Interfaces

Structures of Surfaces/Interfaces

Only mathematical interfaces in geometry can, in fact, be two-dimensional planes separating two different volume elements. This is assumed in the frame of Maxwell boundary conditions entering Mie's theory. In real nanostructured matter, however, the interface regions are three-dimensional layers formed by disturbances of atomic and/or of electronic structures, and/or of chemical compositions. In addition, due to the finite wavelengths of the conduction electrons, the according electron density 'spills out' and, as a consequence, we find electron density beyond the geometric surface. It should be kept in mind that these effects hold, with proper changes, for interfaces in nanostructured materials in general. Hence, an understanding of the interface regions is of key importance for all nano-material.

Electronic Interface States

Up to now, this discussion has been restricted to particles with uncovered, 'clean' surfaces. In most practical examples of heterogeneous materials however, the building units are embedded in a host matrix, deposited upon the surface of a substrate, or are in close contact to neighboring crystallites, etc., to stabilize the samples. Then, special electronic

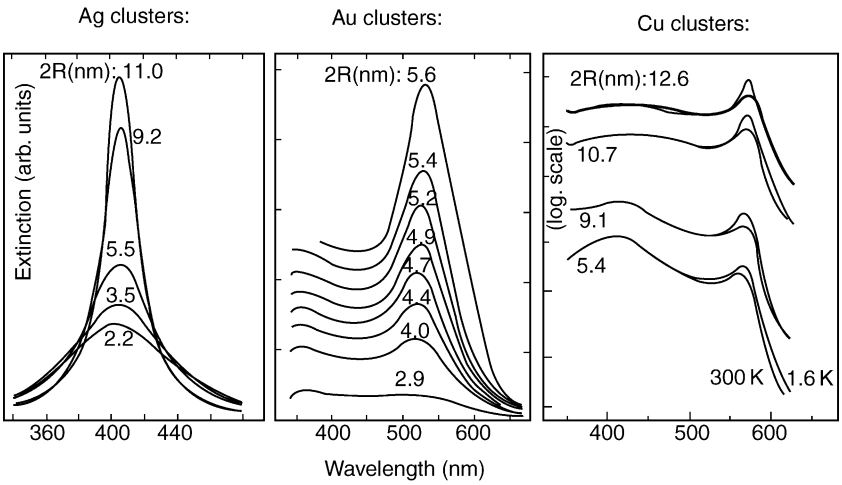


Figure 10 Measured extinction spectra of Ag, Au, and Cu clusters of various mean sizes, grown in glass matrix. In the case of Cu-particles, measurements at 1.5 K are added.

interface states are created by contributions of particle electrons and host electrons, which are (disregarding their density on the surface) similar to adsorbate states.

One might assume that the broad knowledge about these states in surface science, obtained from samples with extended crystallographic planes, should be very useful here. However, the surface of a cluster, in principle, differs from the planar surface of a single crystal by containing a multitude of different local surface structures, atomic sites, and varying atomic coordination numbers. Even if the surface structure is not irregular (as is usual in nanoparticles), the surface consists of several facets of different orientation separated by edges and corners. The coordination numbers of surface atoms, thus, vary strongly (e.g., between 5 at a corner and 9 on a plane), indicating surface sites to have different energies.

We have shown that, if the particles are metallic, the 'particle/host'-interface states give rise to two different electronic charge transfer processes of particle conduction electrons, the static and the dynamic charge transfer. This will be demonstrated qualitatively for metal particles by the example presented in Figure 11.

The Static Interface Charge Transfer and its Effects Upon Mie Resonance Positions

At the left side of Figure 11, the conduction band of the metal cluster close to its surface is shown, which is occupied with electrons up to the Fermi energy E_F. Chemisorption of a free atom (one energy level E_a of which is plotted at the right-hand side) is assumed to occur by approaching the atom to the cluster surface. The resulting energy states are also shown. Let us assume the marked electronic level E_a in the free atom is nonoccupied. During the chemisorption process, this level is shifted and broadened, developing density of nonoccupied states below the particle Fermi energy, which is localized at the three-dimensional interface region. After complete the process is, such a state will be occupied permanently by electrons of the cluster (static charge transfer), and the nanoparticle will have lost electrons, thus its electron density being reduced. The driving impact of this process is the equilibration of the chemical potential in the whole cluster-matrix system.

Now assume we embed a cluster into some surrounding material. Then all those atoms of the latter, which are situated directly at the interface, become chemisorbed at the cluster. Hence remarkable amounts of charges can be transferred into the interlayer region and an electrical double layer is created. As an example, a very strong amount of static charge transfer was observed experimentally when applying solid Fullerite (C_{60} molecules) as embedding matrix for Ag-nanoparticles: here, each C_{60} in direct contact with the Ag particle attracts one electron, forming (C_{60}^-). Hence, in this experiment with 500 atoms large particles, in total about 20% of the conduction electrons of the Ag-particle were transferred into the interlayer. Strictly speaking, such a nanoparticle of the heterogeneous material is no longer a silver particle because of the remarkable changes of electronic properties by static charge transfer.

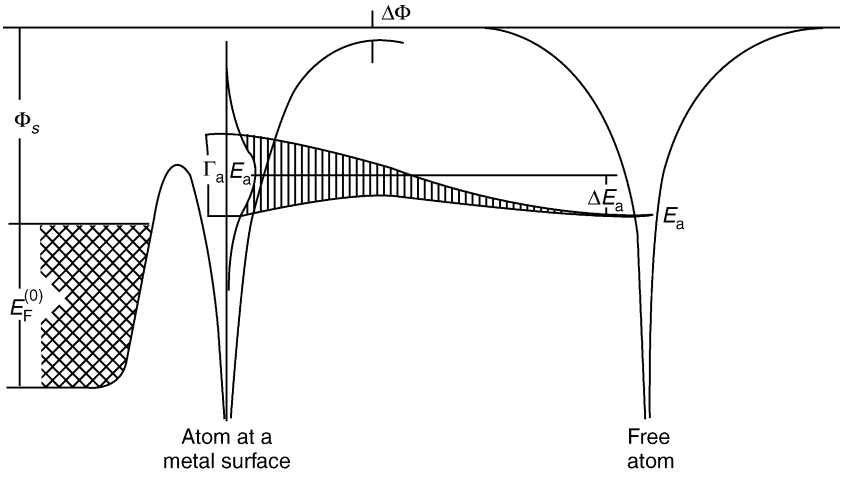


Figure 11 Electron energy scheme for atomic adsorbation at a metal particle surface. The free atom (energy spectrum: right) approaches the particle surface (left; conduction band with Fermi-level E_F). Its energy level E_a is thereby shifted and broadened (including ensemble broadening of many atoms deposited at many clusters). The atoms are assumed to be separated from the metal by a tunnel barrier (see text).

Some consequences are listed in the following:

In a first step, let us assume that the electronic bandstructure is not altered by the charge transfer. Then we expect: (i) changes of level occupation; (ii) changes of Fermi energy and Fermi velocity; (iii) changes of effective masses close to E_F; (iv) changes of surface 'spill out'; (v) additional adsorbate levels and special surface states; and (vi) changes of the optical interband transition edge.

In a second step, we may consider possible changes of electronic bandstructure (e.g., energies) and, finally, also changes of atomic structures (i.e., lattice structures and/or surface structures).

Since plasmon frequencies depend on the conduction electron density, the resulting Mie resonance can monitor and give quantitative information of this effect. We find shifted by the static charge transfer:

$$\Delta\omega_{\text{resonance}} \approx (\sqrt{n_2} - \sqrt{n_1}) \times \sqrt{\frac{e_0^2}{\varepsilon_0 m_{\text{eff}}(2\varepsilon_{\text{matrix}} + 1 + \chi_1^{\text{interband}})}} \quad [13]$$

with n_1 being the conduction electron density in the cluster *in vacuo*, and n_2 being the electron density after contact with foreign material (adsorbate, substrate, or host-matrix). This peak shift of the Mie resonance, however, is superimposed in the experiment by a much larger shift, following simply from the change of Maxwell boundary conditions by the altered dielectric constant of the particle surrounding $\varepsilon_{\text{matrix}}$, when going from vacuum to a matrix material. Hence, the experimental determination of static charge transfer processes is rendered more difficult.

The Dynamic Interface Charge Transfer and its Effects Upon Mie Resonance Widths

After the static charge transfer in Figure 11 has taken place, electrons close to E_F can fluctuate between the cluster and empty adsorbate levels, if present at E_F. The residence times in these levels, which are usually separated from the Fermi sea of the metal cluster by some tunnel barrier, vary statistically. If a plasmon is excited, i.e., mainly a collective, phase ordered motion of all conduction electrons superimposed to their statistical motion, we may assume that, after such electron transfer back and forth, the reminiscence to the primary drift momentum, and the phase of this excitation, has been lost. Thus, the phase relaxation of the plasmon is increased.

The consequences of this dynamic charge transfer effect ('chemical interface damping') have been investigated in detail. Here, we only state that the directly observable main consequence is the broadening of Mie resonances. If the particles are sufficiently small to neglect radiation damping (this is the case for Ag clusters smaller ≈ 15 nm), band broadening is only due to internal energy dissipation and the band width can be modeled quantitatively with the modified relaxation frequency γ in eqn [10] which yields an additive correction term:

$$\Gamma(A, R) \approx \Gamma_0 + \frac{2ne_0^2}{\varepsilon_0 m_{\text{eff}}\omega^3} \times \frac{A}{\sqrt{(\partial\varepsilon_1/\partial\omega)^2 + (\partial\varepsilon_2/\partial\omega)^2}} \frac{v_{\text{Fermi}}}{R} \quad [14]$$

with the A-parameter extended to the additional electronic excitations in the particle/host interlayer:

$$A = A_{\text{size effects}} + A_{\text{interface effects}} \quad [15]$$

The A-parameter of eqn [15] now expresses both the strengths of the particle size effects and of the chemical interface damping, due to statistically varying occupation of interface states near E_F.

An important feature of this model is that the dependence on the kinds of materials is described by specific material parameters ($A_{\text{size effects}}$ and $A_{\text{interface effects}}$), and the direct size dependence effects are given by an $1/R$-dependence, for both the direct size effects and the interface effects. For one fixed cluster size R, the magnitude of $A_{\text{interface effects}}$ directly characterizes the probability of the single temporal charge transfer, including the number and energies of contributing interface states; hence, important properties of the adsorption process can be investigated by determining the A-parameter.

Experimental Results

In the experiments presented in the following, we deal with the interfaces between clusters and a more or less homogeneous solid embedding medium. We selected very small silver nanoparticles of about 2 to 4 nm size with accordingly large ratio of surface- to volume-atoms, in order to optimize the sensitivity of the measurements. Silver is unique in having the most pronounced Mie resonance of all known materials.

In Figure 12 two classes of matrix materials are presented, fluorides and oxides. The absorption spectrum of the free particles with uncovered 'clean' surface (measured in a cluster beam in UHV) is added in both diagrams which is only influenced by size effect. General features are, that embedding in a host medium causes:

- drastic Mie resonance peak shifts toward IR, depending on the dielectric constant of the matrix

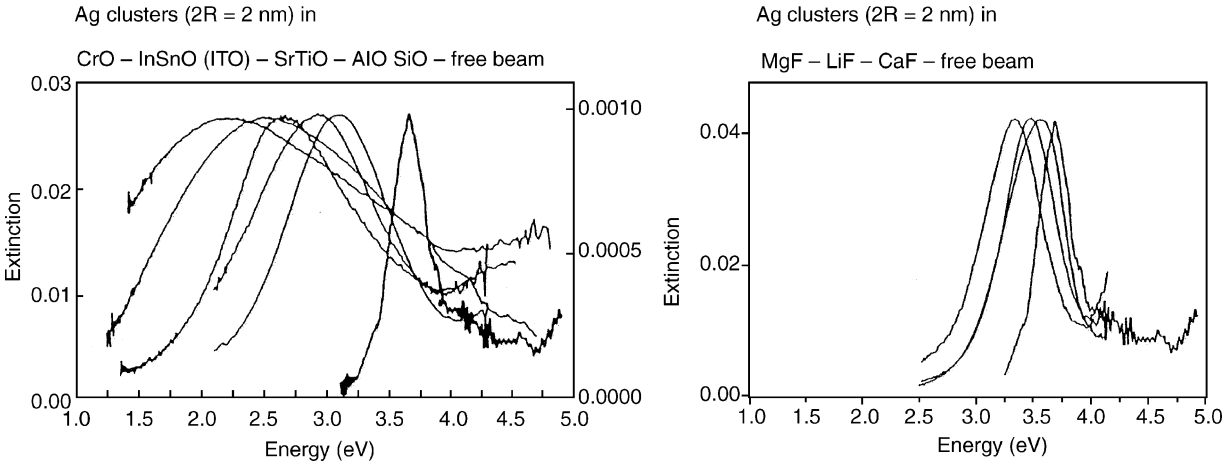


Figure 12 Measured optical extinction spectra of 2 nm Ag nanoparticles, embedded in various different embedding materials. For comparison, the spectrum of the free nanoparticle beam is added. Left side: oxidic materials, right side: fluoridic materials.

$\varepsilon_{\text{matrix}}$, which classically follow from Mie's theory and by comparingly small additional shift due to static charge transfer;

- drastic broadening (exceeding, in some cases, all known size effects), which according to eqn [14], can be described by characterizing values of the *A*-parameter. Recently, these parameters have been experimentally determined for Ag nanoparticles embedded in a broad variety of host materials. Their interpretation is rendered difficult, however, due to inhomogeneous broadening of varying magnitude.

We can immediately judge from Figure 12 that both shift and broadening are larger in the case of oxidic matrices than of fluoridic ones, i.e., oxides possess more interface states close to E_F of the material. In Figure 13, observed peak positions $\hbar\omega_{\max}$ and broadenings expressed in the resulting *A*-parameters, are summarized for a broad variety of different embedding materials.

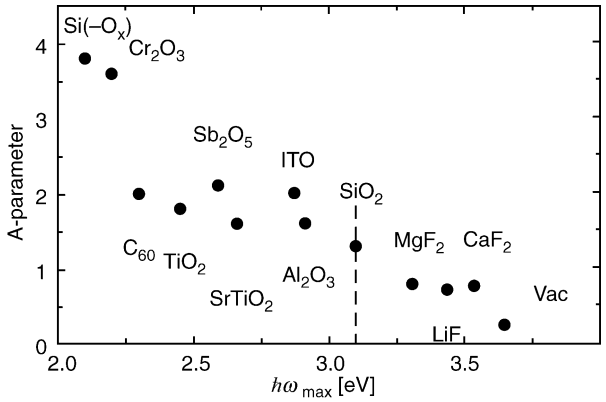


Figure 13 Silver nanoparticles ($2R = 2$ nm) embedded in various solid host matrices: measured Mie peak positions and resonance halfwidths expressed by the *A*-parameters.

Comparison of Plasmon Polariton Lifetimes with Femtosecond Experiments

In the last section, experimental examples were presented to show how fundamental physical properties are changed by forming heterogeneous materials. Because the Mie resonance widths are mainly a consequence of finite plasmon polariton life times τ, (in addition, there are interband contributions, in larger particles radiation damping, and dispersion influence), the latter can be evaluated from measured optical absorption spectra. This supposes the bandwidth to be essentially homogeneous. From a width analysis fitting Mie's theory numerically we obtained for Ag particles, free flying in UHV with uncontaminated free surfaces and 2 nm in size, a plasmon lifetime of $\tau = 6$ fs. Under these experimental conditions, the measured bandwidth is the homogeneous one.

Concerning the size dependences, we can extrapolate to larger particles. The longest lifetimes are expected for 15 nm particles, for which $\tau = 15$ fs was obtained by extrapolating the experimental data for 2 nm particles to this size. The same width analysis was performed for the nanoparticles embedded in matrix materials, and as a result, the lifetimes drastically decrease down to 1–2 fsec. The observed increase of the *A*-parameter, as shown in the previous section, is thus directly correlated to the measured decrease of the plasmon polariton lifetime. Hence, we conclude that the decrease of the lifetime is mainly due to increased phase relaxation caused by the dynamic charge transfer effect.

These results can be compared with recent femtosecond spectroscopy experiments which have recently given much impact on the experimental investigation of the dynamics of Mie resonances in silver and gold nanoparticles. In these experiments data between 6 and 10 fsec were obtained for deposited Ag and Au particles of different sizes and different substrates. Regarding the complexity of the different experiments the correspondence of the data appears convincing.

Obviously the lifetimes evaluated in the frequency domain, from our spectra of particles *in vacuo*, coincide with the lifetimes determined in the time domain. The latter, though, were measured with larger particles deposited on substrates (i.e., not from free particles but including interactions with substrates).

As a concluding remark, it may be pointed out again, that the optical properties and the heterogeneity effects presented here for the model system of nanoparticle/host-matter, are also to be found in other examples of the large field of heterogeneous materials.

List of Units and Nomenclature

Absorption bandwidth	Γ
circular frequency [s^{-1}]	ω
dielectric constant/dielectric function [–]	ε
effective electron mass [kg]	m_{eff}
electric field [$V\,m^{-1}$]	E
(electron) relaxation frequency [s^{-1}]	γ
electron state energies for initial and final state [J]	E_i, E_f
elementary charge [C]	e_0
extinction, scattering, absorption cross section [m^2]	σ_{ext}, σ_{sca}, σ_{abs}
Fermi energy [J]	E_{F}
Fermi velocity of electrons [$m\,s^{-1}$]	v_{Fermi}
intensity [$W\,m^{-2}$]	I
Kronecker energy distribution [J^{-1}]	$\delta(E)$
magnetic field [$A\,m^{-1}$]	H
norm of wave vector of electrons [m^{-1}]	k
number density of conduction electrons [m^{-3}]	n
Riccati-Bessel functions of first kind [–]	$\psi_n(z)$
particle radius [m]	R
(phase-) life time [s]	τ
Planck constant [J_s]	$\hbar$
Riccati-Bessel functions of third kind [–]	$\eta_n(z)$
scalar electromagnetic potentials [V]	Π, Γ
susceptibility [–]	χ
transition matrix element [$kg\,m\,s^{-1}$]	M_{if}
vacuum permittivity [$CV^{-1}\,m^{-1}$]	ε_0
vacuum velocity of light [$m\,s^{-1}$]	c
vacuum wavelength of light [m]	λ
wave vector of electrons [m^{-1}]	$\mathbf{k}$

See also

Scattering: Scattering Theory. **Semiconductor Physics:** Polarons.

Further Reading

Bassani F and Parravicini G (1975) *Electronic States and Optical Transitions in Solids*. Oxford: Pergamon.

Bohren CF and Huffman DR (1983) *Absorption and Scattering of Light by Small Particles*. New York: Wiley.

Born M (1960) *Optik*. Berlin: Springer-Verlag.

Duncan M (ed.) (1998) Advances in Metal and Semiconductor Clusters, vol. IV. London: JAI Press Inc.

Haberland H (ed.) (1994) *Clusters of Atoms and Molecules*. Berlin: Springer Series in Chemical Physics 56.

Nalwa HS (ed.) (2001) *Handbook of Surfaces and Interfaces of Materials*, vol. 3. San Diego: Academic Press.

Hölzl J, Schulte F and Wagner H (1979) *Solid Surface Physics*. Springer Tracts in Modern Physics 85. Heidelberg: Springer.

Hummel R and Wissmann P (eds) (1997) *Handbook of Optical Properties*, vol. II: Optics of Small Particles, Interfaces and Surfaces. Boca Raton, FL: CRS Press.

Jena P, Khanna SN, Rao BK (eds) (2000) *Cluster and Nanostructure Interfaces*. New Jersey: World Scientific.

Kerker M (1969) *The Scattering of Light and Other Electromagnetic Radiation*. New York: Academic Press.

Kreibig U and Vollmer M (1995) *Optical Properties of Metal Clusters*. Berlin: Springer Series in Material Science 25.

Nanometer Scale Science and Technology, Internatl. School of Physics 'Enrico Fermi' (2001) CXLIV. Amsterdam: IOS Press.

Quinten M (1998) *Physical and Chemical Properties of Surfaces in Nano-Particle Systems*. Aachener Beiträge zur Physik kondensierter Materie, vol. 24. Mainz, Aachen: Wissenschaftsverlag.

Ruhbahn HG (2002) *Nanophysik und Nanotechnologie*. Stuttgart: Teubner.
Schmid G (1994) *Clusters and Colloids*. Weinheim: VCH Verlag.
van de Hulst HC (1957) *Light Scattering by Small Particles*. New York: Wiley.
von Plessen G (1998) *Ultrafast Dynamics of the Surface Plasmon in Gold Particles*. Advances in Solid State Physics 253. Berlin: Vieweg.

Lightweight Mirrors

J W Bilbro, George C Marshall Space Flight Center, AL, USA

Published by Elsevier Ltd.

The desire to probe the heavens is as old as humanity; the realization that the human eye could be aided in these investigations is somewhat more recent. In May 1609, Galileo Galilei, while visiting in Venice, heard about the new 'telescope' invented by the Dutch lens-maker Hans Lipershay and decided that it would be useful in his astronomical investigations. He returned to his home in Padua and made his own telescope comprised of 4.2 cm diameter plano-convex and plano-concave lenses. With this telescope, Galileo discovered the satellites of Jupiter; got tenure; had his salary doubled – and the race was on! It did not take too long to realize that 'bigger' was 'better', and from this point forward, there has been a continuous striving to increase collecting area and resolution, thereby enabling the astronomer to 'see' farther and farther into the heavens.

The limitations of refractive optics soon prompted investigations into reflecting optics – mirrors. In 1668, Isaac Newton produced the first reflecting telescope having a 2.5 cm diameter mirror made of speculum (copper (6), tin (2), and arsenic (1)). A little over 100 years later, in 1789, William Herschel constructed his 'Great Telescope', having a 1.22 m diameter mirror made of bronze. Of course as telescopes grew, so did mirror weight. In 1845, William Parsons' 'Leviathan' was built. It had a 1.84 m diameter mirror made of speculum that weighed 3600 kg – a whopping 1354 kg m^{-2}.

Problems with corrosion in speculum and bronze mirrors led to a renewed interest in glass, and the development of highly reflective coatings soon made metal mirrors obsolete. In 1917, the mirror for the Hooker telescope arrived on Mount Wilson. Made by the Saint Gobain glassworks in France from the same material used for wine bottles, the mirror was 2.54 m in diameter, 33 cm thick and weighed over 4000 kg (789 kg m^{-2}).

The increasing weight of larger and larger mirrors produced practical problems for telescope makers and astronomers. Herschel's telescope was so unwieldy that he found it almost impossible to use and returned to smaller telescopes for most of his observations, and the 2.54 m Hooker telescope required a 100 ton mount and drive system. Obviously, if telescope apertures were to increase, steps had to be taken to make mirrors lighter. The first major success in this arena was accomplished for what was to become the Hale Telescope commissioned at the Palomar Observatory in 1948. In 1936, Corning cast a mirror made of a new borosilicate glass blend called Pyrex. The blank was lightweighted by placing hexagonal shaped obstructions in the mold which created corresponding voids in the glass. The finished mirror (**Figure 1**) was 5.08 m in diameter, 61 cm thick at the edge and weighed approximately 13 000 kg (641 kg m^{-2}). Even with this degree of 'lightweighting' the mounting and drive mechanisms still weighed around 182 000 kg.

Although the desire for larger ground-based telescopes provided the initial impetus for developing lightweight mirrors, the advent of the space program and the desire to put telescopes in space became the main driver in the pursuit of 'serious' lightweighting techniques. The cost of launching



Figure 1 5.08 m Pyrex™ mirror made by Corning/Caltech for the Hale Telescope. Reproduced with permission from the Palomar Observatory/California Institute of Technology.

large telescopes into space quickly becomes prohibitive as aperture size increases. Consequently, the current emphasis is on developing ultra-lightweight, deployable telescopes that permit ever-increasing apertures to be placed in orbit. Fortunately, the zero gravity environment of space permits satisfactory operation of such ultra-lightweight telescopes and therefore the extent to which a mirror can be lightweighted is primarily dependent upon the mirror material and the manufacturing techniques utilized.

Manufacturing

Mirror manufacturing can, in some sense, be divided into two general classes, one that removes material (direct generation) and one that deposits material (replication). The two classes have significantly different limitations relative to producing lightweight mirrors. Direct generation encompasses the more traditional methods of mirror manufacturing, blank fabrication, and optical surface generation, i.e., grinding and polishing. Many innovative techniques have been developed over the past decade that have not only made significant strides in improving fabrication times, optical figure, and surface finish, but have also made possible the reduction of weight to unprecedented levels.

Although extremely lightweight mirrors are being fabricated using direct generation, in the end, there are limitations in the fabrication processes that are not present in replication. This process offers the fascinating potential to create mirrors that are only a few molecules thick, thereby bringing a new meaning to the term 'lightweight optics'.

Direct Generation

Mirror blank fabrication

Casting lightweight glass mirror blanks has evolved substantially since 1936. Four 8.2 m diameter Zerodur™ mirror blanks were spun cast for the European Southern Observatory's Very Large Telescope Interferometer. The largest mirror cast to date was the 8.4 m borosilicate mirror cast in a spinning oven for the Large Binocular Telescope on Mount Graham. Although not exactly lightweight by space standards – the blank weighs 15 455 kg (266 kg m^{-2}) – it is substantially less than a conventionally cast solid blank.

It has been postulated that an areal density of around 64 kg m^{-2} probably represents the limit of lightweight glass castings. Therefore, as the demand for lighter and lighter mirrors increased, methods other than casting had to be developed. Investigations in the 1960s indicated that mirrors could be lightweighted to about 180 kg m^{-2} by first fabricating a 'honeycomb' core and then fusing plates on the front and back of the core using a high-temperature fusion process. This process was used to produce the Hubble Space Telescope mirror (Figure 2). Hubble had a 2.5 cm thick face-sheet and backsheet and a honeycomb interior constructed of individual pieces 0.5 cm thick. These pieces were assembled and placed in a large furnace where the temperature was increased until they fused together. The resulting 2.4 m diameter mirror weighed 828 kg (183 kg m^{-2}).

During the 1980s, Frit bonding technology was developed. Unlike the fusion process, Frit allowed the bonding of glass elements at a lower temperature. This allowed the cross-sections of the glass elements to be significantly reduced. The thicknesses of face-plates were typically reduced to around 7 mm and areal densities in the 40–50 kg m^{-2} range became routine. New advances were also made in core manufacturing. Diamond tooling was developed along with computer controlled systems to machine cores from solid blanks. High-pressure water jets were developed that allowed cores to be made much faster than with diamond tooling. In addition, the water jet cutting provided the ability to create very flexible geometries optimized to minimize weight while retaining stiffness. Water jets could machine core walls that could be held to a thickness of around 1.5 mm over a core depth of 20 cm. Computer controlled diamond machining has created even thinner cores (less than 1 mm in thickness) but the process requires longer generation times.

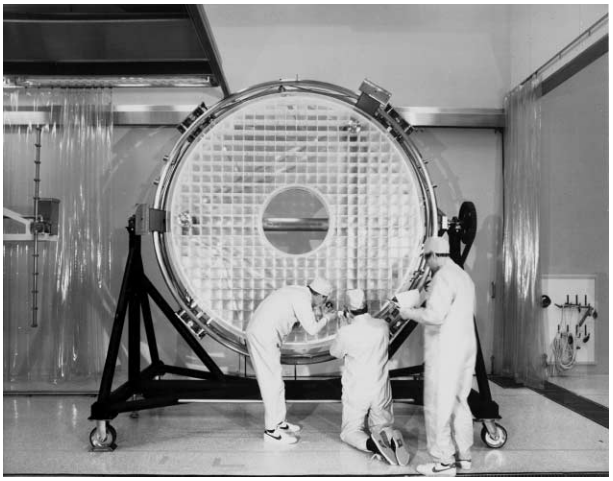


Figure 2 2.4 m ULE™ mirror made by Corning/Perkin Elmer for the Hubble Space Telescope. Photography courtesy of NASA.

Material removal

The evolution in blank manufacturing required a corresponding evolution in the process of material removal. Many new grinding and polishing techniques were developed over the years to more accurately and rapidly generate precision optical surfaces. These included adaptive force grinding, electrolytic in-process dressing (ELID), water jet figuring, computer controlled polishing, stress lap polishing, magneto-rhelogical figuring (MRF), plasma-assisted chemical etching (PACE), ion figuring, and vacuum lap figuring. Most of these processes induce localized forces on the face-sheet in the process of material removal, and for a 'honeycomb' mirror, the face-plate deflects away from the grinding/polishing tool as the tool reaches the center of the cell. This means that less material will be removed in the center of the cell than at the edge of the cell where the core wall holds the face-sheet against the tool. This creates a 'quilted' pattern in the surface that must be removed in subsequent steps. As face-sheets become thinner, quilting becomes more pronounced and the process of removing 'quilting' becomes exacerbated. For a number of years, this was the limiting factor in the development of high-precision lightweight mirrors.

Ion figuring is a noncontact process that has eliminated the problems associated with mechanical material removal. It bombards the surface of the optic with neutralized ions causing atoms to be dislodged. The process creates no deflection of the surface and therefore can remove quilting and other irregularities without generating further distortion in the surface. This has permitted 2–3 mm thick face-plates to be manufactured without concern for quilting effects. Primary limitations now being encountered are in face-plate generation and the various associated handling processes. Great care must be taken to avoid fracture. Spin casting ultra-thin face-plates is being investigated as a means of further reducing problems in the generation process. The 0.75 m mirror made for IKONOS, the high-resolution imaging satellite, represents the state-of-the-art in all-glass lightweight mirrors (28 kg m^{-2}).

Replication

Replication techniques are quite common in low-cost production mirrors and lenses for commercial purposes. However, there are also a number of techniques that hold great potential for precision mirrors as well. Chemical vapor deposition (CVD), plasma vapor deposition (PVD), plasma spray deposition, and electro-forming are all being investigated for producing lightweight precision mirrors. This class of mirror generation does not start with a mirror blank *per se*, rather it replicates a 'mandrel' that has the appropriate surface finish and the inverse of the shape desired in the final mirror. The mandrel can be made very rigid and therefore unperturbed by either gravitational forces or the localized forces generated by grinding and polishing. Extremely thin mirrors with areal densities on the order of 0.1 kg m^{-2} are currently being manufactured using replication techniques. These 'membrane' mirrors are currently being used for solar energy collection; however, recent advances in membrane quality and wavefront correction techniques hold out the promise of producing high-resolution, membrane imaging systems.

The Future?

Theoretically, single-molecule thick mirrors could be manufactured in space using self-replicating nanotechnology systems that 'grow' mirrors according to a mathematical description of the desired optical surface. These 'smart' mirrors could self-correct in the presence of physical disturbances and self-heal when damaged. They could be 'grown' to hitherto unimagined size, being limited only by the availability of raw material. Such mirrors, however, are definitely of the future and should merit re-examination 20 years hence.

Materials

A variety of materials have been developed over the years that have led to lighter and lighter mirrors. Recently, metals are coming back into play as means of fabricating stiff, lightweight mirrors. These metals include beryllium, nickel, nickel alloys, aluminum, and aluminum alloys. Composite materials are also playing a large role in lightweight mirror fabrication, both as a support structure and also as a reflecting surface (with appropriate coatings). Hybrid mirrors comprised of a combination of glass and composites have also been developed, as well as combinations of glass and beryllium. Various forms of silicon carbide have been extensively investigated and lightweight mirrors of pure silicon are being manufactured. Glass still plays a significant role in large lightweight mirrors, and various plastics are under investigation for the production of ultra-thin membrane mirrors.

Glass, Glass Ceramics, and Hybrids

The most common glass/glass ceramic materials used in large telescope mirrors are fused silica, ULE™ (Ultra-Low Expansion) glass, and Zerodur™, a glass ceramic. As was mentioned earlier, there are a variety

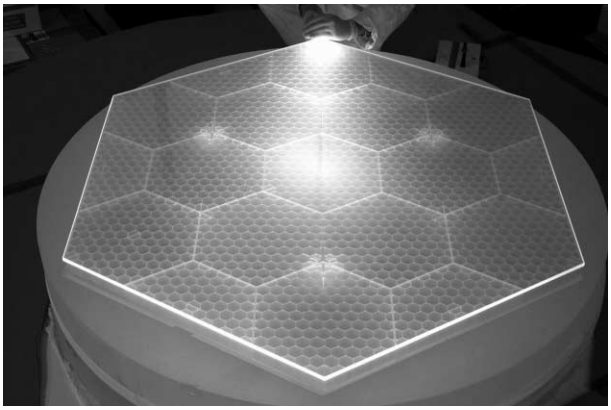


Figure 3 1.5 m ULE™ mirror made by Corning/Kodak for NASA's Next Generation Space Telescope Advanced Mirror System Demonstrator technology program. Photograph courtesy of NASA.

of methods available for the manufacture of glass mirrors depending upon the extent of lightweighting desired. Most of the current effort in lightweight mirrors is centered around hybrid mirrors. One method bonds very thin (2–3 mm) glass face-sheets to composite cores to form 'sandwich' mirrors. Another method attaches face-sheets to composite back-plane structures via actuators. In the latter case, the actuators can be used to actively control mirrors to correct for distortions. These mirrors were developed as part of the James Webb Space Telescope technology program, and typically have areal densities on the order of 15 kg m^{-2} (including back-plane structures). An all glass face-sheet and core is shown in Figure 3.

Silicon Carbide

Silicon carbide has long been a promising material for lightweight mirrors. It has many of the desirable properties of beryllium and is nontoxic. Until recently difficulties were encountered in producing mirrors larger than ~0.5 m in diameter. Of late, several manufacturers have produced large (<1 m) lightweight mirrors using different processes; sintering, liquid silicon infiltration and chemical vapor deposition.

The sintering process consists of preparing a finely ground silicon carbide powder mixed with organic binders and additives and then compacting the powder by isostatically pressing it at high pressure. The resulting compact is then machined to the desired shape and weight and the blank heated to cause agglomeration of the particles. The resulting material is known as reaction-bonded silicon carbide; it can be polished to around 30 angstroms and can be ion figured. If smoother surfaces are required, then a layer of CVD or PVD silicon carbide can be added. A 1.35 m diameter mirror has been fabricated as a demonstration for the European Space Agency's Herschel Space Observatory. The mirror was fabricated in nine segments that were then brazed together to form a unified 26 kg m^{-2} blank. This mirror was not designed to test the limits of lightweight and, consequently, there is no reason to believe that this process could not achieve an areal density comparable to the 7.8 kg m^{-2} of Cesic™ described in the following paragraph. The size of mirrors produced using this technique is currently limited to around 4 m due to the availability of the furnaces required for segment brazing.

The liquid infiltration process is used to create a biphase carbon, silicon carbide material known as Cesic™. This material is manufactured by creating a block of chopped carbon fibers embedded in a phenolic resin. The block is machined to the desired shape and then put into a furnace where the temperature is elevated until the resin is turned to carbon. The temperature is further elevated and liquid silicon infiltrated to produce silicon carbide. Bare Cesic™ can be polished to approximately 100 angstroms. In a subsequent step, a silicon carbide slurry can be applied to the blank that can be polished to around 30 angstroms. It is also possible to add a CVD or PVD silicon carbide layer to the blank that can be polished to obtain angstrom-level surfaces. A 1.04 m diameter Cesic™ mirror has been constructed for the NASA program, Solar Lite. The mirror weighs approximately 60 kg for an areal density of 56 kg m^{-2}. Again, this mirror was not designed to test the limits of lightweighting, but in a separate demonstration, an areal density of 7.8 kg m^{-2} was achieved. The size of mirrors produced using this technique is also currently limited to around 4 m due to the availability of existing furnaces required for infiltration.

The CVD process produces a SiC material that is highly polishable, yielding surface finishes on the order of angstroms. It has been used quite successfully as a coating layer, and has the potential to provide a means of replicating very lightweight mirrors by depositing a thin layer of SiC onto a mandrel. In one approach, the mandrel has the appropriate shape, but not the surface finish. In this case, once the SiC has been deposited, the mandrel is burned away and a SiC mirror blank is left. The blank is then ground and polished to produce the desired figure and finish. If the blank is too thin, this can present a problem in achieving a high-performance mirror. A second approach involves the deposition of SiC onto a figured and polished mandrel. In principle, both the figure and surface finish can be replicated. Success in this approach, however, has been somewhat limited, and although 20-angstrom surface finishes have been

replicated, they have not been repeatable. A 1 m diameter blank was manufactured using the first approach, but it too has not been repeated and the lack of success appears to be due to high stress developed in the deposition process. The residual stress in the blank causes large-scale deformation and frequently fracture.

Beryllium

The extremely high stiffness of beryllium makes it a highly desirable candidate for lightweight mirrors. Unfortunately, it is a highly toxic material and there are only a very few companies involved in the manufacture of beryllium mirrors and even fewer that produce the raw material. Blanks are manufactured by isostatically compressing beryllium powder at high temperature and pressure. The blanks are then machined and subsequently ground and polished. Early beryllium mirrors often exhibited interesting, mysterious changes in figure after seemingly benign handling events. Unpredictable changes in figure also occurred as they were cycled between ambient and cryogenic temperatures. These problems appear to have been resolved with the production of spherical powder. The older powder consisted of irregularly shaped particles and when the powder was compressed, small, irregularly shaped voids were created throughout the material. It is hypothesized that these voids changed in size and shape as the blank was stressed, thereby causing unpredictable dimensional changes in the blank. The spherical powder permits uniform, high-density packing yielding homogeneous blanks that do not appear to exhibit unpredictable dimensional changes.

The lightest-weight beryllium mirror to date ($11.8\ \mathrm{kg\,m^{-2}}$) was made for NASA's James Webb Space Telescope technology program. This technology was selected for the flight mirror. The weight reduction was achieved by machining the blank to create an open back honeycomb structure with an integral face-sheet. The ability to fabricate large beryllium mirrors is limited by the availability of facilities capable of isostatically pressing the material at the requisite pressures and temperatures. Currently, blank sizes are limited by available isostatic presses to around 1.2 m in width and 1.6 m in length. The expense of construction of larger facilities combined with limited need make it highly unlikely that any larger mirrors will be manufactured in the near term. A brazing process, however, has been demonstrated for joining segments to form a large blank. This process could, in principle, permit the fabrication of mirrors several meters in diameter. Bare beryllium can be polished to around 30 angstroms surface finishes. If smoother surfaces are required, beryllium can be vacuum deposited on the blank. Nickel may also be deposited on the blank; this has the added benefit that it can be polished without the special facilities required for the hazardous beryllium polishing operations. The drawback of using nickel is that of the bimetallic effect if the mirror is to be operated over a wide temperature range. The Vavilov Institute in Russia has developed an alternative approach to obtaining good surfaces by creating a glass that matches the coefficient of thermal expansion (CTE) of beryllium over a 50° temperature range. The glass is then fused to the beryllium blank. There have been numerous attempts over the years to produce replicated mirrors using a CVD process to deposit beryllium on a mold. So far these attempts have been unsuccessful, the process producing a material structure that no longer has high stiffness.

Silicon

Silicon as a mirror material has the advantage of being a very homogeneous material that can be polished to very smooth surfaces (angstrom level) very rapidly. A lightweight mirror has been fabricated using a silicon face-sheet brazed to a central core of silicon foam. This mirror 'sandwich' has demonstrated an areal density of $<15\ \mathrm{kg\,m^{-2}}$. Technology has been developed to braze segments together in a manner that has been shown to have no adverse effects on either polishability or cryogenic performance (i.e., no distortion across the braze). Consequently, it should be possible to manufacture lightweight silicon mirrors of 1 m diameter or larger.

Composites

There have been many attempts to manufacture mirrors from composite materials, beginning initially with graphite/epoxy and more recently with graphite/cyanate-ester. These mirror are produced by laying up graphite sheets impregnated with epoxy, or cyanate-ester on a near net shape mold and then curing the material at elevated temperatures. The optical surface is typically created by coating another mold (this time with the desired optical surface) with a passivation layer and then with epoxy and placing the near net shape composite substrate on the epoxy. When the epoxy cures, the completed mirror is separated from the mold and the epoxy surface ion milled to remove residual errors such as fiber print-through. These mirrors can be very lightweight ($\sim 5\ \mathrm{kg\,m^{-2}}$), but typically have problems with large-scale distortion due to inhomogeneities due to

the lay-up process. The epoxy surface replication can also be a problem if the mirror is subjected to temperature extremes. A 3.5 m diameter composite mirror was made as a demonstration for the FIRST program. It has suitable performance for submillimeter and infrared wavelengths, but not for shorter wavelengths. At this time, the accuracy required for large (>1 m) visible imaging systems has not yet been achieved.

Gossamer Mirrors

The term 'gossamer' refers to that category of mirrors that in and of themselves have no structural rigidity beyond spatial scales of fractions of millimeters. Several materials are currently being used for these types of materials including Kapton™ and various other polyimide films. The initial shape of the membrane is formed by a variety of techniques including vacuum forming, pressure forming, spin casting, hot forming, and others. Early mirrors were manufactured by joining together triangular-shaped segments (gores) to form the desired shape. When deployed, these mirrors require some means of applying pressure to achieve the desired shape. Some methods have involved applying force at the edge to put the mirror in tension; some involve using a clear membrane as a face-sheet and then pressurizing the combination to create a 'lens' with a reflecting inner surface. Most recently, techniques have been demonstrated involving various forms of electrostatic deflection to achieve the desired shape after deployment. At present the primary applications are in the area of solar collection, power beaming, and thermal propulsion. The mirrors are approaching the quality required for imaging telescopes, particularly for small sizes. Techniques for correcting these mirrors with 'downstream' active optics have reached the point where it is not unrealistic to imagine that membrane optics could soon reach the accuracy required for imaging. There have been two space-based demonstrations, one by the Russians – a 20 m reflector – and the other by the USA – a 14 m inflatable antenna (Figure 4).

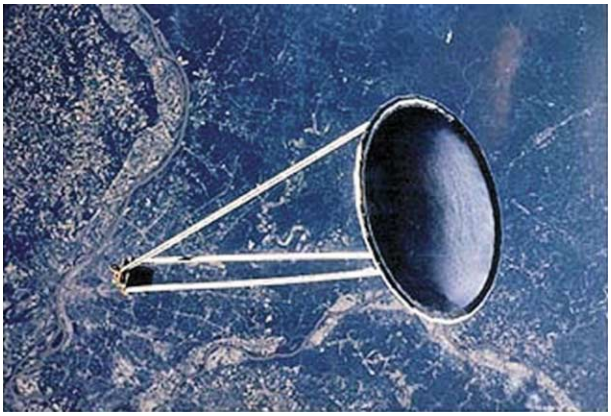


Figure 4 15 m antenna made by L'Garde for NASA's Inflatable Antenna Experiment. Photograph courtesy of NASA.

Grazing Incidence Mirrors

Broadband X-ray telescopes make use of the reflection property of materials at angles of grazing incidence. The most common design is based on the X-ray microscope objective designs of Wolter in the early part of the twentieth century. These mirrors are hyperboloids and paraboloids of revolution. This configuration provides a structural rigidity not present in normal incidence mirrors; consequently, X-ray mirrors can be made much lighter. The largest mirror of the Chandra X-ray Observatory is made from Zerodur™ (Figure 5). It is a 1.2 m diameter, 0.9 m long cylindrical mirror with a wall thickness of only 24 mm. Such a thickness in a normal incidence mirror of this size could be easily deformed.

Lightweight approaches in grazing incidence optics are almost all concentrated in various forms of replication. Recent developments in electroformed nickel alloy mirror technology have resulted in a 0.5 m diameter 0.6 m long mirror that weighs only 1.2 kg (Figure 6). A similar-sized mirror in the Chandra telescope weighs over 100 kg. Extremely lightweight cylindrical mirrors have been made from graphite/epoxy and from CVD SiC. Various investigations are underway in producing segments of cylinders that are then mounted to form a complete



Figure 5 1.2 m Zerodur™ made by Schott Glass Werken/Hughes Danbury Optical Systems for NASA's Chandra X-ray Observatory. Photograph courtesy of NASA.

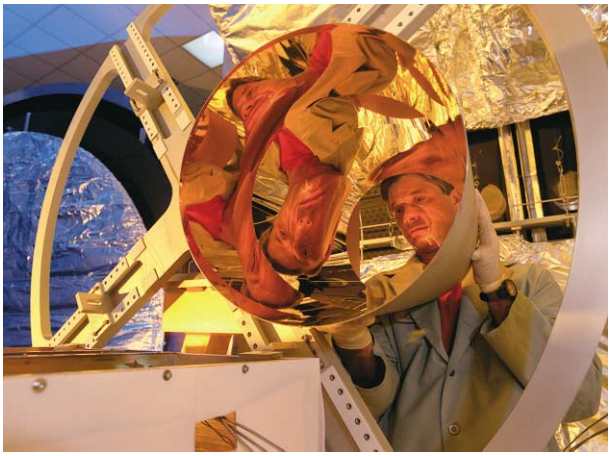


Figure 6 0.5 m electro-formed nickel–cobalt technology mirror made by NASA's George C. Marshall Space Flight Center for NASA's Constellation X-ray Mission technology program. Photography courtesy of NASA.

cylindrical mirror. These techniques involve a variety of materials including aluminum foil, and beryllium plates. In segment replication, the mandrel is passivated and then coated with the desired reflecting material of the final optic (the passivation process prevents the coating from adhering too strongly to the mandrel.) After the mandrel has been coated with the reflecting material, an epoxy coating is applied and a preformed shell segment is placed on the epoxy. After the epoxy cures, the segment can be removed from the mandrel along with the desired reflective surface (typically gold). In this case, the preformed shell does not have to have as precise a figure, or surface as the mandrel, since the epoxy will accommodate minor differences.

See also

Geometrical Optics: Lenses and Mirrors. **Imaging:** Adaptive Optics; Wavefront Sensors and Control (Imaging through Turbulence). **Instrumentation:** Astronomical Instrumentation; Telescopes. **Optical Materials**: Smart Optical Materials.

Further Reading

Bilbro J (2001) Optics in orbit. *SPIE OE Magazine* (Aug. issue).

Breckinridge JB and Jakobsen P (eds) (2000) UV, Optical, and IR Space Telescopes and Instruments. *Proceedings of SPIE* 4013.

Chen PC, Bowers CW, Content DA, *et al.* (2000) Advances in very lightweight composite mirror technology. *Optical Engineering* 39(9).

Ealey MA (ed.) (1995) Silicon carbide materials for optics and precision structures. *Proceedings of SPIE* 2543.

Engelhaupt DE, Ramsey BD, O'Dell SL, *et al.* (2000) New alloys for electroformed replicated X-ray optics. *Proceedings of SPIE* 4138: 22.

Goodman WA and Jacoby MT (2000) Dimensionally stable ultra-lightweight silicon optics for both cryogenic and high-energy laser applications. *Proceedings of SPIE* 4198: 260–270.

Jenkins C (vol. ed.) (2001) Gossamer spacecraft: membrane and inflatable structures technology for space applications. In: Zarchan P (series ed.) *Progress in Astronautics and Aeronautics* 191. Reston, VA: American Institute of Aetronautics and Astronautics.

Kendrick SE, Reed T and Streetman S (2001) In-process status of the 1.4-m beryllium semi-rigid Advanced Mirror System Demonstrator (AMSD). In: Stahl HP (ed.) *Optical Manufacturing and Testing*, vol. IV. *Proceeding of SPIE* 4451: 58–66.

Mathews G (2001) Semi-rigid active mirror technology advancements. *Proceedings of SPIE* 4451.

Müller C, Papenburg U, Goodman WA and Jacoby M (2000) C/SiC high precision lightweight components for optomechanical applications. *Proceedings of SPIE* 4198: 51–57.

Truemper JE and Aschenbach B (eds) (2000) X-Ray optics, instruments, and missions III. *Proceedings of SPIE* 4012.

Measurement of Optical Properties of Solids

P Lucas, University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The interaction of light with solids takes place through different mechanisms, depending on the type of material and the range of wavelength investigated. Insulators or dielectrics are typically transparent to visible light while most semiconductors are opaque to visible light yet transparent to infrared radiations; in contrast metallic solids appear shiny because they reflect all wavelength up to the ultraviolet region. The optical properties of a solid depends on its chemical composition and its structural properties and vary for every material, though, one parameter, the complex refractive index **n**, is sufficient to characterize entirely the optical properties of a specific material.

In this article we will briefly review the physics underlining the nature of the refractive index **n**, in order to understand its relationship to the observable quantities such as absorption, reflection, and transmission routinely measured in optics experiments. A description of the experimental techniques, as well as a physical interpretation of the data, will also be presented.

Origin of Light Interaction with Matter

Light is described as a transverse electromagnetic wave consisting of an electric and magnetic field oscillating perpendicular to the direction of propagation at optical frequencies within $10^{13} \rightarrow 10^{17}$ Hz. This range of frequencies comprises the ultraviolet, visible, and infrared domain. On the other hand, solid mediums are composed of charged particles, negative electrons, and positive ions producing electric dipoles, which can be polarized under the action of an electric field. Hence, when electromagnetic radiation impinges upon a material it interacts by polarizing the molecular units, producing oscillating dipole moments. This interaction results in several observable optical phenomena such as reflection, transmission, absorption, or scattering, which will be described in more detail in a following section. The classical model of light propagation assumes that the oscillating electric field can interact with several different types of dipole oscillators within the material. Different dipoles are usually accessed by light wave from different frequency range depending on their mass.

Electronic Oscillators

Lorentz originally proposed that the electrons bound to the nucleus of an atom behave as oscillators. In this representation, electrons behave as if they were held by springs and subjected to a linear restoring force, therefore oscillating up and down (Figure 1a). This representation of an atom might seem to contradict the conventional picture of electrons orbiting around the nucleus; however, it is valid as far as light–matter interactions are concerned. The resonant frequency of such an oscillator is proportional to the inverse of its reduced mass μ defined as

$$\frac{1}{\mu} = \frac{1}{m_e} + \frac{1}{m_N} \qquad [1]$$

where m_N and m_e are the mass of the nucleus and electrons, respectively. The resonant frequency ω_0 is then related to μ according to

$$\omega_0 \propto \sqrt{\frac{1}{\mu}} \qquad [2]$$

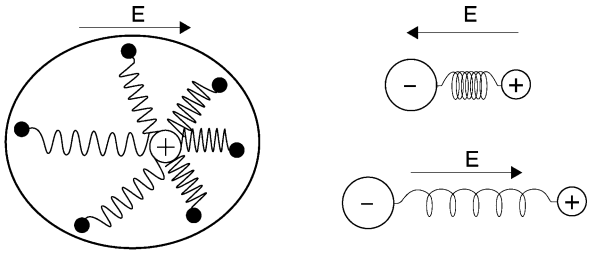


Figure 1 (a) Classical oscillator model of a polyelectronic atom. The electrons depicted as black dots are bound to the positive nucleus by springs, which represent the electrostatic restoring forces between the charged particles. The electronic oscillators are distorted under the effect of an electric field. (b) Classical model of a diatomic vibrational oscillator. The chemical bond is represented as a spring and the charged ions vibrate around their equilibrium position under the effect of an oscillating electric field.

Considering that $m_N \ll m_e$ we can assume that $\mu = m_e$ and show that the smaller mass of the electron determines the resonant frequency of the oscillator. Hence, resonant electronic oscillations are accessed at higher frequencies in the ultraviolet and visible region of the spectrum. In solids, the quantum theory shows that these resonances correspond to excitation of electrons across the gap between a full valence band and an empty conduction band.

Vibrational Oscillators

The second type of dipoles contained in the medium are based on pairs of charged atoms vibrating around their equilibrium position in the solid. Every atom has a different electronegativity, hence the electronic density on atomic pairs is distorted and result in a charge imbalance, which can couple to the electric field of light. This generates oscillating dipoles, which involves the motion of atomic nucleus (Figures 1b). The nuclear mass being several orders of magnitude larger than the electronic mass, we can show by analogy with eqns [1] and [2] that the resonant vibration will occur at lower frequencies in the infrared region. In solids, the quantum theory shows that the lattice atoms vibrate cooperatively and generate quantized lattice waves called phonons.

Free Electron Oscillators

In contrast with the bound electrons from the electronic oscillator model, free electrons can move without being subjected to a restoring force. These correspond to the conduction electrons present in metals and to a smaller extent in doped semiconductors. The free electron oscillator model is therefore

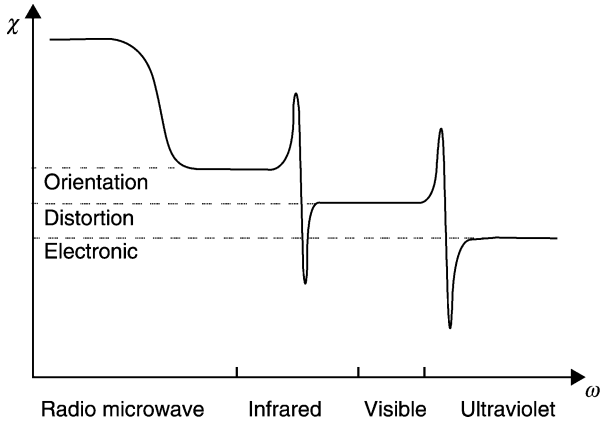


Figure 2 Contributions to the polarizability of a medium as a function of light frequency. At optical frequencies, the molecules cannot re-orient fast enough to follow the reversing electric field and the contribution of dipole orientation is lost.

principally used to describe the optical properties of metals such as reflectivity. The free electrons are well described as conventional oscillator with a natural resonant frequency $\omega_0 = 0$.

The relationship between the applied electric field E and the resulting polarization of the medium P is expressed through the dielectric susceptibility χ as

$$P = \varepsilon_0 \chi E \qquad [3]$$

where ε_0 is the permittivity of vacuum. We can generally recognize three contributions to the dielectric susceptibility or the polarizability of an optical material (Figure 2). Applied fields in the radio or microwave region oscillate slowly enough that polar molecular units can reorient and align in the direction of the field. The characteristic time for orientational motion is fairly long and this contribution is quickly lost when the molecules cannot follow the field oscillating at higher frequency. When the field reaches optical frequency, the two remaining contributions are the vibrational and electronic dipoles mentioned above. The vibrational dipoles involve a distortion of the chemical bond and are the slowest process occurring only in the infrared region. This contribution is lost at visible frequencies and leaves only the fast electronic polarization processes.

If we treat it rigorously, the polarization P is actually expressed as a function of higher order terms according to

$$P = \varepsilon_0\left(\chi E + \chi^{(2)}EE + \chi^{(3)}EEE + \ldots\right) \qquad [4]$$

This gives rise to nonlinear effects such as self-focusing and second or third harmonic generation. However, the higher order susceptibilities $\chi^{(n)}$ are very small and these effects are only significant for very high intensity light source such as lasers. In the following text we will only consider linear optical processes.

The Dipole Oscillator Model (Lorentz Oscillator)

In the Lorentz model, the bound electrons behave as harmonic oscillators and are constrained to vibrate at their natural resonant frequency ω_0. These oscillators are also submitted to a damping force corresponding to a resistance to the motion due to collisional processes. The equation of motion is then

$$m\left(\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \gamma\frac{\mathrm{d}x}{\mathrm{d}t} + \omega_0^2 x\right) = qE \qquad [5]$$

where x is the displacement along E, γ is the damping factor, m is the mass of the electron and q its charge. The passage of an electromagnetic wave through a medium exerts an oscillating force on the electrons causing them to vibrate up and down. In other words these electronic dipoles are submitted to a forced oscillation induced by an electric field varying periodically with time. The electric field felt by the atomic dipoles is written as

$$E = E_0 e^{i\omega t} \qquad [6]$$

where ω is the frequency of the light.

The electric field of the light wave then drives the dipole oscillations at its own frequency ω and the dipoles displacement oscillates according to

$$x = x_0 e^{i\omega t} \qquad [7]$$

Solving eqn [5] for x in terms of E gives the expression for the displacement of the electrons

$$x = \frac{q/m}{\omega_0^2 - \omega^2 + i\gamma\omega}E \qquad [8]$$

The atomic dipole moment p induced by the electrons displacement is therefore

$$p = qx = \frac{q^2/m}{\omega_0^2 - \omega^2 + i\gamma\omega}E \qquad [9]$$

And for a medium with N atoms per unit volume the total polarization P is

$$P = N\frac{q^2/m}{\omega_0^2 - \omega^2 + i\gamma\omega}E \qquad [10]$$

Our result in eqn [10] only contains one natural frequency ω_0 and therefore only accounts for one

single type of oscillator in the medium. However, normal optical mediums are composed of different kinds of atoms, each with several characteristic resonant frequencies ω_j. In our model, we can simply treat these multiple resonances by considering that each oscillator acts separately and add the contributions of all oscillators.

$$P = \left(\frac{q^2}{m} \sum_j \frac{N_j}{\omega_j^2 - \omega^2 + i\gamma_j\omega} \right) E \qquad [11]$$

Comparing eqns [3] and [11] then gives an expression for the polarizability χ of the medium

$$\chi = \frac{q^2}{\varepsilon_0 m} \sum_j \frac{N_j}{\omega_j^2 - \omega^2 + i\gamma_j\omega} \qquad [12]$$

Additionally, the polarization P is associated with Maxwell's displacement vector, D, through the relationship

$$D = \varepsilon_0 E + P \quad \text{and} \quad D = \varepsilon E \qquad [13]$$

where ε is the permittivity of the medium. The permittivity ε is normally expressed relative to the permittivity of vacuum ε_0 in terms of the dimensionless quantity $\varepsilon_r = \varepsilon/\varepsilon_0$. Where ε_r is the familiar dielectric constant of a material, which is related to the refracted index n as $n = \sqrt{\varepsilon_r}$.

Combining eqns [3] and [13] we obtain an important relationship linking χ to ε_r, and in turn χ to n.

$$\varepsilon_r = 1 + \chi \quad \text{or} \quad n^2 = 1 + \chi \qquad [14]$$

Finally, combining eqns [12] and [14] we obtain the expression for n in terms of the oscillator model.

$$n^2 = 1 + \frac{q^2}{\varepsilon_0 m} \sum_j \frac{N_j}{\omega_j^2 - \omega^2 + i\gamma_j\omega} \qquad [15]$$

The Refractive Index

Two important points should be made concerning the result of eqn [15]. The refractive index is a wavelength-dependent quantity and is a complex quantity. The complex refractive index $\mathbf{n}$ is indeed usually expressed as

$$\mathbf{n} = n + i\kappa \qquad [16]$$

where n is the real refractive index also defined as the ratio of the wave velocity in vacuum to the velocity in the medium $n = c/v$. And κ is the extinction coefficient, which is directly related to the absorption coefficient α as we shall see in the next section.

As we mentioned in the introductory section, the complex refractive index $\mathbf{n}$ is sufficient to characterize the optical properties of the solid. The real part of the index describes the change in velocity or wavelength of a wave propagating from a vacuum into a medium while the imaginary part is a measure of the dissipation rate of the wave in the medium.

By inspection of eqn [15] we can predict the form of the variation of n with wavelength. At frequencies ω below the resonant frequency ω_j, the terms ω^2 and $i\gamma_j\omega$ are much smaller than ω_j^2 and to a first approximation n is constant. When ω reaches a resonant frequency ω_j, the term $(\omega_j^2 - \omega^2)$ goes to zero and since the term $i\gamma_j\omega$ is very small, the value of n increases very sharply and exhibits a resonance line as depicted in Figure 3.

On closer inspection of eqn [15] however, we can see that n is not exactly constant in between resonance lines. As ω rises, the denominator slightly decreases and therefore n rises slowly with frequency. The refractive index is then slightly larger for wave of higher frequency. This is why a glass prism bends blue light more strongly than red light. This phenomenon is called dispersion as waves of different frequencies are 'dispersed' by the prism.

Another important point raised by eqn [15] is that because $(\omega_j^2 - \omega^2)$ goes to zero at a resonant frequency, the damping term $i\gamma_j\omega$ dominates and n becomes almost entirely imaginary. $\mathbf{n}$ is then mostly governed by the extinction coefficient κ. The damping associated with the coefficient γ_j corresponds to a friction force causing a loss of energy, which becomes the dominant effect during resonance. The resonance region is then associated with strong attenuation or

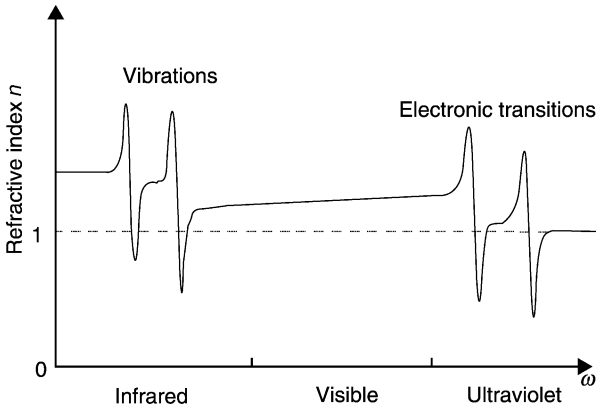


Figure 3 Variation of the refractive index of a hypothetical dielectric solid in the optical frequency range. At high frequency, ω is larger than all resonant frequencies and the refractive index n reaches unity.

absorption of the wave. In contrast, between resonant frequencies, the refractive index is almost entirely real as the imaginary part $i\gamma_j\omega$ is negligible in comparison with $(\omega_j^2 - \omega^2)$. Absorption is then very small and that region corresponds to the transparency domain of the material. In fact, we can assume that $\mathbf{n} = n$ in the transparent region, that is why tables of optical constants of transparent optical materials only list the real part of the refractive index n. However, at resonant frequencies on either side of the transparency domain, absorption becomes significant and the imaginary part of the refractive index must be taken into account so that $\mathbf{n} = n + i\kappa$.

The refractive index $\mathbf{n}$ being a complex quantity it ensues that ε_r must also be complex, since $n = \sqrt{\varepsilon_r}$. Hence we also define the complex dielectric constant as

$$\varepsilon_r = \varepsilon_1 + i\varepsilon_2 \qquad [17]$$

Both parameters describe equivalently the optical properties of a solid so that $\mathbf{n}$ or ε_r are called the optical constants of the material. Note that it is an unfortunate misnomer since they vary distinctly with wavelength (**Figure 3**).

Using equation [17] and $n = \sqrt{\varepsilon_r}$, we can establish the relationships between the real and imaginary parts of $\mathbf{n}$ and ε_r.

$$\varepsilon_1 = n^2 - \kappa^2 \qquad [18]$$

$$\varepsilon_2 = 2n\kappa \qquad [19]$$

We can see that $\mathbf{n}$ and ε_r are not independent variables but that the real and imaginary part of ε_r can be calculated, knowing the real and imaginary part of $\mathbf{n}$ and vice versa.

Moreover, it can be shown that the real and imaginary parts of each individual parameter are not independent either but can be computed from one another with use of the following Kramers–Kronig relations.

$$\varepsilon_1(\omega) = 1 + \frac{2}{\pi}\int_0^\infty \frac{\omega'\varepsilon_2(\omega')}{\omega' 2 + \omega^2}\mathrm{d}\omega' \qquad [20]$$

$$\varepsilon_2(\omega) = -\frac{2\omega}{\pi}\int_0^\infty \frac{\varepsilon_1(\omega')}{\omega' 2 + \omega^2}\mathrm{d}\omega' \qquad [21]$$

In these equations, only the principal value of the integral is calculated so that ε_1 can be computed if ε_2 is known over all frequencies and vice versa. The same equation connect n and κ together.

This analysis shows that it is only necessary to know one real or one imaginary part over a wide frequency range to determine all others. The interaction of light with matter is in fact characterized by a single independent real quantity.

Absorption

All electromagnetic phenomena are governed by Maxwell's equations. The form of an electromagnetic wave propagating through a medium is obtained by solving Maxwell's equation. For a wave propagating along z through a homogeneous, optically isotropic medium the form of the oscillating electric field is given by:

$$E = E_0 e^{i(kz-\omega t)} \qquad [22]$$

where k is the wave vector and ω the angular frequency of the light. For a wave traveling into a transparent medium of refractive index n, the relationship between k and ω is

$$k = \frac{n\omega}{c} \qquad [23]$$

However, for an ordinary optical material there are regions of absorption and the refractive index is therefore complex. We should then rewrite eqn [23] as

$$k = (n + i\kappa)\frac{\omega}{c} \qquad [24]$$

By substituting this value of k in eqn [22] we obtain an expression with two exponential terms

$$E = E_0 e^{-\frac{\omega\kappa z}{c}} e^{i\left(\frac{\omega n z}{c} - \omega t\right)} \qquad [25]$$

The term $e^{i\left(\frac{\omega n z}{c} - \omega t\right)}$ represents a wave traveling at a speed $v = c/n$, while the term $E_0 e^{-(\omega\kappa z/c)}$ represent the amplitude of this wave which decays exponentially with distance z. The intensity I of the wave is proportional to the square of the amplitude so that

$$I \propto e^{-\frac{2\omega\kappa z}{c}} \quad \text{or} \quad I \propto e^{-\alpha z} \qquad [26]$$

where α is the familiar absorption coefficient routinely measured by absorption spectroscopy. Hence we obtain an important relationship relating the observable quantity α to the imaginary part of the refractive index κ according to

$$\alpha = \frac{2\kappa\omega}{c} \qquad [27]$$

Local Field Correction

It should be pointed out that the derivation of our refractive index expression eqn [15] assumes that only the electric field of the light affects the polarization of the dipole oscillators. However, this

assumption is not correct in dense materials because of the contribution from the electric field of neighboring atomic dipoles. The field generated by each dipole will affect all other dipoles in the vicinity. It is possible to approximate this contribution by accounting for the effect of dipoles within a local sphere. The result of this approximation is known as the Clausius–Mossotti equation, which is accurate for isotropic medium such as glasses and cubic crystals

$$\frac{n^2-1}{n^2+2}=\frac{\chi}{3} \qquad [28]$$

Measurable Optical Parameters

Four things can happen to a beam of light propagating through a thick slab of optical material (Figure 4). Some of the light can be reflected at the surface of the solid, some can be absorbed by the sample, some can be scattered in different directions, and some of the light can be transmitted through the sample. The effect on the light beam, resulting from these phenomena, can be quantified by a number of optical coefficients, which characterize the macroscopic properties of the material.

Reflection of light at the surface of a solid is described by the reflectance R defined as the ratio of the reflected intensity I to the incident intensity I_0. For a beam falling perpendicularly on a flat surface, the reflection is called specular and is governed by the complex refractive index according to the Fresnel equation

$$R=\frac{I}{I_0}=\frac{(n-1)^2+\kappa^2}{(n+1)^2+\kappa^2} \qquad [29]$$

This provides us with a second important formula relating a measurable quantity (R) to the optical constant of the material. For measurements performed within the transparency region of a solid, the value of κ is much less than n and eqn [29] reduces to the more familiar form

$$R=\frac{(n-1)^2}{(n+1)^2} \qquad [30]$$

Absorption occurs when the light frequency reaches the natural resonance of some dipole oscillators in the medium. The energy of the electromagnetic wave is partly transferred to the material and generally dissipated in the form of heat. The light intensity is therefore attenuated as it propagates through the material and the attenuation efficiency is quantified by the absorption coefficient α. The intensity decreases exponentially with pathlength z according to Beer's law

$$I(z)=I_0e^{-\alpha z} \qquad [31]$$

where I_0 is the incident intensity. The imaginary part κ of the optical constant can then be directly obtained from measurements of α using eqn [27].

Scattering is the process whereby light is redirected in different directions due to microscopic density fluctuation in the sample produced by defects, impurities, or structural inhomogeneities. The intensity of light propagating in the forward direction is attenuated by the scattering event and it can be quantified in a way equivalent to absorption. The intensity has an exponential dependence on pathlength z analogous to Beer's law

$$I(z)=I_0e^{-Sz} \qquad [32]$$

where S is the scattering coefficient. When the scattering center is smaller than the wavelength of light this phenomenon is called Rayleigh scattering and the scattering coefficient S vary with the inverse fourth power of the wavelength

$$S(\lambda)\propto\frac{1}{\lambda^4} \qquad [33]$$

By measuring light attenuation we cannot tell the difference between absorption and scattering and the total attenuation is $\alpha_{\mathrm{Tot}}=\alpha+S$. However, the scattering contribution is generally much weaker than the absorption and can be neglected so that $\alpha_{\mathrm{Tot}}=\alpha$.

Transmission occurs when the light is neither reflected, absorbed, or scattered. The beam is then transmitted through the sample and the fraction of light exiting the back surface is quantified as the transmittance T. If we disregard scattering in comparison to absorption then conservation of energy

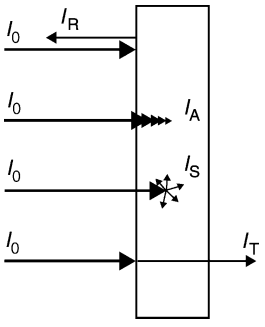


Figure 4 Representation of the four main optical processes happening to a beam as it propagates through a slab of transparent material.

require that

$$R + T + A = 1 \qquad [34]$$

where A is the fraction of light absorbed. The transmittance can then be compiled from R and α for an incident beam I_0 traveling across a sample of thickness l. In this case we must consider the reflection of the light on the front surface as well as on the back surface when it exits the transparent medium. The expression for the transmittance accounting for dual reflection is

$$T = (1 - R)^2 e^{-\alpha l} \qquad [35]$$

It should be noted that a rigorous treatment of eqn [35] should account for multiple reflections between the front and back surface in the interior of the sample. In this case, interference effects complicate the determination of **n** and a modified version of eqn [35] can be computed by summing up the intensities due to the contribution of the multiple reflections.

Transmission and reflection measurement are readily obtained using conventional spectrometers. Two types of spectrometers must be used to cover the entire optical spectral region. FTIR (Fourier Transform Infra Red) spectrometers use a glow bar as the light source, which enable it to cover the infrared and near infrared domain. The UV-VIS (ultraviolet visible) double beam spectrometers use a tungsten and deuterium lamp and cover the spectrum from the ultraviolet down to the near infrared. Using eqns [27], [29] and [35], the reflection and transmission measurements thus obtained are used to calculate the real and imaginary component of the optical constant over the entire optical frequency range.

An example of transmission measurement is shown in Figure 5, for a dielectric and a semiconductor. The two spectra have the same principal features. At short wavelength the edge of the transparency window is due to absorption of energy by the valence electrons. In classical terms this corresponds to the resonance of electronic oscillators, and in quantum mechanical terms to the excitation of electrons from the valence to the conduction band. This edge is due to a sharp increase of κ, which extends over a significant range of frequency corresponding to the domain of opacity or high absorption. The wavelength at the edge determines the minimum energy of the photon necessary to promote an electron across the bandgap E_g. Insulators have a large bandgap and appear transparent to the human eye because visible light is not energetic enough to promote an electron across E_g. In contrast, semiconductors have a smaller bandgap and appear black (opaque) because all the visible light is absorbed to induce electronic transitions. On the low frequency side the transparency window is limited by atomic vibrations in the solid structure. The wavelength of the vibrations increases with the mass of the atomic oscillators. Semiconductor compounds have higher atomic masses and therefore transmit at longer wavelength in the infrared while insulator have lower atomic mass and exhibit their vibrational absorption edge at shorter wavelength.

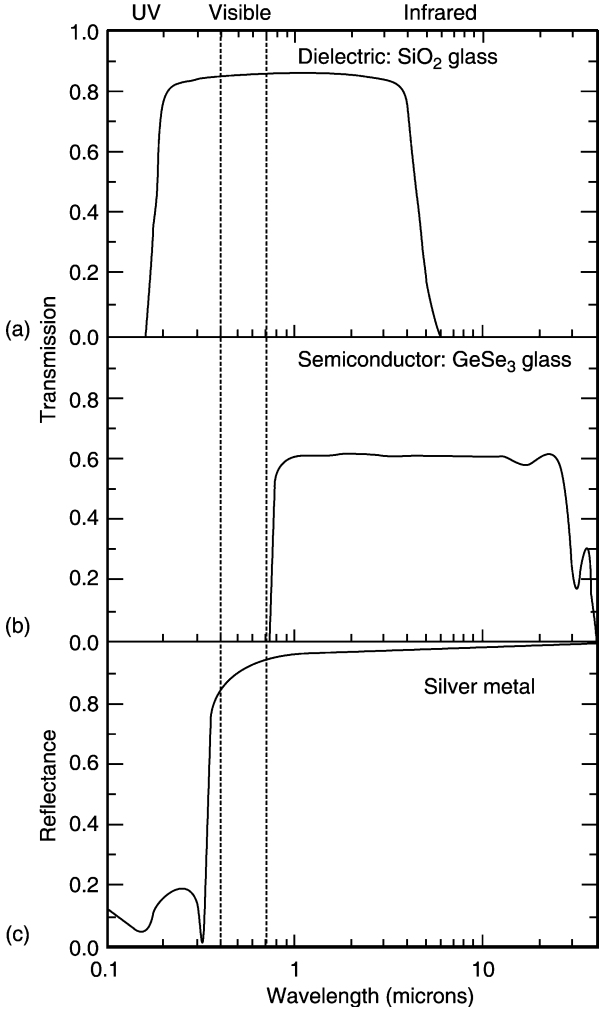


Figure 5 (a) Transmission spectrum of SiO_2 glass, a standard dielectric solid. (b) Transmission spectrum of $GeSe_3$ glass, a semiconductor ($E_g = 1.6$ eV). (c) Reflectance spectrum of silver metal.

In between the two absorption regions, κ is very small and the material is transparent. The refractive index is almost entirely governed by the real part n. Eqn [15] shows that n is proportional to the number N of electronic oscillators. The refractive index will then be higher for heavy atoms with a high electronic

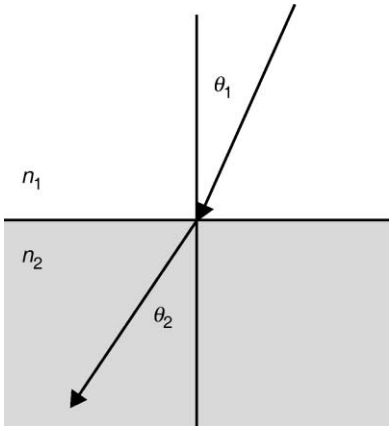


Figure 6 Refraction of a beam propagating across the interface between two mediums of different refractive index. The incident angle and the refraction angle are related to the index of refraction through Snell's Law.

density. Consequently, semiconductors exhibit a maximum transmission of only 60% because of significant surface reflectivity (eqn [35]) resulting from their higher refractive index n (eqn [30]).

Most applications of materials in optics are only concerned with the transparency domain so that the useful optical constant is reduced to the real part n. Hence a number of experimental techniques have been developed to obtain n, often by measuring angles of refraction with the use of Snell's law

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \qquad [36]$$

where θ_1 is the angle of the beam incident in the medium of index n_1, and θ_2 the angle of the beam refracted at the interface with a medium of index n_2 (Figure 6). The basic principle is to compare the sample with a standard glass of known refractive index and measure the refraction angle at their interface. The sample's index can then be obtained following eqn [36]. Since n varies with wavelength, these techniques are normally performed with monochromatic light and provide n at a single wavelength. Among these methods are: the *Abbe refractometer*, using a glass hemisphere as a standard, the *V-block refractometer*, using a V shaped prism as a standard, and the *prism goniometer* based on the relative deviation between a known glass prism and a prism shaped sample.

Ellipsometry is another technique widely used to measure the index n. This technique is normally used on thin films deposited on substrate, as it allows to simultaneously measure the thickness and the refractive index. However, the method can also be applied to bulk samples. The principle of ellipsometry is based on measuring the change in polarization of a beam reflected off the sample surface as a function of incidence angle. The reflected light is elliptically polarized to an extent depending on n. This technique is especially useful to measure n at frequency range above the absorption edge where the sample is highly absorbing.

Metals

The optical properties of metals are mainly characterized by their very high reflectivity, which causes their shiny appearance. All metals reflect light in the infrared and visible region up to a cutoff frequency in the ultraviolet. This critical frequency is called the plasma frequency ω_p. The physical significance of the plasma frequency can be understood using the Drude–Lorentz model of the free electron oscillator. As mentioned previously, the free electron can be described accurately as dipole oscillators with a resonant frequency $\omega_0 = 0$. If we consider a system with negligible damping, we can introduce $\gamma = 0$ and by assuming all free electrons to be equivalent we can consider only one type of oscillators so that eqn [15] reduces to

$$n^2 = 1 - \frac{\omega_p^2}{\omega^2} \qquad [37]$$

where

$$\omega_p^2 = \frac{Nq^2}{\varepsilon_0 m_e} \qquad [38]$$

Equation [37] means that in the low frequency domain $\omega < \omega_p$, the refractive index n must be imaginary ($n^2 < 0$) and in the high frequency domain $\omega > \omega_p$, n is a real positive number. The low frequency region is therefore dominated by the extinction coefficient κ and the reflectance R, given by eqn [29], is essentially unity. In the high frequency limit, R decreases and n is real. These features are shown in Figure 5c. Metals are opaque and highly reflective below the plasma frequency, while they become transparent in the ultraviolet.

List of Units and Nomenclature

A	absorbance
c	light velocity in vacuum
D	electric displacement
E	electric field
I	intensity of light (power per unit surface)
k	wave vector
m_e	mass of electron
m_N	mass of nucleus
$\mathbf{n}$	complex refractive index
n	real refractive index

p	dipole moment
P	polarization
q	charge of electron
R	reflectance
S	scattering coefficient
T	transmittance
v	phase velocity
α	absorption coefficient
ε_0	permittivity of vacuum
ε_r	relative dielectric constant
κ	extinction coefficient
λ	wavelength
μ	reduced mass
χ	dielectric susceptibility
ω	angular frequency
ω_0	resonant frequency
ω_p	plasma frequency

See also

Instrumentation: Ellipsometry; Spectrometers. **Scattering:** Scattering from Surfaces and Thin Films.

Further Reading

Abeles F (1972) *Optical Properties of Solids*. North Holland Publishing Company.
Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge University Press.
Dragoman D and Dragoman M (2002) *Optical Characterization of Solids*. Springer.
Feynman RP (1970). *Lectures on Physics*, vol. I–III. Addison–Wesley.
Fox M (2001) *Optical Properties of Solids*. Oxford University Press.
Hecht E (1998) Optics, 3rd edn. Addison–Wesley.
Klingshirn CF (1997) *Semiconductor Optics*. Springer.
Nazarewick W, Rolland P, Dasilva E and Balkanski M (1962) ABAC chart for fast calculation of the absorption and reflection coefficients. *Applied Optics* 1(3): 369–370.
Palik ED (1985) *Handbook of Optical Constants of Solids*. Academic Press.
Savage JAR (1990) Signals radar. *Optical Materials* 1:323–419.
Simmons CJ and El-Bayoumi OH (1993) *Experimental Techniques of Glass Science*. The American Ceramic Society.
Simmons JH and Potter KS (2000) *Optical Materials*. Academic Press.

Optical Glasses

J E Shelby, Alfred University, Alfred, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

While advances continue in this field, some optical glasses have histories which can be traced to the seventeenth century. These glasses are produced by direct melting of raw materials of relatively high purity. Modern, or nontraditional, optical glasses are more commonly used to transmit light outside the visible region and are often made by nontraditional methods, including many variations on chemical vapor deposition, reactive melt processing, and sol–gel techniques. These glasses are used for ultraviolet and infrared lenses and other optical components, for optical fibers, for glass lasers, and for photonic devices such as optical isolators. Purity requirements and production demands for these glasses are significantly greater than those for traditional optical glasses.

Traditional Optical Glasses

Traditional optical glasses are primarily used as components such as lenses, prisms, or windows for optical systems operating in the visible region. These applications require glasses of high homogeneity, with precisely known refractive indices and dispersions. Glasses, as defined in this article, are nonmetallic, inorganic materials which have no long-range atomic structure and which display a temperature region of time-dependent properties.

Crown and Flint Glasses

Traditional optical glasses are designated as either crown or flint glasses. Crown glasses were originally based on the soda-lime-silica ternary system and were developed for use as window glass. The crown process involved spinning a spherical gob of molten glass to form a flattened disk which was much thicker in the center than at the edges. The edges of these disks were used for window panes, while the 'crown' in the center was often used as a crude lens. Crown glasses have low refractive indices and low to moderate dispersions.

The first flint glasses were potash-lead-silicate glasses prepared using English 'flint' as the source of silica. The presence of lead in these glasses results in

higher refractive indices and dispersions than those of crown glasses.

The designations of crown and flint have since been expanded to include glasses from many other compositional systems. Glasses with lower refractive indices and dispersions are designated as crowns, while those with higher refractive indices and dispersions are designated as flints. Modern crown glasses are based on silicate, phosphate, and borosilicate systems. Additions of fluorine are particularly useful for production of low refractive index, low dispersion glasses, and for adjusting these parameters to meet specific needs. The lowest combinations of refractive index and dispersion are found for glasses based on BeF_2, while the highest combinations of refractive index and dispersion are found in rare earth borate systems. Other common optical glasses contain barium or lanthanum, which can be used to produce either crown or flint glasses, depending on the other components of the glass.

The need for glasses with different combinations of refractive index and dispersion can be traced back to the development of the telescope. Use of lenses made of a single glass made it difficult to obtain sharp images. Newton recognized that this problem was due to divergence of light of different wavelengths while passing through a lens. He felt that this problem was inherent to lenses and could not be eliminated. Others eventually recognized that a compound lens containing a positive crown glass and a negative flint glass would result in cancellation of the divergences and produce a sharper image. While this development did not totally eliminate the colored fringes in telescope images due to the presence of the secondary spectrum, it did lead to major improvements in optical imaging. Binary compound lenses are still fashioned from a combination of crown and flint glasses, with much greater flexibility today due to development of glasses with much more varied combinations of refractive index and dispersion.

Notation for Optical Glasses

Since the refractive index and the dispersion are the two most important characteristics of an optical glass for use as a lens, a notation system has been developed to designate any glass by a six digit number. This notation system requires that the dispersion be expressed by the Abbe number, ν_d, which is defined as

$$\nu_d = (n_d - 1)/(n_F - n_c) \qquad [1]$$

where d, F, and c indicate the value of the refractive index measured at the yellow helium line at 587.6 nm, the blue hydrogen line at 486.1 nm, and the red hydrogen line at 656.3 nm, respectively. Since the Abbe number is a reciprocal dispersion, a large Abbe number indicates a low dispersion, while a small Abbe number indicates a larger dispersion. A glass is then designated by the expression

$$1000(n_d - 1)(10\nu_d) \qquad [2]$$

where each term is rounded to 3 significant digits. A common crown glass known as BK7, for example, which has a n_d of 1.516 800 and ν_d of 64.17 is designated as 517 642. A very high index flint glass, with $n_d = 1.807\,410$ and $\nu_d = 31.61$ would be designated as 807 316. This system can be used for all but the exceptional glasses with $n_d \geq 2.00$. Some tabulations are based on the green mercury line, or e line, at 546.1 nm instead of the yellow helium (d) line. The choice of the d or e line results in a small difference in the reported refractive index and Abbe number for any given glass (Figure 1).

Dispersion Formulae for Optical Glasses

The combination of a refractive index near the middle of the visible region of the spectrum and the Abbe number describe the contribution of an optical glass to the refractive power and chromatic aberration of a lens. This information was sufficient for many early applications of lenses and is still sufficient for the design of optical devices of modest expectations. More sophisticated optical devices require more information for proper design. The simple Abbe number is often insufficient for expressing the dispersion, since it is necessary to provide a

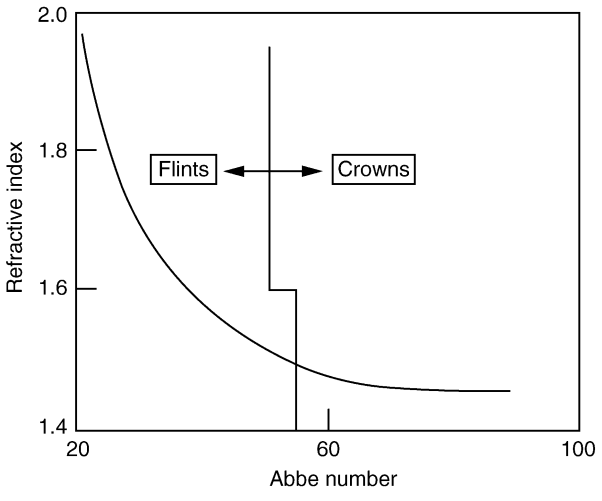


Figure 1 Relation between the refractive index and Abbe number of optical glasses. Commerical optical glasses lie in the regions labeled flint and crown above the curve.

detailed set of values of the refractive index as a function of wavelength across the entire visible region. The refractive index can be listed at a number of wavelengths, or described by an equation covering the entire wavelength range of interest. Although there are a number of expressions for the wavelength dependence of the refractive index, the most commonly used are the Cauchy dispersion formula, or

$$n = a + b\lambda^{-2} + c\lambda^{-4} \quad [3]$$

where a, b, and c are constants derived from a best fit of the data to this expression, and variations of the Sellmeier dispersion formula, which is often simplified to the expression

$$n^2 = A_0 + A_1\lambda^2 + A_2\lambda^{-2} + A_3\lambda^{-4} + A_4\lambda^{-6} + A_5\lambda^{-8} \quad [4]$$

where the six A_n constants are tabulated. The Cauchy dispersion formula is accurate to $\approx 10^{-4}$, while the modified Sellmeier dispersion formula is accurate to $\approx 3 \times 10^{-6}$. The validity of the latter expression can be extended further into the ultraviolet by adding a λ^{-10} term and into the infrared by adding a λ^4 term.

Annealing of Optical Glasses

Much of our knowledge of the behavior of glasses in the glass transformation region is based on studies of the annealing of optical glasses. Glass-forming melts have much higher thermal expansion coefficients than those of solid glasses of the same composition. During cooling of a melt, the density increases rapidly so long as the melt is relatively fluid. As the viscosity increases, the rate of structural rearrangement of the melt slows and eventually becomes fixed at some effective, or fictive, temperature. If the melt is cooled rapidly, the structure of the glass will be representative of a high temperature melt. If the melt is cooled more slowly, the structure will have time to equilibrate at lower temperatures and the glass formed will have the structure of a lower temperature melt and hence will be more dense. Since the refractive index of a solid is related to its density, the refractive index will be less for the glass produced by rapid cooling than for one produced by slower cooling.

There are two ramifications of the cooling effect on refractive index. First, the refractive index will be different for different cooling rates, even though the composition is identical. Statements regarding the refractive index of a glass are only applicable for a known thermal history. Second, but less obvious, the outside of a piece of glass will cool more rapidly than the inside due to the kinetics involved with removing heat from the glass. As a result, a refractive index gradient will exist in the glass, with a lower refractive index at the surface. If the glass is ground and polished to form a lens, the lens will not perform as expected if the shape is based on the assumption that the refractive index is constant through the entire lens. While cooling-induced differences in refractive index between the surface and center of a large block of glass are usually of the order of 0.000 1, they can be ≥ 0.001 for a rapidly cooled block.

Control of the cooling rate is essential for the production of a block of glass of uniform refractive index. The cooling rates used to reduce stresses to acceptable levels in the production of nonoptical glasses are much too fast for the production of optical glasses. Use of a constant cooling rate throughout the glass transformation region will not yield a block of uniform refractive index. Ideally, the cooling rate should be decreased gradually as the temperature decreases throughout the glass transformation region. The production of a fine annealed glass requires a very slow cooling process, with a gradually decreasing cooling rate with decreasing temperature within the temperature region where permanent changes in the structure of the glass can occur.

Athermal Lenses

While the thermal expansion coefficient of many glasses is relatively low, it is not zero. As a result, the dimensions and refractive index of a glass are functions of temperature. Temperature changes in an optical system therefore result in changes in optical path length. While changes in optical path length of this magnitude may be unimportant for optical systems used within a narrow temperature range, those used over ranges exceeding 10 to 20 K must be designed to minimize thermal effects. Production of athermal optical systems, i.e., systems of constant power over a range of temperatures, requires use of glasses where the thermal changes in refractive index are exactly offset by the changes in dimensions, i.e., the optical path length is independent of temperature. This condition is met if the temperature coefficients of the index and the thickness are equal and of opposite sign.

The overall effect of changes in temperature on the optical path length are tabulated in terms of the thermo-optical constant, G, given by

$$G = \alpha(n - 1) + \mathrm{d}n/\mathrm{d}T \quad [5]$$

where α is the linear thermal expansion coefficient, n is the refractive index, and $\mathrm{d}n/\mathrm{d}T$ is the thermal

coefficient of the refractive index. Perfect compensation between changes in index and dimensions are obtained when $G = 0$. Since α is positive for most optical glasses, with values in the range of 3 to 15 ppm/K, this condition requires a negative dn/dT. Negative values of dn/dT are found primarily for glasses containing some fluorine, such as fluoride glasses and those in which oxygen is partially replaced by fluorine, such as fluorosilicates, fluoroborates, etc., or for some phosphate glasses. Silicate glasses rarely, if ever, have a negative value of dn/dT. For the same glassformer, e.g., P_2O_5, use of modifier ions with large diameters and low field strengths favor increasingly negative values of dn/dT.

The value and sign of dn/dT is determined by the competition between the decrease in refractive index induced by an increase in molar volume and the increase in the refractive index induced by an increase in polarizability of the ions in the glass with an increase in temperature. Equation [5] indicates that the absolute value of the refractive index is also important in determining the value of G. If the refractive index is high, then dn/dT must be more negative to yield $G = 0$. As a result, most athermal glasses are crown glasses. The value of dn/dT is also wavelength-dependent, generally decreasing with increasing wavelength. In particular, dn/dT becomes more positive at wavelengths near the UV edge of the glass. An athermal lens is only perfectly athermal at the design wavelength.

A few special glasses and glass-ceramics used for optical applications, e.g., vitreous silica, titanium-doped silica, and some aluminosilicate glass-ceramics, have thermal expansion coefficients which approach zero or are slightly negative. Vitreous silica has a very low refractive index and thermal expansion coefficient and can be used to produce lenses which, while not truly athermal, have relatively low temperature sensitivity. These low thermal expansion materials are often used as substrates for coated reflective optics (mirrors), where dimensional stability is far more important than any change in the refractive index of the substrate.

Production of Traditional Optical Glasses

Quality requirements for optical glasses are much greater than those for other common glasses. A constant refractive index throughout the glass requires an extremely high degree of homogeneity. The glass must be free from defects such as striae (regions of refractive index variation due to density/composition variations) and stones (undissolved material) and the bubble content must be as low as possible. The thermal history of the glass must be such that the refractive index variation, due to differing fictive temperatures, is minimized. Internal stresses must be minimized to yield a low birefringence. The glass must also be free of any species which contribute to absorption in the desired spectral region.

The quality demands of optical glasses require use of much higher quality, and thus more expensive, raw materials. Concentrations of third-row transmission metal ions must be held to rigid specifications, with maximum values usually in the ppm, or even in the ppb range, for the highest quality glasses. Defects are reduced by melting in platinum containers instead of the oxide containers used in the past. Stirring the melt, using platinum wherever possible, is essential for obtaining acceptable homogeneity. Control of the melting atmosphere may be essential for controlling the redox state of the melt. Rigorous annealing schedules must be used to produce glasses with the desired thermal history so that the problems of stress-induced birefringence and refractive index gradients are reduced to acceptable levels.

Non-Traditional Optical Glasses

While most traditional applications of optical glasses involve transmission in the visible, many modern applications require high transmission in the ultraviolet, near-infrared, or infrared regions of the spectrum. Transmission in each of these regions is determined by both the bulk composition of the glass and by the concentration of impurities present. Most nontraditional applications place even more stringent requirements on the purity of the materials used in manufacturing the glass. An extreme example of the demand for purity is found in the glasses used for telecommunication optical fibers, where concentrations of some impurities must be held to the low ppb level.

Ultraviolet-Transmitting Glasses

The ultraviolet transmission of a glass is determined by the intrinsic electronic transitions of the material, by charge transfer absorption bands due to impurities, and by surface and bulk scattering. The first of these is determined by the electronic bandgap of the material, while the other two are controlled by the care taken during production of the glass. These glasses are used in spectroscopy for windows, lenses, UV-lamps, and sample cuvets, for microlithography, and for lenses in excimer laser optical systems.

Vitreous silica is widely used as an ultraviolet transmitting glass for transmission optics at

wavelengths well below 200 nm. The lowest energy absorption band for this glass occurs at 10.2 eV, or 122 nm. Many UV–VIS spectrometers, which normally are capable of measurements extending to 185 to 190 nm, use vitreous silica as optical elements. Additions of alkali or alkaline earth oxides to silica produce nonbridging oxygens in the network. Nonbridging oxygens do not bind the excitable electrons as tightly as bridging oxygens, which shifts the absorption edge to lower energies, i.e., longer wavelenths. These glasses are incapable of high transmission beyond 5 eV, or about 250 nm, and are not used for applications requiring good ultraviolet transmission.

Addition of elements which eliminate nonbridging oxygens favors good ultraviolet transmission. Alumina and/or boric oxide are frequently used in conjunction with alkali or alkaline earth oxides to counter the formation of nonbridging oxygens and preserve the ultraviolet transparency of silicate glasses. This approach can be taken to its logical conclusion by eliminating silica from the composition and producing alkali/alkaline-earth-aluminoborate glasses with good ultraviolet transparency.

In theory, fluoride glasses should provide better ultraviolet transmission than oxide glasses due to the larger bandgaps of fluorides. This assumption is true for single crystals, where fluorides such as CaF_2 are under development for transmission to higher energies than those obtainable with oxide crystals or glasses. In practice, the superiority of fluoride glasses for ultraviolet transmitting glasses has proven difficult to achieve. Vitreous BeF_2 and related glasses have superior ultraviolet transmission when compared to silicate glasses. Unfortunately, BeF_2 is very volatile and quite toxic, which restricts production of glasses based on BeF_2. Some fluoroaluminate and fluorophosphate glasses also have excellent ultraviolet transparency, but difficulties in producing these glasses in large sizes and of optical quality has limited their application.

Impurities are particularly important in production of ultraviolet transmitting glasses. The third-row transition metals, platinum, and cerium are all known to absorb strongly in the region from 200 to 300 nm when dissolved in glasses. Ferric iron has such a strong absorption band at $\approx$230 nm that concentrations of ferric ions must be kept in the low ppm range in glasses used for ultraviolet transmission. Since iron is a very common impurity in the sands used to produce silicate glasses, special care must be taken to minimize the Fe^{3+} content of ultraviolet-transmitting silicate glasses. Since ferrous iron has lower absorption in the ultraviolet, the effect of iron can be reduced by melting under reducing conditions. Platinum can enter melts from the melting unit and can be present as either ions or metallic particles. The ionic form of platinum absorbs ultraviolet light, while the metallic form scatters the same light.

Near-Infrared-Transmitting Glasses

Most glasses have good intrinsic optical transmission in the near-infrared. Absorption in this region is primarily due to impurities. Ferrous iron is especially detrimental to transmission in this region as a result of an extremely broad ligand field absorption band centered at about 1050 nm and extending into the visible and out to over 2000 nm. Reduction of absorption due to ferrous ions is obtained by careful control of raw materials to minimize the amount of iron oxide present and by melting under oxidizing conditions, which converts the ferrous ions to ferric ions. Conversion from ferrous to ferric ions increases the ultraviolet absorption of the glass. If good transmission in both the ultraviolet and near infrared regions is required, the iron content must be held to very low levels.

Although many rare earth ions also have absorption bands in this region, these elements are rarely found as impurities in sufficient levels to impair transmission in the near-infrared region for lenses and other common optical components. A more serious problem occurs for glasses used as optical fibers, where the greater optical path length magnifies the effect of impurities which may be undetectable in the spectra of samples only a few mm thick. The same can be said of absorptions of other ions which have very weak absorption bands in this region. Even higher order overtones/combination bands of infrared vibration absorption bands due to hydroxyl become important in optical fibers. While the hydroxyl absorption in infrared-transmitting glasses used in mm thicknesses is insignificant for concentrations less than 1 ppm, the hydroxyl concentration of optical fibers must be reduced to levels of a few ppb to reduce the effect of the overtone/combination bands to acceptable levels.

Infrared-Transmitting Glasses

Infrared transparency is limited by the intrinsic vibration, or multiphonon, edge of the glass and by the presence of impurities, primarily hydroxyl and carbon dioxide, which absorb in the infrared. The position of the infrared edge is determined by the force constant, which is proportional to the bond strength, and the masses of the ions. The best infrared transmission is found for materials with weak bonds and heavy ions. Infrared transmitting glasses are used in infrared spectrometers, lenses for pyrometers,

infrared transparent domes for aircraft and missiles, and optical components for many systems using infrared lasers.

The fundamental absorptions of B_2O_3, SiO_2, P_2O_5, and GeO_2 lie in the region from 7.5 to 12 μm. These bands have such strong absorptions that the first overtone, at half the wavelength of the fundamental vibration, effectively determines the infrared edge for samples of more than a few hundred micrometers in thickness. The best infrared-transmitting oxide glasses do not contain significant amounts of B_2O_3, SiO_2, or P_2O_5. The lower fundamental vibrational frequency of GeO_2 allows production of some glasses of limited use for infrared transmission to about 5 μm. Comparable transmission is obtained for calcium aluminate glasses. Better transmission has been obtained for glasses combining the oxides of lead, bismuth, or cadmium with gallium oxide. These gallate glasses are termed heavy metal oxide, or HMO, glasses and transmit to ≈7 μm in mm thicknesses. The HMO glasses transmit into the visible, with a cutoff at 400 to 500 nm. Since these glasses are transparent to visible light, they offer advantages over many other infrared-transmitting glasses which are opaque in the visible.

Recently, heavy metal halide (HMH) glasses have been found to exist in a wide range of compositional systems. HMH glasses transmit much further into the infrared than the best oxide glasses, with cutoff wavelengths extending beyond 20 μm. These glasses are much more difficult to prepare than oxide glasses and are highly prone to crystallization, less durable in water, and mechanically weaker than the HMO glasses. Improvements in infrared transmission correlate with degradation in the other properties needed for most applications. Oxide impurities severely degrade the infrared transmission of these glasses. While much greater infrared cutoffs can be obtained in the laboratory, glasses currently produced commercially transmit to ≈8 μm.

Chalcogenide glasses contain either S, Se, or Te as the glassforming anion and are free of oxygen. Commercial compositions transmit to 16 μm, while laboratory samples transmitting beyond 20 μm can be made. These glasses have refractive indices in the infrared in the range of 2 to 3, very low glass transformation temperatures, which limit their temperature range of application, and are, with few exceptions, opaque in the visible. Toxicity is also a problem for Se and Te. Oxygen impurities must be kept below the 1 ppm level to avoid unacceptable absorption bands.

All infrared-transmitting glasses suffer from transmission interference in the region from 2.5 to 6 μm, due to hydroxyl in the glass. Since hydroxyl readily forms in any melt exposed to an atmosphere containing water vapor, great care must be used to reduce the hydroxyl content to acceptable levels. Glasses can be melted in sealed containers, under reactive atmospheres containing fluorine or chlorine, or under vacuum. Some glasses are prepared under dry conditions and then remelted under a vacuum to reduce the hydroxyl content. Similar care must be used to exclude CO_2 from HMH glasses, where a band at 4.25 μm has been assigned to the asymmetric stretching vibration of the dissolved molecules (Figure 2).

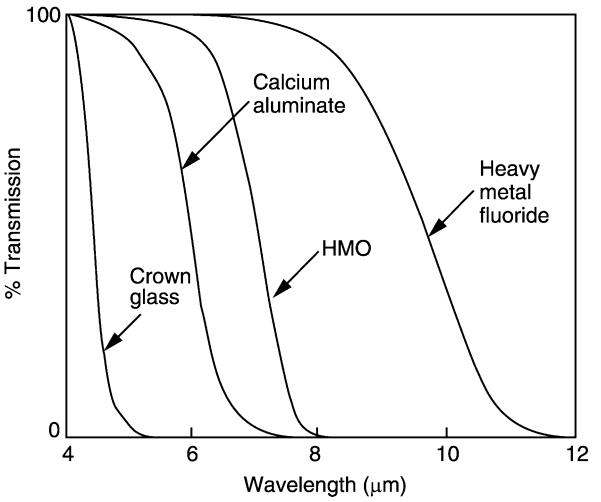


Figure 2 Infrared transmission spectra of several optical glasses showing the effect of glass composition on the infrared cutoff frequency.

Optical Isolators

Optical isolators based on the Faraday effect contain high concentrations of rare earths for paramagnetic devices and lead for diamagnetic devices. The highest concentrations of rare earth ions are obtained in aluminosilicate, aluminoborate, and aluminogermanate systems, where excellent quality glasses can be produced, while the best diamagnetic isolators are obtained using the lead bismuth gallate glasses.

See also

Optical Materials: Color Filter and Absorption Glasses. **Magneto-Optics:** Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping.

Further Reading

Deeg EW (1986) Optical glasses. In: Boyd DC and MacDowell JF (eds) *Commericial Glasses*, pp. 1–34. Columbus, OH: The American Ceramic Society, Inc.

Fanderlik I (1983) *Optical Properties of Glass*. Amsterdam: Elsevier Scientific Publishing Company.
Hoffmann HJ (1998) Differential changes in the refractive index. In: Bach H and Neuroth N (eds) *The Properties of Optical Glass*, pp. 96–123. Heidelberg, Germany: Springer-Verlag.
Lentes FT (1998) Refractive index and dispersion. In: Bach H and Neuroth N (eds) *The Properties of Optical Glass*, pp. 19–58. Heidelberg, Germany: Springer-Verlag.
Liepmann MJ and Neuroth N (1998) Ultraviolet-transmitting glasses. In: Bach H and Neuroth N (eds) *The Properties of Optical Glass*, pp. 288–297. Heidelberg, Germany: Springer-Verlag.
Liepmann MJ and Neuroth N (1998) Infrared-transmitting glasses. In: Bach H and Neuroth N (eds) *The Properties of Optical Glass*, pp. 297–308. Heidelberg, Germany: Springer-Verlag.
Pfander HG (1996) *Schott Guide to Glass*. London: Chapman & Hall.
Rawson H (1980) *Properties and Applications of Glass*. Amsterdam: Elsevier Scientific Publishing Company.
Shelby J (1997) *Introduction to Glass Science and Technology*. Cambridge, UK: Royal Society of Chemistry.
Vogel W (1991) Optical properties of oxide glasses. In: Ullmann DR and Kreidl NJ (eds) *Optical Properties of Glass*, pp. 1–36. Westerville, OH: The American Ceramic Society, Inc.

Plastics

T Bauer, Wahl Optoparts GmbH, Triptis, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Injection molded precision plastic optics in high volumes were first produced during the 1960s. After the development of sophisticated measuring and manufacturing methods in the late 1960s, precise aspheric surfaces were as easy to make as spheric contours in plastics. Today plastic is a widely used low-cost option compared to glass with even more degrees of freedom for optical and component design. Polymer optical systems are used in sensor applications, visual systems, cameras, mobile phones, video-conferencing cameras, scanners, security systems, and so on.

The physical and chemical properties of plastic materials are very different from those of glass. Generally speaking, glass materials are harder, more durable, and more stable against temperature and humidity than plastics. The variety of optical glasses comprises hundreds of different materials. Compared with this the choice for plastic materials is limited only to about 10 different materials (and even less optical parameter variations). However, plastic optics offers other design freedoms that are not achievable with glass optics.

The manufacturing technologies for glass and plastic optics are totally different. Glass lenses are made by a grinding and polishing process whereas precision plastic lenses are usually made by injection molding, compression molding, or casting. Because of the material's characteristics and the manufacturing process, plastic optics have some unique advantages.

High production numbers at low costs

Injection molding is ideal for high-volume production with low-unit costs. Moderate raw material costs and multicavity mold designs (up to 32 cavities) allow large production volumes at a reasonable unit price. In spite of considerable tooling costs, these costs compared to glass design versions, can be relatively low.

Lightweight and hardiness

For a given volume glass is much heavier than plastic (by a factor of 2.3 to 4.9). However, plastic materials are relatively shatter and impact resistant. These features are important for head-mounted systems and other weight-sensitive applications (mobile phones, cameras, etc.).

Design potentials

Injection molding makes it economical to produce sophisticated optical shapes such as aspheres, diffractive optical elements, or even freeform surface structures. From the design point of view, the more sophisticated surface shapes reduce costs or obtain better performance (Fresnel structures, lens arrays, diffractive optical elements, etc.).

Optical systems and component assembly

For typical optical system designs optical components (mirrors, lenses, prisms, etc.) must be fixed in a mounting. With plastic optics it is possible to mold mounting elements, posts, or alignment notches integrally with the optical component. This can reduce part and assembly costs considerably.

Technologies adapted to plastic materials, such as ultrasonic and laser welding, gluing and integrated

snap-in structures, allow fast and cost-efficient automated and manually operated assembling solutions.

History of Transparent Plastic Materials

A Material With No Use

In 1840, the Austrian chemist Redtenbacher (1810–1870), a student of Justus Liebig, discovered during a chemical experiment a malodorous transparent liquidity, the acrylic acid. It was an unwanted byproduct which was not further analyzed at that time. Nevertheless, that discovery can be considered to be the origin of today's plastic optics.

The development of plastics began in the US in 1860 as people were searching for an alternative for ivory in the production of billiard balls.

In 1869 John Wesley Hyatt developed from cellulose nitrate and camphor the first plastic material: celluloid. At first it was used for dental plates and afterwards for shirt collars and as a film base material. Despite its flammability and its lack of light resistance, it was a first economical success.

In the following decades the first synthetical plastics were developed, for example, the material with the trade name Bakelite in 1906.

Otto Röhm's Doctorate

The basis of acrylates was rediscovered in the records of Redtenbacher in 1900. From 1901 onwards, Prof. Pechmann (1850–1902) asked his postgraduates to do more research into the acrylic acid. The acrylic acid is an acetylene compound which can be extracted from carbon, petroleum, or lime.

In 1901 Otto Röhm, a student of Pechmann, wrote his dissertation 'About polymerisation products of the acrylic acid' – the basis of the methacrylate chemistry.

Hermann Staudinger (1881–1965), head of the Institute of Chemistry in Freiburg started theoretical investigations into the structure and natural characteristics of natural and synthetical polymers in 1920. He assumed that the molecules of plastic materials consist of numerous small molecule units. Experiments as proof of the theory led to a rapid development of scientific research and to significant breakthroughs in plastic chemistry. Staudinger won the Nobel Prize for Chemistry for his work in the field of macromolecules in 1953.

In 1928 methyl methacrylate was syntheticized for the first time and patented. Acrylic resin was used as a binder in multi-layer glass.

In the 1920s and 1930s a number of plastic materials were developed, including cellulose acetate CA (photo film, synthetic fiber); polyvinyl chloride PVC (pipes, coatings, isolations); urea formaldehyde resin (dishes, electrical equipment).

Polystyrene resins were commercially produced for the first time in 1937. Polytetrafluoroethene (PTFE) was produced for the first time in 1938, and distributed as 'Teflon' since 1943. Another key development was the synthesis of Nylon, the first technical high-performance plastic material.

Plexiglas Captures the World

The industrial production of methacrylic acid methyl ester began in 1934. With its polymerization to a hard and transparent plastic material, polymethylmethacrylic (PMMA or acrylic) was developed. It was patented in 1936 and caused, under the registered trade name Plexiglas®, a sensation at the world exhibition in Paris in 1936, almost 100 years after the discovery of acrylic acid.

Plates of Plexiglas were manufactured by means of effusion and polymerization between glass plates. At first it was used for watch glass, cockpits, and the construction of street lighting and luminous advertising. People already experimented with Plexiglas injection moulding in 1935. Arthur Kingston, the founder of Combined Optical Industries Limited (COIL) had invented and patented the first plastic lens in 1933.

The First Plastic Lenses in Cameras

The first camera, incorporating plastic optical parts, to be mass produced was the 'Purma Special' introduced by RF Hunter Ltd., in 1937, using 'Perspex' (another brand name for Acrylic), for the direct-vision viewfinder.

Due to the scarcity of natural raw materials at the beginning of World War II, the plastic industry became a source of outstanding substitute materials. This development had a lasting effect far into the post-war period. Results were mainly achieved with technical plastics, such as polycarbonates, acetates, and polyamides. Other plastics were developed as a substitute for metals as well as for hardhats, for high-temperature constructions, and for products which are resistant against acids and bases.

The German chemist Karl Ziegler discovered polyethylene (Polyethen, PE) in 1953 and the Italian chemist Giulio Natta Polypropylene (PP) in 1954.

The Eastman Kodak company is credited with giving birth to the modern-day plastic Fresnel lens. In 1946 Kodak developed a tooling and manufacturing process to mass produce plastic Fresnel lenses.

The circular Fresnel lenses made at that time had spiral grooves like music records.

It took another 10 years of improvement in injection molding techniques to the first plastic lens camera – the Kodak Brownie 44A from 1959.

Between 1953 and 1959 the Polycarbonate polymer was developed by Bayer and General Electrics. Currently acrylics and polycarbonates are still the most important materials for plastic optics.

But not only thermoplastic materials were developed over this time. In May 1940 research work resulted in the monomer allyl diglycol carbonate (ADC). In the following years more than 180 different compounds of this monomer were investigated and examined.

During World War II Columbia Southern Chemical Company (a subsidiary of Pittsburgh Plate Glass) began research into ways of obtaining nonthermoplastic materials. The project was given the name 'Columbia Resins'. The 39th tested (CR 39) turned out to be the most significant due to its unique properties and was used since 1947.

At first, solely the low-refractive CR 39 was available for organic eyeglasses. Only in the mid-1980s did the development begin in the field of high-refractive nonthermoplastic materials.

Properties of Plastics for Optics

As already mentioned, the choice for plastic optical material is limited only to about 10 different types of material. Optical properties (abbe value, refractive index, transparency) as well as mechanical, thermal, and humidity boundary conditions are decisive for the material choice. Different to glass during plastic processing (injection molding, casting, etc.) the process affects not only the geometry but also inner properties like refractive index, transparency, and birefringence.

Optical Properties

Fundamental optical properties are defined by optical transmission, refractive index, and dispersion. For plastic optics birefringence is an important parameter too.

Although the total number of plastic materials has increased in recent years, the range of refractive index and dispersion characteristic is limited to almost two major groups – crown-like materials such as Acrylic (PMMA), Polyolefin (COC, COP) and flint-like materials such as Polystyrene (PS), Polycarbonate (PC) and Styrol-Acrylnitril-Copolymer (SAN).

This strongly limited variety of plastic optical materials significantly restricts the optical design freedom (Figures 1 and 2).

Physical Properties

Important physical properties are weight, impact, and abrasion resistance – and thermal properties like temperature resistance and thermal expansion. Already, during system design, both mechanical and thermal properties have to be taken into consideration. The thermal expansion of optical plastics is approximately ten times higher than that of glass materials. In an optical system this effect has to be compensated for by optical design or mounting.

Typically most optical plastics can withstand temperatures up to 90°C. The maximum service temperature of Polycarbonate and Poly-Olefin materials reaches 120°C.

The specific gravity of plastic optical materials ranges from 1 to 1.3.

Polycarbonate has the highest impact resistance of all optical plastics and so is used for windshields and crash helmets. Acrylic has the best abrasion resistance (Table 1).

Common Materials

Thermoplastics

Thermoplastics are polymers with a linear molecule structure – with or without side chains – which can be deformed reversibly without any modification of their thermoplastic characterization by means of the impact of heat. Optical elements can be manufactured by injection molding or hot embossing.

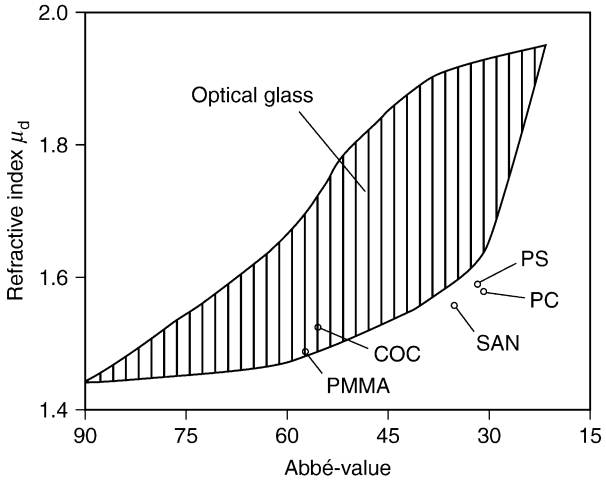


Figure 1 Abbe diagram.

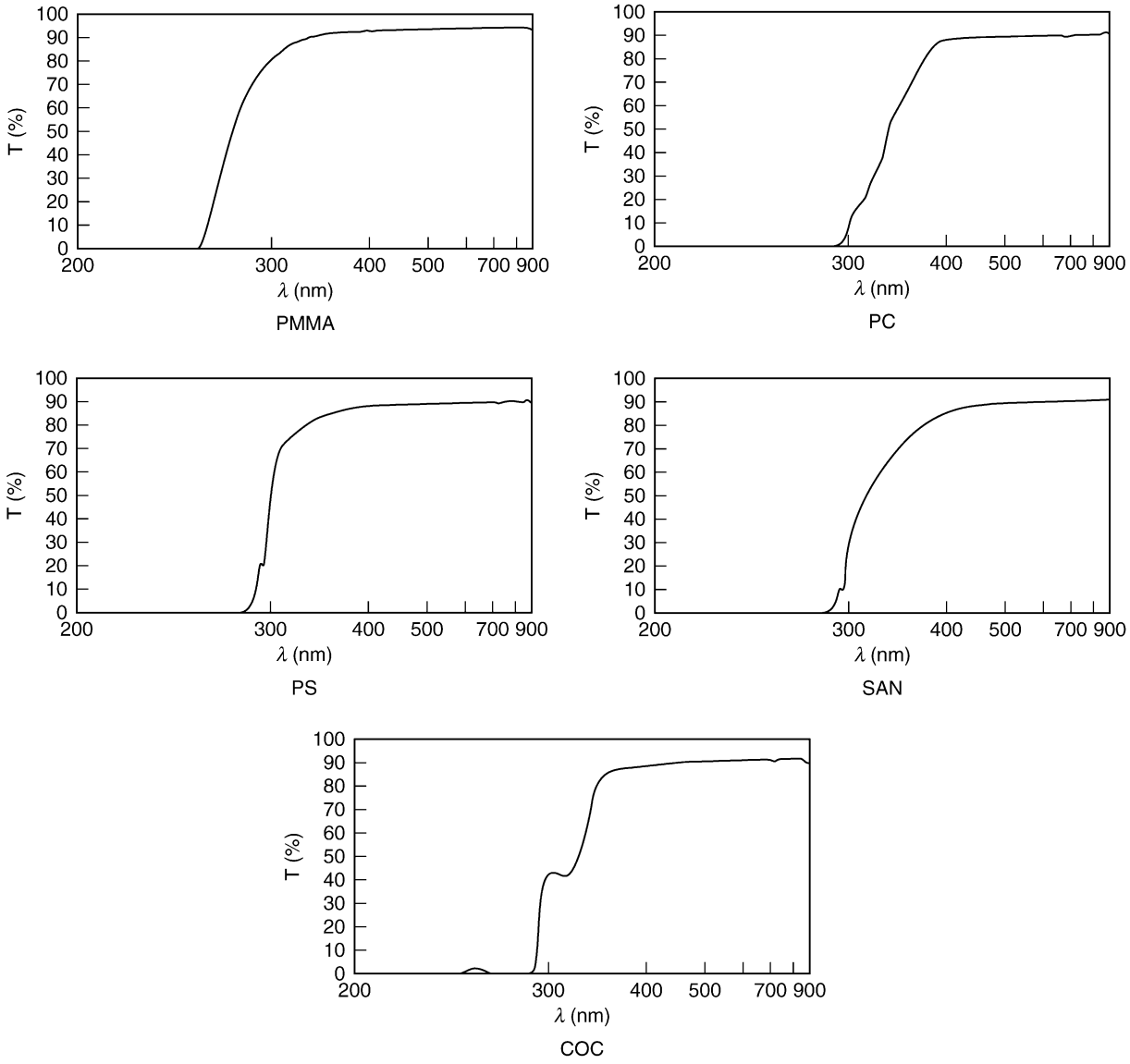


Figure 2 Transmission characteristics.

Typical optical thermoplastic polymers are listed below.

Acrylic or PMMA (Polymethyl methacrylate)
Acrylic is the most commonly used optical plastic. It is moderately priced, easy to mold, scratch resistant, and not very water absorptive. Its transparency is greater than that of most optical glasses throughout the visible range of the spectrum (organic glass). PMMA can replace glass in every application as long as the service temperature is less than 90°C and a low chemical resistance is required. Additives to acrylic (as well as to several other plastics) considerably improve its ultraviolet transmittance and stability.

PC (Polycarbonate)
Polycarbonates are linear polyesters of the carbonic acid which combine many characteristics of metal, glass, and plastics.

Polycarbonate is very similar to styrene in terms of optical properties such as transmission, refractive index, and dispersion. Polycarbonate, however, has a much broader operating temperature band of up to more than 120°C. It is used as the flint material for systems that have to withstand severe thermal conditions.

Another advantage is the high-impact resistance of polycarbonate. Safety glasses and systems requiring durability often consist of polycarbonate. Because of its high ductility polycarbonate is not easily machined.

Table 1 Material properties

Material	*Characteristics*	*Acrylic (PMMA)*	*Polystyrene (PS)*	*Polycarbonate (PC)*	*Styrol-Acrylnitril (SAN)*	*Cyclic Olefin Polymer (COP)*	*Cyclic Olefin Copolymer (COC)*	*Polyether-sulfone (PES)*	*Acrylnitril-Butadien-Styrolco-polymere (ABS)*	*CR39*	*Optical glass (B7)*
Optical	Spectral Passing Band (nm)	390–1600	400–1600	360–1600	395–1600	300–1600					
	Refractive index at 587 nm and 20 °C	1.4918	1.5905	1.5855	1.5674	1.5261		1.6600	1.5380	1.501	1.517
	Abbe value $(nD-1)/(nF-nC)$	57.2 (53)	30.8	34 (29.9)	34.8	55.8	58	19.4		58	64.4
	Transmittance (%) thickness 3.2 mm	92	87–92	85–91	88	92	92	80	85		
	Haze% thickness 3.2 mm	1.3	2–3	1.7	1.5	1.5	1.5	1.5			
Physical	Specific Gravity (g/cm)	1.18	1.06	1.25	1.07	1.01	1.03	1.37	1.05	1.32	2.53
	Max. Service Temperature (°C)	92	82	124	95	123	130	200	90	130	400
	Linear Expansion Coefficient $(1/K)$	6.8×10^{-5}	8.0×10^{-5}	6.6×10^{-5}	7.0×10^{-5}	7.0×10^{-5}	7.0×10^{-5}	5.5×10^{-5}	8.5×10^{-5}		1.1×10^{-4}
	Abrasion Resistance (1–10)	10	4	2		6	6				
Environmental	$dn/dT\ (\times10^{-5})$	−8.5	−12	−11.8…−14.3		−8	−10.1				
	Sensitivity to Humidity	high	low	low	medium	low		high	medium		
	Water Absorbtion (weight %) 23 °C, ISO 61	0.30	0.20	0.15	0.30	<0.01	<0.01	0.70	0.45		0
Manufacturability	Processability	excellent	good	poor	excellent	good					
	Birefrigence	low	high	high		low					
Chemical	Resitance to Alcohol	limited	good	limited				good			
Costs	approx. Material costs (EUR/kg)	3.3	2.5	4.4	4.4	27.1		21.0	3.5		25.0

PS (Polystyrene)

Polystyrene or Styrene is a low-cost material with excellent molding properties. Styrene has a higher index and a lower numerical dispersion value than other plastics. It is often used as the flint element in color-corrected plastic optical systems.

Compared with acrylic, styrene has lower transmission in the UV range of the spectrum. It has a lower resistance to UV than acrylic but is more easily scratched than acrylic.

Because its surface is less durable, styrene is more typically used in nonexposed areas of a lens system. A styrene lens, when paired with an appropriate acrylic lens, offers an effective achromatic solution.

SAN (Styrol-Acrylnitril-Copolymer)

SAN is a copolymer of certain fractions of Polystyrene and Acrylic (typically 70/30), which allows adjusting refractive indexes. It is a glassy polymer with a low tendency to stress cracks. SAN has slightly more chemical durability than PS and is a low-cost material with excellent molding properties.

ABS/MABS (Acryl-Butadien-Styrol-Copolymer)

ABS is developed from polymerization of styrene and acrylnitrit on polybutadien during the emulsion process. Process variations lead to different material characteristics. Compared to PS, the chemical resistance and temperature stability of ABS is considerably better. ABS has a natural color so it cannot be used for every optical application.

PES (Polyethersulfon)

PES shows a remarkable temperature stability, a low transmission against UV-light, a limited transmission with a natural yellow color, and cannot be used for every optical application. Because of its excellent temperature stability, PES is suitable for assembling, even on electronic motherboards by reflow soldering. This process causes temperature stress up to 220°C (Figure 3).

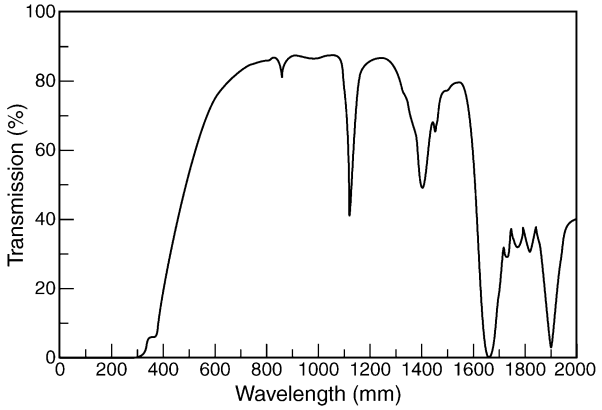


Figure 3 PES.

COP and COC (Cyclic olefin polymer and copolymer)

Cyclic olefin (co-)polymer provides a high-temperature alternative to acrylic. Its refractive index and transmittance are similar but the heat distortion temperature is about 30°C higher than for acrylic. It has a remarkably low water absorption value of less than 0.01% (compared with Polycarbonate 0.2% and PMMA 0.3%).

Cyclic Olefin Polymer (COP) is a family of amorphous plastic resins with low native stress birefringence properties. Zeonex is a resin developed by the Nippon Zeon Company – one grade, E48R, has a heat distortion temperature of about 122°C. Zeon Chemicals has another class of polymers called Zeonor, that can be provided as a lower-cost alternative to Zeonex.

Cyclic Olefin Copolymer (COC), with the trade name of Topas, is another high-temperature alternative to acrylic. Generally, it is a lower-cost alternative to COP. Particularly favorable is the high flowability (good injection moldability) of these thermoplastics.

PMP (Polymethylpenten – TPX®)

PMP has a structure similar to PE, except that only the methyl groups are replaced with isobutyl groups. The chemical resistance can be compared to PP, although it tends to stress crack under the impact of ketones or chlorinated solvents. The main advantages of PMP are its excellent transparency and its good mechanical characteristics, even at high service temperatures of up to 150°C.

PE-HD (high density polyethylene)

If polymerization is controlled catalytically, molecules with a low number of side chains will be received. Compared to PE-LD (low density), the molecules are quite compact with a higher stability and chemical resistance and service temperature can be up to 105°C. PE shows a bad transparency (milky appearance) in the visible spectral range. However, this material is permeable for infrared beams and so is frequently used for motion detectors.

Elastomere

Elastomeres are plastics with loose and networked molecules which are rubber elastic at a normal temperature. The most popular elastomeres are natural rubber and silicone rubber. These materials are optically used as flexible light guides.

Duromeres

Organic glass is a fully synthetic plastic material available in a vitreous state. It consists of macromolecular organic compounds which do not follow any principle of periodic arrangement and are hence amorphous.

Duromeres are plastics with spacial closely networked molecules which are very hard and refractory at a normal temperature. Impact of heat causes an irreversible induration. In most cases, duromers are used to produce plastic lenses made of organic glass. Once they have been thermally treated after production, their shape can no longer be changed.

The well-known plastic CR 39 is one of the organic materials used for plastic lenses.

CR39 plastic is made by polymerization of the allyl diglycol carbonate (ADC). It is a transparent, thermosetting resin which combines an exceptional range of qualities which are not available in other plastic transparent materials. The CR39 plastic is colorless and completely transparent to the visible light and almost completely opaque in the infrared and ultraviolet region of the spectrum. For these reasons, it is largely used for the production of sunglass lenses. They can be colored by surface dyeing or bulk tinting, have high abrasion resistance, and high-quality optical properties. Their weight is about half as much as glass, they keep their excellent optical properties despite long-term exposure to chemicals and resist heat distortion up to 100°C. These qualities, combined with the exceptional optical characteristics, make CR39 the best choice for applications where severe conditions of use exclude all other optical materials.

Additives and Colors

The characteristics of many plastics, particularly of thermoplastic materials, can be changed by adding so-called additives.

UV-filter materials are frequently added to improve the stability of materials in the sunlight. Sometimes nonstabilized plastics tend to yellow (transmittance decreases). Color filters can also be generated by coloring. Infrared coloring permits the production of optical elements which appear black in the visible spectral range. These materials are particularly used for sensors, optical scanners, or remote control front windows. Even characteristics, such as temperature stability, can be influenced by additives.

Manufacturing Methods

The manufacturing processes for glass and plastic optics are totally different. Glass lenses are made by a grinding and polishing process. By contrast, typical manufacturing methods for precise technical plastic optical parts are diamond turning of plastic blocks, injection, and compression molding of granulates. Other methods, such as casting, have their application in special fields like ophthalmics (eyeglasses). In any case, processing transparent plastic materials for optical purposes should take place in a clean environment.

Diamond Point Turning

Diamond point turning is an ultra-precision machining process carried out on special high-performance, ultra-precision diamond CNC-machining systems.

In combination with vibration isolation systems, digital signal processor-based machine control and integrated measuring systems with nanometer programming resolution manufacturing of optical surfaces is possible. This method can be expanded to full 3D-milling systems. With such an arrangement the generation of free-form surface profiles is possible.

Because of the long production time and the high machine costs, this technology is used for prototyping in plastics and nonferrous metals, mold inserts, and series production plastic parts – which cannot be manufactured by injection molding (because of size or precision demands).

Injection Molding

Injection molding produces several parts per shot (molding cycle) in single or multicavity tools. This production method is used for most of the plastic optic parts. Other fabrication techniques are used only when molding is inappropriate.

A plastic injection-molding machine consists of a fixed platen, a moving platen, a clamping unit, and an injection unit. Molding of optical parts requires special machine configurations and auxiliary equipment.

Compression Molding

Compression molding is used in the making of Fresnel lenses or other micro-optical structures. The material is pressed between heated platens with accurately defined temperature cycling during pressing. Mold inserts are formed by electroplated copies, replicated from master structures. These masters can be, for instance, diamond turned structures or diffraction gratings made by holographic methods.

The compression molding process allows realizing small structures with high aspect ratios and tight angular and positioning tolerances.

Manufacturing Process

Optical and System Design

On the one hand, optic design for plastic optics in principle uses the same mathematical algorithms as optic design for glass. On the other hand, designing plastic optics requires a profound understanding of the material properties and the manufacturing processes. Knowledge of production technologies, material characteristics, and assembling methods, together with design expertise, are needed to fully exhaust what precision plastic optics can be. Simply substituting the indices of refraction and re-optimizing, the design will not succeed. Expert design assistance is essential at this stage.

Designing plastic optics with modern design tools allows greater design freedom. The advantage of combining integral mounting structures with the optical surfaces to create mounting flanges, alignment, and snap features, provides the ability to automated assembly. Aspheric, cylindric, or toroidal surfaces are as easy to realize as spherical ones. Microstructures such as diffractive optical elements can be integrated too.

Prototyping

After designing a plastic optical system, lens prototypes can be made by diamond point turning in various plastic materials. The best surface finishes can be obtained with PMMA. Materials such as polycarbonate, pleximid, or Zeonex do not yield smooth surface finishes.

A major problem is the availability of semi-finished plastic blocks in various materials. For PMMA, a wide variety of bars or plates are available. For other materials or colored options, semi-finished items have to be made by injection molding. Because of the high manufacturing costs, diamond point turning is only recommended for making a limited number of prototypes to verify functionality of the optical design and to perform first tests.

In this stage optical systems are normally assembled from single elements. Housing parts often are made from aluminum. The resulting surface quality and system performance cannot be a validation of the manufacturability by injection molding. To get reliable knowledge about this, making a molded prototype from a single-cavity prototype mold is recommended.

The Injection Mold

A high-quality injection mold is obviously essential for precise plastic optic parts. The parts can never be better than the tool – but good tooling, however, does not guarantee good parts. A strong understanding of the whole manufacturing process is the key to producing precision plastic components.

Any injection mold consists of three main parts. The upper half, which is affixed to the injection-side platen, the lower mold half, which is affixed to the ejector-side platen, and the mold-ejection mechanism. Guide pins and taper locks ensure proper alignment of the mold halves.

Production volume and precision required from a tool influences the selection of mold materials and built-in maintenance features.

Single-cavity and multi-cavity molds from two to eight, 16, or even 32 cavitys are used for plastic optic parts.

Thermoplastic materials shrink during cooling in the mold. This geometric effect has to be compensated for in the injection mold. Exact shrink rates can be calculated and the tool can be modified.

For manufacturing optical plastic parts by injection molding, the optical surfaces (plano or aspherical shapes, diffractive, conical, lenticular, and cylindrical surfaces) are generated as separate inserts in the tool.

Aspheric inserts are manufactured by two steps. First a best-fit curve is generated on a stainless steel substrate. The substrate is then subjected to a nickel-plating process (electroless nickel) that deposits a thin layer of nickel (up to 500 μm). In the second step, single-point diamond turning produces the final aspheric or diffractive curve in the nickel. Because the hardness of a nickel-plated insert is less than that of a steel spherical insert, it will be more susceptible to scratch defects (Figure 4).

Pre-Production

A pre-production stage is recommended to check the manufacturing process. Typically this is done by using a single-cavity prototype mold. This mold can be used to find optimal molding conditions.

Optical and mechanical design can be verified with real molded components and design revision is possible to affordable conditions. Often the prototype molds is used to start limited production since production tooling may take much more time.

Series Production

For series production of high volumes, multicavity production molds are required. Depending on quality,

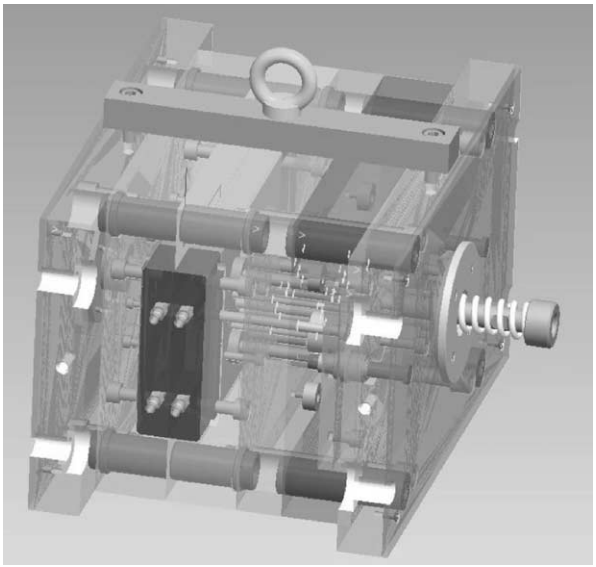


Figure 4 Injection mold.

volume, throughput, and cost, the production tool may have 2, 4, 8 or even up to 32 cavities. The production molds function for at least several hundred of thousands of injection cycles.

Coating and Component Assembly

Coating

Because of their limited temperature and UV resistance, plastic lenses must not be coated in an elevated temperature and radiation environment. During the deposition of thin films onto plastic, the coating chamber temperature is significantly lower than that for glass optics. This requires deposition techniques such as ion-assisted deposition to apply antireflective, conductive, mirror, and beamsplitter coatings.

Today, multi-layer dielectric coatings are routinely deposited on plastic components. Typical broadband antireflection coatings reduce reflection to about 0.5% per surface across the entire visible spectrum. Narrowband, multilayer antireflection coatings can achieve surface reflectances less than 0.2%. Multilayer dielectric coatings can be modified easily to scratch-resistant designs for front lenses and windows.

Several front and back surface reflector coatings are available for plastic substrates. Standard coating metals include aluminum, silver, and gold. Aluminum coatings provide surface reflectances greater than 88% across the visible spectrum and gold coatings greater than 95% for the near-infrared region.

Component Assembly

Many advantages of using plastic optics will come out in manufacturing components. The integration of optical, mechanical, and electronic elements allows building low-cost polymer optical components such as small camera lenses or scanner heads.

Plastic optical assemblies are usually carried out in a clean room environment by manual, semi-, or fully automated processes. The components have to be designed to ensure ease of assembly. Snap-on features, UV-cementing, heat-staking, and ultrasonic and laser welding, can also be employed with plastic materials.

Since most optical tolerances are additive, it is essential to establish quality check points in the manufacturing process to sort out nonconforming sub-assemblies.

Automatic in-line optical performance, monitoring such as MTF testing, can be implemented. SPC techniques should be used here to ensure the process is not drifting out of the controllable range.

Summary

For centuries, glass proved to be the material of choice for all optical applications. During the twentieth century, applications for glass elements and complex optical systems greatly expanded. Early uses of plastic molded elements included toy objectives, gauge windows, and watch crystals. Throughout the 1970s and up to the present day, high-grade polymers were developed specifically for optical applications. These advancements in materials, coupled with improved mold design have enabled plastic optics to replace glass optics in a wide and growing number of applications.

Plastic optic technology permanently expands their traditional limits and fields of application. On the one hand, plastic optic is still limited by material properties but, on the other hand, design and assembling freedoms allow new approaches.

See also

Diffraction: Fresnel Diffraction. **Geometrical Optics:** Lenses and Mirrors. **Optical Coatings:** Thin-Film Optical Coatings. **Optical Materials:** Color Filter and Absorption Glasses.

Further Reading

Corning Precision Lens Inc. (2000) *The Handbook of Plastic Optics*, 2nd edn. Cincinatti, OH: Corning Precision Lens Inc.

OPTICAL MATERIALS

Sculptured Thin Films

K Robbie, Queen's University, Kingston, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical materials, like all materials, ultimately derive their physical properties from their chemical composition and physical form. Aluminum behaves differently from copper because of a chemical difference, while graphite behaves differently from diamond because of a crystallographic difference. For materials that are not infinite perfect crystals, an important contribution to their electromagnetic (optical) response comes from the morphological form of the material. Morphological form describes the physical structure of the material: a perfect crystal is one type of morphological structure, a porous fibrous bundle is another. An example of a morphological difference is that between snow and ice. Snow and ice are chemically identical, composed of solid water, but they exhibit different optical properties because of a difference in morphology. The polished glasses and large perfect crystals that comprise most optical components have morphologies more like ice than snow; their morphologies are dense and crystalline, with few and small disordered 'defects'. Unlike ice, snow is rich with morphological structure (see Figure 1) arising from the complex growth of each flake in the clouds, and the aggregate growth of snow on the ground. The morphology of snow scatters all colors of visible light, giving it an opaque white appearance, while ice is transparent and nearly colorless.

Sculptured thin films (STFs) are materials with controlled and designed morphology that exhibit a rich behavior unlike that of the constituent materials in bulk; they are the snow to most optical materials' ice. Unlike snow, sculptured thin film morphology is designed to scatter light in a controlled manner to produce a desired optical response. A sculptured thin film could be a coating applied to an optical element where the coating converts polarizations of light, circular or linear, but is composed entirely of an optically isotropic material (which doesn't usually affect the polarization of light). The film morphology is designed to create the desired optical properties. Sculptured thin films are also useful for magnetics, where an example of a sculptured thin film is the metal film coated onto plastic tape for video, audio, and data recording. The film is deposited under conditions that impose an elongated shape onto the magnetic domains, altering the magnetic properties. The morphological form of a sculptured thin film defines the way it responds in electromagnetic, chemical, and biological interaction, and there is hope that these unique materials will someday form the core of technologies that make the world a better place.

Structure in Thin Films

As chemical composition and morphology determine the electromagnetic response of materials, it would be wonderful to have a technique where we could build materials atom by atom, placing each exactly where we wished, and varying the element at will. With this level of control it would be possible to design and fabricate materials for almost any conceivable optical application. In some sense this is what is done now to fabricate lasers and optical switches. Microelectronic and micromachining processes have produced fantastic optical devices by controllably mixing materials (mostly semiconductors) and patterning designed structures. These techniques typically employ a repeated series of thin-film depositions and photolithographic patterning. Examples of thin-film coatings include antireflection coatings on eyeglasses, hard coatings on grocery store barcode scanners,

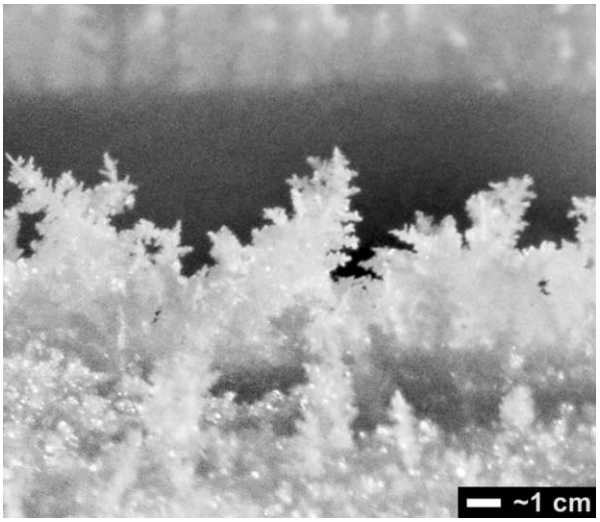


Figure 1 Photograph of snow accumulating at −20 °C. Photo courtesy of Chelsea Elliott.

ultraviolet blocking aluminum inside potato chip bags, and all integrated electronic circuits. Thin films are usually produced by vapor deposition in vacuum, and can be made with most elements and inorganic compounds (and a growing list of organics). Most thin films grow as uniform, dense, two-dimensional (2D) layers, and three-dimensional structure (3D) is produced by removing portions of the film with photolithography and etching. Photolithography is a process where ultraviolet light is projected through a mask and microscope system onto a photosensitive polymer layer atop the thin film. The light hardens areas of the polymer in the pattern of the mask, preventing them from dissolving in a chemical developer solution. Exposed portions of the thin film can then be etched away with chemical or other means. Repeating as desired, many steps can create complex 3D designed structures. Computer chips and other integrated circuits are semiconductor crystals covered with a layered structure of conductors (aluminum, copper, platinum) and insulators (metal and semiconductor oxides) patterned with thin film deposition and lithography. Semiconductor lasers and many other optical devices are made this way as well, usually employing compound semiconductors (GaAs, AlGaAs, GaAsP, InP, etc.). The micrometer-scale patterning made possible with thin-film deposition and etching techniques has produced a wealth of semiconductor optical devices.

Optical lithography, however, is limited in size scale to dimensions comparable to the wavelength of light used to expose the photomask. Modern UV lithography can pattern 100–200 nm structures. Visible light has wavelengths from 400 nm (violet) to 700 nm (red). Shrinking the dimensions of physical structures is a great benefit for optical devices as it allows light to be controlled coherently without strong random scattering. Structures with dimensions comparable to the wavelengths of light strongly scatter light waves; scratches on optical components or imperfections in materials with dimensions comparable to the wavelength are considered defects, and are usually deleterious. However, materials with morphology much smaller than the wavelength (nanometer structure, or nanostructure) are seen by propagating light as homogeneous, with polarization and dispersion effects characteristic of the underlying structure. Sculptured thin films are materials where complex nanostructure has been created by controlling growth parameters dynamically during deposition of a thin film. The simplest sculptured thin film might be one where the temperature is periodically varied during thin-film deposition (although this is a very simple example, lacking the anisotropy central to the sculptured thin-film concept). Variation in the growth process with temperature affects film density, producing a porosity variation (periodic inhomogeneity) as a function of film thickness. Varying porosity produces varying refractive index, and this film would be an optical interference filter. The largest limitation of sculptured thin films is that all attainable structures must be characteristic of the growth process. As all morphological control is accomplished by varying film growth parameters, structure that is not characteristic of thin film growth is not possible. Fortunately, thin film deposition and growth is sufficiently complex to allow the fabrication of a large range of useful structures. As with lithographic patterning, the structural control is two dimensional, but now the third dimension is constructed by continuing the deposition, not by repeating the deposition and lithography cycle. Complex structures can be produced on the scale of tens or hundreds of atoms (approximately tens of nanometers) with controlled 3D morphology. This reduction in size scale compared to lithographic patterning is a huge benefit. Materials can now be designed and created with subwavelength structure to coherently scatter light. Because the morphological structure ($<$100 nm) is much smaller than the wavelength of visible light ($\sim$500 nm), light traveling in a sculptured thin film sees a medium with averaged properties. The material can be designed to control refractive index through a wide range, and birefringence and circular polarization effects are controllable.

The morphological structure of thin films is a vast topic. Thin films are produced with a nonequilibrium growth process, somewhat like the growth of clouds, and the mechanics of the growth determine the structure, which is often very different from that seen in a bulk sample of the same substance. Crystal and morphological structure vary with deposition process conditions, and vary during deposition to record a history of growth much like the rings of a tree, or stratified rock. If chemical composition is varied during the deposition, that change will be reflected in layers in the film. If vapor energy or temperature is varied, variations in density and crystal structure will result. Thin film growth is an aggregation process (the film accumulates over time), and exhibits many of the features of similar processes, sometimes producing materials with structures similar to broccoli or clouds. This self-affine (self-similar) structure is observed over a large range of sizes under some film deposition conditions. Under other deposition conditions, crystallographic effects dominate growth and films become layers of faceted crystals. In sculptured thin films, these growth characteristics are controlled and exploited in a functional manner to

design material properties. For example, magnetic video and audio tapes that have the symbol ME, abbreviating 'metal evaporated', employ a structured magnetic thin film that has superior magnetic response for recording. Similar effects are observed in optical materials, and sculptured thin films appear promising for improving existing optical devices or enabling new ones to be developed.

The most studied form of sculptured thin films is that produced by an asymmetric deposition process. In deposition by vacuum evaporation, a thin film is fabricated by evaporating some source material in vacuum, and condensing a vapor onto a substrate. Eyeglasses, mirrors, television phosphor screens, etc. are coated this way. In most thin films the vapor arrives straight down (perpendicular) to the plane of a flat substrate. The film structure reflects the symmetry of the vapor, and is typically composed of a dense layer with some defects or voids oriented perpendicular to the substrate surface. A thin film exhibiting this morphology is shown in the scanning electron microscope image of Figure 2. When vapor arrives asymmetrically onto a substrate, however, that asymmetry is reflected in the growth, producing structural asymmetry in the resulting film, and altered physical properties. A second scanning electron microscope image is shown in Figure 3, of a sculptured thin film where the orientation of the substrate relative to the vapor is changed twice, producing three layers with different morphologies. The thin uniform gray at the absolute bottom of the image is the edge of the silicon crystal substrate which has been cleaved to reveal an interior section of the film. The bottom and first film layer is inclined to the right, toward the vapor source for the duration of this first part of the deposition. The middle and second film layer has a twisted structure created by slowly rotating the substrate relative to the vapor, while keeping the tilt fixed. The top layer has a vertical structure created by rapidly rotating the substrate. The entire film was produced with the vapor arriving at an angle of 45°, measured from either the plane of the substrate or its perpendicular. Varying the tilt angle has a profound effect on the morphology of thin films. The void structure in these films records the geometry of the deposition. When we cleave a film we are able to observe a 2D slice of this rich morphology. At large tilt angles, glancing angle deposition (GLAD) produces films with highly porous and complex structure. At these angles, the film structure becomes isolated columns of material where growth variations are now reflected in variations in the shape of individual columns. At less oblique vapor angles, growth variations are reflected in the shape of networks of atomic-scale voids within a largely dense film. Indeed, for the best optical devices fabricated to date (with the serial bideposition technique described later) it is difficult to distinguish any variation in structure even using a high-resolution scanning electron microscope. The film deposition process can be designed to control the morphology of the growing film, and to vary the structure as a function of time and thickness. The morphology of these films determines their optical properties; optical birefringence, second harmonic generation, filter bandwidth and attenuation, etc. can all be controlled in sculptured thin films.

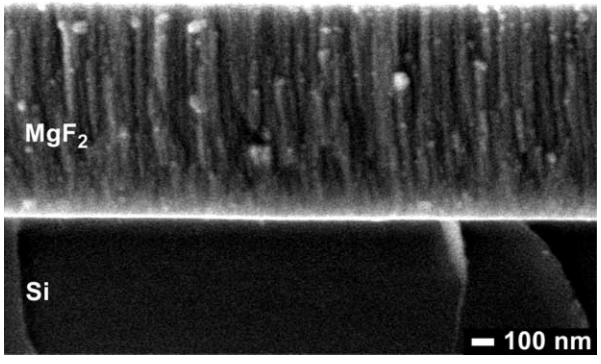


Figure 2 Micrograph of MgF_2 film deposited onto a 75 mm silicon wafer substrate, then fractured to reveal the edge shown.

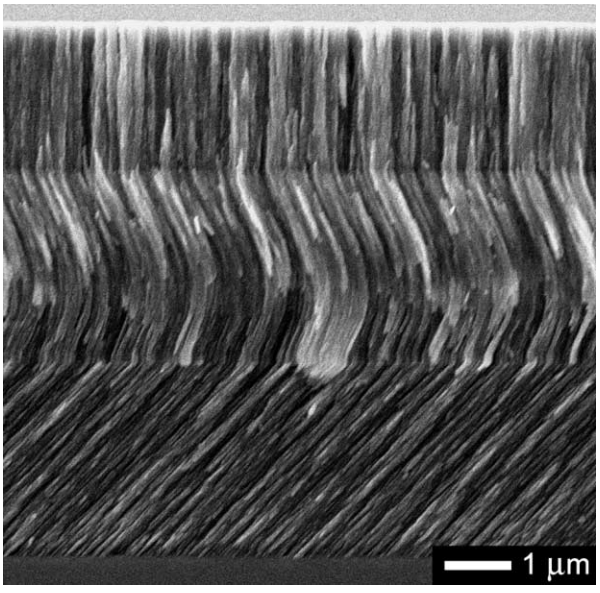


Figure 3 Micrograph of SiO film deposited with a tilt angle of 45° with one layer of no rotation, one layer with one rotation, then one layer with rapid rotation.

Optical Thin Films

Light can be described as a traveling wave with electric and magnetic fields oscillating perpendicular to each other, and to the direction of propagation. If the orientation of the electric field remains constant as the wave propagates, the light is said to be linearly polarized. If the orientation of the electric field rotates through a circular path, the light is circularly polarized, and can be left or right handed. In the most general description, the electric field traces an ellipse during propagation and the light is said to be elliptically polarized. The experimental study of light polarization is called ellipsometry. When light propagates through vacuum or air, the state of polarization remains constant. When light propagates through some types of materials, however, crystallographic or morphological asymmetry (or more correctly, anisotropy) results in different propagation conditions for different polarizations. Upon passage through these materials, the orientation and form of the state of polarization is changed, for example from left circularly polarized (LCP) to right circularly polarized (RCP). When multiple layers of anisotropic materials are combined into an optical circuit, as in the sculptured thin film concept, highly controlled optical response characteristics can be created.

The discussion of the optical properties of thin films begins with definitions of isotropic, uniaxial, and biaxial materials. Isotropic materials have optical properties that are independent of the direction that the light propagates through them. Vacuum and air are isotropic, as is glass. Rotation of the optical material relative to the light propagation direction has no effect on how the light propagates, and the state of polarization remains constant. Uniaxial materials have properties that vary with direction, but with one preferred direction (or axis) where response is independent of rotation. Biaxial materials have properties that are even more dependent on propagation direction, and require two axes to describe their optical properties. For biaxial materials there is no propagation direction where the optical response is independent of rotation. The mineral calcite is an example of a uniaxial material where the refractive index, and therefore the speed of light propagation, varies with direction relative to the crystal structure. This is a direct consequence of the trigonal crystal structure of calcite, and is an example of linear birefringence. Birefringence is the effect caused by anisotropic materials whereby light of different polarizations, even within one ray of light, travel with different speeds. When light propagates in a birefringent material, the orientation of the light's electric field relative to the material's optical axis determines the speed of propagation. Light with different polarization (orientation) propagates at different speeds, and a beam's polarization state is altered on passage through the material. Thin films can be isotropic, uniaxial, or biaxial, depending on the conditions under which the film is deposited. Sculptured thin films are films fabricated in a manner that controllably enhances the anisotropic optical response to produce materials with designed optical properties based on control of their morphological structure. As in calcite where the anisotropic crystal structure yields a uniaxial optical response, anisotropic morphological structure in sculptured thin films yields biaxial optical response. A plan view scanning electron microscope image of an anisotropic thin film is shown in Figure 4. The film was deposited with two vapor sources with the substrate placed between and the vapor arriving from both sides at an angle of 70° from the substrate perpendicular. This film is similar to those produced with the serial bideposition technique, and is optically birefringent. Polarization filters and other optical components can be constructed by designing the morphological structure of the thin film.

As discussed above, when a thin film is deposited from a vapor arriving perpendicular to a flat substrate surface, a structure composed of small (tens of nanometers diameter) columns of film material grows, with the columns oriented perpendicular to the substrate (Figure 2). This film is optically uniaxial, with the preferred axis perpendicular to the substrate, and parallel to the vapor flux. Light propagating along this optical axis (perpendicular to the substrate) travels at one speed regardless of polarization, and independent of rotations of the substrate around this axis. Light propagating along other paths will travel at a speed that depends on the polarization of the light

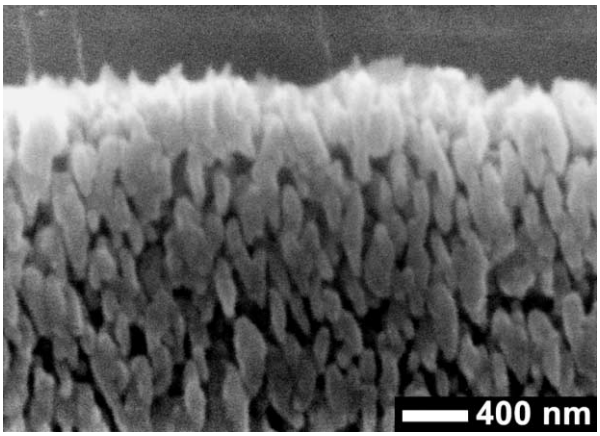


Figure 4 Micrograph of MgF_2 film, viewed from above, deposited with simultaneous growth from two vapor sources.

(orientation of the electric field relative to the propagation direction). The effect is small, and arises from the morphological structure of the film. More pronounced optical response is produced when the vapor arrives onto the substrate at an angle from the substrate perpendicular, producing optically biaxial materials, and enabling exciting polarization and filtering effects. If a thin film of magnesium fluoride is deposited onto a polished piece of glass with the vapor arriving at 45° (relative to an axis perpendicular to the substrate), a slanted columnar film is produced. The morphology of the film is tightly packed columns of polycrystalline magnesium fluoride with diameters between 50 and 200 nanometers (nm), inclined toward the vapor source, and penetrated by a network of atomic-scale voids. Because of the anisotropic nature of the deposition, the voids are anisotropic, and the film is optically biaxial. If the substrate is then held at that tilt angle and is rotated during deposition, the anisotropy in the structure rotates with the deposition, and an important type of sculptured thin film is produced. The electromagnetic description of this type of film is a 'helicoidal bianisotropic medium', describing both the helix shape of the rotation of the structure and the local bianisotropic (biaxial) physical, and optical, structure. Films with helicoidal symmetry are ideally suited for studies involving circular polarizations of light, and many demonstrations of sculptured thin films have involved these films and circularly polarized light. Like many organic molecules and some mineral crystals, these film are chiral, existing in two mirror image forms, called enantiomers. A right-handed structure rotates clockwise traveling away from an observer; a left-handed structure rotates counterclockwise. Also like some organic molecules and mineral crystals, helicoidal films are optically active – they affect the polarization of light. The observed effect depends on the wavelength of light relative to the physical helicoidal structure. If these two lengths match, a strong resonance effect is observed where the light wave coherently scatters from the anisotropic structure. Filters to produce and convert circular polarized light have been constructed with sculptured thin films of this type.

The range of structural control over the morphology of thin optical films limits the ability to tailor optical response for desired applications. While optical systems must always be treated as a whole to determine response properties, the description of some basic elements is useful. A film deposited at normal incidence will be uniaxial with the axis perpendicular to the substrate. For light propagation along this axis (common for optical systems), the material exhibits no birefringence. This structure is often called an isotropic element even though the columnar structure is slightly optically anisotropic. A truly isotropic layer could be produced by ion bombardment or elevated temperature during deposition, either of which would have the effect of eliminating the columnar structure and reducing anisotropy in the film. The second elemental structure is a slanted columnar film that is commonly described as a matchstick morphology. It is produced by changing the vapor incidence away from the substrate normal, leading to an inclination of the column structure toward the vapor source, as described above. Further structures can include: helicoidal structure where the anisotropy rotates through the thickness of the film, and periodically bent nematic structure where the anisotropy tilts in an s-shaped pattern through the film (fabricated with a complex pattern of rotations and depositions). These four elementary microstructures are shown in **Figure 5**. Electrical, optical, and magnetic anisotropy can also be increased by depositing obliquely from two sides instead of one, and this technique (called serial bideposition) can be used to increase the anisotropy in sculptured thin films. At extremely oblique deposition angles, a regime called glancing angle

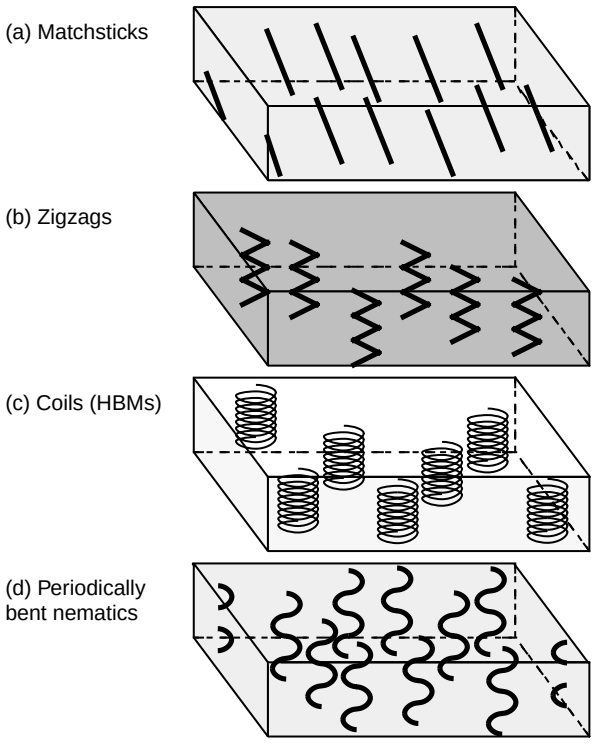


Figure 5 Elementary microstructures of sculptured thin films. From Lakhtakia A, Messier R, Brett MJ and Robbie K (1996) Sculptured thin films (STFs) for optical, chemical and biological applications. *Innovations in Materials Research* 1(2): 165–176, reproduced with permission.

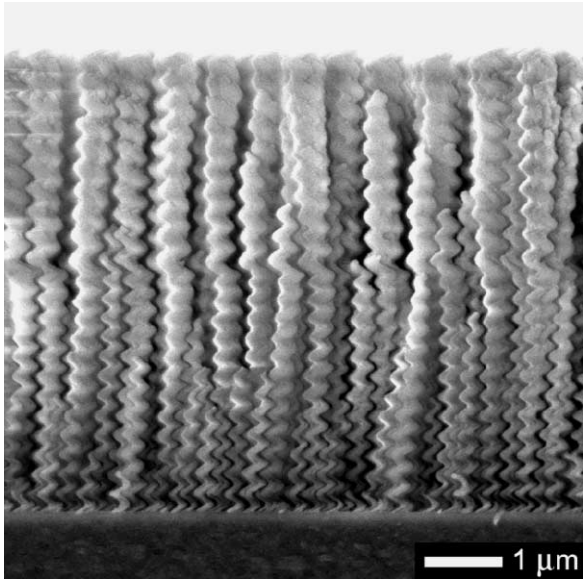


Figure 6 Micrograph of MgF_2 film deposited with a tilt angle of 85° and substrate rotation clockwise for the first half and counterclockwise for the second.

deposition, a rapid increase in porosity with approximately constant column tilt results from a competition between deposition processes. Films deposited here are highly porous and strongly scatter light, limiting their usefulness for conventional optical devices, though some application has been seen in these highly porous films when impregnated with liquid crystals, and switched similarly to liquid crystal displays. An example of a highly porous optical filter with circular polarization effects is shown in Figure 6, where two layers of helicoidal porous films are deposited with different handedness and pitch. Sculptured thin films for most optical devices are deposited at less oblique angles to reduce porosity and therefore diffuse light scatter (haze) as well as sensitivity to humidity and other environmental changes.

Combining the concepts of film growth and optical response to morphology leads to the hoped-for vision of sculptured thin films: optical circuits in a chip. Combining the elementary microstructures described above, and employing powerful computer growth modeling and optical design programs, researchers are currently defining and producing sculptured thin films with increasingly complex optical response characteristics. There are substantial steps required to move from the fairly simple optical filters fabricated today to large-scale devices with complex functionality, but continuing success and increasing interest suggest that sculptured thin films could become a dominant optical technology in years to come.

Current Research

While thin films deposited at off-perpendicular incidence angles have been studied for many years, the last decade has produced significant new insights into these materials, leading to the concept of sculptured thin films. Understanding that morphologically rich materials can be used to produce new optical effects has yielded some impressive demonstrations. Research is also underway to explore these materials for chemical, magnetic, biological, and other applications.

Significant research has been conducted in the past few years to develop and extend the concepts of sculptured thin films. Optical studies, both theoretical and experimental, have revealed a range of remarkable properties including circular birefringence and Bragg reflection, liquid crystal alignment and switching, optical interconnection, and even suggestions of photonic bandgaps. Studies of optical sensing of gases and other materials is progressing, and simulation and modeling of the growth processes is ongoing. The following are two examples of optical materials that have been fabricated with sculptured thin-film techniques; one is a polarization discriminatory light-handedness inverter and the second is a liquid crystal switching cell that can be used to electrically modulate the transmission of polarized light. The handedness inverter converts left circularly polarized light to right circularly polarized within a spectral band, and blocks right circular polarized light. It is constructed by combining a layer with vertical columnar, or matchstick, morphology together with a layer with helicoidal morphology. In keeping with the sculptured thin film concept, the films are deposited sequentially during the same deposition, and optical properties are monitored during growth. The transmission spectrum for the devices is shown in Figure 7, where T_{RL} and T_{LR} are the light transmittances of right circularly polarized light when the incident is LCP, and of LCP when the incident is RCP, respectively. LCP light is almost entirely converted to RCP upon transmission, and RCP light is blocked. The bandwidth of the devices is about 50 nm, determined by matching of the circularly polarized light with the layer with helicoidal morphology.

When films are deposited with helicoidal morphology at glancing angles, i.e., glancing angle deposition, highly porous structures are produced where the columns stand alone and are individually shaped by the deposition process. Optical devices can be constructed by filling these porous films with gases or liquids, and there have been demonstrations of display-type switching with liquid crystals, and discussions of chemical sensing possibilities.

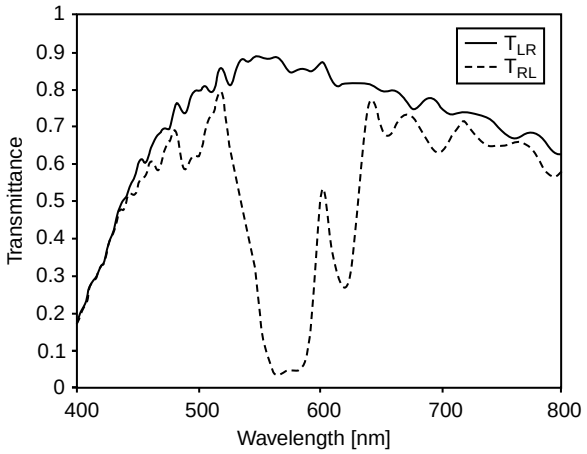


Figure 7 Transmission spectrum for visible light through a polarization-discriminatory light-handedness inverter. Adapted with permission from Hodgkinson IJ, Lakhtakia A and Wu Q-H (2000) Experimental realization of sculptured-thin-film polarization-discriminatory light-handedness inverters, *Optical Engineering*, **39**: 2831–2834.

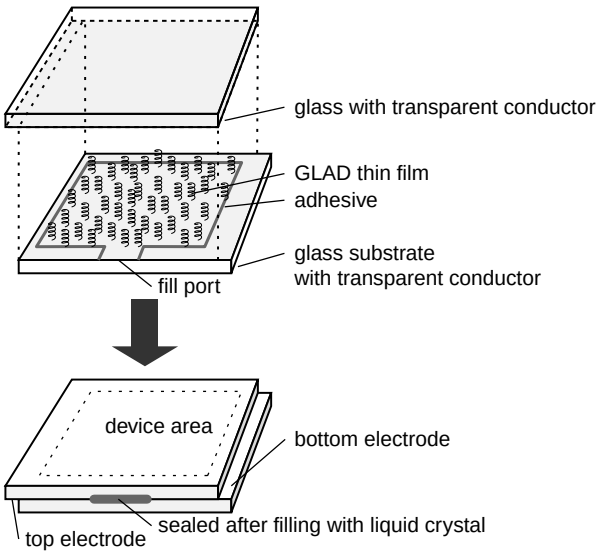


Figure 8 Liquid crystal switching cell schematic with highly porous glancing angle deposited (GLAD) film between glass plates coated with transparent electrodes. Reproduced with permission from Sit JC, Broer DJ and Brett MJ (2000) Alignment and switching of nematic liquid crystals embedded in porous chiral thin films. *Liquid Crystals* 27: 387–391, www.tandf.co.uk/journals.

The device shown in Figure 8 is a liquid crystal switching cell where the space between the electrodes, usually composed solely of liquid crystalline material, is now filled with a highly porous helical structured film. The film structure is somewhat similar to helicoidal film produced for the handedness inverter discussed above, but glancing angle deposition has produced a film of individual formed columns instead of a more dense film with a helicoidal void

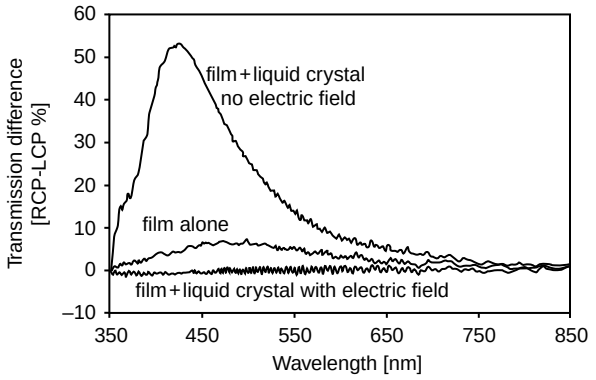


Figure 9 Transmission spectrum for visible light through a liquid crystal cell where the difference between right- and left-circularly polarized light transmission is eliminated with applied electric field. Adapted with permission from Sit JC, Broer DJ and Brett MJ (2000) Alignment and switching of nematic liquid crystals embedded in porous chiral thin films. *Liquid Crystals* 27: 387–391, www.tandf.co.uk/journals.

morphology. When the porous portions of the film are filled with simple nematic liquid crystalline material, the orientation of the rod-like molecules is strongly influenced by the helical columns. An orientational phase occurs in the liquid crystals mirroring the helical structure of the columns in the film, and producing optical properties similar to chiral nematic liquid crystals. The discrimination of left and right circularly polarized light increases relative to the unfilled film. When an electric field is applied across the liquid crystal cell, dipole moments in the liquid crystal molecules force them to align with the field, and chiral optical properties vanish (see Figure 9). Used as an electro-optic filter for circularly polarized light, this cell can switch light transmission on and off. Initial work has not shown clear benefits of this system over conventional liquid crystal display technology, but control of liquid crystal orientation throughout the cell instead of just at the glass plates appears promising. Possibly more complex optical circuits, incorporating layers of porous films filled with liquid crystals, will be useful.

Theoretical and experimental study of sculptured thin films is in its infancy, and many exciting uses for these materials are expected in coming years. Continuing study of thin film growth dynamics will improve and expand the range of structures attainable. Theoretical investigations will aid in the design and understanding of the optical response.

List of Units and Nomenclature

Aluminum (Al)	white metal used in airplanes, mirrors, and microelectronic circuits

Anisotropic	having properties that vary with orientation (rotation)
Biaxial	having anisotropic optical response characteristics that require two axes to describe
Chiral	having a shape that exists in two mirror image forms, called enantiomers
Diamond	carbon crystallographically ordered in a face-centered-cubic lattice, and used in cutting tools and jewelry
Electromagnetic waves	traveling waves of energy that compose light, x-rays, radio-waves, etc.
Epitaxy	growth of a thin film that preserves the crystal structure of the underlying film or substrate
Etching	removing material with chemical or physical atomic scale process
Graphite	carbon crystallographically ordered in hexagonal sheets, and used in pencils and lubricants
Inhomogeneity	varying with translation (position)
Isotropic	having properties that are identical in all directions
Lithography	process where patterns are produced in a sensitive layer by irradiating the layer through a mask with the desired pattern
Magnesium fluoride	ceramic (MgF_2) that is used in thin-film or single-crystal form in many optical systems
Morphology	microstructural shape, or physical arrangement of atoms in a material
Nanometers (nm)	one billionth of a meter, 10^{-9} m
Nematic liquid crystal	optical material composed of rod-shaped organic molecules that usually exhibits uniaxial optical properties
Pitch	physical period for one full rotation of a helical structure
Polarization	orientation and phase of the electric and magnetic fields in a traveling electromagnetic wave; can be linear, circular, or generally, elliptical
Refractive index	the ratio of the speed of light in vacuum to the speed of light in a material
Substrate	object that a thin film is deposited onto, often polished glass or semiconductors
Ultraviolet (UV)	electromagnetic radiation (light) with wavelength shorter than visible light and longer than X-rays
Uniaxial	having optical response characteristics that are anisotropic, but are rotationally symmetric about one axis
Vacuum	absence of matter, typically produced inside sealed glass or steel chambers by removing air with pumps
Vapor flux	moving gas-like collection of material, emitted from a source such as an evaporator or effusion cell, that will condense to form a thin film

See also

Coherence: Overview. **Coherent Lightwave Systems**. **Fiber and Guided Wave Optics:** Overview. **Imaging:** Imaging Through Scattering Media. **Information processing:** All-optical Multiplexing/Demultiplexing; Coherent Analogue Optical Processors; Optical Neural Networks. **Instrumentation:** Ellipsometry. **Optical Processing Systems**. **Optical Coatings:** Anti-Counterfeiting and Decorative Coatings; Optical Black Surfaces; Thin-Film Optical Coatings. **Optical Materials:** Measurement of Optical Properties of Solids. **Polarization:** Introduction; Matrix Analysis. **Scattering:** Scattering from Surfaces and Thin Films; Scattering Theory. **Semiconductor Materials:** Band Structure Engineering. **Terahertz Technology:** Coherent Terahertz Sources.

Further Reading

Ball P (1997) *Made to Measure*. Princeton, NJ: Princeton University Press.

Barabasi A-L and Stanley HE (1982) *Fractal Concepts in Surface Growth*. Cambridge, UK: Cambridge University Press.

Collins PJ (1990) *Liquid Crystals*. Princeton, NJ: Princeton University Press.

de Gennes PG and Prost J (1993) *The Physics of Liquid Crystals*. Oxford, UK: Clarendon Press.

Hodgkinson IJ and Wu Q-H (1997) *Birefringent Thin Films and Polarizing Elements*. Singapore: World Scientific Publishing.

Hodgkinson IJ, Lakhtakia A and Wu Q-H (2000) Experimental realization of sculptured-thin-film polarization-discriminatory light-handedness inverters. *Optical Engineering* 39: 2831–2834.

Hodgkinson IJ, Wu Q-H, Thorn KE, *et al.* (2000) Spacerless circular-polarization spectral-hole fitters using chiral sculptured thin films: theory and experiment. *Optics Communications* 184: 57–66.

Lakhtakia A, Messier R, Brett MJ and Robbie K (1996) Sculptured thin films (STFs) for optical, chemical and biological applications. *Innovations in Materials Research* 2: 165–176.

Liu F, Umlor MT, Shen L, *et al.* (1999) The growth of nanoscale structured iron films by glancing angle deposition. *Journal of Applied Physics* 85: 5486–5488.

Messier R, Venugopal VC and Sunal PD (2000) Origin and evolution of sculptured thin films. *Journal of Vacuum Science and Technology* 18: 1538–1545.

Ohring M (1992) *The Materials Science of Thin Films*. New York: Academic Press.

Robbie K, Brett MJ and Lakhtakia A (1996) Chiral sculptured thin films. *Nature* 384: 616.

Singh ON and Lakhtakia A (eds) (2000) *Electromagnetic Fields in Unconventional Materials and Structures*. New York: John Wiley.

Sit JC, Broer DJ and Brett MJ (2000) Alignment and switching of nematic liquid crystals embedded in porous chiral thin films. *Liquid Crystals* 27: 387–391.

Toader O and John S (2001) Proposed square spiral microfabrication architecture for large three-dimensional photonic band gap crystals. *Science* 292: 1133–1135.

Wang L and Clancy P (2001) Kinetic Monte Carlo simulation of the growth of polycrystalline Cu films. *Surface Science* 473: 25–38.

Smart Optical Materials

P M Martin, Pacific Northwest National Laboratory, Richland, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Smart optical materials are those materials whose optical properties (transmittance, reflectance, absorptance, emittance, fluorescence) can be controlled by external stimuli; either by an applied electric field or voltage, mechanical stress, incident light or electromagnetic field intensity, temperature variation, or time duration. This includes electrochromic, photochromic, thermochromic, nonlinear optical, light emitting, fluorescing, emissive, and piezochromic materials. These materials can be in various states, such as solids, gels, composites, and mesophases. Electrochromic and photochromic materials, however, are used most extensively.

The mechanism for adaptive control of light must be reversible, have a well-defined response time, and have an activation threshold, otherwise, the material is considered passive. The primary mechanisms that can cause variations in optical properties are crystal phase transitions, nonlinear effects in the polarizability, optical energy band transitions and band bending, compositional changes, defect and color center formation, and lattice strain. Crystal phase transitions change the bonding morphology of a material, which affects the energy bandgap. Most optical materials have several possible crystalline phases, depending on temperature. The optical properties of the anatase and rutile phases of TiO_2, for example, are very different, and result from different energy bandstructures (recall that the energy bandstructure depends on lattice spacing and crystalline plane orientation). The crystalline phase of several thermochromic materials changes from the martensite structure to the austenite structure with change in temperature, and an associated change in optical properties. Compositional changes are caused by chemical or physical reactions in the material, such as a metal ion bonding to an atom in the lattice, or loss of an atom in the lattice. The reaction can be caused by an energetic photon or ion driven by an electric field; electrochromic materials fit into this category. If their energy is high enough, incident photons can create defect states in the bandgap of materials, which in turn form color centers that absorb light. Modulated electric and electromagnetic fields can cause changes in the linear and nonlinear polarizability of the atoms in a lattice. Nonlinear optical and piezochromic effects are based on this mechanism.

Photochromic Glass

Photochromic glass is one of the most widely used smart optical materials, and definitely the most widely marketed. It has been around for over 30 years with hundreds of associated patents. The transmission of photochromic materials adjusts to the light level of their environment, but these materials give no output information for data processing or communications. They do, however, require no other external control mechanism other than the intensity of light (characteristics of a true smart material). Photochromic materials darken reversibly when exposed to sunlight or a specific

optical (UV–LWIR) wavelength band. Absorption in specific narrow or broad wavelength bands is due to the formation of color centers, or point defect states in the lattice. The kinetics of darkening has been described by a two-color center model with two rate constants. Energy (usually in the ultraviolet wavelength range) of the incident photons, ionizes atoms in the glass, thus forming 'F' centers. F centers are highly absorptive and the absorption generally occurs in a broad wavelength band. The earliest photochromic glass developed utilized a dispersed and sensitized colloidal silver halide as the darkening agent. Copper oxide was a typical sensitizer and AgCl is also a fairly common impurity used in PC glass. This material darkens by the equation

$$Ag^{I} + Cu^{I} \rightleftharpoons Ag^{0} + Cu^{II} \qquad [1]$$

This type of reaction is fairly typical of these materials. The defect state is created in the silver and copper, creating optical absorption. Photochromic glass can darken down to 22% transmittance in sunlight and become nearly clear indoors, transmitting 85% of visible light. Figure 1 shows the transmittance of the darkened and normal states. Temperature, however, does have an effect on the lenses. The glass will not darken completely at temperatures above 32 °C, and at cold temperatures the lens will get extremely dark in a shorter period of time. Recently, $BaO–R_2O_3–SiO_2$ (R = Al, B, La) nanoparticles have been introduced to enhance photochromicity.

Thermochromic Materials

Thermochromic materials darken and lose transmittance when heated above a critical temperature. They can be as sophisticated as a multilayer thin film optical interference coating or as mundane as a cereal bowl that changes color when heated.

Thermochromic products are used to darken windows of automobiles, for example, when the internal temperature gets above a specific value. The conductivity of these materials also decreases with darkening in many cases because additional electrons are excited into the conduction band due to the increased number of defect states in the bandgap. The most common inorganic thermochromic thin film materials are transition metal oxides, and among these, the vanadium oxide family (V_2O_5, VO_2) shows the best performance. Figure 2 shows the transmittance of a VO_2 thin film before and after heating. Here T_t is the transition temperature. When heated above T_t, the film darkens. Stoichiometric VO_2 darkens at a temperature near $T_t = 68\,°C$ due to a semiconductor to metal transition. The resistivity of VO_2 decreases by three orders of magnitude and the emittance can be increased up to 90% at the transition temperature. This material has promising applications for thermal control of space structures, where surfaces are exposed to intense solar radiation and periods of darkness at very low temperatures. The transition temperature can be engineered (increased or decreased) by doping with metals such as W and Ta.

Loss of transmission can be achieved by absorption or increased reflectance. Reflectance increases with heating are observed in substoichiometric WO_3 films and absorptance increases in stoichiometric films. The reflectance increase is attributed to a decrease in W/O ratio and an associated increase in the refractive index (n) and extinction coefficient (k).

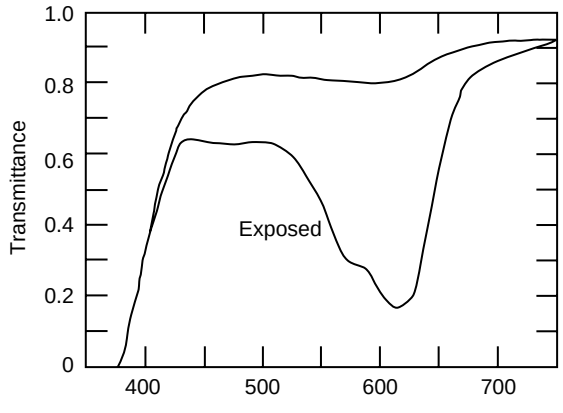


Figure 1 Transmittance of photochromic glass in darkened and normal states. (Reproduced with permission from Jorgenson GV and Lee JC (1985) Optical Materials for Energy Efficiency and Solar Energy Conversion IV, *SPIE Proceedings*, vol. 562. Bellingham, CA: SPIE.)

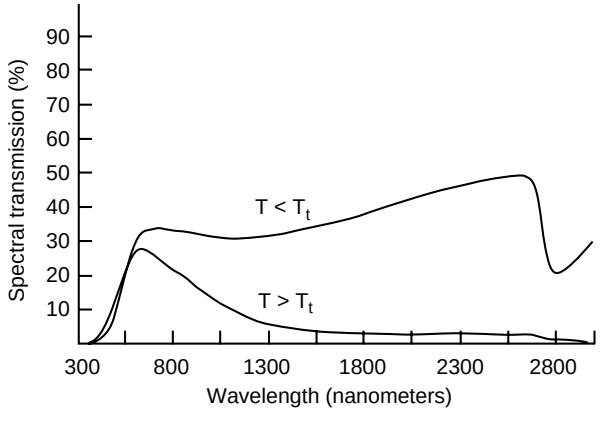


Figure 2 Change in transmittance of a thermochromic VO_2 film with heating. $T < T_t$ for normal transmission and $T > T_t$ when heated. (Reproduced with permission from Jorgenson GV and Lee JC (1985) Optical Materials for Energy Efficiency and Solar Energy Conversion IV, *SPIE Proceedings*, vol. 562. Bellingham, CA: SPIE.)

Electrochromic Materials

Electrochromic materials, also called switchable materials, are usually used in the form of thin film devices. Electrochromic windows, or smart windows, have finally reached maturity and promise to be the next major advance in energy-efficient window technology. These devices are used in architectural glazings, automotive mirrors, rear view mirrors, sunroofs, sunglasses, and other high-end applications. Electrochromic devices are well suited for both solar and thermal control. A window can be controllably darkened during periods of bright sunlight and switched back to transparency during the evening or during cloudy periods. Windows with multilayer thin film electrochromic coatings change color and darken with an applied voltage, and bleach with a reverse bias. The magnitude of the applied voltage determines the degree of coloring and bleaching. In addition to a change in visible transmission, the NIR transmission and reflectance, and the associated thermal properties, can also be controlled. The major advantages of electrochromic devices are: (1) power is used only during switching; (2) switching voltages are small (1–5 V); (3) reflectance and transmittance is always specular; (4) gray scale variations are possible; (5) polarization is minimal (reduces birefringence and distortion); and (6) the memory is adjustable in many cases.

The electrochromic effect occurs in inorganic compounds by dual injection (cathodic) or ejection (anodic) of ions (M) and electrons (e^-). A typical reaction for a cathodic coloring material is

$$WO_3(\text{colorless}) + yM^+ + ye^- \rightleftharpoons MyWO_3(\text{blue}) \quad [2]$$

where M is H^+, Li^+, Na^+, Ag^+, etc.

A typical anodic reaction is

$$Ir(OH)_x(\text{colorless}) \rightleftharpoons IrO_x(\text{black}) + XH^+ + xe^- \quad [3]$$

Movement to the right creates color or darkening, and movement to the left results in bleaching. Figure 3 shows the spectral data for the switching of a tungsten oxide film over the solar spectrum; the spectral response of the coating in the bleached and colored (or 'switched') states, and Figure 4 shows the bleached and darkened states of an actual electrochromic window.

Figure 5 shows the design of a simple electrochromic coating. The basic electrochromic coating consists of a transparent top electrode (usually indium tin oxide – ITO), which also injects electrons into the electrochromic layer beneath. By far the most studied and most promising electrochromic material is tungsten oxide (WO_3), which can have either an amorphous or crystalline microstructure. WO_3 has a high coloration efficiency and capacity. Other inorganic materials that are presently being developed for solar control applications are various forms of WO_3, NiO, $WMoO_3$, MoO_3, and IrO_x. Many of these are doped with a conduction ion such as Li^+ or H^+. The ion conductor layer is used to transport protons (H^+), which are supplied by the ion storage or counter-electrode layer, into the electrochromic material. The most promising ion conductors are certain immobile solvent polymer systems, ionic glasses, and open-channel metal oxide structures such as perovskites. Oxides of Ta and Nb also show promise. The bottom layer is another transparent conducting coating. It is possible to inject both electrons and protons into the electrochromic material, thereby inducing a strong absorption band in a given region, e.g., the visible spectrum, that results in a color change. In an

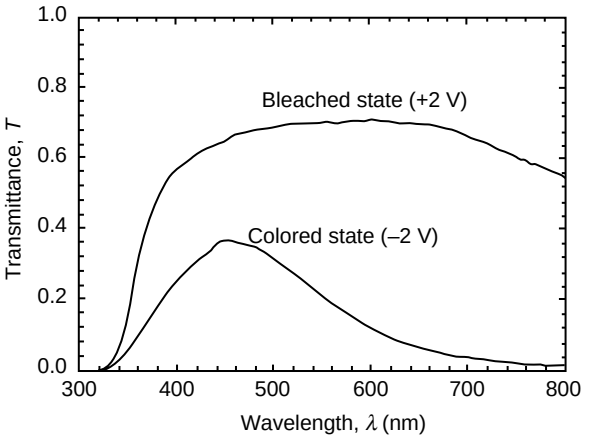


Figure 3 Spectral switching of an electrochromic window for darkened (−2V bias) and bleached (2V bias) states. (Reproduced from Kre J, Topic M, Smole F, *et al.* (2002) *Solar Energy Materials and Solar Cells* 71: 387–395, with permission from Elsevier.)

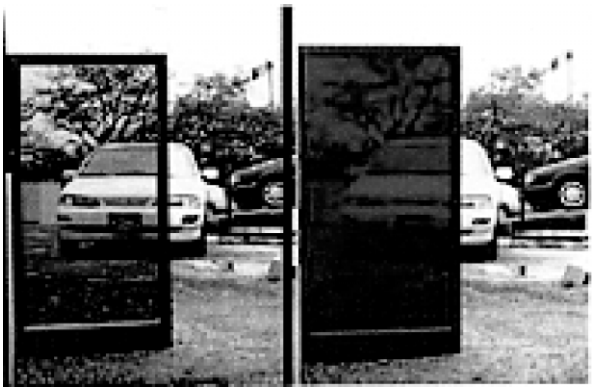


Figure 4 Bleached (left) and darkened (right) states of an electrochromic window.

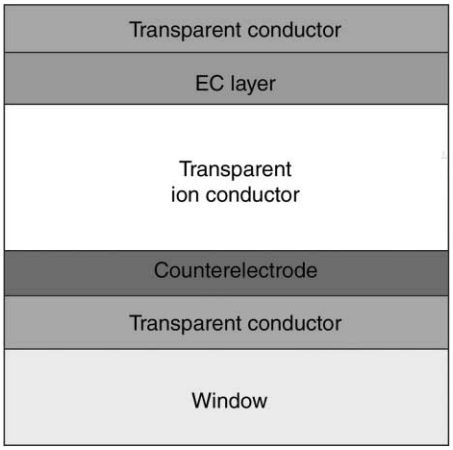


Figure 5 Basic layer design of electrochromic coating for use on a window.

Table 1 Electrochromic device structures

Electrochromic layer	*Ion conductor or electrolyte*	*Ion storage*
a-Li_xWO_3	$LiClO_4$ + PC	Redox Cple/NiO
a-WO_3	Ta_2O_5	$Ir_xSn_yO_2$:F
a-Li_xWO_3	$LiClO_4$ + PC	Prussian blue
a-H_xWO_3	Polymer	Polyaniline
a-H_xWO_3	SiO_2/metal	WO_3
Viologen	PMMA + organic	none
a-WO_3	Ta_2O_5	a-IrO_2
c-WO_3	Li–B–SiO Glass	IC/$Li_xV_2O_5$
a-WO_3	Li–PEO	CeO_x
a-Li_xWO_3	PPG–$LiClO_4$–MMA	$Li_yV_2O_5$
a-H_xWO_3	Ta_2O_5	NiO
a-WO_3	Li–PEO	NiO
NiO	a-PEO copolymer	Nb_2O_5, WO_3:Mo
a-WO_3	a-PEO copolymer	Polymer
Polyaniline	HCl	–

a, amorphous; c, crystalline; IC, ion conductor; MMA, methyl methacrylate; PC, polycarbonate; PEO, poly (ethylene oxide); PMMA, poly (methyl methacrylate); PPG, poly (propylene glycol).

electrochromic device, an externally applied electric field is used to control the injection process (voltage is applied across the two transparent electrodes). The coloration remains for some time, even after the external field is removed. The system returns to the initial state upon reversing the polarity of the external field. It is interesting to note that electrochromic windows being developed by NREL now employ solar cells to facilitate optical switching.

There are two major categories of electrochromic materials: transition metal oxides (including intercalated compounds) and organic compounds (including polymers). Table 1 shows a few of the most common electrochromic materials. Organic electrochromic materials are based on the viologens, anthraquinones, diphthalocyanines, and tetrathiafulvalenues. With organic compounds, coloration is achieved by an oxidation–reduction reaction, which may be coupled to a chemical reaction. The viologens are the most studied of the organic electrochromics. Originally, organic electrochromics tended to suffer from problems with secondary reactions during switching, but recently more stable organic systems have been developed. However, their reliability in harsh environments is questionable, and protective encapsulation may be required.

Electrochromic nickel oxide (NiO) is also being developed. NiO has been combined with manganese oxide, cobalt oxide, and niobium oxide in all alkaline devices.

Nonlinear Optical Materials

Nonlinear optical materials have evolved slowly over the past 20 years. At relatively low light intensities, the properties of most optical materials are independent of intensity of illumination. There is no interaction between the light waves in the medium. However, if the intensity of illumination gets high enough, nonlinear optical effects can occur and the optical properties become intensity dependent. When an electromagnetic field oscillating at optical frequencies ($\sim 10^{13} - \sim 10^{17}$ Hz) is applied to the lattice of a dielectric material, the electrical charges in the lattice become polarized and form electric dipoles. The positive ion cores do not move significantly, but the electrons are displaced in a harmonic motion. The polarization $\mathbf{P}$ is defined as $-Ne\mathbf{X}$, where N is the number of electric dipoles, $\mathbf{X}$ is the displacement of the electrons, and e is the charge on an electron. Here $\mathbf{X}$ is a complex number. Now, $\mathbf{P}$ is related to the electric field E by $\mathbf{P} = \varepsilon_0 \chi E$, where χ is the susceptibility and ε_0 is the permitivity of free space.

The linear dielectric constant is $\mathrm{Re}(1 + \mathbf{X})^{1/2}$, which is just the index of refraction. The imaginary part causes optical absorption.

When the intensity of illumination is high, the elastic limit of the electron's displacement is exceeded and the harmonic restoring force becomes nonlinear in $\mathbf{X}$, which adds quadratic and higher order terms to the expression for $\mathbf{X}$ and $\mathbf{P}$. In addition to the fundamental resonance, the dipoles now also oscillate at twice and thrice their fundamental resonant frequency (second and third harmonics). The polarization then is also nonlinear in displacement and can be expressed by the well-known relation

$$\mathrm{P} = \varepsilon_0(\chi^{(1)}\mathrm{E} + \chi^{(2)}\mathrm{E}^2 + \chi^{(3)}\mathrm{E}^3 + \ldots) \qquad [4]$$

where $\chi^{(2)}$ and $\chi^{(3)}$... are called the nonlinear susceptibilities. The second-order polarization $P^{(2)}$ causes optical mixing phenomena, involving sums and difference frequencies. When mixed with a DC field, the refractive index of the $\chi^{(2)}$ material can be controllably varied by varying the DC field. This is the well-known Pockels effect and is used in optical modulators. Figure 6 shows one method that is used to modulate the refractive index by the photorefractive effect. Here two laser beams (A_1 and A_2) are mixed in the $\chi^{(2)}$ material with the resulting intensity $I = I_o(1 + m\cos(kx))$. The material is polarized in phase with the mixed beams $\rho = \rho_o(1 + m\cos(kx))$, which results in a modulated electric field E_{sc}. The refractive index (Δn in the figure) is modulated in phase with the mixing signal. The $\chi^{(2)}$ materials are also used for frequency doubling.

Barium titanate ($BaTiO_3$), lithium niobate ($LiNbO_3$), KH_2PO_3 (KDP), lithium tantalite ($LiTaO_3$) and potassium niobate ($KNbO_3$) are $\chi^{(2)}$ optical switching and frequency doubling materials that have been studied for decades. Note that because they can be easily polarized, these are also excellent piezoelectric/ferroelectric smart materials.

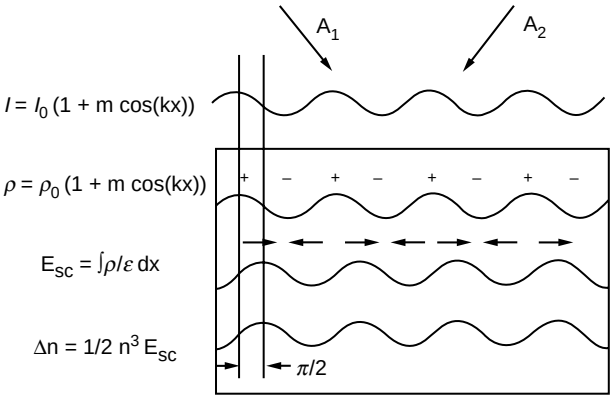


Figure 6 Modulation of the refractive index of a nonlinear optical material with incident laser modulation. From top to bottom: modulated laser intensity I, density ρ, electric field E_{sc}, refractive index Δn, and phase angle $\pi/2$. (Reproduced with permission from Bass M (ed.) (1995) *Handbook of Optics II*. New York: McGraw-Hill.)

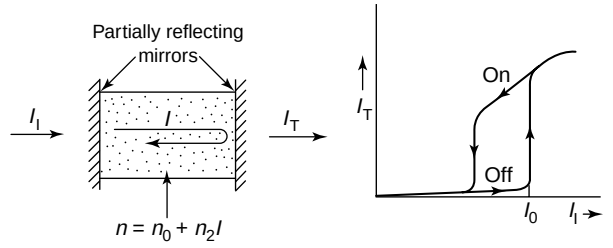


Figure 7 Optical switching behavior of a Fabry–Perot filter. I_i and I_T are the incident and transmitted intensities respectively. (Reproduced with permission from Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.)

The $\chi^{(3)}$ effects are sought after for applications such as laser eye protection, optical sensor protection, optical switching, optical limiting, and signal processing. In this case the refractive index (real and imaginary) depends on the intensity of the incident radiation. Figure 7 shows how this nonlinear effect is used in an optical switching device. It is all based on the change in the optical path length of a Fabry–Perot filter. At normal intensities, the refractive index of the spacer layer is n_0 and the filter is in the 'off' state. The refractive index becomes $n_0 + n_2 I$ and the optical thickness (nt) of the spacer layer changes at high illumination intensities, which in turn shifts the resonant wavelength (frequency) of the Fabry–Perot filter and creates the optical switch ('on' state).

There are a large number of $\chi^{(3)}$ and $\chi^{(2)}$ organic materials, including dyes, dimethylamino nitrostilbene, methyl nitroanaline, poly-BCMU, polydiacetylenes, and urea. Because they are essentially chains, many organic molecules can be easily polarized and possess higher order susceptibilities. The polarizability of organic materials is often enhanced by the mobility of delocalized π electrons in C–C bonds in aromatic rings. Hybrid nonlinear materials have even been formed by combining organic and inorganic components such as metallophthalocyanines which display very strong excited state absorptions. Most of these materials are deposited by spin casting.

Emissive Materials

Two of the newest applications of smart optical materials are emissive displays and electronic ink. This encompasses light-emitting polymers and small molecules, light-emitting diodes, and quantum cascade lasers. Organic light-emitting device (OLED) technology is leading the revolution in the flat panel display industry. This device was first invented by Tang and Van Slike and first reported in *Applied Physics Letters* in 1987. Figure 8a shows the layer structure of the OLED and Figure 8b shows a blue OLED pixel. The device is deposited onto an indium tin oxide (ITO)-coated glass or plastic substrate, with the ITO acting as the anode. The active layers can be either light-emitting polymers or small molecules. About 500 Å of a hole transport material (hole transport layer, HTL), here a tertiary amine, is deposited over the ITO. An aluminum trisquinalate (Alq), which transports electrons (electron transport layer, ETL) is deposited next. The device is topped off with

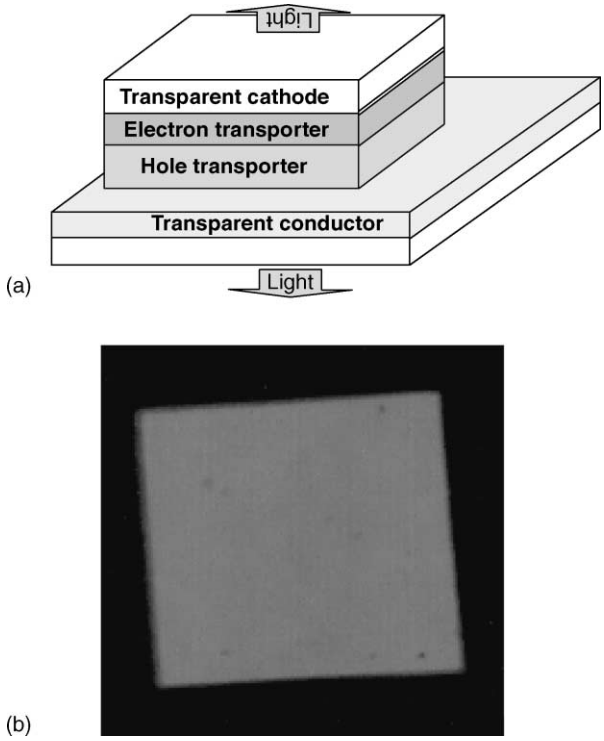


Figure 8 (a) Layer diagram of an OLED. (b) Blue OLED pixel.

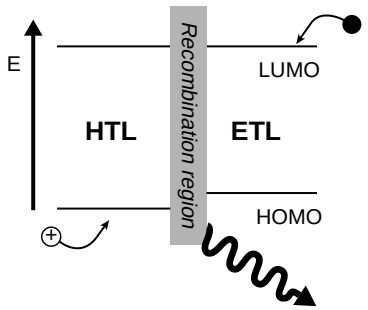


Figure 9 Simple energy level diagram of an OLED showing the relationship between the ETL and LUMO, and HTL and HOMO.

a low work function cathode, usually Ag:Mg or Ca. The entire device thickness is about 1000 Å. Even a dust particle can ruin the operation of an OLED by creating inactive regions or shorting out the electrodes.

The OLED emits light in a blue, green, or red wavelength band when activated by an electric current, i.e., electron and hole injection. The device functions by injecting holes from the ITO into the HTL. As shown in the simplified energy band diagram (Figure 9) for an OLED, the hole moves into the HOMO (highest occupied molecular orbital), which loosely corresponds to the valence band of a semiconductor. Electrons are injected from the cathode into the ETL, where it is transported into the LUMO (lowest unoccupied molecular orbital), which loosely corresponds to the conduction band of a semiconductor. When a voltage is applied between the anode and cathode, an electron flows and the holes and electrons combine in the ALQ layer and form excitons. Light is emitted when the excitons decay back a lower energy state. The color of the emitted light (green, red, or blue) can be changed by changing the HTL and ETL material composition. Light is emitted by fluorescence processes, from both singlet and triplet states if the material chemistry is correct, which results in 100% internal quantum efficiency. Twenty percent external quantum efficiency results from using all available exciton states. Photons can be lost through absorption in the layers, scattering, reflection between layers, and waveguide modes. These factors must be minimized to maximize the external quantum efficiency.

OLED video displays now have outputs near 100 cd/m (20 lm/W), which is about twice the normal video brightness. Red, green, and blue OLEDs now emit with very high efficiencies. For comparison, a 100 W light bulb emits at 17 lm/W and the best fluorescent light emits at 80 lm/W. Today, OLEDs are already sitting at intensities of 50–60 lm/W, and could easily push 150 lm/W in the near future. Another advantage of OLED displays is their wide viewing angle, up to 160°.

There are several types of OLED display configurations. Figure 10 shows a passive matrix display geometry and placement of the OLED layers. Here the OLED is sandwiched between crossed cathode and anode contacts. By injecting current between sets of crossed anode and cathode rows, the OLED will light up, each contact point forming a pixel. Displays can typically have as many as 3 million pixels. OLED displays seem brighter and livelier to the eye because the colors are additive, rather than subtractive as with LCD displays. For a full-color passive matrix display green, red, and blue pixels must be located very close together. Sony has developed an active matrix 13-inch diagonal display with 800×600 pixels (~150 000 functioning OLEDs).

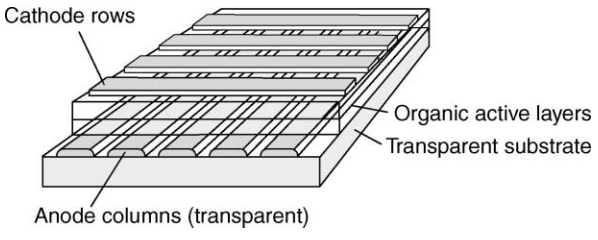


Figure 10 Geometry of a passive matrix display.

Among the more promising pixel geometries is the SOLED, the stacked OLED, where, red, green, and blue pixels are stacked on top of each other. It is mandatory that the pixels and the electrical contacts be transparent for this design, or light will not be transmitted through each layer. Very thin Mg contacts are used with a top sputtered ITO cathode. The emitted color is simply changed by applying current across the pixels, with white being obtained when all pixels are activated.

The future of OLED displays is plastics, or the flexible OLED (FOLED). Plastic substrates are lightweight, low cost, and rugged. However, they have an inherent problem of being very permeable to oxygen and water vapor, both of which severely degrade performance of the OLED by reacting with the anode layer. For displays to be practical, they must have lifetimes in excess of 10 000 hr and preferably 20 000 hr. Barrier coatings are now being developed for plastics to prevent oxygen and water ingress and extend the life of the OLED to this goal.

Electronic ink is a display medium that can be formatted using a passive matrix. It consists of millions of tiny negatively charged microcapsules, about 100 microns in diameter, floating inside a liquid medium. Each microcapsule is filled with hundreds of pigmented chips. One half of each microcapsule is white and the other half is dyed blue. When an electric charge is applied to the microcapsules, the chips are drawn either to the bottom or top of the microsphere. The viewer sees dark areas when they are drawn to the top. In another version being developed by Xerox Corp., half the chip is white and half the chip is dark. When an electric charge is placed on the chip, the dark side is oriented up and the white side down, creating dark regions where charge is placed. This can be easily erased and reformatted into text, symbols, signs, and pictures.

Future smart optical materials will simulate biological systems such as an insect's eye and will employ quantum structures to achieve optical switching, emission, and control of optical properties.

List of Units and Nomenclature

Absorptance (A)	The amount of light absorbed by an optical element divided by the incident intensity. $A = 1 - T - R$.
Active matrix display	A display geometry in which the voltage to the pixel is controlled individually by a transistor.
Candela (cd)	Formerly known as a candle. The luminous intensity equal to 1/60 of the luminous intensity per square meter of a black body radiator operating at the temperature of freezing platinum.
Exciton	An electron–hole pair, usually generated by a photon.
FOLED	Organic light emitting device on a flexible plastic substrate.
LCD	Liquid crystal diode.
Lumen (lm)	A unit of luminous flux equal to the luminous flux emitted in a solid angle of one steradian for a point source having an intensity of one candela.
NIR	The near-infrared spectral region, consisting of wavelengths between 700 and 2500 nm.
Passive matrix display	A display geometry consisting of a grid of horizontal and vertical wires. At the intersection of each grid is an active element (LCD or OLED) which constitutes a single pixel, either letting light through or blocking it.
Piezoelectric	Capable of generating a voltage when a force or stress is applied.
Polarization (coulomb/m^3)	A vector quantity equal to the electric dipole moment per unit volume.
Quantum efficiency	The number of electrons, holes, or electron–hole pairs generated per incident photon at a given optical wavelength.
Reflectance (R)	The intensity of light reflected from an optical element divided by the incident intensity. $R = 1 - T - A$.
Susceptibility (no units)	A scalar quantity equal to the ratio of the electric polarization of a material to the electric field strength.
Transmission (T)	The intensity of light transmitted through an optical element divided by the incident intensity. Note that transmission + reflection + absorption = 1 ($T + R + A = 1$).
UV	The ultraviolet spectral region consisting of wavelengths between 200 and 400 nm.
Visible	The spectral region consisting of wavelengths between 400 and 700 nm.
Wavelength (μm or nm)	The length of one oscillation of a light wave.

See also

Materials for Nonlinear Optics: Organic Nonlinear Materials. **Optical Coatings:** Thin-Film Optical Coatings. **Optical Materials:** Optical Glasses.

Further Reading

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Chemla DS and Zyss J (1987) *Nonlinear Optical Properties of Organic Molecules and Crystals*, vol. 1. Orlando, FL: Academic Press.

Chemla DS and Zyss J (1987) *Nonlinear Optical Properties of Organic Molecules and Crystals*, vol. 2. Orlando, FL: Academic Press.

Clark DT (1994) Second European Conference on Smart Structures and Materials, *SPIE Proceedings*, vol. 2361. Bellingham, CA: SPIE.

Crowson A (1996) Smart Materials Technology and Biomimetics, *SPIE Proceedings*, vol. 2716. Bellingham, CA: SPIE.

Gobin P-F and Friend CM (2000) Smart Structures and Materials, *SPIE Proceedings*, vol. 4073. Bellingham, CA: SPIE.

Janocha H (1999) *Adaptronics and Smart Structures, Basics, Materials, Designs, and Applications*. Berlin: Springer.

Jorgenson GV and Lee JC (1985) Optical Materials for Energy Efficiency and Solar Energy Conversion IV, *SPIE Proceedings*, vol. 562. Bellingham, CA: SPIE.

Goff AH-L, Granqvist C-G and Lampert CM (1992) Optical Materials Technology for Energy Efficiency and Solar Energy Conversion XI: Chromogenics for Smart Windows, *SPIE Proceedings*, vol. 1728. Bellingham, CA: SPIE.

McDonuch A, Gardiner APT and McEwen KS (1994) In: Culshaw B (ed.) Second European Conference on Smart Structures and Materials, *SPIE Proceedings*, vol. 2361. Bellingham, CA: SPIE.

Zerbi G (ed.) (1993) *Organic Materials for Photonics*, European Materials Research Society Monographs, V6. Amsterdam: North-Holland.

Sol-Gel Materials

Lisa C Klein, Rutgers the State University of New Jersey, Piscataway, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Sol-gel optics represents a methodology for fabricating optical materials in fiber, thin film, and bulk geometries. Defined broadly, sol-gel optics encompasses a variety of optical devices that contain sol-gel processed components. The optical device may be based wholly on sol-gel processing or incorporate a sol-gel material, along with conventional optical materials. Principally, the optical devices prepared by sol-gel processing or assembled with sol-gel materials are traditional optics or, in some cases, opto-electronics. Among the current sol-gel optical materials, by far the most common are oxides. Further, the majority of sol-gel optics is for transparent components, which typically means glasses, and, consequently, silicates.

In fact, most sol-gel materials have been developed to duplicate conventional silica optical materials. While glass melting technology is highly advanced for making quality lenses, an equivalent material can be made using a sol-gel route. Similarly, ultralow-loss optical fiber can be made by chemical vapor deposition (CVD). Now sol-gel processing can be used to prepare comparable fibers for telecommunications. In cases where sol-gel processing produces an exact duplicate of optical materials produced other ways, the perceived advantage is that sol-gel precursors have a higher purity and better level of mixing than powder raw materials. Perhaps, a more significant advantage is that sol-gel formulations can be designed to promote a particular geometry in the final material, essentially making near-net shape materials.

While sol-gel processing is not designed to produce new optical phenomena, it does offer new ways to assemble materials to produce new optical designs, such as filters, Bragg gratings, amplifiers, and beam splitters. One noted case of sol-gel materials in nonlinear optics is the case where semiconductor particles are trapped in nanometer-sized pores. This leads to so-called quantum dots. The microporosity of the gel, no doubt, assists in this effect, but it is not the gel itself that has produced the optical effect. The optical effect is derived from the nanometer size of the particles. The gel serves as a transparent host that restricts the size of the particles and provides for their uniform distribution.

Sol-Gel Chemistry of Oxide Optical Materials

The sol-gel process for oxides is the name given to any one of a number of processes involving a solution or

sol that undergoes a sol-gel transition. Solution refers to a single-phase liquid, while a sol is a kinetically stabilized suspension of colloidal particles. At the transition, the solution or sol becomes a rigid, porous mass by destabilization, precipitation, or supersaturation. Once the solution or sol goes through a sol-gel transition, it is a rigid two-phase system of solid and solvent filled pores. The sol-gel transition is not reversible.

In using a sol-gel process, it is important to be familiar with the reagents, the intermediates and the products. The most important reagent in this case is a hydrolyzable organometallic which is called a metal alkoxide, $M(OR)_x$ where (OR) is an alkoxy group. Alkoxides are the precursors for SiO_2, GeO_2 or B_2O_3, among others. Hydrolysis occurs when the metal alkoxide and water are mixed in a mutual solvent.

The first step is choosing the right reagents. Since most of the glasses are silicates, we start with the silica precursor. Of the available silicon alkoxides, tetraethylorthosilicate (TEOS) is used most often, because it reacts slowly with water and comes to equilibrium as a complex silanol. The colorless liquid $Si(OC_2H_5)_4$ has a density of about 0.9 g/cm^3, is easy to handle safely and is extremely pure when distilled. Most metal alkoxides are prepared by the reaction of halogenated species or pure metals with alcohol.

The two other ingredients in a typical formulation are alcohol and distilled water. In the case of TEOS, ethanol serves as the mutual solvent for TEOS and water. Once TEOS is dissolved in ethanol to make it soluble in water, the chemical reactions hydrolyzation and polymerization begin. The chemical reactions are approximately:

Hydrolyzation

$$-\overset{|}{\underset{|}{Si}}-O-C_2H_5 + H_2O \rightarrow -\overset{|}{\underset{|}{Si}}-OH + C_2H_5OH \qquad [1]$$

Polymerization

$$-\overset{|}{\underset{|}{Si}}-O-C_2H_5 + Si-OH \rightarrow -\overset{|}{\underset{|}{Si}}-O-\overset{|}{\underset{|}{Si}} + C_2H_5OH \qquad [2]$$

An acid is usually added to control the rates of these reactions. The temperature is often ambient, and mixing is continued for between 1 and 3 hours. Complete hydrolysis does not occur. Instead, condensation occurs between silanol and ethoxy groups to give bridging oxygens or siloxane groups. The intermediates that remain soluble in the alcohol–water medium are silanols, ethoxysilanols, and polysiloxanes.

The equipment needed for laboratory-scale production consists of a narrow-necked glass flask with magnetic stir bar on a stirring hot plate. The reactants are added volumetrically to the flask in the order: solvent, TEOS, water, and acid. When everything is dissolved and the solution is clear, the hot plate may be turned on to accelerate reactions or a reflux condenser may be attached.

The parameters that influence the chemical reactions are temperature, pH, amount of water, solvent, and precursor. More complex compositions are easier to mix in longer chain alcohols, and longer chain organic groups give slower reaction rates. Low pH favors a polymerization scheme that gives linear molecules and the solution remains clear. High pH favors cluster growth and the solution may become cloudy. At high pH or high water, solutions produce cross-linked polymers or branched clusters. A small volume fraction of branched clusters restricts flow, when the same volume fraction of linear polymers does not. At a higher volume fraction, the linear polymers become tangled, and then low water solutions will gel. The result of these reactions under all conditions is an increase in the molecular weight of the oxide polymer. Eventually, the solution reacts to a point where the molecular structure is no longer reversible, and the sol-gel transition has been reached.

Shapes, Compositions and Precursor Concentrations

A sampling of compositions and precursors is given in Table 1. Some compositions such as SiO_2 and TiO_2–SiO_2 have been prepared in bulk, to duplicate materials made by melting for lenses and mirror blanks. Other compositions such as B_2O_3–SiO_2 have been prepared for thin-film applications. The distinguishing aspect between the solutions formulated for bulk objects and those formulated for thin films is the oxide concentration in the solution. An indication of the solution concentration is the molar ratio of water to alkoxide. The higher ratios are used when bulk objects are desired.

Table 2 gives a sampling of the geometries for sol-gel silica. For each geometry, an example of a shape is given, with an application for each shape and the approximate dimensions of the sol-gel material.

In terms of near-net shape processing, when the material goes through its sol-gel transition, it proceeds to dry and shrink. This phase is characterized either by isotropic shrinkage or anisotropic shrinkage. Monoliths fall into the first category with isotropic shrinkage following molding. Thin films on substrates fall into the second category with anisotropic shrinkage following dipping. Fibers also show

Table 1 Typical compositions of sol-gel optical materials

Oxides	*Precursors*	*Compositions in mol %*	R^*
Single oxide			
SiO_2	$Si(OC_2H_5)_4$	100 SiO_2	16
Two oxides			
B_2O_3–SiO_2	$Si(OC_2H_5)_4$ $B(OCH_3)_3$	80 SiO_2–20 B_2O_3	1.5
GeO_2–SiO_2	$Si(OCH_3)_4$ $Ge(OC_2H_5)_4$	90 SiO_2–10 GeO_2	4
TiO_2–SiO_2	$Si(OC_2H_5)_4$ $Ti(OC_3H_7)_4$	94 SiO_2–6 TiO_2	50
P_2O_5–SiO_2	$Si(OC_2H_5)_4$ H_3PO_4	90 SiO_2–10 P_2O_5	16
Three oxides			
Li_2O–Al_2O_3–SiO_2	$Si(OC_2H_5)_4$ $Al(OC_4H_9)_3$ $LiNO_3$	82 SiO_2–3 Al_2O_3–15 Li_2O	8
Na_2O–B_2O_3–SiO_2	$Si(OC_2H_5)_4$ $B(OCH_3)_3$ $NaNO_3$	82 SiO_2–12.4 B_2O_3–5.6 Na_2O	10

*Approximate ratio of water to alkoxide, e.g., $R = 16$ means 16 moles of H_2O per mole of $Si(OC_2H_5)_4$.

Table 2 Geometries and approximate dimensions for selected applications of sol-gel silica

Geometry	*Shape*	*Dimensions*	*Application*
Monolith	Flat disk	10 cm diameter 1 cm thick	Lens
	Cylinder	0.5 cm diameter 2 cm long	Rigid host for chromophores
Thin film	Film on Si wafer	500 nm thick	Antireflection
	Film on Hollow tube	~10 000 nm thick	Solar collector panel
Fiber	Rod	0.8 cm diameter 25 cm long	Preform for drawing
	Filament	150 micron diameter/100 m length	Lightguide

anisotropic shrinkage, though it depends on whether or not the fiber is drawn from a preform many times its diameter or produced from a filament that is closer to its diameter.

Sol-Gel Monolith Processing

Monoliths formed from an alkoxide solution are achieved only in dilute solutions. Therefore, alkoxide gels are somewhat difficult to dry because of their small pores (<10 nm). To avoid cracking in alkoxide gels, different treatments have been tried. Aerogels are dried in an autoclave by hypercritical techniques. That is, the solvent is removed above its critical point. The resulting gel is about 10% dense and shows no shrinkage. Xerogels are dried by natural evaporation. Xerogels are 60% dense and have reductions 40–60% in volume.

Alternatively, monoliths can be formed from colloidal solutions. These particles coalesce at gelling. Colloidal gels are easily dried, and they are less likely to crack than alkoxide xerogels. While alkoxide gels have small pores, colloidal gels have larger pores or voids between particles. For the same reason, colloidal gels densify at higher temperatures than alkoxide xerogels. The higher temperature for colloidal gels may lead to undesirable crystallization.

By either route, a serious problem for sol-gel processing is the presence of water. Especially when monoliths are used for fiber drawing, residual OH ions have a deleterious effect on optical properties. That is why non-aqueous solvents, as well as fluorination or chlorination treatments, have been investigated to reduce the OH content.

Sol-Gel Fiber Processing

While monolithic pieces need high water content, fibers can be drawn out of low water content

solutions. The sol-gel process allows one to bait and draw a string of gel about the same diameter as the desired fiber directly from the solution. This has been shown where the water to alkoxide ratio was controlled to give the property of spinnability. The time and manner in which the system loses fluidity depend on composition and catalyst. Low water solutions show a gradual increase in viscosity before losing fluidity, while high water solutions lose fluidity more abruptly from a lower viscosity. A problem with this method is the shape of the fiber, which tends not to be cylindrical.

Another approach to making near-net shape fibers from a sol-gel process involves coating a shell or concentric shells of sol-gel glass on a filament of textile fiber. This approach is referred to as the volatile host method, where a plastic filament is used as a template. Multiple dipping or adjustment of the viscosity can lead to different thicknesses. Upon exposure to atmospheric moisture, the films gel. After gelling, a rigid, continuous shell runs the length of the filament.

Once the shell has hardened, solvent and water can escape through interconnected pores. The shell shrinks in the radial dimension, but remains continuous along its axis. During heat treatment and desiccation, little if any densification occurs at temperatures well below the glass transition. The interconnected pore structure remains open until the volatile host is eliminated and undesirable gases are removed. This method is capable of producing optical fibers, but is more suitable for hollow fibers or ion exchange media.

Sol-Gel Thin Film Processing

Like sol-gel fibers, sol-gel thin films are prepared from low water content solutions. The process is exceedingly simple. A solution containing the desired oxide precursors is applied to a substrate by dipping or spinning.

At this time, the majority of sol-gel coatings are applied by dipping. This is an approach that allows the properties of the solution to control the deposition. A substrate is lowered into a vessel containing the solution. A meniscus develops at the contact of the liquid and the substrate. As the substrate is withdrawn, the meniscus generates a continuous film on the substrate. The process is able to apply a coating to the inside and outside of complex shapes simultaneously.

The equipment is inexpensive, especially in comparison to any deposition techniques that involve vacuum. Coatings can be applied to metals, plastics, and ceramics. Typically, the coatings are applied at room temperature, although most need to be calcined and densified with heating. Dip films are typically one micron, uniform over large areas and adherent. The time-to-gel is especially important when it comes to coatings because film formation, drying, and creation of pores must be rapid. Optimum film formulations correspond to those solutions that lose tackiness quickly.

In most cases, the effect of the viscosity is far greater than the surface tension. Two simple relationships are that the film thickness increases with increasing withdrawal rate and, for a given withdrawal speed, the film thickness increases with an increase in oxide content. The process is frequently a batch process, certainly on the laboratory scale. However, the batch process can be scaled up. Repeated dipping builds up a thicker film.

For one-side coating, a technique that is often used is spin coating. In this case, the substrate, usually a silicon wafer, is placed on a spinner, rotated at perhaps 200 rpm while solution is dripped on the center of the substrate. As with dip coatings, a film between 50 and 500 nm results.

After the Sol-Gel Transition

Having selected a shape and designed the formulations accordingly, there are several further steps common to monoliths, fibers, and films. First of all, the gels must be dried. Using the example of silica, TEOS is less than 30% by weight silica and the solution is even less. From the time the solution is applied to the time it gels, there is about 50% weight loss. Then, as the gel dries there is another reduction in weight by one-half. To go along with the weight loss there is about 70% volume reduction.

Films can be dried quickly in air because of the one thin dimension. It has been shown repeatedly that all shrinkage is taken up in the thin dimension and not in the plane of the substrate. Yet, the films remain adherent and continuous and maintain complete surface coverage. To go from a tacky film to a hard film may take only a few minutes.

Following drying and shrinkage comes the heat treatment. It would be inaccurate to say that glass has been formed at room temperature by simply reacting, gelling, and drying the solutions. In this state, the materials have many of the characteristics of the corresponding glass, but they are more or less porous. Water and solvent escape through interconnected pores that remain open at the surface until the gels are fired to temperatures well above 600°C. When needed to fulfill the requirements of an application, the hard gel is heated to various degrees of collapse.

The microporosity in silica xerogels is not removed entirely until 1000°C.

Applications of Sol-Gel Optical Coatings

Sol-gel processing has a special relevance to optical coatings. In most cases, the applications are substitutions of a sol-gel coating for a coating obtainable by other means. Most optical coatings achieve their effect by selective absorption and interference. By depositing 1–14 or more layers of dielectric thin films, using two or more different compositions, a variety of antireflection (AR), half-mirror, hot-mirror and cold-mirror coatings can be designed. An early design was for rear-view mirrors for cars, where the coatings consisted of titania-silica-titania interference filters that gave the effect of total reflectance and replaced conventional metallizing. Another early application was solar reflecting films for windows. These coatings consisted of Pd-containing titania films which show selective absorption.

A sampling of optical coatings is given in Table 3. The application, the mechanism, and the oxide composition are listed. Generally, these coatings can be prepared by other deposition techniques. Those applications which are specific to sol-gel coatings relate to the microporous nature of the sol-gel processed films. In Table 3, there are antireflective coatings of several types. Both borosilicates and titania silicates have been developed for this application. Changes in the composition, changes in the microstructure, and changes in the porosity can be used to grade the index of refraction. Antireflection in the ultraviolet range has been accomplished with thoria and hafnia films. Broadband antireflection has been used for laser optics. The sol-gel coating technique provides control over composition and deposition thickness. The layers, being quarter wavelength in thickness, require precise control.

Table 3 Thin-film applications

Optical application	*Mechanism*	*Typical composition*
Color filters	Selective Absorption	Cr_2O_3–SiO_2
Mirror	Interference	14-layer TiO_2: SiO_2
Antiglare	Absorption	CoO–SiO_2
Antireflection		
Narrow	Graded porosity	SiO_2
Broadband	Interference	Three-layer
	1/4 wavelength	SiO_2–TiO_2:ZrO_2: SiO_2
	Graded porosity	SiO_2
IR reflector	Interference	Three-layer TiO_2:Ag:TiO_2

Sol-Gel Rigid Hosts

An exciting way of looking at sol-gel materials is to use the gel as a host. The important point to note is that the sol-gel materials are transparent. While most of the materials are porous, the porosity is on a scale that does not scatter visible light. Therefore, most light is transmitted. The appeal of the bulk shape host is that it can be molded to near-net shape and can be cut and polished.

Hybrids

Several approaches have been taken to prepare so-called hybrid gels incorporating organic and inorganic components. Those methods that mix the components on the nanometer scale include the infiltration of previously formed oxide gels with monomers, reacting of alkoxide and organic monomers, or organic polymers with alkoxides in mutual solvents. The products of these processes are variously called ormosils (organically modified silicates), ceramers (ceramic polymers), interpenetrating networks (IPNs) and nanocomposites.

The original route to hybrids was one where a gel was formed first as a rigid matrix. Monomer was infiltrated into the open porosity of the shaped gel and polymerized inside the gel using UV curing, for example. Since silica gel is the easiest gel to prepare in monolithic form, one combination that has been investigated widely is silica gel/polymethyl methacrylate (PMMA).

Silica and PMMA, which both have good visible light transmission and similar index of refraction, have been used in hybrids, although it turns out that the match in index of refraction is less critical on the nanoscale. They are also glassy materials with similar strength. PMMA has much higher fracture energy than silica and is often used as glazing in applications where impact resistance is a priority. However, PMMA has a much lower hardness than silica and is less resistant to abrasion. Ideally, silica–PMMA hybrid nanocomposites have the strength and impact resistance of PMMA, and the hardness of silica in one. One obvious application is transparent, lightweight windows.

Encapsulation

Sol-gel hosts have been used to encapsulate a variety of organic and inorganic molecules, chromophores, and biological agents. Whatever is incorporated into the gel can act in two ways, to

give information about the gel host or to be influenced by the gel in its interaction with light. For example, spiropyrans have been incorporate for photochromics, CdS for nonlinear optics, Nd for tunable lasers, and various enzymes for sensors. Early on, silica gel hosts were used to encapsulate the electro-optically active organic compounds 2-methyl-4-nitroaniline (MNA). In comparison to conventional optical cells, MNA doped silica hosts retained optical quality at higher temperature and were more abrasion resistant. Since this discovery, silica hosts, in particular, and sol-gel processing, in general, has been used extensively for encapsulating molecules and species that cannot be heated, but need to be protected in a rigid matrix.

Future Directions

As more and more is learned about the sol-gel process, the greater the number of possibilities for useful optical devices. In the selected examples given above, a key feature is that the sol-gel material is transparent. In addition, the sol-gel material is suitable for application to a variety of substrates. Also, the process is a low-temperature process that is compatible with polymers and organic chromophores. In summary, sol-gel processing has had a number of successes in its application to optical materials, and its use is likely to expand.

See also

Optical Coatings: Thin-Film Optical Coatings.

Further Reading

Andrews MP and Najafi SI (eds) (1997) *Sol-Gel and Polymer Photonic Devices*. Bellingham, CA: SPIE Critical Reviews, CR68-03.

Brinker CJ and Scherer GW (1990) *Sol-Gel Science: The Physics and Chemistry of Sol-Gel Processing*. Boston, CA: Academic Press.

Dunn B and Zink JI (1991) Optical properties of sol-gel glasses doped with organic molecules. *Journal of Material Chemistry* 1: 903–913.

Klein LC (1993) Sol-gel optical materials. *Annual Reviews of Material Science* 23: 437–452.

Klein LC (ed.) (1994) *Sol-Gel Optics: Processing and Applications*. Dordrecht: Kluwer Academic.

Najafi SI (ed.) (1998) *Selected Papers on Sol-Gel for Photonics*. Bellingham, CA: SPIE Milestones Series, MS148.

Riesfeld R and Jorgenson CK (1992) Optical properties of colorants or luminescent species in sol-gel glasses. *Structure and Bonding* 77: 208–256.

Sol-Gel Optics. SPIE vols. 1328, 1758, 2288 and 3943, 1990–2000. Bellingham, CA: SPIE.

OPTICAL MICROLENSES

H Ottevaere and H Thienpont, Vrije Universiteit Brussel, Brussel, Belgium

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In today's world of information processing the role of optics and opto-electronics is expected to become increasingly important as the performance of communication, processing, sensing, and display technologies is continuously evolving. Making these photonics technologies faster and smaller requires, at the same time, the introduction of massive parallelism and miniaturization. As a result, high-quality, high-precision, and low-cost microlens arrays are becoming indispensable components. Moreover, the introduction of micro-optics provides new degrees of freedom for the system design and, therefore, also for the functionality. It allows solutions which are not possible with purely conventional optics such as the integration of different micro-optical components and systems, hence reducing the additional mechanical assembly cost as required in conventional optics.

Since the beginning of the 1990s, several research groups and industrial research labs have been focusing their attention on the development of fabrication techniques for refractive and diffractive microlenses. Microlenses and microlens arrays are used for beam shaping purposes such as collimation and focusing (e.g., in combination with laserdiode arrays, detector arrays, and fibers), for illumination (e.g., in display systems and projection systems), and for imaging purposes (e.g., in photocopiers).

In this article we will mainly focus on refractive microlenses and we intend to provide an introductory guide to those basic geometrical and optical parameters that are used to characterize refractive microlenses. Indeed, some elementary principles of geometrical optics (the geometrical laws for optical propagation without the inclusion of diffraction) and wave optics (in wave optics the imaging is studied as a result of the propagation of a wavefront to the image

plane; diffraction effects can be explained on the basis of this wave theory) are required to understand the characterization and performance evaluation of these lenses. We, therefore, bring forward the basic nomenclature of refractive micro-optics and will highlight how the performance of an optical lens is affected when reducing its physical dimensions.

We start this article by introducing the imaging properties of both thick and thin lenses in the paraxial approximation. In the section on lens aberrations, we discuss the different primary aberrations that influence the imaging properties of a lens. Combined effects of aberrations and diffraction give rise to some frequently used lens performance criteria which we discuss in the section on diffraction. Different aspects have to be considered when discussing the quality of a lens, because each type of microlens application requires specific performances. It is, therefore, not possible to come up with a single figure of merit which satisfactorily evaluates the overall quality of a microlens. Moreover, one needs specific information about the characteristics of the light sources and detectors that are used in the system and also details about the imaging system geometry and the required imaging contrast. We review them in the section on quality criteria.

In the section on interference, we describe these phenomena and focus on two-beam interference that forms the basis of the photonic instrumentation that can be used to characterize discrete microlenses and microlens arrays. For a complete evaluation of a microlens it is essential to measure both its geometrical properties (diameter, sag, surface profile, etc.) and its optical performance (focal length, wave aberrations, etc.). In the final section of this article we introduce three interference-based instruments which are crucial to the full quantitative characterization of refractive microlenses: a noncontact optical profiler based on a Mirau interferometer, a transmission Mach–Zehnder interferometer, and a Twyman–Green interferometer.

Geometrical Optics

Introduction

Within the paradigm of geometrical optics or ray-optics, light travels from the source along straight lines or rays. When a light ray traverses an optical system consisting of several homogeneous media in sequence, the optical path is a chain of straight-line segments. Discontinuities in the line segments occur each time the light is refracted or reflected. The laws of geometrical optics that describe the direction of the rays are described below.

The law of refraction, also called Snell's law
The refraction of a ray at a 'well-polished' surface between two isotropic media, with indices of refraction n_1 and n_2, is given by

$$n_1 \sin \varphi_1 = n_2 \sin \varphi_2 \qquad [1]$$

where φ_1 and φ_2 are the angles between, respectively, the incident and the refracted ray and the normal to the surface.

The law of reflection
When a ray is reflected on a flat 'polished' interface dividing two uniform media, the reflected ray remains within the plane of incidence and the angle of reflection equals the angle of incidence.

Geometrical wavefronts can then be constructed from the optical rays propagating through the system. Locally they are orthogonal to the rays and the optical path lengths along all rays from one wavefront to another are equal. In regions where the wavelength is considered to be negligible, as compared to the dimensions of the relevant components of the optical system, geometrical wavefronts are close approximations of the physical wavefronts.

The Paraxial Approximation

The basic optics of image location and magnification is demonstrated by the application of paraxial or first-order optics, also called Gaussian optics. In the paraxial approximation, only those points and rays will be considered which lie in the immediate vicinity of the optical axis, so that all angles φ are small ($\varphi \leq 15°$). This assumption is true in lens systems with small object and image field sizes and small apertures of pupils. The terms involving squares and higher powers of the off-axis distances or of the angles, which the rays make with the optical axis, will not be discussed here. This leads to:

$$\sin \varphi = \tan \varphi = \varphi + O_{\geq 2}(\varphi) \approx \varphi \qquad [2]$$

With this assumption we can write the refraction law by the linear relation:

$$n_1 \varphi_1 = n_2 \varphi_2 \qquad [3]$$

Refraction at a spherical surface
In Figure 1 a spherical surface with a radius of curvature R forms an interface between two materials with indices of refraction n_1 and n_2. Two rays are shown emanating from an axial object point S. One is an axial ray, normal to the surface of the lens at its vertex V. The vertex of a lens is defined as the point on the refracting surface at the center of its free aperture. This assumes that the aperture is circular

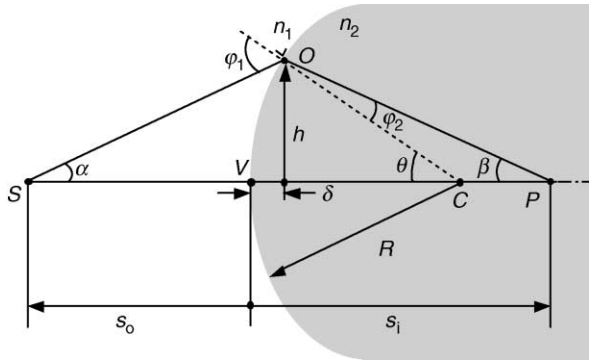


Figure 1 Refraction at a spherical surface.

and centered and the surface is spherical. More generally, if the surface is not spherical but has rotational symmetry, the vertex is the point where the axis of symmetry intersects the optical surface. It is refracted without any change in direction. The other ray is an arbitrary ray incident at O and refracting there according to Snell's Law.

We can rewrite the paraxial refraction law (eqn [3]) and derive a relation between the object and image distances, s_o and s_i, respectively.

From the triangles SOC and POC, we have:

$$\varphi_1 = \alpha + |\theta| \qquad |\theta| = |\beta| + \varphi_2 \qquad [4]$$

where $|\theta|$ and $|\beta|$ represent the absolute values of θ and β.

We also have:

$$\tan\alpha = \frac{h}{|s_o| + \delta} \qquad \tan|\beta| = \frac{h}{s_i - \delta}$$
$$\tan|\theta| = \frac{h}{R - \delta} \qquad [5]$$

As discussed above, for paraxial rays we may approximate the sine and tangent of an angle by the angle itself and one can show that the small distance δ may be ignored. In accordance with the sign convention the object distance, being to the left of the refracting surface, is to be taken as negative. Inserting eqns [4] and [5] into [3] results in:

$$\frac{n_2}{s_i} - \frac{n_1}{s_o} = \frac{n_2 - n_1}{R} \qquad [6]$$

This equation does not contain the angle α, so the image distance s_i is the same for all paraxial rays starting from S.

The spherical surface power $\wp$ is now defined as:

$$\wp \equiv \frac{n_2 - n_1}{R} \qquad [7]$$

We can relate this power of the surface $\wp$ to the object and image distance, by combining eqns [6] and [7]:

$$\frac{n_2}{s_i} - \frac{n_1}{s_o} = \wp \qquad [8]$$

The distance from the object point to the vertex of the refractive surface is called the front focal length f_{FFL}, if the image is at infinity ($s_i = \infty$). Equations [6] and [8] then give:

$$f_{FFL} = -\frac{n_1}{n_2 - n_1}R = -\frac{n_1}{\wp} \qquad [9]$$

Similarly, if the image is placed at the back focal length f_{BFL}, such that the object distance equals infinity ($s_o = \infty$), then:

$$f_{BFL} = \frac{n_2}{n_2 - n_1}R = \frac{n_2}{\wp} \qquad [10]$$

Thick lenses

A thick lens can be treated as a component consisting of two spherical refracting surfaces separated by a distance d between their vertices V_1 and V_2, as shown in Figure 2. The determination of the focal length as a function of the object and image distance and of the radii of curvature can also be solved using the matrix formalism.

The first and second focal points or the object and image foci, F_o and F_i, can be measured from the two vertices or poles and are called the front and the back focal length denoted by f_{FFL} and f_{BFL}, respectively (Figure 2). The incident and the emerging rays will virtually meet at points that belong to two curved surfaces. These surfaces, approximating planes in the paraxial region, are termed the primary and secondary principal planes, PP_1 and PP_2. Points where the primary and secondary principal planes intersect the optical axis are known as the first and second principal points, H_1 and H_2, respectively.

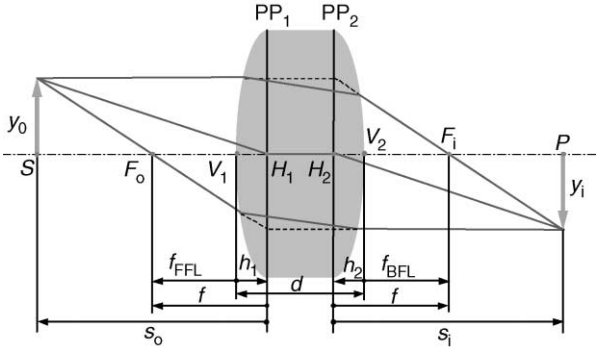


Figure 2 The thick lens geometry.

They provide a set of very useful references from which to measure several of the system parameters. To fully describe a system by its cardinal points, two more points are needed in addition to the two focal points, F_o and F_i and the two principal points, H_1 and H_2. These points are called the nodal points. However, they coincide with the principal points when the lens is surrounded on both sides by the same medium, for example, air. Note that h_1 and h_2 determine the positions of the principal points relative to the vertices V_1 and V_2, while f determines the focal point position relative to the principal points. One can show that these distances are given by:

$$h_1 = \frac{-R_1 n_m d}{n_l(R_2 - R_1) + (n_l - n_m)d}$$

$$h_2 = \frac{R_2 n_m d}{n_l(R_2 - R_1) + (n_l - n_m)d} \qquad [11]$$

$$f = \frac{n_m}{(n_l - n_m)\left[\frac{1}{R_1} - \frac{1}{R_2} + \frac{d(n_l - n_m)}{n_l R_1 R_2}\right]}$$

We now derive the Gaussian formulae that describe imaging through two surfaces which are rotationally symmetric about the same axis. We assume here that the lens, with refractive index n_l, faces the same medium of refractive index n_m on both sides (**Figure 3**). We can correctly use this hypothesis since the microlenses which we discuss throughout this work are mainly surrounded by air. Let R_1 and R_2 be the radii of curvature of the surfaces at their axial points, measured as positive when the surface is convex towards the incident light.

The image of a given object, formed by refraction at the first surface with radius of curvature R_1, becomes the object for refraction at the second surface. The image formed by the second surface with R_2 as radius is then the final image due to the action of the thick lens. We know from eqn [6] that the paraxial rays issuing from S at a distance s_{o1} from V_1 will meet at P', at a distance which we call s_{i1}. This gives:

$$\frac{n_l}{s_{i1}} - \frac{n_m}{s_{o1}} = \frac{n_l - n_m}{R_1} \qquad [12]$$

Thus, at the second spherical surface, eqn [6] becomes:

$$\frac{n_m}{s_{i2}} + \frac{n_l}{(s_{i1} - d)} = -\frac{n_m - n_l}{R_2} \qquad [13]$$

As discussed above, we have:

$$-\frac{1}{f_{FFL1}} = \frac{n_l - n_m}{n_m}\frac{1}{R_1}$$
$$\frac{1}{f_{BFL2}} = -\frac{n_m - n_l}{n_m}\frac{1}{R_2} \qquad [14]$$

where f_{FFL1} is the front focal length of the first refracting surface and f_{BFL2} the back focal length of the second refracting surface.

Eliminating s_{i1} in eqns [12] and [13] and inserting [14] yields:

$$d = \frac{n_l}{n_m\left(\frac{1}{s_{o1}} - \frac{1}{f_{FFL1}}\right)} - \frac{n_l}{n_m\left(\frac{1}{f_{BFL2}} - \frac{1}{s_{i2}}\right)} \qquad [15]$$

where s_{o1} is measured with respect to V_1 and s_{i2} is measured with respect to V_2. After some calculations this equation may also be put in the form:

$$s_{o1}s_{i2} - \frac{f_{BFL2}(n_l f_{FFL1} + d n_m)}{n_l(f_{FFL1} + f_{BFL2}) + d n_m}$$
$$s_{o1} - \frac{f_{FFL1}(n_l f_{BFL2} + d n_m)}{n_l(f_{FFL1} + f_{BFL2}) + d n_m} \qquad [16]$$
$$s_{i2} + \frac{d n_m f_{FFL1} f_{BFL2}}{n_l(f_{FFL1} + f_{BFL2}) + d n_m} = 0$$

Obviously eqn [16] is much too cumbersome to find image locations. However, we can transform and reduce the latter equation to the form:

$$\frac{1}{s_{i2} - h_2} - \frac{1}{s_{o1} - h_1} = \frac{1}{f} \qquad [17]$$

where h_1, h_2 and f are the unknowns.

Indeed, rewriting eqn [17] gives:

$$s_{o1}s_{i2} - (f + h_2)s_{o1} + (f - h_1)s_{i2} + f(h_1 - h_2) + h_1 h_2 = 0 \qquad [18]$$

Equations [16] and [18] will be identical if the coefficients of their corresponding terms are equal;

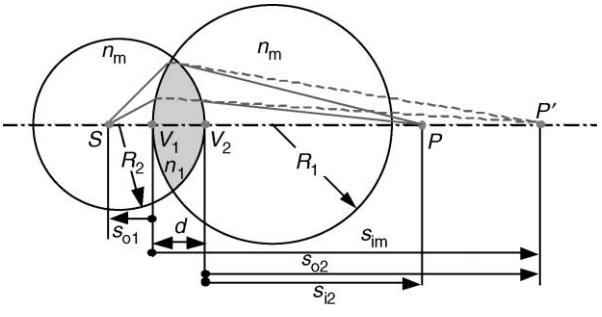


Figure 3 Refraction at both spherical surfaces of a thick lens.

that is when:

$$f + h_2 = \frac{f_{\mathrm{BFL2}}(n_{\mathrm{l}} f_{\mathrm{FFL1}} + dn_{\mathrm{m}})}{n_{\mathrm{l}}(f_{\mathrm{FFL1}} + f_{\mathrm{BFL2}}) + dn_{\mathrm{m}}}$$

$$f - h_1 = -\frac{f_{\mathrm{FFL1}}(n_{\mathrm{l}} f_{\mathrm{BFL2}} + dn_{\mathrm{m}})}{n_{\mathrm{l}}(f_{\mathrm{FFL1}} + f_{\mathrm{BFL2}}) + dn_{\mathrm{m}}} \qquad [19]$$

$$f(h_1 - h_2) + h_1 h_2 = \frac{dn_{\mathrm{m}} f_{\mathrm{FFL1}} f_{\mathrm{BFL2}}}{n_{\mathrm{l}}(f_{\mathrm{FFL1}} + f_{\mathrm{BFL2}}) + dn_{\mathrm{m}}}$$

These three equations may be solved for the three unknowns h_1, h_2, and f:

$$f = \frac{n_{\mathrm{m}}}{(n_{\mathrm{l}} - n_{\mathrm{m}})\left[\frac{1}{R_1} - \frac{1}{R_2} + \frac{d(n_{\mathrm{l}} - n_{\mathrm{m}})}{n_{\mathrm{l}} R_1 R_2}\right]}$$

$$h_1 = \frac{-R_1 n_{\mathrm{m}} d}{n_{\mathrm{l}}(R_2 - R_1) + (n_{\mathrm{l}} - n_{\mathrm{m}})d} \qquad [20]$$

$$h_2 = \frac{R_2 n_{\mathrm{m}} d}{n_{\mathrm{l}}(R_2 - R_1) + (n_{\mathrm{l}} - n_{\mathrm{m}})d}$$

Thus we see that the relation between the object position and the image position (eqn [16]) is expressed in the more convenient form of eqn [17] with the values of h_1, h_2, and f given by eqn [20]. Here h_1, h_2 and f depend on the constants of the system where, as shown in eqn [11], h_{i} represents the distance between the vertex V_{i} and the principal point H_{i}, and f is the focal length measured from a principal point on the axis and not from the vertex of the lens. If we now assume that:

$$s_{\mathrm{o}} = s_{\mathrm{o1}} - h_1, \qquad s_{\mathrm{i}} = s_{\mathrm{i2}} - h_2 \qquad [21]$$

we can write the thick lens equation as:

$$\frac{1}{s_{\mathrm{i}}} - \frac{1}{s_0} = \frac{1}{f} \qquad [22]$$

where the object and image distance, s_{o} and s_{i}, respectively, are measured from the principal points of the lens. We can conclude that the formula that describes imaging through two surfaces may be simplified to the thick lens formulation given in eqn [22] by choosing the principal points H_{i} instead of the vertices V_{i}, from which S and P are to be measured.

Thin lenses

The preceding eqn [20] takes a particularly simple form when the lens is so thin that the axial thickness of the lens d may be neglected. But obviously there exists no real lens with a thickness equal to zero. A thin lens is only a concept which is an extremely useful tool in optical system design. When a lens or an optical system has a zero thickness, the object and image calculations can be greatly simplified. In this case, the two principal points are coincident and are located where the thin lens is situated.

As a further simplification, we assume that the surrounding medium of the lens is air which means that $n_{\mathrm{m}} \approx 1$. When $n_{\mathrm{l}} = n$, we find:

$$\frac{1}{f} = \frac{1}{s_{\mathrm{i}}} - \frac{1}{s_{\mathrm{o}}} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \qquad [23]$$

These equations are often called the thin-lens equation and the lensmaker's formula, respectively, because the latter predicts the focal length of a lens, fabricated with a given refractive index and radii of curvature. This formula is very useful in the optical design process as a first-order approach to calculate the features of the thin lenses and results in a set of lens powers (or focal lengths) for the different lenses and in the spacings between them. Once the thin lens optical system is solved and the primary aberrations are corrected for, one can design the thick lens system. The corresponding thick lens system must have the same component powers and the lenses must be spaced apart by the same distances as for the thin lens design, but the thick lens spacings must be measured from the principal points of the lenses instead of from the vertices of the lenses.

Other lens characteristics

Another basic parameter of a refractive lens is its focal number $f\#$. The f-number, also called the speed of the lens, is defined as the ratio of its effective focal length and the lens aperture diameter D (Figure 4):

$$f\# = \frac{f}{D} \qquad [24]$$

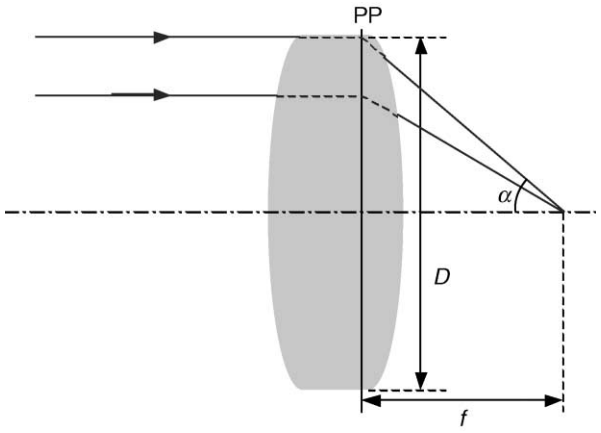


Figure 4 A lens with an aperture diameter D, a focal length f, and a semi-angle α.

The effective focal length f of a lens or often called the focal length is defined from a principal point on the axis of the lens and not from its vertex. Typical focal numbers for refractive microlenses range between 0.8 and 30.

The numerical aperture (NA), is another way of defining the same characteristic. It is described as the sine of the angular semi-aperture α in the object space multiplied by the refractive index n of the object space (Figure 4):

$$\mathrm{NA} = n \sin \alpha \qquad [25]$$

Both the focal number ($f\#$) and the NA are related in the paraxial approximation by

$$f\# = \frac{n}{2\mathrm{NA}} \qquad [26]$$

Lens Aberrations

Introduction to Third-Order Monochromatic Aberrations

An exact ray trace or a measurement on a real lens will certainly disclose inconsistencies with the corresponding paraxial description. Such deviations from the idealized Gaussian optics are known as aberrations. For monochromatic light there are five third-order or Seidel aberrations: three of them, namely spherical aberration, coma, and astigmatism, deteriorate the image by blurring it, while the remaining two, curvature of field and distortion, deform the image. An additional aberration, chromatic aberration, results from the wavelength dependence of the imaging properties of an optical system. We will not discuss the latter as this aberration can be minimized by selecting a lens material with a refractive index that only varies by a negligible amount over the wavelength range of interest.

For a brief discussion of aberrations we consider the case of a perfect lens focusing a collimated light beam. In this case, the lens transforms the incoming planewave into a spherical wave, also called the reference sphere. For a lens with aberrations, however, the real wavefront behind the lens shows deviations from the ideal spherical shape (Figure 5). The optical path length [AB] may be called the wavefront aberrations $\psi(x, y)$ and can be calculated as:

$$\psi(x, y) = W(x, y) - R(x, y) \qquad [27]$$

where $W(x, y)$ represents the spherical wavefront in the Gaussian approximation and $R(x, y)$ depicts the real wavefront. The wave aberration is given by the optical path difference along a ray between the

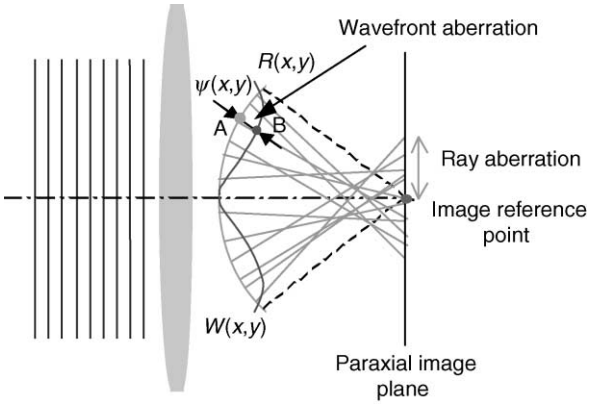


Figure 5 Relation between wavefront and ray aberration.

reference sphere and the real wavefront behind the lens. Alternatively, the aberrations can be described in terms of the ray aberrations. In this case, not all the rays intersect the paraxial image plane at the same point and the distribution of the ray intercepts around the paraxial image point can be used to measure the amount of light blur. These two ways of quantitatively describing the aberrations are apparently not independent of each other, the ray aberrations being the derivate of the wavefront aberrations.

There are several sets of suitable polynomials over which the wavefront aberrations can be expanded. The Zernike polynomials will be further used in this work (see, for example, the Mach–Zehnder and Twyman–Green interferometers in the section on Microlens testing).

For rotational symmetric optical systems, the most common approach is to expand the aberration function ψ to the fourth order as a power series in polar pupil coordinates. This corresponds to the third-order ray aberrations or primary aberrations. The different terms in the aberration polynomial ψ are to the fourth power of the object coordinate h' and the pupil coordinate r. In that case, the wavefront aberration polynomial ψ, as a function of the object coordinate h' and the polar pupil coordinates r and θ (Figure 6), is given by

$$\psi(h', r, \theta) = {}_0C_{40} r^4 + {}_1C_{31} h' r^3 \cos\theta + {}_2C_{22} h'^2 r^2 \cos^2\theta + {}_2C_{20} h'^2 r^2 + {}_3C_{11} h'^3 r \cos\theta \qquad [28]$$

The ${}_iC_{jk}$ coefficients are subscripted by the numbers that specify the powers of the term dependence on h', r and $\cos\theta$, respectively. These terms comprise the five monochromatic or Seidel aberrations as shown in Table 1.

Each aberration is characterized by its dependence on the deviation from the optical axis h', on the aperture of the refracting surface r, and on the

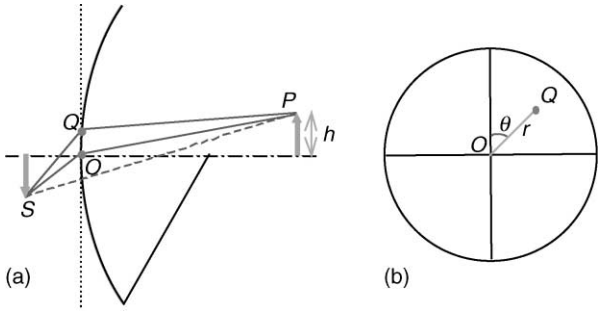


Figure 6 (a) Imaging of an off-axis point P; (b) frontal view of a portion of the wavefront.

Table 1 The five monochromatic or Seidel aberrations

r^4	Spherical aberration
$h'r^3 \cos\theta$	Coma
$h'^2 r^2 \cos^2\theta$	Astigmatism
$h'^2 r^2$	Curvature of field
$h'^3 r \cos\theta$	Distortion

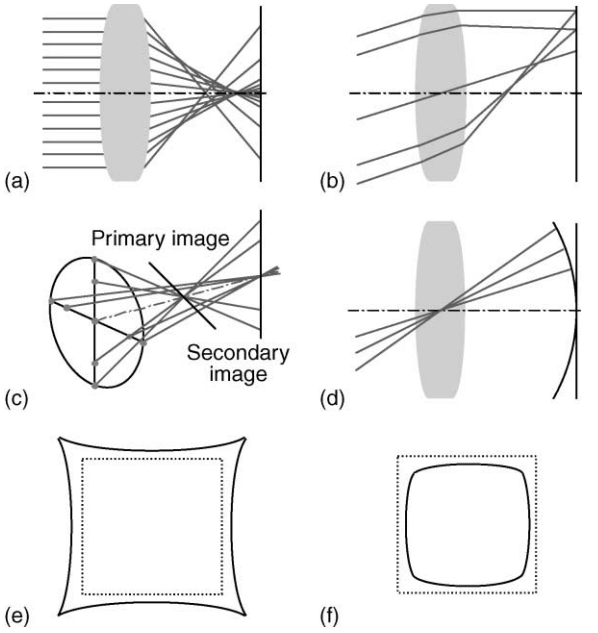


Figure 7 The primary aberrations: (a) spherical aberration – negative SA: the rays through the outer zones of the lens focus closer to the lens than the rays through the central zones; (b) coma – positive coma: the rays through the outer zones of the lens form a larger image than those going through the center; (c) astigmatism – negative astigmatism: the tangential or vertical fan of rays is focused to the left of the horizontal or sagittal fan; (d) curvature of field – negative curvature of field: images farther from the axis focus nearer to the lens than the on-axis images; (e) positive distortion – the magnification increases as the field angle increases; (f) negative distortion.

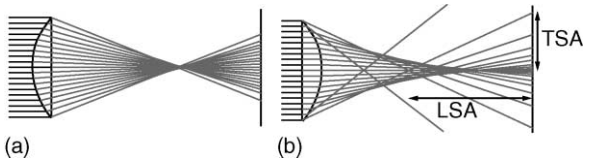


Figure 8 Ray diagram of a planoconvex lens ($D = 200$ μm; $f = 249$ μm) with planewave illumination incident on (a) the curved surface and (b) on the plane surface. LSA (longitudinal SA) represents the difference between the paraxial and the marginal focus while TSA (transversal SA) shows the radius of the spot in the paraxial focal plane.

symmetry around the optical axis θ. Equation [28] can be used to estimate how the aberration blurs change if the aperture or field of view of a lens is changed. In the following section we briefly describe each of these aberrations in terms of their visual effects.

Spherical Aberration

The first term in eqn [28], ${}_0C_{40}r^4$, is the only term in the third-order wave aberration that does not depend on the deviation h' from the optical axis. Thus, spherical aberration (SA) is the only monochromatic aberration that occurs on the optical axis and that exists even for axial object and image points. As shown in Figure 7a, SA results from the rays that enter via different zones of the aperture and are focused at different distances from the lens. It is evident from Figure 7a that the best focus is not at the paraxial plane. The best focus, however, is found at the location where the circular image blur is the same over the entire field of view and is called the circle of least confusion.

Because most applications consist of on-axis illuminated planoconvex microlenses, we will firstly focus our attention on how microlenses have to be positioned to minimize SA. When we consider a planoconvex lens with the curved surface towards the infinite conjugate for an incident parallel beam, both surfaces are contributing approximately equal amounts of ray bending. Figure 8a shows the ray diagram for a planoconvex lens as obtained with the ray-tracing program Solstis. Solstis is a commercial photonics design software, consisting of different modules, each of which is dedicated to a certain optical modeling approach. If, however, the flat surface is turned toward the incident rays, then the first surface does not bend the rays and all the ray-bending work is done by the second surface (Figure 8b). We can conclude here that the best orientation is that which distributes the ray bending equally between the two surfaces of the lens.

Because the ray blur varies with the third power of the lens aperture, lowering the aperture diameter will cause a rapid decrease in the extent of the spherical

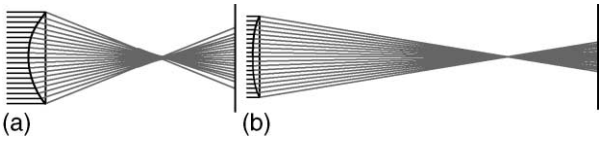


Figure 9 Ray diagram of a planoconvex lens (a) with a focal length *f* and (b) with a focal length *2f*.

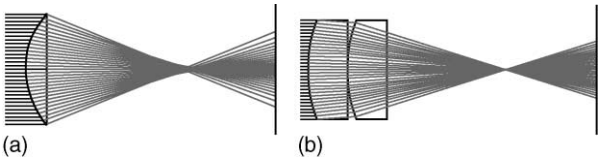


Figure 10 (a) Ray diagram for a planoconvex lens ($D = 200$ μm; $f = 249$ μm). (b) The spherical aberration is reduced with a factor 4 by splitting the optical power of the original lens and using a doublet ($D = 200$ μm; $f = 621$ μm for both lenses).

aberration blur. For a given aperture of the lens, the spherical aberration is a function of the focal length of the lens and increases with a decreasing focal length or an increasing NA (Figure 9).

A change in SA, contributing to the wave aberration, for a variation of the lens diameter D or the focal length f can be estimated with the following rules:

$$\text{Spherical aberration} \div D^4 \quad \text{for a constant } f$$

$$\text{Spherical aberration} \div 1/f^3 \quad \text{for a constant } D$$

However, by preserving the focal length and hence the optical power of the lens, the SA can be changed by varying the radii of the surfaces. This would be the same as physically bending a lens made of flexible plastic. In Figure 10 we have reduced the spherical aberration with a factor of four by splitting a lens into two elements, each with a longer focal length, keeping the total optical power of the elements the same as the original lens. This means that, in practice, high NA microlenses can be replaced by a number of microlenses reducing the SA significantly.

However, in optical systems, where only a small number of surfaces are allowed, aspherical surfaces can be used to correct the SA. In Figure 11 we show that making the lens shape elliptical while maintaining the other characteristics of the microlens ($D = 200$ μm; $R = 150$ μm) results in a reduction of the SA. The most common form of an aspheric surface is a rotationally symmetric surface with the sag, defined as:

$$z_{\text{conic}} = \frac{R^{-1}(x^2 + y^2)}{1 + \sqrt{1 - (1 + K)R^{-2}(x^2 + y^2)}} \qquad [29]$$

where R is the radius of curvature, K the conic parameter, and x and y the coordinates in the substrate plane. The meaning of K is given in Table 2.

In eqn [28] we have shown that the wavefront aberrations can be written as the sum of the Seidel or third-order aberrations. Furthermore, we have seen that only the SA contributes to the wavefront aberration when we illuminate the microlens on-axis ($h' = 0$). When we limit our experiments to on-axis illumination during the optical characterization of the microlenses, the wavefront aberrations are a measure for the SA. To prove this, we have calculated and plotted in Figure 12, the SA and the wavefront aberrations as a function of the focal length for a 200 μm lens ($n = 1.48$).

Both the SA and wavefront aberration values are obtained with the ray-tracing software Solstis. The wavefront aberration is calculated here as the optical path difference [BA] between a real wavefront surface and a spherical reference surface. From these results we can observe that for microlenses with an $f > 330$ μm or an NA lower than 0.3 only the spherical aberration contributes to the root-mean-square (RMS) wavefront aberration. The RMS or standard deviation σ of a number N of data x_i ($i = 1, \ldots, N$) with a mean value X can be written as $\sigma = \sqrt{\sum_{i=1}^{N} (X - x_i)^2 / N}$. For smaller focal length microlenses, however, we notice that higher-order aberrations appear and compensate the SA, resulting in a lower wavefront aberration compared to the SA. We can conclude that the wavefront aberration values represent the optical performance of the microlenses with a main contribution from the SA for on-axis illumination.

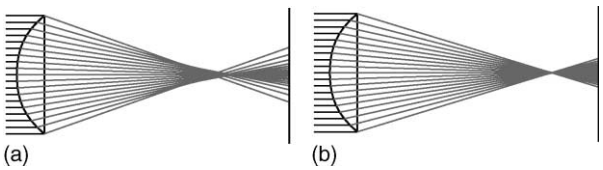


Figure 11 (a) Ray diagram of a spherical planoconvex lens ($D = 200$ μm; $R = 150$ μm; $K = 0$) with planewave illumination. (b) The spherical aberration is reduced by making the lens shape elliptical ($K = -0.57$).

Table 2 The surface type for different conic parameters *K*

Conic constant K	*Surface type*
$K = 0$	Sphere
$K < -1$	Hyperboloid
$K = -1$	Paraboloid
$-1 < K < 0$	Ellipsoid
$K > 0$	Oblate ellipsoid

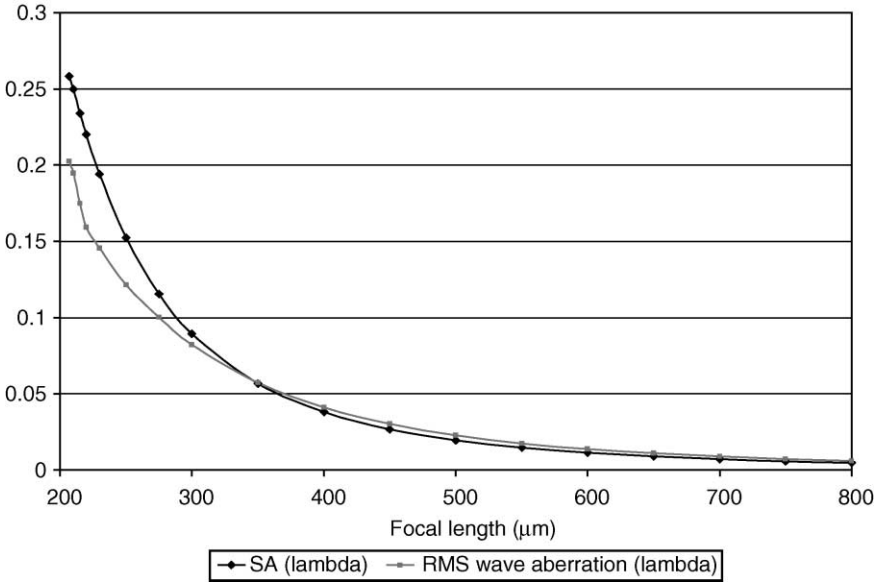


Figure 12 The spherical aberration SA and the RMS wave aberration as a function of the focal length for a 200 μm lens ($n = 1.48$).

Other Seidel Aberrations

Coma is an image-degrading aberration that increases rapidly with the lens aperture r. It indicates an off-axis aberration ($h' \neq 0$) that is nonsymmetrical around the optical axis ($\cos\theta \neq$ constant). Figure 7b illustrates the aberration due to a vertical fan of parallel rays refracted by a spherical lens. It results from a lens having different magnifications for rays passing different zones of its aperture. Each circular zone of the lens forms a circular image called the comatic circle. Every other fan of rays forms images that complete the comatic circle. The combination of all these circles causes a comet-shaped flare.

Astigmatism is the imaging of an off-axis point as two perpendicular lines at different distances. In this aberration the rays from a point object converge at a certain distance from the lens to a line which is called the primary image. This line lies in the sagittal plane and is perpendicular to the plane defined by the optical distance and the object point. At a somewhat different distance from the lens they converge to a second line, called the secondary image, which is parallel to the tangential plane. This effect is shown in Figure 7c. The elimination of astigmatism requires that the tangential and sagittal surfaces be made to coincide. The resulting surface is then called the Petzval surface. Although in that case astigmatism has been eliminated, the associated aberration, called curvature of field, remains. So off-axis images are focused on a curved saucer-shaped surface as drawn in Figure 7d, instead of on an ideal flat image surface.

The last of the five primary aberrations is distortion and causes straight lines which do not intersect the axis to be imaged as curved lines. This effect is not related to a lack of sharpness of the image but results from the fact that the magnification varies across the field and induces the image of a square object to be bowed outward (Figure 7e) or sagged inward (Figure 7f).

We have shown that three of the five Seidel aberrations (namely SA, coma, and astigmatism) are responsible for a lack of sharpness of the image. The other two (i.e., curvature of field and distortion) are related to the position and the form of the image. In general, it is impossible to design a system which is free from all the primary as well as the higher-order aberrations, and therefore, a suitable compromise as to their relative magnitude has to be made. In some cases, the effects of the Seidel aberrations are reduced by balancing them against aberrations of higher orders. In other cases, one has to eliminate certain aberrations completely, even at the price of introducing aberrations of other types. So far we have studied the aberration effects on the basis of geometrical optics only. However, if the aberrations are very small (wave aberrations of the order of a wavelength or less), diffraction starts to play an important role. That is why, in the next section, we will discuss the role of diffraction on the imaging quality of a lens.

Diffraction

Introduction

So far we have used simple geometrical optics to evaluate the effects of aberrations – deviations of the

wavefront behind the lens from the ideal shape – that influence the quality of an image generated by a lens. However, we know that even for an aberration-free lens, blur of the spot in the image plane occurs. This means that diffraction effects at the lens aperture can also influence the spot size and the point spread function (PSF), certainly because diffraction gains importance when miniaturizing optical components. Therefore, we have to introduce diffraction before we can bring forward the most important figures of merit for microlens performance and study its effects on imaging quality.

Fraunhofer and Fresnel Diffraction

In certain regions of space occupied by our optical lens system, the simple geometrical model of energy propagation is inadequate. In particular, deviations from this model can be expected in the immediate vicinity of the boundaries of shadows and in regions where a large number of rays meet. These regions are of great practical interest, as they include that part of the image space in which the optical image is situated. These deviations are manifested by the appearance of diffraction fringes. Also, the phenomena involving the superposition of coherent wavefronts cannot be treated with a geometrical model of energy propagation. Thus a more accurate treatment is necessary where the wave nature of light and diffraction are taken into account. Wave optics is such an approach with which the actual light intensity, that will be observed as a result of the propagation of a wavefront to the image plane, can be constructed. According to Huygens' construction, every point of a wavefront may be considered as a center of secondary disturbance, which gives rise to spherical wavelets. The wavefront at any later instance may be regarded as the envelope of these wavelets. Fresnel was able to account for diffraction by supplementing Huygens' construction with the postulate that the secondary wavelets mutually interfere. We discuss diffraction here, on the basis of wave theory, by the application of Huygens' construction together with the principle of interference. The latter principle, also called the Huygens–Fresnel principle, sufficiently describes the propagation of light in free space.

This Huygens–Fresnel principle may be regarded as an approximate form of the integral theorem of Helmholtz and Kirchhoff. Kirchhoff's theory makes an approximation to the diffraction of scalar waves. This scalar theory is adequate enough for the treatment of problems in instrumental optics where polarization effects are of minor importance and can be neglected. If we consider a strictly monochromatic electric field $\varepsilon(x, y, z, t) = E(x, y, z)\cdot e^{-i\omega}$ in any ideal isotropic dielectric, the space-dependent part will satisfy the homogeneous time-independent wave equation: $(\nabla^2 + k^2)E = 0$, also known as the Helmholtz equation where $k = \omega/c$. This theorem expresses the solution E of the Helmholtz equation at an arbitrary field point P, in terms of its solution E and its first derivatives $\partial E/\partial n$ at all points on an arbitrary closed surface surrounding P, provided the medium is a homogeneous dielectric:

$$E(P) = -\frac{1}{4\pi}\oiint_S \left[E\overline{\nabla}\left(\frac{e^{-iks}}{s}\right) - \frac{e^{-iks}}{s}\overline{\nabla}E \right]\bar{n}\,dS \qquad [30]$$

Here, s is the distance from point P to the point on surface S, n is the outward normal to S, and $k = \omega/c$ is the wave number in vacuum.

When we consider a monochromatic wave, from a point source P_o propagating through an aperture Ω in a plane opaque screen, and P is the point at which the light disturbance has to be calculated, the theorem can be reduced to an approximate but much simpler form (**Figure 13a**) on condition that we can assume that the linear dimensions of the aperture Ω, although large compared to the wavelength, are small compared to the distances of both P_o and P from the screen.

As a result of Kirchhoff's theorem we obtain:

$$E(P) = \frac{iA}{2\lambda}\iint_\Omega \frac{e^{-ik(r+s)}}{rs}[\cos(n,r) - \cos(n,s)]dS \qquad [31]$$

where A is a constant, λ the wavelength, and r and s are the distances from element dS on Ω to P_o and P, respectively. Equation [31] is known as the Fresnel–Kirchhoff diffraction formula. In case a more general field E_o is incident at the aperture Ω, replacing the point source $A\,e^{-ikr}/r$, an equivalent formulation, known as the Huygens integral, can be found:

$$E(P) = \frac{i}{\lambda}\iint_\Omega E_o(x_o, y_o, z_o)\frac{e^{-iks}}{s}\cos(n,s)dS \qquad [32]$$

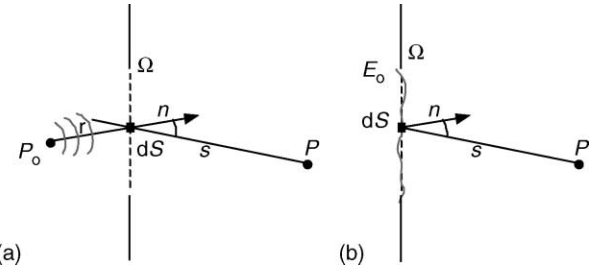


Figure 13 Geometry for evaluating (a) the Fresnel–Kirchhoff diffraction formula and (b) the Huygens' integral.

In the case of the Huygens integral an alternative Green's function is chosen, namely $G(P) = \exp(-jks)/s - \exp(-jkr)/r$ instead of $G(P) = \exp(-jks)/s$ used in the Fresnel–Kirchhoff diffraction formula. This then results in an obliquity factor $\cos(n,s)$ instead of $\cos(n,s) - \cos(n,r)$. If the distances of the points P_o and P are large compared to the dimensions of the aperture, the factor $[\cos(n,r) - \cos(n,s)]$ in eqn [31] will not vary appreciably over the aperture and may be replaced by 1. In this case, the optical beams are close to the optical axis and this is considered as the paraxial diffraction theory. In the phase factors, we can expand $r(x,y,z)$ and $s(x,y,z)$ as power series. If the quadratic and higher-order terms may be neglected in the series expansion ($r,s \gg \lambda$), one speaks of Fraunhofer diffraction, when the previous assumption is not fulfilled ($r,s \geq \lambda$), the quadratic terms cannot be neglected and one speaks of Fresnel diffraction. The regions where the Fresnel and Fraunhofer diffraction integral may be applied are called the near-field (NF) and far-field (FF) diffraction regions respectively. The Rayleigh range z_R marks the boundary between the NF and FF regions for the beam emerging from the aperture (Figure 14).

As an example, we discuss the diffraction of a uniform planewave by a circular aperture of diameter $2a$. The Rayleigh range for this circular aperture is given by

$$z_R = \frac{\pi a^2}{\lambda} \qquad [33]$$

The FF diffraction pattern ($z \gg z_R$) for a uniformly illuminated circular aperture is the well-known Airy disk pattern and is given by

$$E(r,z) \sim \frac{2J_1(2\pi a\theta/\lambda)}{(2\pi a\theta/\lambda)} \qquad [34]$$

where J_1 is the first-order Bessel function. This pattern has a single dominant central lobe surrounded by a series of increasingly weaker circular rings. The first zero of this pattern occurs at a half angle θ_1 or a radius r_1 in the FF, given by

$$\theta_1 = \frac{r_1}{z} \approx \frac{1.22\lambda}{2a} \qquad [35]$$

The NF or Fresnel diffraction patterns for a uniformly illuminated circular aperture ($z < z_R$) consist of a series of circular rings modulating a constant-amplitude background. When moving even closer to the aperture, the beam profile becomes more and more square and the frequency of the Fresnel ripples increases.

We now know that light is diffracted at the apertures of optical elements (lenses) and that this diffraction can affect its performance. Ideally, the focus of a planewave should be infinitely small, being the image of a point source located at infinity. But even in the ideal case, when no geometrical aberrations are introduced, diffraction will limit the lens performance and the focus of such diffraction-limited lenses will have a finite extension.

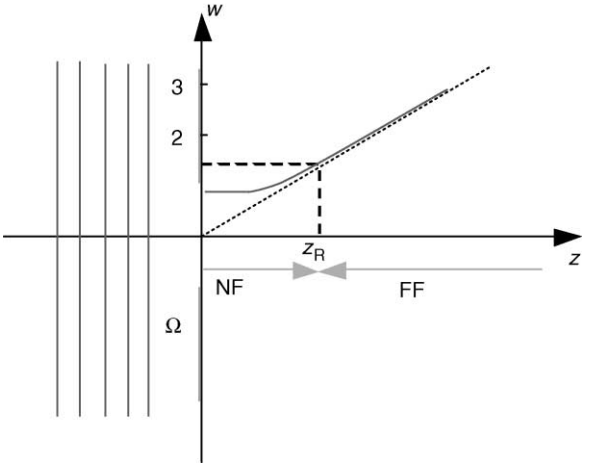


Figure 14 The Rayleigh range z_R marks the boundary between the NF and FF regions for a beam emerging from the aperture Ω.

Quality Criteria for Lens Performance

In this section we discuss some of the most important figures of merit for diffraction-limited lens performance. This survey raises the problems related to this issue. A more detailed overview can be found in the literature.

Diffraction in a Lens System: The Point Spread Function (PSF)

Diffraction at the lens aperture D causes a blur of the focus. The light distribution in the focal plane of an aberration free lens essentially arises from Fraunhofer diffraction at the aperture of the lens. From eqn [34] we know that the diffraction image for an aberration-free lens with a circular aperture D is the scaled Airy disk, given by

$$E(r,z) \sim \frac{\pi J_1(rD/\lambda f)}{(rD/\lambda f)} \qquad [36]$$

where λ denotes the wavelength of the illuminating light beam and f the focal length.

The normalized intensity of this pattern is called the diffraction-limited (PSF) of the lens:

$$\mathrm{PSF}(r,z) = |E(r,z)|^2 \qquad [37]$$

Although irradiance replaces nowadays the word intensity in optics, we will still use intensity to nominate the flow of energy per unit area per unit time. We can define the irradiance as $I \equiv \frac{1}{2}\varepsilon_0 c\langle \mathrm{E}^2\rangle$, with ε_0 and c the electric permittivity in vacuum and the speed of light, respectively. The radius of this pattern is known as the Rayleigh diffraction limit or Rayleigh resolution and is compared to the measured spot size in the image plane.

To include the effects of aberrations in the diffraction computation, we have to include both the phase profile and the amplitude of the real wavefront in the pupil of the lens. The pupil function $T(x, y)$ may be written as:

$$T(x,y) = E(x,y)\mathrm{e}^{\mathrm{i}k\psi(x,y)} \qquad [38]$$

where $E(x, y)$ represents the amplitude distribution over the exit pupil and $\psi(x, y)$ is the wavefront deformation on this pupil. Starting from the Huygens–Fresnel approach, the diffraction integral can be rewritten in the form of a Fourier transform:

$$E(r,z) \sim FT(\mathrm{e}^{\mathrm{i}k\psi(x,y)}) \qquad [39]$$

The PSF at the focal position can then be found by calculating the Fourier transform of the wavefront error ψ or the pupil function $T(x, y)$ of the lens and taking the complex square of the latter. Figure 15 shows the shape of the PSF of an ideal or diffraction-limited lens. When the lens aperture D decreases, the extension of the PSF increases proportionally, which means that it becomes easier to achieve diffraction-limited performance. Since the extension of the diffraction-limited PSF becomes larger, the constraints on the shape of the phase profile become less stringent. This has important consequences on the scaling behavior of lenses. In the extreme case, the microlens diameter becomes so small, that diffraction-limited imaging is possible without any specific phase profile in the lens pupil. This is also one of the reasons why small water droplets often form relatively good lenslets. For systems including micro-optic lenses, we therefore will have to find better criteria than the PSF of a diffraction-limited lens to evaluate or compare these optical components.

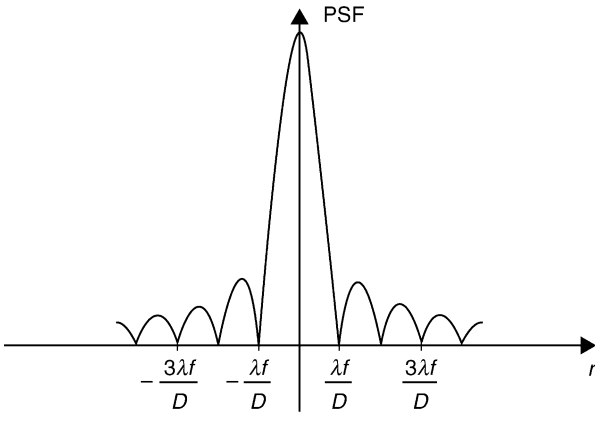


Figure 15 The 1D point spread function of an ideal lens.

The Rayleigh and Maréchal Criterion

The widely used Rayleigh criterion states that a system is diffraction-limited, as long as the maximum wavefront deviation $|\psi_{\max}(x)|$ is less than one quarter of a wavelength (Figure 16). It was shown by Rayleigh that for a system which suffers primarily from SA such that the wavefront deviation is less than a quarter wavelength, the intensity at the Gaussian focus is diminished by less than 20%; a loss of light that can usually be tolerated. Also in the presence of other commonly occurring Seidel aberrations is the quality of the image not seriously affected when the deviation is less than $\lambda/4$. This criterion, which determines the amount of aberration that can be tolerated in an image-formation system is, of course, only a rough guideline since the light distribution in the image does not only depend on the maximum deformation but also on the shape of the wavefront, and even more on the type of aberration. Moreover, the loss of light that can be tolerated depends on the particular use to which the optical component is put and more stringent tolerances have to be imposed for certain applications.

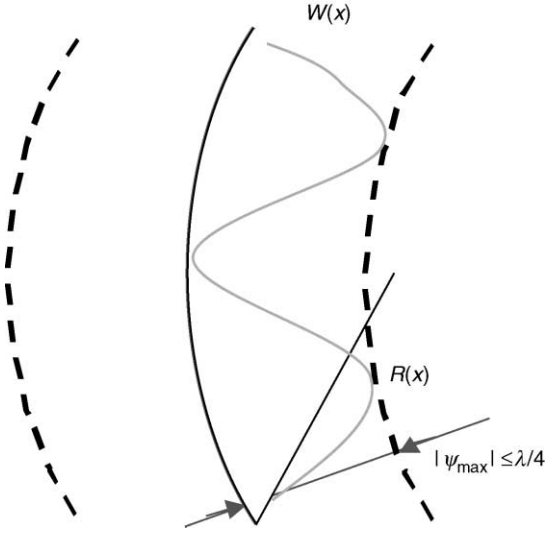


Figure 16 Illustration of the Rayleigh criterion.

Low-order aberrations merely cause a shift of light from the main lobe of the PSF to the side lobes. Higher-order aberrations additionally cause a broadening of the PSF and the correlation between the image quality and the maximum value of the wavefront aberration is usually lost. In the latter case, it seems more appropriate to formulate tolerance criteria restricting the average amount of aberration. Criteria of this type were considered by Maréchal, who used the relation that exists between the intensity at the center of the reference sphere and the RMS deviation of the wavefront from a spherical form.

This Maréchal criterion uses the following RMS wavefront aberration ψ_{RMS} to define the lens quality:

$$\psi_{\mathrm{RMS}} = \sqrt{\int |\psi(x)|^2 \, \mathrm{d}x - \left[\int |\psi(x)| \mathrm{d}x\right]^2} \leq \frac{\lambda}{14} \qquad [40]$$

This criterion is sensitive to statistical noise on the measured phase profile, while a criterion such as the Rayleigh criterion is little affected by it.

Strehl Ratio

For some applications not the imaging quality but the maximum deposited energy is the most essential feature. In such a situation, the Strehl ratio S is a suitable figure of merit. The Strehl ratio is defined as the normalized peak intensity of the PSF of the lens:

$$S = \frac{I_{\mathrm{real}}(0,0)}{I_{\mathrm{ideal}}(0,0)} = \left| \int\int \mathrm{e}^{\mathrm{i}k\psi(x,y)} \mathrm{d}x \, \mathrm{d}y \right|^2 \qquad [41]$$

$I_{\mathrm{real}}(0,0)$ and $I_{\mathrm{ideal}}(0,0)$ denote the intensities at the center of the real point image and the ideal PSF without aberrations, respectively. Even for small aberrations, which do not affect the extension of the PSF, the light intensity in the center peak can drop significantly (Figure 17). A Strehl ratio $S \geq 0.8$ is generally considered to correspond to diffraction-limited performance.

By a Taylor expansion of eqn [40], the relation between the Strehl ratio S and the Maréchal criterion can be written as:

$$S = 1 - \left(\frac{2\pi}{\lambda} \psi_{\mathrm{RMS}}\right)^2 \qquad [42]$$

If only low-order aberrations are taken into account, Rayleigh's $\lambda/4$ criterion also results in a Strehl ratio of $S = 0.8$. For large aberrations ($S \leq 0.3$), however, the behavior of the Strehl ratio becomes uncorrelated to the image quality due to interference effects.

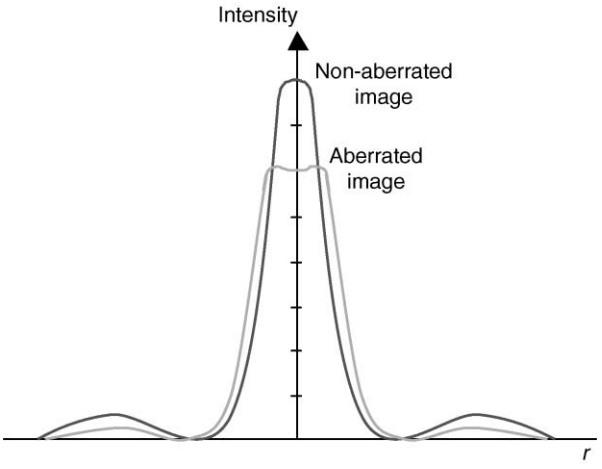


Figure 17 Strehl ratio criterion.

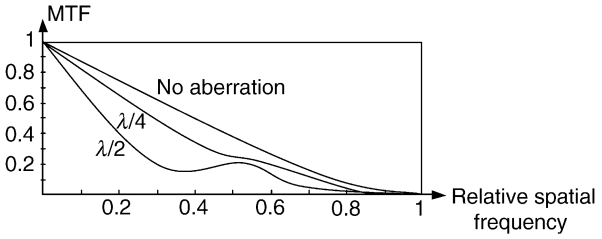


Figure 18 Modulation transfer function for a perfect lens and for small amounts of spherical aberration ($\lambda/2$ and $\lambda/4$).

Transfer Functions

Transfer functions are most commonly used for evaluating the image quality in lens design. As discussed above, the PSF is the Fourier transform of the wavefront error ψ. For simplicity we can assume that the modulus of the pupil function is 1. The PSF then follows from the Fourier transform:

$$PSF(\mu, \nu) = |\mathrm{FT}(\mathrm{e}^{\mathrm{i}k\psi(x,y)})|^2 \qquad [43]$$

The optical transfer function (OTF) is calculated from the PSF data by means of an inverse Fourier transformation:

$$\mathrm{OTF}(\vartheta, \rho) = \mathrm{FT}^{-1}[\mathrm{PSF}(\mu, \nu)] \qquad [44]$$

The OTF is a spatial frequency-dependent complex quantity whose modulus is the modulation transfer function (MTF) and whose phase is, of course, the phase transfer function (PTF). The former is a measure for the reduction in contrast from the object to the image over the whole spectrum. Figure 18 shows the normalized MTF for both a perfect lens and a lens subjected to some small amounts of spherical aberration. The PTF is, in most cases, not so conclusive concerning the optical behavior.

Therefore the wave aberrations, the PSF, the MTF, and the Strehl ratio are the characteristics that are usually provided as an output of lens design programs.

Interference

Many of the processes for the manufacturing of microlenses are not fully understood nor perfectly controlled. It occurs that one generates a microlens array and then measures its performance, hoping that by accurately repeating the processing conditions, a consistent result can be achieved. To achieve optimum conditions, it is therefore crucial to measure the geometrical and optical parameters of the microlenses. Most of the microlens measurement instruments are based on two-beam interference, the phenomena where light from a source is divided into two beams which are later on superposed again. The intensity in the region of superposition is found to vary from point to point between maxima which exceed the sum of the beam intensities and minima which may be zero.

Two-Beam Interference

We consider the interference of two waves, a reference and an object wave, represented by $\mathbf{E}_{\text{ref}}$ and $\mathbf{E}_{\text{obj}}$ respectively, where we take into account the vector properties of the electric fields. In the case of interference, both waves typically originate from a single source and recombine after traveling along different paths. For the moment we will consider only monochromatic linearly polarized waves of the form:

$$\mathbf{E}_{\text{ref}} = \mathbf{E}_{0\text{ref}}\, e^{i\Phi_{\text{ref}}(r) - i\omega t}, \quad \mathbf{E}_{\text{obj}} = \mathbf{E}_{0\text{obj}}\, e^{i\Phi_{\text{obj}}(r) - i\omega t} \qquad [45]$$

The total electric field intensity $\mathbf{E}$ at the output of the interferometer, according to the principle of superposition, is given by:

$$\mathbf{E} = \mathbf{E}_{\text{ref}} + \mathbf{E}_{\text{obj}} \qquad [46]$$

The total intensity at the output, being the time average of the square of the wave amplitude, can be written as

$$I = I_{\text{ref}} + I_{\text{obj}} + I_{\text{ref/obj}} \qquad [47]$$

with $I_{\text{ref}} \sim \langle \mathbf{E}_{\text{ref}}^2 \rangle; I_{\text{obj}} \sim \langle \mathbf{E}_{\text{obj}}^2 \rangle$ and $I_{\text{ref/obj}} \sim 2\langle \mathbf{E}_{\text{ref}} \cdot \mathbf{E}_{\text{obj}} \rangle$. Notice that if both $\mathbf{E}$-fields are perpendicular to one another, the interference term $I_{\text{ref/obj}}$ vanishes, while if we assume that both fields are parallel, we can restrict our discussion to scalar theory. Under the latter condition we can write the total intensity as:

$$I = I_{\text{ref}} + I_{\text{obj}} + 2\sqrt{I_{\text{ref}} I_{\text{obj}}} \cos(\Phi_{\text{obj}} - \Phi_{\text{ref}}) \qquad [48]$$

where the presence of the third term is indicative of the wave nature of light which can produce enhancement or diminution of the intensity through interference.

We can rewrite eqn [48] in Michelson's manner of writing:

$$I = I_0(1 + V \cos(\Phi_{\text{obj}} - \Phi_{\text{ref}}))$$
$$\text{where } I_0 = I_{\text{ref}} + I_{\text{obj}}, \quad V = \frac{2\sqrt{I_{\text{ref}} I_{\text{obj}}}}{I_{\text{ref}} + I_{\text{obj}}} \qquad [49]$$

with V the visibility or fringe contrast indicating the degree of modulation.

We can conclude that the measured intensity of a two-beam interferogram depends on the mean intensity I_0, the visibility V, and the phase difference $\Phi_{\text{obj}} - \Phi_{\text{ref}}$. There are a number of techniques to measure the three unknowns: the mean intensity, the visibility, and the interesting spatial phase variation. In the optical instrumentation, further discussed in this article, phase shifting interferometry (PSI) is used which requires a measurement of the intensity in each point of the aperture for at least three different reference phase values, in order to obtain three equations for the three unknowns (I_0, V, and $\Phi_{\text{obj}} - \Phi_{\text{ref}}$). In practice, the reference phase values are obtained by a piezodriven transducer (PZT) mirror and should be equally spaced over one phase cycle of 2π. It has been shown that the five-phase algorithm is the optimal choice as the calculation still remains simple and the second-order nonlinear response of the detector is eliminated. A charge coupled device (CCD) sensor then records the intensity of the resulting interference pattern for different relative phase shifts and converts the intensity data to wavefront phase data through integration.

Microlens Testing

In microlens (lens with a diameter lower than several millimeters) testing, it is essential to measure in a first step the shape or the surface profile of the microlenses or in cases where gradient-index lenses are concerned, their refractive index distribution. In addition, the paraxial parameters, such as the focal length, and the more complex properties like the wave aberrations and surface quality have to be characterized. As discussed in the introduction, massive parallelism and miniaturization ask for the fabrication of microlenses and microlens arrays. The most critical features of these lenslet arrays are the uniformity of the microlenses and their mutual positioning accuracy. As a variation of these properties deteriorates the optical

performance in most applications, dedicated test instrumentation is needed.

In the following, we discuss different measurement instruments designed to characterize all the geometrical and optical parameters of microlenses. First, we introduce an optical noncontact profiler which can be used to determine the diameter and the sag (the maximum height of the spherical lenscap) of the surface profile microlenses. In addition, it allows us to quantify the surface roughness on the top of the microlenses. With a transmission Mach–Zehnder (MZ) interferometer the wave aberrations and the focal length of the microlenses can be measured, while a Twyman–Green (TG) interferometer can be used for the measurement of the surface deviations of a microlens from an ideal spherical shape and for the measurement of the radius of curvature of the surface.

Surface Profile Measurements

Optical noncontact profiler

Principle. The surface profile of a microlens can be measured with a Wyko NT2000 noncontact optical profiler (Figure 19a). This surface profiler system uses two approaches to measure a wide range of surface heights. The PSI mode allows measuring fairly smooth and continuous surfaces (0.1 nm $<$ heights $<$ 160 nm) while the vertical scanning interferometry (VSI) mode can measure rough surfaces and heights ranging between 160 nm and 2 mm. We only discuss here the vertical scanning mode because the principle on which PSI is based has been explained above. As shown in Figure 19b, this noncontact profiler is based on a Mirau interference microscope where a white light beam passes through a beamsplitter which reflects half of the incident beam to a reference surface and transmits the other half onto the sample. The light reflected from the sample and from the reference surface then recombine at the beamsplitter to form interference fringes. The system measures the degree of the fringe modulation or the fringe contrast. Because white light has a short coherence length, interference fringes are present only over a very shallow depth for each focus position. The coherence length can be described here as the optical path distance (OPD) between two wavefronts of an interferometric system for which the phase remains correlated. Or in other words it is a measure of the range of heights over which the instrument will be able to obtain measurable interference fringes. Fringe contrast at a single sample point reaches a peak as the sample is translated through focus. During a measurement, the reference arm containing the interferometric objective moves vertically to scan the surface at varying heights. A linearized PZT precisely controls this motion. As the system scans downward, an interference signal for each point on the surface is recorded. Finally, a series of advanced computer algorithms is used to demodulate the envelope of the fringe signal and to extract the surface information.

The vertical resolution is approximately 3 nm RMS for a single measurement. The lateral resolution is a function of the total magnification, of the cursor position accuracy, and of the detector array size.

Measuring the geometrical lens dimensions. This profiler provides us with information on geometrical dimensions of microlenses. For each microlens we can derive from the contour plot (Figure 20) the sag and the diameter of the microlenses along the x and the y direction. As an example, we show in Figure 21, the

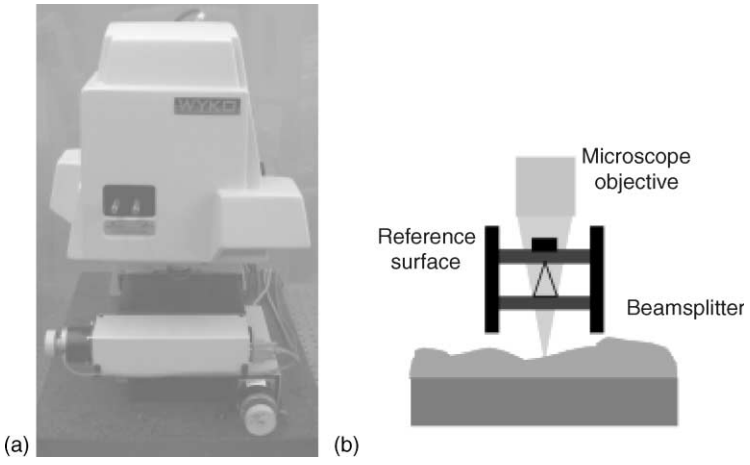


Figure 19 (a) Vertical scanning noncontact profiler (WYKO NT2000); (b) Working principle.

profiles of the cross-section of a 200 μm fused silica microlens ($NA = 0.05$; $D = 200$ μm) along the x and y directions.

The surface roughness on top of the microlens can also be derived from measurements with the noncontact profiler. As an example, we show the fringes on the vertex area (65×61 μm^2) of a 200 μm diameter lens as well as the average and RMS roughness on the vertex of the lens (Figure 22).

Measurement of the Optical Lens Performance

To investigate the optical properties of discrete microlenses one of the most accurate and reliable techniques is interferometry. It is well known that microlenses can be tested in transmitted light using instruments with a single pass geometry (e.g., MZ interferometer) as well as in reflected light interferometers using a double pass configuration (e.g., TG interferometer). The interferometers proposed in this section have been especially designed and constructed for microlens testing.

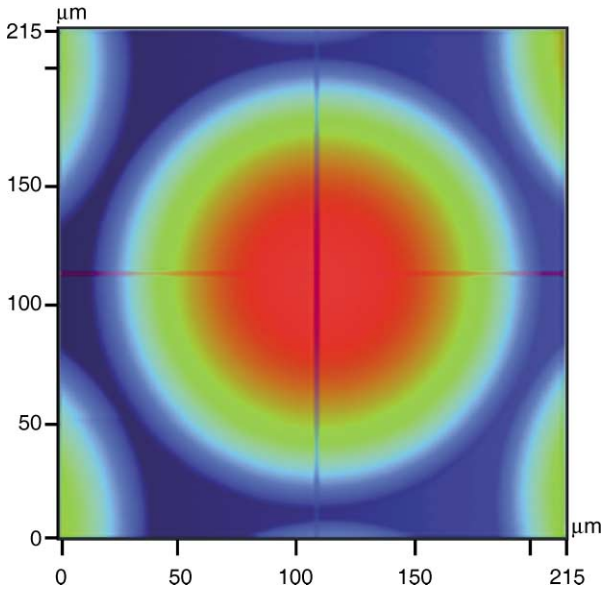


Figure 20 Contour plot of a 200 μm fused silica microlens as obtained with the noncontact profiler.

Mach–Zehnder (MZ) interferometer for transmitted light measurements

Principle. A MZ interferometer, constructed at the Erlangen–Nürnberg University, is used in this work for the measurement of microlenses in transmitted light. The interferometer is installed in a commercial Carl Zeiss Jena microscope, while fiber-optics is used for beam delivery. The scheme is shown in Figure 23. A polarizing beamsplitter (PBS) cube splits the beam of a HeNe laser while a halfwave plate at the entrance enables the balancing of the two interfering beams. The fibers are polarization preserving in order to keep the polarization in the two beams of the interferometer. The planewave of the object arm is either focused by an illumination microscope objective so that a spherical wave impinges onto the microlenses (this is the so-called spherical wave illumination) or the first microscope objective is removed so that a planewave illuminates the object (this is the so-called planewave illumination). By using a telescope consisting of the imaging microscope objective and an achromatic doublet, the expanded planewave then enters a second beamsplitter and is superimposed on a reference planewave generated by the second fiber and an expansion lens. This second beamsplitter is a nonpolarizing one. To force the two beams to interfere, the plane of polarization of the reference beam must be rotated by 90° which is achieved by twisting the fiber end accordingly. This solution has been chosen to avoid a polarizer in front of the CCD camera as it has been shown that the latter would decrease the optical quality. In addition one of the mirrors in the object

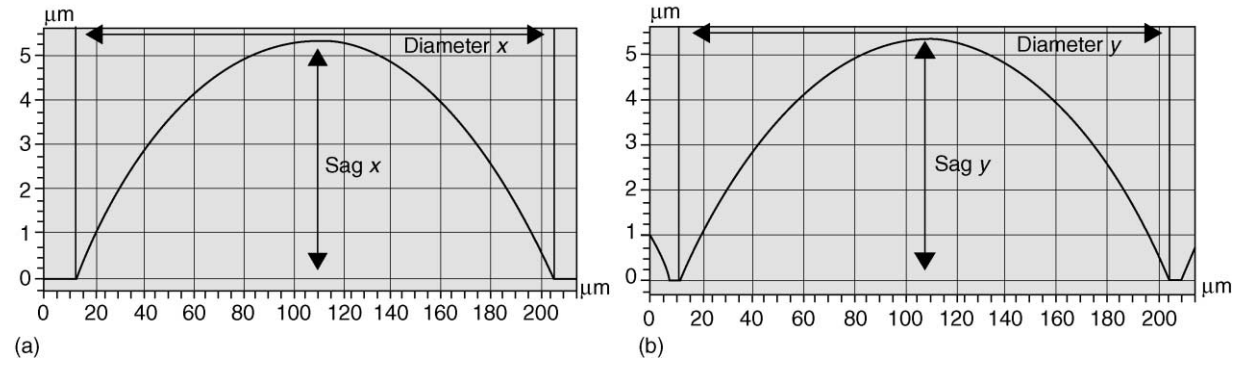
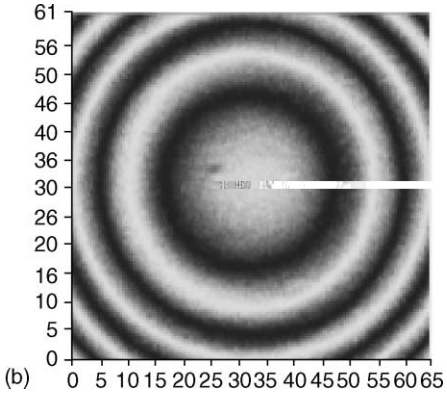


Figure 21 (a) 2D lens profile of the microlens shown in Figure 20 along the x direction and (b) along the y direction.

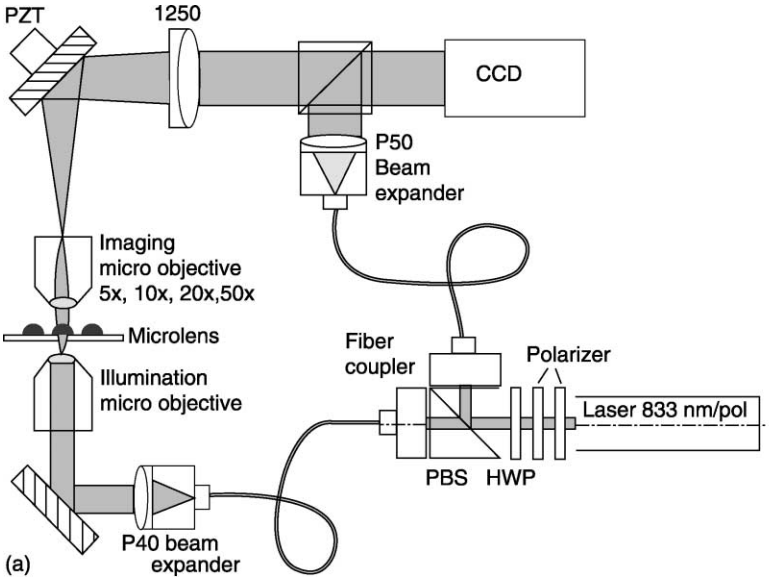
Roughness values	(nm)
Average roughness R_1	1.55
RMS roughness R_q	1.94

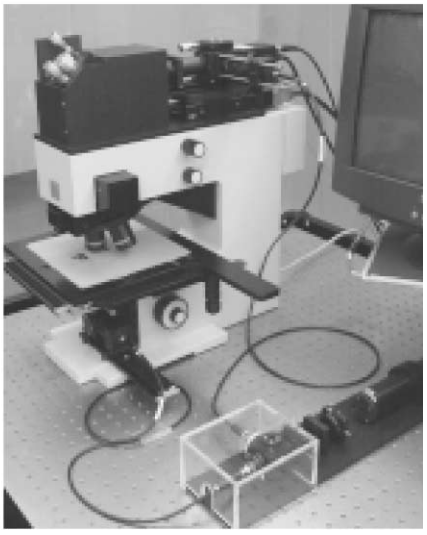
(a)



(b)

Figure 22 (a) Different roughness values on the vertex area of a 200 μm lens; (b) Fringes in this vertex area (65 × 61 μm^2).





(b)

Figure 23 (a) Scheme and (b) picture of the Mach–Zehnder interferometer for measurements in transmitted light.

arm of the MZ is mounted on a PZT driver to enable phase shifting interferometry.

Focal length and array uniformity through plane-wave illumination. The MZ interferometer with planar wavefront illumination offers for slow lenses (NA < 0.1) the opportunity to measure also the total phase lag introduced by the lenses, as shown in Figure 24. This makes it possible to determine the focal length and hence the NA of the microlenses.

For certain applications of arrays of microlenses it is also important to know the uniformity of the lens array or at least of a part of the lens array. There are several different test methods (e.g., Smartt test, shearing test) which can provide a measure of the uniformity. One can even obtain the total phase retardation for a subsection of the microlens array (see Figure 25b).

Measuring the aberrations with spherical wave illumination. Wave aberrations of refractive microlenses can be determined with a spherical wavefront illumination (Figure 26). The measurement of the wave aberrations is carried out by using a PSI algorithm. This requires a measurement of the intensity in each point of the aperture for five different reference phase values. The result of a PSI evaluation is the phase distribution $\Phi(x, y)$ in the exit pupil:

$$\Phi(x, y) = \frac{2\pi}{\lambda} \psi(x, y) \qquad [50]$$

where $\psi(x, y)$ are the wave aberrations of the lens under test.

From these data, the software calculates the unwrapped phase distribution and by subtracting the tilt and defocus, the deviation from an ideal wavefront (phase/lambda) can be found. Fitting this

deviation to a fourth degree Zernike polynomial then gives the aberrations (phase/lambda). Besides the aberrations we also derive from the measured phase distribution other optical lens characteristics such as the modulation transfer function and the PSF.

Finally, we emphasise that the measurements should preferably be done in a cleanroom environment as the smallest dust particle on the lens surface decreases the optical performance.

Twyman–Green interferometer

Principle. For most applications microlenses have to feature a spherical or nearly spherical shape. The measurement of the deviations of the microlens surface from an ideal sphere is therefore very important. The basic principle comes down to a planewave which is focused by a high NA microscope objective in such a way that the focus coincides with the center of curvature of the microsphere under test. The light rays are reflected, and behind the microscope objective a nearly planewave carrying the surface deviations and misalignment errors, results. The vertex of the test surface is then imaged onto a CCD camera and the nearly plane object wave interferes with the plane reference wave. The evaluation here is also done with phase shifting interferometry by axially shifting the reference mirror with a piezo transducer.

The use of laser illumination allows us to obtain high contrast fringes. However, the high spatial and temporal coherence of the laser light also has important drawbacks, such as speckle noise and spurious fringes across the field. It is convenient to divide coherence into two classifications, temporal and spatial. The former relates directly to the finite bandwidth of the source, the latter to its finite extent in space. Therefore, polarization beamsplitter optics is commonly used to reduce the contribution of spurious reflections. The polarizing beamsplitter group additionally allows balancing the intensity between the object and the reference arm of the interferometer. Even so, the remaining scatter leads to poor and rough interference fringes. To further increase the fringe quality, the spatial coherence has to be reduced while the temporal coherence of the laser is preserved. This can be done by introducing a rotating scatterer in the light beam which gives a smoothing effect of the interference fringes.

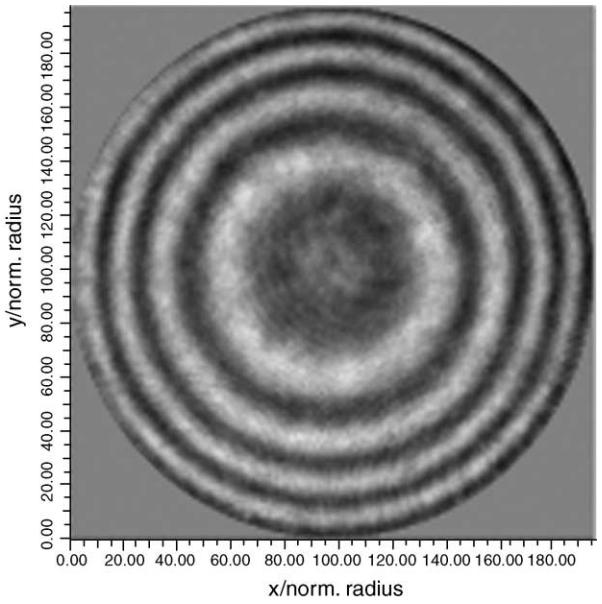


Figure 24 Interferogram of a 200 μm lens with an NA of 0.05 under planewave illumination.

Measurement of the deviations from sphericity. To determine the surface deviation of a microlens from a perfect sphere we make use of a TG

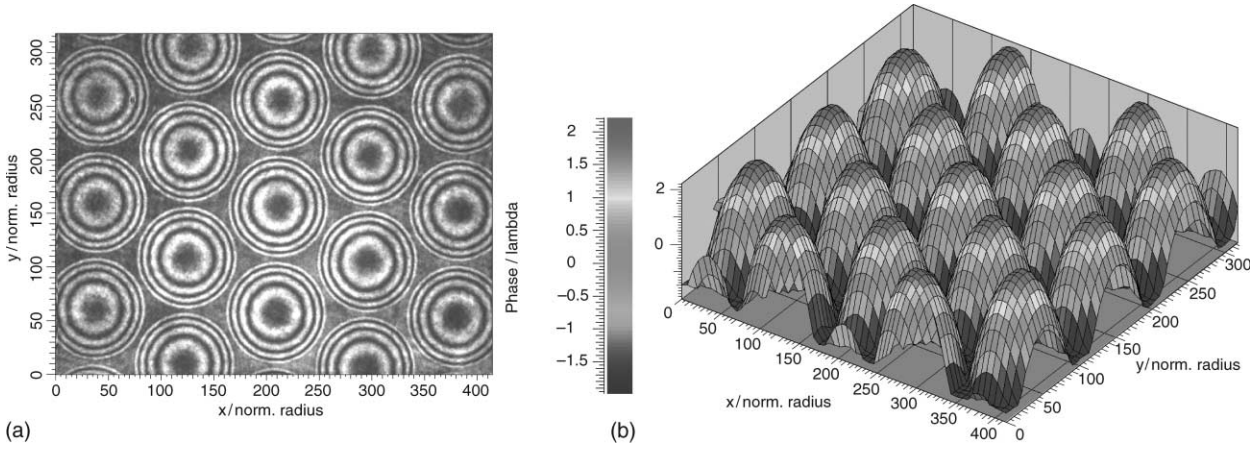


Figure 25 (a) Interferogram of an array of 200 μm diameter microlenses with an average NA of 0.05 and illuminated with a plane wavefront; (b) 3D representation of the phase lag (RMS 1.26λ, *PV* 4.37λ) for the same microlens array.

interferometer, constructed by the Erlangen-Nürnberg University (Figure 27). In this interferometer, the light of a HeNe laser is imaged onto a rotating scatterer by a microscope objective and the scattered light is collimated by an achromat. The collimated laser beam then passes a field stop which is sharply imaged by a 4f system onto the lens under test. The image of the field stop can be focused onto the surface under test by moving the field stop along the optical axis. Once the image of the field stop is sharply imaged onto the CCD array, one can translate the reference mirror until the reference image of this stop in its turn becomes sharp onto the CCD camera. Thus, two conditions have to be met to achieve high contrast fringes: the surface under test and the image of the field stop of both arms of the interferometer have to be sharp on the detector. The half wave plate in front of the PBS adjusts the splitting ratio at the PBS while the polarizer after the beamsplitter is necessary to force the two orthogonally polarized beams to interfere. Finally, the optics at the output of the interferometer gives a sharp image of the intermediate image of the lens surface onto the CCD camera. Analogous to the MZ interferometer, five phase images are grabbed and subjected to different algorithms, leading to the desired information.

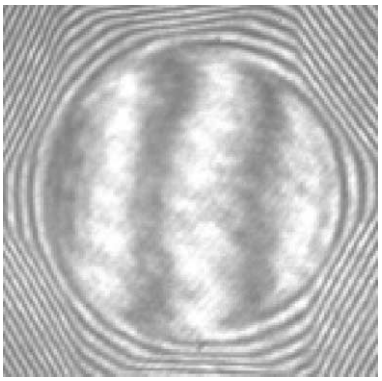


Figure 26 Interferogram of a microlens under spherical wavefront illumination: the resulting planewave is a measure for the wave aberrations.

A condensor objective transforms the radius of curvature of the spherical wavefront such that it matches the radius of curvature of the microlens. Hence, the wavefront impinges everywhere perpendicularly on the lens surface. This means that the sensitivity to surface deviations is independent on the surface position leading to high accuracy measurements. The TG interferometer allows us to test the sphericity of microlenses with diameters ranging

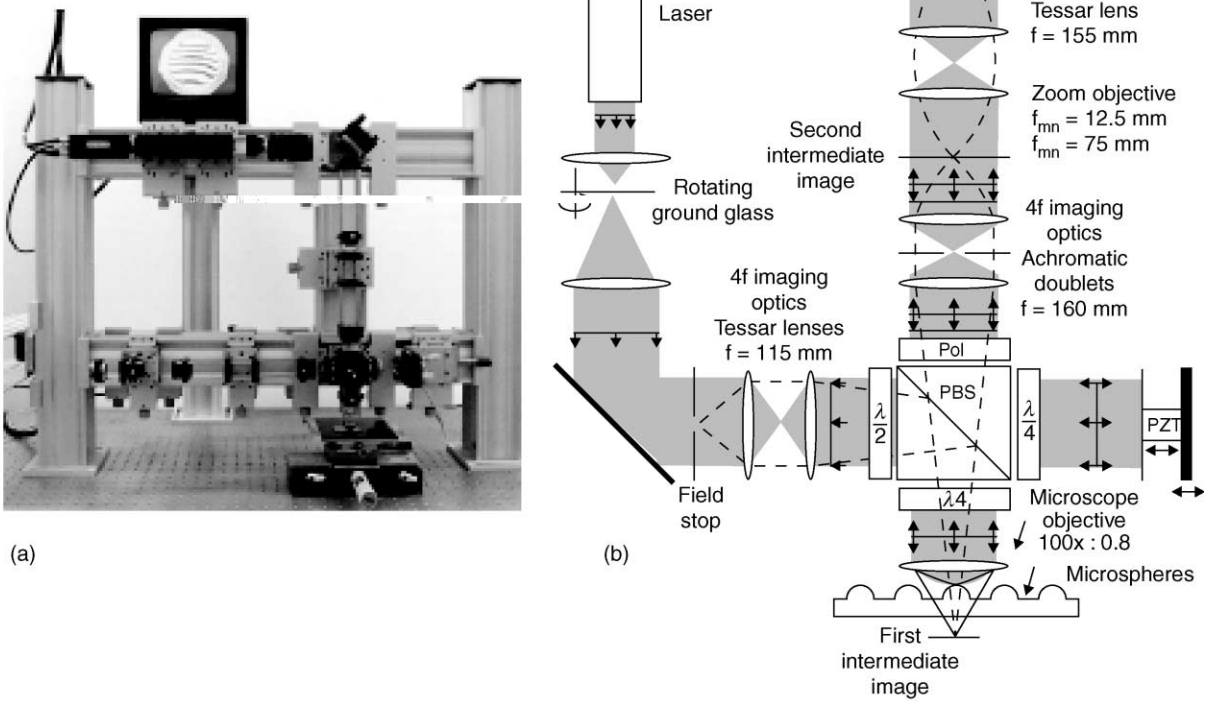


Figure 27 (a) Twyman–Green interferometer setup. (b) Detailed schematic overview of this interferometer.

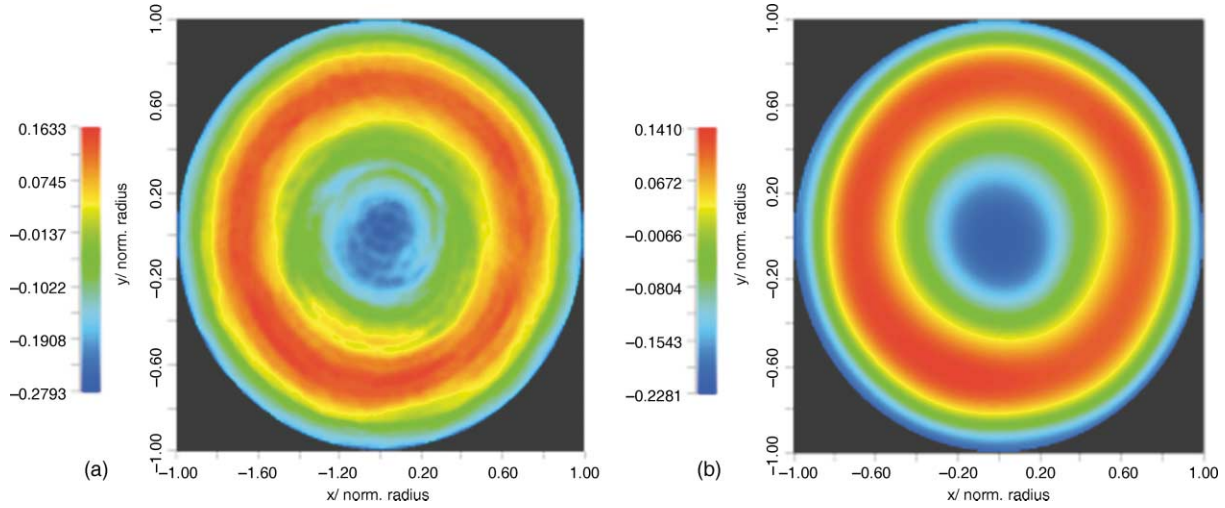


Figure 28 (a) Measured deviation from an ideal sphere (phase/lambda) for a 200 μm lens with an NA of 0.05 as obtained with a Twyman–Green interferometer. (b) Polynomial fitted surface deviation (phase/lambda) of this microlens.

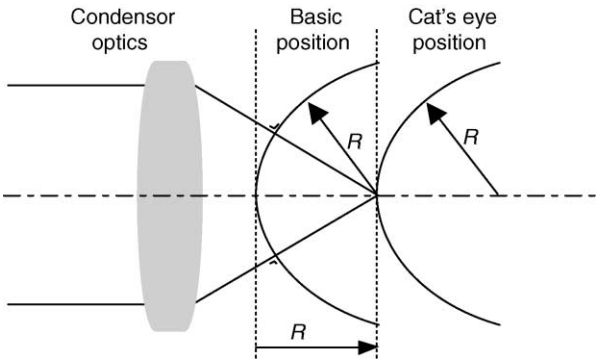


Figure 29 Necessary adjustments of the spherical microlens for radius measurements.

from 50 to 1000 μm and with a maximum radius of curvature of 1960 μm.

An example of a 200 μm fused silica microlens, with an NA of 0.05, is shown in Figure 28. The deviation from an ideal sphere can be derived from the unwrapped phase after subtracting the tilt and defocus. Fitting the deviation to a fourth degree Zernike polynomial then gives the surface deviation.

Measuring the radius of curvature. Using the TG interferometer it is also possible to determine the radius of curvature R of a microlens. To do this, the test object has to be shifted axially from the basic position, where the light rays impinge perpendicularly onto the surface and the focal length of the condensor objective coincides with the center of the sphere under test, to the so-called cat's eye position, where the focus of the impinging spherical wave is at the vertex of the sphere (Figure 29). For both positions straight, parallel, and equidistant interference fringes are obtained, only perturbed by aberrations due to the lens surface and the condensor objective. This forms the criterion for the adjustment of the two positions. The difference between these two positions then gives the radius of the microsphere. It should be mentioned that the cat's eye position causes an inversion of the object wavefront with respect to the reference wavefront, which results in a vanishing contrast with the partially coherent illumination. The latter can only be avoided by providing the necessary coherence. This implies that the rotating scatterer will have to be removed.

Summary of the Interferometric Instrumentation

Figure 30 summarizes the various instruments we discussed in this section: a noncontact optical profiler, a MZ and a TG interferometer. For each of these instruments, we highlight the measurands, the assumptions that are made, and the resulting optical lens characteristics. From this figure, it is clear that one can obtain a measure for the focal length and hence the numerical aperture and the focal number of a refractive microlens, with all three instruments. However, only the focal length obtained with the MZ interferometer is a direct measurement. The focal length obtained with the optical profiler is derived from the sag of the lens under the assumption of an ideal sphere while the radius of curvature measured with the TG interferometer allows to calculate the focal length as well. Nevertheless, the three values make an interesting comparison possible.

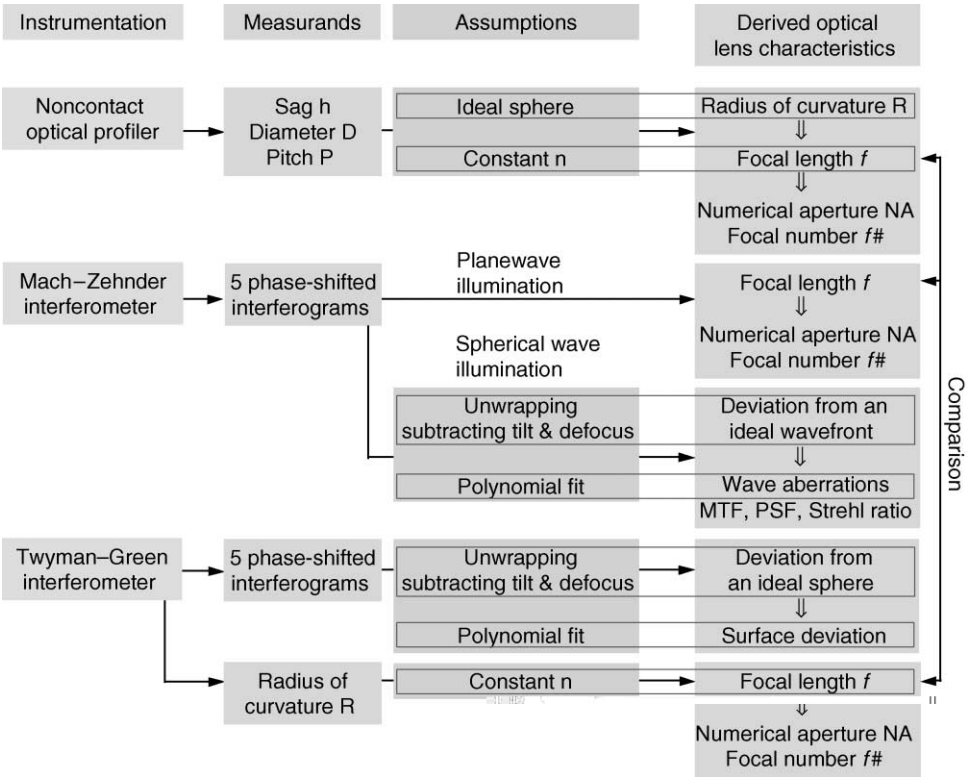


Figure 30 Overview of the instruments that we will use to quantitatively characterize the refractive microlenses: the measurands and the resulting optical lens characteristics.

Conclusions

In this article we introduced geometrical optics to discuss the imaging properties and the aberrations of refractive lenses, while wave optics was used to explain the effects of diffraction on the imaging properties of lenses. Next we put forward some commonly used quality criteria to assess lens performance and we pointed out the unique properties of microlenses. Finally, interference was briefly introduced to clarify the working principle of various instruments to quantitatively characterize refractive microlenses and microlens arrays: a noncontact profiler, a transmission MZ interferometer and a TG interferometer.

We can conclude that micro-optics is a key technology with many applications in various areas of tomorrow's information society. It enables new functionalities and applications previously unachievable with conventional optics and improves product performance, stability and robustness while reducing cost, volume, and weight. While in conventional optics discrete lenses are mounted to form optical systems using mechanical alignment equipment, for microlenses new integration approaches are available. For most of the applications, the integration of micro-optical components into fully functional systems is very important.

Thus micro-optics is a generic technology that allows the manipulation of light with micron and submicron scale structures and components. Micro-optics is therefore the technology to help provide the necessary tools to interface the macroscopic world we live in with the microscopic world of opto- and nano-electronic data processing. Micro-optic processing and fabrication technologies can be adapted to a wide variety of specialty optical materials such as plastics, polymers, semiconductors, sol-gels, and doped glasses. The application of these modified materials permits the widening of the field of photonic functionalities to virtually any region of the optical spectrum and beyond, such that a wealth of novel optical and photonic applications can be realized. Micro-optic processing technologies can be made compatible with scalable industrial mass-manufacturing and replication techniques.

Acknowledgment

This research has been made possible through the financial support of the Fund for Scientific Research-Flanders (FWO-Vlaanderen), the IAP Photon

Network, the Bilateral Scientific and Technological Cooperation Flanders-Poland, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the IWT GBOU project 'Generic Technologies for Plastic Photonics', the Research Council (OZR) of the Vrije Universiteit Brussel and the EC Network of Excellence NEMO.

List of Units and Nomenclature

[AB]	optical path length
c	speed of light
d	lens thickness
D	lens aperture diameter
$\mathbf{E}$	electricfield
f	focal length
f_{BFL}	back focal length
f_{FFL}	front focal length
$f\#$	focal number
F_{i}	image focal point
F_o	object focal point
h_1	position of the first principal point relative to the vertex V1
h_2	position of the second principal point relative to the vertex V2
h'	deviation from the optical axis
H_1	first principal point
H_2	second principal point
I	intensity
k	wave number in vacuum
n	index of refraction
n_{i}	index of refraction of medium i
O	object point
P	arbitrary field point
P	image point
P	lens pitch
$\wp$	spherical surface power
r	aperture of the refracting surface
R	radius of curvature
$R(x, y)$	real wavefront
S	axial object point
S	Strehl ratio
s_0	object distance
s_i	image distance
V	lens vertex
V	visibility or fringe contrast
$W(x, y)$	spherical wavefront in the Gaussian approximation
y_{o}	object height
y_{i}	image height
z_{R}	Rayleigh range
α	angular semi-aperture
θ	symmetry around the optical axis
λ	wavelength of light
σ	standard deviation
φ_1	angle between the incident ray and the normal to the flat surface
φ_2	angle between the refracted ray and the normal to the flat surface
$\Phi_{\mathrm{obj}} - \Phi_{\mathrm{ref}}$	phase difference
$\Phi(x, y)$	phase distribution
$\psi(x, y)$	wavefront aberration
ω	angular frequency
Ω	aperture in a plane opaque screen
BFL	Back focal length
CCD	Charge coupled device
FF	Far-field
FFL	Front focal length
MTF	Modulation transfer function
MZ	Mach–Zehnder
NA	Numerical aperture
NF	Near-field
OPD	Optical path difference
OTF	Optical transfer function
PBS	Polarizing beamsplitter
PP	Principal plane
PSF	Point spread function
PSI	Phase shifting interferometry
PTF	Phase transfer function
PZT	Piezo-electric transducer
RMS	Root mean square
SA	Spherical aberration
VSI	Vertical scanning interferometry

See also

Diffraction: Fraunhofer Diffraction. **Diffractive Systems:** Aberration Correction with Diffractive Elements. **Geometrical Optics:** Aberrations; Lenses and Mirrors. **Interferometry:** Overview; Phase Measurement Interferometry.

Further Reading

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge, UK: Cambridge University Press.

Hecht E (1987) *Optics*. San Francisco, CA: Addison-Wesley Publishing Company.

Herzig HP (1997) *Micro-Optics: Elements, Systems and Applications*. London: Taylor & Francis.

Jenkins FA and White HE (1981) *Fundamentals of Optics*, 4th edn. New York: McGraw-Hill.

Malacara D and Malacara Z (1994) *Handbook of Lens Design*. New York: Marcel Dekker.

Malacara D, Servin M and Malacara Z (1998) *Interferogram Analysis for Optical Testing*. New York: Marcel Dekker.

Mouroulis P and Macdonald J (1997) *Geometrical Optics and Optical Design*. New York: Oxford University Press.
Pedrotti FL and Pedrotti LS (1993) *Introduction to Optics*, 2nd edn. Englewood Cliffs: Prentice-Hall International.
Shannon RR (1997) *The Art and Science of Optical Design*. Cambridge, UK: Cambridge University Press.
Siegman AE (1986) *Lasers*. Sausalito: University Science Books.
Sinzinger S and Jahns J (1999) *Microoptics*. Weinheim, Germany: Wiley-VCH GmbH.
Smith WJ (1997) *Practical Optical System Layout and Use of Stock Lenses*. New York: McGraw-Hill.
Smith W (2002) Engineering an optical system. *SPIE OE Magazine* 7: 49–54.
Young HD (1992) *Physics*, 8th edn. San Francisco, CA: Addison-Wesley Publishing Company.

OPTICAL PARAMETRIC DEVICES

Contents

Overview
Optical Parametric Oscillators (Continuous Wave)
Optical Parametric Oscillators (Pulsed)

Overview

B J Orr, Macquarie University, Sydney, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Many forms of optical parametric device (OPD) play significant roles as three-wave nonlinear-optical sources of tunable coherent light for laser-based science and technology. Any such device will fall into one of three categories, in increasing order of sophistication: optical parametric generator (OPG); optical parametric amplifier (OPA); and optical parametric oscillator (OPO). This article concentrates largely on nanosecond-pulsed OPOs, with particular emphasis on designs that facilitate their wavelength control and their optical power- or energy-efficiency. The two other important temporal manifestations of OPOs (ultra-fast pulsed and continuous-wave) are briefly outlined. OPDs serve as versatile sources of tunable, intense, coherent radiation covering the infrared, visible, or ultraviolet spectral regions. Their most obvious uses are associated with laser wavelength-conversion and/or tunability, resulting in a variety of spectroscopic applications. Four-wave nonlinear-optical parametric contributions, such as those involved in stimulated Raman scattering and highly nonlinear optical fibers, are also briefly considered.

What is an Optical Parametric Device?

An optical parametric device (OPD) is a source of coherent, laser-like radiation that is typically intense and tunable over a wide range of wavelengths. Its mechanism invariably depends on nonlinear optics, usually through a three-wave mixing process mediated by the nonlinear-optical susceptibility $\chi^{(2)}$ in a noncentrosymmetric crystalline medium. Three forms of OPD are illustrated in Figure 1, and also illustrated is a closely related (but distinct) nonlinear-optical device. Coherent light waves are represented by arrows, with their associated optical frequency ω_j and wave vector $\mathbf{k}_j$ (which are defined below). In Figure 1, input and output waves are shown as arrows on the left and right, respectively, with their breadth intended to indicate typical relative intensities.

An optical parametric generator (OPG) is the simplest form of OPD, with a single input wave (pump, P) from a laser generating two coherent output waves (signal, S; idler, I; $\omega_S \geq \omega_I$), as depicted in Figure 1a. The nonlinear-optical process itself is initiated by low-intensity parametric spontaneous emission/noise/fluorescence processes that occur naturally, effectively 'splitting' a pump photon into two new photons. Once a signal and/or idler wave has been generated, it can be coherently amplified by passing it through an optical parametric amplifier (OPA) together with input pump radiation, as depicted in Figure 1b. A further order of sophistication is reached in an optical parametric oscillator (OPO), as depicted in Figure 1c, where the

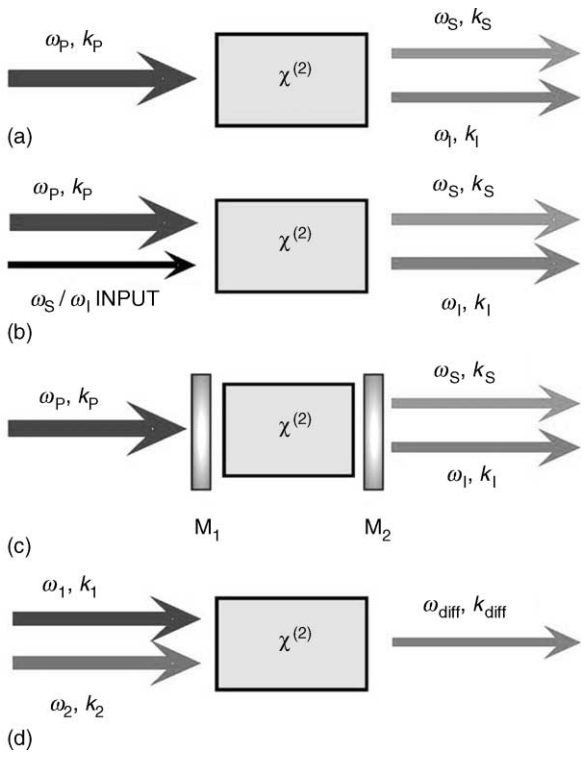


Figure 1 Schematic diagrams of three forms of optical parametric device (OPD): (a) optical parametric generator (OPG); (b) optical parametric amplifier (OPA); (c) optical parametric oscillator (OPO). Note that, by convention, optical frequencies of the signal (S) and idler (I) output waves are defined such that $\omega_S \geq \omega_I$. Also shown is a fourth closely related device: (d) difference-frequency generator (DFG). Nonlinear-optical media are denoted by their susceptibility $\chi^{(2)}$. Arrows are used to represent input and output waves, together with their optical frequencies ω_j and wave vectors $\mathbf{k}_j$. An OPO requires an optical resonator, comprising at least two aligned reflectors (M_1, M_2).

functions of an OPG and an OPA are combined by multipassing the optical waves involved inside a resonant optical cavity, formed by two or more appropriately aligned reflectors (M_1, M_2).

A difference frequency generator (DFG), as depicted in Figure 1d, is not an OPD, although the DFG source term is central to the nonlinear-optical mechanism of an OPD. In a DFG, two input waves (labeled 1 and 2) are combined coherently to generate a third output wave at the difference frequency of the two input waves: there are now two relatively high-power driving waves, rather than one, and photons are effectively 'joined', rather than 'split' as in the OPD case. Nevertheless, the outcome and utility of a DFG can be similar to that of an OPD. For instance, if coherent radiation is required at a particular infrared (IR) wavelength, it can be generated either as the idler wave of an OPD, with frequency $\omega_I = (\omega_P - \omega_S)$ *or* as the output wave of a DFG, with frequency $\omega_3 = |\omega_1 - \omega_2|$.

Many of the desirable attributes of OPDs in general, and OPOs in particular, arise from the fact that any such instrument is a nonlinear-optical device, not a laser. This allows versatile, flexible design features, such as methods of temporal and wavelength control that are not possible with lasers. The latter generally depend on population inversion of an optical gain medium, with associated optical lifetime and saturation limitations. On the other hand, optical parametric gain, oscillation and amplification are more amenable to modular system design because (as is explained below) they depend on nonlinear-optical coefficients and phase-matching conditions.

Nonlinear-Optical Origins of OPDs

Nonlinear optics involves interaction of a number ($\sigma > 2$) of optical waves in a medium with a nonlinear-optical susceptibility tensor $\chi^{(\sigma-1)}$. For inelastic optical processes, these waves (each with an angular frequency $\omega_1, \omega_2, \ldots, \omega_\sigma$) obey two conservation conditions, one for energy (or frequency):

$$\omega_1 + \omega_2 + \ldots + \omega_\sigma = 0 \qquad [1]$$

The other conservation condition (expressed in terms of wave vectors $\mathbf{k}_i$, with $i = 1, 2, \ldots, \sigma$, that have magnitudes $k_i = n_i\omega_i/c = 2\pi n_i/\lambda_i$, where n_i is the refractive index at vacuum wavelength λ_i and c is the speed of light) is for momentum:

$$\mathbf{k}_1 + \mathbf{k}_2 + \ldots + \mathbf{k}_\sigma + \Delta\mathbf{k} = 0 \qquad [2]$$

where $\Delta\mathbf{k}$ is the phase-mismatch between the σ interacting waves. Each frequency component and wavevector is ascribed a positive or negative sign, according to their phase relationships. Equation [2] defines phase-matching conditions in which $\Delta\mathbf{k}$ must be minimized to optimize the efficiency of the nonlinear-optical process of interest.

We shall consider below the two specific three-wave nonlinear-optical processes that are particularly relevant to this article, namely, those for an OPD (i.e., OPG, OPA, or OPO) or a DFG. Each of these is mediated by the nonlinear-optical susceptibility tensor $\chi^{(2)}$, which is nonzero in a crystalline medium only if it is noncentrosymmetric. Many such crystals are available. For example, lithium niobate ($LiNbO_3$) has been popular since the early days of pulsed tunable OPOs. Subsequent interest and activity in OPD technology have been stimulated by the availability of new nonlinear-optical materials such as BBO (β-barium borate, BaB_2O_4) and KTP (potassium titanyl phosphate, $KTiOPO_4$).

Additional recent impetus has come from quasi-phase-matched (QPM) nonlinear-optical media, such as periodically poled lithium niobate (PPLN) and PPKTP, tailored for specific wavelengths by periodic optical structuring of domains. QPM media offer compact, efficient, low-threshold alternatives to conventional birefringently phase-matched (BPM) media. Characteristics of BPM and QPM nonlinear-optical crystalline media are accessible *via* Arlee Smith's SNLO software package, accessible by internet at www.sandia.gov/imrl/XWEB1128/xxtal.htm.

For a three-wave OPD, which is of principal interest in this article, the energy and momentum conservation conditions of eqns [1] and [2] become:

$$\omega_P - \omega_S - \omega_I = 0, \qquad \mathbf{k}_P - \mathbf{k}_S - \mathbf{k}_I - \Delta\mathbf{k} = 0 \qquad [3]$$

where a laser input wave ('pump', frequency ω_P, wave vector $\mathbf{k}_P$) yields two coherent output waves ('signal', frequency ω_S, wave vector $\mathbf{k}_S$; 'idler', frequency ω_I, wave vector $\mathbf{k}_I$; $\omega_S \geq \omega_I$). Note that the idler frequency ω_I equals the difference $|\omega_P - \omega_S|$ between pump and signal frequencies. Equation [3] should be viewed in the context of sections (a–c) of Figure 1.

Equation [2] and the second half of eqn [3] apply strictly only to the conventional case of BPM media. In such media, the phase-matching condition $\Delta\mathbf{k} \approx 0$ is achieved by adjusting the angle and/or temperature of a birefringent nonlinear-optical crystal *via* its ordinary- and extraordinary-ray refractive indices. Such adjustments are used to optimize OPD conversion efficiency and thereby control the output signal and idler wavelengths, λ_S and λ_I. If it is assumed that the three waves are collinear and that $\Delta\mathbf{k}$ is exactly zero, then the signal frequency/wavelength is given simply in terms of the pump frequency/wavelength and the refractive indices n_i $(i = P, S, I)$ as:

$$\omega_S = \omega_P(n_P - n_I)/(n_S - n_I); \qquad \lambda_S = \lambda_P(n_S - n_I)/(n_P - n_I) \qquad [4]$$

An alternative to birefringent phase matching is the quasi-phase-matching approach first suggested by Nobel Laureate Nicolaas Bloembergen and co-workers in 1962. This approach has been realized experimentally only in the last few years, *via* QPM nonlinear-optical media such as PPLN or PPKTP. Equation [2] and the second half of eqn [3] need to be slightly modified to include QPM grating contributions.

In the corresponding case of a DFG (which, we repeat, is not an OPD), the energy and momentum conservation conditions of eqns [1] and [2] become:

$$|\omega_1 - \omega_2| - \omega_{diff} = 0, \qquad \mathbf{k}_1 - \mathbf{k}_2 - \mathbf{k}_{diff} - \Delta\mathbf{k} = 0 \qquad [5]$$

where two laser input waves (frequencies ω_1, ω_2; wave vectors $\mathbf{k}_1$, $\mathbf{k}_2$) yield one coherent output wave at the difference frequency $\omega_{diff} = |\omega_1 - \omega_2|$ (with wave vector $\mathbf{k}_{diff}$), as depicted in Figure 1d. As above, phase matching is defined by $\Delta\mathbf{k} \approx 0$ for BPM media and, in the case of QPM media, there is an additional grating contribution in the second half of eqn [5].

Finally, we mention briefly two other important forms of nonlinear-optical device arising from four-wave mixing processes that are mediated by the nonlinear-optical susceptibility $\chi^{(3)}$ which can be non-zero, even in isotropic or centrosymmetric media such as gases, liquids, optical fibers, and all classes of crystal. Optical parametric processes of this type contribute to stimulated Raman scattering (SRS), involving an optical medium with Raman-active resonance frequencies ω_R that coincide with the difference between two optical frequencies. This is particularly useful in providing a relatively straightforward source of coherent radiation, Raman-shifted at discrete intervals from the frequency ω_L of an input pump laser (either tunable or fixed-wavelength). Four-wave parametric processes are involved, such as that in eqn [1] with $|\omega_1| = |\omega_3| = \omega_L, |\omega_2| = \omega_L - \omega_R$, and $|\omega_4| = |\omega_1| + |\omega_3| - |\omega_2| = \omega_L + \omega_R$, where ω_R is the characteristic Raman frequency of the nonlinear-optical medium used in the Raman shifter (e.g., high-pressure hydrogen, deuterium or methane gas). These Raman-shifted intervals, both added to (anti-Stokes) and subtracted from (Stokes) the laser frequency ω_L, are integer multiples of ω_R. Other nonlinear-optical Raman parametric processes give rise to various forms of nonlinear Raman spectroscopy, such as CARS (coherent anti-Stokes Raman spectroscopy), and to Raman fiber-optical amplifiers, used in optical telecommunications.

Another promising area of developing OPD technology entailing $\chi^{(3)}$ nonlinearities concerns OPGs, OPAs, and OPOs based on highly nonlinear optical fibers, with either pulsed or cw pump lasers. Such processes use two pump waves (P) to generate tunable signal (S) and idler (I) output waves, such that $\omega_I = 2\omega_P - \omega_S$.

Nanosecond-Pulsed OPOs – Design and Wavelength Control

Here we consider pulsed OPOs that operate on nanosecond (ns, 10^{-9} s) time-scales, and design

Table 1 Some elements of OPO applications to spectroscopy

Key properties	*Options*	*Comments on instrument/technique*
Wavelength range: what forms of spectra need to be measured?	UV/visible (0.2–0.7 μm) Near IR (0.7–4.0 μm) Longwave IR (>4.0 μm)	Many nonlinear-optical OPO materials are available, but less well developed for longwave IR applications
OPO phase matching: BPM or QPM?	Birefringent phase matching (BPM) Quasi-phase matching (QPM)	BPM is well established and preferred for high-power operation. Various low-threshold QPM media are now available for both pulsed and cw OPOs
Temporal: continuous-wave (cw) or pulsed?	Pulsed for power and timing Cw for narrowest bandwidth	Ultrafast (ps, fs) OPO output pulses are relatively broadband, owing to the Fourier transform limit
Optical bandwidth: broad or narrow?	Broadband (free-running) Single longitudinal mode	The Fourier transform limit is 44 MHz ($0.0015\ cm^{-1}$) for an ideal 10 ns pulse Cw OPOs offer even lower $\Delta\nu$. Ultrafast-pulsed OPOs are inevitably broadband: e.g., $\Delta\nu \geq 0.44$ THz ($\geq 15\ cm^{-1}$) for a 1 ps pulse
Mode of recording of spectra: continuous tuning or multiplex?	Scan narrowband signal/idler OPO output wavelength Free-running OPOs operate broadband in multiplex case	Wavelength control yields continuously tuned narrowband spectra. Multiplex spectra use dispersed detection or multi-wavelength spectroscopic tailoring
Wavelength control: intracavity elements or injection seeding?	Intracavity gratings or étalons Injection seeding of signal or idler by a tunable low-power coherent light source	Intracavity-element designs yield broad tunability but can be complicated. Injection seeding facilitates narrowband, mode-hop-free spectra and tailored multi-wavelength experiments

The above options determine ways to use OPOs for spectroscopic applications, such as: (1) linear absorption (e.g., with a multipass cell); (2) cavity ringdown (CRD) absorption spectra; (3) high-resolution spectra ($\Delta\nu \approx$ MHz or kHz); (4) nonlinear-optical (e.g., coherent Raman); (5) atmospheric remote sensing (e.g., DIAL); (6) fast (μs, ns) and ultrafast (ps, fs) processes.

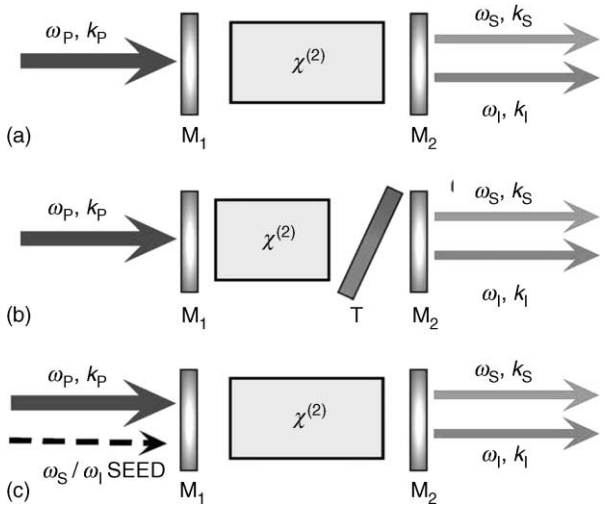


Figure 2 Schematic diagrams of three forms of optical parametric oscillator: (a) free-running OPO (with no active wavelength control), similar to **Figure 1c**; (b) OPO with an intracavity tuning element (T); (c) injection-seeded OPO.

features that make them fit for one of the principal applications of OPDs, namely, spectroscopy. Options that determine ways to use OPOs for spectroscopic applications are summarized in Table 1. Some design and wavelength-control features used in ns-pulsed OPOs are depicted schematically in Figure 2. Some of these applications and features are shared with other forms of OPO, notably the ultrafast-pulsed OPOs and cw OPOs that are considered much more briefly in later sections.

The first OPO, demonstrated by Giordmaine and Miller in 1965, was ns-pulsed; it was based on $LiNbO_3$, tunable over the range ~0.96–1.16 μm, spanning a signal- and idler-wavelength range of ±0.1 μm on either side of the degeneracy point defined by the 529 nm pump radiation (from a frequency-doubled, Q-switched Nd:$CaWO_4$ laser). This advance occurred soon after lasers were discovered and the potential of nonlinear optics had been realized through harmonic-generation processes such as frequency doubling. Pioneering work by Stephen Harris, Robert Byer and their co-workers at Stanford University soon established ns-pulsed OPOs as practical sources of tunable coherent light for significant applications such as spectroscopic sensing of chemical processes, in industrial or environmental diagnostics and in basic science. Such ns-pulsed OPOs had an impressive impact in the 1970s, followed by sporadic progress in much of the 1980s, and a dramatic resurgence of interest and activity since the late 1980s.

This irregular rate of progress is attributable to the availability of suitable high-quality nonlinear-optical crystals (as mentioned above), as well as high-performance pump lasers and advanced tunable OPO system designs.

In a typical ns-pulsed OPO, the pump laser delivers sufficient optical energy to enable parametric gain to build up from noise during the pump–pulse duration, exceeding the threshold for oscillation. To maximize gain, the parametric (signal and/or idler) waves are amplified by multipassing them during the pump pulse in an optical resonator, as depicted in Figures 1c and 2. Light travels ~3 m during a 10 ns pump pulse, so that the parametric waves can make ~15 round trips of a 10 cm linear OPO cavity.

A simple free-running OPO, comprising an optical cavity with input and output mirrors M_1 and M_2 but with no wavelength-selective elements, is depicted in Figure 2a. The output radiation from such an OPO generally has a relatively broad optical bandwidth – typically ~5 cm^{-1} (~150 GHz) or more – depending on several factors: refractivity, dispersion, and absorption of the OPO medium; wavelengths λ_P, λ_S, and λ_I; type of phase matching (BPM or QPM, whether collinear or not); crystal dimensions and orientation; cavity reflectivity and effective number of passes of the resonated wave; optical bandwidth, divergence, pulse duration and pulse energy of the pump radiation. Free-running pulsed OPOs represent one extreme of operational simplicity, yielding relatively broadband output radiation suitable for low-resolution or multiplex spectroscopy. Additional OPO wavelength-control measures are usually necessary for higher-resolution spectroscopic applications.

At the other extreme of operational complexity, intracavity wavelength-selective elements, such as gratings and/or étalons, provide a traditional way to control OPO output wavelength and to achieve narrow optical bandwidths. One such approach is depicted in Figure 2b, where a tuning element T (in this case, a tilted étalon or filter; in other designs, an intracavity diffraction grating replacing the output cavity mirror M_2) is inserted in the cavity. Such approaches were used in early pulsed $LiNbO_3$ OPO designs that were continuously tunable in the near IR with an optical bandwidth of ~0.1 cm^{-1} (~3 GHz), but these were often difficult to operate and were damage-prone (with intracavity losses from gratings and étalons causing the operating threshold to approach the damage threshold of OPO nonlinear-optical materials such as $LiNbO_3$).

Continuous narrowband tunability, preferably on a single-longitudinal-mode (SLM) basis, is a desirable performance characteristic of pulsed OPOs amenable to high-resolution spectroscopy. A commercially viable approach to this ideal was achieved by Walter Bosenberg and Dean Guyer in their advanced KTP OPO/OPA system, which includes a master OPO stage with a grazing-incidence grating. This is continuously tunable under computer control in the near IR (1.3–4 μm) with narrow optical bandwidth – ~0.02 cm^{-1} (~600 MHz) or better. Such a high-performance tunable coherent IR source has been used for various forms of laser spectroscopy (cavity ringdown, degenerate four-wave mixing, long-path absorption, etc.), including investigations of chemically reactive media, combustion diagnostics, and studies of processes in molecular beams.

A popular alternative approach to OPO wavelength control is injection seeding by a low-power tunable coherent source such as a tunable diode laser (TDL). This approach is depicted in Figure 3; in practice, a ring cavity is often used to avoid

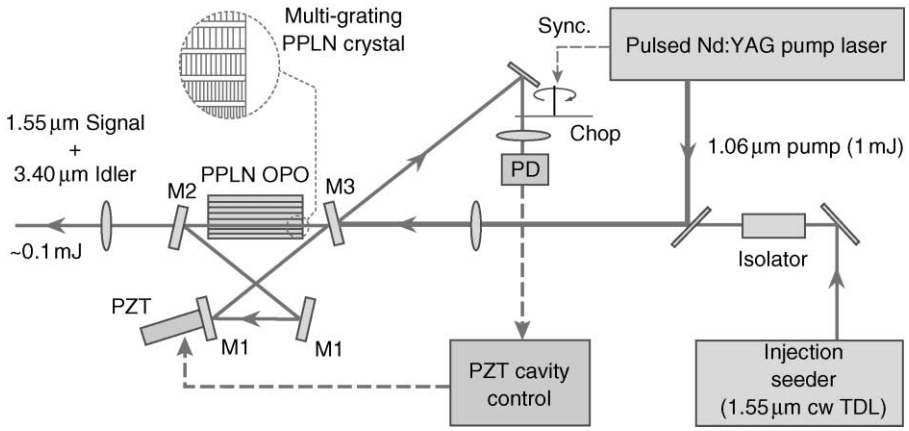


Figure 3 Schematic diagram of an injection-seeded tunable PPLN OPO system with active cavity control, developed in the author's laboratory. Legend: PD = photodetector, PZT = piezoelectric translator, TDL = tunable diode laser, M_{1-3} = cavity reflectors; the inset shows the QPM multi-grating structure of the PPLN nonlinear-optical crystal.

feedback from the OPO to the seed laser. A significant advantage of injection seeding is that OPO construction is simplified by putting the wavelength-control function into a module that is effectively separate from the optical generation and amplification functions. For example, a high-performance, TDL-seeded PPLN OPO system developed in the author's laboratory is illustrated in Figure 3. It features a multigrating QPM gain medium that is temperature-adjustable up to 250°C, and is pumped at a wavelength of 1.064 μm by a ns-pulsed, SLM Nd:YAG laser. An active cavity control scheme devised by Yabai He enables the SLM near-IR signal and idler outputs from this OPO (at ~1.5 μm and ~3.5 μm, respectively) to be tuned continuously without mode hops over optical frequency ranges of ~250 cm^{-1} (~7.5 THz) with a SLM optical bandwidth of <0.004 cm^{-1} (<120 MHz), as has been confirmed in assorted applications to molecular spectroscopy. The PPLN OPO can by itself generate only moderate output pulse energy (~0.1 mJ, approximately 70% of which is signal radiation), since the 1.064 μm Nd:YAG pump laser pulse energy is limited to ~1 mJ to avoid the optical damage threshold of the PPLN crystal. For higher-power applications, a further portion of the Nd:YAG laser energy (~200 mJ/pulse) can be used to pump an OPA stage based on BPM $LiNbO_3$, thereby generating SLM signal and idler output pulse energies of ~1.3 mJ and ~0.7 mJ, respectively.

An alternative approach to injection-seeded tuning of ns-pulsed OPOs (used extensively by the author's group in the context of BPM media such as BBO and $LiNbO_3$) entails passive control of the OPO cavity by slightly misaligning one of its reflectors. This facilitates continuous tuning of the injection-seeded OPO signal and idler outputs by decreasing the effective finesse of the OPO cavity, so that it is then not necessary to lock the OPO cavity length to the seed wavelength. This mode of operation is simpler optically and electronically than active control of the OPO cavity. The method depends on the OPO cavity having a series of high-order transverse modes to smooth out the sharp, widely separated resonances that occur when the OPO cavity is well aligned; a resulting disadvantage is that the multiple transverse modes tend to cause some degradation of output beam quality. Nevertheless, this approach has proved useful for many applications of tunable OPOs, with seeding by either pulsed dye lasers or SLM TDLs.

This passive, misaligned-cavity approach to injection seeding of ns-pulsed OPOs is particularly well suited for spectroscopic applications requiring a coherent source that simultaneously generates two or more adjustable output wavelengths. Such a TDL-seeded, dual-wavelength BBO OPO has been used in the author's laboratory for coherent-Raman spectroscopic measurements of nitrogen (N_2) in furnace air; the relative intensity of the two spectroscopic features provides an instantaneous, turbulence-immune estimate of the temperature of gas in the furnace, within a single shot of the OPO's pump laser (i.e., within ~10 ns). This is an example of so-called spectroscopic tailoring, in which an OPO generates a structured set of two or more discrete output wavelengths, each of which is set to be on- or off-resonance with characteristic features in spectra of molecular target species of interest.

Dual-wavelength injection seeding of ns-pulsed OPOs is relevant to atmospheric remote sensing by DIAL (differential absorption lidar), with TDL-seeded OPO output switched between on- and off-resonance IR wavelengths on alternate shots of the pump laser; range-resolved measurements have been made of key atmospheric species such as methane (CH_4) and water vapor (H_2O). Figure 4 illustrates a proposed extension of the spectroscopic tailoring concept, applicable to dual- and multi-wavelength remote sensing applications by long-path absorption or DIAL. This multiplex system comprises a multi-wavelength passive-cavity pulsed OPO, injection-seeded by a set of single-mode TDLs and a computer-controlled fiber-optic switch to code and decode the resulting multiplex spectroscopic signals.

A significant frontier for narrowband pulsed OPOs entails extending the spectral output range beyond the present readily attainable near-IR limit of ~4 μm (as provided by materials such as BBO, KTP, $LiNbO_3$, PPLN, and PPKTP). This extension yields high-performance tunable coherent spectroscopic sources at longwave-IR wavelengths where many of the strongest molecular fundamental IR absorption bands are located. OPOs that target the longwave IR depend on availability of suitable nonlinear-optical materials and coherent pump sources at sufficiently long wavelengths (typically >2 μm). One of the most promising longwave-IR OPO materials is zinc germanium phosphide ($ZnGeP_2$), pumped at 2–3 μm by solid-state crystal lasers such as those based on holmium or erbium. An étalon-tuned narrowband (~0.1 cm^{-1}) $ZnGeP_2$ OPO, itself pumped at 2.55 μm by a broadband (~15 cm^{-1}) Nd:YAG-pumped $LiNbO_3$ OPO, is able to cover the wavelength range 3.7–8 μm and has been used to record spectra of H_2O vapor in the 6.2 μm region. Pulsed OPOs based on gallium arsenide (GaAs) also have great potential; they can be pumped at wavelengths as short as 1 μm and

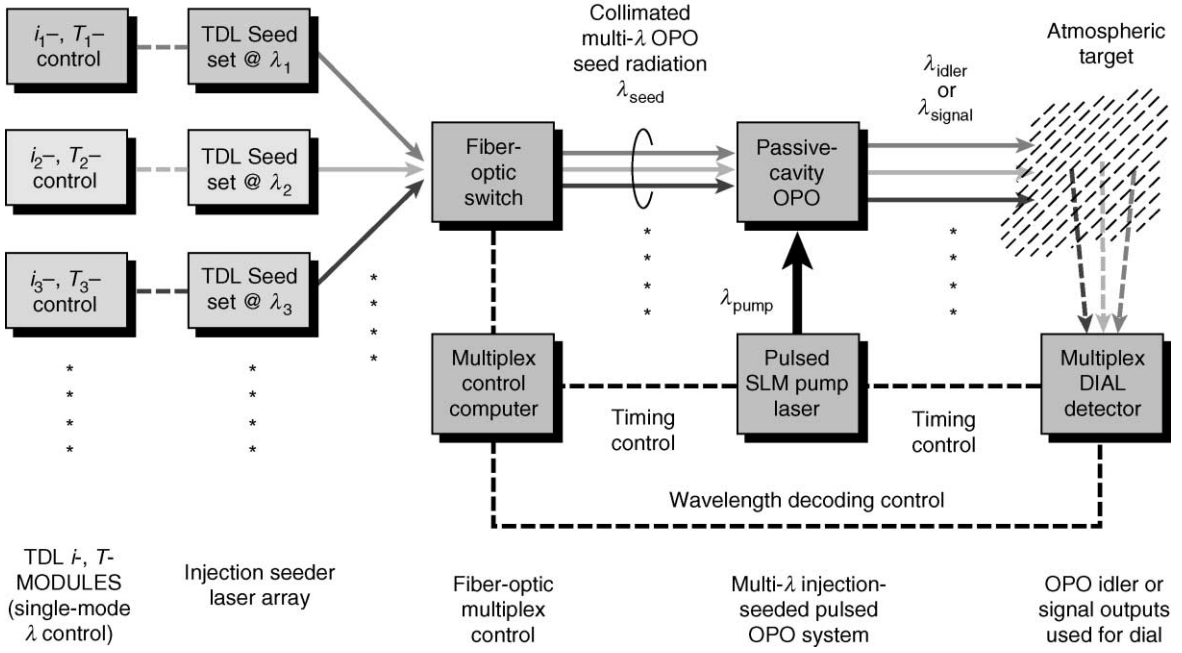


Figure 4 Schematic diagram of a multi-wavelength pulsed OPO system employing spectroscopically tailored injection seeding. This design has been proposed by the author for atmospheric remote sensing by DIAL (differential absorption lidar) or long-path absorption.

offer (so far unrealized) tunability out to ~16 μm. Despite its cubic structure and lack of birefringence, GaAs serves as a QPM nonlinear-optical medium in the form of orientation-patterned films.

Another approach to operating continuously tunable OPOs is to vary the wavelength of the pump laser, while fixing other phase-matching parameters such as temperature of the gain medium and crystal angle or QPM grating pitch. Suitable tunable pump laser sources that have been used to pump ns-pulsed OPOs include Ti:sapphire and tunable fiber lasers.

The various wavelength-control strategies outlined above illustrate some of the distinctive attributes of OPOs. It should be recognized that the signal and idler wavelengths of an OPO are distinctly different from that of the pump laser itself, thereby opening up fresh regions of the electromagnetic spectrum. Even if an OPO is injection-seeded (e.g., by a cw TDL) at its signal or idler wavelength, it still generates complementary idler or signal radiation at a new wavelength and (in the case of a pulsed OPO) it also pulse-amplifies the original seed light. Moreover, the nonlinear-optical parametric relationships between the pump, signal, and idler waves in an OPO allow various wavelength-control approaches (e.g., spectroscopic tailoring by multi-wavelength injection seeding) that are simply not attainable in tunable lasers.

Ultrafast-Pulsed OPOs

Given the relatively high peak power of very short (sub-ns) pulses of coherent laser light, it might seem quite easy to use such light to pump the gain medium of an OPD and to exceed its OPO threshold. However, such a task is not at all straightforward, because an ultrafast light pulse, on the time-scale of picoseconds (ps, 10^{-12} s) or femtoseconds (fs, 10^{-15} s), does not stay in one place long enough to enable a coherent parametric wave to build up from noise in the nonlinear-optical medium. Light travels a distance of only ~0.3 mm in 1 ps, which does not allow it to make multiple traversals of the nonlinear-optical medium and build up coherent signal and idler waves. This problem is overcome by synchronous pumping (also used in some ultrafast lasers), in which a train of many consecutive ultrafast pulses from a pump laser interact sequentially with a single signal or idler pulse circulating within the OPO cavity. The mode-locking interval of the pump laser (essentially the round-trip transit time in the laser cavity) must therefore equal the round-trip time of the down-converted (signal or idler) pulse in the OPO cavity. The situation is aggravated when sub-ps (e.g., ~100 fs) pump pulses are used, because the pump and signal (or idler) waves have sufficiently different group velocities that they undergo

'temporal walk-off', becoming separated in time after traversing a relatively short length (typically 1–10 mm for 100-fs pulses) of the nonlinear-optical medium.

Practical ultrafast OPOs were introduced around 1990 and their utility grew as improved pump lasers, nonlinear-optical materials (e.g., KTP, PPLN), and OPO cavity designs became available. Research on ps- and fs-OPOs is now a hot topic, and numerous broadly tunable commercial systems are on the market. Ultrafast OPO-based spectroscopy is becoming extremely important in applications to biology and medicine, where many key processes occur on ps- and fs-time-scales.

Continuous-Wave OPOs

The first cw OPO, reported in 1968, operated in the near-IR and was based on barium sodium niobate ($Ba_2NaNb_2O_5$ – known popularly as 'banana'); it was soon followed by visible cw OPOs based on $LiNbO_3$. However, there was not much more substantial progress on practical cw OPOs for another 20 years, until several technological advances had been made: low-threshold QPM nonlinear-optical gain media, such as MgO:$LiNbO_3$ and KTP; efficient, wavelength-selective cavity designs; all-solid-state cw pump lasers (e.g., diode-pumped crystal lasers or MOPA diode systems). More recently, QPM media (such as PPLN and PPKTP) have provided a further stimulus to efficient, compact cw OPOs.

Most ns-pulsed OPOs use singly resonant (SROPO) designs in which the cavity resonates at either the signal or idler wavelength (but not both), doubly resonant (DROPO) schemes in which both signal and idler resonate are often needed to meet the more demanding pump threshold requirements of cw OPOs. Continuous tuning of a DROPO without longitudinal mode hops needs multiparameter approaches, either dual-cavity (with separate signal- and idler-resonant arms) or concerted tuning of pump and signal/idler wavelengths. Efficient cavity designs for cw OPOs include those that are pump-enhanced (i.e., the cavity is resonant at the pump wavelength as well as the signal and/or idler) and intracavity arrangements in which the laser and OPO gain media share a common cavity. Progress is also being made on cw OPOs based on highly nonlinear optical fibers. With such developments, cw OPOs now have great potential as tunable spectroscopic sources, through their narrow optical bandwidth and their compactness.

OPGs and DFGs – Dispensing with the Optical Cavity

Most of this article has focused on OPOs, in which parametric processes are generated by pumping a nonlinear-optical medium in an optically resonant cavity. As already explained, such a cavity is used to enhance the optical parametric gain by multipassing the signal and/or idler radiation. In some cases, the cavity also contributes to wavelength control, temporal characteristics, beam quality, etc. However, not all optical parametric and related devices need a cavity, as is evident from the OPG, OPA, and DFG schematics in **Figures 1a, b,** and **d**.

The use of one or more OPA stages, to supplement the intracavity amplification already present in an OPO itself, is widespread in tunable OPO/OPA systems. Moreover, in some applications (e.g., spectroscopic measurements where optical power is not at a premium) of pulsed OPDs, it is found that a single pass of a (relatively long) nonlinear-optical crystal results in a useful OPG (without an OPO cavity). For instance, Scott Bisson, Thomas Kulp, Peter Powers and co-workers have developed and used several OPD systems of this type. A cavity-free OPG/OPA system that is pumped by a 1 kHz Nd:YAG laser, based on 'fan' PPLN (tuned by continuous transverse variation of the QPM grating pitch), and spectrally filtered by a high-finesse étalon, generates ~10 μJ/pulse of idler radiation. This OPG/OPA radiation has been used to measure narrowband ($\Delta\nu < 0.1\ cm^{-1}$) cavity-ring-down spectra of CH_4 gas.

Likewise, various tunable DFG devices (closely related, but not identical, to OPDs) have been used in many spectroscopic applications. For instance, a DFG system developed by Frank Tittel and colleagues at Rice University took the near-IR output of two room-temperature TDLs (at 1.55 μm and 1.32 μm, respectively), mixed them in a BPM nonlinear-optical medium such as $AgGaSe_2$, and generated 8.7 μm mid-IR radiation for long-path absorption studies of sulfur dioxide (SO_2), an important atmospheric pollutant. The same group has also used QPM media (e.g., PPLN) for DFG mixing of TDL outputs in the 0.8 μm–1.6 μm range, yielding output at ~3.5 μm to monitor gases such as H_2O, CH_4, and formaldehyde (H_2CO); such DFG devices have been used in NASAs Lunar-Mars Life Support Test program.

Relative to OPGs and DFGs, OPOs and OPAs offer higher output power and coherence, both of which are advantageous in some applications (e.g., remote spectroscopic sensing of the atmosphere). Nevertheless, OPGs and DFGs still have a useful role to play, as indicated above.

Concluding Remarks – When to Use an OPD?

Coherent light sources based on optical parametric generation, amplification, and oscillation represent a significant application of nonlinear optics. Such devices can in turn address as many applications as lasers themselves, at least in principle. However, the fact that OPDs need to be pumped by a laser tends to limit their use to situations that take advantage of their wavelength-conversion and tunability capabilities. The most obvious areas of application of an OPO or any other OPD are therefore those in which the wavelength of the output radiation is critical (e.g., spectroscopy, as indicated in various parts of the above text).

From time to time, there have been suggestions that OPOs would largely displace established tunable lasers such as dye and Ti:sapphire lasers. In the early 1990s, one major laser manufacturer ventured to advertise its new OPO-based instrument with the slogan, "Goodbye to Ti: and Dye". However, such a projected monopoly for OPOs has not eventuated and most laser spectroscopy facilities now depend upon a balanced combination of OPOs together with various forms of tunable laser (including semiconductor lasers, which are almost ubiquitous).

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(2)}$-harmonic Generation; $\chi^{(3)}$ – Third-harmonic Generation; Four-wave Mixing. **Optical Parametric Devices:** Optical Parametric Oscillators (Continuous Wave); Optical Parametric Oscillators (Pulsed). **Scattering:** Stimulated Scattering. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Barnes NP (1995) Optical parametric oscillators. In: Duarte FJ (ed.) *Tunable Lasers Handbook*, pp. 293–348. San Diego, CA: Academic Press.

Byer RL (1975) Optical parametric oscillators. In: Rabin H and Tang CL (eds) *Quantum Electronics: A Treatise*. Nonlinear Optics, part B, chap. 9, vol. 1, pp. 587–702. New York: Academic.

Byer RL and Herbst RL (1977) Parametric oscillation and mixing. In: Shen Y-R (ed.) *Nonlinear Infrared Generation*, pp. 81–137. New York: Springer-Verlag.

Dmitriev VG, Gurzayan GG and Nikogosyan DN (1997) *Handbook of Nonlinear Optical Crystals*, 2nd edn, pp. 339–362.

Dunn MH and Ebrahimzadeh M (1998) Optical parametric oscillators. In: Finlayson DM and Sinclair BD (eds) *Advances in Lasers and Applications*, pp. 61–82. Bristol and Philadelphia: Institute of Physics Publishing.

Ebrahimzadeh M and Dunn MH (2001) Optical parametric oscillators. In: Bass M, Enoch JM, Van Stryland EW and Wolfe WL (eds) *Handbook of Optics, Fiber Optics and Nonlinear Optics*, vol. 4, pp. 22.1–22.73. New York: McGraw-Hill.

Harris SE (1969) Tunable optical parametric oscillators. *Proceedings of the IEEE* 57: 2096–2113.

Orr BJ (2000) IR Lidar applications in air monitoring. In: Myers RA (ed.) *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation, Applications of Instrumental Methods*, vol. 3, pp. 2007–2032. Chichester: Wiley.

Orr BJ, Johnson MJ and Haub JG (1995) Spectroscopic applications of pulsed tunable optical parametric oscillators. In: Duarte FJ (ed.) *Tunable Laser Applications,* chap 2, pp. 11–82. New York: Marcel Dekker.

Sorokina IT and Vodopyanov KL (eds) (2003) *Solid-State Mid-Infrared Sources*. New York: Springer-Verlag.

Sutherland RL (1996) Optical parametric generation, amplification, and oscillation. *Handbook of Nonlinear Optics,* chap. 3, pp. 111–206. New York: Marcel Dekker.

Optical Parametric Oscillators (Continuous Wave)

S Schiller, Heinrich-Heine Universität Düsseldorf, Düsseldorf, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A continuous-wave optical parametric oscillator (cwOPO) is a frequency conversion device in which a continuous-wave (cw) laser field (pump) of frequency ω_p is partially converted into two cw waves of lower frequencies ω_s, ω_i (historically denoted by signal and idler) inside an optical cavity. The three frequencies are related by $\omega_i + \omega_s = \omega_p$. The cavity contains a nonlinear medium with a second-order susceptibility and the mirror reflectivities are chosen such that at least one of signal and idler resonates in the cavity. The conversion efficiency from pump power to signal or idler power or both can be larger than 10%. The frequencies of signal and idler waves can be tuned by means similar to conventional lasers, i.e., by changing the spectral dependence of the cavity

loss and by tuning the cavity length. Additional tuning parameters specific to OPOs are the pump frequency and the refractive indices of the nonlinear medium for pump, signal and idler. CwOPOs provide a unique combination of features that makes them suitable sources for a variety of spectroscopic applications. Today's cwOPOs can exhibit the following features:

- spectral coverage from the near-infrared (0.7 μm) to the mid-infrared (5.2 μm); for a single device two optical octaves can be covered without change of optical components;
- high output power (up to several W);
- high pump power conversion efficiency (>30%) and good wall-plug efficiency;
- nearly diffraction-limited output beams (modes are close to lowest-order transverse electromagnetic (TEM_{00}) modes)
- single-frequency output, absence of frequency jumps (mode hops);
- narrow free-running linewidth (<6 kHz);
- high absolute frequency stability (<100 MHz drift per hour);
- continuous frequency tuning exceeding 10 GHz;
- rapid, stepwise tuning over several THz; and
- transportability.

cwOPOs complement their pulsed relatives (*see* **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed)) nanosecond and synchronously pumped pulsed parametric oscillators (*see* **Optical Parametric Devices:** Overview).

Basic Principle

The operation of a cwOPO is most simply described for a singly resonant OPO (SRO) (see Figure 1 and Figure 2a).

Without loss of generality we take the signal wave to be the resonant wave. When a higher-frequency pump wave (power P_p, angular frequency ω_p) and a lower-frequency (signal) wave (power P_s, angular frequency ω_s) traverse a $\chi^{(2)}$ medium (*see* **Fiber and Guided Wave Optics:** Nonlinear Optics), a transfer of power from the pump wave to the signal wave occurs, described by a power gain:

$$\Delta P_s = GP_s = EP_pP_s \qquad [1]$$

This parametric amplification expression is valid as long as $\Delta P_s \ll P_s, P_p$, i.e., in the limit of small gain and small pump depletion. The (unsaturated) gain coefficient E is a function of: the medium's indices of refraction, orientation and length L, the frequencies

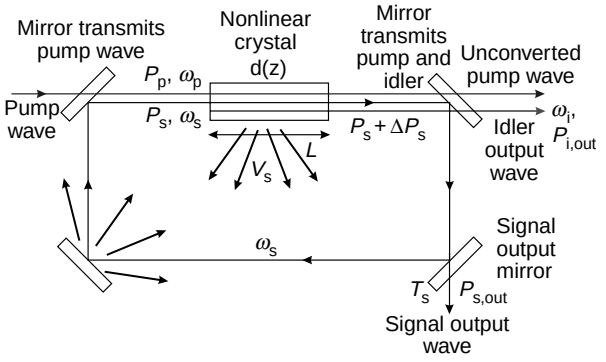


Figure 1 Schematic of a singly resonant cwOPO (SRO) in ring configuration. The signal wave (ω_s) is resonated, the idler wave (ω_i) is not. In the nonlinear crystal, the pump wave overlaps with the signal and idler waves; here they are shown displaced with respect to each other for clarity. The signal wave's attenuation coefficient per roundtrip, S_s, is the sum of loss inside the nonlinear crystal, V_s (indicated by arrows), and transmission through mirrors, here represented by a single mirror of transmission T_s. Scatter and absorption losses at mirrors (indicated by arrows) are included in V_s.

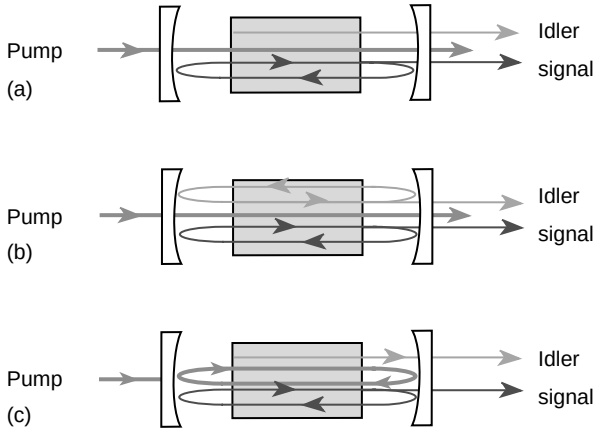


Figure 2 Common types of cwOPOs. (a) basic SRO. (b) Doubly resonant OPO (DRO). Standing-wave cavities are shown. (c) SRO with enhancement of the pump wave (PR-SRO); a common-cavity configuration is shown.

ω_s, ω_p, and the spatial modes and overlap of the two beams.

If the wave ω_s is made to circulate inside a cavity and the power attenuation per roundtrip, S_sP_s is smaller than the power gain per roundtrip (eqn [1]), a sustained oscillation is possible. This is satisfied when the input pump power P_p exceeds the threshold:

$$P_{p,in}^{th} = S_s/E \qquad [2]$$

As the pump power P_p is increased above $P_{p,in}^{th}$, conversion of pump power to idler power occurs, implying that the pump power decreases within the medium. Qualitatively, the average pump power $\langle P_p \rangle$

along the medium remains at the level $P_{p,in}^{th}$, so that the roundtrip fractional power gain remains constant, $G \simeq E\langle P_p\rangle \simeq EP_p^{th} \simeq S_s$. This ensures that the device reaches a stationary state. The circulating signal power P_s increases steadily with increasing pump power, and so does the signal power emitted from the OPO, T_sP_s, which is equal to the 'useable' fraction of the roundtrip power gain, $(T_s/S_s)GP_s$.

The parametric process that allows the signal wave to experience gain (eqn [1]), also generates an idler wave (difference frequency generation). Its power at the end of the medium is given by the Manley–Rowe relation:

$$P_i = \frac{\omega_i}{\omega_s}\Delta P_s \qquad [3]$$

An OPO also requires a process that starts the oscillation. Spontaneous parametric fluorescence provides this trigger. Due to this quantum effect, a pump wave traversing a $\chi^{(2)}$ medium will generate a flux of signal photons even if there is no signal wave present at the crystal input face. In the photon picture, a tiny fraction (on the order 10^{-9} for a 1 W pump) of the pump photons spontaneously annihilate, creating pairs of signal and idler photons (photon splitting). The idler photons have a frequency $\omega_i = \omega_p - \omega_s$, so that photon energy conservation holds. The process may be thought of as being stimulated by the vacuum fluctuations of the electromagnetic field at the signal and idler frequencies. Some of the signal photons are emitted into the cavity mode, so that after a roundtrip they can serve as a seed for further (stimulated) photon splitting. The process repeats, and the signal wave builds up exponentially in time to a macroscopic level if the pump power is above threshold. The spontaneous (as well as the parametric amplification) process is strongest when momentum conservation is essentially fulfilled, $k_p \simeq k_s + k_i$, where $k = \omega n(\omega)/c$.

In the photon picture, the relation (eqn [3]) is a statement that signal and idler photons are always created in pairs.

Types of OPOs

OPO types may be classified according to the number of waves that resonate within cavities. Typically, a realizable roundtrip loss coefficient is on the order of $S_s = 1\%$ (to which the linear absorption of the nonlinear crystal and mirror transmissivities contribute) and the gain coefficient $E \sim 0.1\%/\mathrm{W}-1\%/\mathrm{W}$ for Gaussian waves, depending on the pump wavelength, crystal type, etc. This leads to SRO threshold powers on the order of 1 to 10 W. Until the early 1990s, solid-state pump lasers of this power level and single-frequency output were unavailable. This led to the interest in and development of cwOPOs in which more than one of the three waves is resonated (Figure 2). The corresponding external threshold powers are

$$\begin{aligned} P_{p,in}^{th} &\simeq S_sS_i/4E \text{ for a doubly-resonant OPO} \\ &\qquad (\text{DRO, Figure 2b}) \\ P_{p,in}^{th} &\simeq S_sS_p/E \text{ for a pump-resonant SRO} \\ &\qquad (\text{PR-SRO, Figure 2c, see eqn [35]}) \\ P_{p,in}^{th} &\simeq S_sS_iS_p/E \text{ for a triply-resonant OPO (TRO)} \end{aligned} \qquad [4]$$

Here $S = T + V$ is the fractional power attenuation per roundtrip, arising from the mirror transmission T and absorption/scatter loss V. While the mirror transmission maintains power in a given spatial mode, the loss V removes optical power completely and is the fundamental cause for the nonunity efficiency of the frequency conversion process. Typically, each additional resonant enhancement lowers the threshold power by one to two orders.

Examples for threshold values reported in the literature are: 1.9 W for a 920 nm-pumped SRO, 4 mW for a 0.5 μm-pumped DRO, and 140 mW for a 1 μm-pumped common-cavity PR-SRO. DROs can therefore be pumped even by low-power diode lasers. TROs (where pump, signal, and idler are resonated) are not of importance due to their higher complexity.

Output Power

As an example, the signal and idler output powers for a PR-SRO that resonates the signal are (for a derivation, see Appendix B):

$$P_{s,out} \simeq T_sP_s(0), \qquad P_{i,out} \simeq (\omega_i/\omega_s)S_sP_s(0) \qquad [5]$$

where the circulating signal power is

$$P_s(0) \simeq \frac{\omega_s}{\omega_p}\frac{4T_p}{S_sS_p}P_{p,in}^{th}\left(\sqrt{\frac{P_{p,in}}{P_{p,in}^{th}}} - 1\right) \qquad [6]$$

T_p is the input mirror transmission for the pump wave, and T_s is the output mirror transmission of the signal wave. Signal and idler output powers have the same dependence on pump input power and increase monotonically with $P_{p,in}$. The conversion efficiencies reach a maximum four times above threshold, i.e., when $P_{p,in} = 4P_{p,in}^{th}$. There, the signal and idler photon (quantum) conversion efficiencies $\omega_pP_{s,out}/P_{p,in}\omega_s$ and $\omega_pP_{i,out}/P_{p,in}\omega_i$ exceed 25% and 50%, respectively, if the mirror transmissions T_p, T_s for pump and signal

are chosen larger than the respective loss V_p, V_s. This can be achieved in practice, of course, at the expense of a higher threshold.

Nonlinear Materials

While the fundamental developments in cwOPOs were performed using birefringently phase-matched nonlinear crystals ($MgO:LiNbO_3$, LBO, KTP, $KNbO_3$) (*see* **Materials for Nonlinear Optics:** Liquid Crystals for NLO), and substantial spectral coverage is possible with these crystals, most cwOPOs are now operated with quasi-phase-matched (QPM) nonlinear crystals (*see* **Nonlinear Optics, Application:** Phase Matching). The most commonly used material is periodically poled (PP) $LiNbO_3$ (PPLN), other crystals used are PPKTP and $PPLiTaO_3$, crystals. Their use has led to a substantial widening of the spectral coverage, opening in particular the wavelength range $\lambda > 1.6$ μm, a reduction of the required pump powers thanks to their much higher nonlinearities compared to birefringently phase-matched crystals, and a simplification of (gross) tuning through the use of multigrating structures (containing several sections of differing Λ periods side by side within the same crystal). The above PP crystals exhibit a loss of 0.1%/cm at wavelength around 1 μm, and can be coated with broadband dielectric antireflection coatings, in order to maximize the overall transmission. Another important aspect is that the crystals withstand long irradiation times with high-power focused cw light. For PPLN, operation temperature above 100°C is chosen in order to prevent photorefractive effects.

Spectral Coverage

An example of particularly wide spectral coverage reported in the literature is a cwOPO pumped by a frequency-doubled $Nd:YVO_4$ laser ($\lambda_p = 532$ nm) using a multigrating PPLN crystal. With grating periods $\Lambda = 6.5–9.6$ μm, the range 660–1030 nm (signal wave) and 1100–2800 nm (idler wave) was covered. A cwOPO with a Nd:YAG pump source emitting at the fundamental wavelength ($\lambda_p = 1064$ nm) and a multigrating PPLN crystal ($\Lambda = 25.5–31.2$ μm) achieved a range of 1.45–1.99 μm (signal) and 2.3–4.0 μm (idler).

For idler wavelengths beyond 4 μm, idler absorption by the crystals types above becomes relevant, leading to a reduction in the gain G (this effect is not considered in the treatment given in the Appendix) and to an increase of the threshold. To a certain extent, this can be compensated with powerful cw pump lasers. For example, a 1064 nm pumped PPLN-SRO achieved idler emission in the range 3.6–4.7 μm at 11 W pump power. Also, an 850 nm pumped PR-SRO covered an idler range 4.3–5.3 μm when pumped with 750 mW.

In the future, cwOPOs will certainly profit from the development of novel QPM materials with wider IR transparency range.

Frequency Control

The small-signal gain $G(\omega_s)$ of a parametric interaction in a cwOPO is analogous to the homogenously broadened gain in a laser medium. In contrast to it, however, spatial hole-burning cannot occur. Therefore, mode competition is generally strong in a cwOPO, which will oscillate on a single longitudinal mode: the one for which the difference (or ratio) between 'unsaturated' parametric gain G and total loss S is largest (Figure 3). This is equivalent to the condition of minimum threshold.

The functions $G(\omega_s)$ and $S(\omega_s)$ are therefore important characteristics of a cwOPO. Oscillation can only occur for a frequency ω_s that is very close to a cavity mode frequency $\omega_q = 2\pi qc/OPL(\omega_q)$, where q is the integer mode number, and $OPL(\omega_s)$ is the round-trip optical path length at the frequency ω_s.

To tune a cwOPO to a particular signal/idler frequency pair, a coarse tuning is initially performed

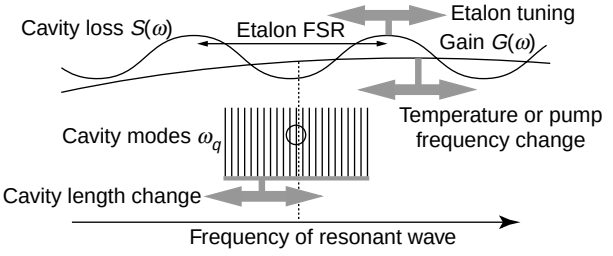


Figure 3 Frequency selection and tuning in a SRO (or PR-SRO) containing an etalon. Oscillation occurs on the signal cavity mode (circled) closest to the frequency where the difference between the small-signal gain G and the loss S is largest (dashed line). The frequency of this mode or the mode number q can be changed by changing the parameters indicated near the thick double arrows. The parameter changes shift the line(s) 'attached' to the corresponding double arrow. Changing the temperature leads to a rough setting of the oscillation frequency. Tilting the etalon allows mode-hop tuning of signal and idler frequencies. Fine tuning by a small amount can be done by changing the cavity length. The frequency of the idler is not shown here; it is always given by the difference between pump frequency and the frequency of the oscillating mode. Thus, tuning of the pump frequency generally leads to a tuning of the idler frequency. For modest pump tuning the amount of signal tuning however depends on SRO type: no signal tuning in a basic SRO and in a dual-cavity PR-SRO, while in a common-cavity PR-SRO pump and signal frequencies are correlated and thus pump tuning leads to signal tuning. FSR: free spectral range.

by tuning the corresponding phase mismatch ΔkL (L is the crystal length) approximately to zero. Taking into account a modulation of the nonlinear susceptibility of the crystal with period Λ, the quantity $\Delta k = k_p - k_s - k_i \pm 2\pi/\Lambda$ describes the deviation from momentum conservation in the photon splitting process. When using a multigrating QPM crystal, the first tuning consists in the selection of a grating of appropriate period Λ, which is moved into the pump wavepath. This is followed by tuning the photon momenta via the refractive indices of the nonlinear medium, which is usually accomplished by temperature tuning, or alternatively by angle-tuning of the crystal. These steps place the maximum of $G(\omega_s)$ to the vicinity of the desired frequency. Options for tuning the emission exactly to the desired frequency are described below for the different OPO types.

Because external perturbations change parameters that enter in $G(\omega_s)$, $S(\omega_s)$, and ω_q, the value of the oscillating frequency will also change in time. The change will be continuous over a certain range of the parameters, but if their perturbation is large enough, a mode hop can occur: the oscillation on the initial mode q will eventually cease and another mode q', for which $G(\omega_{q'}) - S(\omega_{q'})$ is larger, will oscillate instead. A variety of parameters is affected by perturbations: pump power, pump beam direction, cavity length, crystal temperature, and pump frequency.

When such perturbations are applied on intentionally, the OPO output frequencies can be tuned, scanned, or adjusted. The tolerable change in the various parameters before a mode-hop occurs, and the sensitivity of the output frequencies to changes in the parameters depends on the particular cwOPO type. One important approach to suppress mode-hops consists in adding elements into the cavity, such as an etalon, that modify the function $S(\omega)$, typically by modulating it spectrally.

SRO

In the following, we assume that the signal wave is resonated. The SRO is the simplest cwOPO type. The oscillation nominally occurs at the cavity frequency that is closest to the gain maximum, as described above. In practice, environmental disturbances will cause mode-hops on a time-scale of minutes. For this reason, an etalon is added to the cavity. Mode-hop-free oscillation, for several hours, is then achieved.

Typical nonlinear crystals exhibit parametric gain with spectral widths of hundreds of GHz. In contrast, the cavity mode spacing (free spectral range, FSR) c/OPL is on the order of 1 GHz. The tuning range achievable by a cavity length change (using, e.g., a piezo translator) is therefore, at most, one free spectral range (equal but opposite for signal and idler); then a mode-hop to the neighboring cavity mode occurs, with a signal/idler frequency jump by one FSR in the opposite direction to the previous tuning. This process repeats itself as the cavity length is tuned further. It is, therefore, not possible to easily and reliably access a large frequency range by using the cavity length as tuning parameter.

As shown in Figure 3, if an etalon of linewidth smaller than the gain bandwidth is added to the cavity, it becomes the dominant influence (etalon FSRs on the order of one hundred to a few hundred GHz have been used). This means that shifting the etalon's transmission maximum (i.e., the cavity loss S) in frequency space (by tilting a solid etalon or changing the spacing of an air-spaced etalon) as indicated by the double-headed arrows in Figure 3, forces the signal to mode-hop to an adjacent mode, with a corresponding idler frequency hop in the opposite direction. As the etalon is tuned further this occurs repeatedly. This discontinuous 'mode-hop tuning' is useful for tuning the SRO output frequencies over a significant range (e.g., 100 GHz) and allows to reach a desired frequency within one cavity free spectral range.

Continuous tuning of both signal and idler waves is obtainable, if the pump laser is not frequency tunable, by changing the cavity length and synchronously tuning the etalon. This can be achieved by a feedback or feedforward system that controls the etalon angle or spacing. Up to 38 GHz tuning for signal and idler were achieved in this way.

When a tunable pump is available, one can tune the pump frequency while keeping the cavity length and the etalon angle fixed. This keeps the signal cavity mode fixed, and thus the nonresonant wave's frequency (usually the idler) is tuned by an amount equal to the pump frequency change. Continuous tuning ranges exceeding 50 GHz have been obtained in this way. For such pump tuning ranges, the shift of the gain curve, indicated by a double-arrow in Figure 3, is usually sufficiently small that it can be neglected. The use of a widely tuneable pump laser is also possible. A (discontinuous) idler tuning of 9 THz around 3.3 μm has been demonstrated, covered in just 0.3 ms thanks to a large pump tuning rate. In such a case of large pump tuning range, the gain shift is important, causing a significant tuning also of the signal.

The advantages of the basic SRO as compared to other cwOPOs are mainly the simpler tuning behavior, the possibility of continuous tuning without necessitating a tunable pump laser, and the possibility to achieve a very wide tuning range without optics

change, since only the spectral range covered by one of the two parametric waves must be provided as a high reflectivity coating on the cavity mirrors. Moreover, in the simplest implementation, it is possible to operate without any electronic servo control of cavity elements, at the expense of a reduced power and frequency stability. In terms of spectral properties (linewidth, frequency stability, time between undesired mode-hops) the SRO does not offer particular advantages compared to other cwOPO types, even if actively stabilized (see below). The disadvantage of the basic SROs based on Gaussian modes is the high threshold, necessitating expensive high-power (several W) cw lasers. This disadvantage is likely to disappear with progress in laser development. The alternative of waveguide SROs is mentioned below.

PR-SRO

In a PR-SRO, the pump wave is also resonated, leading to a substantial reduction of threshold. For example, it is possible to generate mid-IR radiation (3 μm range), even with much more affordable 1 W-level Nd:YAG pump lasers.

One implementation is the common-cavity PR-SRO (Figure 4). The cavity length is either locked to the pump laser frequency or vice-versa. If a frequency-stable pump laser is used, the first solution is favorable, since then both signal and idler waves have a stability comparable to that of the pump laser. With an etalon (coated so as to be transparent for the pump wave) inside the cavity, mode-hop-free operation for several hours has been demonstrated.

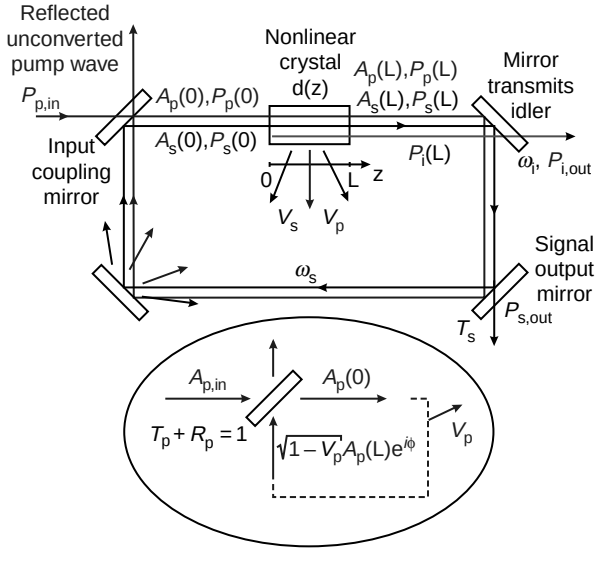


Figure 4 Schematic of a pump-resonant SRO with a common cavity for pump and signal waves. The inset shows the pump wave fields incident on and leaving the input coupling mirror. The arrow symbolizes pump wave loss during the round trip.

Mode-hop tuning in a PR-SRO with cavity-to-laser lock can be achieved:

(i) by tuning the etalon at constant pump frequency: the signal mode-hops from mode to mode, the idler tunes equally but in opposite direction;
(ii) by tuning the pump frequency at constant etalon position: the cavity length must follow and therefore the signal tunes by $\Delta\omega_s = (\omega_s/\omega_p)\Delta\omega_p$, while the idler frequency tunes by $\Delta\omega_i = (\omega_i/\omega_p)\Delta\omega_p$. This continues until the signal must mode-hop back, causing an equal but opposite frequency jump in the idler. The pump tuning is thus completely transferred to the idler.

Continuous signal and idler tuning is performed by tuning the pump frequency and synchronously adjusting the etalon angle. Thus, a limitation of the common-cavity PR-SRO concept is that continuous tunability necessitates a tunable pump laser.

In a dual-cavity, PR-SRO pump and signal waves resonate in separate cavities. The pump cavity is usually stabilized to the pump frequency and an etalon is placed into the signal cavity for better stability against mode-hops. This configuration combines the tuning simplicity of a basic SRO with a low-threshold power. For example, a threshold for a 1 μm-pumped PPLN dual-cavity PR-SRO at the 0.4 W level was obtained.

Mode-hop tuning (in opposite directions) of the frequencies of both parametric waves in steps of the cavity's FSR can be performed by changing the etalon angle (Figure 3). A range exceeding 50 GHz (at constant output power) has been achieved. Continuous signal and idler wave tuning (by equal but opposite amounts) is possible by changing the signal cavity length (Figure 3). However, the range is limited to the FSR of the cavity, typically on the order of 500 MHz, then the frequencies mode-hop back. It is expected that wider continuous tuning should be possible by synchronous etalon and cavity length scan. If a tuneable pump laser is employed, continuous tuning of only the idler frequency is possible by keeping the etalon and signal wave cavity length fixed; the pump tuning range is fully transferred to the idler.

DRO

In contrast to an SRO, this device type requires both parametrically generated waves to be resonant in the cavity. Here we consider only DROs with a common cavity for both signal and idler waves (dual-cavity DROs have also been demonstrated).

Because of the added resonance condition, mode-hops of the oscillating mode pair occur already upon a very small change (~1 nm) in the cavity length. Under normal environmental conditions such changes are unavoidable and therefore a 'free-running' DRO does not produce frequency-stable output. However, by active control of one additional parameter of the OPO system the mode-hops can be effectively suppressed. A practical solution is to actively control the length of the cavity via the position a of one of the cavity mirrors so that a particular mode pair, q_s, q_i, continues to oscillate; this is achieved by maximizing the OPO output power or by minimizing the detuning of one of the two resonant waves from its respective cavity mode frequency. This control system ensures that energy conservation and the two resonance conditions are simultaneously satisfied:

$$\begin{aligned} &\omega_s + \omega_i = \omega_p \\ &\omega_s = 2\pi c q_s/\mathrm{OPL}(d, \omega_s, T) \\ &\omega_i = 2\pi c q_i/\mathrm{OPL}(d, \omega_i, T) \end{aligned} \qquad [7]$$

Experimentally, mode-hop-free oscillation exceeding 10 h was achieved by such active stabilization.

Another system parameter may then be used to tune both signal and idler frequencies. The tuning coefficient for the crystal temperature T is small, but tuning the pump frequency is an efficient way. The tuning coefficient is approximately $\Delta\omega_s = (\omega_s/\omega_p)\Delta\omega_p$, and accordingly for ω_i.

Continuous tuning of signal and idler frequencies over 10 GHz by pump tuning was achieved. This range was limited by the available continuous tuning range of the pump laser. Mode-hop tuning can be induced by mechanical perturbations: by tapping on the DRO cavity, mode-hop tuning over 220 GHz has been demonstrated.

Linewidth and Frequency Stability

For high-resolution spectroscopy a narrow linewidth for signal or idler is desirable. One contribution to the linewidth of a cwOPO arises from the unavoidable spontaneous parametric fluorescence, similar to the Schawlow–Townes limit in lasers due to spontaneous emission. This contribution is negligible if the OPO operates at power levels of practical interest. Thus, other sources of classical noise dominate the linewidth of cwOPOs. As discussed above, the frequency of cwOPOs depends (on a fine scale) on pump frequency and cavity optical path length. Both parameters fluctuate: the pump laser has a finite linewidth as well as frequency jitter, and the cavity optical path length fluctuates due to acoustic noise affecting the mirror positions, to air pressure fluctuations, temperature fluctuations of the crystal and of the cavity structure.

With careful construction and shielding, many of these noise sources can be minimized. Furthermore, it is highly favorable to use pump lasers of narrow linewidth and high intrinsic frequency stability (e.g., diode-pumped monolithic solid-state lasers), or actively frequency-stabilized lasers. Note that this special requirement on the pump source does not exist for lasers.

In an SRO, the linewidth and frequency stability of the signal are mostly determined by the stability properties of the cavity. The idler spectral properties then follow entirely from the signal and pump spectral properties through $\omega_i = \omega_p - \omega_s$. Since the idler is usually the wave of interest, this implies that the SRO cavity and the pump laser must both have a stability appropriate to the spectroscopic application.

If the SRO cavity is not sufficiently stable (e.g., because of insufficient temperature stability of the oven containing the nonlinear crystal), the frequency fluctuation/drift of the signal and idler can be substantial, tens to hundreds of MHz within minutes being typical.

To avoid mode-hops also on long time-scales, the cavity length (or the etalon angle) can be actively stabilized so as to maximize the idler's output power. This ensures that the cavity mode position and the frequency of maximum difference between gain and loss are kept equal.

If a frequency-stable pump wave is used, it is possible to transfer these characteristics by stabilizing the SRO cavity length to the pump wave frequency. In an SRO without resonant pump one can make use of the residual reflection of the pump wave by the cavity mirrors that leads to interference with the pump wave reflected from the pump input mirror. This was implemented in a SRO pumped by a high-power diode-pumped internally doubled Nd:YVO_4 laser (532 nm) of high intrinsic frequency stability. Signal and idler drifts of less than 50 MHz/h were achieved. At the same time, a very small 20 kHz linewidth of the (resonant) signal radiation was obtained. Similar levels were obtained for a PR-SRO with a common cavity for pump and signal wave.

In a PR-SRO with dual-cavity configuration, the same requirements of minimization of disturbances to the parametric wave's cavity arise as in a basic SRO. Using a highly frequency-stable kHz linewidth monolithic pump laser, together with the above-mentioned idler power maximization control, frequency drift values for both parametric waves at the

level of 30 MHz/h were obtained and a short-term linewidth below 6 kHz.

In a single-cavity DRO, the linewidth and frequency stability of the parametric waves are essentially determined by the pump wave properties (see discussion in the section on DRO above). Here, too, the use of narrow-linewidth, high-frequency-stability pump lasers has led to excellent signal/idler characteristics. For example, less than 40 kHz linewidth and <40 MHz per hour drift were obtained with a DRO pumped by an externally doubled monolithic Nd:YAG laser of 10 kHz linewidth.

If further narrowing of linewidth or reduction of frequency drift is desired, the OPO output frequencies can be stabilized to a stable cavity, atomic, or molecular transition. Locking to a cavity has been demonstrated for a DRO.

Applications

CwOPOs are most interesting as sources of radiation in the IR range beyond 2 μm, especially in the 3–5 μm range, where vibrational transitions of many molecules occur. In comparison to other sources of coherent radiation (color center laser, quantum cascade laser, lead-salt diode laser, CO overtone laser, difference frequency generation), cwOPOs provide a unique combination of desirable features.

A simple application of cwOPOs would be as a source providing a power level/output wavelength/mode quality combination not available from other sources. Possible applications are vibrational spectroscopy in the condensed phase, where a narrow linewidth is not required. The excellent beam quality provides a high spatial resolution.

Concerning high-resolution spectroscopy (mostly in the gas phase), methods demonstrated in conjunction with cwOPOs include:

- Absorption spectroscopy;
- Doppler-free saturation spectroscopy;
- Photo-acoustic spectroscopy (PAS);
- Hole-burning spectroscopy;
- Coherent atomic spectroscopy;
- Cw cavity ring down spectroscopy (cw-CRDS).

Trace gas detection using PAS or cw-CRDS is a particular application where cwOPOs have reached excellent sensitivity (minimum detectable concentrations as low as 1 part in 10^{11}). The high output power and/or the narrow linewidth are specific features that make such sensitivity possible.

Quantum-Optical Properties

The pairwise production of signal and idler photons is at the heart of the parametric process. This perfect quantum correlation is to a certain extent also maintained in the output waves of a cwOPO. In particular, the intensity fluctuations of signal and idler waves emitted by a DRO are strongly (albeit never perfectly) quantum correlated if their output coupler transmissions significantly exceed their respective roundtrip losses. For these 'twin beams' the spectral density of the fluctuations of their power difference can be less than the spectral density of the fluctuations of the power of an individual wave (squeezing). Making use of this correlation permits to increase the obtainable sensitivity in spectroscopic measurements for a given power level of the wave interacting with the sample.

Device Development Trends

CwOPOs using solid-state pump lasers have become commercially available. In the near future, some likely developments will be miniaturization and optimization in order to reduce the influence of perturbations, further extension of continuous and mode-hop tuning ranges and the wider use of diode lasers as pump lasers, leading to cost reduction. To access a larger range of molecules in spectroscopy, extension of the emission range well beyond 5 μm is also of significant interest. Novel nonlinear materials and pump lasers of wavelength longer than the often employed Nd:YAG lasers will become of importance in this development.

Two specific lines of development complementary to the bulk cwOPOs described above are waveguide and intracavity cwOPOs (ICSRO). If pump, signal and idler waves are confined in a waveguide (*see* **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguides) the parametric gain coefficient is a few orders higher than for free-space Gaussian waves, due to the small mode cross-section and the scaling with the square of the crystal length, compared with the linear scaling in case of Gaussian beams (see eqn [36]). Values are on the order of $E \sim 1/W$ at $\lambda_p = 1.5$ μm leading to low thresholds even for SROs, with pump lasers of comparatively long wavelength. For example, in a SRO consisting of a 9 cm long titanium:PPLN waveguide, a threshold below 300 mW for a 1.5 μm pump was achieved. The idler emission range was 3.1–3.4 μm. The potential for mass production and, therefore, cost minimization is clear.

An ICSRO consists of an SRO internal to the cavity of a cw laser, where the high circulating laser power is taken advantage of for pumping. An ICSRO is thus similar to a PR-SRO but there is no need to implement a control system for the frequency lock between pump laser and pump cavity nor to mode-match the pump wave to the SRO cavity mode.

Appendix A

Three-Wave Mixing in a $\chi^{(2)}$ Medium with Focused Waves

The wave equation in a nonlinear medium reads

$$\Delta\vec{E} - \mu_0\mu_r\frac{\partial^2}{\partial t^2}\vec{D}_{\mathrm{L}} = \mu_0\mu_r\frac{\partial^2}{\partial t^2}\vec{P}_{\mathrm{NL}} \qquad [8]$$

The displacement field $\vec{D}_{\mathrm{L}}$ is related to the electric field $\vec{E}$ by linear response. $\vec{P}_{\mathrm{NL}}$ is the nonlinear polarization. We deal with nonmagnetic media ($\mu_r = 1$) and choose the principal axis system as coordinate system. The Fourier transform of, for example, the x-component of the wave equation then reads:

$$\begin{aligned}\Delta E_x(\omega,\vec{r}) + \mu_0\omega^2\varepsilon_0\varepsilon_x(\omega)E_x(\omega,\vec{r}) \\ = -\mu_0\omega^2 P_{\mathrm{NL},x}(\omega,\vec{r})\end{aligned} \qquad [9]$$

In the following, we take into account only second-order nonlinear effects, so the nonlinear polarization reduces to $P_{\mathrm{NL}} = P^{(2)}$. The fields to be considered are monochromatic, propagate in the z-direction and are linearly polarized along x. The notation can then be simplified to:

$$E_x(\omega,\vec{r}) = E(\vec{r})\mathrm{e}^{\mathrm{i}kz}, \quad P_{\mathrm{NL},x}(\omega,\vec{r}) = P^{(2)}_{\omega}(\vec{r}) \qquad [10]$$

with the wavevector $k = \omega n(\omega)/c$ and the refractive index for the polarization under consideration $n^2_{\omega} = n(\omega)^2 = E_x(\omega)$. The electric field is obtained as $\vec{E}(t,\vec{r}) = \frac{1}{2}\mathrm{Re}(E(\vec{r})\mathrm{e}^{\mathrm{i}(kz-\omega t)})\hat{x}$.

Since the parametric gains are small in cw conversion, the derivative of the envelope $E(\vec{r})$ changes only little over a propagation distance on the order of a wavelength $2\pi/k$. Inserting eqn [9] into eqns [6] and [8] and neglecting the term $\sim \mathrm{d}^2E(\vec{r})/\mathrm{d}z^2$, yields the paraxial or slowly varying envelope wave equation:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + 2\mathrm{i}k\frac{\partial}{\partial z}\right)E(\vec{r}) = -\mu_0\omega^2 P^{(2)}_{\omega}(\vec{r})\mathrm{e}^{-\mathrm{i}kz} \qquad [11]$$

The electric field envelope can be expanded in a complete set of Gaussian TEM_{mn} – mode functions ψ_{mn}:

$$E(\vec{r}) = \sqrt{\frac{\omega}{n_{\omega}}}\sum_{m,n=0}^{\infty} A_{mn}(z)\psi_{mn}(\vec{r}) \qquad [12]$$

These mode functions are defined to satisfy the paraxial wave equation in absence of nonlinearity (vanishing right-hand side in eqn [11]) and the orthonormality properties

$$\int\int_{-\infty}^{+\infty}\psi^*_{mn}(\vec{r})\psi_{m'n'}(\vec{r})\mathrm{d}x\,\mathrm{d}y = \delta_{mm'}\delta_{nn'} \qquad [13]$$

A particular mode function set is characterized by position (along z) and size of the waist, which can be chosen arbitrarily. The functions $A_{mn}(z)$ are slowly varying amplitude functions. The power in a particular mode can be calculated from:

$$P_{mn}(z) = \frac{\omega|A_{mn}(z)|^2}{2\mu_0 c} \qquad [14]$$

where c is the speed of light *in vacuo*. Inserting eqn [12] into eqn [11], multiplication with $\psi^*_{m'n'}(\vec{r})$, and integration over x, y yields:

$$\frac{\mathrm{d}A_{m'n'}(z)}{\mathrm{d}z} = \mathrm{i}\frac{\mu_0 c}{2}\sqrt{\frac{\omega}{n_{\omega}}}\mathrm{e}^{-\mathrm{i}kz}\int\int_{-\infty}^{\infty}P^{(2)}_{\omega}(\vec{r})\psi^*_{m'n'}(\vec{r})\mathrm{d}x\,\mathrm{d}y \qquad [15]$$

If all electric fields are linearly polarized, the Fourier amplitudes of the nonlinear polarization can be written as:

$$\begin{aligned}P^{(2)}_{\omega_{\mathrm{s}}}(\vec{r}) &= 2\varepsilon_0 d(z)E_{\omega_{\mathrm{p}}}(\vec{r})E^*_{\omega_{\mathrm{i}}}(\vec{r}) \\ P^{(2)}_{\omega_{\mathrm{i}}}(\vec{r}) &= 2\varepsilon_0 d(z)E_{\omega_{\mathrm{p}}}(\vec{r})E^*_{\omega_{\mathrm{s}}}(\vec{r}) \\ P^{(2)}_{\omega_{\mathrm{p}}}(\vec{r}) &= 2\varepsilon_0 d(z)E_{\omega_{\mathrm{s}}}(\vec{r})E_{\omega_{\mathrm{i}}}(\vec{r})\end{aligned} \qquad [16]$$

where d is the tensor element appropriate to the particular combination of linear polarizations of the three waves. (In eqn [16], $\omega_{\mathrm{i}} + \omega_{\mathrm{s}} = \omega_{\mathrm{p}}$.)

The mode amplitude equations for the three interacting waves follow from eqns [10], [15], and [16] as:

$$\begin{aligned}\frac{\mathrm{d}A_{mn}\begin{Bmatrix}\mathrm{p}\\ \mathrm{s}\\ \mathrm{i}\end{Bmatrix}(z)}{\mathrm{d}z} &= \mathrm{i}\frac{d(z)}{c}\begin{Bmatrix}\sqrt{\omega_{\mathrm{p}}/n_{\mathrm{p}}}\mathrm{e}^{-\mathrm{i}\Delta kz}\\ \sqrt{\omega_{\mathrm{s}}/n_{\mathrm{s}}}\mathrm{e}^{+\mathrm{i}\Delta kz}\\ \sqrt{\omega_{\mathrm{i}}/n_{\mathrm{i}}}\mathrm{e}^{+\mathrm{i}\Delta kz}\end{Bmatrix}\int\int\mathrm{d}x\,\mathrm{d}y \\ &\quad\times\begin{Bmatrix}\mathrm{E_s}(\vec{r})\mathrm{E_i}(\vec{r})\\ \mathrm{E_p}(\vec{r})\mathrm{E_i}(\vec{r})^*\\ \mathrm{E_p}(\vec{r})\mathrm{E_s}(\vec{r})^*\end{Bmatrix}\psi_{mn}\begin{Bmatrix}\mathrm{p}\\ \mathrm{s}\\ \mathrm{i}\end{Bmatrix}(\vec{r})^{*+\infty}\end{aligned} \qquad [17]$$

The quantity $\Delta k = k_{\mathrm{p}} - k_{\mathrm{s}} - k_{\mathrm{i}}$ is the wavevector mismatch. Here, the envelope of each wave has an expansion of the form in eqn [12] with its respective (independent) set of mode functions.

Appendix B

Steady-State Description of the cwOPO

The general eqns [17] are applicable to any type of cwOPO. Below and at threshold at least two of the three waves have given modes, i.e., the corresponding function $E(\vec{r})$ is known up to a slowly z-dependent amplitude: the resonant parametric wave has a mode defined by the cavity geometry and mirrors, while the pump wave mode is defined by the used focusing optics. This knowledge allows to obtain an analytic expression for the threshold of any OPO type and for an arbitrary cavity.

Above threshold two cases can be distinguished:

(i) in a basic SRO the pump mode itself becomes a function of pump power, due to its depletion by signal and idler wave, and only a numerical solution can be given; and
(ii) if two or more waves resonate in one or more cavities (DRO, PR-SRO, TRO), an analytic solution is possible even above threshold.

In the following, the PR-SRO is treated. Figure 4 shows a schematic of a common cavity PR-SRO and the notation; the treatment below also holds for the dual cavity PR-SRO. Without loss of generality, the signal wave is taken to be the signal. For the sake of simplicity, below a further assumption about the spatial mode of the idler will be made.

The pump and signal waves are given by TEM_{00} modes of the cavity (or cavities):

$$E_p(\vec{r}) = \sqrt{\frac{\omega_p}{n_p}} A_{00p}(z)\psi_{00p}(\vec{r})$$
$$E_s(\vec{r}) = \sqrt{\frac{\omega_s}{n_l}} A_{00s}(z)\psi_{00s}(\vec{r}) \qquad [18]$$

In the case of a common-cavity PR-SRO, pump and signal resonate in the same cavity and therefore their modes have the same waist position and Rayleigh ranges $z_R = kw^2/2$, (w is the waist). In the case of a dual-cavity PR-SRO the modes are in principle independent.

Although the idler wave is in general not purely TEM_{00}, it may be approximated by a TEM_{00} mode, whose mode parameters are determined by an optimization criterion. We can then introduce the overlap function:

$$O(z) = \int\int dx\, dy(\psi_{00p}(\vec{r}))^* \psi_{00s}(\vec{r})\psi_{00i}(\vec{r})_{-\infty}^{+\infty} \qquad [19]$$

Under these assumptions, the infinite set of mode eqns [17] reduces to just three coupled equations ($A \equiv A_{00}$):

$$\frac{dA_p(z)}{dz} = -i\xi(z)A_s(z)A_i(z)e^{-i\Delta kz}O(z) \qquad [20]$$

$$\frac{dA_s(z)}{dz} = -i\xi(z)A_p(z)A_i^*(z)e^{i\Delta kz}O(z)^* \qquad [21]$$

$$\frac{dA_i(z)}{dz} = -i\xi(z)A_p(z)A_s^*(z)e^{i\Delta kz}O(z)^* \qquad [22]$$

where

$$\xi(z) = \frac{d(z)}{c}\sqrt{\frac{\omega_p\omega_s\omega_i}{n_p n_s n_i}} \qquad [23]$$

Two phasematching cases have to be considered: for birefringent phasematching, $d = \text{const}$ and ξ is independent of z. In QPM, the d-coefficient is modulated periodically along z with period Λ. It suffices to consider only the particular term of its Fourier series expansion that gives the smallest wavevector quasi-mismatch. When only the first term is relevant, this is referred to as first-order QPM and $d(z) \simeq \frac{1}{2}\hat{d}(g_1 e^{i2\pi z/\Lambda} + \text{c.c.})$. This case can be taken into account by replacing $d(z)$ in eqn [23] by $d_{eff} = \frac{1}{2}\hat{d}g_1$, and in eqns [20]–[22] the relevant wavevector mismatch becomes $\Delta k = k_p - k_s - k_i \pm 2\pi/\Lambda$, where, for usual materials, the minus sign applies.

We solve eqns [20]–[22] by expanding in powers of the nonlinearity ξ. This is appropriate due to the small gain and implies that the resonant pump and signal wave amplitudes do not change substantially (relative to the respective average amplitude) along the crystal length and around the cavity.

Integrating eqn [22] after approximating the amplitudes on the right-hand side by their values at $z = 0$ yields:

$$A_i(z) = -i\xi z A_p(0)A_s^*(0)I^*(z)$$
$$I(z) = \frac{1}{z}\int_0^z dz' O(z')e^{-i\Delta kz'} \qquad [24]$$

We insert this result into eqns [20] and [21] and approximate $A_s(z)$ and $A_p(z)$ by their values at $z = 0$. The result is

$$A_p(z) = A_p(0) - \frac{1}{2}(\xi z)^2 D^*(z)A_p(0)|A_s(0)|^2 \qquad [25]$$

$$A_s(z) = A_s(0) + \frac{1}{2}(\xi z)^2 D(z)A_s(0)|A_p(0)|^2 \qquad [26]$$

$$D(z) = \frac{2}{z^2}\int_0^z \mathrm{d}z' z' I(z') O(z')^* \, \mathrm{e}^{\mathrm{i}\Delta k z'} \qquad [27]$$

This completes the calculation of the propagation through the nonlinear medium. Now we proceed to calculate the propagation around the cavity.

Since the signal wave is resonant with the cavity, we can immediately write the self-consistency equation for the roundtrip:

$$|A_s(0)| = \sqrt{R_s(1 - V_s)}|A_s(L)| \qquad [28]$$

The roundtrip cavity reflectivity and roundtrip loss for the signal wave are denoted by R_s and V_s. At the input mirror (see lower part of Figure 4), characterized by a transmission $T_p = 1 - R_p$, the interference of the injected pump wave $A_{p,in}$ and the internally circulating pump wave must be considered:

$$A_p(0) = \sqrt{T_p}A_{p,in} + \sqrt{R_p(1 - V_p)}A_p(L)\mathrm{e}^{\mathrm{i}\phi} \qquad [29]$$

V_p is the pump roundtrip loss. Evaluating eqn [26] at $z = L$ and comparing with eqn [28] yields:

$$|A_p(0)|^2 = \left[\frac{1}{\sqrt{R_s(1 - V_s)}} - 1\right]\frac{1}{\frac{1}{2}(\xi L)^2 \mathrm{Re}\, D(L)} \qquad [30]$$

Here the imaginary part of $D(z)$ was neglected since $|\xi L A_p(0)|^2 \ll 1$. Expression [30] shows that the intracavity pump power is clamped at a fixed value, equal to the value at threshold, independent of the input power.

In eqn [29] we can assume $A_{p,in}$ to be real without loss of generality, yielding the relation:

$$\sqrt{T_p}\frac{A_{p,in}}{A_p(0)} = 1 - \sqrt{R_p(1 - V_p)}\mathrm{e}^{\mathrm{i}\phi}\frac{A_p(L)}{A_p(0)} \qquad [31]$$

where $\phi = \omega_p \mathrm{OPL}(\omega_p, T)/c$. The pump wave is assumed to be exactly resonant; this means that $A_p(0)$ is maximum. The phase ϕ must therefore be such that:

$$\mathrm{e}^{\mathrm{i}\phi}\frac{A_p(L)}{|A_p(0)|} \in \mathbb{R} \qquad [32]$$

Together, with eqn [25], we obtain:

$$\left|\frac{A_p(L)}{A_p(0)}\right|^2 = \left|1 - \frac{1}{2}(\xi L)^2 D^*(L)|A_s(0)|^2\right|^2 \qquad [33]$$

The right-hand side is evaluated neglecting the contribution proportional to ξ^4; inserting the result in eqns [31] and [32] gives:

$$\sqrt{T_p}\frac{A_{p,in}}{|A_p(0)|} = 1 - \sqrt{R_p(1 - V_p)} + \sqrt{R_p(1 - V_p)}\frac{1}{2}(\xi L)^2 \mathrm{Re}\, D|A_s(0)|^2 \qquad [34]$$

Equations [30] and [34] represent the steady-state solution of the PR-SRO: they give the circulating pump and signal powers, from which all desired quantities can be computed.

The external threshold is found by setting $A_s(0) = 0$ and assuming small losses, T_p, $V_p \ll 1$

$$P_{p,in}^{th} \simeq \frac{T_p + V_p}{T_p}\cdot(T_p + V_p)\cdot\frac{(T_s + V_s)}{E} \qquad [35]$$

with the nonlinearity

$$E = \frac{2\mu_0 d^2}{\pi c^2}\frac{\omega_s \omega_i}{n_p n_s n_i}L^2 \,\mathrm{Re}\, D(L) \qquad [36]$$

Note that $\mathrm{Re}\, D(L) = |I(L)|^2$. The second factor in eqn [35] is responsible for the reduction of threshold due to pump resonance. The last factor is the internal pump threshold and constant circulating pump power level:

$$P_p(0) \simeq \frac{(T_s + V_s)}{E} \qquad [37]$$

which is also equal to the threshold of a basic SRO. The nonlinearity may be written as

$$E \simeq \frac{4\mu_0 d^2}{\pi c^2}\frac{\omega_s^2 \omega_i^2}{n_p^2 \omega_p} L h_{PA} \qquad [38]$$

with the dimensionless focusing function h_{PA}. A detailed calculation shows that the value of h_{PA} cannot exceed 1.1 and is close to its maximum when the phase mismatch $\Delta k L$ is approximately zero and the pump and signal focusing by the cavity is such that their Rayleigh range $z_R \simeq L/2$.

The OPO will select the signal cavity mode frequency ω_s and idler spatial mode that minimizes the threshold eqn [35]. Note that this is not necessarily equivalent to a maximization of E, since T_s and V_s may also be dependent on signal frequency, for example, because of the presence of an etalon inside the cavity.

The circulating signal power as derived from eqns [30] and [34], is given in eqn [6].

The signal power extracted from the cavity is $P_{s,out} = T_s P_s(L) \simeq T_s P_s(0)$, while the (fully extracted) idler power is given by:

$$P_{i,out} = P_i(L) = \frac{\omega_i}{\omega_s} E P_s(0) P_p(0) \qquad [39]$$

See also

Nonlinear Optics, Applications: Phase Matching. **Optical Parametric Devices:** Overview.

Further Reading

Breitenbach G, Schiller S and Mlynek J (1995) 81% conversion efficiency in frequency-stable continuous-wave parametric oscillation. *Journal of the Optical Society of America B* 12: 2095–2101.

Ebrahimzadeh M and Dunn MH (2000) Optical parametric oscillators. In: Bass M, Enoch JM, van Stryland EW and Wolfe WL (eds) *Handbook of Optics IV*, pp. 2201–2272. New York: McGraw-Hill.

Ebrahimzadeh M, Turnbull GA, Edwards TJ, Stothard DJM, Lindsay ID and Dunn MH (1999) *Journal of the Optical Society of America B* 16: 1499–1511.

Gibson GM, Dunn MH and Padgett MJ (1998) Application of a continuously tunable, cw optical parametric oscillator for high-resolution spectroscopy. *Optics Letters* 22: 40–43.

Klein ME, Laue CK, Lee D-H, Boller K-J and Wallenstein R (2000) Diode-pumped singly resonant continuous-wave optical parametric oscillator with wide continuous tuning of the near-infrared idler wave. *Optics Letters* 25: 490–493.

Klein ME, Gross P, Boller K-J, *et al.* (2003) Rapidly tunable continuous-wave optical parametric oscillator pumped by a fiber laser. *Optics Letters* 28: 920–923.

Müller F, Popp A, Kühnemann F and Schiller S (2003) Transportable, highly sensitive photoacoustic spectrometer based on a continuous-wave dual-cavity optical parametric oscillator. *Optics Express* 11: 2820–2825.

Schiller S, Schneider K and Mlynek J (1999) *Journal of the Optical Society of America B* 16: 1512–1524.

Schreiber G, Hofmann D and Grundkötter W, *et al.* (2001) Nonlinear integrated optical frequency converters with periodically poled Ti:$LiNbO_3$ waveguides. *Proceeding SPIE* 4277: 144.

Special issue (1998) Continuous-wave optical parametric oscillators: materials, devices, and applications. *Applied Physics B* 66(6).

Strössner U, Meyn J-P, Wallenstein R, *et al.* (2002) Single-frequency continuous-wave optical parametric oscillator system with an ultrawide tuning range of 550 to 2830 nm. *Journal of the Optical Society of America B* 19: 1419–1424.

Yariv A and Yeh P (2002) *Optical Waves in Crystals*. Hoboken, NJ: Wiley-Interscience.

Optical Parametric Oscillators (Pulsed)

H Giessen and X P Zhang, University of Bonn, Bonn, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Advantages and Applications of Ultrashort Pulsed OPOs

The advantages of ultrashort pulsed optical parametric oscillators (OPOs) are twofold: on the one hand, it is possible to generate tunable laser radiation with a pulse duration in the femtosecond and picosecond time domain; on the other hand, the short pulses give rise to a high peak intensity of the electric field of the pump wave inside the nonlinear medium. Due to the fact that the OPO single-path gain coefficient reads as:

$$\eta = \frac{P_{\text{Signal}}}{P_{\text{Pump}}} = \frac{8\pi^2 d_{\text{eff}}^2 L^2 I_{\text{P}}}{\varepsilon_0 n_{\text{P}} n_{\text{S}} n_{\text{I}} c \lambda_{\text{S}}^2} \frac{\sin^2(\Delta k L/2)}{(\Delta k L/2)^2} \qquad [1]$$

where I_{P} is the pump intensity, Δk is the phase mismatch between pump, signal, and idler, n_{P}, n_{S}, n_{I}, are the refractive indices at the pump, signal, and idler wavelengths, respectively, and L is the length of the nonlinear medium with an effective nonlinear coefficient d_{eff}, the efficiency grows proportional to the intensity of the pump.

Ultrashort pulsed OPOs can provide widely tunable femtosecond or picosecond pulses, covering the spectral range from the visible to the mid-infrared, which have many applications, such as time-resolved spectroscopy, pump-probe measurements, semiconductor analysis, photochemistry, optical ranging, data transfer, and data processing with high carrier frequency, white-light generation, etc.

Synchronous Pumping

The key issue in operating ultrashort-pulse OPOs is synchronization between the pump pulses and the signal pulses. Consider the following sequence: a pump pulse enters the OPO cavity and generates spontaneously downconverted light at the signal wavelength (see Figure 1, inset). These signal photons travel inside the OPO cavity, entering the nonlinear medium again after exactly one cavity round trip. This light will act as seed light for another downconversion process if a second pump pulse enters the nonlinear medium at exactly the same instance (see Figure 1). The presence of the signal pulse will lead to a much more efficient stimulated downconversion process.

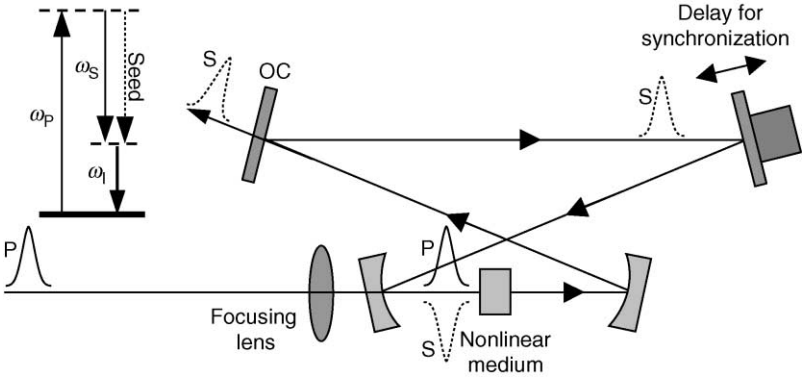


Figure 1 Synchronous pumping of a typical ring-cavity OPO: P, pump pulses; S, signal pulses; OC, output coupler; ω_P, ω_S, ω_I, frequencies of the pump, the signal, and the idler, respectively. Inset: energy level diagram for optical parametric oscillation.

This requires the repetition rate of the pump oscillator and the OPO to be synchronized precisely. Therefore, this pump scheme is called *synchronized pumping*. If the temporal separation between the pump pulses does not exactly equal the roundtrip time of the signal pulses inside the OPO cavity, the generation efficiency will decrease dramatically as the temporal pulse overlap is reduced.

Dispersion in Nonlinear Media and Optical Glasses

Maintaining synchronicity between the pump pulses and the signal pulses is a necessary, but not sufficient prerequisite for successful operation of an ultrashort-pulse OPO. The signal pulses will be temporally broadened during their repeated trip through the nonlinear medium. This will lead to problems, as the newly generated signal pulses are still rather short (as they have been generated by a short pump pulse), but the previously generated signal pulses, which should act as seed pulses and ideally have a similar pulse duration as the pump pulses, have already been broadened temporally, due to dispersion in the nonlinear medium and other intracavity optical elements. Meanwhile, temporal walk-off between the signal (idler) and the pump pulses will take place, resulting in reduced conversion efficiency and broadened signal pulses. Group velocity dispersion (GVD) and group velocity mismatch (GVM) are normally used to characterize the dispersion in nonlinear medium.

The group velocity of the laser pulses propagating in a nonlinear optical medium is defined as:

$$v_G = \frac{c}{n - \lambda \dfrac{dn}{d\lambda}} \qquad [2]$$

where c is the velocity of light in vacuum, n is the refractive index of the nonlinear medium at the center wavelength λ of the laser pulses. Then the GVD can be written as:

$$\frac{dv_G}{d\lambda} = \frac{\lambda v_G^2}{c} \frac{d^2 n}{d\lambda^2} \qquad [3]$$

In most cases, we evaluate the GVD of the optical materials using the group velocity dispersion parameter (k''), which is defined as:

$$k'' = \frac{d^2 k}{d\omega^2} = \frac{\lambda^3}{2\pi c^2} \frac{d^2 n}{d\lambda^2} \qquad [4]$$

where $k = n\omega/c$ is the wave vector, ω is the frequency of the incident wave, and consequently:

$$\frac{dv_G}{d\lambda} = \frac{2\pi c v_G^2}{\lambda^2} \frac{d^2 k}{d\omega^2} \qquad [5]$$

Due to GVD in the nonlinear medium, the spectral components in the broadband spectrum of an ultrashort pulse will travel at different group velocities and will consequently be delayed differently in the crystal, therefore the pulses will become temporally broadened or become chirped. This linear chirp induced pulse broadening effect can be characterized quantitatively by the following formula:

$$\tau_P = \tau_{P0} \sqrt{1 + \left(\frac{z}{z_d}\right)^2} \qquad [6]$$

where τ_{P0} is the incident pulse duration, τ_P is the pulse duration after the pulse propagates in the nonlinear medium over a distance of z, and z_d is called the characteristic length and defined as:

$$z_d = \frac{\tau_{P0}^2}{4 \ln 2 |k''|} \qquad [7]$$

In frequency conversion processes, such as second-harmonic generation (SHG), sum-frequency

generation (SFG), difference frequency generation (DFG), optical parametric oscillation (OPO), or optical parametric amplification (OPA), three waves interact with each other while propagating through the nonlinear medium. Due to dispersion of the material, different waves propagate at different group velocities. This effect is generally called group velocity mismatch. In cases of interaction on the femtosecond scale, the interacting pulses may get separated after propagating some distance in the medium, which means that there is a reduced effective interaction length. The GVM is a quantitative evaluation of this effect, which is defined as:

$$\Delta v_{\mathrm{G}} = \left(\frac{1}{v_{\mathrm{G,i}}} - \frac{1}{v_{\mathrm{G,j}}} \right)^{-1} \qquad [8]$$

where $v_{\mathrm{G,i}}$ and $v_{\mathrm{G,j}}$ are the group velocities of the two interacting waves i and j. The effective interaction length L_{eff} for wave i and wave j can thus be calculated by the following formula:

$$L_{\mathrm{eff}} = \tau_{\mathrm{P}} |\Delta v_{\mathrm{G}}| = \tau_{\mathrm{P}} \left| \left(\frac{1}{v_{\mathrm{G,i}}} - \frac{1}{v_{\mathrm{G,j}}} \right)^{-1} \right| \qquad [9]$$

where τ_{P} is the incident pulse duration at full width at half maximum (FWHM). It is assumed that wave i and wave j have a similar pulse duration. This temporal walk-off will lead to two effects. First, the interacting pulses will get separated and the conversion efficiency will be reduced. The second effect is the broadening of the signal pulses. In the case of sub-100 fs operation, to restrict the broadening in pulse length to less than ~10%, the crystal length should be smaller than $L_{\mathrm{eff}}/2$.

Figure 2 shows the GVD of some typical OPO crystals, including KTP, KTA, RTA, CTA, PPLN, PPKTP, and LBO. Their Sellmeier equations are as summarized in Appendix A. In addition to the wavelength of the incident beam, the GVD is also dependent on the polarization of the incident beams and the type of phase matching.

We give the GVM values between the pump and the signal (P–S), the signal and the idler (S–I), and between the pump and the idler (P–I) for a KTP crystal in Figure 3.

Intracavity Dispersion Compensation and Extra Cavity Pulse Compression

In order to tackle the problem of intracavity dispersion, corrective elements have to be placed inside the cavity in order to compensate the chirp. There have been several solutions to this problem, namely the combination of intracavity prism pairs, chirped mirrors, chirped grating periods (for quasi-phase matching) with oppositely chirped pump pulses, and glass plate dispersion compensators. Certainly, a shorter crystal length will yield a smaller GVM and GVD, but will also give a smaller nonlinear conversion efficiency due to the shorter crystal length (see eqn [1]).

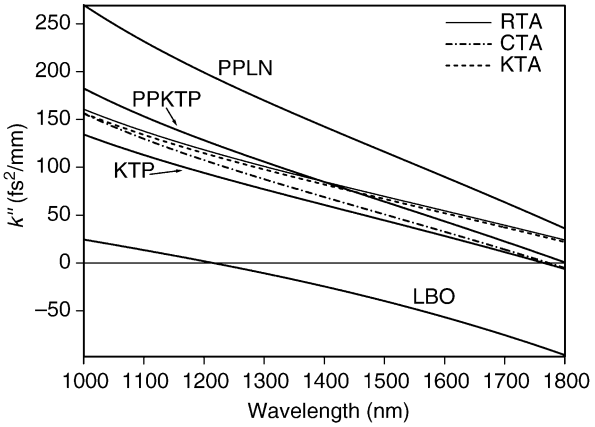


Figure 2 Group velocity dispersion characteristics of some typical OPO crystals.

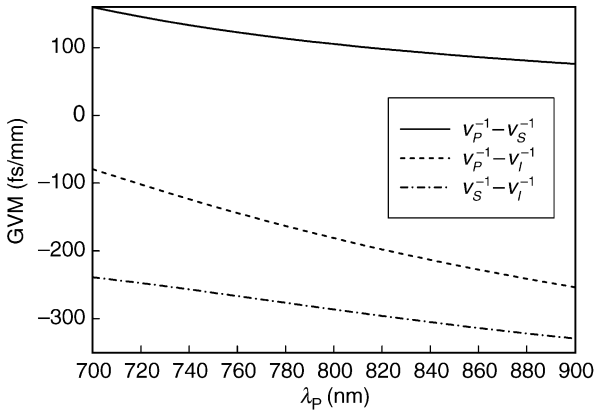


Figure 3 GVM between the interaction waves of a KTP OPO versus pump wavelength (λ_{P}).

Pump Sources

As pump pulses, mode-locked femtosecond or picosecond lasers are used. The requirements on pulse durations are determined by the desired pulse durations of the OPO signal and by the phase-matching bandwidth of the correspondingly used crystal. Shorter pump pulses normally lead to higher pump power thresholds due to shorter temporal overlap between the interacting pulses. It can also lead to higher instability of the OPO operation due to its higher sensitivity to the cavity length fluctuations. Generally, larger pump bandwidth means larger bandwidth of the signal. Consequently, a potentially shorter signal pulse duration can be achieved if the

intracavity dispersion is perfectly compensated. Sometimes it is advantageous for stability to operate the OPO in the slightly negative total GVD regime. In a quasi-phase-matched (QPM) OPO, the phase-matching bandwidth is so large that only a narrow pump bandwidth is required to phase-match a large bandwidth of the signal. Therefore, a pump depletion hole can be commonly observed with the oscillation of the OPO.

Design of Ultrashort Pulsed OPOs

Design of an OPO device may involve the design of the cavity length, cavity mirrors, focal length for focusing the pump beam, intracavity dispersion compensation scheme, frequency stabilization module, crystal length, and some other design depending on the special configuration of the OPO, for example, a heating module (an oven) is required for a PPLN OPO and temperature control module is required for a temperature phase-matched LBO OPO. For an ultrashort pulsed OPO, the most important designs include the design of the crystal length, the design of intracavity dispersion compensation components, and the frequency stabilization control.

Crystal Length

Several factors have to be taken into account when designing a crystal length. The design of a crystal length should be a compromise of all of these considerations. What is special for a femtosecond or picosecond OPO is the dispersion and temporal walk-off between the interacting pulses in the crystal:

1. Effective interaction length: Pump pulse duration, estimated signal pulse duration, temporal walk-off between the pump and the signal give a limit of the crystal length that is characterized by eqn [9].
2. Pulse broadening due to GVM and GVD from the OPO crystal and other intracavity optical glass elements should be estimated. It is reasonable that shorter crystal length is always favorable for ultrashort pulse generation. The crystal length should be as short as possible while maintaining an acceptable conversion efficiency. When designing the crystal length, the available dispersion compensation device should be considered simultaneously.

Dispersion Compensation

For a crystal with a given material and length, the intracavity dispersion can be calculated. The compensation device should introduce an amount of negative GVD such that the total intracavity GVD is equal to or slightly smaller than zero at the desired center wavelength of the signal pulses. Meanwhile, the compensation should be relatively uniform (for instance having a GVD fluctuation of $<50\ \mathrm{fs}^2/\mathrm{mm}$) over the bandwidth of the expected signal spectrum. Then a compensation method (as described above) should be chosen and corresponding device should be designed.

Piezo-Driven Mirror and Frequency Stabilization

To achieve fine tuning of the cavity length, one of the cavity mirrors should be mounted onto a piezoelectric translation stage. When feeding back the signal output of the OPO to this piezo-driven mirror, forming a closed loop to compensate for the drift of the cavity length, a stabilization of the OPO repetition rate is achieved. If the signal is dispersed by a grating, a change in cavity length will also lead to a change in the output spectrum. Using a split photodiode, the difference signal of the dispersed spectrum can be utilized as the measured signal in the feedback loop.

Tuning Characteristics and Methods

Tunability is one of the most important advantages of the OPO over other ultrashort pulsed lasers. A large variety of tuning methods make the OPO more practical and versatile in many applications. For birefringently phase-matched OPOs, the tuning of the signal wavelength can be achieved through changing the pump wavelength (pump wavelength tuning), using an intracavity wavelength selection element (such as a birefringent filter), changing the temperature of the crystal when its refractive index is sensitive to the temperature change (temperature tuning), or changing the orientation of the OPO crystal (angle tuning). Some more methods can be utilized for a QPM OPO, for example through changing the poling period of the crystal (grating period tuning). A special method for an ultrashort pulsed OPO is the so-called cavity length tuning.

As has been discussed above, in synchronous pumping, the OPO must have the same cavity length as the pump laser, so that the signal pulse arrives simultaneously with the pump pulse at the crystal to get the most efficient interaction. Because of dispersion of the nonlinear crystal, the signal will be delayed differently at different wavelengths. In other words, signal pulses have different round-trip times in the cavity at different center wavelengths. When the cavity length of the OPO is changed slightly, the signal has to change its wavelength to keep constant optical path length or constant round-trip time. Cavity length tuning can be achieved easily by tuning

the voltage on a piezoelectric transducer (PZT), onto which one of the cavity mirrors is mounted. Therefore, this method is the most common and frequently used to tune the signal wavelength of a synchronously pumped OPO. This is also the basis for stabilizing the operation of a synchronously pumped OPO.

High-Repetition-Rate (HRR) Operation

Synchronously pumping an OPO at high repetition rate produces a high-repetition-rate signal pulse train which extends the application fields of the OPO. Using high-repetition-rate laser pulses (e.g., >1 GHz) as the pump is a straightforward method to run the OPO at a high repetition rate. But a high repetition rate means an extremely short cavity length. For instance, the cavity length is about 30 cm at 1 GHz repetition rate. Such a short cavity length leads to a large mode of the signal beam waist in the crystal. Mode matching between the signal and the pump beams requires that the transverse mode of the pump beam should be slightly larger than that of the signal beam within the interaction zone to maintain stable and efficient oscillation of the OPO. Large modes of the signal and the pump beams mean low interacting intensities, increasing the pump threshold significantly. Using smaller curvature radius of the curved mirrors and shorter focal length of the focusing lens might reduce the mode sizes at the focus in the nonlinear medium, but this will introduce extremely serious astigmatism when the signal beam is reflected from a strongly tilted curved mirror with a quite small curvature radius. In addition to exact synchronous pumping, there are at least two methods to realize high-repetition-rate operation of an OPO. First, pumping at low repetition rate while using a shorter cavity length of the OPO (L_{OPO}) than the pump laser (L_P) with $L_{OPO} = L_P m/N$, where m and N are both integers and $m < N$, as demonstrated in Figure 4a, which is called lower-order synchronous pumping. However, applying this technique, the signal pulses experience more round-trip losses, also resulting in a much higher pump threshold and periodically varying intensities of the output pulses. Second, pumping the OPO at high repetition rate while using a much longer cavity length of the OPO than the pump laser with $L_{OPO} = L_P N/m$, where m and N are both integers and $m < N$, is called higher-order pumping, as demonstrated in Figure 4b. In higher-order synchronous pumping, the signal pulse train has the same repetition rate as the pump, while the mode-matching problem is removed through a long OPO cavity.

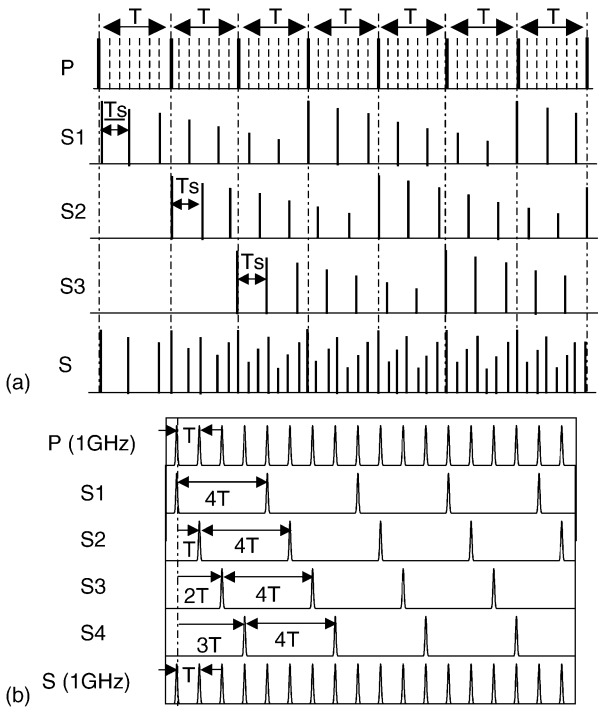


Figure 4 (a) 3/7-order synchronous pumping. Three signal pulse trains (S1–S3) are produced, each having a repetition period (Ts) 3/7 that of the pump pulse train (P): Ts = 3/7 *T*. The pump repetition rate is multiplied by a factor of 7 in the signal pulse train (S). (b) Fourth-order synchronous pumping at 1 GHz. The signal pulse train (S) is a combination of four pulse trains (S1–S4), each having a repetition period (4*T*) 4 times that of the pump pulse train (P).

See also

Optical Parametric Devices: Optical Parametric Oscillators (Continuous Wave); Overview.

Appendix A

Sellmeier Equations for some Typical Crystals and Optical Glasses

In the equations, n denotes the refractive index; λ is the wavelength of incident light in μm; $A, B, C, D, A_1, A_2, A_3, A_4, A_5, A_6, B_1, B_2, B_3, B_4, \ldots$, are the Sellmeier coefficients; the subscripts X, Y, Z, stand for the directions of the principal axes of the biaxial crystals.

KTP

$$n^2 = A + \frac{B}{1 - \left(\frac{C}{\lambda}\right)^2} - D\lambda^2$$

	A	B	C	D
n_X	2.1146	0.89188	0.20861	0.01320
n_Y	2.1518	0.87862	0.21801	0.01327
n_Z	2.3136	1.00012	0.23831	0.01679

KTA

$$n^2 = A + \frac{B}{\lambda^2 - C} - D\lambda^2$$

	A	B	C	D
n_X	3.1413	0.04683	0.04055	0.01023
n_Y	3.1593	0.04828	0.04710	0.01049
n_Z	3.4435	0.06571	0.05435	0.01460

RTA

$$n^2 = A + \frac{B}{1 - \left(\frac{C}{\lambda}\right)^2} - D\lambda^2$$

	A	B	C	D
n_X	2.04207	1.17785	0.20157	0.01035
n_Y	2.14941	1.09267	0.21462	0.01067
n_Z	2.18962	1.30103	0.22809	0.01390

CTA

$$n^2 = A + \frac{B}{1 - \left(\frac{C}{\lambda}\right)^2} - D\lambda^2$$

	A	B	C	D
n_X	2.34498	1.04863	0.22044	0.01483
n_Y	2.74440	0.70733	0.26033	0.01526
n_Z	2.53666	1.10600	0.24988	0.01711

$LiNbO_3$

$$n_e^2 = A_1 + B_1 f + \frac{A_2 + B_2 f}{\lambda^2 - (A_3 + B_3 f)^2} + \frac{A_4 + B_4 f}{\lambda^2 - A_5^2} - A_6\lambda^2$$

$$f = (T - T_0)(T + T_0 + 546) \text{ and } T_0 = 24.5\,^\circ\text{C}$$

Coefficients	*Values*
A_1	5.35583
A_2	0.100473
A_3	0.20692
A_4	100
A_5	11.34927
A_6	1.5334×10^{-2}
B_1	4.629×10^{-7}
B_2	3.862×10^{-8}
B_3	-0.89×10^{-8}
B_4	2.657×10^{-5}

LBO

$$n^2 = A + \frac{B}{\lambda^2 + C} + D\lambda^2$$

	A	B	C	D
n_X	2.4543	1.1413×10^{-2}	-9.4981×10^{-3}	-1.3900×10^{-2}
n_Y	2.5382	1.2830×10^{-2}	-1.1387×10^{-2}	-1.7034×10^{-2}
n_Z	2.5854	1.3065×10^{-2}	-1.1617×10^{-2}	-1.8146×10^{-2}

$$\frac{dn_X}{dT} = 2.0342 \times 10^{-7} - 1.9697 \times 10^{-8} T - 1.4415 \times 10^{-11} T^2$$

$$\frac{dn_Y}{dT} = -1.0748 \times 10^{-5} - 7.1034 \times 10^{-8} T - 5.7387 \times 10^{-11} T^2$$

$$\frac{dn_Z}{dT} = -8.5998 \times 10^{-7} - 1.5476 \times 10^{-7} T + 9.4675 \times 10^{-10} T^2 - 2.2375 \times 10^{-12} T^3$$

BBO

$$n_o^2 = 1.7018379 + \frac{1.0357554\lambda^2}{\lambda^2 - 0.01800344} + \frac{1.2479989\lambda^2}{\lambda^2 - 91}$$

$$n_e^2 = 1.5920433 + \frac{0.7816893\lambda^2}{\lambda^2 - 0.016067891} + \frac{0.8403893\lambda^2}{\lambda^2 - 91}$$

Sapphire

$$n^2 = 1 + \frac{A_1\lambda^2}{\lambda^2 - B_1} + \frac{A_2\lambda^2}{\lambda^2 - B_2} + \frac{A_3\lambda^2}{\lambda^2 - B_3}$$

A_1	A_2	A_3	B_1	B_2	B_3
1.023798	1.058264	5.280792	0.00377588	0.0122544	321.3616

Quartz

$$n_{o,e}^2 = A + \frac{B\lambda^2}{\lambda^2 - C} + \frac{D\lambda^2}{\lambda^2 - E}$$

	A	B	C	D	E
n_o	1.28604141	1.07044083	0.0100585997	1.10202242	100
n_e	1.28851804	1.09509924	0.0102101864	1.15662475	100

Fused Silica, BK7, SF2, SF5, SF10, SF11, SF18

$$n^2 = 1 + \frac{A_1\lambda^2}{\lambda^2 - B_1} + \frac{A_2\lambda^2}{\lambda^2 - B_2} + \frac{A_3\lambda^2}{\lambda^2 - B_3}$$

	A_1	A_2	A_3	B_1	B_2	B_3
Fused silica	0.6961663	0.4079426	0.8974794	0.0046791	0.01351206	97.9340025
BK7	1.03961212	0.23179234	1.01046945	0.0060007	0.0200179	103.560653
SF2	1.40301821	0.23176750	0.93905659	0.0105796	0.04932270	112.405955
SF5	1.46141885	0.24771302	0.94999583	0.0118261	0.05085946	112.041888
SF10	1.61625977	0.25922933	1.07762317	0.0127535	0.05819840	116.607680
SF11	1.73848403	0.31116897	1.17490871	0.0136069	0.06159605	121.922711
SF18	1.56441436	0.29141358	0.96030788	0.0121864	0.05355680	111.451201

Further Reading

Agrawal GP (1989) *Nonlinear Fiber Optics*. Boston, San Diego: Academic Press.

Albrecht TF, Sandmann JHH, Feldmann J, Stolz W, Göbel EO, Hillmer H, Lösch R and Schlapp W (1995) Design and application of a femtosecond optical parametric oscillator for time-resolved spectroscopy of semiconductor heterostructures. *Applied Physics* B 60: 459.

Arbore MA, Marco O and Fejer MM (1997) Pulse compression during second-harmonic generation in aperiodic quasi-phase-matching gratings. *Optics Letters* 22: 865–867.

Arigas D and Reid DT (2002) Efficient femtosecond optical parameteric oscillators based on aperiodically poled nonlinear crystals. *Optics Letters* 27: 851–853.

Bass M, Enoch JM, Van Stryland EW and Wolfe WL (2001) *Handbook of Optics: Volume IV, Fiber Optics and Nonlinear Optics*. New York: McGraw-Hill.

Beddard T, Ebrahimzadeh M, Reid TD and Sibbett W (2000) Synchronously pumped optical parametric oscillator driven by a femtosecond mode-locked fiber laser. *Optics Letters* 25: 1052.

Bierlein JD and Vanherzeele H (1997) Potassium titanyl phosphate: properties and new applications. *Journal of the Optical Society of America B* 6: 622.

Burr KC, Tang CL, Arbore MA and Fejer MM (1997) High-repetition-rate femtosecond optical parametric oscillator based on periodically poled lithium niobate. *Applied Physics Letters* 70: 3341.

Diels J-C and Rudolph W (1996) *Ultrashort Laser Pulse Phenomena: Fundamentals, Techniques, and Applications on a Femtosecond Time Scale*. San Diego, CA: Academic Press.

Dmitriev VG, Gurzadyan GG and Nikogosyan DN (1991) *Handbook of Nonlinear Optical Crystals*. New York: Springer-Verlag.

Dunn MH and Ebrahimzadeh M (1999) Parametric generation of tunable light from continuous-wave to femtosecond pulses. *Science* 286: 1513.

Fenimore DL, Schepler KL, Zelmon D, Kuck S, Ramabadran, Von Richter P, and Small D (1996) Rubidium titanyl arsenate difference-frequency generation and validation of new Sellmeier coefficients. *Journal of the Optical Society of America B* 13: 1935.

Ghost G (1995) Temperature dispersion of refractive indices in β-BaB_2O_4 and LiB_3O_5 crystals for nonlinear optical devices. *Journal of Applied Physics* 78: 6752.

Ghost G (1999) Dispersion-equation coefficients for the refractive index and birefringence of calcite and quartz crystals. *Optics Communications* 163: 95.

Hebling J, Giessen H, Linden S and Kuhl J (1997) Mirror-dispersion-compensated femtosecond optical parametric oscillator. *Optical Communications* 141: 229.

Jundt DH (1997) Temperature-dependent Sellmeier equation for the index of refraction, ne, in congruent lithium niobate. *Optics Letters* 22: 1553.

Kato K, Umemura N and Tanaka E (1997) 90° phase-matched mid-infrared parametric oscillation in undoped $KTiOAsO_4$. *Japanese Journal of Applied Physics 2* 36(4A): L403.

Loza-Alvarez P, Reid DT, Faller P, Ebrahimzadeh M and Sibbett W (1999) Simultaneous femtosecond-pulse compression and second-harmonic generation in aperiodically poled KTiOPO4. *Optics Letters* 24: 1071.

Matuschek N, Gallmann L, Sutter DH, Steinmeyer G and Keller U (2000) Back-side-coated chirped mirrors with ultra-smooth broadband dispersion characteristics. *Applied Physics B* 71: 509.

Mokhtari A, Fini L and Chesony J (1987) Efficient frequency mixing of a cw femtosecond laser synchronously pumped by a frequency doubled Nd:YAG laser. *Optics Communications* 61: 421.

Nebel A, Fallnich C, Beigang R and Wallenstein R (1997) Noncritically phase-matched continuous-wave mode-locked singly resonant optical parametric oscillator synchronously pumped by a Ti:sapphire laser. *Journal of the Optical Society of America B* 10: 2195.

Schubert M and Wilhelmi B (1978) *Nonlinear Optics and Quantum Electronics*. New York: John Wiley & Sons.

Seres J and Hebling J (2000) Nonstationary theory of synchronously pumped femtosecond optical parametric oscillators. *Journal of the Optical Society of America B* 17: 741.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: John Wiley & Sons.

Sutherland RL (1996) *Handbook of Nonlinear Optics*. New York: Marcel Dekker Inc.

Szipöcs R and Kohazi-Kis A (1997) Theory and design of chirped dielectric laser mirrors. *Applied Physics B* 65: 115.

Tang Y, Cui Y and Dunn MH (1995) Thermal dependence of the principal refractive indices of lithium triborate. *Journal of the Optical Society of America B* 12: 638.

Tomaru T and Petek H (2000) Femtosecond Cr4+ YAG laser with an L-fold cavity operating at a 1.2-GHz repetition rate. *Optics Letters* 25: 584.

Tomaru T and Petek H (2000) Thermal dependence of the principal refractive indices of lithium triborate. *Journal of the Optical Society of America B* 18: 388.

Velsko SP, Webb M, Davis L and Huang C (1991) Phase-matched generation in lithium triborate (LBO). *IEEE Journal of Quantum Electronics* 27: 2182.

Yariv A (1984) *Optical Waves in Crystals: Propagation and Control of Laser Radiation*. New York: John Wiley & Sons.

Yariv A (1989) *Quantum Electronics*. New York: John Wiley & Sons.

Zernike F and Midwinter JE (1973) *Applied Nonlinear Optics*. New York: John Wiley & Sons.

Zhang XP, Hebling J, Kuhl J, Rühle WW, Palfalvi L, Giessen H (2002) Femtosecond near-IR optical parametric oscillator with efficient intracavity generation of visible light. *Journal of the Optical Society of America B* 19: 2479.

Zhang XP, Kuhl J, Hebling J, Bartels A, Nau D, Rühle XP, Giessen H (2002) 1-GHz repetition-rate femtosecond optical parametric oscillator. *Applied Physics Letters* 80: 1873.

OPTICAL PROCESSING SYSTEMS

D Arbel and N S Kopeika, Ben-Gurion University of the Negev, Beer Sheva, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Modern optical systems find many applications in sensing, recording, storage, and transmission of data. These applications stimulate the development of many new optical devices, components, materials, and applied technologies. The basic element of an optical system is the light. The physical implications of the light in optical systems are very analogous to those of electricity, hence, light signals are quite similar to electric signals. High-speed spatial light components are being developed for more stringent applications in coherent optical processing systems, where they may be used in the input plane or in the Fourier transform plane. Research in this area is directed towards the design of robust and compact correlators.

Novel transforms for information processing (also known in electronics) include Fourier and wavelet transforms. Such optical transformations serve as the basic platform for linear optical system processing. Optical devices and systems for information processing include spatial light modulators, which are used to implement optical information processing systems. There are holographic applications in information processing, display systems, memory systems, optical computing systems, and imaging systems. Such devices and systems are based on applications of novel materials for information processing. Optical and digital techniques are used for optical system design, optical computing, image processing, and even for encryption and security systems.

The imaging system is an optical system and integrates hardware as well as software. Figure 1 presents a general model of an extensive imaging system. The starting point of the system is the object (or the light source). An imaging device, such as camera or optical sensor detects the energy radiated in electromagnetic waves from the light source. The image (or series of images – video) formed in the imaging device depends on the received irradiance of the image propagated through the imaging path and is correspondingly diminished from that in the object plane. The properties of the optical medium (transmission path) such as atmospheric properties, background light, motion, and vibration in both the object and the image planes, and the quality of the imaging device, all affect the quality of the images.

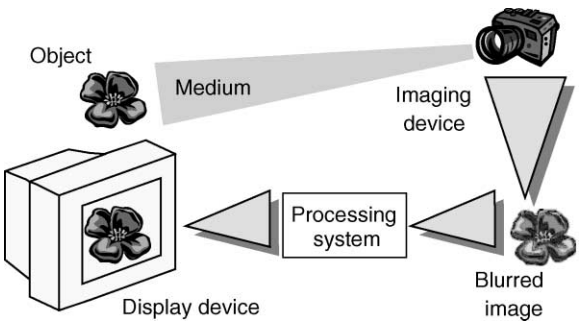


Figure 1 General description of an imaging system.

There are atmospheric phenomena that give rise to attenuation of the irradiance of the propagating image, thus reducing the contrast of the final image, as well as atmospheric phenomena that cause blurring of detail. Both types of phenomena degrade image quality.

Image blurring caused by vibration is a factor whose influence on resolution is often significant in imaging systems that involve mechanical motion. In order to obtain a well processed imaged, as similar as possible to the ideal image (the source), image processing techniques such as image enhancement, restoration, and filtering are implemented in an optical processing system which is a combination of software (algorithms) and hardware (electronic and optical components). The processing procedure involves mathematical and especially optical transformations. The output of the processing system is presented by using a display device or hardcopy. The development of imaging systems (imaging devices and displays) which include integrated optical components and processing units, assist in improving image quality at early stages, and as a result reduce the need of unique processing systems and may serve as the standalone device.

The key issues which must be addressed in the design, evaluation, and implementation of practical systems for optical processing, are provided here in a detailed review. Thus, the review is focused on imaging where considerations such as optical transformations, optical computing, effects of diffraction, scatter, absorption, and design on the performance of optical systems, imaging and the optical transfer function (OTF) for image restoration, and filtering, are all discussed in detail.

Digital Processing and Control

The Fourier transform properties of Fraunhofer diffraction permit data processing operations carried out optically that can, in many situations, be more advantageous than digital signal processing. The primary advantages of optical techniques over electronic analog and digital signal processing, are best realized when the information to be processed has two degrees of freedom. Optical systems correspondingly can have two degrees of freedom. Thus, for handling pictorial data, they can be intrinsically superior to electric or electronic systems, which have only time as an independent variable. This two-dimensional (2D) processing property can be particularly valuable in radar applications where previously velocity and position data had to be processed separately in order for sufficient accuracy in the estimation of each set of data to be obtained.

Many of the basic operations that can be performed by optical computers can also be done with electronic systems. However, if large quantities of wideband data are to be processed, the electronic techniques can often be prohibitively expensive and time consuming. The optical computer is faster than a digital computer because it operates on signals that are functions of positions, not time. The input to an optical data processor is a pattern of light irradiance that varies over some area, not light irradiance that varies as a function of time.

Thus, optical data processing can be very appropriate for such lengthy operations as the calculation of complicated Fourier transforms and the processing of side-looking radar signals. In addition to mathematical operations such as differentiation, integration, and producing Fourier transforms, the two classes of applications for which optical signal processing is most often used are filtering and correlation.

Optical Computing

Optical computing (OC) means the use of light as a primary means for carrying out numerical calculations, reasoning, artificial intelligence, etc. The history of OC is linked to that of radar systems. OC systems received a great push from the invention of laser in 1960. In the 1960s, the first schemes for all-optical digital computers were proposed. Since the 1990s, researchers have been pursuing the development of computers based on optical technology, and emphasis has shifted to optical interconnection of arrays of semiconductor smart pixels. OC provides the first in-depth review of the possibilities and limitations of optical data processing. The partial listing of those points that make OC appealing are direct image processing, massive parallelism and connectivity, immunity to electromagnetic interference (EMI), speed, size, and cost. There are three distinct trends in OC: special purpose analog optical systems, general purpose digital optical systems, and hybrid optical/electronic systems.

Innovations such as the electro-optic spatial filter, the 2D spatial light modulator, and the variable electro-optic mirror provide the fundamental building blocks for both analog optical computing (AOC) and digital optical computing (DOC) technologies. Optical techniques can provide solutions and a number of ways of extending the information processing capability of electronics. The optical processors have to be compatible with existing electronic systems. Free space digital optics is one

direction that provides some valuable solution. DOC requires the use of nonlinear optics.

All-optical processing refers to optically based processors employed nonlinear optical resources either of liquid crystal spatial light modulators (SLMs) or nonlinear interference filters (NLIFs). Electro-absorption devices are a bridge between all-optical demonstrators and optically interconnected smart pixels (logic and local on-chip interconnection is electrical whereas chip to chip interconnection is optical). Optically interconnected smart pixels have several advantages like faster data acquisition and low power consumption. A major breakthrough in nonlinear optics is essential to increase the use of optics within computing and the manufacture of optical digital computers.

An optically bistable device used for photonic switching is the self electro-optic effect device (SEED). It is the analogous component for the transistor in optical processing. A SEED consists of multiple elements, biased by an external voltage, which creates an external field that shifts the wavelength of the onset of absorption, thus varying the intensity of the transmitted light. SEEDs rely on changes in the optical transmission of a semiconductor induced by an applied electric field. SEEDs can be configured in pairs (connected electrically) so that a beam of light switching one device can cause a complementary switch in the transmission of the other. Hence, a small change in the intensity of one beam can cause a large change in intensity of the other. This kind of configuration is known as symmetric SEEDs, and acts like an electronic flip-flop and permits logic operations NAND and NOR to be carried out on pairs of optical input signals. With symmetric SEEDs, the higher the input optical power, the faster the switching speed. Cellular logic image processor (CLIP) is implemented using the symmetric SEEDs. CLIP computer architecture is designed to permit parallel information processing, in which logic operations are performed on each element of the array simultaneously.

Holography

Holography is a photographic method of recording information (two mutually coherent beams) about an object, in which one beam is the object beam and the other a reference beam, which enables the construction of objects in three dimensions. Holography relies on the encoding of the object formation in a set of complex interference fringes formed by the interaction of a plane coherent reference wave with the wavelets diffusely scattered by the object.

The microscopic interference fringes recorded on a high resolution photographic plate after development and processing become a hologram. Hologram applications assist the study of small deformations of objects, to replace conventional optical elements such as lenses and beamsplitters, for data/information storage, and use of photo-refractive materials in analog computing, object recognition, and correlation, etc.

The purpose of recording a hologram is not optical data processing but rather a lensless method of forming an image. The uniqueness of a hologram as compared to an ordinary photographic image, for example, is that the latter contains amplitude information only. A hologram, because it involves interference of the complex field amplitude (CFA) of a given scene with a reference beam, records both amplitude and phase of the CFA of the given scene as seen within its field of view (FOV) by a square-low photosensitive device such as a film.

The film is exposed within its region of linearity whereby transmission is proportional to irradiance for a given time exposure. The transparency resulting from the film is the hologram. To view the holographic image, the hologram is illuminated with a reference wave, which causes the CFA of the given scene (both amplitude and phase) to be reconstructed. Since hologram reconstruction involves the CFA of the given scene in its entirety, depth information is included as well. Furthermore, it is even possible to view the holographic scene from different physical points of view and to see around objects in the scene. This ability, however, is limited by the solid angle subtended by the object when the hologram is recorded. Hologram reconstruction cannot supply more information than that recorded in the hologram itself.

Image Processing

Figure 1 demonstrates a general behavior of an imaging process. The 'Processing System' block in Figure 1 is expanded and presented in more detail by Figure 2. Figure 2 demonstrates in a flow chart, a general process occurring within the processing system. The input to the processing system is a degraded image (or set of images or video signal) such as blurred image, noisy image, image with clutters, multi-spectral image, etc. The input depends on exterior (environmental conditions) and interior (optical, electronic, and mechanical) effects of the imaging system. It may be generated as an analog signal or digital signal. As a result, the output of the imaging system is expected to be a processed (filtered, enhanced, estimated, restored, cleared, etc.) image.

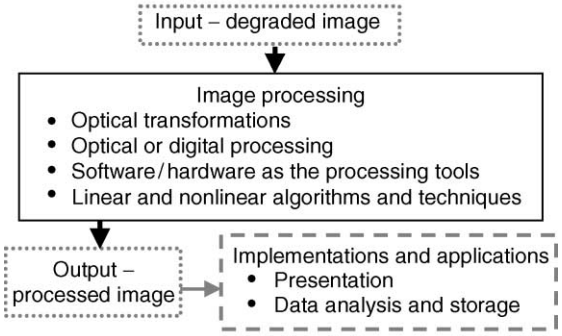


Figure 2 Image processing.

Implementations and applications depend on the output obtained by the imaging system. The output can be presented by a display device and/or hardcopy. The output can be stored for future assignments and may also serve for data analysis such as detection, recognition, and identification of patterns and other details of interest (mapping, target acquisition, defects and imperfections in the human body, meteorology, security, intelligence, etc.).

Spatial units are reciprocal to time, and are cycles/seconds or Hz. The spatial units are also reciprocal to the domain of the information being sought. Spatial signals are 2D or even 3D on occasions. 2D spatial transforms may actually exist physically in the spatial plane. Novel transforms for information processing include Fourier and wavelets transforms. Such optical transformations serve as the basic platform for linear optical system processing. Image processing techniques are based on the image spatial frequencies. In most techniques the degraded image is transformed using an optical transformation, as a primary stage before any other process is being conducted. Optical (inverse) transformations also serve as the last processing tool before the output. Practically, optical transformations are involved and integrated in most processes related to image processing.

Image processing refers to both digital and optical image processing, and also considers algorithms and techniques, which are implemented by an integration of software and hardware. Image processing is an extremely broad field. It considers digital image processing, 2D and 3D image processing and pattern recognition, real-time processing, neural networks, image enhancement/improvement, image restoration, estimation, filtering for noise removal, linear and nonlinear filtering techniques for variety of implementations, optimum filtering, etc. Image processing also include varieties of algorithms such as linear and nonlinear algorithms, algorithms for pattern recognition, linear and nonlinear techniques for pattern recognition, neural networks based algorithms, etc. Both optical and electronic implementations (hardware) are still under investigation. Nevertheless, image processing based on software presents reliable results and hence, can be transformed into hardware components, which may become an integrated part of the processing hardware.

Optical Transformations

Mathematical transformations are applied to signals to obtain further information from the signal that is not readily available in the raw signal. Most of the signals, in practice, are time-domain signals in their raw format. That is, whatever that signal is measuring, is a function of time. In many cases, the most important information is hidden in the frequency content of the signal. The frequency spectrum of a signal is basically the frequency components (spectral components) of that signal. The frequency spectrum of a signal shows what frequencies exist in the signal.

There are many transforms that are widely used, often by engineers and mathematicians, in solving problems in science and engineering. Fourier, Laplace, Hilbert, Wigner distributions, Radon, and wavelet transforms, constitute only a small portion of a huge list of transforms that are available. Every transformation technique has its own area of application, with advantages and disadvantages.

Fourier Transform

The Fourier transform (FT) is probably the most popular linear transform being used. The FT, a pervasive and versatile tool, is used in many fields of science as a mathematical or physical tool to alter a problem into one that can be more easily solved. It is used in linear systems analysis, antenna studies, optics, random process modeling, probability theory, quantum physics, and boundary-value problems and has been successfully applied to restoration of astronomical data. Some scientists understand Fourier theory as a physical phenomenon, not simply as a mathematical tool. In some branches of science, the FT of one function may yield another physical function. The FT, in essence, decomposes or separates a waveform or function into sinusoids of different frequency which sum to the original waveform. It identifies or distinguishes the different frequency sinusoids and their respective amplitudes. There are

functions for which the Fourier transform does not exist; however, most physical functions have a FT, especially if the transform represents a physical quantity. Other functions can be treated with Fourier theory as limiting cases. Many of the common theoretical functions are actually limiting cases in Fourier theory.

Wavelets

Wavelets were developed independently in the fields of mathematics, quantum physics, electrical engineering, and seismic geology. Interchanges between these fields have led to many new wavelet applications including image processing. Wavelets are mathematical functions that cut up data into different frequency components, and then study each component with a resolution matched to its scale. Wavelet transforms comprise an infinite set. The different wavelet families make different trade-offs between how compactly the basis functions are localized in space and how smooth they are. Within each family of wavelets are wavelet subclasses distinguished by the number of coefficients and by the level of iteration.

The wavelet transform (WT), which is also a linear operation, provides the time-frequency representation. Often a particular spectral component occurring at any instant can be of particular interest. In these cases it may be beneficial to know the time intervals at which these particular spectral components occur. WT is capable of providing the time and frequency information simultaneously, hence giving a time-frequency representation of the signal. The WT was developed as an alternative to the short-time FT (STFT). The WT was developed to overcome some resolution related problems of the STFT. The frequency and time information of a signal at some certain point in the time-frequency plane cannot be known. What spectral component exists at any given time instant is a mystery. Therefore, it is best to investigate what spectral components exist at any given interval of time. This is a problem of resolution, and it is the main reason why researchers have switched to WT from STFT. STFT gives a fixed resolution at all times, whereas WT gives a variable resolution as follows: higher frequencies are better resolved in time, and lower frequencies are better resolved in frequency. This means that a certain high frequency component can be located better in time (with less relative error) than a low frequency component. On the contrary, a low frequency component can be located better in frequency, compared to high frequency component.

Wavelet versus Fourier Transforms

Both transforms are linear and the mathematical properties of the matrices involved in the transforms are similar. The inverse transform matrix for both transforms is the transpose of the original. Wavelets have advantages over traditional Fourier methods in analyzing physical situations where the signal contains discontinuities and sharp spikes. The most interesting dissimilarity is that individual wavelet functions are localized in space, Fourier sine and cosine functions are not. This localization feature, along with wavelets localization of frequency, makes many functions and operators using wavelets 'sparse' when transformed into the wavelet domain. This sparseness, in turn, results in a number of useful applications such as data compression, detecting features in images, and removing noise from time series. One way to see the time-frequency resolution differences between the FT and the WT is to look at the basis function coverage of the time-frequency plane. In order to isolate signal discontinuities, some very short basis functions are needed. At the same time, in order to obtain detailed frequency analysis, some very long basis functions are needed. A way to achieve this is to have short high-frequency basis functions and long low-frequency ones. This happy medium is obtained with WT. FT utilizes just the sine and cosine functions (single set of basis functions). Therefore, its new domain contains basis functions that are sines and cosines. On the other hand, WT includes an infinite set of possible basis functions. Hence, its new domain contains more complicated basis functions called wavelets, mother wavelets, or analyzing wavelets. The basis functions are localized in frequency, making mathematical tools such as power spectra and scalegrams useful at picking out frequencies and calculating power distributions. Thus wavelet analysis provides immediate access to information that can be obscured by other time-frequency methods such as Fourier analysis.

Physical Properties and Analog/Digital Mitigation Processing in Nonideal Systems

The weakest link in most imaging systems is time varying distortions, such as the Earth's atmosphere, motion, or vibrations. To obtain clear imaging through time varying distortions, it is useful to characterize the physical degradation phenomena using an OTF. It is often important to use the appropriate filter based on the OTF magnitude, modulation transfer function (MTF), to correct

(restore) the degraded image (using software or hardware). On the other hand, it may be useful to incorporate real-time phase correction into the imaging system such as adaptive optics.

Optical Transfer Function for Incoherent Imaging

In electronics, the transfer function is redefined as the Fourier transform of the impulse response. The OTF is defined similarly, but is normalized to its own maximum value which normally occurs at zero spatial frequency, that is:

$$\begin{aligned}\mathrm{OTF} &= \tau(\omega) = \tau(\omega_x, \omega_y) \\ &= \frac{\int\int_{-\infty}^{\infty} s(x',y')\exp[-j(\omega_x x' + \omega_y y')\mathrm{d}x'\,\mathrm{d}y']}{\int\int_{-\infty}^{\infty} s(x',y')\mathrm{d}x'\,\mathrm{d}y'} \\ &= \frac{S(\omega_x,\omega_y)}{S(0,0)} = \mathrm{MTF}\,\exp(j\mathrm{PTF}) \qquad [1]\end{aligned}$$

where $S(\omega_x, \omega_y)$ is the FT of the spread function or impulse response $s(x', y')$. Ordinarily, for artificial imaging systems, the OTF is maximum at zero frequency. The magnitude of the OTF is called the MTF. The name refers to spatial modulation of light irradiance. If there is no such modulation, irradiance is uniform and there is no image. Thus, spatial modulation is related to image quality and MTF is a measure of the ability of an imaging component or system to transfer such spatial modulation from input (object plane) to output (image plane).

There is also a phase transfer function (PTF) whose importance in resolution is not nearly as widespread as that of MTF but which, nevertheless, cannot be neglected. Spatial phase determines image position and orientation, rather than amount of detail. If a target is displaced bodily in the image plane such that each part of the target image is displaced by the same amount, the target image is not distorted. However, if portions of the target image are displaced more or less than other portions, then the target image is distorted. This information is contained in the PTF.

The physical implications of the OTF are very analogous to those of electronics transfer functions. Both permit determination of output for any given input. In both cases, since the FT of the delta function is a constant (unity) amplitude, a delta function input permits indication of the frequency response of the system, that is, which frequencies pass through the system unattenuated and which are attenuated and by how much. Perfect fidelity or resemblance between input and output requires infinite bandwidth so that an impulse function is obtained at output for impulse function input. In imaging, an impulse formation at output is a point image. This means that, to obtain a point image for a point object, infinite spatial frequency bandwidth is required of the imaging system. Physically, this means that because of diffraction effects optical elements such as lenses, field stops, and mirrors must be of infinite diameter so as to eliminate diffraction at the edges. This, of course, is not attainable.

Pixel Modulation Transfer Function

To measure the point spread function (PSF), it is important to determine which size object can be considered a point object. This is affected primarily by instrumentation pixel size. The pixel is the smallest element recordable in image space. The brightness value represents average irradiance over that small portion of the image scene. Pixel size is often related to detector size. If only a portion of the detector is illuminated, the output current is equivalent to that obtained for the same total radiant power absorbed by the detector but averaged over the entire detector area. No detail smaller than a pixel can be resolved in the image. In practice, dead space between pixels violates isoplanaticism (spatial stationarity), which is a requirement for linear systems. Nevertheless, dead space is small and such effects are often neglected in image system characterization. However, it is possible to overcome distortions in MTF measurements arising from this lack of isoplanaticism by using a 'white' spatial noise random object. This gives rise to a white noise random image. The ratio of the image power spectral density to object power spectral density is equal to the square of the system MTF. In this way the lack of isoplanaticism is overcome by the spatial randomness of the object and image.

Pixel size strongly affects system MTF. If a pixel represents a point image, then pixel size and shape are minimum spread function. A best case OTF for such an imaging system is then a normalized FT of the pixel shape. For example, a square pixel yields a two-dimensional sinc MTF whose width increases as pixel size decreases.

Optical Transfer Functions for Image Motion and Vibration

Image blurring caused by vibrations is a factor whose influence on resolution is often significant in imaging systems that involve mechanical motion. A stabilized system, which obtains image motion corrections by moving optical elements so as to counteract sensor motions, has a residual error typically on the order of 2500 μrad/s involving pitch, roll, yaw, and

forward motion compensation errors. The greatest errors usually result from rotational vibration. A stabilized mount-type system has a combined error typically on the order of 100 μrad/s. Even where no motion is involved, as in large fixed-position astronomical telescopes, vibrations deriving from thermal gradients in the walls and other parts of the telescope can be the limiting factor in image resolution.

As a consequence of such image motion, in many high-resolution systems, despite the use of high-quality sensors, resolution is limited by image motion and, as a result, the high-resolution capability of the sensor may be wasted. It can be a significant waste of means to use an expensive high-resolution sensor in such a situation unless the blur can be corrected with image restoration. In imaging system design, a convenient engineering tool is OTF. The overall system OTF is limited by the OTF of the weakest link. The formulation of such image motion blur into an OTF format is thus very convenient for system design and system analysis purposes, as well as for image restoration.

OTFs can describe image quality for many forms of image motion such as linear (uniform) motion, sinusoidal vibrations at high vibration frequencies, sinusoidal vibrations at low vibration frequencies, acceleration, etc. The OTF for any type of image motion can be calculated numerically in real time provided the motion function is known. This is usually determined easily via a gyroscope or microelectronic accelerometer.

In general, the motion of interest is not the actual relative motion between object and imager, but rather the relative motion between imager and image plane. The latter motion is that observed in the image plane. Both of these motions are related to each other by the image system magnification. Hence, image plane relative motion is usually much less than the actual relative motion. However, any motion during an exposure gives rise to blurring of image plane detail. One question to be considered is how much blur can take place as a result of motion and still not be noticeable because of inherent blur deriving from the PSF of the rest of the imaging system. In other words, assuming that image motion and vibration blur cannot be removed completely, how much is tolerable before resolution and target acquisition capability are affected? This depends on focal length and system zoom.

A general method of OTF calculation is obtained by the line spread function (LSF) derived from image motion transverse to the optical axis. The MTF is derived as the modulus of the MTF or FT of the LSF, and the PTF is derived as the phase of the OTF. The LSF is the probability density function (PDF) of the histogram of the relative transverse displacement between object and sensor beginning at time t_s and ending at time $t_s + t_e$ where t_s is measured from the instant the sensor is first exposed and t_e is exposure time.

Recently a method of calculating motion OTF analytically has been developed for any type of motion. Also, methods to calculate motion OTF without a motion sensor have been derived for a sequence of consecutive images and even from a single motion-blurred image. All three methods have been used successfully in image restoration.

Optical Transfer Functions for the Atmosphere

Many properties of the atmosphere affect the quality of images propagating through it. There are atmospheric phenomena that give rise to attenuation of the irradiance of the propagating image, thus reducing the contrast of the final image. There are also atmospheric phenomena that cause blurring of detail. Both types of phenomena prevent small detail from being resolved in the final image, thus degrading image quality.

Electromagnetic (EM) wave absorption and scattering by the constituent gases and particulates of the atmosphere and airborne particulates give rise to attenuation. Scattering of photons by airborne particulates is manifested as deflections of the photons to directions other than that of the original propagation. If such scattering causes the deflected photons to miss the imaging receiver, then the scattering is manifested as attenuation. The received irradiance of the image propagated through the atmosphere is correspondingly diminished from that in the object plane. However, if the light scattering is at very small angles with respect to the original direction of the propagation, and several such small-angle scattering events take place, then forward-scattered radiation can take round-about paths and still be received by the imaging system together with the unscattered radiation. The net effect is image blurring caused by a multitude of angles of arrival at the imaging receiver of radiation emanating from the same point in the object plane. Many multiple-scattering paths give rise to a relatively large blurred-point image rather than a fine sharp-point image. Adjacent image plane points can then appear for a single object plane point, thus degrading image resolution. Such aerosol scatter blur is known as the adjacency effect.

Another effect of light scattering, particularly at large angles, is increased path radiance. The atmospheric background irradiance at wavelengths less

than 2–4 μm, consists through the day of scattered background light such as sunlight. This atmospheric background radiation is imaged over the same image space as the received irradiance from the target plane. The result is decreased contrast of the target plane scene. At wavelengths larger than 4 μm, most of the path radiance is not scattered sunlight, but rather thermal emission of the atmospheric constituents, gases in particular. This thermal path radiance decreases contrast in thermal imaging systems, particularly in the longwave infrared (8–13 μm).

Turbulence results from variations in the atmospheric refractive index, which are random in both time and space. These refractive index fluctuations are caused by local fluctuations in atmospheric temperatures, pressure, humidity, etc., all of which are affected by wind speed, which in itself is also random in time and space. Generally, the larger the temperature and humidity gradients, the more serious the image resolution degradation caused by turbulence. The lower the wind speed, the stronger the impact of turbulence. Scattering and turbulence are basically two very different physical mechanisms, each exhibiting quite different properties. Time dynamic properties such as scintillations and image dancing derive from turbulence, while aerosol scatter light changes only slowly with time.

A point that must be strongly emphasized is that there is no such thing as a purely turbulent, purely scattering, or purely absorption atmosphere. All three properties occur simultaneously in the real-world atmosphere. Modeling and interpretation of experimental results must relate to all three processes. Any attempt to pretend otherwise is not realistic. In general, blur is caused by turbulence and small-angle scatter of light by aerosols and particles. Attenuation is caused by large-angle scatter and by absorption.

Generally turbulence blur is dominant for images near the ground, where ground heating gives rise to the turbulence. Aerosol blur is usually dominant at higher elevations, even as little as several meters above ground, depending on weather conditions and wavelength.

The product of both aerosol and turbulence MTFs (in the spatial frequency domain, after FT) yields approximately atmospheric MTF. The atmospheric MTF can be predicted according to atmospheric (meteorological) parameters. Aerosol MTF calculation is based on optical depth, which includes parameters such as scattering and absorption coefficients through the atmosphere over the imaging path. Turbulence MTF calculation is based on the refractive index structure coefficient, which includes parameters such as temperature, humidity, wind speed, etc.

By referring to the edge response in a given scene the atmospheric MTF may be calculated. The edge separating the wide white and black bars is imaged from the scene. The image of the edge is not a step function because of atmospheric blur. The gradient of the edge image is the LSF, from which the overall atmospheric MTF can be obtained after the FT of the LSF is divided by the hardware MTF.

Spatial Filtering

Spatial filtering is a process by which we can alter properties of an optical image by selectively removing certain spatial frequencies that make up an object, for example, filtering video data received from satellite and space probes, or removal of raster from a television picture or scanned image. In digital image processing, the term image refers to a two-dimensional light-intensity function, where amplitude at spatial coordinates (x, y) gives the intensity (brightness) of the image at that point.

Digital image processing techniques are oriented to transform, enhance, restore, or encode an image. Digital image restoration is on the basis of the OTF or even MTF alone describing the image degradation. Resolution is limited usually not by electronics or optics, but by image motion and vibration. For long-range imaging, the image quality may be limited by atmospheric effects.

By image restoration filters we refer primarily to spatial frequency filters placed in the spatial frequency plane in the optical data processor. Although data processing is done optically, the same mathematical concept is usually carried out with a digital computer. It is also possible to do restoration in the spatial domain. The simplest image restoration filter is the inverse filter. When noise becomes a problem, there are often used three classes of minimum mean square error estimations (MMSEE), which are the noncausal Wiener filtering, causal Wiener filtering, and Kalman filtering. Kalman and Wiener filters are proposed to obtain the optimum solution for the MMSEE problem in linear systems. Wiener filters are applied in linear time invariant systems, and Kalman filters in linear systems which may be either time invariant or time variant. A novel area of application for Wiener filtering is image restoration. On the other hand, if the estimation processes are partial and an ideal image is not obtainable, a Kalman filter may present better restoration results than a Wiener filter.

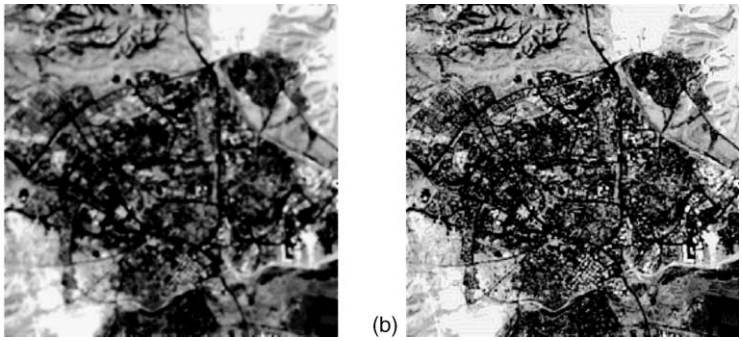


Figure 3 An example for satellite image restoration. (a) Original image of Beer-Sheva, Israel, from Landsat 7 satellite; (b) restored image after Kalman filter using aerosol and turbulence MTFs. Reproduced with permission of the American Society for Photogrammetry and Remote Sensing from Arbel D, Cohen E, Citroen M, Blumberg DG and Kopeika NS (2004) Landsat TM satellite image restoration using Kalman filters. *Photogrammetric Engineering and Remote Sensing* (PE&RS) 70(1): 91–100.

Figure 3 demonstrates the restoration result for a satellite image. Figure 3a presents an original Landsat 7 enhanced thematic mapper plus (ETM+) image of the city of Beer-Sheva, Israel. The image was recorded from 750 km satellite elevation. The image was recorded using an ETM+ sensor with 15 m ground resolution, and its wavelength range is 520–900 nm (Panchromatic band). Figure 3a is of 256×256 pixel size, which is a segment defined as a region of interest (ROI) in the original image. The scene was recorded on August 7, 1999 and was degraded (blurred) as a result of the meteorological conditions, which were measured and recorded (as data) at the same time. Figure 3b shows the restored image using the Kalman filter, which was based on the atmospheric MTF, which is obtained by using aerosol MTF with turbulence MTF. The restored image (Figure 3b) is more detailed. For example, roads and buildings were deblurred and became more accurate than in the original image (Figure 3a).

See also

Environmental Measurements: Optical Transmission and Scatter of the Atmosphere. **Fourier Optics**. **Holography, Techniques:** Computer-Generated Holograms. **Quantum Optics:** Quantum Computing with Atoms.

Further Reading

Arbel D, Greenberg S, Hadar O and Kopeika NS (2003) Criteria for satellite image restoration success. *Optical Engineering* 42(9): 2607–2613.

Arbel D, Cohen E, Citroen M, Blumberg DG and Kopeika NS (2004) Landsat TM satellite image restoration using Kalman filters. *Photogrammetric Engineering and Remote Sensing (PE&RS)* 70(1): 91–100.

Daniels A, Boreman GD and Ducharme AD (1995) Random transparency targets for modulation transfer function measurement in the visible and infrared. *Optical Engineering* 34(3): 860–868.

Gabor D (1948) A new microscope principle. *Nature* 161: 777.

Gabor D (1949) Microscopy by reconstructed wavefronts. *Proceeding of the Royal Society A* 197: 454.

Gabor D (1951) Microscopy by reconstructed wavefronts: II. *Proceeding of the Royal Society B* 64: 449.

Hadar O, Dror I and Kopeika NS (1994) Image restoration limits resulting from mechanical vibrations, part IV: real-time numerical calculation of optical transfer functions and experimental verification. *Optical Engineering* 33: 566–578.

Hadar O, Robbins M, Novogrozky Y and Kaplan D (1996) Image motion restoration from a sequence of images. *Optical Engineering* 35: 2898–2904.

Kopeika NS (1998) *A System Engineering Approach to Imaging*. Bellingham, WA: SPIE Press.

O'Donnell E (1987) Image steadiness: a critical factor in system performance. In: Wright R (ed.) *Electron Optical Imaging System Integration. Proceedings SPIE* 762: 51–56.

Stern A and Kopeika NS (1995) Analytical method to calculate optical transfer functions for image motion and vibrations using moments. *Journal of the Optical Society of America A* 14: 388–396.

Yitzhaky Y and Kopeika NS (1997) Identification of blur parameters from motion blurred images. *CVGIP: Graphical Models and Image Processing* 59(5): 321–332.

OPTICAL TWEEZERS

A Gajraj and J C Meiners, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The concept of radiation pressure or pressure from the propagation of light is not new. In the early seventeenth century, Johannes Kepler proposed that radiation pressure from the sun accounted for the fact that tails of comets always point away from the sun. In 1873, James Maxwell, on the basis of electromagnetic theory, showed that pressure should result from the absorption or reflection of a beam of light. However, early studies in the beginning of the twentieth century that attempted to detect and quantify the effects of radiation pressure were unsuccessful, mainly frustrated by thermal forces from temperature gradients that tended to obscure the weak effects of radiation pressure. It was only in the 1960s, that the advent of lasers provided a light source with the power to generate sufficient radiation pressure to overcome forces due to thermal effects.

As lasers became generally available in the early 1970s, Arthur Ashkin from the Bell Labs pioneered the development of optical tweezers. He showed that the forces of radiation pressure from focused laser beams could be used to affect the dynamics of small transparent micrometer-sized particles. According to Ashkin, his early experiments were motivated by simple back of the envelope calculations of the magnitude of the radiation pressure force of light on a totally reflecting mirror. Let us consider such a calculation: a photon of frequency ω has a momentum of $\hbar\omega/c$. For a total power of I, we have a flux of $I/\hbar\omega$ photons per second incident on the mirror. When the photons are reflected, they transfer a momentum of $2I/c$ to the mirror. The force exerted on the mirror for $I = 1$ W, is approximately 10^{-9} N. If the mirror is a 1 μm size particle weighing 10^{-15} kg, this force will induce an acceleration of 10^6 g!

In the following decade, Ashkin demonstrated that by focusing laser light using a high numerical aperture microscope objective, strong intensity gradients could be created to optically trap small dielectric particles such as microspheres. Ashkin named this trapping scheme a 'single-beam gradient force trap', now more commonly called optical tweezers.

Today optical tweezers are well-established and extremely versatile tools for the physical and life sciences. Using optical tweezers, forces up to 200 pN can be applied with a resolution of 100 aN on particles with dimensions on the order of the wavelength of light. Even particles as small as 5 nm can be trapped, rotated, positioned, and arrayed. Optical tweezers have been used in single molecule studies of biopolymers (DNA, RNA), in the manipulation of various types of cells and in cell sorting, in microsurgical procedures, as well as in viscoelastic studies of membranes, and protein. Optical tweezing has also been the key tool in the understanding of the workings of molecular motors such as myosin, kinesin, RNA polymerases, etc.

Theory

Optical tweezers trap dielectric particles using radiation pressure generated by a tightly focused beam of light from a laser. Depending on the size of the particle in relation to the wavelength of the trapping beam, there are two different mechanisms responsible for trapping small particles in a single beam gradient trap. For particles in the Mie regime, where the particle size is large compared to the laser wavelength, a simple geometric-optics model serves best. Figures 1a–c, illustrate possible cases of a Mie particle in an optical trap. Here, the particle is assumed to be a sphere that is transparent with a refractive index higher than that of the surrounding medium. The particle behaves as a convergent lens refracting and deflecting the beam when its center is displaced from the laser focus. In Figure 1a, the sphere is axially displaced from the focus, away from the microscope objective. Ray A is then refracted toward the surface normal as it enters and exits the sphere, causing a change in the momentum of the ray. Due to conservation of momentum, the sphere reacts with a force F_a in the opposite direction. The situation is similar for ray B. The gradient force, F, obtained from the vectorial addition of F_a and F_b, is directed toward the light source. At the same time, the scattering force, arising from back and scattered reflections, is directed away from the light source. It can be seen that above the focal point, the scattering and gradient forces act in the same direction, pushing the particle away from the light source (Figure 1b). For a particle in the focal plane, but transversely displaced from the trap center, the gradient force acts in a restoring direction pushing the bead into the trap focus (Figure 1c). At a location just below the focus, the scattering force balances the gradient force and stable optical trapping is achieved.

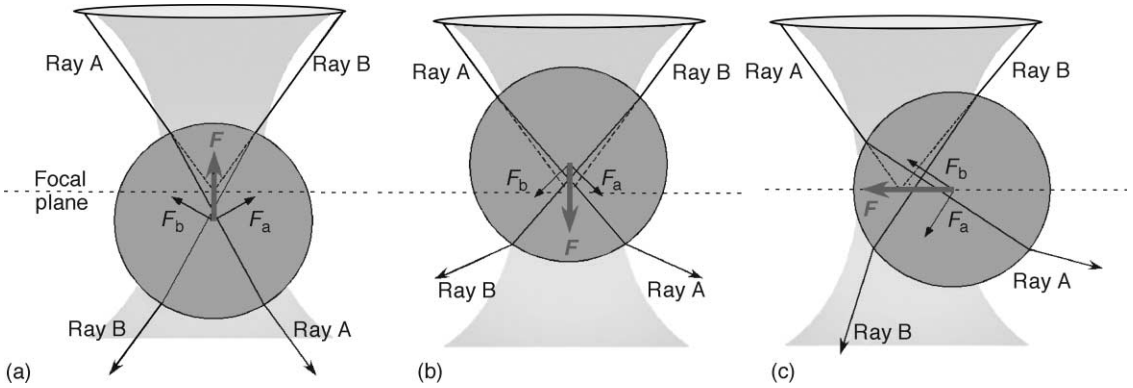


Figure 1 Ray diagram sketches showing three cases of particle trapping in the Mie regime. Rays A and B border the cone of light focused by a high numerical aperture microscope objective. The refraction of these rays gives rise to forces F_a and F_b, the resultant force F tends to restore axial and transverse displacements of the particle from the trap focus (at the intersection of dotted rays A and B).

This simple ray optics model can also be used to estimate the trapping force, an example of such a calculation is shown in **Box 1**.

For dielectric particles in the Rayleigh regime, where the size of the particle is small compared to the wavelength of the incident light, the particle acts as a simple dipole. The electromagnetic light wave polarizes the particle, inducing a set of oscillating dipoles. The dipoles respond to the electric field of the incident light in the direction of the gradient, resulting in the movement of the particle towards the focal point. Additionally, there is also a scattering force, acting in the direction of beam propagation, from light that is scattered or absorbed. While the scattering force is proportional to the electromagnetic field, the gradient force is proportional to the gradient of the electromagnetic field. Therefore, stable optical trapping can be achieved for a field with a gradient large enough to overcome the scattering force. This can be obtained by using a microscope objective with a high numerical aperture.

Stable trapping is achieved when the energy associated with the trapping significantly exceeds the thermal energy of the trapped particle, $\frac{1}{2}k_BT$ per degree of freedom. Since the optical gradient is weakest in the axial direction of the trap, the trap stiffness in this direction limits its stability. Furthermore, the trap is weakened in the axial direction by the scattering force, which pushes the microsphere away from the objective. To overcome this, two counterpropagating laser beams can be used, not only to increase the axial gradient of the trap, but also to provide a compensating scattering force in the opposite direction.

Force Measurement

Optical tweezers can be used as picotensiometers or highly sensitive force sensors. Measuring the displacement of a particle in a trap with known stiffness k, yields the force F given by

$$\mathbf{F} = -k \cdot x \qquad [1]$$

The displacement can be measured in different ways, ranging from video microscopy to light-scattering techniques. Some of these methods are discussed in the instrument design section below.

The trap stiffness is generally not known *a priori* as it depends on the shape and intensity of the trap, as well as the relative refractive index, shape, and position of the object in the trap. Since theoretical approximations are usually not accurate enough, the trap stiffness must be empirically determined and calibrated.

Various force calibration strategies have evolved over the years, however, they can be broadly divided into two types: (i) schemes that rely on the interaction of the optical trap with viscous forces; and (ii) those that exploit the thermal fluctuations of the trapped object.

Force Calibration by Viscous Drag

The conceptual basis for calibration by viscous drag is simple. A known force is exerted on a trapped object and the deviation in its position is measured as a function of the applied force. The trap stiffness can then be calculated from eqn [1]. Force can be exerted by either flowing liquid past the trapped object, translating the microscope stage while keeping the trap fixed, or translating the optical trap while keeping the sample fixed.

In a situation where the Reynolds number is small, inertial forces can be neglected and Stokes' Law gives the viscous drag force on a sphere of known radius r:

$$F_d = (6\pi\eta r)v = \beta v \qquad [2]$$

Box 1

Optical Tweezers in the Ray Optics Model

In a simple ray optics model, the trapped microsphere acts like a small lens on the incident laser beam. If the microsphere is not centered on the beam axis, the beam will be refracted. The corresponding change in the momentum of the laser beam gives rise to a restoring force on the microsphere.

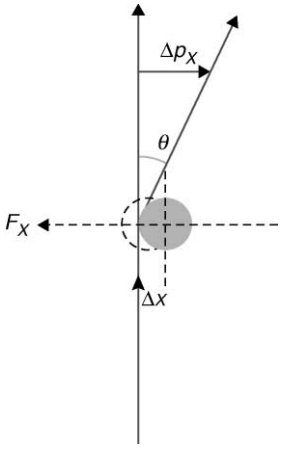


Figure B1

A photon of frequency ω in a medium of refractive index n_1 that is deflected by an angle θ, as shown in Figure B1, acquires a lateral momentum of

$$\Delta p_x = \frac{\hbar \omega n_1}{c} \sin \theta \qquad [B1]$$

Momentum conservation requires a counteracting force on the microsphere, which is for a laser beam of intensity I

$$F_x = \frac{I n_1}{c} \sin \theta \qquad [B2]$$

Assuming that the deflection angle θ is given by the deflection of a beam that is incident on biconvex lens with curvature R and refractive index n_2 at an off-axis displacement Δx, we find for the restoring force on the microsphere:

$$F_x = \frac{2I}{Rc}(n_2 - n_1)\Delta x \qquad [B3]$$

which describes a linear spring with a stiffness of the optical trap of

$$k = \frac{2I}{Rc}(n_2 - n_1) \qquad [B4]$$

Here, η, the viscosity of the sample fluid, is known and immediately allows the determination of the friction coefficient β of the sphere. The fluid velocity, v, can be measured in the case of external fluid flow by either tracking the trapped object when the trap is switched off or measuring the speed of some other marker in the sample. If the stage is moved to generate the fluid flow, the velocity of the fluid surrounding the trapped object can be estimated from the motion of the stage. In either case, the drag coefficient, from eqn [2] may need to be corrected if the trapped object is close to the walls of the sample cell. This correction is well known as Faxen's Law, which accounts for hydrodynamic interactions between the trapped particle and the wall. These effects are particularly significant when the center of the trapped particle is less than a diameter away from the wall. The advantage of these fluid-flow

calibration methods is that the trap stiffness at regions far from the trap center can be measured. The drawback is that, in general, only transverse trapping forces can be measured and the trapped objects have to be approximated as spherical or ellipsoidal.

Thermal Motion Methods

These methods measure the Brownian motion of a trapped particle to deduce the trap stiffness. A trapped particle experiences three forces: (i) the thermal forces due to collisions with surrounding solvent molecules; (ii) forces from viscous drag; and (iii) the restoring force of the trap. For low Reynolds numbers, where the inertial effects can be neglected for the microsphere, the equation of motion is

$$\beta\dot{x} + kx = F(t) \qquad [3]$$

Where β is the viscous drag, k is the trap stiffness, and $F(t)$ is a randomly fluctuating Langevin force due to thermal fluctuations. This motion of the microsphere in the trap at a temperature T is characterized by a Lorentzian power spectrum of the form:

$$x_f^2 = \frac{k_b T}{2\pi^3 \beta(f_0^2 + f^2)} \qquad [4]$$

where $f_0 = k/(2\pi\beta)$ is the corner frequency, above which the motion of the particle is Brownian.

To calibrate the optical trap by the corner frequency method, the motion of a trapped particle is observed and its power spectrum, which is the Fourier transform of the autocorrelation function, is calculated. Fitting the power spectrum to a Lorentzian function will yield the corner frequency f_0, which can be directly related to the trap stiffness k, provided that the drag coefficient β can be estimated.

Alternatively, the trap stiffness can be obtained directly from the time constant of the exponential decay of the autocorrelation function of the particle motion. The spring constant is the ratio of the drag coefficient to the time constant of the trap.

Another approach known as the mean square displacement method does not require the estimation of the drag coefficient but does require a well-calibrated position detector. In this method, the mean square displacement $\langle x^2 \rangle$ of the fluctuations of a trapped particle is computed from position data and then by the Equipartition theorem:

$$\langle x^2 \rangle = k_b T/k \qquad [5]$$

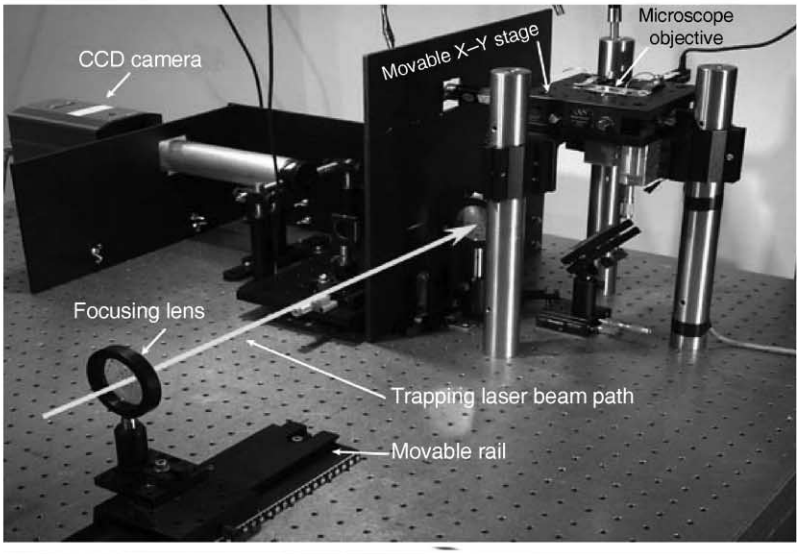
This enables a simple, straightforward calculation of k, the trap stiffness.

Design

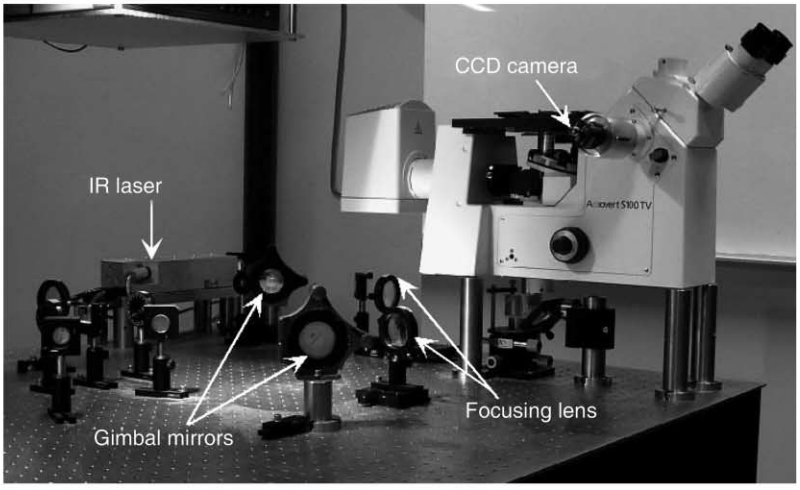
Optical tweezing stations broadly fall into three categories: (i) stand-alone systems; (ii) systems incorporating an optical microscope; and (iii) completely configured commercially available systems. Stand-alone systems feature optical tweezers built around a microscope objective and suitable optical components. These systems provide the maximum flexibility and can be easily reconfigured to adapt to different experimental needs. Furthermore, they also tend to provide the best mechanical stability, if configured properly. Figure 2a shows a picture of a stand-alone optical tweezing setup from our laboratory. This system is capable of the basic tasks of trapping, holding, and positioning microscopic objects. The most common optical tweezing stations are integrated into commercial light microscopes as sketched in Figure 3. The advantage of these hybrid systems is that the commercial microscope provides a flexible, optimized optical imaging platform, in which the optical trapping system can be integrated with only moderate difficulty. Figure 2b shows a picture of a typical optical tweezing station built around a commercial inverted microscope and mounted on a vibration–isolation table. This system includes an acousto-optic modulator that enables beam scanning for line-tweezing experiments. At the time of this writing, optical tweezing systems range from less than $10 000 for basic demonstration stations to between $50 000 to $100 000 for more sophisticated hybrid systems.

Commercial fully integrated systems are available from P.A.L.M. Microlaser Technologies AG (Bernried, Germany), Cell Robotics, Inc (Albuquerque, NM, USA), and Arryx (Chicago, IL, USA). Figure 2c shows a commercial optical tweezers set-up available from P.A.L.M. Microlaser Technologies.

When designing an optical tweezing station, the first item to consider is the choice of laser system. A simple single beam tweezing system can be built around a laser with less than 100 mW of power. Even though the minimum power required for trapping is a few milliwatts in the laser focus, there is considerable power loss in the tweezing and microscope optics. Also, the trap stiffness and thus the maximum trapping force decreases linearly with intensity, therefore most optical tweezers typically use lasers with an output power of 1 W. In some special applications, when beam scanning or multiple traps are required, the power requirement can be significantly larger. The wavelength of the laser also affects the capability of the trap in that the optimum wavelength depends on the size of the objects that are to be trapped. In general, the strongest traps occur


CCD camera
Movable X–Y stage
Microscope objective
Focusing lens
Trapping laser beam path
Movable rail

(a)


CCD camera
IR laser
Focusing lens
Gimbal mirrors

(b)



(c)

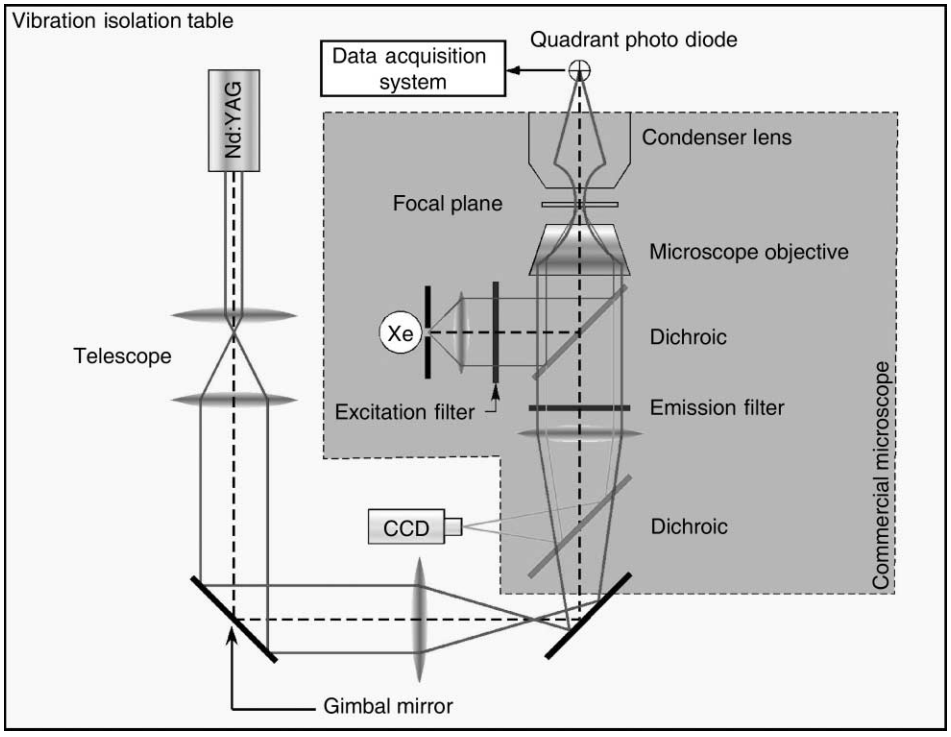


Figure 3 A sketch showing the instrumentation of a typical optical tweezing system incorporated into a commercial microscope.

for particles of size comparable to the wavelength of the laser. Another important concern in the choice of wavelength is possible damage to the sample from absorption and heating. In general, biological material is more transparent to the longer wavelengths of the IR spectrum. Studies on living cells have shown that generally, a range of wavelengths of 700 nm to 1100 nm, termed 'the optical window' of cells and tissue, is most suitable. Importantly, at higher wavelengths there is a reduced possibility of damage from multiphoton effects. DNA (260 nm) and proteins (280 nm) are strong absorbers in the ultraviolet range. Due to multiphoton effects, wavelengths between 700–800 nm have been seen to affect cell metabolism and lead to cell death within seconds. However, at wavelengths greater than 1000 nm, the simultaneous absorption of four photons required is unlikely at typical laser intensities used in trapping experiments.

The beam quality of the laser also restricts the choice of laser system. Diode lasers, in particular, have poor beam quality since the beam is often elliptical in shape and astigmatic, leading to a larger inefficient trap. Over the years, Nd-YAG lasers, at 1064 nm, have emerged as the lasers of choice, providing the best balance between price, power, and quality.

To achieve stable trapping it is essential to create an optical potential in which the gradient forces dominate the scattering force as discussed in the Theory section above. This suggests that the higher the numerical aperture (NA) of the microscope objective the stronger and more stable the trap will be. Oil or water-immersion microscope objectives with $NA > 1.2$ are commonly used. Further, this requirement also implies that the back aperture (rear entrance) of the microscope objective needs to be completely filled or, in the case of a Gaussian beam, overfilled to a certain degree to achieve optimal trapping efficiency. Since most commercial objectives are corrected for achromatic and spherical aberrations in the visual region of the spectrum, only

Figure 2 (a) A stand-alone tweezing station built around a microscope objective mounted on a vibration isolation table. (b) A typical tweezing station incorporated into a standard commercial microscope. This setup features dual-steerable optical traps and includes a CCD camera. Position detecting instrumentation can be mounted on the existing microscope stage. (c) The commercially available PALM® Combisystem utilizes an IR laser (wavelength 1064 nm) and a UV laser (wavelength 377 nm) coupled via the epi-fluorescence path to a research microscope. The system advertises automated and noncontact laser microdissection as well as trapping, sorting, and positioning. (**Figure 2:** (c), courtesy of P.A.L.M. Microlaser Technologies, Bernried, Germany.)

a small fraction of the working distance is usable for viable trapping. Normally this is limited to within 40 μm of the cover slip.

In addition to providing a beam large enough to overfill the back aperture, there are two other criteria that the external optics have to satisfy: first, the optical trap should be parfocal with the specimen plane. This enables the visualization of trapped objects. Second, the laser beam should be able to pivot around the back aperture without moving laterally to maintain the same degree of overfilling. This allows for a stable two-dimensionally movable trap. In practice, these requirements are fulfilled by using a telescopic system of lenses and Gimbal mounted mirrors. The latter are special mounts that enable a mirror to be tilted around its center position without lateral displacement. The telescope expands the incoming beam to fill the back aperture of the microscope objective and also images the back aperture onto the center of the Gimbal mounted mirror. As seen in Figure 4, since the back aperture of the microscope objective and the Gimbal mounted mirror are in conjugate planes, a small tilting of the mirror will only result in a change in the angle of the laser beam entering the microscope objective. The position of the beam will be unchanged, as will the degree of overfilling.

More sophisticated optical tweezer designs use acousto-optic modulators and galvanometer scanning mirrors to create scanning line tweezers and/or multiple traps. The basic approach is to deflect the laser beam to create multiple traps or to scan a single trap to create a one-dimensional potential well. Additionally, a new generation of holographic optical tweezers, capable of creating three-dimensional arrays of trapped particles, has emerged. Computer generated holograms that modify the phase and amplitude of the trapping laser are reverse engineered from the desired two- or three-dimensional optical trapping patterns. Furthermore, constructing such holograms from liquid crystal light modulators allows for dynamic control of the trapping pattern, yielding a fully computerized reconfigurable trapping system. An interesting example of the application of holographic optical tweezers to colloidal transport is shown in Figure 5, the first demonstration of 'kinetically locked-in states' where particles traveling across a potential mesh display a preference for certain paths.

It is to be noted that the optical tweezer is fundamentally a constant-extension technique, that is, it attempts to maintain the extension of a stretched molecule whereas the force is varied. However, it is of interest to study the action of molecular motors under constant force conditions. A basic optical tweezers setup can be modified to include an electronic feedback loop to provide constant force conditions. More recently, scanning-line optical tweezers have been adapted to implement all-optical constant force schemes that dispense with the need for electronic feedback.

An important aspect of any force measurement is the ability to accurately image and measure the position of a trapped particle. Standard digital video microscopy techniques yield a resolution of ~5 nm but have limited bandwidth and thus cannot be used for systems that employ feedback control of laser intensity or trap position. In these systems, the forward- or back-scattered trapping laser light from the trapped particle can be imaged onto a calibrated quadrant photodiode to provide sensitive position detection at high bandwidths (~1 kHz). If the photodiode is fixed in a plane conjugate to the back focal plane of the microscope objective, the trap can

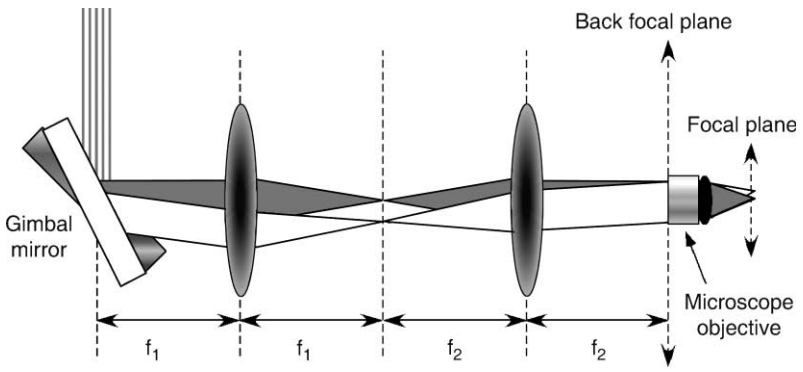


Figure 4 A sketch showing the working of a gimbal mounted mirror in two positions (shown as shaded and as unshaded). For small tilts of the mirror, the beam is seen to 'pivot' about the back aperture of the microscope objective resulting in the displacement of the optical trap in the focal plane. Note that for both beams the degree of overfilling of the back aperture of the microscope objective is unchanged, a crucial requirement for movable traps.

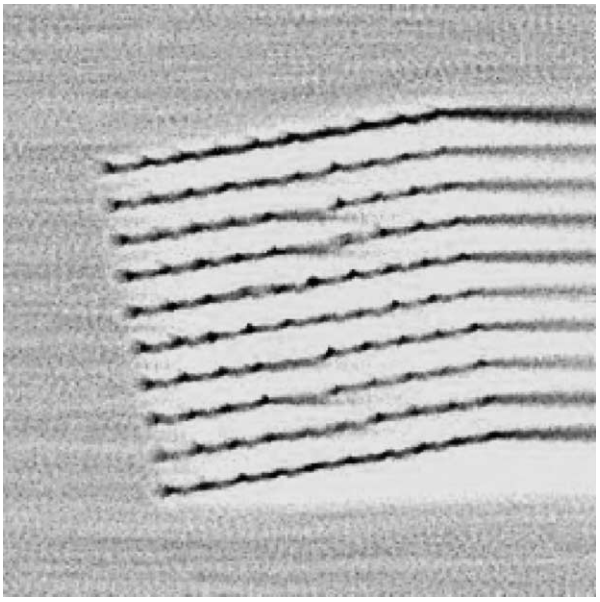


Figure 5 Data showing the superimposed paths of approximately 18 000 micron-sized particles flowing across a 10×10 grid of optical traps created by holographic optical tweezers. The particles are deflected laterally assuming a path dictated by the potential energy landscape and display kinetically locked-in states. This technique can be used to sort particles based on size, refractive index or charge. (Picture courtesy of D. Grier.)

be moved without having to move or recalibrate the photodiode detector. Using the quadrant photodiode method 2- and 3-dimensional position sensing is possible. A further improvement is found in schemes that use an additional laser dedicated to position sensing. The most sensitive method, however, for position sensing is interferometry. In this scheme, the trapping laser beam is split into two orthogonally polarized beams that are recombined after passing through the microscope objective and the sample. For a particle perfectly centered in the trap, the resulting combined beam has a circular polarization. In other cases, the bead displacement is sensitively reflected in the ellipticity of the polarization.

Applications

Colloids

Optical tweezers have been widely used to probe the interaction in colloids and other complex fluids. In fact, an optical tweezer apparatus was included in the instrumentation on the International Space Station for use with colloidal crystallization experiments in space. Various aspects of colloidal interactions have been studied using optical tweezers including mutual- and self-diffusion coefficients, the effects of hydrodynamic and electrostatic interactions, and depletion layer effects. More sophisticated scanning-line tweezing setups have been used to probe entropically driven self-assembly and interactions of colloids in polymer suspensions.

Blinking optical tweezers provide an elegant method for positioning colloidal particles in configurations of interest in a solution. When the tweezers are switched off the colloids respond to their environment, reacting to thermal forces and interacting either with other colloids or interfaces. The trajectories of the released colloids can be tracked and analyzed in 2D using digital video microscopy or 3D if combined with an evanescent illumination scheme. Switching on the tweezers again results in a return to the initial configuration, yielding a reproducible and tightly controlled experiment to examine colloidal behavior.

Biopolymers

The study of biopolymers has benefited from the development of optical tweezers. By chemically attaching a trappable latex microsphere to a biopolymer of interest, the molecule can be stretched, positioned, or moved. Measuring the response of the biopolymer to these external forces yields useful information such as elasticity, rigidity, and domain unfolding characteristics. This information obtained at a single molecule level has provided invaluable insight into the relationships between the structure and function of biopolymers. For example, in an early tweezing experiment, characterizing the intrinsic elasticity of λ-phage DNA, it was discovered that the double-stranded DNA undergoes a reversible structural transition when stretched at forces greater than 65 pN, extending to about 1.7 times its crystallographic contour length. These studies complement well-established structural techniques, such as X-ray crystallography and NMR, which provide information in a force-free environment. Furthermore, by anchoring one strand of DNA to a surface and pulling on the other strand, unzipping forces for DNA have been measured. Optical tweezers have also been combined with other mechanical tools, such as a micropipette, and used as sensitive tensiometers. Protein folding and unfolding experiments are generally harder, as their contour length is shorter than that of typical DNA fragments. Nevertheless, the folding and unfolding behavior of single giant muscle protein, titin has been probed by attaching a different latex bead to either end of the molecule. The first bead is held by a movable micropipette and the second bead is optically trapped. As the micropipette is moved, the position of the optically trapped bead provides information about the forces required to unfold the membrane domains of the protein.

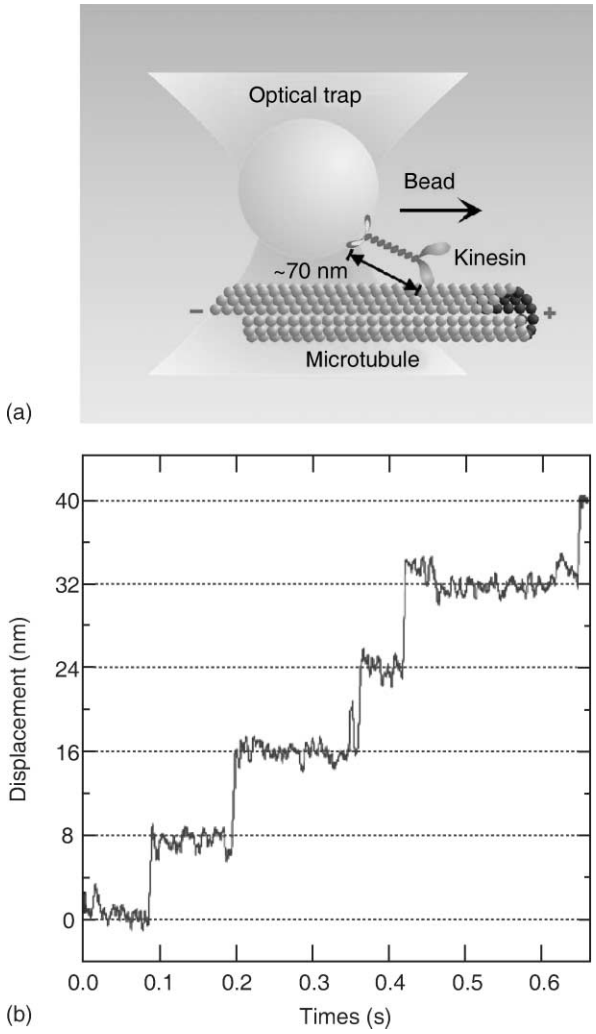


Figure 6 Observation of the stepping motion of the molecular motor kinesin. (a) A microsphere attached to a kinesin molecule is optically trapped; as the kinesin steps along a microtubule track which is immobilized onto a glass substrate, it pulls on the microsphere, whose motion is followed by an electronic feedback mechanism. (b) Motion of a single kinesin motor stepping against a 5 pN force. The data shows the discrete 8 nm steps of kinesin. (Figures courtesy of S. Block.)

Molecular Motors

Molecular motors (proteins that transduce chemical potential energy in to mechanical motion) have been the subject of intense attention since the early 1990s. Optical tweezers have played a crucial role in furthering the understanding of the mechanisms of molecular motors. The characteristics of molecular motor systems that have been studied with optical tweezers include speed, step-size, processivity, ATPs consumed per stroke, stall force, and static force. In fact, many of the advances in optical tweezers technology have been motivated by the study of molecular motors and have lead to spectacular results. An impressive example is the observance of the stepping motion of kinesin, as shown in Figure 6. Other experiments on molecular motors with optical tweezers have measured the stall force of motors as they move along their polymer track, or observed the transcription of DNA and RNA polymers, yielding interesting observations of the pauses during the transcription process.

Cells

The application of optical tweezers to cell biology was first demonstrated by Ashkin and co-workers in the late 1980s, by successfully trapping living objects: large viruses and bacteria. Since then, optical tweezers have been widely used to study various aspects of cellular biology. These include cell reproduction, growth, adhesion, mobility, and membrane elasticity. Additionally, the noninvasive nature of optical tweezers enables manipulations in the interior of an unopened object, allowing for the study of intracellular mechanisms by subcellular manipulations.

Optical tweezers have also been combined with various techniques to yield application-oriented tools for cell biologists. For instance, the combination of optical tweezers and state-of-the-art image recognition algorithms has been used to develop microfluidic cell sorters. In reproductive medicine, two of the major problems are penetration of the egg and sperm motility. A promising new solution combines laser zona drilling to pierce the egg envelope and optical tweezing to transport the sperm cells to the egg.

See also

Modulators: Acousto-Optics.

Further Reading

Ashkin A (2000) History of optical trapping and manipulation of small-neutral particle, atoms and molecules. *IEEE Journal of Selected Topics in Quantum Electronics* 6(6): 841–856.

Greulich KO (1999) *Micromanipulation by Light in Biology and Medicine: The Laser Microbeam and Optical Tweezers*. Basel; Boston; Berlin: Birkhauser.

Grier DG (2003) A revolution in optical manipulation. *Nature* 424: 810–816.

Lang MJ and Block SM (2003) Resource letter: LBOT-1: Laser-based optical tweezers. *American Journal of Physics* 71(3).

Sheetz MP (1998) *Methods in Cell Biology. Volume 55: Laser Tweezers in Cell Biology*. San Diego, CA: Academic Press.

Svoboda K and Block SM (1994) Biological applications of optical forces. *Annual Review of Biophysics Biomolecular Structure* 23: 247–285.

P

PHASE CONTROL

Contents

Phase Conjugation and Image Correction
Wavefront Coding

Phase Conjugation and Image Correction

E N Leith, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Phase conjugation of a wave is a process that has long been recognized. However, the technique for producing high-quality phase conjugation is relatively new. Also, the principal use of phase conjugation, the correction of phase errors in waves that have propagated through phase-distorting media, is relatively new, having originated in holography in the 1960s. Today, phase conjugation is an important area of modern optics. This article describes the conjugation of a wavefield by the process of holography and the holographic phase conjugation method of removing phase distortions from a wavefield, thereby permitting high-quality images to be produced from phase-distorted waves.

The basic meaning of a conjugated wave is explained with the aid of Figure 1. Figure 1a shows a wavefield propagating to the right. For simplicity, we show only one wavefront, i.e., one contour of constant phase. It is understood that a wavefield consists of a sequence of such wavefronts, with each wavefront differing in phase by 2π from the adjacent one, and separated by the wavelength λ. Of the two points, A and B, on the wavefront, we say that the wavefront at A lags that at B, since when the wavefront crosses a plane normal to the direction of travel, point A of the wavefront crosses that plane behind point B. The wavefield produces, at the plane P, a field $ae^{i\phi}$, where a is the amplitude and ϕ is the phase of the wavefront at the plane P.

We place a phase conjugator in the path of the wave. The wave passes through the conjugator, emerging as a phase conjugated wave, in which the point B, which formerly led point A in phase, now lags behind it by a corresponding amount. This illustrates the basic meaning of the term 'phase conjugation'. This device could be considered as a forward phase conjugator, since the phase conjugated wave travels in the same direction as the original wave.

Next (Figure 1b) we describe a backward phase conjugator. The conjugate wave travels in a direction opposite to that of the incident wave. The conjugate wave emitted by the phase conjugator seems to be

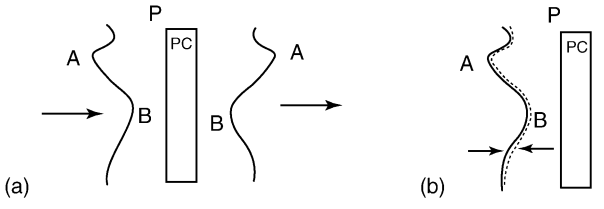


Figure 1 A phase conjugator. (a) shows a forward phase conjugator; (b) shows a backward phase conjugator. For clarity, in (b) the conjugated wave is shown with dotted lines. PC is phase conjugator, and P is the entrance plane of the conjugator.

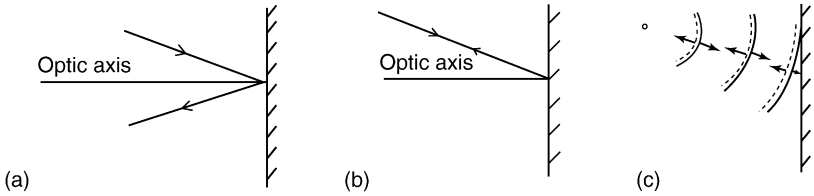


Figure 2 Light reflected from mirrors. (a) Light falling on a conventional mirror and being reflected; (b) light reflected from a conjugating mirror; (c) as in (b), except that illustration is with wavefronts instead of rays.

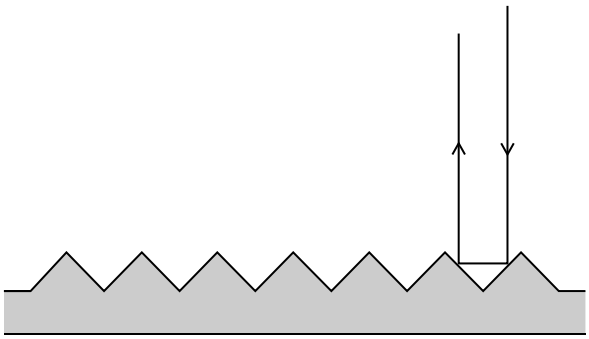


Figure 3 A ray of light reflected from a retroreflector (or corner cube).

identical to the incident wave; we show the two waves to be in coincidence, or more precisely, for purposes of clarity, almost in coincidence. The points A on the two wavefronts coincide, as do the points B. As before, on the incident wave, point A lags behind point B. On the exiting wave, since the wave is traveling in the opposite direction, point B is now lagging point A, i.e., is lagging in phase. In either case, forward or backward conjugation, the wave emerging from the conjugator has the complex amplitude $a(x, y)e^{-i\phi(x,y)}$.

A phase conjugator (backward variety) is like a mirror, in that it obeys the basic mirror law, that the angle of reflection equals the angle of incidence, but in a rather different way, as Figure 2 shows. In fact, the reflected ray travels back on itself. A simple way to generate a conjugate wave, and one that has been known for a very long time, is with the use of an array of retroreflectors, or corner cubes. As shown in Figure 3, each ray undergoes two reflections and emerges traveling in exactly the opposite direction, although with some displacement. If we considered the third dimension, there could be three reflections. The amount of displacement depends on where the ray strikes; if it strikes near the apex of the corner cube, the ray displacement will be slight; if at the edge, the displacement is greatest. Thus, the image formed in reflection is degraded. This geometrical degradation is greater for larger corner cubes. Thus, why not make the cubes very small? Then, however, the image is degraded by diffraction effects, since each corner cube constitutes a small aperture. Thus, there is an optimum size for the cubes; if too large, geometrical factors degrade the image; if too small, diffraction effects degrade the image.

Phase Conjugation by Holography

With holography, however, came a method of conjugating a wavefront that yields high-resolution imagery. The holographic process inherently produces a phase conjugate image, as was noted by Gabor in his early papers on holography. The holographic process is illustrated in Figure 4. For simplicity, we assume the object to be a planar transparency $s(x, y)$. A coherent light wave, generally a plane or spherical wave, passes through the transparency and a diffracted field u is produced at the recording plane P, as in Figure 4a. The field u is typically the Fresnel diffraction pattern of the object s, but if the propagation distance is sufficiently great, u could be the Fraunhofer pattern of s. Of course, lenses can be placed between the object and recording planes, thus modifying the diffraction pattern.

The field u is combined with a reference wave u_o, which is generally an off-axis planar or spherical wave. The two beams produce an interference pattern with intensity

$$I = |u_o + u|^2 = |u_o|^2 + |u|^2 + 2Re\, u_o^* u \qquad [1]$$

where the final term can be written as $u_o^* u + u_o u^*$; the * denotes complex conjugate. The process of forming the interference of the two beams, i.e., forming the intensity of their sum, has generated the conjugate term u^*. When the intensity is recorded, as on photographic film, for example, both terms are stored on the film. Suppose for simplicity that u_o is a plane wave $a_o e^{i2\pi f_o x}$, where a_o is a constant, and $f_o = (\sin\theta)/\lambda$, where θ is the angle that the reference wave makes with the optic axis and λ is the wavelength of the light. Also, u can be

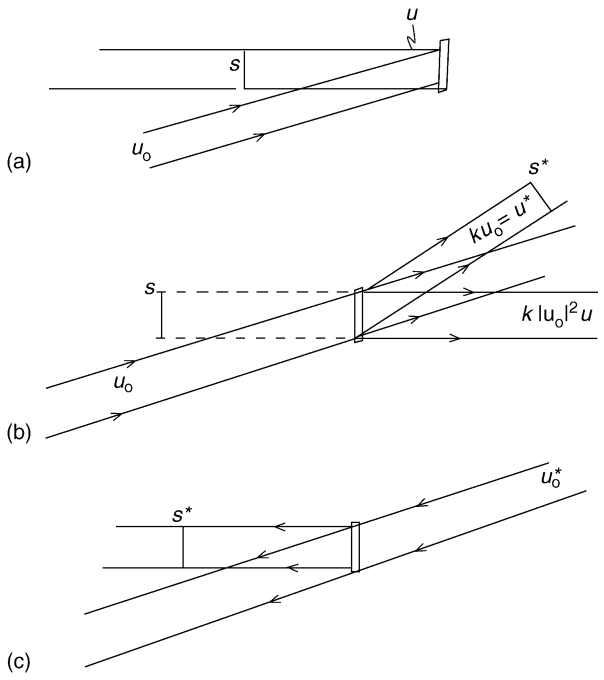


Figure 4 The holographic process. (a) Making the hologram; (b) reading out the hologram, using a readout beam that duplicates reference beam; (c) reading out with reversal of reference beam.

written as a product of an amplitude term a and a phase term $e^{i\phi}$, or $u = ae^{i\phi}$. Let the photographic plate, after exposure and development, have a transmittance t proportional to the intensity I, or $t = kI$, with k a constant; this is an oversimplification of the recording process, but preserves the essence of the holographic process. Now let the photographic plate, which after development becomes a hologram, be illuminated with a readout wave that, for simplicity, we let be u_o, a duplicate of the reference wave. Then, the field emerging from the hologram is

$$u_e = u_o k\left[|u_o|^2 + |u|^2 + u_o^* u + u_o u^*\right] \qquad [2]$$

Only the last two terms are of interest. These can be written as $k|u_o|^2 u$ and $ku_o^2 u^*$, respectively, or alternatively, as $ka_o^2 a e^{i\phi}$ and $ka_o^2 a e^{i(4\pi f_o x - \phi)}$. The term $k|u_o|^2 u$ represents a wave that is, to within a constant, identical to the object wave u; the other term represents a wave that is, to within a constant and a linear phase term, a conjugate of the original wave. The factor $e^{i4\pi f_o x}$ has the physical meaning that the wave is propagating in a direction 2θ given by $\sin 2\theta/\lambda = 2f_o$. The readout waves are shown in Figure 4b.

The so-called true wave u and the conjugate wave u^* thus propagate angularly away from each other and from the direct transmitted beam, which carries the uninteresting components $|u_o|^2$ and $|u|^2$.

However, a more suitable way to form the conjugate wave is to illuminate the hologram with a readout wave that is, again, a duplicate of the original reference wave, but propagating in the opposite direction, as in Figure 4c. The conjugate wave is now propagating along the z axis, along the path of the original object wave. It is readily shown that this counterpropagating wave is identical in shape to the original object wave recorded by the hologram, except that it travels in the opposite direction, exactly the situation portrayed in Figure 2. This wave propagates so as to form an image that is, to within a constant, identical to the original object, except for the counterpropagation, which produces a conjugate image $s^*(x, y)$. The image intensity, however, is $|s|^2$, which is identical to the original image intensity.

This conjugate wave was, in the early days of holography, a detriment, since without the use of the off-axis method (reference wave introduced at an angle to the object wave) the two images were inseparable, and the true image could be observed only against the background of the defocused conjugate image. Thus, an out-of-focus image overlaid the true image, making it noisy.

In 1962, the same year that the off-axis reference wave for separating the two images was introduced, a possible use for the conjugate image was suggested. If the original propagation path of the object wave were aberrated, as would occur if the wave propagated through an irregular medium, then the conjugate wave propagating through the medium would undergo a correction and would arrive at the original object position perfectly corrected, resulting in a sharp image, a process sometimes called wavefront healing. The idea lay dormant for 3 years, until in 1965, experimental demonstration of this idea was reported. In one case, the object was a transparency containing opaque lettering against a transparent background and the inhomogeneity was a piece of frosted glass, which resulted in extremely severe aberrations, so that the object wave, having traversed the diffuser, was no longer capable of forming an image. However, when the distorted wave was recorded holographically, conjugated, and sent back through the same diffuser, the result was a highly legible image.

The principle of wavefront healing is illustrated in Figure 5, when the inhomogeneity is an extended, or three-dimensional, structure. A point object O sends a spherical wave into the inhomogeneity. As the wavefront propagates, it becomes distorted, and loses its sphericity. When it emerges from the inhomogeneity, the wavefront is highly distorted.

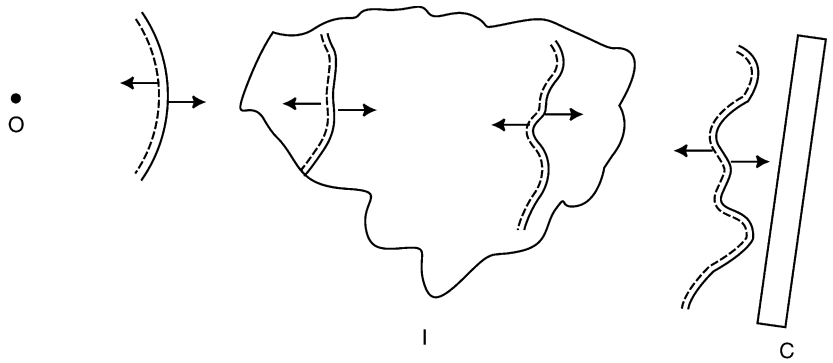


Figure 5 The wavefront healing process. C is the conjugator, I is the inhomogeneity, and O is a point object.

The wave is conjugated at the conjugator C and emerges as a back-propagating wave that retraces the original wave path through the inhomogeneity, gradually losing its aberrations and looking more like a spherical wave. When it emerges from the inhomogeneity, it is a perfect spherical wave that converges to a point image that coincides with the original object. At each plane throughout the inhomogeneity, the back-propagating conjugate wave (dashed lines) has the same shape as did the original wave when it crossed that plane (solid line).

This interesting wavefront healing property of holographic phase conjugation was the beginning of a new branch of optics, called phase conjugation, giving rise to a number of applications. However, the holographic method had the problem that it did not occur instantaneously, but had to await the processing of the photographic film on which the hologram was recorded. This delay limited the range of potential applications. Later, in the 1970s, nonlinear crystals became available that did the equivalent of holography, but in real time, or near real time – a process known as four wave mixing. There are yet other types of real-time phase conjugators with nonlinear devices.

Pseudoscopic Image

As long as the object is only two-dimensional, a function of x and y only, with no extent along the axial dimension, z, the intensity distribution of the true and conjugate images is identical. The conjugateness, being a pure phase phenomenon, disappears when the observable is the image intensity $|s|^2$, and not the complex amplitude s. Since the intensity is always the observable, whether the image is viewed by eye, or with a camera, the conjugateness generally is of no significance.

However, when the object for holography is three-dimensional, the situation is different, and the conjugate image exhibits some strange properties. These arise from conflicting cues pertaining to the three-dimensionality of the image. To explain, we review some basics about how a three-dimensional world is perceived by a viewer. First, there is the stereo effect, resulting from the viewer having two eyes, with each eye seeing an object from slightly different positions, a phenomenon called binocular disparity. The mind uses this stereo effect to perceive three-dimensionality. But there are other cues that also contribute to perception of three-dimensionality. Another is parallax; as the viewer moves his head, objects farther from the observer change their angular position more slowly than objects close by. The observer can sense the three-dimensionality of an object, even with only one eye, by moving the head, thereby inducing parallax. Also, there is interposition, whereby near object portions block the viewing of object portions lying behind them, again giving a sense of depth. There are still other cues, more subjective, such as perspective, parallel lines converging toward infinity, shadowing, etc. With these observations in mind, we consider the hologram under illumination and the two images that it forms (Figure 6). As depicted in Figure 6, these two images lie in mirror symmetrical positions relative to the hologram. Both images have the same side facing the hologram. It must be this way, because when the hologram was recorded, it was side **a** that faced the recording plane. And both images must therefore have side **a** closer to the hologram.

When a viewer observes the true image (a virtual image), he views it looking through the hologram, and he sees the image at the position where the original object was when the hologram was recorded, and it is, in its visual properties, indistinguishable from the actual object; it has all of the three-dimensional and the parallax properties that the original object would exhibit. However, when the viewer observes the conjugate image, he must view it

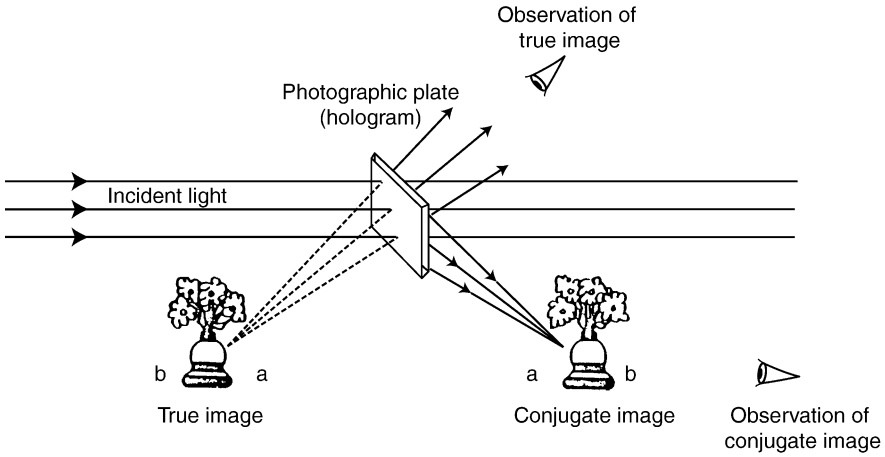


Figure 6 Formation of the true and conjugate images from a hologram.

from the side of the image opposite from the hologram, i.e., he views the image from the back side. When a viewer observes the conjugate image, he sees it from the back side (side **b**). All of the above cited cues are in agreement about the orientation of the image, viz., that side **b** is closer to the observer than is side **a**. That is, all cues but one. The conjugate image is formed as the hologram recording plate saw it, with side **a** being closer. Thus, all interpositions are formed with the same relation that the object had with respect to the recording plate. The side **a** is the part of the object that was closer to the hologram, so that when interpositions occur, side **a** blocks side **b**. Since the observer thus sees side **a** partially obscuring side **b**, he must conclude that side **a** is closer than is side **b**. However, this observation is in direct conflict with the other cues, especially the stereo and the parallax cues, which are themselves strong cues, as is the interposition cue. The conflict of these cues, with two of them interpreting the image as being formed with one depth orientation, the other with the opposite, gives a number of strange effects. For example, as the viewer moves his head so as to induce parallax, the object appears to rotate. If the image is that of a human head, for example, the observed effect is that as the viewer changes his position, the head appears to rotate in synchronism with the observer, as if the image had its gaze fixed on the viewer, an effect that is often perceived as being eerie.

Similarly, if the object were two flat plates, (Figure 7) in the situation depicted, the planes have no overlap, i.e., no interposition, as seen by the observer. And the observer unequivocally perceives plate A as being closer than plate B. However, if the viewer moves his head to a position where

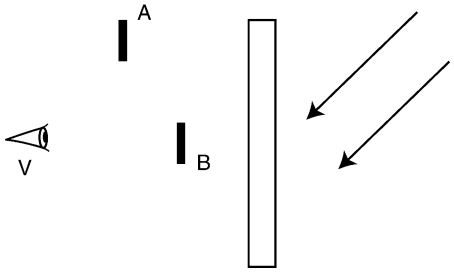


Figure 7 Pseudoscopic conjugate image formed from an object consisting of two plates. The viewer V perceives plate A as being closer than B and there are no conflicting cues, since there is no interposition. However, if the observer moves his head upward, there is partial interposition, thus conflicting cues, and the sense of three-dimensionality is lost.

interposition occurs, the conflict of cues arises, A becomes partially blocked by B, which is quite correct as observed at the position of the hologram, but is quite incorrect from the viewer's vantage point. The result of this conflict is that the viewer, who previously perceived the three-dimensionality, now sees the three-dimensionality perception suddenly collapse, and the sense of three-dimensionality is lost. Other strange perception anomalies can also be produced. Such images, with conflicting depth cues, are called pseudoscopic, as opposed to orthoscopic images, where the cues do not conflict.

One Way Phase Conjugation

A significant limitation of phase conjugation aberration correction is that the image and object are on the same side of the inhomogeneity. One would usually want them to be on opposite sides. In a modified form

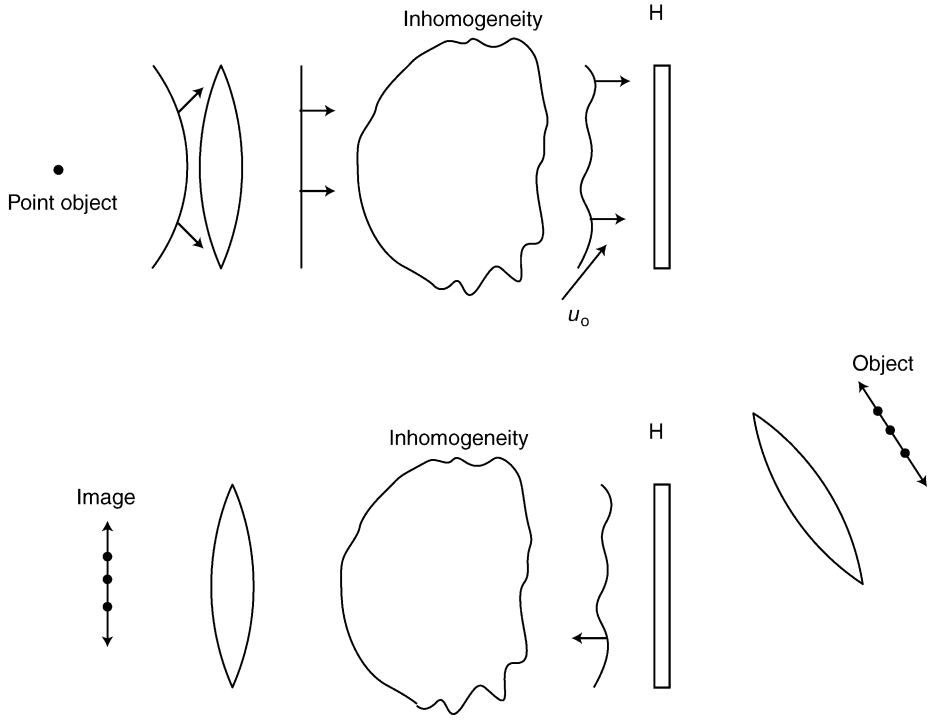


Figure 8 One-way phase conjugation. Top: Making the hologram. Bottom: The viewing process. We show three points on the object. The center point forms a well-corrected image, but the other two points produce imperfectly corrected images.

of the phase conjugation process, shown in Figure 8, this result is achieved. The first step of the process is carried out by passing a plane wave, instead of the object wave, through the inhomogeneity. This plane wave originates from a point source located at the focal plane of a collimating lens. The plane wave thus carries information about the inhomogeneity to the recording plate, where it is combined with a reference wave and a hologram of the distorted plane wave is recorded. Thus, information about the phase defect is stored on the hologram.

Next, the hologram is read out, not with a counterpropagating replica of the reference wave as before, but instead with an object-bearing wave that passes through the hologram. This wave emerges from the hologram and now has on it the recorded aberration information. As it counterpropagates along the path of the original object wave, it passes through the inhomogeneity, whereupon the aberration is cancelled and a corrected image forms at the focal plane of the collimating lens, where previously the point source had originated. The goal is achieved; the object and the corrected image are now on opposite sides of the inhomogeneity.

This correction is only approximate. To explain, we think of the object as being a collection of many points. The radiation from each point is converted into a plane wave by the lens, with different object points producing plane waves propagating at different angles. These plane waves then impinge on the hologram and emerge with each one bearing the aberration stored on the hologram. They continue on their paths and enter the inhomogeneity. The plane wave component, whose path direction matches that of the reference wave used in recording the hologram, will emerge from the inhomogeneity as a perfectly corrected wave, thus forming a well-corrected point image at the position of the original point radiator. However, another plane wave component, traveling in an incorrect direction, will be laterally displaced from the proper position for entering the inhomogeneity, and the aberration correction will be imperfect, the degree of imperfection being in direct relation to this displacement error. The displacement error is the product of the angular error and the propagation distance between hologram and inhomogeneity. If the inhomogeneity is slowly varying as a function of x and y, and if the object does not have a large extent, so that the range of angles is small, the image degradation will be small.

If the last two conditions do not apply, two options are available. First, the distance between inhomogeneity and hologram can be reduced, either

by bringing the hologram closer to the inhomogeneity, or by imaging the hologram onto the inhomogeneity. The correction is now valid over a wider extent of the image, especially if the inhomogeneity can be treated as planar rather than extending over a volume. Alternatively, a scanning process can be carried out on the object beam falling on the hologram, so that in sequence, each plane wave component assumes the proper angle to counterpropagate along the path for perfect correction, and will form a well-corrected image point. This output spot then falls on a pinhole positioned to separate it from other, imperfectly corrected points. Then, over time, all points of the object will form well-corrected image points, and the entire object will thus be well-imaged. The one way phase conjugation method was first used to correct the spherical aberration of a lens, and the result was quite successful.

Phase conjugation has become an important branch of modern optics. The process is often carried out using photographic materials. More often nowadays, it is carried out using nonlinear crystals. Some of the crystal methods are essentially holography in real time, but others, such as phase conjugation by Brillouin scattering, go beyond holography.

See also

Nonlinear Optics, Basics: Nonlinear Optical Phase Conjugation. **Phase Control:** Wavefront Coding.

Further Reading

Caulfield HJ (ed.) (1979) *Handbook of Optical Holography*. New York: Academic Press.

Goodman JW (1996) *Introduction to Fourier Optics*, 2nd edn. New York: McGraw-Hill, chap. 9. This chapter is devoted to holography, with section 9.12.3 dealing with phase conjugation aberration correction.

Kogelnik H (1965) Holographic image projection through inhomogeneous media. *Bell Systems Technology Journal* 48: 2451–2452.

Leith EN and Upatnieks J (1965) Holograms: their properties and uses. *SPIE Journal* 4: 3–6. This is the earliest publication demonstrating the principle of wavefront healing by holographic phase conjugation.

Smith HM (1975) *Principles of Holography*, 2nd edn. New York: Wiley.

Wavefront Coding

W T Cathey, University of Colorado, Boulder, CO, USA

E R Dowski, CDM Optics, Inc. Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

We describe a new paradigm for designing imaging systems that have performance that is otherwise unobtainable. These are hybrid systems that use the optics of the system to acquire an 'intermediate' image, which is processed to produce the final image. The optical components of the system are designed jointly with signal processing to optimize overall performance. It is possible to extend the depth of field of an imaging system, for example, by a factor of eight to ten. On the other hand, by concentrating on the increase of the depth of focus in the detector plane, it is possible to use that increase in the depth of focus to make the system invariant to focus-related aberrations. By doing this, it becomes possible to design a single-lens imaging system with small f-number and large field of view. Another example of the sort of new imaging system that can be designed is a two-lens zoom lens that images with a large field and a very short overall length.

First, the term 'Wavefront Coding' is defined as employed with imaging systems, and a description is given of how such coding is achieved. Examples of means of applying Wavefront Coding are given along with the analysis and design tools that are useful in designing hybrid optical/digital imaging systems. Finally, some potential future applications of Wavefront Coding in imaging systems are suggested.

Background and Basic Concepts

Some early suggested modifications to imaging systems can be viewed as wavefront modification, even though they were not described in that manner at the time. These include the work of Ojeda-Castañeda that described masks to block portions of the lens to extend the depth of field. In addition, it has long been known that the depth of field is slightly greater in lenses that have spherical aberration. In both of these cases, the wavefront is modified, in one case by changing the wave amplitude, and in the other by a deviation from the traditional ideal phase front. However, in both cases, the increase in the depth of

field was only a few percent. Häusler extended the depth of field by requiring that the focus be continuously changed during the exposure time of the camera. This resulted in a modulation transfer function (MTF) for the imaging system that was the incoherent superposition of a spread of MTFs for the different foci during the exposure. To form an acceptable image, signal processing was used.

The use of Wavefront Coding here refers to new optical systems arrived at by designing the modulation of the wave and signal processing to achieve previously unobtainable imaging modalities. These new imaging systems can have more than an order of magnitude increase in depth of field, for example. The coding of signals so that particular information is optimally conveyed is not new. For example, in radar, coding of the pulses is done to optimally provide information concerning a target's range. At the radar receiver, signal processing is required to extract the range information. In a similar manner, an imaging system can have a coded wavefront in order to provide range or misfocus information in a manner such that signal processing can easily extract the desired information in the scene. Alternatively, the wavefront can be modified so that the imaging system is invariant to object depth or to misfocus of the system, so that, after processing, good resolution can be obtained over a large image depth.

A primary feature of Wavefront Coding in imaging is that the optical image-gathering portion of the system can be designed in combination with the design of the signal processing. This is normally not done in the design of imaging systems. It is done, however, in tomography, coded aperture imaging, and sometimes, interferometric imaging. In 1984, a group pointed out the potential of increasing the performance of imaging systems by jointly designing the optics and the signal processing. Since that time, major advances have been made in doing this. Now, several analysis and design tools are available, and are described in the following section.

Theory and Applications of Wavefront Coding

Wavefront Coding refers to a modification of the imaging system by placing specialized phase plates (or phase and amplitude masks) in the aperture stop, or in an image of the aperture stop, of the imaging system. This could be a phase plate as a separate optical element, or, if a surface of a lens is close to the aperture stop, the Wavefront Coding surface can be superimposed onto the normal lens surface. Processing to decode the image that is formed by the detector array (the intermediate image) is related to the coding that is impressed on the wavefront. The term 'signal processing' also includes the spatial integration provided by the width and height of the pixels. The final image is hence a function of the optics that form the intermediate image and the signal processing that produces the final image.

In the following section, we discuss the Woodward function as used in analysis and design of hybrid imaging systems, and show visualizations of the effect of Wavefront Coding on the image performance of two dramatically different types of systems. Then, other design techniques are discussed for imaging systems using Wavefront Coding and several examples of systems are given where the depth of field of imaging systems is extended by at least an order of magnitude. By concentrating on the increase in the depth of focus, rather than the depth of field, very simple imaging systems can be designed where Wavefront Coding makes the imaging system invariant to focus-related aberrations, and the shape of the lens is used to minimize the other aberrations.

Wavefront Coding Theory

The Woodward function

P. M. Woodward showed that the capability of a radar system to estimate both the range and the velocity of a target can be described by a mathematical function that became known as the Ambiguity Function. The Ambiguity Function shows how well systems that use different types of radar pulses can measure the transit time of the pulse (hence the range to the target) and the Doppler shift in the frequency of the returned pulse (hence the velocity of the target). Because the parameters, range, and Doppler, are measured with different accuracy, there is ambiguity in the measurement, and the Ambiguity Function is a display of the tradeoffs that are possible, Brenner, Lohmann, and Ojeda-Castañeda showed that the same mathematical function, with differently defined variables, can represent the Optical Transfer Function (OTF) of a one-dimensional imaging system.

The Ambiguity Function for a radar system is given by

$$A(u,v) = \int P(x+u/2)P^{*}(x-u/2)\exp(\mathrm{i}2\pi vx)\mathrm{d}x \quad [1]$$

where $\mathrm{i} = \mathrm{sqrt}(-1)$, * denotes complex conjugate, u is time, v is frequency, x is a dummy variable, and $P(x)$ is a complex function representing the transmitted waveform of the radar system. The OTF of a one-dimensional imaging system for different values of

misfocus can be written as

$$H(u,\psi) = \int \left\{P(x+u/2)\exp(j(x+u/2)^2\psi)\right\} \times \left\{P^*(x-u/2)\exp(-j(x-u/2)^2\psi)\right\}\mathrm{d}x \quad [2]$$

where $P(x)$ represents the complex pupil function of the imaging system. The misfocus parameter ψ can be shown to be equal to the physical parameters:

$$\psi = (\pi L^2/(4\lambda))(1/f - 1/d_o - 1/d_i) = 2\pi W_{20}/\lambda = kW_{20} \quad [3]$$

where L is the one-dimensional length of the aperture, and λ is the center wavelength of the narrowband illumination. The distance d_o is measured between the object and the first principal plane of the lens, and d_i is the distance between the second principal plane of the lens and the image plane. The focal length of the lens is given by f. The wavenumber is given by k and the traditional misfocus aberration constant is given by W_{20}. A comparison of the form of the ambiguity function and the OTF shows that:

$$H(u,\psi) = A(u, u\psi/\pi) \quad [4]$$

The mathematics of the Woodward function as used for the Ambiguity Function has been studied extensively. Hence one can apply the theory that was worked out for the Ambiguity Function to the study of imaging systems. This gives a tool that is, in many ways, more powerful than other forms of three-dimensional optical transfer functions. For example, it can be shown that the squared volume under the Woodward function is a constant. This means that whatever one does to modify the Woodward function, there is a constraint on what can be done. Another constant is the power under a vertical cut of the Woodward function at a particular value of spatial frequency. This means that there is constant power for all misfocus values at any given spatial frequency. It can be bunched up about the normal in-focus position, or it can be spread out as shown below, but the squared area does not change.

Examples of the Woodward function for different imaging systems

A display of the Woodward function for a high-quality, one-dimensional lens is shown in Figure 1a. The horizontal line and the slanted line represent in-focus and out-of-focus slices through the Woodward function. A projection onto the horizontal axis gives

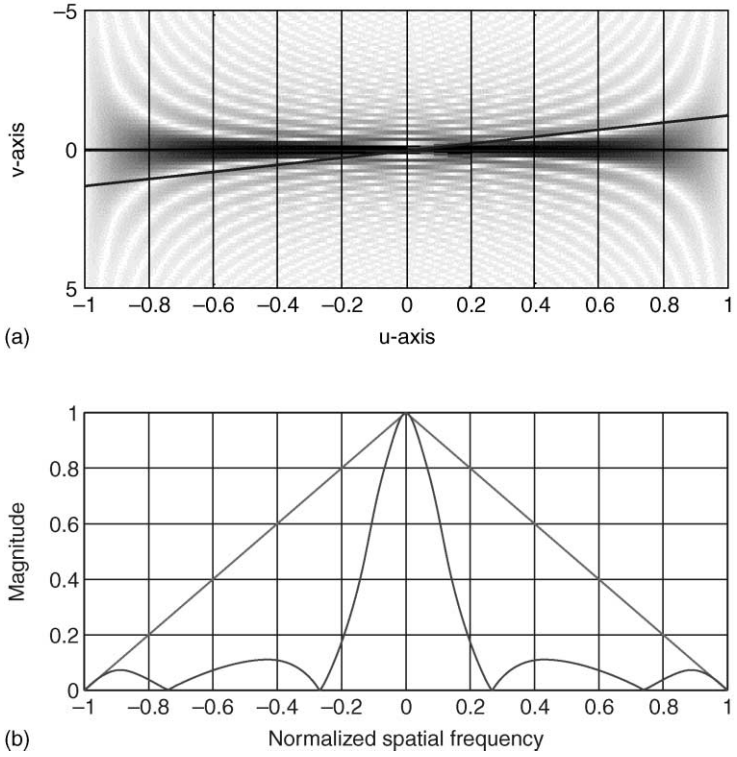


Figure 1 (a) The Woodward function for a high-quality one-dimensional lens with slices that represent in-focus and out-of-focus imaging. (b) The corresponding values of the modulation transfer function for those two cases.

the OTFs for those two focal conditions. The magnitudes of those two OTFs, the MTF, is shown in Figure 1b. Note that the out-of-focus MTF goes through a zero, where there is a phase shift, which gives a contrast reversal in the image.

Extended depth of field. Figure 2a shows the corresponding display of the Woodward function for a one-dimensional lens with a cubic phase function, $\phi = \alpha x^3$ added in the aperture stop. The appearance of the Woodward function immediately shows that this imaging system will have an increased depth of field, because the Woodward function shows that the energy is spread over a larger misfocus region. The slices in the Woodward function are taken for the same misfocus values as were used for the images in Figure 1. Notice that the MTFs of Figure 2b are very similar, and that the MTF for the misfocused case does not go through a zero. This implies that the imaging system will have an extended depth of focus.

The analogy to the Ambiguity Function of radar was key in selection of the cubic function for early demonstrations of extending the depth of field of an imaging system. Work in the radar field had shown that a radar pulse, where the carrier frequency increases quadratically with time, has an Ambiguity Function that is identical to the distribution in Figure 2a. Because a quadratic in time is a cubic in phase, the cubic phase function:

$$\phi = \alpha x^3 \qquad [5]$$

was selected as one that will provide the same Woodward function, and hence an increase in the depth of field. Many families of phase distributions have now been found that extend the depth of focus of an imaging system.

Antialiasing. When the resolution of the optics of an imaging system exceeds the resolution of the image detector array, the high spatial frequencies cannot be detected. They are not lost, however. When a spatial frequency is sampled below the Nyquist sampling rate, the high spatial frequency shows up, or aliases, as a lower spatial frequency. This aliased frequency is equal to the sampling frequency (determined by the pixel spacing) less the difference between the high frequency that has been undersampled and the sampling frequency. This results in an image with broad bands (low spatial frequencies) across the image as is often seen in television images of patterns with high spatial frequency content, such as striped shirts. Because the signal power at a given spatial

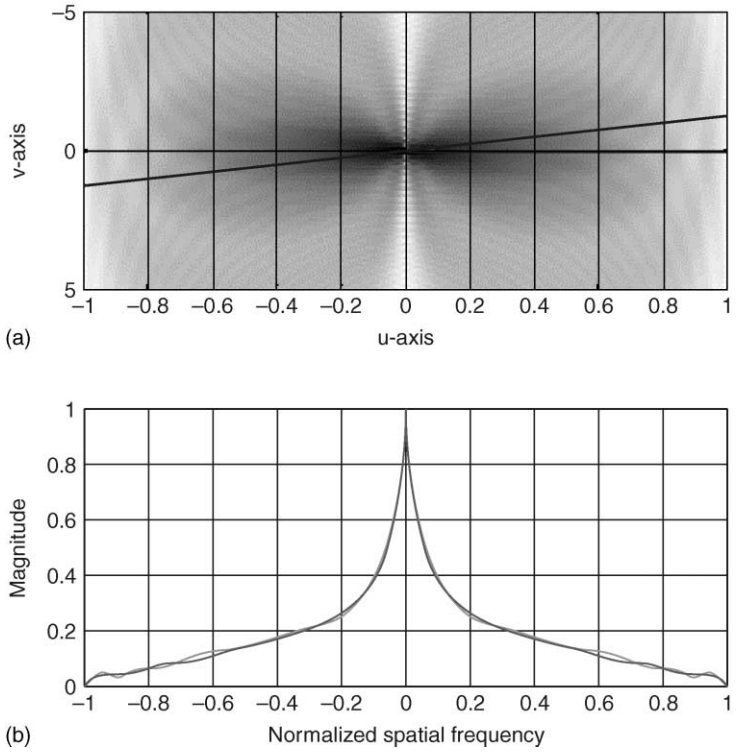


Figure 2 The Woodward function and the modulation transfer functions for a high-quality one-dimensional lens that has a phase plate with a cubic profile in the aperture stop. (a) The Woodward function showing a broadened pattern that is indicative of an increased depth of field. (b) The MTFs that are associated with the slices that represent in focus and out of focus.

frequency over all misfocus (the squared area under a vertical slice of the Woodward function at that spatial frequency) is constant, these frequency components cannot be removed. They can, however, be reduced at a given focus value by spreading the Woodward function vertically. Figure 3 shows a Woodward function that would be ideal. Below the cutoff spatial frequency, all of the signal power is in the focal plane, and beyond the cutoff frequency, the power is spread as much as possible over a large misfocus region. There is no assurance that a phase distribution can be found to give a desired Woodward function, however. Figure 4 shows the effect of one phase function, given by

$$S(r, \theta) = r^3 \cos(3\theta) \qquad [6]$$

upon the MTF of the imaging system. The dashed vertical line shows the cutoff spatial frequency that is due to the sampling rate of the pixels. Note that in Figure 4, the magnitude of the Wavefront Coded spatial frequencies that are aliased have lower values than do those that are passed by a traditional full aperture F/2.5 system. Stopping the traditional system down to F/25 reduces the aliased MTF to an amount close to that obtained with the wide aperture and Wavefront Coding, but it also reduces the amount of light by a factor of 100.

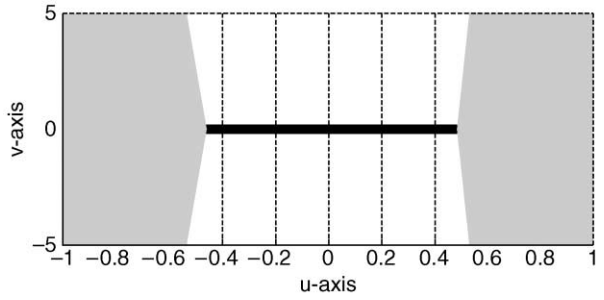


Figure 3 The Woodward function for an ideal antialiasing filter.

Wavefront Coding in Imaging Systems with Extended Depth of Field or Focus

Figure 2 shows the Woodward function for one phase distribution (cubic in x and y for a two-dimensional imaging system) that increases the depth of field of an imaging system. The cubic function was selected by using analogies to the Ambiguity Function of radar. Even though it is doubtful that this phase distribution could have been found by using conventional lens design techniques, it is instructive to show the ray traces and point spread functions that are obtained when this phase distribution is used in the aperture stop.

Ray trace analysis

Figure 5 shows the ray trace from a point object through a conventional (one-dimensional) lens and through a lens that has been modified with the cubic profile of eqn [5]. Figure 5a shows the rays through a conventional lens coming to a focus, and Figure 5b shows the corresponding rays as they go through a lens with a cubic profile added. In the latter case, the

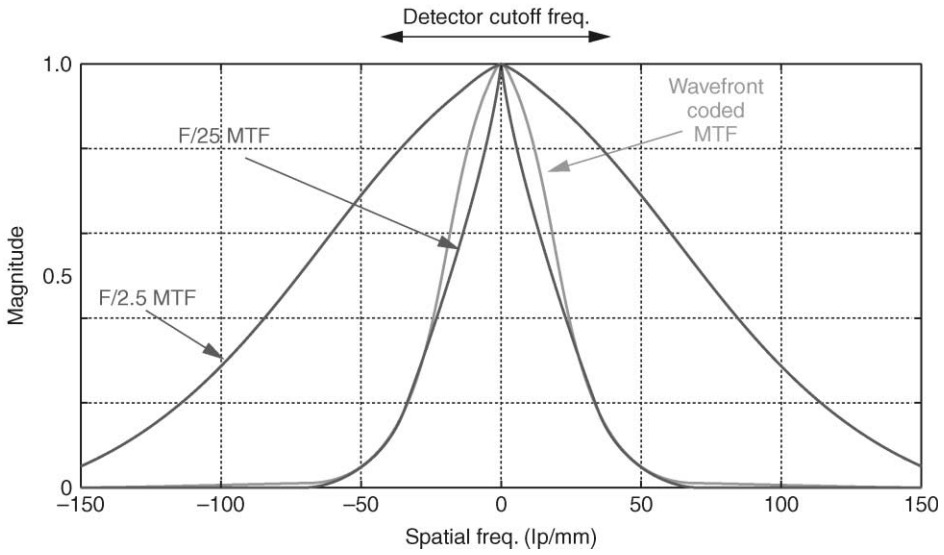


Figure 4 Normalized MTFs for traditional optics and Wavefront Coding. The full aperture F/2.5 system has a large amount of optical power beyond the detector cutoff frequency of 50 lp/mm. With Wavefront Coding the aliased optical power can be greatly reduced. The traditional system aperture needs to be reduced to F/25 to compare to the antialiasing performance of the Wavefront Coded system. This reduction reduces the light level by a factor of 100 in comparison with the F/2.5 system.

rays cross, but have a relatively constant ray density in the region of what was the focal region. Figure 6 shows experimentally acquired point spread functions (PSF) for both cases. Note that the PSF changes very little over the region of misfocus. After signal processing, the PSF can be brought back to that of the conventional imaging system when in focus. Figure 7 illustrates the effect of the signal processing upon the intermediate image that is acquired by the optics. The object is a tilted bar pattern, so that a portion is in focus, and the remainder of the bar pattern is so far out of focus that there is a contrast reversal in the image. With the system that has a coded wavefront to give an extended depth of field, the entire image appears to have the same blur, no matter which part of the tilted bar pattern is observed. This is because the same PSF is convolved with the object distribution for all regions of misfocus. Signal processing removes the effect of the modified PSF, resulting in the image of Figure 7b. Figure 7c shows a trace through the image of Figure 7b. The amount of the misfocus that can be tolerated is a design parameter, and is discussed in the next section.

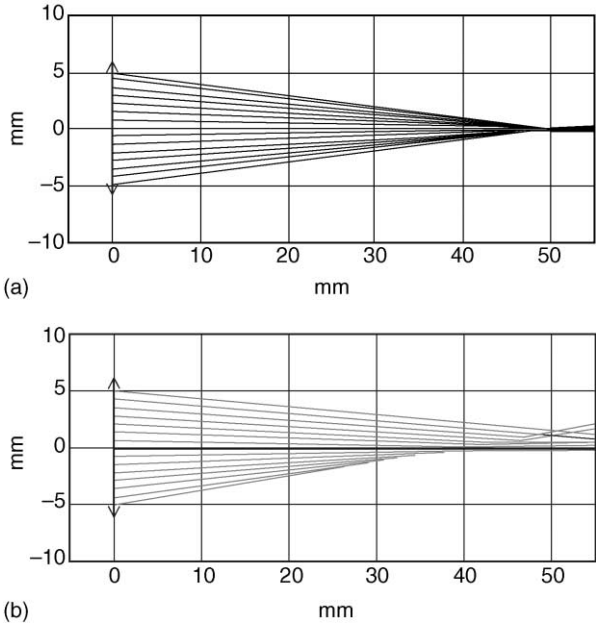


Figure 5 Ray traces from a point object through (a) a conventional one-dimensional lens, and (b) a lens that has been modified to have a cubic profile on the surface.

Signal processing requirements

It is necessary to do signal processing when using an imaging system that employs Wavefront Coding; the signal processing is needed to 'decode' the image that is formed by the detector array. This intermediate image is not a clear image. To obtain an intermediate image, the object distribution is convolved with a PSF, or impulse response, that has a shape similar to that

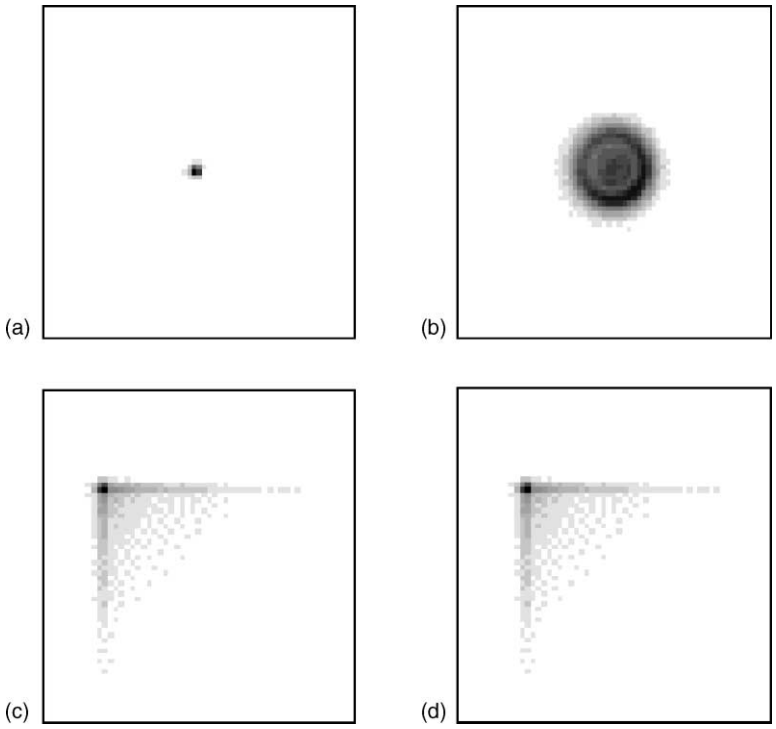


Figure 6 Point spread functions for a conventional lens in focus (a) and with large misfocus (b), and a lens that has been modified to have a cubic profile on the surface when in focus (c) and with a large misfocus (d).

of Figure 6c and d. To obtain the final image, the intermediate image must go through a deconvolution with the PSF. Because of the square law detection that occurs in the detector array, this deconvolution does not simply undo the effects of the convolution with the PSF that describes image formation. A straightforward digital convolution is required in most cases, and has been shown to run at video frame rates in software.

Parameter trade offs in imaging systems with extended depth of field

Depth of field and signal-to-noise ratio. The greater the phase modulation by the phase plate (with a cubic profile in this example), the greater the extension of the depth of field. The greater the phase modulation for extending the depth of field, the larger the PSF, and the greater the drop in the MTF, as seen in Figure 2b, which show the MTF of the imaging system. The use of the phase function to extend the depth of field causes portions of the MTF to sag as compared with the MTF of the in-focus conventional system. This reduces the signal-to-noise ratio. If there is a greater phase alteration, the MTF sags more.

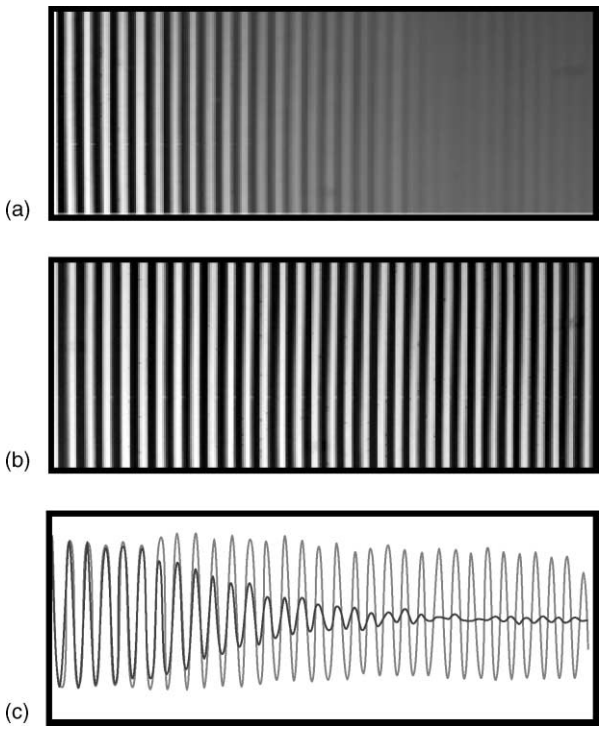


Figure 7 Image of a tilted bar pattern with an extended depth of field imaging system. (a) Image with a conventional imaging system. (b) Final image from the imaging system with Wavefront Coding. (c) Traces through the images to give quantitative comparisons of the image quality.

Focus budget and focus-related aberrations. Another set of trade offs can be seen by considering the imaging systems from the detector plane. Here, we see an increase in the depth of focus of the system. All of this increase in the focal depth can be used to increase the depth of field if the optics of the system are of high quality. Otherwise, the increase in the depth of focus can be used to compensate for focus-related aberrations. For example, chromatic aberration occurs when different colors focus to different planes. Figure 8 shows that if the depth of focus is extended, then there is a region where red, blue, and green, for example, are in focus. Figure 9 shows the red, blue, and green images of a US Air Force test pattern, as formed with a two-element imaging system with severe chromatic aberration, and with the same system when a Wavefront Coding element is used to extend the depth of focus.

Compensation can also be seen for the case of curvature of field, where the image of a plane object falls onto a curved surface. If an extension of the depth of focus causes the focal region to resemble a curved bar instead of a curved line, then it is possible to place a flat detector array in the curved region, and therefore obtain a good image after signal processing. Similar arguments can be made for other focus-related aberrations. This leads to the concept of a focus budget, where a portion of the increased depth of focus can be allocated to extend the depth of field by some amount, and other portions of the depth of focus allocated to provide invariance to focus-related aberrations.

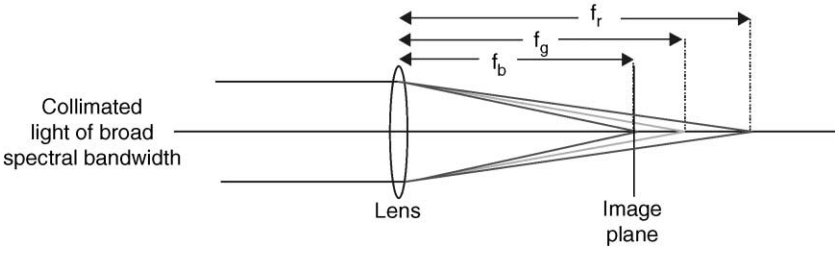


Figure 8 Ray traces showing different focal planes for red, blue, and green, and the effect of extending the depth of focus for each color.

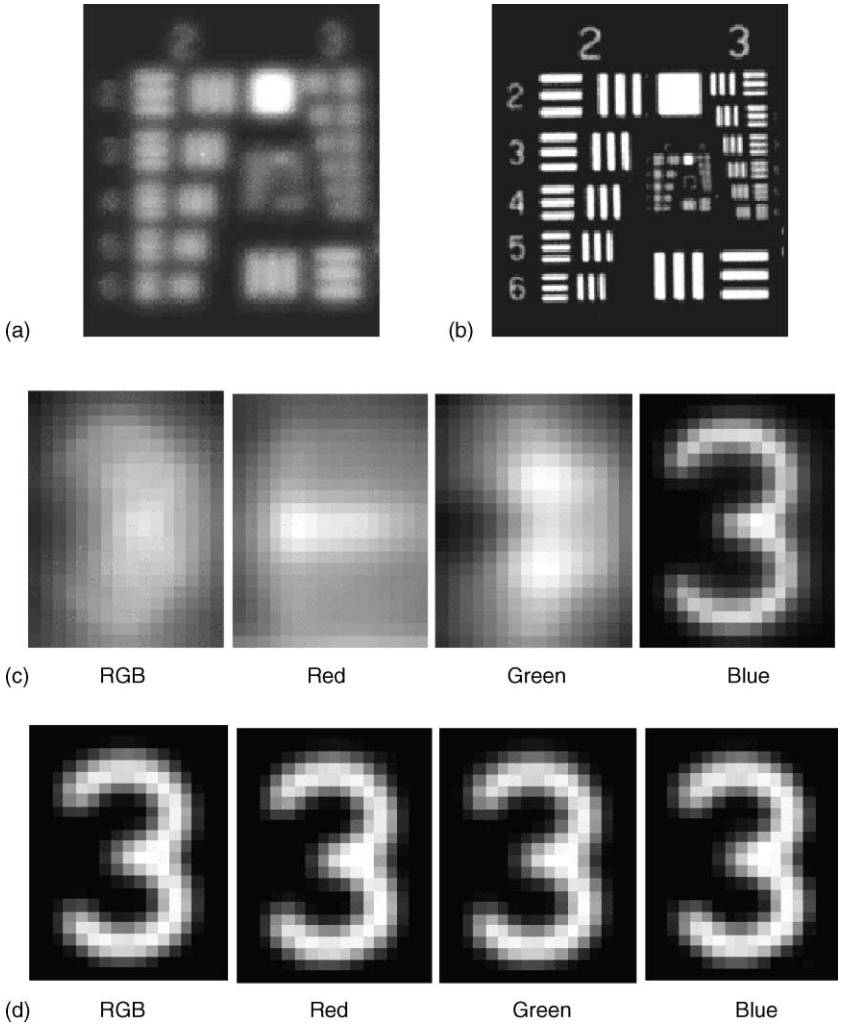


Figure 9 Color images of a US Air Force test pattern. (a) Images from a two-element lens of the center portion of the test pattern. (b) Images from an extended-depth-of-focus imaging system of the center portion of the test pattern. (c) Images from a two-element lens of a small number 'three' on the test pattern. (d) Images from an extended-depth-of-focus imaging system of a small number 'three' on the test pattern. Some residual effects remain due to different magnifications at different wavelengths.

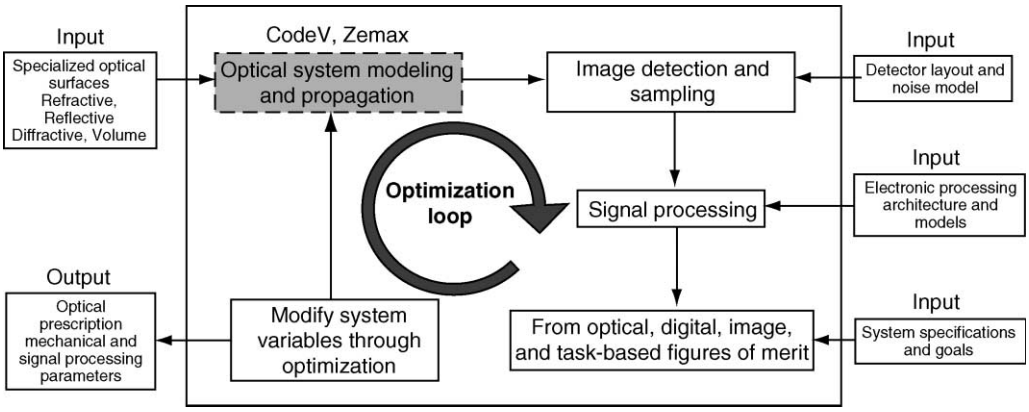


Figure 10 Schematic of a tool for designing imaging systems with Wavefront Coding. A ray trace determines the distribution in the exit aperture, and software determines the effect of the detector array and the signal processing. Then the ray trace and signal processing can be modified to optimize the entire system.

New design tools for imaging systems with wavefront coding

The Woodward function is a very good tool for analyzing imaging systems with a coded wavefront and is a good guide for someone who is expert in designing such systems. However, another approach is better for those designers who are more familiar with conventional lens design tools. If one combines a conventional ray-tracing tool with custom software, it is possible to jointly design the optical and the signal processing portions of the system. A schematic of such a system is shown in Figure 10. The ray-tracing tool is used to trace the rays and determine the distribution in the exit aperture. Various software tools then are used to take into account the effect of pixel size and spacing, and the signal processing. The criteria for optimizing the design can include the size of the digital filter, acceptable noise gain, and other parameters related to signal processing as well as the parameters of the optical elements.

Example applications of hybrid imaging systems with an extended depth of field or focus

Barcode and label readers. The images of Figure 7 show that image-based bar code reading is a good application for imaging systems that are based on Wavefront Coding. A bar code or label reader need not be positioned nearly as accurately as without the

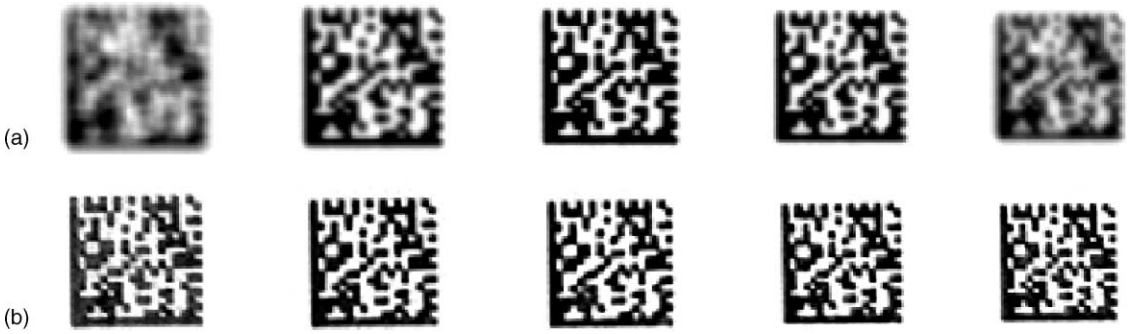


Figure 11 Images of a 2D bar code as the object position is varied for (a) conventional images and (b) Wavefront Coded images after signal processing for the same object positions as in (a).

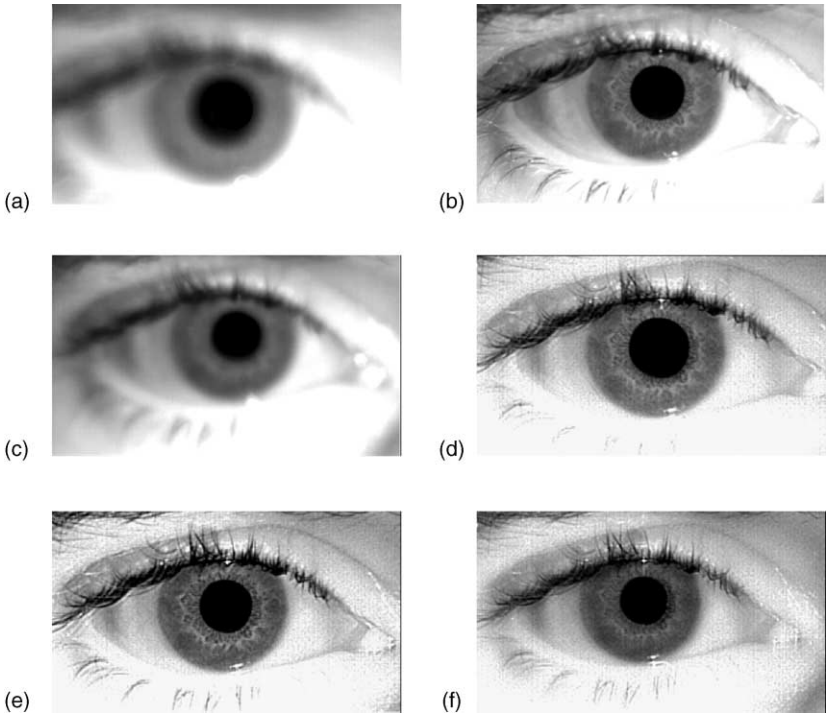


Figure 12 Images of an iris taken with infrared illumination. The top row shows images with a traditional CMOS camera. The lower row is taken with a camera using wavefront coding to extend the depth of field. The misfocus distance in the right and left columns is the same.

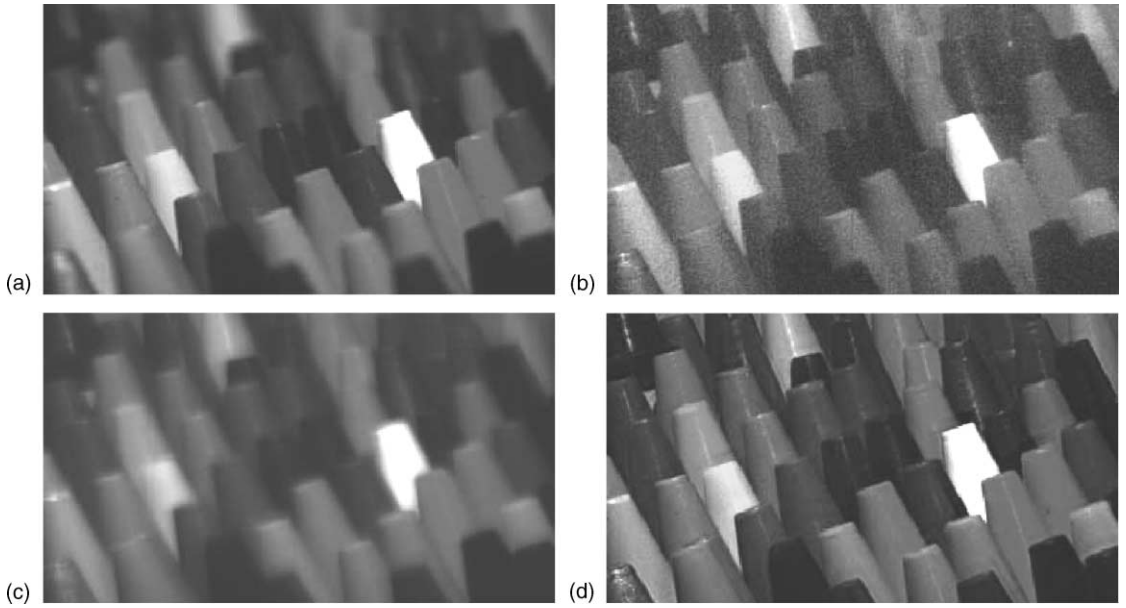


Figure 13 Images from a digital camera (a) Conventional camera with the lens wide open and focused on the white crayon. (b) Conventional camera with the lens stopped down to increase the depth of field, with the resulting increase in noise due to a lower light level. (c) Intermediate image of a camera that was modified with Wavefront Coding to have an extended depth of field. (d) Final image with the modified camera.

extended depth of field. Images of 2D bar codes imaged with and without Wavefront Coding are shown in Figure 11. Images in Figure 11a show the image of a 2D bar code formed by a conventional imaging system as the position of the object is varied. As with most imaging systems with fast optics and moderate magnification, the depth of field of the system is smaller than the application would prefer. Images in Figure 11b show the Wavefront Coded images of a 2D bar code after signal processing for the same object positions as (Figure 11a). These images are sharp and clear over a very large change in the distance to the bar code.

Biometric imaging systems. Biometric imaging applications, such as fingerprint or iris recognition, provide the means to eliminate passwords, secure E-commerce, and securely identify humans or animals. In traditional imaging technology, quite often the trade-off between light throughput and depth of field makes conventional biometric systems difficult to use or not sufficiently reliable. By using Wavefront Coding, the depth of information gathered can greatly increasing the ease of use and possibly reducing the cost of the systems.

Figures 12a,b, and c shows example images of an iris taken with near IR illumination and a traditional CMOS imaging system at three different planes of focus. At best focus (Figure 12b) the detail of the iris is sharp and clear. This iris detail is used by iris recognition algorithms to accept or reject the person.

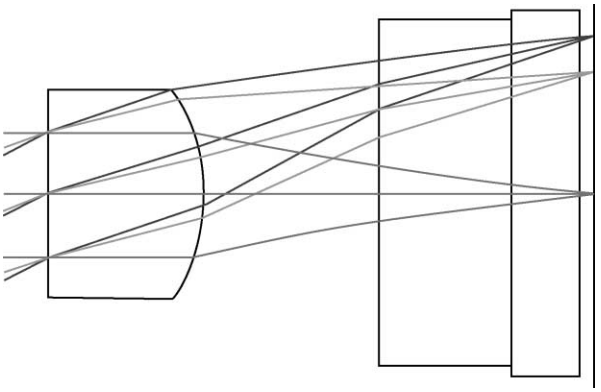


Figure 14 A schematic of a fast single-lens imaging system that has good performance over a wide field of view and the full visible range.

With slight movement towards (Figure 12a) or away (Figure 12c) from the imaging system the detail of the iris is lost. Figures 12d,e, and f show example iris images taken with the same near-IR illumination and a CMOS Wavefront Coded imaging system. At best focus (Figure 12e), the iris detail is as high as that from the traditional system. Movement towards (Figure 12d) or away (Figure 12f) from the imaging system results in images with essentially a constant level of iris detail. The size of the iris changes through magnification change. Increasing the range at which clear images can be made extends the useful image capture volume, enabling a more flexible system that is easier to use.

Color images with digital cameras. A conventional digital camera has a small depth of field as shown in Figure 13a, where the camera was focused on the white crayon. In a conventional imaging system, the only way to increase the depth of field is to stop down the lens. As seen in Figure 13b, this reduces the light that is available, and results in a very noisy image. The intermediate image of an imaging system that has an extended depth of field is shown in Figure 13c, where it is seen that the result of convolving the object distribution with the modified PSF is an intermediate image that appears to have the same blur over the entire image. After signal processing of the intermediate image to obtain the final image, a clear image is seen with good signal-to-noise ratio.

Inexpensive imaging systems with high-quality imaging characteristics. By exploiting the extension of the depth of focus that is made possible with Wavefront Coding techniques, one can design imaging systems that have fewer optical elements than traditional designs that do not use Wavefront Coding. In high-volume applications, reduction in the number of optical elements can lead to greatly reduced costs. Figure 14 shows a schematic of a fast, single-lens Wavefront Coded imaging system that is very compact and has the performance of a longer lens that traditionally requires two or more optical elements. In this example, the entire focus budget was used to compensate for focus-related aberrations, and the shape of the lens was used to compensate for distortion and coma. Even zoom systems can be greatly simplified with the new design techniques. Figure 15 shows a fast two-element Wavefront Coded zoom lens that images over a wide field with performance that is traditionally obtained only with three or more optical elements. Both of these imaging systems use non-rotationally symmetric forms of phase functions and are different from the ones of eqns [5] and [6].

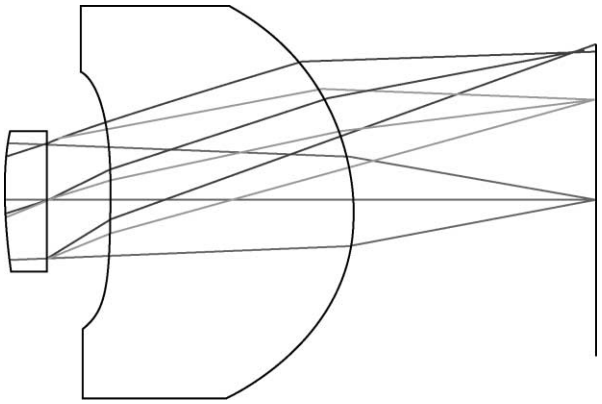


Figure 15 A schematic of a fast two-element zoom system with good performance over a 2.5 × zoom and a wide viewing angle.

Large depth of focus microscopes. A microscope can be modified to employ Wavefront Coding by either modifying the surface of one of the lenses in the objective, or by inserting a phase plate into the system at some appropriate position. One easy means of doing the modification is to place a Wavefront Coding phase plate in the slider position of a microscope, as shown in Figure 16. In this case, the optics is of high quality, and the entire focus budget was allocated to increasing the depth of field. A modified Zeiss microscope was used to obtain the images shown in Figure 17. An image of a diatom in Figure 17 shows the comparison with a conventional microscope in (a) and as with a Wavefront Coded microscope in (b). Notice that the greater depth of field gives more of an appearance of depth in the image.

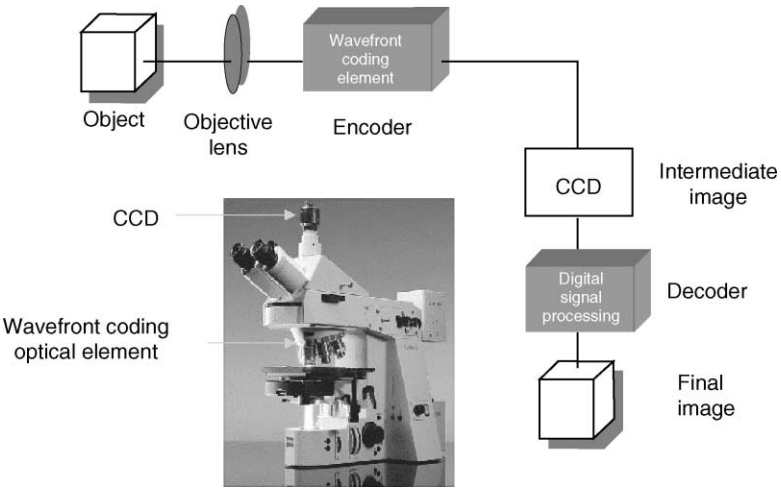


Figure 16 Schematic of a modified Zeiss microscope to employ Wavefront Coding to extend the depth of field.

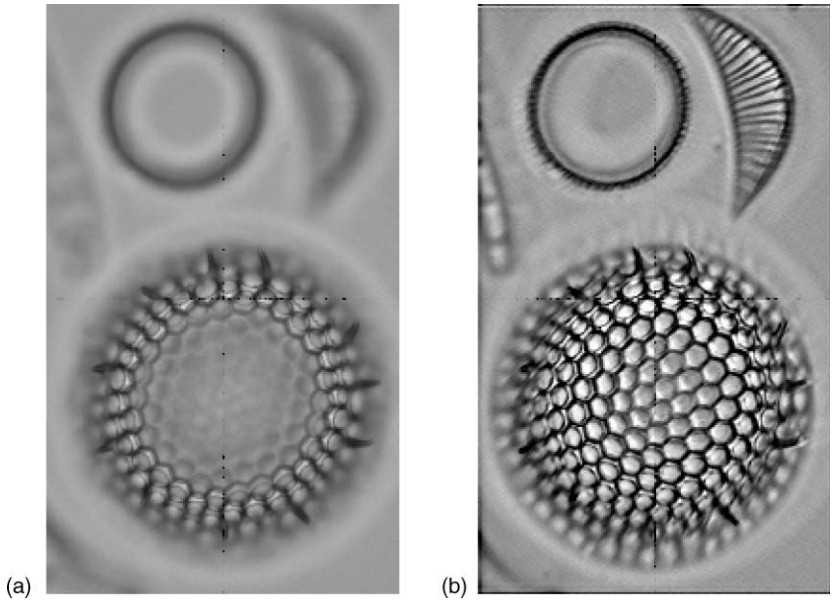


Figure 17 A 100 × image of diatoms made with (a) conventional and (b) modified Zeiss microscope to give an extended depth of field.

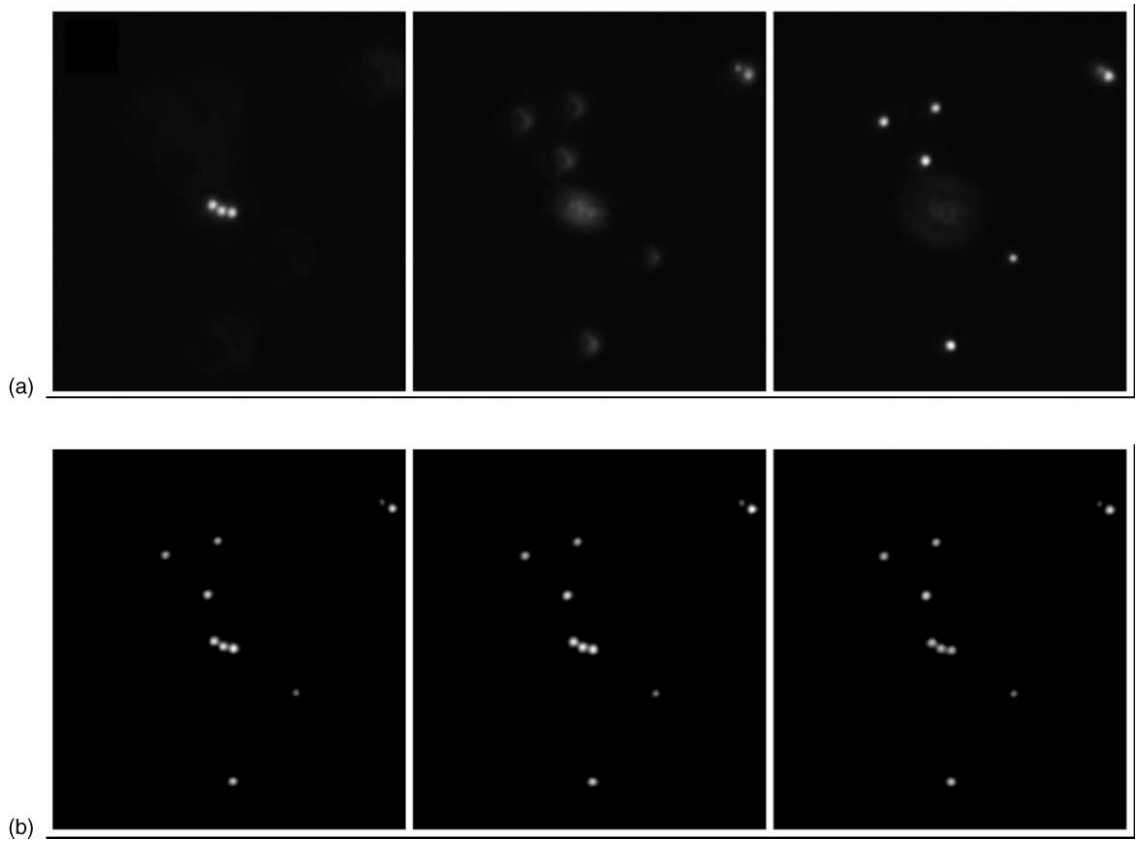


Figure 18 Imaging 0.5 micron diameter fluorescent beads at different sample distances. The images were taken with a 63×, NA = 1.4 imaging system in one micron steps. (a) Three image planes taken using a traditional system. They show that individual beads are sharp only in a single plane. The images in (b), which were taken with a Wavefront Coding system at the same image planes, show all beads are imaged clearly at each plane.

Other modes of microscopy are also possible. In fluorescence microscopy, ultraviolet light illuminates objects that have fluorescing compounds in them. The images are formed using the fluorescence that is excited by the ultraviolet illumination. Figure 18 shows images of fluorescing beads having a 0.5 micron diameter. The beads are at various depths within the sample. The traditional system, operating at 63 × with an immersion objective having a NA of 1.4, has a depth of field that is too small to image all of them. The images in Figure 18a were taken with the same microscope but with a wavefront coding phase element placed in the slider and signal processing of the captured images. All beads are clearly seen at all three of the focus positions.

Potential Future Applications

Wavefront Coding should be useful in many imaging system where signal processing is to be done anyway; in that case, optimizing the optics and the signal processing to the purpose of the imaging system. Examples where an extended depth of field may be desired are endoscopes, video cameras, and lithography. In designing an imaging system to take advantage of a large depth of focus to produce inexpensive digital cameras, high-quality images can be formed with fewer optical elements and much lighter cameras. In lithography, a larger process window is possible.

Because an increase in the depth of focus can permit wide spectral imaging, wide band infrared images can be formed. In systems that are used to provide images for machine interpretation, it is possible to design the optics and signal processing so that the desired process can be optimized globally, with the optics and the signal processing. For example, if the object is to classify cells as cancerous or noncancerous, it may be that the optics should be modified so that the information of primary importance is acquired and transmitted to the detector array. This modification would be done in conjunction with the optimization of the signal processing.

Conclusion

The mathematics and concepts that underlie a new approach to lens design were summarized. The technique has thus far been useful in enhancing the performance of some imaging systems such as microscopes, and in making simpler systems for inexpensive cameras. This new approach requires that some signal processing be done, but in digital cameras, some signal processing is being done already.

List of Units and Nomenclature

Modulation transfer function	The magnitude of the optical transfer function of an imaging system.
Optical transfer function	The response of an imaging system as a function of the spatial frequency of a sinusoidal object distribution.
Spatial frequency	A measure of spatial variation in the form of sinusoids in cycles per millimeter.

See also

Imaging: Wavefront Sensors and Control (Imaging Through Turbulence). **Microscopy:** Imaging Multiple Photon Fluorescence Microscopy; Overview.

Further Reading

Barrett HH and Swindell W (1981) *Radiological Imaging: The Theory of Image Formation, Detection, and Processing*. New York: Academic Press.

Brenner K, Lohmann A and Ojeda-Castañeda J (1983) The ambiguity function as a polar display of the OTF. *Optics Communications* 44: 323–326.

Cathey WT and Dowski ER (2002) New paradigm for imaging systems. *Applied Optics* 41: 6080–6092.

Cathey WT, Frieden BR, Rhodes WT and Rushforth CK (1984) Image gathering and processing for enhanced resolution. *Journal of the Optical Society of America A* 1: 241–249.

Cushman SC (2001) *Finite-Dimensional Matrix Treatment of Partially Coherent Optical Systems*. University of Colorado thesis, Boulder, Colorado.

Dowski ER and Cathey WT (1995) Extended depth of field through wavefront coding. *Applied Optics* 34: 1859–1866.

FitzGerrell AR (1993) *Applications of the Ambiguity Function to Optics*. Thesis, University of Colorado, Boulder, Colorado.

Häusler G (1972) A method to increase the depth of focus by two step image processing. *Optics Communications* 6: 38–42.

Ojeda-Castañeda J, Tepichin E and Diaz A (1989) Arbitrary high focal depth with quasioptimum real and positive transmittance apodizer. *Applied Optics* 28: 2666–2670.

Papoulis A (1974) Ambiguity function in Fourier optics. *Journal of the Optical Society of America* 64: 779–788.

Rihaczek AW (1969) *Principles of High Resolution Radar*. New York: McGraw-Hill.

Van der Gracht J, Dowski ER, Taylor MG and Deaver DM (1996) Broadband behavior of an optical-digital focus-invariant system. *Optics Letters* 21: 919–921.

Wach HB, Dowski ER and Cathey WT (1998) Control of chromatic focal shift through wave-front coding. *Applied Optics* 37: 5359–5367.

Woodward PM (1953) *Probability and Information Theory with Applications to Radar*. New York: Pergamon Press.

PHOTON PICTURE OF LIGHT

S J Bentley, Adelphi University, Garden City, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The photon nature of light is the central topic in the field of quantum optics, which has developed into a leading area of research. The possible technological impacts of this field to many areas, such as imaging, computing, and lithography, have greatly expanded interest into the nature of the photon. A photon description of light goes well beyond the particle nature of light, to include any optical phenomena that cannot be adequately described by classical physical optics. This article will introduce the nature of the photon through a discussion of many such phenomena, covering topics from the beginning of quantum mechanics, including black-body radiation and Compton scattering, to areas of modern interest, such as squeezed light and entangled photons.

The Origins of Quantum Optics

One of the earliest theories in the development of quantum physics was the description of black-body radiation by Max Planck (1858–1947). Using the model of a radiating body as a collection of oscillators, he postulated that the amount of energy emitted or absorbed was proportional to its frequency, ν, and that these oscillators could only have energies corresponding to integer multiples of such a packet of energy, where one unit of energy was given by

$$E = h\nu \qquad [1]$$

where $h = 6.626 \times 10^{-34}$ Js is the universal constant, now known as Planck's constant. He then assumed a Maxwell–Boltzmann distribution to describe the probability of such oscillator states being occupied:

$$P(n) = P_0 e^{-nh\nu/kT} \qquad [2]$$

where $k = 1.381 \times 10^{-23}$ JK^{-1} is Boltzmann's constant and T is the temperature of the blackbody. The energy density in the radiation field, at a given frequency, is given by the oscillator density (at that frequency) times the average energy of the oscillators. Previous models had not only failed to quantize the oscillator energies, but in doing so also assumed equal occupation probabilities that led to a prediction of infinite energy densities at short wavelengths, the so-called ultraviolet catastrophe. The oscillator density is given by

$$n(\nu) = 8\pi\nu^2/c^3 \qquad [3]$$

where $c = 3 \times 10^8$ m s^{-1} is the speed of light. This value can be found by a mode density calculation within a box, as is often done in an introductory electromagnetics text. The average energy of the oscillators is found by calculating the sum of the occupation probability times the state energy and dividing by the sum of the occupation probabilities:

$$E_{\text{av}} = \sum_{n=0}^{\infty} [P(n)h\nu m] \Big/ \sum_{n=0}^{\infty} P(n) \qquad [4]$$

Evaluating the above expression for E_{av} and multiplying by $n(\nu)$, one finds the Planck radiation law:

$$I(\nu) = \frac{8h\pi\nu^3}{c^3(e^{h\nu/kT} - 1)} \qquad [5]$$

The black-body curve described by Planck's law is shown in Figure 1. By introducing quantization to solve a problem that classical physics had failed to explain, Planck paved the way for quantum physics.

Albert Einstein (1879–1955) made use of Planck's quantization for two important results that will be described here. The first is the photoelectric effect, a process by which electrons are released from a metal surface when exposed to light. The process has several traits that could not be explained by

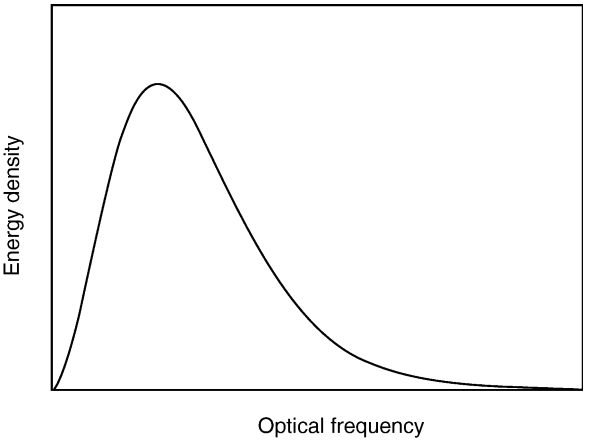


Figure 1 The spectral density emitted by a blackbody at a given temperature as given by Planck's law.

classical physics. A given metal will only release electrons if the incident light is above a certain frequency, regardless of the intensity. However, once sufficient frequency components are present, electrons are immediately released, with their number dependent on the intensity of the light and independent of its frequency. Contrarily, the maximum potential energy of the released electrons depends only on the maximum frequency of light present, independent of intensity. Einstein was able to satisfy all of these conditions by saying that the light energy comes in quantized amounts as described by Planck, and that the electrons in the metal require a minimum energy, or work function W, to be ejected. Then for each quantum of light, which we now call a photon, with energy greater than W that strikes an electron, the electron will be released with kinetic energy given by

$$K = h\nu - W \qquad [6]$$

It is historically important to note that, although Planck introduced the idea of a quantized release of radiation from oscillators, he still believed the light field itself to be continuous and classical. Planck's approach of quantizing energy levels of matter while treating light classically has since become known as the semiclassical method, and is adequate for solving many problems. Thus, it was also in quantizing light itself that Einstein took the next step, and thus can be viewed in many respects as the father of the modern photon, although the term photon was not used until two decades later.

The second area in which Einstein used the idea of energy quantization is in predicting radiative transition probabilities. A radiative transition is a process by which a photon is either absorbed or emitted to satisfy conservation of energy when an atom moves to a higher or lower energy state. Einstein assumed three such processes existed, as shown in Figure 2: absorption of a photon raising the atom to a higher energy level; spontaneous emission of a photon, where the atom emits a photon in going to a lower energy level; and stimulated emission, a process in which a photon interacts with the atom and causes it to emit a second photon (in phase with the first) and thereby go to a lower energy level. For each of these processes he defined a corresponding coefficient, now known as the Einstein A and B coefficients. They are labeled B_{12} for absorption, A_{21} for spontaneous emission, and B_{21} for stimulated emission. Note that there is no A_{12} as absorption by its very nature is a stimulated process. Einstein further stated that the rate of spontaneous emission was only dependent on his material dependent coefficient and the number of

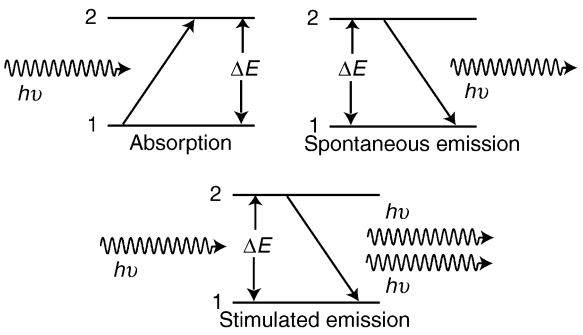


Figure 2 Energy level diagrams representing the processes of absorption, spontaneous emission, and stimulated emission.

atoms in the upper energy state, with the rate given by

$$R_{sp} = A_{21}N_2$$

However, he said that the rates of the stimulated processes were also dependent upon the strength of the optical field seen by the atoms, $I(\nu)$, and were given by

$$R_{ab} = I(\nu)B_{12}N_1 \qquad [7]$$

for absorption and

$$R_{st} = I(\nu)B_{21}N_2 \qquad [8]$$

for stimulated emission. The full absorption rate is given by R_{ab} as defined above, whereas the full emission rate is the sum of the spontaneous and stimulated rates, given by

$$R_{em} = [A_{21} + B_{21}I(\nu)]N_2 \qquad [9]$$

Einstein was primarily interested in a system in equilibrium, for which $R_{ab} = R_{em}$. If one then looks at the limit where the optical field is very large such that spontaneous emission is negligibly small and the population is equally divided between the upper and lower state, the relationship $B_{12} = B_{21}$ is found. Applying the same Maxwell–Boltzmann occupation statistics that Planck used to find his black-body expression, one can find that in thermal equilibrium the ratio of probability of finding an atom in state 1 to finding an atom in state 2 is

$$N_1/N_2 = e^{(E_2-E_1)/kT} = e^{h\nu/kT} \qquad [10]$$

Note that for a two-level system in thermal equilibrium, the population in the upper level can at most equal, never exceed, the population in the lower level, regardless of the strength of the applied field. This is important in showing that the B coefficients are equal, and will also be a crucial point when discussing the theory of the laser later in this article.

Appropriately combining the above expressions leads to

$$I(\nu) = A_{21}/[B_{21}(e^{h\nu/kT} - 1)] \qquad [11]$$

Equating this with Planck's radiation law yields the relation:

$$A_{21}/B_{21} = 8\pi h\nu^3/c^3 \qquad [12]$$

Thus, to fully specify the transition probabilities for a given system, one needs only to find A or B. It is rather straightforward using only atomic quantization to calculate B (such a derivation can be found in many introductory quantum mechanics textbooks). Direct calculation of A is more involved, requiring quantization of both the optical field as well as the atomic system.

Arthur Compton (1892–1962) used the idea of quantized radiation to explain the presence of longer wavelength components present when X-rays are scattered from matter, an effect which now bears his name. The amount of scattering is independent of the scattering material and the wavelength of the source, depending only on the angle of scattering, θ. If the scattering is treated as a collision between a particle-like photon and a massive particle (taken to be initially at rest) as shown in Figure 3, the result can then be found simply by applying conservation of energy and momentum, where the photon energy is as given above and the photon momentum is given by

$$p = E/c = h/\lambda \qquad [13]$$

Recalling that the momentum is a vector quantity, the conservation laws provide a set of three coupled equations. Solving them directly one finds that

$$\Delta\lambda = (h/c)(1 - \cos\theta)(2K)/[p^2 - (K/c)^2] \qquad [14]$$

where K and p correspond to the final kinetic energy and momentum of the massive particle. It is important to allow for the possibility of very high velocities for the massive particle, thus implying that relativistic expressions for K and p are needed. In terms of the kinetic energy, the relativistic expression for momentum is

$$p = \sqrt{2Km + (K/c)^2} \qquad [15]$$

where m is the mass of the scattering particle. Using this relation, the above expression can be simplified, giving the wavelength shift to be

$$\Delta\lambda = (1 - \cos\theta)h/mc \qquad [16]$$

The scaling factor is known as the Compton wavelength:

$$\lambda_c = h/mc \qquad [17]$$

For the case of an electron, the Compton wavelength has the value 0.02426 angstroms. For a massive particle, the Compton wavelength is often used when wanting to discuss and analyze wavelike properties. As a final note on this process, if the calculation would have been done nonrelativistically, there would have been an extra factor of $[1 - 0.25(\nu/c)^2]^{-1}$, where ν is the velocity of the particle. The appearance of the factor (ν/c) in any result implies the need for a relativistic treatment, though often (as in this case) taking the limit of $\nu/c = 0$ yields the proper result.

Another area in which the photon nature of light is revealed is a form of inelastic scattering of light now known as the spontaneous Raman effect. Discovered by CV Raman (1888–1970), this is a process in which a material illuminated by a light source scatters light into frequency components not found in the source. The frequency shift is due to the energy difference, ΔE, between two energy states in the material. There are two possible cases, each of which is represented in Figure 4. In the first, known as Stokes scattering, an atom in the lower state absorbs a source photon, raising it to the upper state via a virtual level, during which it emits a photon at the Stokes frequency, ν_s, as represented by the relation

$$h\nu_i = h\nu_s + \Delta E \qquad [18]$$

where ν_i is the frequency of the incident photon from the source. The second case, anti-Stokes scattering, has an atom in the upper state absorb the source photon and simultaneously fall to the lower state, again through the virtual level during which it emits a photon at the higher anti-Stokes frequency, ν_a,

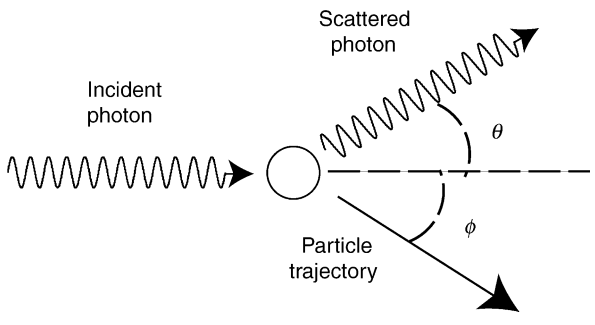


Figure 3 Schematic representation of Compton scattering.

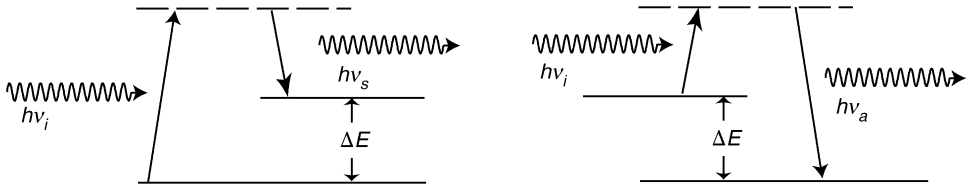


Figure 4 Energy level diagrams for Stokes and anti-Stokes Raman scattering.

given by

$$h\nu_i = h\nu_a - \Delta E \qquad [19]$$

Since the anti-Stokes scattering requires the atom to be initially in the upper state, the probability of such scattering is generally much less than that for Stokes scattering. This can be quantified by the same occupation statistics used above, giving the ratio of probabilities as

$$P_{as}/P_s = e^{-h\Delta E/kT} \qquad [20]$$

It is important to note that the energy difference can either be based on the electronic structure of an atomic system as has been discussed thus far, or other energies such as the vibrational energies in a molecular system. The ability to probe a variety of energy structures without having to tune the source to a particular wavelength has led to wide use of Raman scattering for spectroscopy. A related effect, known as stimulated Raman scattering, will be discussed later.

The Photon in Modern Physics

As an indication of the widespread importance of the results discussed thus far, they led to four Nobel prizes in physics: Planck in 1918; Einstein in 1921; Compton in 1927; and Raman in 1930. Our initial understanding of the photon comes from these results, but it was with the development of modern quantum mechanics by Erwin Schrödinger (1887–1961), Werner Heisenberg (1901–1976), and Paul Dirac (1902–1984); quantum electodynamics by Richard Feynman (1918–1988) and others; and the invention of the laser by Charles Townes (1915–) and others, that the field of quantum optics and our modern view of the photon began to fully take shape. In this section, the basics of modern quantum optics will first be introduced. A brief discussion of the theory of the laser as related to photon transitions will then be given, followed finally by examples of important areas of experimental quantum optics that have been made accessible by the laser.

It is now common to treat the photon with a very similar formalism as that originally created for massive particles in the development of quantum mechanics. That is, one describes the photon by a wave function, $|\Psi\rangle$, which contains all the information about the photon and must satisfy the Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}|\Psi\rangle = H|\Psi\rangle \qquad [21]$$

where $\hbar = h/2\pi = 1.054 \times 10^{-34}$ Js and H is the Hamiltonian for the photon. Just as in classical electromagnetics, the Hamiltonian is related to an electric and magnetic field. Clearly, for this treatment to be in agreement with the well-established classical electromagnetic theory, these fields must satisfy Maxwell's equations. There are several possible representations of the Hamiltonian, but for the purposes of this discussion a convenient choice is

$$H = h\nu(\hat{n} + 0.5) \qquad [22]$$

where $\hat{n}$ is called the number operator as it returns the average number of photons in the state, $\langle\hat{n}\rangle$, represented by the operation

$$\langle\hat{n}\rangle \equiv \langle\Psi|\hat{n}|\Psi\rangle \qquad [23]$$

Explanations of the notation used can be found in any elementary quantum mechanics text. Note that in defining the Hamiltonian, a single frequency ν was used. This implies a single mode field. For a more general case, one would simply sum over the modes corresponding to various frequencies.

In addition to obeying the Schrödinger equation, the photon is also restricted by the Heisenberg uncertainties as for massive matter, given in a general form as

$$\Delta q \Delta p \geq \hbar \qquad [24]$$

where q and p are any set of canonically conjugate variables, such as position and momentum or energy and time. Note that, depending on the exact definitions used for q and p, there can be extra factors such as 0.5, and sometimes normalized

variables are used such that the inequality is with respect to 1. However, it is interesting to note that, unlike for massive particles that were considered well localized in classical theory, the idea of uncertainty for light had been accepted in some form long before the development of quantum mechanics. Neglecting the scaling factor $\hbar$, which is purely quantum in origin, an uncertainty principle can be derived for any wavelike phenomena simply using the relationship between Fourier transform pairs.

Using the basic ideas of a photon state and the uncertainty principle, some important properties of light will now be explored. There are several possible forms the wave function can take. It is often convenient to write the wave function in terms of states containing a certain number of photons. Such states are known as Fock or number states and are represented as $|n\rangle$, with n being the number of photons in the state and are said to be eigenstates of the number operator. The energy of a Fock state is given by

$$E_n = h\nu(n + 0.5) \qquad [25]$$

Notice that this predicts a background, or vacuum, energy for the electromagnetic field, even when the average number of photons in the field is zero. This vacuum energy is given by $0.5\,h\nu$. Correspondingly, due to the uncertainty principle, there must be some uncertainty in this energy as well as with any energy. The fluctuations in the vacuum energy are known as vacuum fluctuations. Many processes, such as spontaneous emission that cannot be adequately described by a simple semiclassical approach, where matter is treated quantum mechanically and light is treated classically, can be treated by the addition of vacuum fluctuations, without requiring a full quantum mechanical treatment of the electromagnetic field. Such an approach is used in many texts, but there are several important physical phenomena that do require the fully quantized approach discussed here.

The state most resembling a classical field is known as the coherent state, represented here as a sum of Fock states:

$$|\Psi\rangle = |\alpha\rangle = e^{-|\alpha|^2/2} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |n\rangle \qquad [26]$$

where α is a parameter describing the field, with the average number of photons in the field given by

$$\langle \hat{n} \rangle = |\alpha|^2 \qquad [27]$$

One key property of coherent states is that they exhibit minimal uncertainty. That is, the uncertainty relations become equalities. Also, the uncertainty is equally divided between quadrature components.

Before applying the above principles to explore some of the dramatic results of quantum optics, a brief discussion of the laser will be given due to its importance in the field. The theory of operation of the laser is based directly on Einstein's transition probabilities. For the transition of interest, one wants to generate a large number of coherent photons (that is, photons in phase with one another). Recall that in the process of stimulated emission, one photon leads to two coherent photons, whereas spontaneous emission leads to photons which are not phase referenced to anything else in the system. Also, if a significant number of the photons are to escape the region where they are generated, it is important to minimize the absorption. Thus, a laser requires that the stimulated emission rate exceed both the spontaneous emission rate and the absorption rate. The first condition places a lower limit on the optical field density present at the lasing frequency. This density can be reasonably large, but presents no fundamental difficulties. In practice, mirrors are placed around the emission region to confine the photons and counteract losses, thereby building up the energy density until the above condition is satisfied. The second condition, however, can be written as

$$N_2 > N_1 \qquad [28]$$

and is the more restrictive requirement. Such a condition is referred to as inversion and, as was previously mentioned, cannot be sustained in a two-level system once equilibrium is reached. Therefore, to be able to maintain lasing, additional levels are needed. Two typical laser energy diagrams are shown in **Figure 5**. In such multilevel systems, it is possible to create an inversion in the levels of interest and thus allow lasing. Further details on lasers (both in theory and practice) are widely available, though it should be noted that the output of a laser is a good approximation of a coherent state described above.

The invention of the laser revolutionized optics, as well as the whole of modern physics, by making readily available electromagnetic fields of a vastly different nature than that from any previous source of light. This quickly led to the birth of a new field known as nonlinear optics. Nonlinear optics describes interactions between light and matter which have properties dependent on the field strength. Although some nonlinear processes can be described adequately without the use of a photon picture, most nonlinear process can be best described as some combination of annihilation (absorption) and creation (emission) of photons. Two such

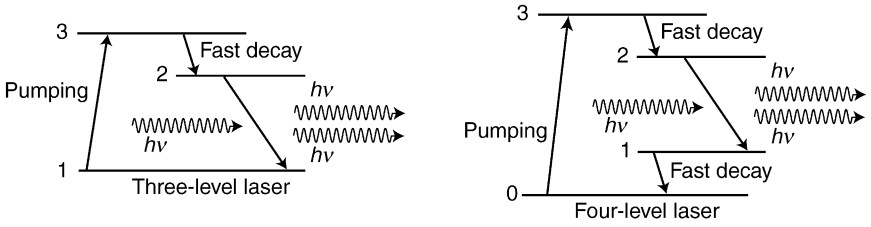


Figure 5 Energy level diagrams for three-level and four-level lasers.

processes, that are of great interest in quantum optics, are stimulated scattering and parametric down conversion.

Earlier, the process of spontaneous Raman scattering was discussed. For such a process, it is assumed that the incident field driving the process is relatively weak such that the scattering process generates very low photon densities in the various modes of the electromagnetic field. However, just as there are both stimulated and spontaneous emission processes in the radiative transitions discussed previously, in the presence of strong fields there can also be stimulated scattering. The total Raman scattering rate can be expressed as

$$R = \eta I_i(I_R + 1) \qquad [29]$$

where η is a scaling factor, including the gain of the process for a given material, I_i is the strength of the incident optical field (generally a laser field to be sufficiently strong to cause stimulated scattering), and I_R is the Raman field strength. Note that for I_R, the rate is just linear with the applied field, representing the spontaneous process. However, with a strong pump, the generated field will quickly become large enough such that the stimulated process will dominate. The spontaneous scattering is not directional dependent, but since the stimulated process grows along the length of the scattering material in the path of the incident beam, it will tend to be generated along the path of the incident laser. Due to linear and nonlinear dispersion effects, however, the light will form in a narrow cone about the pumping laser rather than exactly collinear. Unlike the spontaneous scattering, a significant percentage of the incident photons can be converted to the Raman frequencies, making stimulated Raman scattering an extremely powerful spectroscopic tool.

Possibly the most important nonlinear optical process for quantum optics is parametric down conversion. This process is interesting both in the nature of the process itself, as well in the properties of the generated light. In parametric down conversion, shown schematically in Figure 6, one photon (known as a pump photon) of frequency ν_p is annihilated

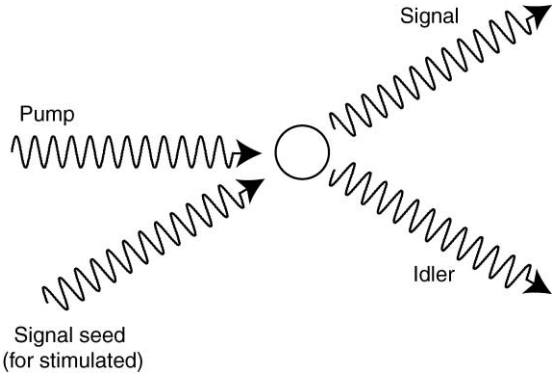


Figure 6 Schematic representation of parametric down conversion.

and two new photons (generally referred to as signal and idler) are created with frequencies ν_s and ν_i respectively, given by the conservation of energy relationship:

$$h\nu_p = h\nu_s + h\nu_i \qquad [30]$$

In the case where $\nu_s = \nu_i$, the process is referred to as degenerate. Down conversion can either be spontaneous, in which only pump photons are externally applied, or stimulated, in which both pump photons and signal photons are applied. In the stimulated case, the signal is amplified and idler photons are generated as a byproduct. Parametric down conversion is one of the leading sources of nonclassical states of light. The term nonclassical is used rather than quantum, because all states of light can be accurately described by a quantum picture, but while most can also be described very well with a classical picture the states referred to as nonclassical cannot. Two of the most common nonclassical states of light are entangled states and squeezed states.

Entangled states are multiphoton states in which the state cannot be written as a product of the states of the individual photons. For simplicity, only two-photon entangled states (commonly referred to as EPR states for Einstein, Podolsky, and Rosen) will be discussed. More extensive and general material, including discussions of higher-order entanglement,

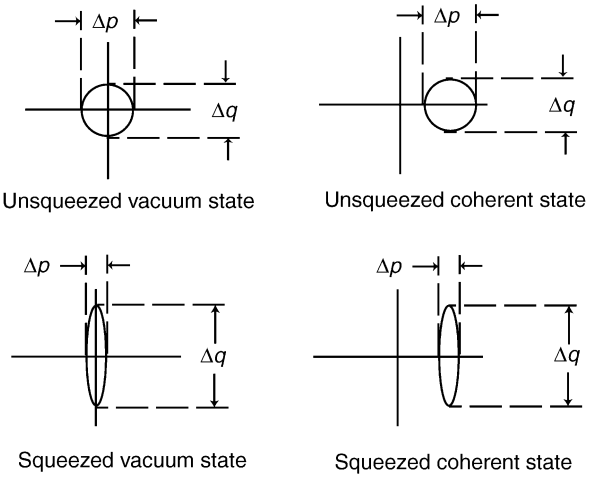


Figure 7 Uncertainty diagrams for squeezed and unsqueezed vacuum and coherent states.

such as the three-photon entangled states (commonly referred to as GHZ states for Greenberger, Horne, and Zeilinger), can be found in a variety of textbooks on modern quantum optics. One of the simplest ways to generate a two-photon entangled state is by the process of spontaneous parametric down conversion described above. If the product of the pumping intensity, generation length, and generation gain is of an appropriate level, individual pairs of photons are emitted. The states of these photon pairs are entangled, with the position, momentum, wavelength, polarization, and all other properties of the photons completely correlated and inseparable. Such states have been a powerful tool for the experimental verification of many of the predictions of quantum mechanics and quantum optics. Entangled states can also act as a source for many of the newer technologies in quantum optics, including cryptography and teleportation.

Another important form of nonclassical light is known as squeezed light. The principle behind squeezed light is that, unlike for coherent states where the uncertainty is equally distributed, the uncertainty in one quadrature can be reduced at the expense of increased uncertainty in the other quadrature. This is shown schematically in Figure 7. As opposed to entangled states which tend to be very weak, containing only a few photons, squeezed states of light can be quite intense. This makes squeezed light more desirable for many applications.

In conclusion, it should be noted that while the photon description of light can explain a wide variety of physical processes and is now leading to many exciting technological developments, much is still unknown about the photon at the fundamental level. Any accurate description of the photon must be in agreement with experimental evidence, including that discussed in this article, but one should be aware that the accepted picture of the photon nature of light is still constantly evolving and growing.

See also

Nonclassical Light. **Nonlinear Optics, Applications:** Raman Lasers. **Optical Amplifiers:** Raman, Brillouin and Parametric Amplifiers. **Quantum Optics:** Entanglement and Quantum Information. **Scattering:** Raman Scattering. **Spectroscopy:** Raman Spectroscopy.

Further Reading

Boyd RW (2002) *Nonlinear Optics*, 2nd edn. San Diego, CA: Academic Press.

Cohen-Tannoudji C, Diu B and Laloë F (1977) *Quantum Mechanics*. New York: Wiley-Interscience Publications.

Hecht E (2001) *Optics*, 4th edn. Reading, MA: Addison-Wesley Publishing.

Heisenberg W (1949) *The Physical Principles of the Quantum Theory*. New York: Dover Publications.

Lipson SG, Lipson H and Tannhauser DS (1995) *Optical Physics*, 3rd edn. Cambridge, UK: Cambridge University Press.

Loudon R (2000) *The Quantum Theory of Light*, 3rd edn. Oxford, UK: Oxford University Press.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge, UK: Cambridge University Press.

Milonni PW and Eberly JH (1988) *Lasers*. New York: Wiley-Interscience Publications.

Roychoudhuri C and Roy R (eds) (2003) *OPN Trends – The Nature of Light: What is a Photon?* Washington, DC: Optical Society of America.

Schwinger J (ed.) (1958) *Selected Papers on Quantum Electrodynamics*. New York: Dover Publications.

Scully MO and Zubairy MS (1997) *Quantum Optics*. Cambridge, UK: Cambridge University Press.

Yariv A (1989) *Quantum Electronics*, 3rd edn. New York: John Wiley & Sons.

PHOTONIC CRYSTALS

Contents

Atomic Physics
Electromagnetic Theory
Microwave Photonic Crystals
Nonlinear Optics in Photonic Crystal Fibers
Photonic Crystal Lasers, Cavities and Waveguides
Self-Assembled and Functionalized Photonic Crystals

Atomic Physics

G Kurizki and A G Kofman, Weizmann Institute of Science, Rehovot, Israel

D Petrosyan, Institute of Electronic Structure and Laser, Heraklion, Greece

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The fabrication and study of dielectric structures whose refractive index is periodically modulated on a micron or submicron scale, known as photonic crystals (PCs), are attracting considerable interest at present. One-, two-, or three-dimensional (1D-, 2D-, or 3D-) periodic PCs exhibit photonic band gaps (PBGs), analogously to electronic band gaps in ordinary crystals, and can incorporate defects, designed to form localized narrow-linewidth (high-Q) modes at PBG frequencies. The advent of such structures opens up new perspectives in atomic physics and quantum optics, since they are expected to allow an unprecedented control over the spectral density of modes (DOM) and the spatial modulation of narrow-linewidth (high-Q) modes, in the microwave, infrared, and optical domains. An interesting situation arises when foreign atoms or ions – dopants – with transition frequencies within the PBG, are implanted in the PC. Then light near one of these frequencies resonantly interacts with the dopants and is concurrently affected by the PBG dispersion. Consequently, highly nonlinear processes with a rich variety of unusual PC-related features are anticipated. Such nonlinear optical processes, involving near-resonant transitions in PCs, undergo basic modifications as compared to the corresponding processes in free space, which are attributed to the strong suppression of the DOM within PBGs, to sharp bandedges and to intra-gap narrow lines associated with high-Q defect modes.

Results are detailed below for spontaneous emission of radiation in PCs and for the resonant interaction of atoms with the field of a single high-Q defect mode. These results stem from the failure of perturbation theory and the onset of strong field-atom coupling near sharp bandedges or narrow defect-mode lines in 2D and 3D PCs. 3D-PBGs are needed, in order to extinguish spontaneous emission in all possible directions of propagation and dipole orientations. If only one polarized atomic transition is involved in the spontaneous emission, 2D-PCs suffice for its suppression. For controlling strictly unidirectional field propagation, it is sufficient to resort to PBGs in 1D-periodic structures (Bragg reflectors or dielectric multi-layer mirrors).

As a further example of field–atom interactions in doped PCs, we will show that two ultraweak electromagnetic fields (or photons), coupled to appropriate transitions of the dopants in a PC, can mutually induce large phase shifts or drastic changes in absorption. Appreciable photon–photon correlations can then be established, which can be utilized in both classical and quantum optical communications.

Radiative Decay and Photon–Atom Binding in a PC

The interaction of a two-level atom with an arbitrary field-mode continuum can be described by the following second-quantized Hamiltonian in the

rotating-wave approximation (RWA)

$$H = \hbar\omega_a|e\rangle\langle e| + \hbar\sum_{\vec{k}}\omega_{\vec{k}}a^{\dagger}_{\vec{k}}a_{\vec{k}} + \hbar\sum_{\vec{k}}[\kappa(\omega_{\vec{k}})a^{\dagger}_{\vec{k}}|g\rangle\langle e| + \text{hermitian constant}] \quad [1]$$

Here $|e\rangle$ and $|g\rangle$ are the excited and ground atomic states, respectively, ω_a is the atomic transition frequency, $a^{\dagger}_{\vec{k}}$ and $a_{\vec{k}}$ are the creation and annihilation operators of a field mode with wavevector $\vec{k}$, and frequency $\omega_{\vec{k}}$, and $\hbar\kappa(\omega_{\vec{k}})$ is the resonant coupling energy of this mode with the atomic dipole. The $\vec{k}$-summation can be replaced by an integral over the continuum DOM in the frequency domain, $\sum_{\vec{k}} \rightarrow \int_0^{\infty} d\omega\rho(\omega)$, avoiding, for the sake of simplicity, the study of direction-(angle-)dependent DOM effects.

We consider a two-level atom that is excited at time $t = 0$ in an empty, periodic dielectric structure. The wavefunction of the combined system field + atom can be cast in the general form:

$$|\Psi(t)\rangle = \alpha(t)|e, \{0_\omega\}\rangle + \int \beta_\omega(t)|g, 1_\omega\rangle\rho(\omega)d\omega \quad [2]$$

where $\{0_\omega\}$ signifies the completeness of field modes in the vacuum state and $|1_\omega\rangle$ denotes single-photon occupation of the ω-mode. The corresponding Schrödinger equation can be solved with the initial condition $|\Psi(0)\rangle = |e, \{0_\omega\}\rangle$ by means of the continuum spectral response:

$$G(\omega) = |\kappa(\omega)|^2\rho(\omega) \quad [3]$$

The analysis of eqn [2], with $G(\omega)$ appropriate for photonic bandstructures, is aimed at revealing the prominent features of the atomic excitation decay $\alpha(t)$ and the corresponding emission spectrum.

In what follows, we consider a photonic bandstructure with several PBGs, separated by allowed bands. Each PBG is labeled by index i and has lower and upper cutoff frequencies ω_{Li} and ω_{Ui}, respectively. Then $G(\omega) = 0$ for $\omega_{Li} < \omega < \omega_{Ui}$. Naively, one might expect that an excited atom would either be stable against single-photon decay, if ω_a is within a PBG, or decay completely at $t \rightarrow \infty$, if ω_a is anywhere in an allowed band. However, both statements turn out to be inaccurate.

Incomplete decay of $\alpha(t \rightarrow \infty)$ occurs if there is a stable eigenvalue (energy level) $\hbar\omega_i$ of the total (field-atom) Hamiltonian eqn [1]. Such an energy level is possible only if $G(\omega_i) = 0$, i.e., for ω_i in a PBG. It must satisfy:

$$\omega_i = \omega_a + \Delta(\omega_i) \quad [4]$$

$$\Delta(\omega_i) = \int_0^{\omega_{Li}} \frac{G(\omega')}{\omega_i - \omega'}d\omega' + \int_{\omega_{Li}}^{\infty} \frac{G(\omega')}{\omega_i - \omega'}d\omega' \quad [5]$$

Here the integrals are the frequency shifts of the atomic resonance $\hbar\omega_a$ induced by the spectral parts of the reservoir situated, respectively, below and above the ith PBG (Figure 1). If eqn [4] holds, then the corresponding term in $\alpha(t)$ is proportional to $\exp(-i\omega_i t)$, and does not decay. Physically, such a stable level can be interpreted as representing the binding of the photon to the atom (photon-dressed atom), without the ability to leave the atomic vicinity, due to Bragg reflection in the PBG.

Assuming that ω_i occurs in the ith PBG, we observe that the first shift is positive whereas the second shift is negative, i.e., each part of the continuum repels the atomic level from its PBG edge. Equation [4] has a solution if ω_a falls between the minimum and maximum values assumed by its left-hand side in the ith PBG, i.e., if

$$\omega_{Li} - \Delta(\omega_{Li}) < \omega_a < \omega_{Ui} - \Delta(\omega_{Ui}) \quad [6]$$

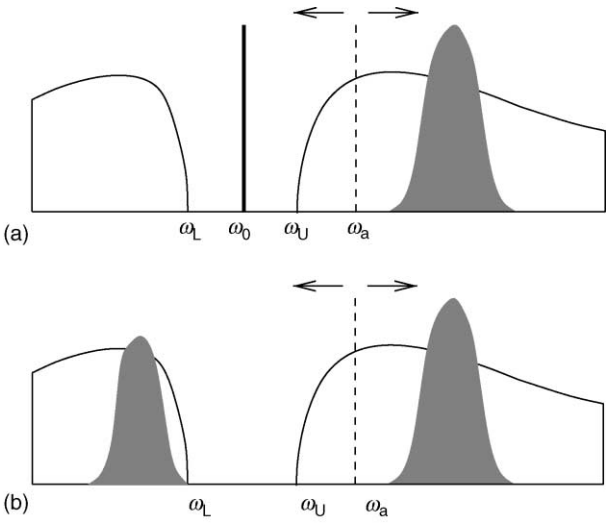


Figure 1 Frequency shifts of the atomic resonance ω_a (dashed line) due to 'repulsion' (arrows) by asymmetric spectral parts of the continuum below and above the PBG: (a) Incomplete decay for ω_a in allowed band. The interaction with the continuum splits the state into a superposition of a stable part in the PBG (solid line) and a decaying part above ω_a (shaded peak). (b) Complete decay for same ω_a, due to larger shifts that split the state into decaying parts in two allowed bands.

Incomplete decay occurs if this condition holds at least for one PBG. There may be, at most, one stable state in a single PBG (Figure 1). However, a two-level atom in a periodic structure can have several stable dressed states simultaneously, when conditions hold for several PBGs. It is quite extraordinary that the conditions (eqn [6]) for incomplete decay can be fulfilled with ω_a, not only in a PBG, but even in an allowed band, e.g., when $\Delta(\omega_{Ui})$ is negative or when $\Delta(\omega_{Li})$ is positive (Figure 1a,b).

Equally counterintuitive is the converse possibility, of complete decay for ω_a within a PBG. When there is a single PBG and one of the inequalities in eqn [6] is violated, complete decay will occur for $\omega_{Li} < \omega_a < \omega_{Li} - \Delta(\omega_{Li})$, if $\Delta(\omega_{Li}) < 0$, or $\omega_{Ui} - \Delta(\omega_{Ui}) < \omega_a < \omega_{Ui}$, if $\Delta(\omega_{Ui}) > 0$.

The possibility that one or several discrete, stable, states exist yields the corresponding wavefunction in the form:

$$|\Psi(t)\rangle = \sum_i \sqrt{c_i}|\psi_i\rangle e^{-i\omega_i t} + |\Psi_c(t)\rangle \qquad [7]$$

Here the summation is over all discrete atomic photon-dressed states with energies $\hbar\omega_i$, $|\psi_i\rangle$ is a discrete-(dressed-)state eigenfunction of the Hamiltonian (eqn [1]) normalized to unity, and weighted by the amplitude $\sqrt{c_i} = [1 + \int_0^\infty d\omega G(\omega)/(\omega - \omega_i)^2]^{-1/2}$. The dressed state $|\psi_i\rangle$ consists of an excited-state component and a photon-bound ground-state component.

The population of the excited state for long times:

$$|\alpha(t)|^2 = \sum_{i,i'} c_i c_{i'} \cos(\omega_i - \omega_{i'})t \quad \text{for} \quad t \to \infty \qquad [8]$$

is a nonzero time-constant if there is one discrete stable state, whereas in the case of several such states, it undergoes beats at the frequencies corresponding to their energy differences. The splitting of an excited state $|e\rangle$ into superposed stable states, oscillating at bandgap frequencies ω_i, whose amplitudes c_i, and eigenfrequencies ω_i are controllable by the atomic transition detuning from cutoff, constitutes spontaneous coherence control.

If, however, ω_a is far from the cutoff, then eqn [2] results in an exponentially decaying amplitude:

$$\alpha(t) \approx e^{-(\gamma_a + i\tilde{\omega}_a)t} \qquad [9]$$

where $\tilde{\omega}_a = \omega_a + \Delta_a$, Δ_a, and γ_a being the effective spectral shift and width of the decaying atom. This regime holds for a locally smooth $G(\omega \approx \omega_a)$ such that

$$|\gamma_a'|, |\Delta_a'| \ll 1; \; |\gamma_a'\Delta_a| \ll \gamma_a; \; \gamma_a, |\Delta_a| \ll |\omega_a - \omega_g| \qquad [10]$$

where ω_g is the bandedge frequency nearest to ω_a and the primes denote differentiation with respect to ω_a.

We now apply the foregoing general results to a model DOM distribution. This distribution is derived on keeping the lowest term in the Taylor expansion of the dispersion relation $\omega(\vec{k})$ near a photonic bandedge ω_U (the effects of the further PBG edge are neglected). This yields the effective mass approximation:

$$\omega \approx \omega_U + \sum_{i=x,y,z} (k - k_U)_i^2/m_i \qquad [11]$$

with $1/m_i = (1/2)(\partial^2\omega/\partial k_i^2)|_{\omega=\omega_U}$. In a structure with period L, k_U satisfies the Bragg condition $k_U = \pi/L$. The corresponding DOM in a 3D-periodic structure with an allowed symmetry may be approximated as

$$\rho(\omega) \sim (\omega - \omega_U)^{(1-D)/2}\theta(\omega - \omega_U) \qquad [12]$$

where θ is the step function and D is the dimension of the Brillouin-zone surface spanned by bandedge modes with vanishing group velocity. In realistic photonic structures $D \leq 2$, $D = 2$ corresponding to completely isotropic dispersion (spherical Brillouin-zone surface) and $D = 0$ corresponding to an anisotropic three-dimensional Brillouin zone. Both cases can be represented, respectively, as the limits $\varepsilon \to 0$ and $\varepsilon \to \infty$ of the function:

$$G(\omega) = \frac{C}{\pi}\frac{\sqrt{\omega - \omega_U}}{\omega - \omega_U + \varepsilon}\theta(\omega - \omega_U) \qquad [13]$$

where ε is the 'cutoff-smoothing' parameter and C is the continuum coupling constant. Depending on whether $C^{2/3}/\varepsilon$ is greater or less than one, the continuum is close to the case $\varepsilon = 0$ ($D = 2$) or $\varepsilon \to \infty$ ($D = 0$), respectively.

Using the properties of the model DOM (eqn [13]), we can infer the criteria for the two regimes discussed above:

(i) The conditions for incomplete decay, eqn [6], are now (Figure 2) $\Delta_c \equiv \omega_a - \omega_U < C\varepsilon^{-1/2}$. Hence, the abrupt, singular cutoff of the DOM with $D = 2$ ($\varepsilon \to 0$) implies the existence of a discrete state for any ω_a, either inside or outside the PBG. The energy of the discrete state, $\hbar\omega_0$, which must lie in the PBG, is found to be a real and positive root of the equation $\omega_a - \omega_0 = C/(\sqrt{\omega_U - \omega_0} + \sqrt{\varepsilon})$.

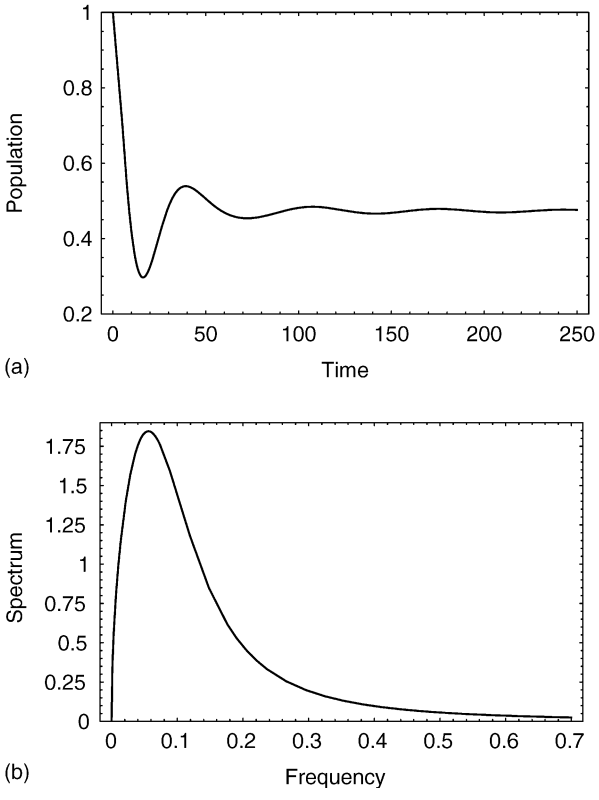


Figure 2 (a) Incomplete decay of the excited state population as a function of $\gamma_c t$, where $\gamma_c = C\varepsilon^{-1/2}$, for $\varepsilon = 10^{-3}$, at cutoff $\Delta_c = 0$. The beat period is $2\pi/(\omega_U - \omega_0)$ and the nondecaying probability is $c_0^2 = 4/9$. (b) The corresponding spectrum. The frequency is normalized to γ_c.

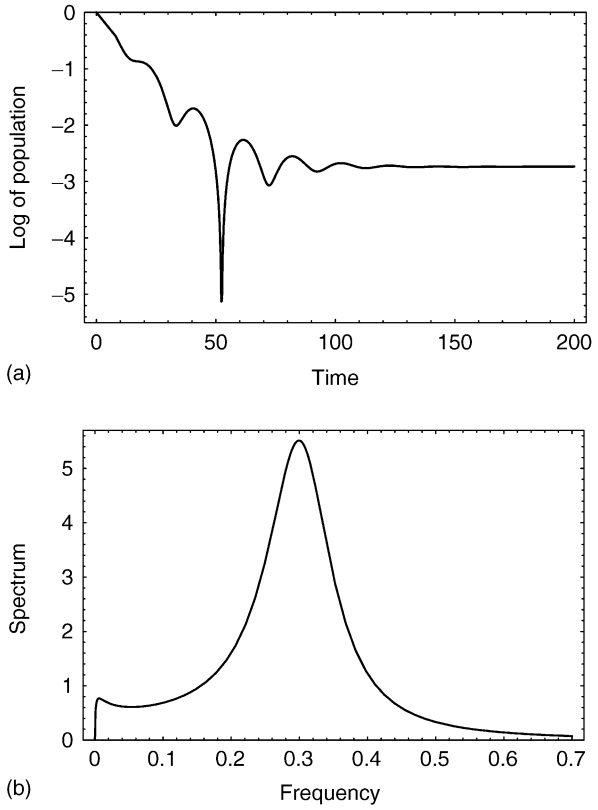


Figure 3 As in Figure 2, for a large detuning from the sharp cutoff, $\varepsilon = 10^{-3}$, $\Delta_c = 0.3$. (a) $\log_{10}|\alpha|t|^2$ is plotted as a function of *i*. The nearly complete decay ($c_0^2 \sim 10^{-3}$) is modulated by beats with frequency Δ_c. A power tail obtains for $t \gg \gamma_a^{-1} \approx \Delta_c^{1/2}/C$, decreasing as $t^{-1/2}$ for $t \ll \Delta_c^2/C^2$ and as $t^{-3/2}$ for $t \gg \Delta_c^2/C^2$. (b) The nearly Lorentzian spectrum, w(ω), has a small peak near ω_U, due to the sharply peaked DOM near the cutoff.

(ii) The conditions for nearly exponential decay, eqn [10], can be shown to reduce now to the requirement that ω_a be in the allowed zone sufficiently far from cutoff, $\Delta_c \gg \min\{C^{2/3}, C\varepsilon^{-1/2}\}$. Under this condition, the intermediate-time exponential decay of the excited state (Figure 3) is given to first approximation by eqn [9], with $\gamma_a = C\sqrt{\Delta_c}/(\varepsilon + \Delta_c)$ and $\Delta_a = -C\sqrt{\varepsilon}/(\varepsilon + \Delta_c)$.

The resulting atomic frequency shift Δ_a is negative in the present model, vanishing for $\varepsilon \to 0$ ($D = 2$). The long-time behavior of $\alpha(t)$ can be shown to exhibit a tail decaying as $t^{-3/2}$ (or $t^{-1/2}$ at $\Delta_c = C\varepsilon^{-1/2}$) and oscillating at the cutoff frequency ω_U (Figure 3).

With the increase of ε, the smoothing inhibits the decay more strongly for ω_a at cutoff, $\Delta_c = 0$, because $G(\omega_a \simeq \omega_U)$ is now weaker and the stable-state probability c_0 (eqn [7]) is correspondingly larger. By contrast, at a large detuning from cutoff Δ_c, the large-ε smoothing allows the decay to become complete, and the corresponding spectrum is entirely Lorentzian.

A defect in the periodic structure can produce a narrow-linewidth local mode in the PBG, whose spectral response is describable by a Lorentzian:

$$G_d(\omega) = \frac{\gamma_d}{\pi} \frac{\Gamma_d^2}{\Gamma_d^2 + (\omega - \omega_d)^2} \qquad [14]$$

where γ_d characterizes the coupling strength of the atomic dipole with the defect field, whereas ω_d and Γ_d stand for the line center and width, respectively. The presence of a nonvanishing DOM in the PBG, due to a defect, causes spontaneous emission in the PBG spectral range. This broadens the discrete state ω_0, which becomes metastable (see Figure 4). The oscillator strength of the line at ω_0 is then proportional to c_0.

EIT and Cross-Coupling of Photons in Doped PCs

Nonlinear effects, whereby one light beam influences another, require strong fields or else light confinement

in a high-Q cavity. The analysis above has shown that by choosing an appropriate detuning of an atomic transition from the PBG cutoff, the spontaneous coherence is established between the states of an initially excited atom. The ability to control this coherence, by varying the detuning, opens interesting perspectives for optical processes in PCs containing multilevel atoms with a resonant transition near a bandedge. From among such processes, we discuss here electromagnetically induced transparency (EIT) and its applicability to nonlinear photon switching or giant cross-phase modulation.

Let us examine the nonlinear coupling of two weak (single-photon) optical fields E_a and E_b with the frequencies ω_a and ω_b, respectively, propagating along the z-axis in a PC dilutely doped with identical four-level atoms (Figure 5a). These fields interact with the atoms via the transitions $|1\rangle \rightarrow |2\rangle$ and $|3\rangle \rightarrow |4\rangle$, respectively, while the transition $|2\rangle \rightarrow |3\rangle$ is coupled to the structured mode-continuum $\rho(\omega)$ in the PC (Figure 5b). Initially the atoms are in the ground state $|1\rangle$ and the continuum is in the vacuum state $\{0_\omega\}$. Then the wavefunction of the system at the position z_l of the lth atom reads:

$$|\Psi(z_l, t)\rangle = \alpha_1 |1, \{0_\omega\}\rangle + \alpha_2 |2, \{0_\omega\}\rangle + \int \beta_{3,\omega} |3, 1_\omega\rangle \rho(\omega) \mathrm{d}\omega + \int \beta_{4,\omega} |4, 1_\omega\rangle \rho(\omega) \mathrm{d}\omega \qquad [15]$$

The consecutive terms in eqn [15] denote the atom being in states $|1\rangle$, $|2\rangle$, $|3\rangle$, or $|4\rangle$, with zero $\{0_\omega\}$ or one $|1_\omega\rangle$ photon in mode ω whose DOM is $\rho(\omega)$, and α_1, α_2, $\beta_{3,\omega}$, or $\beta_{4,\omega}$ are the corresponding probability amplitudes. Upon making the weak-field linear-response approximation, we can set $\alpha_1 \simeq 1$ and solve the set of equations for the slowly varying (compared to an optical cycle) probability amplitudes α and β, using the perturbation theory. Under these conditions, we effectively obtain a free-space propagation of the E_b field. By contrast, the evolution of the E_a field in the slowly varying envelope approximation, is given by $E_a(z,t) = E_a(0, t - z/v_g)\exp(ipz)$. We thus see that the real part of the macroscopic complex polarizability p is responsible for the phase shift ϕ_a of the E_a field, $\phi_a = \mathrm{Re}(p)z$, while the probability of the absorption $\mathscr{A}$ of the field depends on the imaginary part of p, $\mathscr{A} = 1 - \exp[-2\mathrm{Im}(p)z]$. The polarizability is expressed by

$$p = a_0 \frac{i\gamma_2/2}{\gamma_2/2 - i\Delta_a + I(\Delta_a)} \qquad [16]$$

where $a_0 \equiv \sigma_0 N$ is the linear resonant absorption coefficient on the atomic transition $|1\rangle \rightarrow |2\rangle$, with σ_0 the resonant absorption cross-section and N the density of doping atoms, γ_2 is the radiative width of

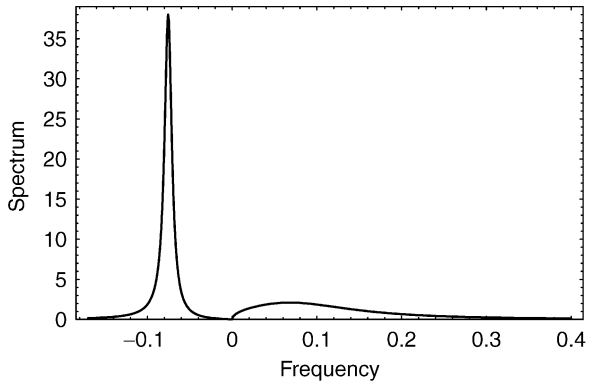


Figure 4 Spectrum w(ω)(same units as in Figures 2 and 3) for ω_a at cutoff, $\Delta_c = 0$, in the presence of a defect in the PBG, $\omega_d = -10$, $\gamma = 0.1$, $\Gamma_d = 3$, $\varepsilon = 10^{-3}$. The oscillator strength is now roughly equally shared between the distribution above cutoff (same as in Figure 2b) and the defect peak at ω_0.

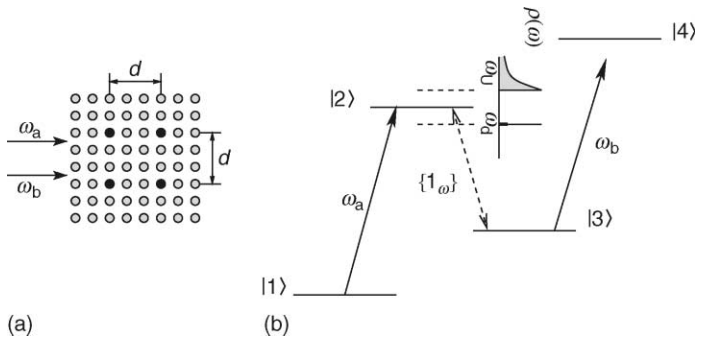


Figure 5 (a) Photonic crystal dilutely doped with atoms located at black dots. (b) Four-level atom coupled to a structured continuum $\rho(\omega)$ near the band-edge or defect mode frequencies (DOM plotted) via the intermediate transition $|2\rangle \rightarrow |3\rangle$ and interacting with two weak fields E_a and E_b at the sideband transitions $|1\rangle \rightarrow |2\rangle$ and $|3\rangle \rightarrow |4\rangle$, respectively.

state $|2\rangle$, $\Delta_a = \omega_a - \omega_{21}$ is the detuning from the atomic resonance ω_{21}, and $I(\Delta_a)$ is the integral of the saturation factor over the structured DOM. The group velocity v_g is expressed as $v_g = \partial_{k_a}\omega_a = [n_a/c + \partial_{\omega_a}\mathrm{Re}(p)]^{-1}$, where n_a is the (averaged) refraction index at the frequency ω_a.

To calculate $I(\Delta_a)$, we assume the isotropic PBG model, eqn [13] with $\varepsilon \to 0$, with the atoms doped at the positions of the local defects in the PC separated by a distance d from each other. These defect modes in the PBG are localized around each atomic site in a volume $V_d \simeq (rL)^3$ of several $(r)^3$ lattice cells L^3, with $L \simeq \pi c/\omega_{23}$, and serve as effective high-Q cavities, eqn [14] with $\Gamma_d \ll 1$. Assuming $|\Delta_b| \gg |\Delta|$, $|\Delta_a|$, γ_4, where $\Delta_b = \omega_b - \omega_{43}$ and $\Delta = \omega - \omega_{23}$, the integration leads to

$$I(\Delta_a) = \frac{B_d^2}{\gamma_{31} - i(\Delta_a - \Delta_d - s_3)} - \frac{B_U^{3/2}}{\sqrt{i\gamma_{31} + (\Delta_a - \Delta_U - s_3)}} \qquad [17]$$

where $\Delta_{d,U} = \omega_{d,U} - \omega_{23} \ll \omega_{23}$ are the detunings of the defect-mode and PBG-edge frequencies from the atomic resonance ω_{23}, γ_{31} the $|1\rangle \leftrightarrow |3\rangle$ decoherence rate, $s_3 = (\mu_{34}^2/\hbar^2\Delta_b)|E_b|^2$ is the E_b field-induced ac Stark shift of level $|3\rangle$ (μ_{ij} is the atomic dipole matrix element on the transition $|i\rangle \to |j\rangle$), and B_d and B_U are the coupling constants of the atom with the structured reservoir, whose main contributions are near ω_d and ω_U.

To illustrate the results of the foregoing analysis, we plot in Figure 6 the imaginary and real parts of the polarizability (eqn [16]). Consider first the case of one incident field $E_a(E_b = 0)$. Clearly, two frequency regions, $\Delta_a \sim \Delta_d$ and $\Delta_a \sim \Delta_U$, where the absorption vanishes and, at the same time, the dispersion slope is steep, are of particular interest. One can see that there is, however, a substantial difference between the spectra in the foregoing frequency regions, for the following physical reasons. First, in the vicinity of Δ_d, the atom interacts with the defect mode as in a high-Q cavity. This strong interaction 'dresses' the atomic states $|2\rangle$ and $|3\rangle$, thereby splitting the spectrum around $\Delta_a \simeq \Delta_d$ by the amount equal roughly to $2B_d$. Near the two-photon Raman resonance $\Delta_a = \Delta_d$, the two alternative transition paths $|1\rangle \to |2\rangle$ (the direct transition) and $|1\rangle \to |2\rangle \to |3\rangle \to |2\rangle$ (transition via state $|3\rangle$) interfere distractively with each other, cancelling thus the absorption of the E_a field and the medium becomes transparent to the radiation. This effect has been widely studied in atomic vapors, where the transition $|2\rangle \to |3\rangle$ is strongly driven by a coherent laser field, and is called electromagnetically induced transparency (EIT). The transparency window is rather broad and is given by the inverse Lorentzian (see the first term on the right-hand side of eqn [17]). Due to the steepness of the dispersion curve, the corresponding group velocity is much smaller than the speed of light, $v_g \simeq 2B_d^2/(\gamma_2 a_0) \ll c$, which leads to a large delay time $T_{del} = \zeta/v_g$ at the exit $z = \zeta$ from the medium. One has to keep in mind, however, that the absorption-free propagation time is limited by the EIT decoherence time $T_{del} < \gamma_{31}^{-1}$, which imposes a limitation on the length ζ of the active PC medium. Second, in the vicinity of Δ_U, the strong interaction of the atom with the continuum near the bandedge ω_U causes the Autler–Townes splitting of level $|2\rangle$ into a doublet with a separation equal roughly to B_U. One component of this doublet is shifted out of the PBG, while the other one remains

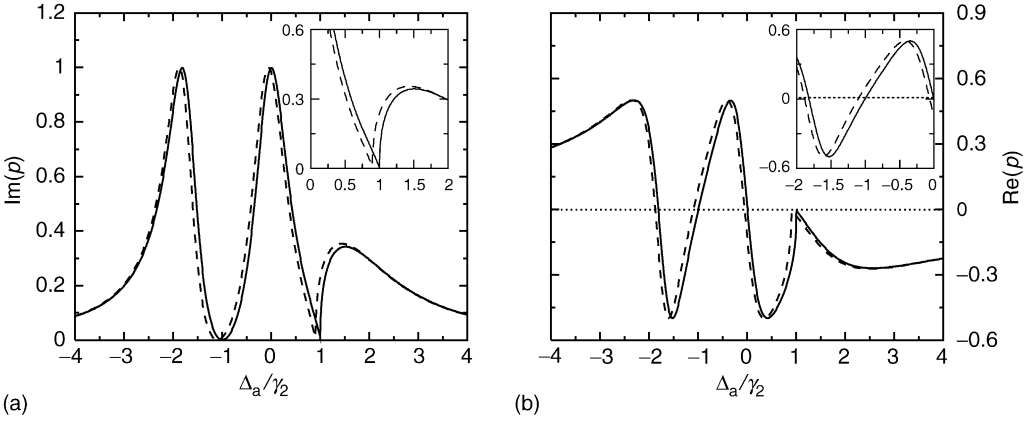


Figure 6 (a) Imaginary and (b) real part of the complex polarizability p as a function of the detuning Δ_a for the case $E_b = 0$ (solid lines) and $E_b \neq 0$ (dashed lines). The parameters (normalized by γ_2) are: $\Delta_d = -1$, $\Delta_U = 1$, $\gamma_{31} = 0.001$, $B_d = B_U = 1$, $s_3 = -0.1$, and $a_0 = 1\ \mathrm{cm}^{-1}$. The insets magnify the important frequency regions.

within the gap and forms the photon–atom bound state. Consequently, there is vanishing absorption and rapid variation of the dispersion at $\Delta_a \simeq \Delta_U$. Since the transparency region is very narrow with a width $\delta\omega \sim \gamma_{31}(\ll \gamma_2, B_U)$, for an absorption-free propagation of the E_a pulse, its temporal width τ_a should satisfy the condition $\tau_a > \pi/\delta\omega$. Simultaneously, a small deviation from the condition $\Delta_a = \Delta_U$ will lead to a strong increase in the absorption of the E_a field.

Let us now switch on the E_b field. As seen from eqn [17], its effect is merely to shift the spectrum by the amount equal to s_3 (Figure 6). This shift, however, will have different implications in the two frequency regions distinguished above: if $s_3 \ll B_d$, i.e., the Stark shift is smaller than the width of the EIT window at Δ_d, the medium will still remain transparent for an E_a field with the detuning $\Delta_a = \Delta_d$, but its phase will experience an appreciable nonlinear shift ϕ_a, given by $\phi_a = Re(p)z \simeq -\gamma_2 a_0 s_3 z/(2B_d^2)$. On the other hand, for an E_a field with the detuning $\Delta_a = \Delta_U$, the medium, which is transparent for $E_b = 0$, $\mathrm{Im}(p) \ll a_0$, will become highly absorptive (opaque) even for such a small frequency shift as s_3 (provided $s_3 < 0$ and $|s_3| > \Delta\omega$), $\mathrm{Im}(p) \simeq \gamma_2 a_0 \sqrt{|s_3|}/(2B_U^{3/2})$ and thus acting as an ultrasensitive, effective switch.

The remaining question is how to maximize the interaction between the E_a pulse, which propagates with a small group velocity, and the E_b pulse, which propagates with a velocity close to the speed of light. The interaction between the fields is maximized if: they enter the medium simultaneously; the transverse shapes of their envelopes overlap completely; and the pulse length l_b of the E_b field satisfies the condition $(l_b + \zeta)/c \leq \zeta/v_g$. Then the E_b pulse leaves the medium not later than the E_a pulse. The effective interaction length between the two pulses is, therefore, $z_{eff} \sim l_b v_g/c \leq \zeta$, after which the two pulses slip apart. Thus, the presence of the E_b (control) field induces either strong absorption or a large phase shift of the E_a (signal) field, depending on the frequency region employed.

The effects surveyed above reveal unusual features of spontaneous emission and photon–atom binding in PCs, atomic interaction with the field of a high-Q defect mode, and nonlinear coupling of two fields via four-level dopant atoms in PCs. These effects are of fundamental interest. In addition, they can serve as the basis for highly efficient optical communications and data processing, in either the classical or the quantum domain, by providing two key elements: ultrasensitive nonlinear phase-shifters and photon switches.

List of Units and Nomenclature

Atomic or photonic eigenstates [dimensionless]	$\|\ldots\rangle$
Complex polarizability [cm^{-1}]	p
Coupling constant [s^{-1}]	κ, $C^{2/3}$, B
Decay rate [s^{-1}]	γ
Density of modes [s]	ρ
Frequency [s^{-1}]	ω, Δ, s
Phase [dimensionless]	ϕ
Probability amplitudes [dimensionless]	α, β, c
Reservoir spectral response [s^{-1}]	G
Resonant absorption coefficient [cm^{-1}]	a_0
Spontaneous emission spectrum [s]	W
Wave vector [m^{-1}]	k

See also

Electromagnetically Induced Transparency. **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguides.

Further Reading

Joannopoulos JD, Meade RD and Winn JN (1995) *Photonic Crystals: Molding the Flow of Light*. Princeton, NJ: Princeton University Press.

John S and Wang J (1991) Quantum optics of localized light in a photonic band gap. *Physics Review B* 43: 12772–12789.

Kofman AG, Kurizki G and Sherman B (1994) Spontaneous and induced atomic decay in photonic band structures. *Journal of Modern Optics* 41: 353–384.

Lambropoulos P, Nikolopoulos GM, Nielsen TR and Bay S (2000) Fundamental quantum optics in structured reservoirs. *Reports on Progress in Physics* 63: 455–503.

Loudon R (1983) *The Quantum Theory of Light*. Oxford, UK: Clarendon Press.

Meystre P and Sargent M III (1991) *Elements of Quantum Optics*. Berlin: Springer Verlag.

Petrosyan D and Kurizki G (2001) Photon–photon correlations and entanglement in doped photonic crystals. *Physics Review A* 64: 023810-1-6.

Sakoda K (2001) *Optical Properties of Photonic Crystals*. Springer Series in Optical Sciences, 80. Berlin: Springer Verlag.

Scully MO and Zubairy MS (1997) *Quantum Optics*. Cambridge, UK: Cambridge University Press.

Electromagnetic Theory

S G Johnson and J D Joannopoulos,
Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Photonic crystals are periodically structured electromagnetic media, generally possessing photonic bandgaps: ranges of frequency in which light cannot propagate through the structure. This periodicity, whose lengthscale is proportional to the wavelength of light in the bandgap, is the electromagnetic analog of a crystalline atomic lattice, where the latter acts on the electron wavefunction to produce the familiar band gaps, semiconductors, etc., of solid-state physics. The study of photonic crystals is likewise governed by the Bloch–Floquet theorem, and intentionally introduced defects in the crystal (analogous to electronic dopants) give rise to localized electromagnetic states: linear waveguides and point-like cavities. The crystal can thus form a kind of perfect optical 'insulator', which can confine light around sharp bends, in lower-index media, and within wavelength-scale cavities, among other novel possibilities for control of electromagnetic phenomena. Below is introduced the basic theoretical background of photonic crystals in one, two, and three dimensions (schematically depicted in Figure 1), as well as hybrid structures that combine photonic-crystal effects in some directions with more-conventional index guiding in other directions. (Line and point defects in photonic crystals are discussed in another article.)

Electromagnetic wave propagation in periodic media was first studied by Lord Rayleigh in 1887, in connection with the peculiar reflective properties of a crystalline mineral with periodic 'twinning' planes (across which the dielectric tensor undergoes a mirror flip). These correspond to one-dimensional photonic crystals, and he identified the fact that they have a narrow bandgap prohibiting light propagation through the planes. This bandgap is angle-dependent, due to the differing periodicities experienced by light propagating at non-normal incidences, producing a reflected color that varies sharply with angle. (A similar effect is responsible for many other iridescent colors in nature, such as those of butterfly wings and abalone shells.) Although multilayer films received intensive study over the following century, it was not until 100 years later, when Yablonovitch and John, in 1987, joined the tools of classical electromagnetism and solid-state physics, that the concepts of omnidirectional photonic bandgaps in two and three dimensions was introduced. This generalization, which inspired the name 'photonic crystal', led to many subsequent developments in their fabrication, theory, and application, from integrated optics to negative refraction to optical fibers that guide light in air.

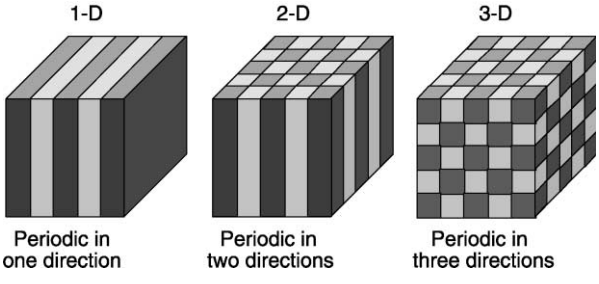


Figure 1 Schematic depiction of photonic crystals periodic in one, two, and three directions, where the periodicity is in the material (typically dielectric) structure of the crystal. Only a 3d periodicity, with a more complex topology than is shown here, can support an omnidirectional photonic bandgap.

Maxwell's Equations in Periodic Media

The study of wave propagation in three-dimensionally periodic media was pioneered by Felix Bloch in 1928, unknowingly extending an 1883 theorem in one dimension by Gaston Floquet. Bloch proved that waves in such a medium can propagate without scattering, their behavior described by a periodic envelope function multiplied by a planewave. Although Bloch studied quantum mechanics, leading to the surprising result that electrons in a conductor scatter only from imperfections and not from the periodic ions, the same techniques can be applied to electromagnetism by casting Maxwell's equations as an eigenproblem in analog with Schrödinger's equation. By combining the source-free Faraday's and Ampere's laws at a fixed (angular) frequency ω, i.e., time dependence $e^{-i\omega t}$, one can obtain an equation in only the magnetic field $\vec{H}$:

$$\vec{\nabla} \times \frac{1}{\varepsilon} \vec{\nabla} \times \vec{H} = \left(\frac{\omega}{c}\right)^2 \vec{H} \qquad [1]$$

where ε is the dielectric function $\varepsilon(x, y, z)$ and c is the speed of light. This is an eigenvalue equation, with eigenvalue $(\omega/c)^2$ and an eigen-operator $\vec{\nabla} \times (1/\varepsilon)\vec{\nabla} \times$ that is Hermitian (acts the same to the left and right)

under the inner product $\int \vec{H}^* \cdot \vec{H}'$ between two fields $\vec{H}$ and $\vec{H}'$. (The two curls correspond roughly to the 'kinetic energy' and $1/\varepsilon$ to the 'potential' compared to the Schrödinger Hamiltonian $\nabla^2 + V$.) It is sometimes more convenient to write a generalized Hermitian eigenproblem in the electric field $\vec{E}$, $\vec{\nabla} \times \vec{\nabla} \times \vec{E} = (\omega/c)^2 \varepsilon \vec{E}$, which separates the kinetic and potential terms. Electric fields that lie in higher ε, i.e., lower potential, will have lower ω; this is discussed further in the context of the variational theorem of eqn [3].

Thus, the same linear-algebraic theorems as those in quantum mechanics can be applied to the electromagnetic wave solutions. The fact that the eigen-operator is Hermitian and positive-definite (for real $\varepsilon > 0$) implies that the eigenfrequencies ω are real, for example, and also leads to orthogonality, variational formulations, and perturbation–theory relations that are discussed further below. An important difference compared to quantum mechanics is that there is a transversality constraint: one typically excludes $\vec{\nabla} \cdot \vec{H} \neq 0$ (or $\vec{\nabla} \cdot \varepsilon \vec{E} \neq 0$) eigensolutions, which lie at $\omega = 0$; i.e., static-field solutions with free magnetic (or electric) charge are forbidden.

Bloch Waves and Brillouin Zones

A photonic crystal corresponds to a periodic dielectric function $\varepsilon(\vec{x}) = \varepsilon(\vec{x} + \vec{R}_i)$ for some primitive lattice vectors $\vec{R}_i$ ($i = 1, 2, 3$ for a crystal periodic in all three dimensions). In this case, the Bloch–Floquet theorem for periodic eigenproblems states that the solutions to eqn [1] can be chosen of the form $\vec{H}(\vec{x}) = e^{i\vec{k}\cdot\vec{x}} \vec{H}_{n,\vec{k}}(\vec{x})$ with eigenvalues $\omega_n(\vec{k})$, where $\vec{H}_{n,\vec{k}}$ is a periodic envelope function satisfying

$$\left(\vec{\nabla} + i\vec{k}\right) \times \frac{1}{\varepsilon}\left(\vec{\nabla} + i\vec{k}\right) \times \vec{H}_{n,\vec{k}} = \frac{\omega_n(\vec{k})}{c}^2 \vec{H}_{n,\vec{k}} \qquad [2]$$

yielding a different Hermitian eigenproblem over the primitive cell of the lattice at each Bloch wavevector $\vec{k}$. This primitive cell is a finite domain if the structure is periodic in all directions, leading to discrete eigenvalues labeled by $n = 1, 2, \ldots$. These eigenvalues $\omega_n(\vec{k})$ are continuous functions of $\vec{k}$, forming discrete 'bands' when plotted versus the latter, in a 'band structure' or dispersion diagram – both ω and $\vec{k}$ are conserved quantities, meaning that a band diagram maps out all possible interactions in the system. (Note also that $\vec{k}$ is not required to be real; complex $\vec{k}$ gives evanescent modes that can exponentially decay from the boundaries of a finite crystal, but which cannot exist in the bulk.)

Moreover, the eigensolutions are periodic functions of $\vec{k}$ as well: the solution at $\vec{k}$ is the same as the solution at $\vec{k} + \vec{G}_j$, where $\vec{G}_j$ is a primitive reciprocal lattice vector defined by $\vec{R}_i \cdot \vec{G}_j = 2\pi\delta_{i,j}$. Thanks to this periodicity, one need only compute the eigensolutions for $\vec{k}$ within the primitive cell of this reciprocal lattice – or, more conventionally, one considers the set of inequivalent wavevectors closest to the $\vec{k} = 0$ origin, a region called the first Brillouin zone. For example, in a one-dimensional system, where $R_1 = a$ for some periodicity a and $G_1 = 2\pi/a$, the first Brillouin zone is the region $k = -\pi/a \ldots \pi/a$; all other wavevectors are equivalent to some point in this zone under translation by a multiple of G_1. Furthermore, the first Brillouin zone may itself be redundant if the crystal possesses additional symmetries such as mirror planes; by eliminating these redundant regions, one obtains the irreducible Brillouin zone, a convex polyhedron that can be found tabulated for most crystalline structures. In the preceding one-dimensional example, since most systems will have time-reversal symmetry ($k \to -k$), the irreducible Brillouin zone would be $k = 0 \ldots \pi/a$.

The familar dispersion relations of uniform waveguides arise as a special case of the Bloch formalism: such translational symmetry corresponds to a period $a \to 0$. In this case, the Brillouin zone of the wavevector k (also called β) is unbounded, and the envelope function $\vec{H}_{n,k}$ is a function only of the transverse coordinates.

The Origin of the Photonic Bandgap

A complete photonic bandgap is a range of ω in which there are no propagating (real $\vec{k}$) solutions of Maxwell's eqn [2] for any $\vec{k}$, surrounded by propagating states above and below the gap. There are also incomplete gaps, which only exist over a subset of all possible wavevectors, polarizations, and/or symmetries. Both sorts of gaps are discussed in the subsequent sections, but in either case, their origins are the same and can be understood by examining the consequences of periodicity for a simple one-dimensional system.

Consider a one-dimensional system with uniform $\varepsilon = \bar{\varepsilon}$, which has planewave eigensolutions $\omega(k) = ck$, as depicted in Figure 2(left). This ε has trivial periodicity a for any $a \geq 0$, with $a = 0$ giving the usual unbounded dispersion relation. One is free, however, to label the states in terms of Bloch envelope functions and wavevectors for some $a \neq 0$, in which case the bands for $|k| > \pi/a$ are translated (folded) into the first Brillouin zone, as shown by the dashed lines in Figure 2(left). In particular, the $k = -\pi/a$ mode in this description now lies at an equivalent wavevector to the $k = \pi/a$ mode, and at the same frequency; this accidental degeneracy is an artifact of

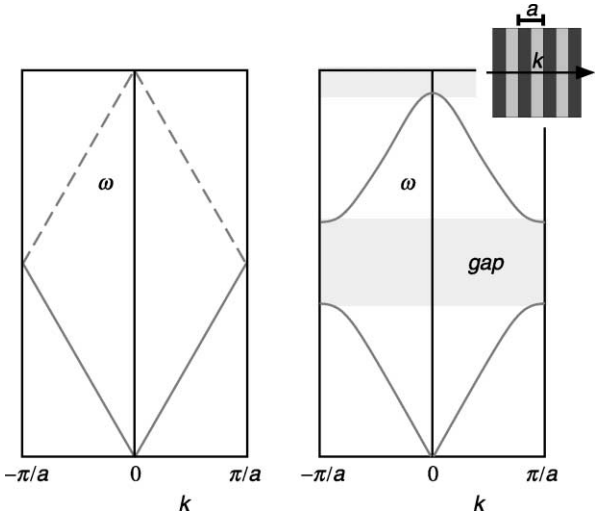


Figure 2 Left: Dispersion relation (band diagram), frequency ω versus wavenumber k, of a uniform one-dimensional medium, where the dashed lines show the 'folding' effect of applying Bloch's theorem with an artificial periodicity a. Right: Schematic effect on the bands of a physical periodic dielectric variation (inset), where a gap has been opened by splitting the degeneracy at the $k = \pm\pi/a$ Brillouin-zone boundaries (as well as a higher-order gap at $k = 0$).

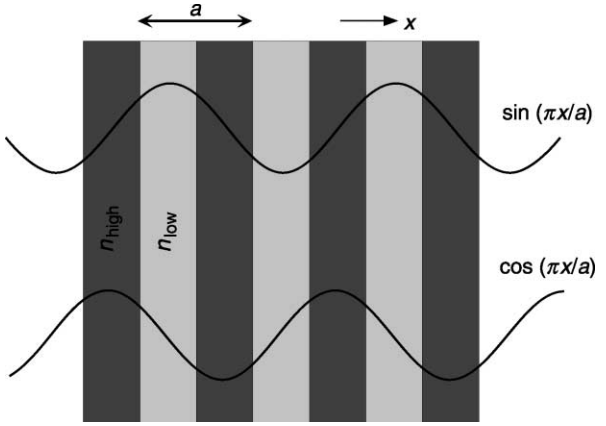


Figure 3 Schematic origin of the band gap in one dimension. The degenerate $ka = \pm\pi/a$ planewaves of a uniform medium are split into $\cos(\pi x/a)$ and $\sin(\pi x/a)$ standing waves by a dielectric periodicity, forming the lower and upper edges of the bandgap, respectively – the former has electric-field peaks in the high dielectric (n_{high}) and so will lie at a lower frequency than the latter (which peaks in the low dielectric).

the 'artificial' period that has been chosen. Instead of writing these wave solutions with electric fields $\vec{E}(x) \sim e^{\pm i\pi x/a}$, one can equivalently write linear combinations $e(x) = \cos(\pi x/a)$ and $o(x) = \sin(\pi x/a)$ as shown in Figure 3, both at $\omega = c\pi/a\sqrt{\bar{\varepsilon}}$. Now, however, suppose that one perturbs ε so that it is nontrivially periodic with period a; for example, a sinusoid $\varepsilon(x) = \bar{\varepsilon}\cdot[1 + \Delta\cdot\cos(2\pi x/a)]$, or a square wave as in the inset of Figure 2(right). In the presence of such an oscillating 'potential', the accidental degeneracy between $e(x)$ and $o(x)$ is broken: supposing $\Delta > 0$, then the field $e(x)$ is more concentrated in the higher-ε regions than $o(x)$, and so lies at a lower frequency. This opposite shifting of the bands away from the mid-gap frequency $\omega \cong c\pi/a\sqrt{\bar{\varepsilon}}$ creates a bandgap, as depicted in Figure 2(right). (In fact, from the perturbation theory described subsequently, one can show that for $\Delta \ll 1$ the bandgap, as a fraction of mid-gap frequency, is $\Delta\omega/\omega \cong \Delta/2$.) By the same arguments, it follows that any periodic dielectric variation in one dimension will lead to a bandgap, albeit a small gap for a small variation; a similar result was identified by Lord Rayleigh in 1887.

More generally, it follows immediately from the properties of Hermitian eigensystems that the eigenvalues minimize a variational problem:

$$\omega^2_{n,\vec{k}} = \min_{\vec{E}_{n,\vec{k}}} \frac{\int \left|\left(\vec{\nabla} + i\vec{k}\right)\times \vec{E}_{n,\vec{k}}\right|^2}{\int \varepsilon |\vec{E}_{n,\vec{k}}|^2} c^2 \qquad [3]$$

in terms of the periodic electric field envelope $\vec{E}_{n,\vec{k}}$, where the numerator minimizes the 'kinetic energy' and the denominator minimizes the 'potential energy' Here, the $n > 1$ bands are additionally constrained to be orthogonal to the lower bands:

$$\int \vec{H}^*_{m,\vec{k}} \cdot \vec{H}_{n,\vec{k}} = \int \varepsilon \vec{E}^*_{m,\vec{k}} \cdot \vec{E}_{n,\vec{k}} = 0 \qquad [4]$$

for $m < n$. Thus, at each $\vec{k}$, there will be a gap between the lower 'dielectric' bands concentrated in the high dielectric (low potential) and the upper 'air' bands that are less concentrated in the high dielectric: the air bands are forced out by the orthogonality condition, or otherwise must have fast oscillations that increase their kinetic energy. (The dielectric/air bands are analogous to the valence/conduction bands in a semiconductor.)

In order for a complete bandgap to arise in two or three dimensions, two additional hurdles must be overcome. First, although in each symmetry direction of the crystal (and each $\vec{k}$ point) there will be a bandgap by the one-dimensional argument, these bandgaps will not necessarily overlap in frequency (or even lie between the same bands). In order that they overlap, the gaps must be sufficiently large, which implies a minimum ε contrast (typically at least 4/1 in 3d). Since the 1d mid-gap frequency $\sim c\pi/a\sqrt{\bar{\varepsilon}}$ varies inversely with the period a, it is also helpful if the periodicity is nearly the same in different directions – thus, the largest gaps typically arise for hexagonal lattices in 2d and fcc lattices in 3d, which have the most nearly circular/spherical Brillouin zones. Second, one must take into account

the vectorial boundary conditions on the electric field: moving across a dielectric boundary from ε to some $\varepsilon' < \varepsilon$, the inverse 'potential' $\varepsilon|\vec{E}|^2$ will decrease discontinuously if $\vec{E}$ is parallel to the interface ($\vec{E}_\parallel$ is continuous) and will increase discontinuously if $\vec{E}$ is perpendicular to the interface ($\varepsilon\vec{E}_\perp$ is continuous). This means that, whenever the electric field lines cross a dielectric boundary, it is much harder to strongly contain the field energy within the high dielectric, and the converse is true when the field lines are parallel to a boundary. Thus, in order to obtain a large bandgap, a dielectric structure should consist of thin, continuous veins/membranes along which the electric field lines can run – this way, the lowest band(s) can be strongly confined, while the upper bands are forced to a much higher frequency because the thin veins cannot support multiple modes (except for two orthogonal polarizations). The veins must also run in all directions, so that this confinement can occur for all $\vec{k}$ and polarizations, necessitating a complex topology in the crystal.

Ultimately, however, in two or three dimensions there are only rules of thumb for the existence of a bandgap in a periodic structure, since no rigorous criteria have yet been determined. This made the design of 3d photonic crystals a trial and error process, with the first example by Ho *et al.* of a complete 3d gap coming three years after the initial 1987 concept. As is discussed by the final section below, a small number of families of 3d photonic crystals have since been identified, with many variations thereof explored for fabrication.

Computational Techniques

Because photonic crystals are generally complex, high index-contrast, two- and three-dimensional vectorial systems, numerical computations are a crucial part of most theoretical analyses. Such computations typically fall into three categories: time-domain 'numerical experiments' that model the time-evolution of the fields with arbitrary starting conditions in a discretized system (e.g., finite-difference); definite-frequency transfer matrices wherein the scattering matrices are computed in some basis to extract transmission/reflection through the structure; and frequency-domain methods to directly extract the Bloch fields and frequencies by diagonalizing the eigenoperator. The first two categories intuitively correspond to directly measurable quantities such as transmission (although they can also be used to compute e.g., eigenvalues), whereas the third is more abstract, yielding the band diagrams that provide a guide to interpretation of measurements as well as a starting-point for device design and semi-analytical methods. Moreover, several band diagrams are included in the following sections, and so the frequency-domain method used to compute them is briefly outlined here.

Any frequency–domain method begins by expanding the fields in some complete basis, $\vec{H}_{\vec{k}}(\vec{x}) = \sum_n h_n \vec{b}_n(\vec{x})$, transforming the partial differential eqn [2] into a discrete matrix eigenvalue problem for the coefficients h_n. Truncating the basis to N elements leads to $N \times N$ matrices, which could be diagonalized in $O(N^3)$ time by standard methods. This is impractical for large 3d systems, however, and is also unnecessary – typically, one only wants the few lowest eigenfrequencies, in which case one can use iterative eigensolver methods requiring only $\sim O(N)$ time. Perhaps the simplest such method is based directly on the variational theorem eqn [3]: given some starting coefficients h_n, one iteratively minimizes the variational 'Rayleigh' quotient using e.g., preconditioned conjugate–gradient descent. This yields the lowest band's eigenvalue and field, and upper bands are found by the same minimization while orthogonalizing against the lower bands ('deflation'). There is one additional difficulty, however, and that is that one must at the same time enforce the $(\vec{\nabla} + i\vec{k}) \cdot \vec{H}_{\vec{k}} = 0$ transversality constraint, which is nontrivial in three dimensions. The simplest way to maintain this constraint is to employ a basis that is already transverse, for example planewaves $\vec{h}_{\vec{G}} e^{i\vec{G}\cdot\vec{x}}$ with transverse amplitudes $\vec{h}_{\vec{G}} \cdot (\vec{G} + \vec{k}) = 0$. (In such a planewave basis, the action of the eigen-operator can be computed via a fast Fourier transform in $O(N \log N)$ time.)

Semi-analytical Methods: Perturbation Theory

As in quantum mechanics, the eigenstates can be the starting point for many analytical and semi-analytical studies. One common technique is perturbation theory, applied to small deviations from an ideal system – closely related to the variational eqn [3], perturbation theory can be exploited to consider effects such as nonlinearities, material absorption, fabrication disorder, and external tunability. Not only is perturbation theory useful in its own right, but it also illustrates both old and peculiarly new features that arise in such analyses of electromagnetism compared to scalar problems such as quantum mechanics.

Given an unperturbed eigenfield $\vec{E}_{n,\vec{k}}$ for a structure ε, the lowest-order correction $\Delta\omega_n^{(1)}$ to the eigenfrequency from a small perturbation $\Delta\varepsilon$ is given by

$$\Delta\omega_n^{(1)} = -\frac{\omega_n(\vec{k})}{2} \frac{\int \Delta\varepsilon \left|\vec{E}_{n,\vec{k}}\right|^2}{\int \varepsilon \left|\vec{E}_{n,\vec{k}}\right|^2} \qquad [5]$$

where the integral is over the primitive cell of the lattice. A Kerr nonlinearity would give $\Delta\varepsilon \sim |\vec{E}|^2$, material absorption would produce an imaginary frequency correction (decay coefficient) from a small imaginary $\Delta\varepsilon$, and so on. Similarly, one can compute the shift in frequency from a small $\Delta\vec{k}$ in order to determine the group velocity $d\omega/dk$; this variation of perturbation theory is also called $k \cdot p$ theory. All such first-order perturbation corrections are well known from quantum mechanics, and in the limit of infinitesimal perturbations give the exact Hellman–Feynman expression for the derivative of the eigenvalue. However, in the limit where $\Delta\varepsilon$ is a small shift Δh of a dielectric boundary between some ε_1 and ε_2, an important class of geometric perturbation, eqn [5] gives a surface integral of $|\vec{E}|^2$ on the interface, but this is ill-defined because the field there is discontinous. The proper derivation of perturbation theory in the face of such discontinuity requires a more careful limiting process from an anisotropically smoothed system, yielding the surface integral:

$$\Delta\omega_n^{(1)} = -\frac{\omega_n(\vec{k})}{2} \frac{\int\int \Delta h \left(\Delta\varepsilon_{12}|\vec{E}_{\parallel}|^2 - \Delta\varepsilon_{12}^{-1}|D_{\perp}|^2\right)}{\int \varepsilon |\vec{E}_{n,\vec{k}}|^2}$$

where $\Delta\varepsilon_{12} = \varepsilon_1 - \varepsilon_2, \Delta\varepsilon_{12}^{-1} = \varepsilon_1^{-1} - \varepsilon_2^{-1}$, and $\vec{E}_{\parallel}/D_{\perp}$ denotes the (continuous) interface parallel/perpendicular components of the unperturbed electric/displacement eigenfield. A similar expression is required in high index-contrast systems to employ, e.g., coupled-mode theory for slowly-varying waveguides or Green's functions for interface roughness.

Standard perturbation–theory techniques also provide expressions for higher-order corrections to the eigenvalue and eigenfield, based on an expansion in the basis of the unperturbed eigenfields. This approach, however, runs into immediate difficulty because the eigenfields are also subject to the transversality constraint, $(\vec{\nabla} + i\vec{k}) \cdot \varepsilon \vec{E}_{\vec{k}} = 0$, and this constraint varies with ε and $\vec{k}$ – the eigenfields $\vec{E}_{\vec{k}}$ are not a complete basis for the eigenfields constrained at a different ε or $\vec{k}$. For ε perturbations, this problem can be eliminated by using the $\vec{H}$ or $\vec{D}$ eigenproblems, whose constraints are independent of ε. For $\vec{k}$ perturbations, one can employ a transformation by Sipe to derive a corrected higher-order perturbation theory (for e.g., the group-velocity dispersion), based on the fact that all of the non-transverse fields lie at $\omega = 0$. Such completeness issues also arise applying the variational eqn [3], as was noted in the previous section: in order for a useful variational bound to apply, one must operate in the constrained (transverse) subspace.

Two-Dimensional Photonic Crystals

After the identification of one-dimensional bandgaps, it took a full century to add a second dimension, and three years to add the third. It should therefore come as no surprise that 2d systems exhibit most of the important characteristics of photonic crystals, from nontrivial Brillouin zones to topological sensitivity to a minimum index contrast, and can also be used to demonstrate most proposed photonic–crystal devices. The key to understanding photonic crystals in two dimensions is to realize that the fields in 2d can be divided into two polarizations by symmetry: *TM* (transverse magnetic), in which the magnetic field is in the (xy) plane and the electric field is perpendicular (z); and *TE* (transverse electric), in which the electric field is in the plane and the magnetic field is perpendicular.

Corresponding to the polarizations, there are two basic topologies for 2d photonic crystals, as depicted in Figure 4(top): high index rods surrounded by low index (top) and low-index holes in high index (bottom). Here, a hexagonal lattice is used because, as noted earlier, it gives the largest gaps. Recall that a photonic band gap requires that the electric field lines run along thin veins: thus, the rods are best suited to TM light ($\vec{E}$ parallel to the rods), and the holes are best suited to TE light ($\vec{E}$ running around the holes). This preference is reflected in the band diagrams, shown in Figure 4, in which the rods/holes (top/bottom) have a strong TM/TE band gap. For these diagrams, the rod/hole radius is chosen to be $0.2a/0.3a$, where a is the lattice constant (the nearest-neighbor periodicity) and the high/low ε is 12/1. The TM/TE bandgaps are then 47%/28% as a fraction of mid-gap frequency, but these bandgaps require a minimum ε contrast of 1.7/1 and 1.9/1, respectively. Moreover, it is conventional to give the frequencies ω in units of $2\pi c/a$, which is equivalent to a/λ (λ being the vacuum wavelength) – Maxwell's equations are scale-invariant, and the same solutions can be applied to any wavelength simply by choosing the appropriate a. For example, the TM mid-gap ω in these units is 0.36, so if one wanted this to correspond to $\lambda = 1.55\ \mu m$ one would use $a = 0.36 \cdot 1.55\ \mu m = 0.56\ \mu m$.

The Brillouin zone (a hexagon) is shown at left-center, with the irreducible Brillouin zone shaded (following the sixfold symmetry of the crystal); the corners (high symmetry points) of this zone are given canonical names, where Γ always denotes the origin $\vec{k} = 0$, K is the nearest-neighbor direction,

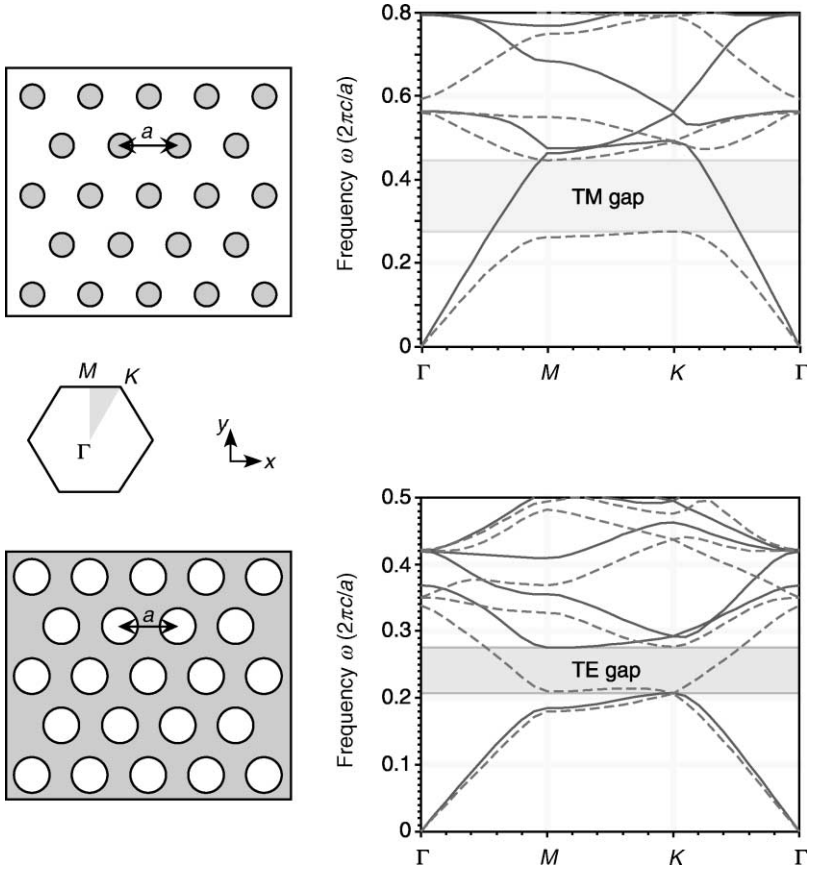


Figure 4 Band diagrams and photonic band gaps for hexagonal lattices of high dielectric rods ($\varepsilon = 12$, $r = 0.2a$) in air (top), and air holes ($r = 0.3a$) in dielectric (bottom), where a is the center–center periodicity. The frequencies are plotted around the boundary of the irreducible Brillouin zone (shaded triangle, left center), with solid/dashed lines denoting TE/TM polarization (electric field parallel/perpendicular to plane of periodicity). The rods/holes have a gap in the TM/TE bands.

and M is the next-nearest-neighbor direction. The Brillouin zone is a two-dimensional region of wavevectors, so the bands $\omega_n(\vec{k})$ are actually surfaces, but in practice the band extrema almost always occur along the boundaries of the irreducible zone (i.e., the high-symmetry directions). So, it is conventional to plot the bands only along these zone boundaries in order to identify the bandgap, as is done in Figure 4.

Actually, the hole lattice can display not only a TE gap, but a complete photonic bandgap (for both polarizations) if the holes are sufficiently large (nearly touching). In this case, the thin veins between nearest-neighbor holes induce a TE gap, while the interstices between triplets of holes form 'rod-like' regions that support a TM gap overlapping the TE gap.

Photonic-Crystal Slabs

In order to realize 2d photonic-crystal phenomena in three dimensions, the most straightforward design is to simply fabricate a 2d-periodic crystal with a finite height: a photonic-crystal slab, as depicted in Figure 5. Such a structure can confine light vertically within the slab via index guiding, a generalization of total internal reflection – this mechanism is the source of several new tradeoffs and behaviors of slab systems compared to their 2d analogs.

The key to index guiding is the fact that the 2d periodicity implies that the 2d Bloch wavevector $\vec{k}_{\|}$ is a conserved quantity, so the projected band structure – all states in the bulk substrate/superstrate (the uniform regions far below/above the slab) versus their in-plane wavevector component (projected wavevector) – creates a map of which states can radiate vertically. If the slab is suspended in air, for example, then the eigensolutions of the bulk air are $\omega = c\sqrt{|\vec{k}_{\|}|^2 + k_{\perp}^2}$, which when plotted versus $\vec{k}_{\|}$ forms the continuous light cone $\omega \geq c|\vec{k}_{\|}|$, shown as a shaded region in Figure 5. Because the slab has a higher ε (12) than the air (1), and frequency goes as $1/\sqrt{\varepsilon}$, discrete guided bands are 'pulled down' in frequency from this continuum – these bands, lying beneath the light cone, cannot couple to any vertically

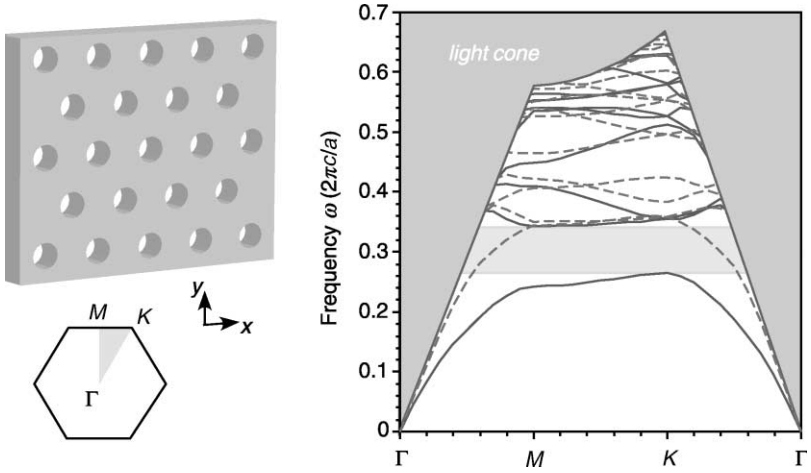


Figure 5 Projected band diagram for a finite-thickness (0.5*a*) slab of air holes in dielectric (cross section as in **Figure 4** bottom), with the irreducible Brillouin zone at lower left. The shaded region is the light cone: the projection of all states that can radiate in the air. Solid/dashed lines denote guided modes (confined to the slab) that are even/odd with respect to the horizontal mirror plane of the slab, whose polarization is TE-like/TM-like, respectively. There is a 'bandgap' (region without guided modes) in the TE-like guided modes only.

radiating mode by the conservation law and so are confined to the slab (exponentially decaying away from it). If the horizontal mid-plane of the slab is a mirror symmetry plane, then just as there were TM and TE states in 2d, here there are two categories of modes: even (TE-like) and odd (TM-like) modes under reflections through the mirror plane (which are purely TE/TM in the mirror plane itself). Because the slab here is based on the 2d hole crystal, which had a TE gap, here there is a 26% 'bandgap' in the even modes: a range of frequencies in which there are no guided modes. It is not a complete photonic bandgap, not only because of the odd modes, but also because there are radiating (light cone) modes at every ω. The presence of these radiating modes means that if all in-plane translational symmetry is broken by a localized change in the structure, say a waveguide bend or a resonant cavity, then vertical radiation losses are inevitable; there are various strategies to minimize the losses to tolerable levels, however. On the other hand, if only one direction of translational symmetry is broken, as in a linear-defect waveguide, ideally lossless guiding can be maintained.

Photonic-crystal slabs have two new critical parameters that influence the existence of a gap. First, it must have vertical mirror symmetry in order that the gaps in the even and odd modes be treatable separately – such symmetry is broken by the presence of a substrate that is not the mirror image of the superstrate, but in practice the symmetry breaking can be weak if the index contrast is sufficiently high (so that the modes are strongly confined in the slab). Second, the height of the slab must not be too small (or the modes will be weakly confined) or too large (or higher-order modes will fill the gap); the optimum height is around half a wavelength λ/n_{eff} (relative to an average/effective index n_{eff} that depends on the polarization). In **Figure 5**, a height of $0.5a$ is used, which is near the optimum (with holes of radius $0.3a$ and $\varepsilon = 12$ as in the previous section).

Three-Dimensional Photonic Crystals

Photonic-crystal slabs are one way of realizing 2d photonic-crystal effects in three dimensions; an example of another way, lifting the sacrifices imposed by the light cone, is depicted in **Figure 6**. This is a 3d-periodic crystal, formed by an alternating hole-slab/rod-slab sequence in an ABCABC stacking of bilayers – equivalently, it is an fcc (face-centered cubic) lattice of air cylinders in dielectric, stacked and oriented in the 111 direction, where each overlapping layer of cylinders forms a rod/hole bilayer simultaneously. Its band diagram is shown in **Figure 6** along the boundaries of its irreducible Brillouin zone (from a truncated octahedron, inset), and this structure has a >21% complete gap ($\Delta\omega$ as a fraction of mid-gap frequency) for $\varepsilon = 12/1$, forbidding light propagation for all wavevectors (directions) and all polarizations. Not only can this crystal confine light perfectly in 3d, but because its layers resemble 2d rod/hole crystals, it turns out that the confined modes created by defects in these layers strongly resemble the TM/TE states created by corresponding defects in two dimensions. One can therefore use this crystal to directly transfer designs from two to three dimensions while retaining omnidirectional confinement. Its fabrication, of

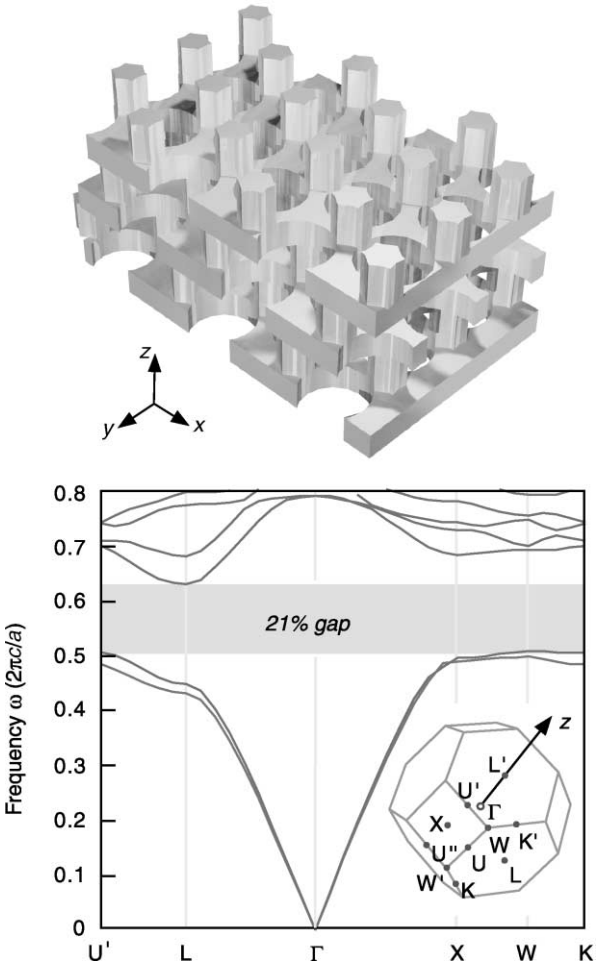


Figure 6 Band diagram (bottom) for 3d-periodic photonic crystal (top) consisting of an alternating stack of rod and hole 2d-periodic slabs (similar to **Figure 4**), with the corners of the irreducible Brillouin zone labeled in the inset. This structure exhibits a $\Delta\omega/\omega_{\text{midgap}} = 21\%$ omnidirectional bandgap.

course, is more complex than that of photonic-crystal slabs (with a minimum ε contrast of 4/1), but this and other 3d photonic crystal structures have been constructed even at micron (infrared) lengthscales, as described below.

There are three general dielectric topologies that have been identified to support complete 3d gaps for $\varepsilon = 12/1$ (e.g., Si:air) contrast: diamond-like arrangements of high dielectric 'atoms' surrounded by low dielectric, which can lead to $>20\%$ gaps between the 2nd and 3rd bands; fcc 'inverse opal' lattices of nearly-touching low dielectric spheres (or similar) surrounded by high dielectric, giving gaps around 10% between the 9th and 10th bands; and cubic 'scaffold' lattices of rods along the cube edges, giving $\sim$7% gaps between the 2nd and 3rd bands. It is notable that the first two topologies correspond to fcc lattices, which have the most nearly spherical Brillouin zones in accordance with the rules of thumb given above. Many variations on these topologies continue to be proposed – for example, the structure of **Figure 6** is diamond/graphite-like – mainly in conjunction with different fabrication strategies, such as the following three successful approaches. First, layer-by-layer fabrication, in which individual crystal layers (typically of constant cross-section) are deposited one-by-one and etched with a 2d pattern via standard lithographic methods (giving fine control over placement of defects, etc.); **Figure 6** can be constructed in this fashion (as well as other diamond-like structures with large gaps). Second, colloidal self-assembly, in which small dielectric spheres in a fluid automatically arrange themselves into close-packed (fcc) crystals by surface forces – these crystals can be back-filled with a high-index material, out of which the original spheres are dissolved in order to form inverse–opal crystals with a complete gap. Third, holographic lithography, in which a variety of 3d crystals can be formed by an interference pattern of four laser beams to harden a light-sensitive resin (which is then back-filled and dissolved, as with colloids, to achieve the requisite index contrast). The second and third techniques are notable for their ability to construct large-scale 3d crystals (thousands of periods) in a short time.

See also

Photonic Crystals: Atomic Physics; Photonic Crystal Lasers, Cavities and Waveguide.

Further Reading

Ashcroft NW and Mermin ND (1976) *Solid State Physics*. Philadelphia: Holt Saunders.

Cohen-Tannoudji C, Din B and Laloë F (1977) *Quantum Mechanics*. Paris: Hermann.

Joannopoulos JD, Meade RD and Winn JN (1995) *Photonic Crystals: Molding the Flow of Light*. Princeton: Princeton University Press.

Johnson SG and Joannopoulos JD (2002) *Photonic Crystals: The Road from Theory to Practice*. Boston: Kluwer.

Johnson SG, Ibanescu M, Skorobogatiy M, Weisberg O, Joannopoulos JD and Fink Y (2002) Perturbation theory for Maxwell's equations with shifting material boundaries. *Physical Review E* 65: 066611.

Sakoda K (2001) *Optical Properties of Photonic Crystals*. Berlin: Springer.

Sipe JE (2000) Vector $k \cdot p$ approach for photonic band structures. *Physical Review E* 62: 5672–5677.

Microwave Photonic Crystals

D F Sievenpiper, HRL Laboratories LLC, Malibu, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Photonic Crystals

Photonic crystals are materials that have a periodically varying dielectric constant that results in a bandgap, or range of frequencies where electromagnetic waves are unable to propagate. A simple one-dimensional example is the Bragg mirror, or quarter-wave stack, in which periodic layers of alternating dielectrics reflect photons within a frequency band determined by the thickness and refractive index of the layers. Photonic crystals are the extension of this concept into two or three dimensions. They are constructed as a lattice of two different dielectrics, and examples include a triangular lattice of circular rods, or a diamond lattice of spheres. Certain geometries will produce a bandgap, within which propagation is impossible in any direction. These materials are the photonic analog of semiconductors, in that their periodic dielectric function provides a bandgap for photons in much the same way that the periodic atomic potential of a crystal lattice can provide a bandgap for electrons.

While most photonic crystal research has focused on optical wavelength structures, many of the first experimental studies were performed on scale models at microwave frequencies, where fabrication and measurements were simpler. Operating at microwave frequencies also allowed a new class of materials to emerge, based on metals and lumped circuit elements. The nearly infinite dielectric contrast provided by the use of metals, and the ability to make complex subwavelength scale structures at radio frequencies, allows photonic crystals with ultra-wide bandgaps to be produced. Furthermore, the use of lumped circuit elements allows the electromagnetic properties to be easily tuned, resulting in unique microwave structures.

Wire Mesh Photonic Crystals

At optical frequencies, conductive losses preclude the use of metals for many applications. However, at microwave frequencies, metals are nearly ideal materials, since they have low loss and are easily formed into complicated geometries. The simplest metallic photonic crystal is a three-dimensional wire mesh, shown in Figure 1. This example resembles a diamond crystal lattice, in which metal wires have been substituted for atomic bonds. It can be constructed using zig-zag-shaped metal strips that fit together to form a continuous conductive network.

The electromagnetic properties of this wire mesh can be understood by considering a traditional microwave analog: an array of metal waveguides. Open tunnels permeate the mesh in various directions, and it behaves as an interpenetrating fabric of waveguides, whose walls are defined by the wires. Shorter wavelengths, that fit between the wires, can propagate through the crystal, while longer waves see the mesh as a continuous metal surface, and are reflected. As for a conventional waveguide, the cutoff frequency occurs approximately where one half wavelength fits between the rows of wires. Similarly, for a diamond lattice of wires, the cutoff frequency is

$$\omega_{\text{cutoff}} = \frac{\pi c}{na} \qquad [1]$$

where n is the refractive index of the material between the wires, and a is the length of the unit cube of the diamond lattice. Because the cutoff frequency is primarily determined by the diameter of the open tunnels, and it is largely independent of the detailed arrangement of the wires, other mesh geometries can

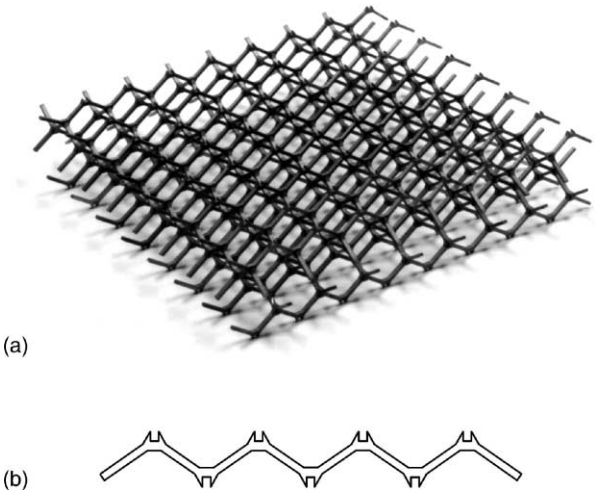


Figure 1 A wire mesh photonic crystal with diamond geometry. (a) The mesh forms continuous conductive structure that is permeated with open tunnels. (b) A diagram of one of the interlocking metal parts that fit together to form the crystal. (Reproduced with permission from Sievenpiper D, Sickmiller M and Yablonovitch E (1996) 3D wire mesh photonic crystals. *Physics Review Letters* 76: 2480. Copyright (1996) by the American Physical Society.)

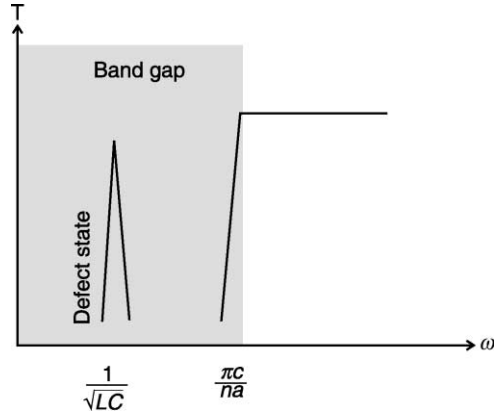


Figure 2 The transmission magnitude through a conductive wire mesh photonic crystal as a function of frequency, ω. Shorter wavelengths can propagate through the open tunnels, while longer wavelengths see the mesh as a continuous block of metal. The cutoff frequency is determined largely by the lattice constant, a, and the refractive index, n, of the material between the wires. A cut wire forms a defect state within the band gap.

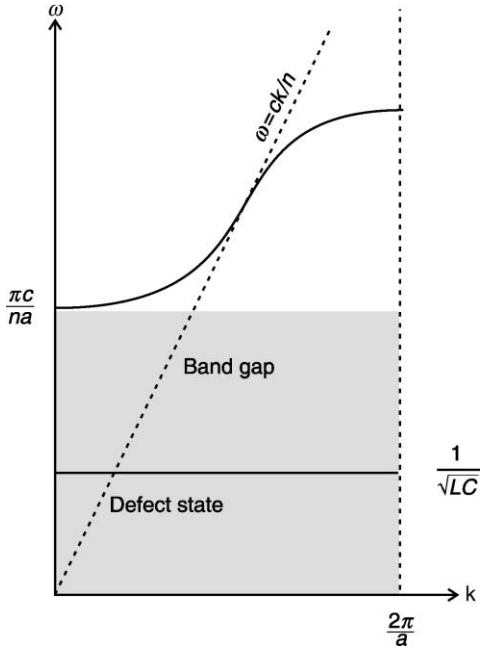


Figure 3 The dispersion diagram for a wire mesh photonic crystal, plotted as frequency, ω, versus wave vector k. The bandgap extends to zero frequency. The lower bandedge of the allowed band is determined largely by the lattice constant. Like other photonic crystals, the wire mesh also supports higher frequency bands, which are not shown. A point defect such as a cut wire creates a mode at a single frequency, which contains all wave vectors.

be expected to give similar results. However, the cutoff frequency can vary by a constant of order unity, depending on the lattice and the unit cell chosen.

The transmission magnitude, as a function of frequency for such a mesh, is depicted in Figure 2. Experimentally, it is found that a typical wire mesh structure attenuates roughly 6 dB per layer for frequencies below cutoff. However, because the exact transmission magnitude above cutoff depends on details of the measurement, such as impedance matching at the crystal boundary, the vertical scale shown here is arbitrary, and the intent is simply to give the qualitative behavior of these and similar crystal structures. Waves just below the cutoff frequency can penetrate a crystal with finite thickness, as depicted by the finite slope of the transmission edge.

We can better understand the behavior of waves within the crystal by examining the dispersion diagram, shown in Figure 3, in which we plot the allowed frequencies as a function of wave vector $k = 2\pi/\lambda$, the spatial frequency of waves inside the material. A band of allowed frequencies is bounded by upper and lower band edges, corresponding to high and low cutoff frequencies, beyond which propagation is forbidden. The reflective nature of the crystal below cutoff is shown as a bandgap that extends down to zero frequency. The slope of the dispersion curve corresponds to the group velocity of waves within the allowed band, and states at the bandedges represent standing waves. The upper edge of this band occurs where the wavelength matches the periodicity of the crystal, and for a diamond mesh, it is approximately

$$\omega_{\text{lattice}} = \frac{2\pi c}{na} \qquad [2]$$

Like traditional dielectric photonic crystals, these wire mesh structures also have higher order bands and gaps that occur at higher frequencies. The nature of the higher band structure depends on the detailed geometry of the mesh to a much greater extent than the lowest order gap, which is our primary interest for radio frequency materials.

Defect States

As in semiconductor crystals, we can alter the behavior of photonic crystals by adding defects to create localized states within the bandgap. For the wire mesh photonic crystal, a defect may be a cut wire, or a lumped circuit element inserted into the mesh. For example, replacing one of the wires with a capacitor creates a parallel resonant circuit, which acts as a small electromagnetic cavity within the mesh. Its resonance frequency is given by

$$\omega_{\text{defect}} = \frac{1}{\sqrt{LC}} \qquad [3]$$

where C is the capacitance of the defect, and L is the inductance of the surrounding wires. The inductance depends logarithmically on the wire diameter, so it is largely fixed for a given mesh lattice. However, the capacitance can be easily changed, since it is a simple lumped circuit element. Thus, by adding capacitive defects we can create tunable local states within the bandgap. The presence of a defect is seen as a narrow transmission band within the bandgap, shown in Figure 2. Because a point defect contains all wave vectors, it is shown as a horizontal line in the dispersion diagram in Figure 3.

If two defects are located in close proximity, their modes overlap, and they interact with each other. For example, two neighboring defects with identical capacitance will exhibit a split pair of resonances, corresponding to a lower frequency symmetric mode, and a higher frequency anti-symmetric mode. The modes are highly localized at the defects, and this interaction is limited to defects that are within one or two lattice constants of each other. If one member of a pair of nearby defects is tuned with respect to the other, the frequencies the two modes exhibit are an avoided crossing behavior, as is typical in many systems of coupled oscillators. The resonance frequencies for more complex arrangements of defects can be calculated by assigning an interaction energy to each neighboring pair, and solving the resulting Hamiltonian matrix to find the energy levels of the complete system. In this way, the interaction among defects in small clusters is analogous to atoms in simple molecules, and the resulting energy level structure is similar.

As more defects are added periodically throughout the mesh, they form a superlattice, and their states merge into an additional band within the original bandgap. The upper frequency of this band depends on the LC resonance frequency of the defects. The lower frequency depends on the density of the remaining wires, since waves at the lower bandedge must be able to fit through the open regions in the mesh. In the extreme case, where every wire is cut and replaced by a capacitor, the mesh is no longer a continuous conductive network. These structures have an allowed band that extends to zero frequency, and they will be explored in greater detail in the next section. Not every wire must be cut to reach this limit, and the conditions for this to occur are the subject of percolation theory.

Capacitive Mesh Photonic Crystals

In a capacitive mesh photonic crystal, small isolated metal islands are distributed on a lattice, and they are capacitively coupled to their neighbors. This represents the extreme case of the wire mesh photonic crystal, where every wire has been cut, and replaced by a capacitor. An example of such a structure, built as a stack of printed circuit boards, is shown in Figure 4. A diamond lattice can be approximated using planar circuit boards, with printed tetrapodal structures consisting of two orthogonal metal strips connected by a metal plated via. Each metallic island couples to its four nearest neighbors through their overlapping capacitor plates.

The capacitive mesh photonic crystal is transmissive at low frequencies because it is not traversed by conducting paths, like the continuous wire mesh version. The crystal supports a lower band that extends from zero frequency, up to an LC resonance frequency:

$$\omega_{\text{lower}} = \frac{1}{\sqrt{LC}} \qquad [4]$$

which marks the lower edge of the bandgap. This is determined by the capacitance between the metal islands, and the inductance associated with each island. At higher frequencies, the capacitors appear as shorts, and the crystal behaves similarly to the conductive wire mesh structure described previously, with an upper edge given by

$$\omega_{\text{upper}} = \frac{\pi c}{na} \qquad [5]$$

At this frequency, the waves can again fit through the open areas between the metallic islands. The transmission magnitude, as a function of frequency for the capacitive mesh photonic crystal, is shown in Figure 5, and the additional band is illustrated in the dispersion diagram in Figure 6. Like the conductive wire mesh, the capacitive mesh also has a rich upper band structure, that is outside the scope of this article.

With these capacitive mesh structures, it is possible to create materials with finite, but arbitrarily large bandgaps, because the upper and lower edges of the bandgap are controlled independently by the lattice constant and the capacitance, respectively. In conventional dielectric photonic crystals, the width of the bandgap depends on the dielectric contrast between the constituent materials. Waves at the upper edge primarily reside in the lower dielectric material, and waves at the lower edge inhabit the higher dielectric material. This reasoning can also be applied to capacitive mesh structures. The lower dielectric is the substrate material that fills the open spaces between the metal islands, and modes at the upper edge of the gap primarily fill these open spaces. However, for modes at the lower edge of the gap, the electric field is primarily concentrated within the capacitors. The metal regions and the capacitors can

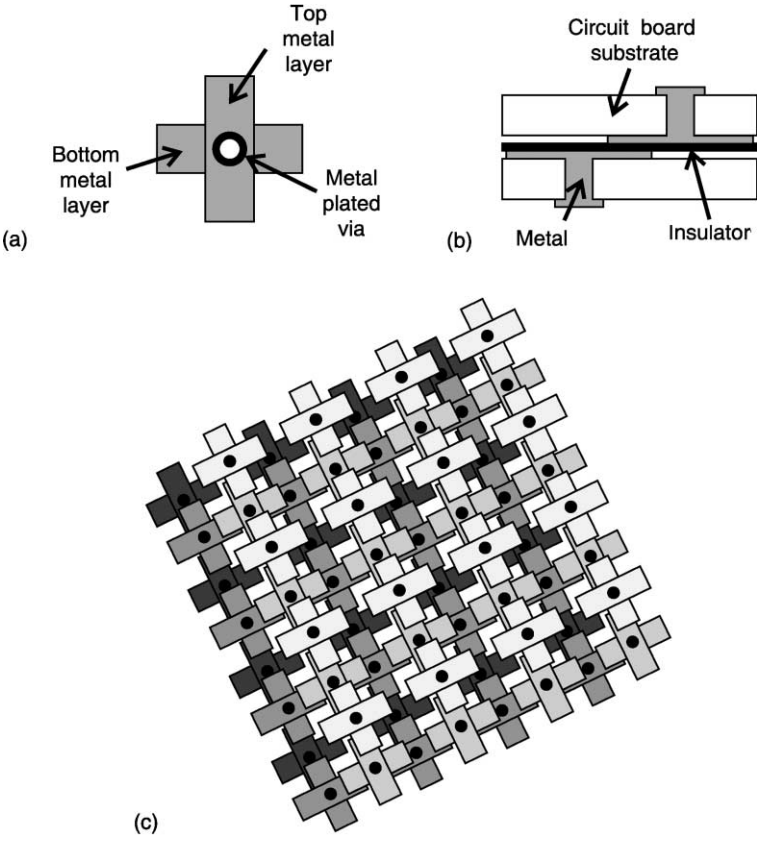


Figure 4 A capacitive mesh photonic crystal, built as a diamond lattice of capacitively coupled metallic islands. It can be built by stacking printed circuit boards. (a) One 'atom' of the crystal, which is made of two small orthogonal metal strips connected by a metal-plated via. (b) A side view, showing how the metal strips overlap to form a capacitor. (c) Several periods of the completed crystal structure, shown without the substrate. (Reproduced with permission from Sievenpiper D, Yablonovitch E, Winn J, et al. (1998) 3D metallo-dielectric photonic crystals with strong capacitive coupling between metallic islands. Physics Review Lett. 80: 2829. Copyright (1998) by the American Physical Society.)

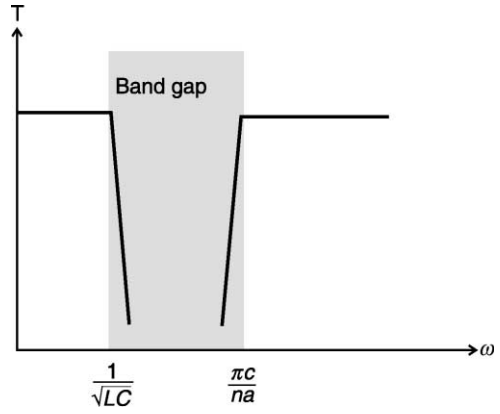


Figure 5 The transmission magnitude through a capacitive mesh photonic crystal, as a function of frequency. The crystal is transparent at low frequencies, because the conducting paths are broken by capacitors. At higher frequencies, the capacitors appear as shorts, and crystal behaves like a solid wire mesh.

be considered as an effective medium, or artificial dielectric whose properties depend on the density, arrangement, and values of the capacitors. By using large capacitors, the dielectric contrast can be made arbitrarily large, and the bandgap arbitrarily wide.

Within the bandgap, these materials can support surface waves, which are modes that are bound to the interface between the crystal and free space. Surface waves propagate along the crystal surface, and they have their own band structure. In the following sections, we examine these waves, and we will introduce periodically textured surfaces that suppress them, much like three-dimensional photonic crystals suppress internal plane waves.

Surface Waves

Electromagnetic waves that are bound to the interface between two materials are known as surface waves. They exist on ordinary metals, and at optical frequencies they are often called surface plasmons. They also exist on photonic crystals within the

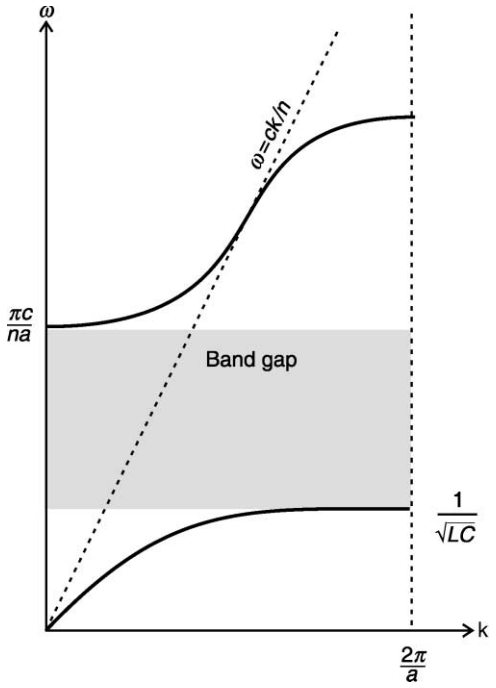


Figure 6 The dispersion diagram of the capacitive mesh photonic crystal. The two edges of the bandgap can be adjusted independently by varying the lattice constant and the capacitance.

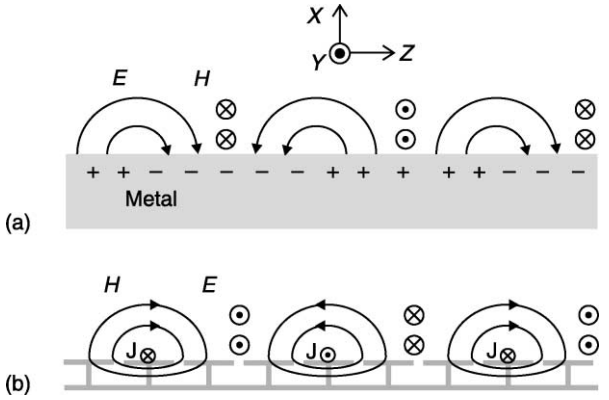


Figure 7 Surface waves can be classified into two polarizations. (a) A TM wave, shown on a metal surface; (b) A TE wave, shown on a textured surface. TM waves consist of electric charges on the metal surface that oscillate along the direction of propagation. They have nonzero electric field components both perpendicular to the surface, and parallel to the direction of propagation. TE waves are associated with transverse currents in the surface. They are similar to TM waves, but with the electric and magnetic fields exchanged.

bandgap region, as well as on other impenetrable materials. They cannot propagate into the crystal because of Bragg reflection, yet they are prevented from escaping into free space by total internal reflection. They typically interact strongly with the top layer of the crystal, and they can often be manipulated or suppressed by adjusting the geometry of that layer. At radio and microwave frequencies, these waves are usually called surface currents, because the associated fields are only loosely bound to the surface, and extend a great distance into free space. Their presence can be seen in antenna radiation patterns and other phenomena, and by suppressing them we can make a variety of unique radio frequency devices. However, we must first understand the nature of these waves.

To understand the nature of these surface currents, we begin by examining surface waves on metals, where we assume a wave confined to a dielectric interface, and then insert the complex dielectric constant of a metal. Consider a surface in the YZ plane, where $X < 0$ is filled with dielectric ε, and $X > 0$ is filled with vacuum. A wave propagates in the Z direction with propagation constant k. For transverse magnetic (TM) polarized waves, which are illustrated in Figure 7, the transverse electric field $E_y = 0$. The electric field has the form

$$E_1 = (\hat{x}E_{1x} + \hat{z}E_{1z})e^{i\omega x - ikz - \alpha x} \qquad [6]$$

for $X > 0$ and

$$E_2 = (\hat{x}E_{2x} + \hat{z}E_{2z})e^{i\omega x - ikz + \gamma x} \qquad [7]$$

for $X < 0$, where α and γ are the attenuation constants into free space and dielectric, respectively. The fields of such a wave are shown in Figure 7. We can solve the wave equation for the electric field:

$$\nabla \times \nabla \times \vec{E} = \varepsilon_r \frac{\omega^2}{c^2} \frac{\partial^2 \vec{E}}{\partial t^2} \qquad [8]$$

by matching E_x and $\varepsilon_r E_z$, at the interface to yield

$$k = \sqrt{\frac{\varepsilon}{1+\varepsilon}} \frac{\omega}{c} \qquad [9]$$

$$\alpha = \sqrt{\frac{-1}{1+\varepsilon}} \frac{\omega}{c} \qquad [10]$$

$$\gamma = \sqrt{\frac{-\varepsilon^2}{1+\varepsilon}} \frac{\omega}{c} \qquad [11]$$

A solution for transverse electric (TE) waves, where $H_y = 0$, can be found by applying the principle of duality, by exchanging the electric and magnetic fields, and substituting μ for ε.

For surface waves on a metal surface, we insert the dielectric constant of a metal, which is

$$\varepsilon = 1 - \frac{i\sigma}{\omega\varepsilon_0} \qquad [12]$$

The conductivity, λ, is

$$\sigma = \frac{nq^2\tau/m}{1 + i\omega\tau} \qquad [13]$$

where τ is the mean electron collision time, q is the electron charge, m and n are the effective mass and the density, respectively, of the conduction electrons. For frequencies much lower than $1/\tau$, which includes microwave and radio frequencies, we can substitute eqns [12] and [13] into eqn [9] to find

$$k \approx \frac{\omega}{c} \qquad [14]$$

Thus, surface waves on metals propagate at nearly the speed of light in vacuum, and they travel for many wavelengths with little attenuation. We can also solve for the decay constants into free space using eqn [10]:

$$\alpha = (1 - i)\frac{\omega}{c}\sqrt{\frac{\omega\varepsilon_0}{2\sigma}} \qquad [15]$$

and into the metal using eqn [11]:

$$\gamma \approx (1 + i)\sqrt{\frac{\omega\mu_0\sigma}{2}} = \frac{(1 + i)}{\delta} \qquad [16]$$

where δ is recognized as the skin depth. Thus, we have derived the skin depth of a metal from the penetration depth of surface waves.

From the skin depth, we can derive the impedance of a metal sheet. Assuming that E_0 is the electric field at the surface, the surface current is given by

$$J_z(x) = \sigma E_z(x) = \sigma E_0 e^{-x(1+i)/\delta} \qquad [17]$$

The magnetic field at the surface is found by integrating along a path that surrounds the surface current:

$$H_0 = \int_0^{-\infty} J_z(x)\mathrm{d}x = \frac{\sigma\delta}{1 + i}E_0 \qquad [18]$$

The surface impedance of the metal is therefore

$$Z_s = \frac{E_z}{H_y} = \frac{1 + i}{\sigma\delta} \qquad [19]$$

We can see that the resistance of a metal surface is accompanied by an equal amount of inductance, because the real and imaginary parts of the surface impedance are equal, and have the same sign.

The part of the wave that extends into vacuum is very loosely bound to the metal surface, but the currents in the metal occur in a well-defined skin depth. For example, copper has a surface impedance of $0.03(1 + i)$ ohms/square at 10 GHz. Using eqns [15] and [16], a surface wave at that frequency extends about 70 meters, or 2300 wavelengths into free space, but less than 1 micron into the metal. For this reason, surface waves are often described simply as surface currents at microwave and radio frequencies, although they are the same phenomenon as the more tightly bound surface plasmons that occur at optical frequencies. Nonetheless, their presence can be seen in a variety of phenomena at RF frequencies, such as antenna radiation patterns.

The above discussion can be generalized to a simple impedance surface, and it can be shown that TM waves are supported on inductive surfaces having an impedance of

$$Z_s(\mathrm{TM}) = \frac{i\alpha}{\omega\varepsilon} \qquad [20]$$

while TE waves are supported on capacitive surfaces, having an impedance of

$$Z_s(\mathrm{TE}) = \frac{-i\omega\mu}{\alpha} \qquad [21]$$

Thus, while ordinary metals, which are inductive, support TM waves, they do not support TE waves, which require a capacitive surface.

Surface waves on metals can be suppressed over a finite frequency band by applying various kinds of periotic textures. For example, at optical frequencies, a lattice of small bumps on a metal surface can create a surface wave bandgap. Surface waves are scattered by the bumps, and form standing waves when one-half wavelength fits between the rows of bumps. Two modes occur, in which the wave nodes lie either on the bumps or between them. These two modes have slightly different frequencies, separated by a narrow bandgap.

At microwave frequencies, more elaborate structures can be built to control the surface properties. One example is a corrugated metal surface, shown in Figure 8. The corrugations can be considered as transmission lines running into the slab, and shorted at the bottom. They are typically one-quarter wavelength deep, so that the short circuit at the bottom is transformed into an open circuit at the top. Because of these resonant structures built into the surface, it has high impedance for waves that are polarized with the electric field transverse to the corrugations, but it has low impedance for waves that are polarized in the orthogonal direction, with the electric field parallel to the metal ridges. For waves propagating transverse to or parallel to the corrugations, the surface can be described as hard or soft, respectively, by analogy to the acoustic surfaces of the same names. Both of these boundary conditions are used in antennas to control scattering properties or

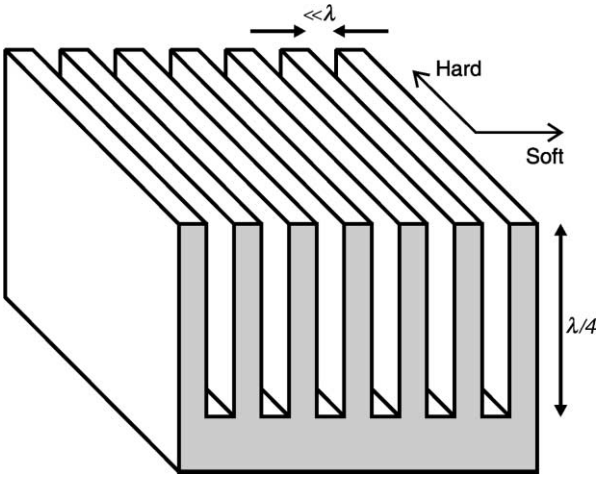


Figure 8 A corrugated metal surface containing quarter-wave resonant grooves presents a soft boundary condition to waves propagating transverse to the grooves, and a hard boundary condition to waves propagating along the grooves.

Table 1 Four basic boundary conditions in electromagnetics, and the surface impedance for each polarization

Boundary condition	Z_s(TM)	Z_s(TE)
Electric conductor	0	0
Soft surface	∞	0
Hard surface	0	∞
Magnetic conductor	∞	∞

field profiles. These two surfaces, along with the electric and magnetic conductors, describe four basic boundary conditions for electromagnetics, which are summarized in Table 1. Like the soft and hard surfaces, the magnetic conductor does not exist in nature, but it can be produced artificially over a finite frequency band, using a kind of two-dimensional photonic crystal known as a high-impedance surface.

High-Impedance Surfaces

Just as three-dimensional photonic crystals forbid internal plane waves, two-dimensional photonic crystals can be designed to suppress surface waves. Like the capacitive mesh described earlier, they can be constructed with lumped circuit elements. They can be understood as corrugated surfaces, in which the resonant corrugations have been folded up into resonant LC circuits, and distributed on a two-dimensional lattice. An example of such a structure is shown in Figure 9. It can be built using printed circuit boards, in which flat metal patches are connected to a common ground plane by metal plated vias.

When the plates are small compared to the wavelength, we may treat the surface as an effective

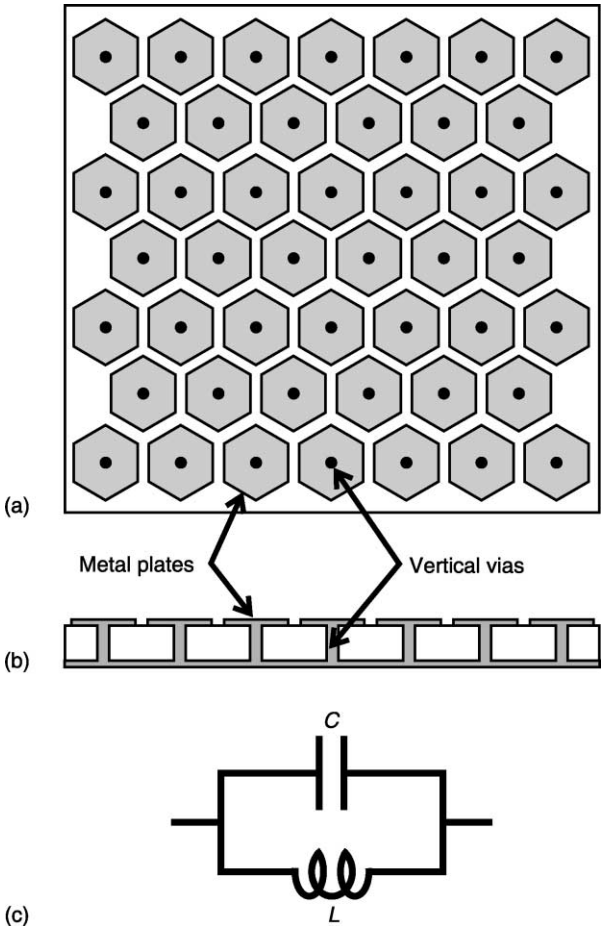


Figure 9 A high-impedance surface, constructed as a printed circuit board. (a) The top view shows an array of hexagonal metal patches; (b) the side view shows metal-plated vias that connect the patches to the ground plane; (c) the effective surface impedance can be modeled as a parallel resonant LC circuit.

medium, with a surface impedance defined by a parallel resonant LC circuit. The proximity of the neighboring metal plates provides capacitance, while the current paths, through the vias and the ground plane, provide inductance. The surface impedance is given by

$$Z_s = \frac{i\omega L}{1 - \omega^2 LC} \qquad [22]$$

and the resonance frequency is

$$\omega_0 = \frac{1}{\sqrt{LC}} \qquad [23]$$

The sheet capacitance, C, is equal to the value of the individual capacitors, multiplied by a geometrical factor that is related to their arrangement. It can be shown that for nonmagnetic materials, the sheet

inductance L is equal to

$$L = \mu_0 t \qquad [24]$$

where t is the total thickness of the high-impedance surface. The surface is inductive for frequencies below resonance, and capacitive for frequencies above resonance. Near the resonance frequency, the magnitude of the surface impedance is very high ($Z_s >> \eta$, where $\eta = 377\Omega$ is the impedance of free space).

For a normally incident wave, the reflection coefficient of the surface is

$$R = \frac{Z_s - \eta}{Z_s + \eta} \qquad [25]$$

The phase $\phi = \mathrm{Im}(Ln(R))$ of the reflection coefficient is plotted as a function of frequency in Figure 10. While a low impedance surface such as a conventional electric conductor reflects with a phase shift of π, a high-impedance surface or artificial magnetic conductor reflects with a phase shift of 0. This can be used to build low-profile antennas, because radiating elements can lie very near to a high-impedance ground plane without being shorted out. For example, a thin wire antenna can be placed within a small fraction of a wavelength of the high-impedance surface, yet retain high radiation efficiency. While image currents in a metal ground plane would normally oppose the antenna current, the currents in a high-impedance ground plane reinforce the antenna current because of the opposite reflection phase. This effect occurs when the reflection phase falls between $\pi/2$ and $-\pi/2$, and the corresponding frequency range defines the practical bandwidth of a low-profile antenna on such a surface.

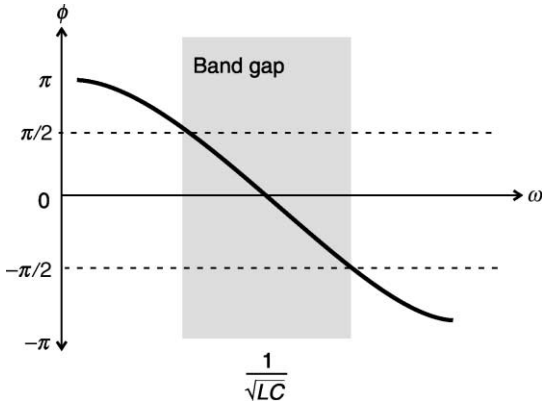


Figure 10 The reflection phase of a high-impedance surface crosses through 0 at the resonance frequency, where the surface appears as an artificial magnetic conductor. The reflection phase falls between $\pi/2$ and $-\pi/2$ for frequencies corresponding roughly to the edges of the surface wave bandgap.

Experimentally this bandwidth is found to be roughly equal to the surface wave bandgap, to be described later.

The reflection phase properties of these materials can also be used to build unique microwave reflectors. Since the phase depends on the resonance frequency, one can vary the local resonance frequency in order to tune the reflection phase as a function of position on the surface. For example, a linear phase gradient can be used for beam steering, or a parabolic phase function can be used for focusing. The resonance frequency can be adjusted by tuning the capacitance of the individual unit cells, using varactors or mechanical motion, to build actively controlled microwave reflectors.

The other important property of these textured ground planes is their surface wave bandgap, which can be used to control the radiation pattern of antennas near the surface. We examine the surface wave properties by considering a wave that decays into free space, as in eqn [6]. From Maxwell's equations, we can obtain

$$k^2 = \mu_0 \varepsilon_0 \omega^2 + \alpha^2 \qquad [26]$$

which can be combined with eqs [20] and [21] to obtain the dispersion relations for TM and TE waves on a general impedance surface, respectively. For TM waves, which occur where $\omega < \omega_0$, we have

$$k(\mathrm{TM}) = \frac{\omega}{c}\sqrt{1 - \frac{Z_s^2}{\eta^2}} \qquad [27]$$

and for TE waves, which occur where $\omega > \omega_0$ we have

$$k(\mathrm{TE}) = \frac{\omega}{c}\sqrt{1 - \frac{\eta^2}{Z_s^2}} \qquad [28]$$

These functions describe, to first order, the dispersion of waves on a high-impedance surface. However, they do not predict the bandgap, which can only be obtained by including the effects of radiation damping.

Consider an oscillating sheet of current that sees the impedance of the textured surface on one side, and the impedance of free space on the other side. At very low, or very high frequencies, the current is shunted through the surface. Near the resonance frequency, where the surface impedance is high, power is delivered to free space through radiation damping. The half power bandwidth of this radiation is defined by the frequencies where the magnitude of the surface impedance is equal to the impedance of

free space:

$$\left|\frac{i\omega L}{1-\omega^2 LC}\right| = \eta \qquad [29]$$

We solve for λ to yield the frequencies of the two bandedges:

$$\omega^2 = \frac{1}{LC} + \frac{1}{2\eta^2 C^2} \pm \frac{1}{\eta C}\sqrt{\frac{1}{LC} + \frac{1}{4\eta^2 C^2}} \qquad [30]$$

The terms in $1/\eta^2 C^2$ are typically small compared to the terms in $1/LC$, so we neglect them to obtain

$$\omega \approx \omega_0 \sqrt{1 \pm \frac{Z_0}{\eta}} \qquad [31]$$

where

$$Z_0 = \sqrt{\frac{L}{C}} \qquad [32]$$

is the characteristic impedance of the surface. For structures that are thin compared to the wavelength, the characteristic impedance is small compared to η. We can expand the square root in eqn [31], to approximate the fractional bandwidth of the radiation as

$$\frac{\Delta\omega}{\omega_0} \approx \frac{Z_0}{\eta} = \frac{\sqrt{L/C}}{\sqrt{\mu_0/\varepsilon_0}} \qquad [33]$$

Using eqn [23], this can be rearranged to obtain

$$\frac{\Delta\omega}{\omega_0} \approx t\frac{2\pi}{\lambda_0} \qquad [34]$$

where λ_0 is the free space wavelength at the resonance frequency. This defines a surface wave bandgap that is centered at the resonance frequency, between a lower-frequency TM band and a higher-frequency TE band. From eqn [34], we see that the bandwidth of a high-impedance surface, or of an antenna on that surface, is limited by its thickness with respect to its free space wavelength at resonance. The term $\lambda_0/2\pi$ is recognized as the radian length, described by Wheeler, who showed that the fractional bandwidth limit of a small resonant antenna is given by its volume with respect to a notional sphere of this radius. Equation [34] represents the equivalent bandwidth limit for a two-dimensional resonant surface.

The surface wave bandgap can be measured experimentally with a pair of small probes placed near the surface. A surface wave probe can be constructed from the end of a coaxial cable with a short section of the inner conductor exposed. Fields from the probe excite surface waves when they are placed near the surface, and the transmission magnitude between the probes can be used to analyze the surface wave band structure. For a high impedance surface, the bandgap can be seen as a range of frequencies with a significant reduction in surface wave transmission magnitude, shown in Figure 11. Within the bandgap, waves are not bound to the surface, and currents on the surface readily radiate energy into free space. The upper edge of the bandgap

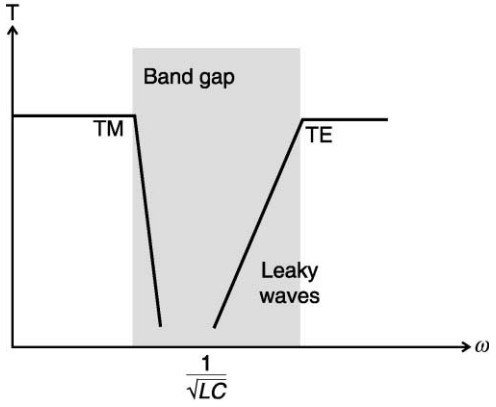


Figure 11 The surface wave transmission magnitude as a function of frequency for a high-impedance surface. TM waves are supported at low frequencies, and TE waves are supported at high frequencies. Leaky TE waves, which are damped by radiation, can occur within the bandgap.

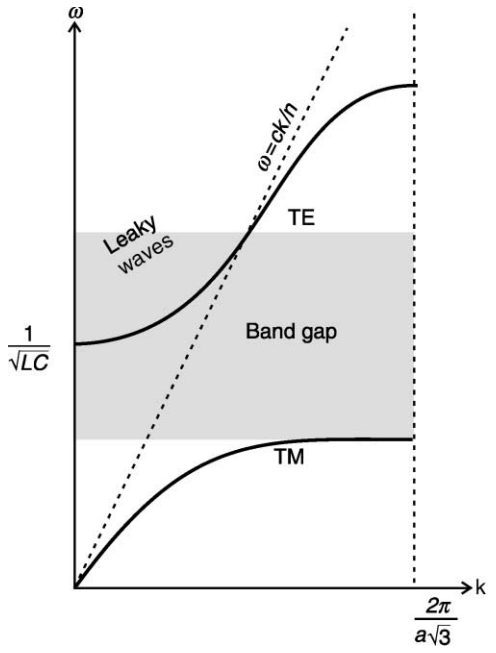
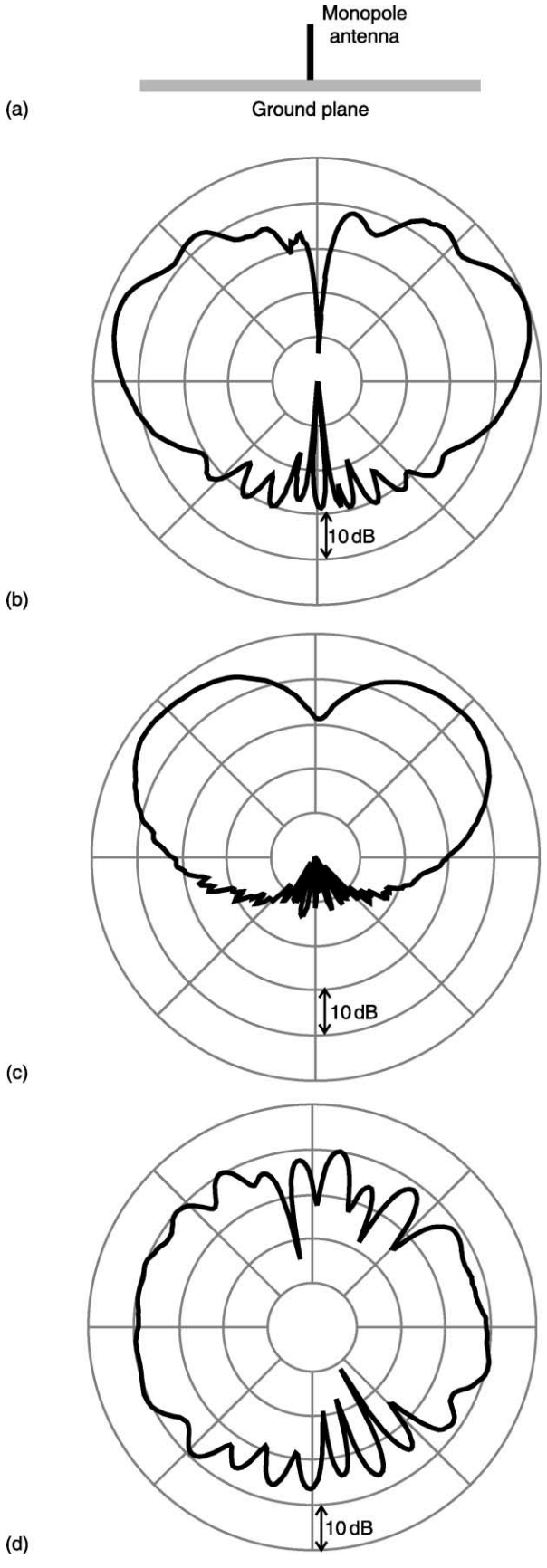


Figure 12 The dispersion diagram for surface waves on a high-impedance surface. The bandgap spans from the upper edge of the TM band, to the point where the TE band crosses the light line. Within the bandgap, electromagnetic waves are not bound to the surface. Modes to the left of the light line represent leaky waves.



is typically less abrupt than the lower edge, due to the presence of leaky TE waves.

The dispersion diagram for surface waves on a high-impedance surface is illustrated in Figure 12. The TM band begins at zero frequency, and bends over flat below the resonance frequency. The TE band begins at the resonance frequency, and slopes upward through the light line. The bandgap, which is the region where bound surface waves are not supported, extends from the upper edge of the TM band to the point where the TE band crosses the light line. This boundary, where $\omega = ck/n$, represents waves in free space propagating parallel to the surface. TE modes, that lie to the right of the light line, are bound to the surface while those that lie left of the light line lose energy to radiation damping. They are not bound to the surface, and are described as leaky waves. The approximate dispersion functions of eqns [27] and [28] are similar to the curves shown in Figure 12, but because they neglect radiation damping, they do not describe these leaky modes. Furthermore, like other photonic crystals, high-impedance surfaces also have higher-order bands and gaps that are not explained by the effective medium model.

The suppression of propagating surface modes can be used for a variety of applications, such as antennas and other microwave structures. One example is to change the radiation pattern of an antenna by covering its ground plane with a textured surface. Figure 13 shows a simple monopole antenna, along with the radiation patterns produced by the monopole when it is placed on a flat or textured ground plane. These plots show the relative magnitude of radiation from the antenna as a function of angle. TM surface currents on the finite metal ground plane form standing waves, which scatter from the edges and corners of a finite ground plane. This causes ripples in the radiation pattern as radiation from the edges interferes with the direct radiation from the antenna. Scattering of the surface currents from the ground plane edges also allows the antenna to radiate power into the backward hemisphere, which would otherwise be shadowed by the ground plane. On the high-impedance ground plane, the suppression of surface

Figure 13 (a) A monopole antenna on a ground plane, used for testing the effect of a textured surface on the radiation pattern. (b) The E-plane radiation pattern of the monopole on a flat metal ground plane shows ripples due to standing surface waves. (c) On a high-impedance ground plane, within the bandgap, surface waves are forbidden, and the radiation pattern is smooth. (d) Outside the bandgap, in the TM band, there is a high density of surface modes.

currents within the bandgap results in a smooth radiation pattern in the forward hemisphere, and very little power in the backward hemisphere. For frequencies outside the bandgap, such as within the TM band, the high density of TM modes can be seen as fine ripples in the pattern, and significant power in the backward hemisphere.

The monopole is useful for illustrating the suppression of TM surface modes, but many other antenna designs are possible. For example, the suppression of both TM and TE surface waves can be used to build patch antennas or aperture antennas with a high degree of symmetry between the E and H planes. Because these materials behave as artificial magnetic conductors, they can also be used to design a variety of thin antennas that would be impractical on conventional metal ground planes. Finally, by incorporating active elements into the surface, tunable or steerable antennas are possible.

List of Units and Nomenclature

a	lattice constant	meters
c	speed of light in vacuum	meters/second
C	capacitance	Farads
E	electric field	Volts/meter
H	magnetic field	Amps/meter
i	unit imaginary number	–
J	current density	Amps/meter^2
k	wave vector	1/meters
L	inductance	Henrys
m	effective electron mass	kilograms
n	refractive index	–
n	electron density	number/meter^3
q	electron charge	Coulombs
r	radius	meters
R	reflection coefficient	–
t	thickness	meters
T	transmission coefficient	–
Z	surface impedance	Ohms/square
α	decay constant into free space	1/meters
γ	decay constant into material	1/meters
δ	skin depth	meters
ε	electric permittivity	Farads/meter
η	impedance of free space	Ohms
λ	wavelength	meters
μ	magnetic permeability	Henrys/meter
π	diameter/circumference ratio	–
σ	electrical conductivity	1/Ohms
τ	mean electron collision time	seconds
ω	angular frequency	radians/second
∇	vector differential operator	–
∂	scalar differential operator	–
∞	infinity	–

See also

Babinet's Principle. **Diffractive Systems**: Diffractives in Animals; Microstructure Fibers; Omnidirectional Surfaces and Fibers. **Photonic Crystals:** Atomic Physics; Electromagnetic Theory; Nonlinear Optics in Photonic Crystal Fibers; Photonic Crystal Lasers, Cavities and Waveguide; Self-Assembled and Functionalized Photonic Crystals.

Further Reading

Ashcroft N and Mermin N (1976) *Solid State Physics*. Orlando, FL: Saunders College Publishing.

Brillouin L (1953) *Wave Propagation in Periodic Structures: Electric Filters and Crystal Lattices*, 2nd edn. New York: Dover Publications.

Brown E, Parker C and Yablonovitch E (1993) Radiation properties of a planar antenna on a photonic-crystal substrate. *Journal of the Optical Society of America B* 10: 404.

Collin R (1991) *Field Theory of Guided Waves*, 2nd edn. New York: IEEE Press.

Fan S, Villeneuve P and Joannopoulos J (1996) Large omnidirectional band gaps in metallodielectric photonic crystals. *Physics Review B* 54: 11245.

Feynman R, Leighton R and Sands M (1964) *The Feynman Lectures on Physics*, vol. 3. Reading, MA: Addison-Wesley.

Joannopoulos J, Meade R and Winn J (1995) *Photonic Crystals: Molding the Flow of Light*. Princeton, NJ: Princeton University Press.

Kildal P-S (1990) Artificially soft and hard surfaces in electromagnetics. *IEEE Transactions on Antennas and Propagation* 38: 1537.

Kurizki G and Haus J (eds) (1994) Photonic band structures. *Journal of Modern Optics* 41, special issue.

Lee S, Zarrillo G and Law C (1982) Simple formulas for transmission through periodic metal grids or plates. *IEEE Transactions on Antennas and Propagation* 30: 904.

Pendry J, Holden A, Stewart W and Youngs I (1996) Extremely low frequency plasmons in metallic mesostructures. *Physics Review Letters* 76: 4773.

Ramo S, Whinnery J and Van Duzer T (1984) *Fields and Waves in Communication Electronics*, 2nd edn. New York: John Wiley and Sons.

Sievenpiper D (1999) *High-Impedance Electromagnetic Surfaces*. Ph.D. Thesis, Los Angeles: University of California.

Sievenpiper D, Sickmiller M and Yablonovitch E (1996) 3D wire mesh photonic crystals. *Physics Review Letters* 76: 2480.

Sievenpiper D, Yablonovitch E, Winn J, *et al.* (1998) 3D metallo-dielectric photonic crystals with strong capacitive coupling between metallic islands. *Physics Review Letters* 80: 2829.

Sigalas M, Chan C, Ho K-M and Soukoulis C (1995) Metallic photonic band-gap materials. *Physics Review B* 52: 11744.

Smith D, Schultz S, Kroll N, et al. (1994) Experimental and theoretical results for a two-dimensional metal photonic band-gap cavity. *Applied Physics Letters* 65: 645.

Soukoulis C (ed.) (1996) Photonic band gap materials. *Proceedings of the NATO ASI on Photonic Band Gap Materials, Elounda, Crete, Greece, June 18–30, 1995*. The Netherlands: Kluwer Academic Publishers.

Wheeler H (1947) Fundamental limitations of small antennas. *Proceedings of the Institute of Radio Engineers* 35: 1479.

Yablonovitch E (1987) Inhibited spontaneous emission in solid-state physics and electronics. *Physics Review Letters* 58: 2059.

Nonlinear Optics in Photonic Crystal Fibers

J E Sharping, Cornell University, Ithaca, NY, USA

P Kumar, Northwestern University, Evanston, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Photonic crystal fibers (PCFs) are very similar to normal optical fibers in that they consist of a core surrounded by cladding, such that light is guided within the core of the fiber. The primary difference between PCF and standard optical fibers is that PCFs feature an air–silica cross-section, whereas standard optical fibers have an all-glass cross-section. An electron micrograph of a typical PCF is shown in Figure 1. The air holes extend along the axis of the fiber for its entire length and the core of the fiber is formed by a defect, or missing hole, in the periodic structure. The core is formed of solid glass, whose refractive index is that of pure silica (or whatever other glass is chosen), and the cladding is formed by the air–glass mixture, whose effective refractive index depends on the ratio of air-to-glass, also known as the air-fill fraction, that comprises the structure. The resulting effective-index of the cladding will be lower compared with that of the core and, as such, will provide the refractive index variation necessary to support total internal reflection at the core-cladding boundary, and guide light in a manner similar to that of standard optical fibers. The fiber design (i.e., size, shape, and the air-fill fraction) dictates solutions to Maxwell's equations for light propagating within the fiber. Valid solutions are referred to as 'modes' which propagate along the fiber in a known manner, and have a well-defined shape in the transverse direction (i.e., they have a well-defined transverse mode structure).

Nonlinear-optical effects in fibers result from the interaction of optical fields with the glass via the $\chi^{(3)}$, or Kerr nonlinearity. The phenomenon of nonlinear refractive index is a manifestation of a light–material interaction mediated by $\chi^{(3)}$. The magnitudes of the components of the third-order susceptibility tensor in glass, $\chi^{(3)}$, are generally quite small compared with the analogous second-order ($\chi^{(2)}$) terms for materials exhibiting such nonlinearities (e.g., lithium niobate, beta-barium borate (BBO), etc.). The relatively small $\chi^{(3)}$ nonlinearity in optical fibers makes them ideal for wavelength-division multiplexed optical communication where light propagation subject to a minimum of nonlinear effects is critical. Nonlinearity does, however, eventually become an issue in wavelength-division multiplexed systems as the launched optical power increases and as the channel spacing decreases. On the other hand, one can utilize nonlinear-optical effects in soliton communication systems and to build useful photonic devices. Despite the weak $\chi^{(3)}$

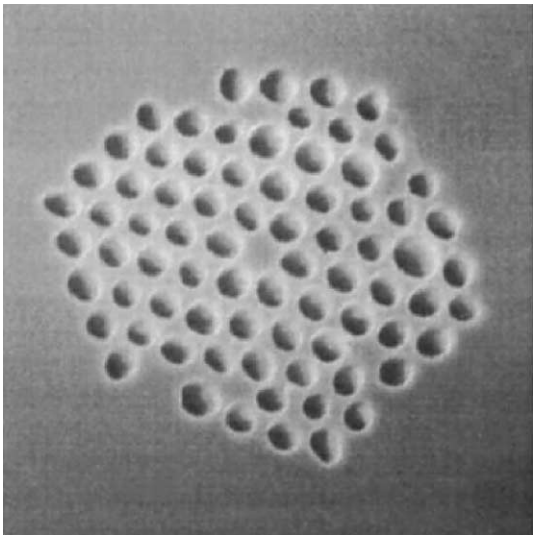


Figure 1 An electron micrograph showing the periodic microstructure of a typical PCF. The core is formed by the 'missing hole' in the center of the microstructure. (Reproduced with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optical Letters* 25: 25–27. ©2000 Optical Society of America, courtesy of OFS.)

nonlinearity, the net nonlinear-optical effect in fibers can be large due to the ability to tightly confine intense fields within the core of an optical fiber and maintain the interaction over a long distance as the guided fields propagate through the fiber.

The study of nonlinear-fiber optics has benefited from dramatic improvements in optical fiber and fiber-optic device fabrication. The importance of understanding nonlinear-fiber optics is driven by the need to develop fiber-integrated devices, and also by the need to understand and mitigate the problems that these nonlinearities cause in optical communication systems.

This section introduces the unique linear- and nonlinear-optical properties of PCFs in order to understand the reasons why nonlinear-optical effects are often enhanced in such fibers. These discussions pertain to PCFs which are 'highly nonlinear'. It is essential to clarify that 'highly nonlinear' in this context does not mean that the $\chi^{(3)}$ is any larger than that of standard telecommunication fibers, rather that the effect of this nonlinearity is enhanced due to the fiber's very small core.

PCF Properties

Photonic crystal fibers feature a variety of interesting properties. From the standpoint of nonlinear-fiber optics there are four very useful fundamental properties of PCFs:

- a mechanically robust optical fiber can be fabricated with an extremely small core (a few μm^2);
- a fiber can be made to guide in a single transverse mode over an extremely broad wavelength range (370 nm–1600 nm);
- there are new degrees of freedom that allow one to manipulate the fiber's group-velocity dispersion (GVD) properties; and
- many, but not all, PCFs are polarization maintaining as a result of form birefringence present in the core.

The fact that small-core PCFs can be fabricated is clear from Figure 1 by taking note of the fact that the center defect region which comprises the core is about 1.7 μm in diameter. Photonic crystal fibers with even smaller cores have been fabricated.

Transverse Mode Structure

A widely accepted model used to describe the transverse modal behavior of PCFs is called the effective-index model. The effective-index model can be used to understand why some PCFs are 'endlessly single mode', meaning that the fiber guides in a single transverse mode over an exceptionally wide wavelength range (370 nm–1600 nm). In the effective-index model, the refractive index of the core, $n_{co}(\lambda)$, is that of glass, and the refractive index of the cladding, $n_{eff}(\lambda)$, assumes a value in between that of glass and air. In the context of PCFs, one makes a modification to the standard expression describing single-mode behavior in step-index fibers:

$$V_{eff} = \frac{2\pi\Lambda}{\lambda}\sqrt{n_{co}(\lambda)^2 - n_{eff}(\lambda)^2} < V_{cutoff} \qquad [1]$$

where Λ is the spacing between air holes, λ is the wavelength of light, and V_{cutoff} is the cutoff condition for the PCF. A similar expression for the V parameter is commonly used to understand the modal behavior of standard fibers where the larger the V is, the more transverse modes are supported within the fiber. In standard fibers the cutoff condition below which only a single mode can propagate within the core of a fiber is given by $V_{cutoff} < 2.405$. In the case of PCFs, a numerical method should be used to determine V_{cutoff}. Mechanically robust PCFs can be fabricated where the dispersion in $n_{eff}(\lambda)$ (i.e., the variation of n_{eff} with λ) offsets dispersion in $n_{co}(\lambda)$ and compensates for the $2\pi\Lambda/\lambda$ coefficient in eqn [1]. Therefore, the light within the fiber propagates in a single, Gaussian-like mode because for all wavelengths $V_{eff} < V_{cutoff}$. A graph of the variation of V_{eff} with Λ/λ is shown in Figure 2, where d represents the size of an air hole.

Conceptually, the effective index model can be understood by noting that at short wavelengths the mode field is confined well within the all-silica core, but as λ increases the mode field extends further into the air–glass cladding and V_{eff} and $n_{eff}(\lambda)$ both decrease.

Dispersion in PCF

Other critical differences between PCFs and standard optical fibers lie in the dispersion properties. When light propagates through a fiber its behavior depends on the light's optical frequency:

$$E(t,z) = A(t,z)e^{i[\omega t - \beta(\omega)z]} \qquad [2]$$

Equation [2] describes the mode as it propagates through the fiber. It is decomposed into a slowly varying envelope, $A(t,z)$, and a rapidly varying exponential component where ω is the frequency of the mode, t is time, z is the position along the length of the fiber, and $\beta(\omega)$ is called the mode-propagation constant. The general term used in describing the

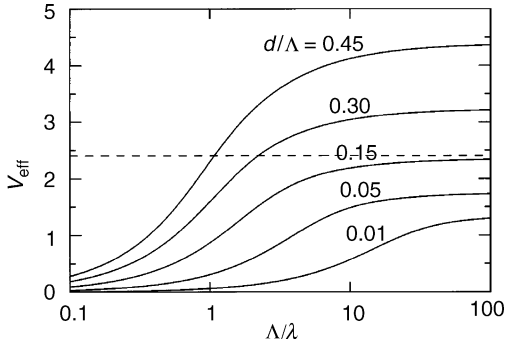


Figure 2 Variation of V_{eff} for different relative hole diameters d/Λ. The calculation assumes a fiber with an air–glass cross-section where the refractive index of air and glass was taken to be 1 and 1.45, respectively. The dashed line marks $V_{eff} = 2.405$, the cutoff value for a step-index fiber. (Reproduced with permission from Birks TA, Knight JC and Russell PStJ (1997) Endlessly single-mode photonic crystal fiber. *Optical Letters* 22: 961–963. ©1997 Optical Society of America.)

frequency dependence of β is chromatic dispersion, which includes contributions from the material as well as the waveguide. Other types of dispersion present in optical fibers include multi-modal (arising from multiple guided transverse modes) and polarization-mode dispersion.

One way to understand the chromatic dispersion of a mode propagating through an optical fiber is to study the Taylor series expansion of the mode-propagation constant, β, about the center frequency of the field, ω_0:

$$\beta(\omega) = \beta_0 + \beta_1(\omega - \omega_0) + \frac{1}{2}\beta_2(\omega - \omega_0)^2 + \cdots \qquad [3]$$

where $\beta_i = \mathrm{d}^i\beta/\mathrm{d}\omega^i$. The physical significance of the various β_i in eqn [3] are as follows: the phase-fronts of the electric field move at a speed given by $\omega/\beta_0 = v_p$, and the envelope, $A(t,z)$, moves at a group velocity given by $1/\beta_1$. A GVD term, which governs temporal spreading of the envelope, is given by β_2. Higher-order β terms are usually negligible for propagation of pulses of $\geq$1 ps duration in optical fibers and are lumped into the category of 'higher-order chromatic dispersion'.

The notation 'β_2', as defined above, is often used in the literature with dimensions of ps^2/km. However, another expression is frequently used because of its direct relationship to measured quantities. It is straightforward to measure the relative delay, T, between two pulses having different center wavelengths. Choosing a particular wavelength as a reference, one can then measure relative delay as a function of an injected pulse's center wavelength. The first derivative with respect to λ of the relative delay curve gives the GVD according to

$$D = \frac{\partial\left(\frac{1}{v_g}\right)}{\partial\lambda} = \frac{1}{L}\frac{\mathrm{d}T}{\mathrm{d}\lambda} = -\frac{2\pi c}{\lambda^2}\beta_2 \qquad [4]$$

where v_g is the group velocity, and L is the length of the fiber under test. The dimension commonly used for D is ps/(nm km).

Chromatic dispersion in single-mode optical fibers results from two different wavelength-dependent fiber parameters. The medium itself, glass in this case, has a wavelength-dependent refractive index. This 'material' contribution has the same magnitude regardless of the various parameters associated with the waveguide. A second contribution has to do with the design of the optical fiber. This 'waveguide' contribution to dispersion arises from the fact that the wavelength-dependent mode depends on the properties of the waveguide (i.e., the core size and refractive index contrast between the core and cladding). Empirical models can be used to describe the material, waveguide, and total GVD for standard communication fibers. Such a set of curves is given in Figure 3 where it can be seen that it is possible to have positive, negative, or zero values for D. For historical reasons, the regions where D is negative (β_2 is positive) exhibit 'normal GVD', while those where D is positive (β_2 is negative), exhibit 'anomalous GVD'. The wavelength corresponding to $D = 0$ is referred to as the zero-dispersion wavelength (λ_0), which for most silica glass fibers is about 1,300 nm.

In contrast with standard optical fibers, where the waveguide contribution to D is always less than zero, small-core PCFs can be fabricated where the waveguide contribution to GVD is positive and quite large. As such, in PCFs, λ_0 can be shifted to wavelengths shorter than the intrinsic dispersion zero of glass. Control over the GVD is essential for phase matching certain nonlinear-optical interactions involving light

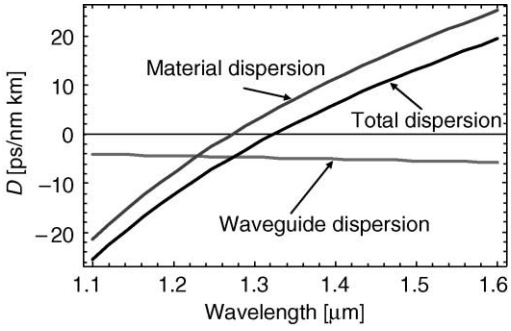


Figure 3 Plots of the theoretical dispersion coefficient, D, as a function of wavelength for a standard optical communication fiber.

of different colors co-propagating within a fiber. Indeed, several exciting applications of nonlinear optics in PCF require a fiber with a $\lambda_0 \sim 800$ nm. The GVD is also of great importance when working with pulsed light in PCF, because GVD results in temporal pulse broadening. It also governs pulse temporal walkoff effects, limiting the effective interaction length between pulses of different colors. This new flexibility to manipulate the GVD curve, by varying waveguide design parameters, is a key advantage associated with using PCFs for nonlinear optics.

Birefringence in PCF

The core of an optical fiber often exhibits some amount of anisotropy. The core may be elliptical in form (shape) which leads to a phenomenon referred to as form birefringence. Since mode propagation depends on the fiber structure, a fiber with an elliptical core will exhibit mode propagation that depends on the electric field's polarization with respect to the axes of the elliptical core. As a result of birefringence, the polarization of the mode varies as it propagates through the fiber (unless care is taken to align the polarization of the injected light with respect to a principal axis of the birefringence).

Polarization-maintaining (PM) fibers are designed to include birefringence in a particular axis of the fiber. By including a well-defined birefringence throughout the length of a fiber that is larger than that induced by external perturbations, fast and slow axes of the optical fiber are created for all guided wavelengths, giving two orthogonal 'polarization modes'. If light is injected into one of the polarization modes (i.e., with its linear polarization along one of the axes) it remains linearly polarized along that axis as it propagates along the fiber. The two polarization modes generally have different group velocities, so pulsed light in each mode will take a different amount of time to propagate through a given segment of fiber. Most PCFs exhibit strong birefringence due to a slightly elliptical core combined with a large core-cladding index difference, and so they behave similarly to PM fibers. Special care must be taken when working with PCF to be sure that the polarization of the light launched into the fiber is aligned with one of the birefringent axes.

In practice, there are a few other features of PCF that are of importance when discussing nonlinear-optic interactions:

- propagation losses are generally larger in PCFs than in standard optical fibers; and
- free-space coupling and splicing are difficult and can result in large coupling losses.

Nonlinear Phenomena

The basic principles determining nonlinear effects in PCFs are the same as those for standard optical fibers (*see* **Fiber and Guided Wave Optics:** Nonlinear Effects (Basics)). It is the new flexibility in PCFs to obtain transverse-modal and GVD behavior different from that of standard optical fibers that makes PCFs truly interesting for nonlinear optics. The relevant nonlinear-optical effects are: self-phase modulation (SPM); cross-phase modulation (CPM); third-harmonic generation (3HG); four-wave mixing (FWM); Raman scattering; and Brillouin scattering.

Self-phase modulation (also known as the optical Kerr effect) refers to the self-induced phase shift experienced by an optical field as it propagates through a fiber. It becomes particularly important for the case of pulses of light propagating through optical fibers. In small core PCFs, SPM is enhanced due to the high-intensity light propagating within the core. Self-phase modulation can lead to substantial spectral broadening of pulsed light propagating along an optical fiber.

When a pulse of light experiences normal GVD (i.e., $D < 0$) as it propagates, the longer-wavelength components travel faster than the shorter-wavelength components. Anomalous GVD (i.e., $D > 0$) leads to the opposite, short-wavelength components traveling faster than the long-wavelength components. Group-velocity dispersion generally leads to temporal broadening of pulses as they propagate along a fiber. Under ideal conditions, however, SPM in combination with anomalous GVD, leads to pulses which propagate without any temporal or spectral broadening. These self-sustaining pulses are called 'optical solitons'.

When waves of light having different wavelengths co-propagate along a fiber, CPM can occur. It can be understood as a phase shift induced on one wave, due to the presence of the other wave. Cross-phase modulation also leads to spectral broadening and solitonic pulse propagation.

In 3HG and FWM, one or more photons are destroyed and others are created. In 3HG, three 'fundamental' photons are destroyed to create one with three times the energy of the fundamental photons. In FWM, two fundamental photons are destroyed while two others are created. While it is straightforward to conserve energy in 3HG and FWM, these interactions must be 'phase-matched', meaning that the interacting waves must be made to propagate in-phase over a meaningful length. Such phase-matching conditions need to be carefully considered when studying 3HG and FWM. Nevertheless, 3HG and FWM can be used to obtain

frequency shifts and all-optical amplification. In comparison with other types of optical fibers, PCFs are particularly useful for 3HG and FWM applications. The small core of the PCF allows interactions to occur at much lower input powers, and the new flexibility associated with the GVD properties permits phase matching in cases which are not possible using standard optical fibers. Finally, the fiber's endlessly single-mode behavior permits very good transverse mode overlap between interacting waves having widely different center wavelengths.

Self-phase modulation, CPM, 3HG, and FWM are photon–photon interactions wherein no energy is exchanged with the medium itself. In contrast, Raman scattering and Brillouin scattering result from photon–phonon interactions. The differences between Raman and Brillouin scatterings lie in the energy of the phonons involved and the direction in which the interactions occur. Raman scattering is an interaction between a photon and an optical–phonon mode of the molecules making up the material. In the case of Raman scattering in glass, the energy shift, associated with molecular vibrational (Raman) modes, corresponds to frequencies of 1–12 THz. Raman scattering occurs in the forward and backward directions. With pulsed light, stimulated Raman scattering can occur when the lower-frequency spectrum of the pulse overlaps with the spectrum of the Raman resonances excited by the higher-frequency spectrum. When this happens, energy can be efficiently shifted in spectrum towards the peak of the Raman resonance. In the forward direction, this 'Raman self-frequency shift' builds up. Additionally, if the Raman self-frequency shift occurs in the presence of anomalous dispersion, a 'Raman soliton self-frequency shift' can result. As the injected power is increased, the spectral shift between the injected pulse and the resulting Raman soliton increases. The principal advantage associated with PCF is the ability to generate Raman solitons for a broader range of wavelengths than was previously possible.

For Brillouin scattering, interaction with the acoustical phonons results in frequency shifts of about 10 GHz and the interaction only occurs in the backward direction. Brillouin scattering is generally a nuisance in fiber-based devices, leading to intensity noise and other problems. The interesting feature of Brillouin scattering in PCFs is that the threshold intensity where problems begin to occur is higher for PCFs than for standard optical fibers. The higher threshold permits further optimization of fiber-based devices wherein Brillouin scattering limits the performance.

Experiment Examples

In the following subsections, a selection of experiments, demonstrating a few of the relevant nonlinear-optical phenomena, are briefly described.

Supercontinuum Generation

One of the most exciting demonstrations of nonlinear optics in PCF is that of supercontinuum generation. In a typical experiment, 100 femtosecond pulses from a mode-locked Ti:Sapphire laser operating at a wavelength of 800 nm were injected into the PCF. As the injected power was increased, a broad continuum of spectrum was generated from wavelengths of 400 nm up to 1,600 nm. Typical data showing the input and output optical spectra are given in Figure 4, where it can be seen that well over an octave of frequency spectrum is generated from an input pulse whose spectral width is only about 10 nm.

It is widely accepted that supercontinuum generation results from a combination of linear and nonlinear optical effects conspiring to generate the broad spectrum. As the pulse propagates through the PCF, SPM, FWM, and Raman scattering are all likely to occur with relative efficiencies depending heavily on the input pulse's spectral and temporal characteristics, as well as the fiber's properties. Indeed, efficient supercontinuum generation can occur in PCFs for pump wavelengths lying near the zero-dispersion point in the normal or anomalous dispersion regime of the PCF, and for fibers as short as a few millimeters in length.

Spectral broadening, due to SPM, is common in standard optical fibers, but what makes this

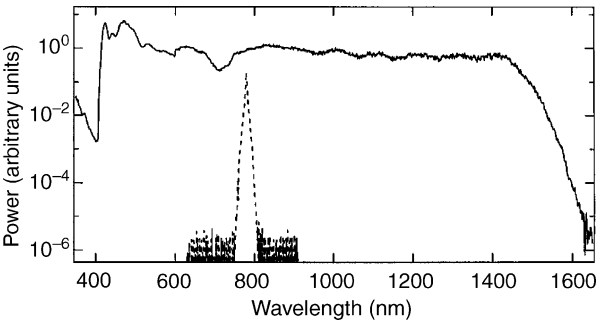


Figure 4 In supercontinuum generation one observes a broad continuum generated after short pulses of light from a Ti:Sapphire laser propagate through a 75 cm section of PCF. The spectrum of the input pulse is shown as a dashed curve, while the output is a solid curve. (Reproduced with permission from Ranka JK, Windeler RS and Strentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optical Letters* 25: 25–27. ©2000 Optical Society of America.)

particular experiment interesting is the remarkable width of spectrum generated with a short piece of fiber (~1 m) and with comparatively small optical powers.

Optical Switching in PCF

Cross-phase modulation can be used to build an all-optical switch. Such a switch can be implemented as a three-port device where the output port for a given optical bit (the signal pulse), is determined by the presence of an optical control (the pump pulse). Switching can then be achieved by dividing the signal pulse equally on two arms of an interferometer and injecting the strong pump pulse only on one arm. Because the pump pulse co-propagates with only one of the two signal pulses, there exists a CPM-induced phase difference $\phi_{NL} = 2\gamma P_p L$ at the output of the interferometer, where P_p is the peak power of the pump pulse and L is the interaction length. By varying the intensity of the pump pulses one can vary the magnitude of this phase difference. If a π-phase shift is achieved, one can switch the interference from destructive to constructive, or vice-versa, thus realizing an all-optical switch.

Figure 5 shows an experimental setup used to observe switching near 1,550 nm (a similar apparatus using bulk-optic rather than fiber-optic components can be used to conduct experiments near 780 nm). The pump and signal are synchronous few-ps-duration pulses with a tunable wavelength separation

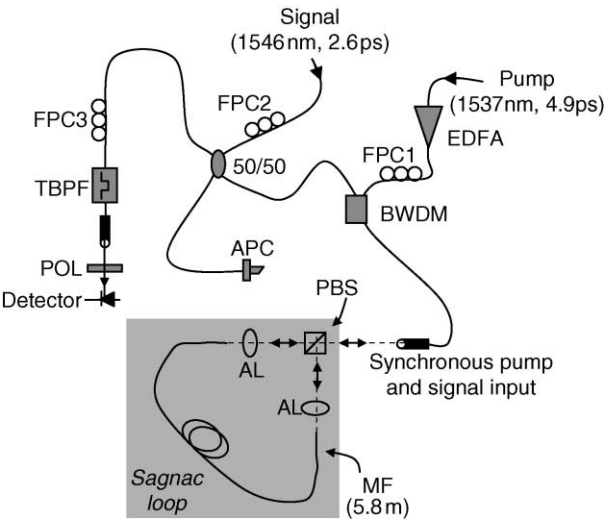


Figure 5 Experimental setup used to demonstrate all-optical switching near 1,550 nm. (EDFA, erbium-doped fiber amplifier; BWDM, bandpass wavelength-division multiplexor; PBS, polarization beamsplitter; FPC, fiber polarization controller) (Reproduced with permission from Sharping JE, Fiorentino M, Kumar P and Windeler RS (2002) All-optical switching based on cross-phase modulation in microstructure fiber. *IEEE Photonics Technology Letters* 14: 77–79. ©2002 IEEE).

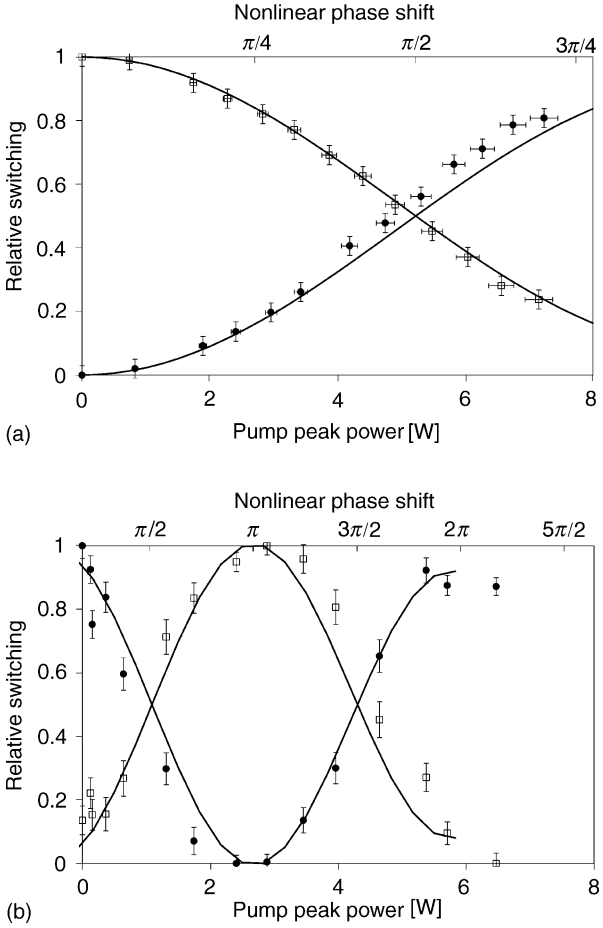


Figure 6 Switching curves (open boxes and filled circles), showing the relative power measured in each port of the switch vs. the pump peak power, for experiments conducted near (a) 1,550 nm and (b) 780 nm. The curves accompanying the data are generated from numerical solutions of coupled wave equations for CPM. (Reproduced with permission from Sharping JE, Fiorentino M, Kumar P and Windeler RS (2002) All-optical switching based on cross-phase modulation in microstructure fiber. *IEEE Photonics Technology Letters* 14: 77–79. ©2002 IEEE.)

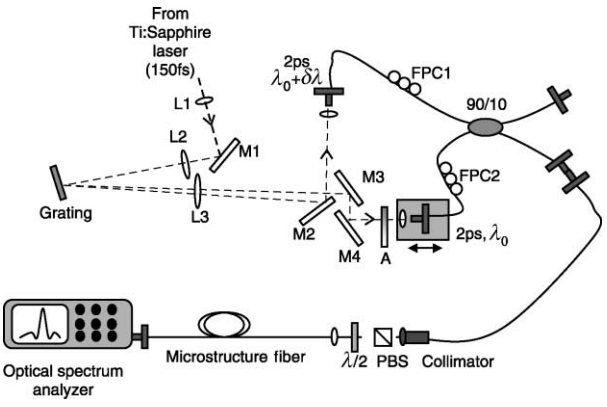


Figure 7 A schematic of the experimental setup used to investigate FWM in PCFs. (Reproduced with permission from Sharping JE, Fiorentino M, Coker A, Kumar P and Windeler RS (2001) Four-wave mixing in microstructure fiber. *Optical Letters* 26: 1048–1050. ©2001 Optical Society of America.)

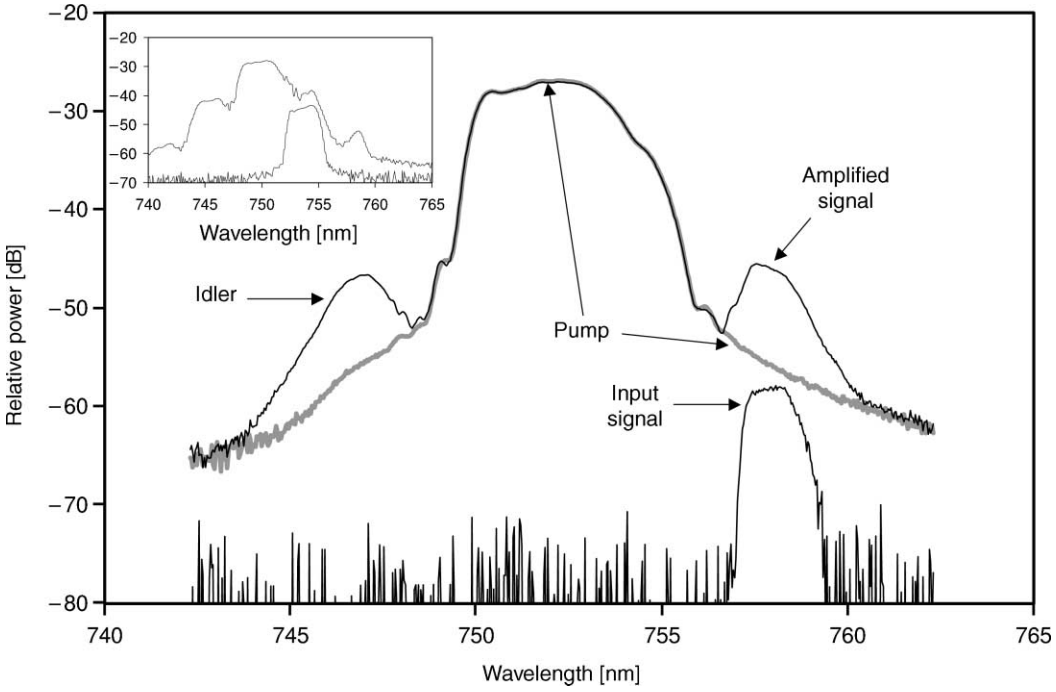


Figure 8 A typical FWM spectrum observed at the output of the microstructure fiber. The inset shows a spectrum where higher-order cascaded mixing is evident. (Reproduced with permission from Sharping JE, Fiorentino M, Coker A, Kumar P and Windeler RS (2001) Four-wave mixing in microstructure fiber. *Optical Letters* 26: 1048–1050. ©2001 Optical Society of America.)

of 5–15 nm. The switching characteristics for this implementation are shown in Figure 6. The apparatus has the advantage of requiring short fiber lengths, low switching powers, and allows switching of weak pulses. It demonstrates the feasibility of using nonlinear optics in PCF to perform essential functions in high-speed all-optical processing.

Parametric (Mixing) Processes

The first set of experiments with controlled FWM in a PCF achieved nondegenerate parametric gains over a 30 nm range of pump wavelengths near the λ_0 of the PCF. The experiments also confirmed the wavelength dependence of the GVD coefficient of the PCF near λ_0. Since the dispersion characteristics of these fibers can be adjusted during the fabrication process, the experiments demonstrate the potential for the use of PCFs in broadband parametric amplifiers, wavelength shifters, and other optical communication devices.

The experimental setup used to demonstrate phase-matched FWM in PCF is shown in Figure 7. The pump and the input signal are two synchronous pulsed beams having 3–5 nm wavelength separation with the center wavelength tunable over a 720–850 nm range. The maximum peak power of the pump pulses is ≃12 W. The two synchronous beams are then combined and injected into the PCF. The pump and signal's optical paths are adjusted to obtain temporal overlap in the PCF and their polarizations are aligned by fiber polarization controllers.

Figure 8 shows a typical FWM optical spectrum at the output of a 6.1 m long PCF. Here the strong pump beam and the weak signal beam have wavelengths of 753 nm and 758 nm, respectively. The spectrum shows the undepleted pump, the amplified signal, and the generated idler at 747 nm. The spectra in Figure 8 show that large gain is achievable for a pump-to-signal spacing of 5 nm. Gain values of more than 20 (13 dB) were obtained.

Conclusion

In summary, the advantages of using photonic crystal fibers for demonstrating nonlinear-fiber optical effects arise from four novel properties:

- the nonlinear coefficient is enhanced in small-core PCFs (core area of a few μm^2);
- PCFs can support a single transverse mode over an extremely broad wavelength range (370 nm–1600 nm);
- PCF design parameters allow one to manipulate the fiber's GVD properties; and
- nonlinear interactions are enhanced due to the polarization maintaining properties of PCFs which result from form birefringence present in the core.

These four properties combine to allow efficient interactions to occur in PCFs which are either inefficient or not possible at all in standard optical fibers.

See also

Optical Amplifiers: Raman, Brillouin and Parametric Amplifiers. **Photonic Crystals:** Electromagnetic Theory. **Scattering:** Raman Scattering.

Further Reading

Agrawal GP (2000) *Nonlinear Fiber Optics*, 3rd edn. San Diego, CA: Academic Press.

Birks TA, Knight JC and Russell PStJ (1997) Endlessly single-mode photonic crystal fiber. *Optical Letters* 22: 961–963.

Hansryd J, Andrekson PK, Westlund M, Li J and Hedekvist P (2002) Fiber-based optical parametric amplifiers and their applications. *IEEE Journal of Selected Topics in Quantum Electronics* 8: 506–520.

Monro TM, Richardson DJ, Broderick NGR and Bennett PJ (2000) Modeling large air fraction holey optical fibers. *Journal of Lightwave Technology* 18: 50–56.

Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optical Letters* 25: 25–27.

Russell PStJ (2003) Photonic crystal fibers. *Science* 299: 358–362.

Sharping JE, Fiorentino M, Coker A, Kumar P and Windeler RS (2001) Four-wave mixing in microstructure fiber. *Optical Letters* 26: 1048–1050.

Sharping JE, Fiorentino M, Kumar P and Windeler RS (2002) All-optical switching based on cross-phase modulation in microstructure fiber. *IEEE Photonics Technology Letters* 14: 77–79.

Photonic Crystal Lasers, Cavities and Waveguides

J O'Brien and W Kuang, University of Southern California, Los Angeles, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Photonic bandstructure engineering, in which the electromagnetic dispersion relations are intentionally modified, can be accomplished by creating one-, two-, and three-dimensionally periodic dielectric materials. These periodic materials, called photonic crystals, can be used to confine and guide light. This article describes the formation of optical resonant cavities and waveguides in these materials. Most of the discussion will focus on two-dimensional photonic crystals. One-dimensionally periodic systems are treated elsewhere in this volume. This article will begin with a short review of the relevant electromagnetic properties of photonic crystals. The formation and properties of resonant cavities will follow this, and the article will conclude with waveguides formed by linear defects in two-dimensional photonic crystals.

Electromagnetics of Photonic Crystals

A two-dimensional photonic crystal consists of a dielectric material in which the dielectric constant is periodic in two dimensions. The period of these materials is on the order of a half-wavelength of the operating optical wavelength. In this section, we briefly review the electromagnetic properties of two-dimensional photonic crystals. For more details, see the article entitled **Spectroscopy:** Raman Spectroscopy in this volume.

There are five Bravais lattices in two dimensions. Most of the research on photonic crystal resonant cavities and waveguides has focused on square and triangular lattices, and in this article we will consider examples based on the triangular lattice. Since these structures are usually defined lithographically however, there is, in principle, no need to confine the investigation to naturally occurring Bravais lattices. Due to the limitations of current nanofabrication technology, much of the research has focused on photonic crystals that have a finite thickness. The important feature of photonic crystals is that the bandgaps in the electromagnetic spectrum opened up by all of the Bragg planes overlap spatially and spectrally, so that a frequency region exists in which no electromagnetic propagation is allowed. These electromagnetic bandgaps can be used to confine optical modes in very small-volume resonant cavities and in waveguides. In high-contrast dielectric systems, such as a semiconductor/air periodic system, only a few lattice periods may be necessary to confine an electromagnetic mode. In the case of finite thickness photonic crystals, this periodicity has the effect of limiting the bandgap formed to the in-plane

directions and in the case of a two-dimensional photonic crystal in a high-index dielectric slab to the formation of a bandgap in the guided modes of the slab. In this case there are still radiation modes of the slab.

Consider a two-dimensional photonic crystal that is periodic in the $x-y$ plane. To start, we can consider materials that are uniform and infinitely extended in z. Electromagnetic fields propagating in the $x-y$ plane are Bloch states. These can be written as:

$$\vec{H}_{k_z,k_\parallel}(\vec{\rho},z,t) = \mathrm{e}^{\mathrm{i}(\omega t-(\vec{k}_\parallel\cdot\vec{\rho}+k_z z))}\vec{u}_{k_z,k_\parallel}(\vec{\rho}) \qquad [1]$$

where $\vec{\rho}$ labels the in-plane coordinates and $\vec{k}_\parallel$ labels the in-plane wavevectors. $\vec{u}_{k_z,k_\parallel}$ is a periodic function in the $x-y$ plane. Modes propagating in the $x-y$ plane in such a system can be classified as transverse electric (TE) waves or transverse magnetic (TM) waves. TE waves have nonzero E_x, E_y, and H_z field components, and TM waves have nonzero E_z, H_x, and H_y field components. A triangular lattice of air holes patterned into a high refractive index dielectric can be used to create a bandgap for both the guided TE and TM modes of the membrane for a range of hole radii. The TE and TM bandgaps that are formed can be over different spectral ranges, however. For laser applications this is unimportant since in practice the emission occurs from electron-hole recombination in a semiconductor quantum well, and for unstrained or compressively strained quantum well materials this emission is TE polarized. In most cases, waveguides have also been designed to work for a single optical polarization. It is generally true that a photonic lattice that consists of a connected high dielectric region is likely to exhibit a TE bandgap, while a lattice formed by disconnected high dielectric regions is more likely to exhibit a TM bandgap. The bandgap in the TE modes is formed between the first and second bands. The first band is called the dielectric band because the field at the Brillouin zone boundary is a standing wave with its intensity concentrated in the high dielectric regions. The second band is called the air band because at the Brillouin zone boundary this field is a standing wave with its intensity localized in the low dielectric regions. Over a reasonable range of lattice parameters, the bandwidth of the TE bandgap can be changed by changing r/a where r is the radius of the holes and a is the lattice constant of the triangular lattice. As r/a gets larger the dielectric band moves up in frequency. This can be thought of as being due to the fact that this mode has a decreasing effective index as r/a increases. The air band frequency increases as r/a increases. Since more of the electric field of the air band is located in the low dielectric regions than the field of the dielectric band, the air band increases in frequency with increasing r/a faster than the dielectric band and the bandgap therefore increases with increasing r/a. This trend holds in photonic crystals of finite thickness as well.

For photonic crystals of a finite thickness, the ability to simply classify modes as either TE or TM is lost. In the most simple finite thickness photonic crystals, that of a high index dielectric slab in which a two-dimensional photonic crystal has been patterned surrounded above and below by the same low index material which may be air of sapphire, or silicon dioxide, the modes can be classified as being either even or odd modes with respect to the mid-plane of the high index slab. In this mid-plane, however, modes can be classified as being either TE or TM, with each containing the same nonzero vector field components as in the infinite case. Away from this mid-plane, however, modes have in general six nonzero electromagnetic field components. The photonic bandgap that is created in this case is in the guided mode spectrum of the high index dielectric slab. There is no gap formed in the radiation modes of the slab. The radiation modes consist of modes with a ray vector that is not totally internally reflected at the high index slab/low index cladding interfaces. Figure 1 shows the dispersion relation, plotted over the irreducible Brillouin zone, of the even modes for dielectric slab with a refractive index of 3.4 in which a two-dimensional triangular lattice of holes has been patterned. The surrounding material is air. The figure shows an electromagnetic bandgap

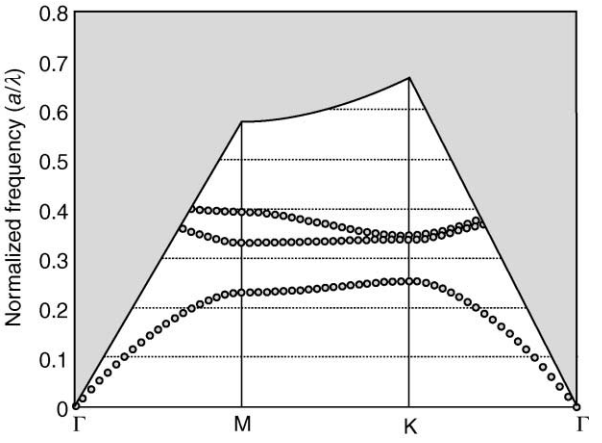


Figure 1 The photonic band diagram for a photonic crystal slab with a triangular lattice of air holes perforating a high-index dielectric slab. The slab, which assumes a refractive index $n = 3.4$, is suspended in the air with a thickness of $d/a = 0.6$. The air holes have radii of $r/a = 0.3$, where a is the lattice constant. Only the lowest three eigenmodes are shown in the plot.

formed between the first and the second guided modes of the slab. Radiation modes of the slab have real valued propagation vectors in the z-direction. These modes lie above the in-plane dispersion relation for air and exist inside the shaded region in the figure. This dividing line between the guided modes and the radiation modes is referred to as the light line.

If the upper and lower cladding layers of the high index slab are different and therefore have different indices of refraction, then the ability to classify modes as being either even or odd about the mid-plane of the slab disappears and it is possible to lose the bandgap in the guided modes altogether. In cases where the refractive index of the bottom cladding is not significantly different from that of the top cladding, such as a case where air serves as the top cladding layer and a material such as sapphire or silicon dioxide forms the bottom cladding, it is found that the even and odd modes describe the field reasonably accurately. To emphasize the approximation involved in the model, these modes are most often referred to in the literature as even-like or odd-like. Overall, there are three primary effects of having asymmetric cladding layers such as an air top cladding and sapphire bottom cladding case. These are a reduction in the effective bandgap width, an increase in the area of the radiation modes on the dispersion relation, and the loss of a rigorous bandgap in the guided modes of the slab. The first two of these effects are the most serious for device designers.

Photonic Crystal Resonant Cavities and Lasers

Photonic crystal resonant cavities can be formed at frequencies inside the bandgap or at the bandedges. Modes in the bandgap are formed as a result of a defect in the lattice. Modes at the bandedge are also sometimes used because the fields with wavevectors at Brillouin zone boundary are standing waves. These standing waves are analogous to the resonant modes that are formed in distributed feedback (DFB) laser structures. Here we will focus on modes confined in the bandgap. Allowed modes in electromagnetic bandgaps in photonic crystal slabs can be introduced by introducing one or more defects into the lattice. By perturbing a single lattice site, we can permit a localized mode or modes that have a frequency in the gap. This phenomenon is similar to the formation of deep levels in an electronic bandgap of a solid due to impurities and identical to the formation of resonant modes in vertical cavity surface emitting lasers (VCSELs) and phase-shifted DFB lasers. In each of these last cases, a defect in the periodicity of a precise thickness occurs between two one-dimensional distributed Bragg reflectors. Defect modes in photonic crystal resonant cavities can be engineered to radiate in a particular direction by trading-off the in-plane versus out-of-plane losses, and the resonant frequency is determined by the parameters of the lattice. Very small mode optical mode volumes can also be obtained in these resonant cavities leading to the possibility of modifying spontaneous emission.

Simple models exist which illustrate the formation of localized modes inside the bandgap as a result of defects in the otherwise periodic lattice. One of these was applied to the formation of deep levels in solids. This model can also be applied to the electromagnetics of defects in periodic structure. For a simple one-dimensional case, assuming that the defect exists at a single unit cell and that the dispersion of the band is dominated by nearest neighbor interactions, this can be solved exactly. Figure 2 shows the allowed frequencies of a ten unit cell chain with a single defect located at the center unit cell. Notice that one mode, called a donor mode, drops out of the band for positive values of $\Delta\varepsilon$ and one mode, called an acceptor mode, rises out of the band for negative values of $\Delta\varepsilon$. This is true in general. A perturbation of the lattice in which the defect has a higher index than the unperturbed lattice will cause a mode to drop out of the air band while a perturbation that has a lower dielectric constant than the background lattice will cause a mode to rise into the bandgap out of the dielectric band. In practice, resonant modes and frequencies in photonic crystals are modeled numerically.

A great deal of variety exists in resonant modes formed by defects in two-dimensional photonic crystal slabs. The simplest examples consist of a single missing hole in a two-dimensional square or

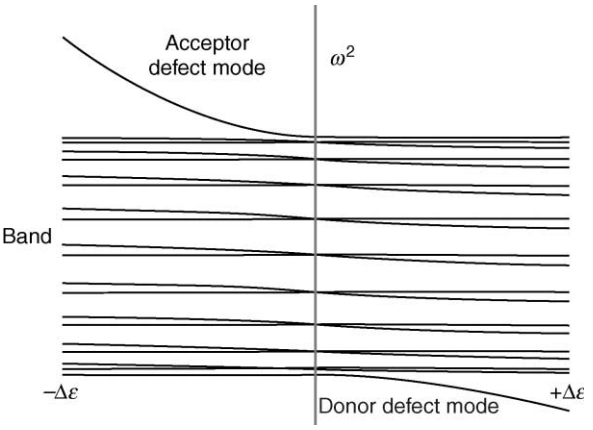


Figure 2 A one-dimensional Slater–Koster model for a single defect in a 10-unit-cell chain, showing the allowed states as a function of the defect perturbation.

triangular lattice patterned into a high-index dielectric slab. However, the resonant mode size, shape, frequency, and polarization can be engineered by tailoring the local dielectric in the region of the resonant mode. Losses in these cavities can be conceptually separated into losses in-plane through the photonic crystal and radiation losses out-of-plane. The in-plane losses are a result of having only a finite number of lattice periods surrounding the localized mode. A finite number of periods results in a finite loss due to tunneling of the fields through the lattice. This loss can in principle be made arbitrarily small by increasing the number of lattice periods surrounding the resonant mode. For a fixed number of lattice periods, the in-plane radiation loss will be reduced as the bandwidth of the photonic bandgap is increased because the magnitude of the in-plane decay constant increases with increasing bandgap width. A resonant mode formed by a defect in a truly three-dimensional photonic crystal will have losses limited by the number of achievable lattice periods and the bandwidth of the photonic crystal bandgap.

Photonic crystal cavities formed by two-dimensional photonic crystals of finite thickness suffer from an out-of-plane radiation loss. Optical confinement in the direction perpendicular to the two-dimensional photonic crystal is due to total internal reflection of the mode that occurs in the high index photonic crystal slab surrounded by lower index materials. Out-of-plane losses are due to the presence of wavevector components in the resonant mode that lie above the light line inside the radiation cone of the cladding. Figure 3 shows the Fourier transform of the magnetic field at the mid-plane of photonic crystal slab in which the resonant mode is formed by a single missing hole in a two-dimensional triangular lattice. In this cavity geometry, a doubly degenerate pair of modes is introduced into the photonic bandgap. The figure shows the Fourier transform of the field for one mode of this double degenerate pair. The boundary of the radiation cone in the figure is marked by the solid circle. Components of the mode inside this circle are not confined to the slab and contribute to out-of-plane radiation losses. These losses can be reduced by engineering the mode so that it has a smaller overlap with the radiation cone. This can be accomplished by allowing the mode to expand in real space in directions in which the wavevectors of the mode extend into the radiation cone. This reduces the spread of the Fourier transform of the mode and reduces the overlap of the mode with the radiation cone.

Symmetry can also be used to reduce the coupling to the radiation fields of the slab. Reducing the resonant frequency of the mode can also reduce out-of-plane radiation losses. This has the effect of reducing the size of the radiation cone in k-space. One way to accomplish this is to reduce r/a in the photonic crystal lattice. This will generally reduce frequencies of all of the photonic crystal modes. It will also, however, generally be accompanied by a reduction in the bandwidth of the photonic bandgap which will increase the in-plane losses. As a result, care must be taken in designing photonic crystal resonant cavities. Nevertheless, in the smallest mode volume photonic crystal resonant cavities, the optical losses will generally be dominated by out-of-plane radiation loss.

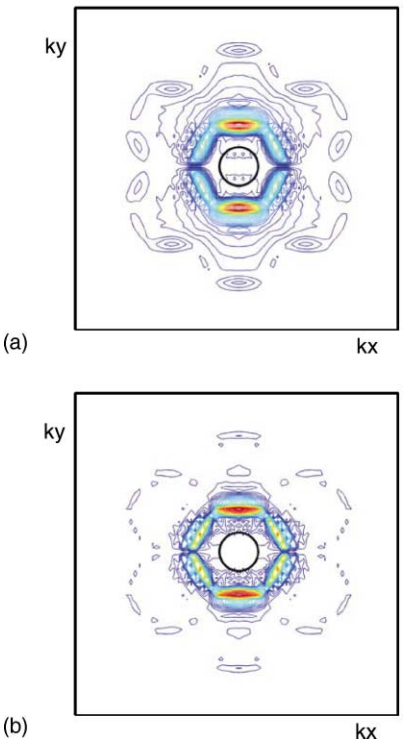


Figure 3 (a) The spatial Fourier transform of the magnetic field component, Hz, of a defect cavity mode at the mid-plane of a triangular-lattice suspended membrane single-defect photonic crystal microcavity. The inner circle indicates the light cone inside which mode components radiate vertically. (b) The spatial Fourier transform of a modified defect cavity mode. The overlap of the mode with light cone has been reduced by modifying the defect cavity.

Quality factors have been predicted to exceed 100 000 in carefully designed photonic crystal resonant cavities in which the optical mode volume is on the order of a few cubic half-wavelengths in the material. Larger resonant cavities formed by introducing multiple defects in the lattice are often less well confined in the real space lattice leading to a reduction in the components overlapping the radiation cone in k-space and a reduction in out-of-plane

radiation loss. These cavities also are much more likely to support multiple resonant modes for a given index perturbation.

In analyzing resonance modes of photonic crystal resonant cavities, one of the most important parameters is the quality factor Q of the cold cavity. Practically, a theoretical prediction of the quality factor requires the use of numerical methods and finite-difference time-domain and finite element techniques are commonly used. There are basically three numerical approaches that can be used to calculate the quality factor of the cavity modes. Two of these methods calculate the quality factor from the time domain fields while the third is a frequency domain approach. The first method is to calculate the slope of the exponential decay of the energy of a given cavity mode with time. This decay of energy from the cavity is described by $\exp(-t/t_{\text{ph}})$ where t_{ph} is the photon lifetime which is related to the quality factor Q, by $Q = \omega t_{\text{ph}}$. This method is most useful for relatively low Q modes where the slope of energy decay is visibly greater than zero.

Another method is to calculate the ratio of the full width at half magnitude (FWHM) of the cavity resonance in the frequency domain, $\Delta\omega$, to the center frequency, ω_0. However, distortion to the spectrum is often introduced because the numerical simulation terminates before the impulse response is fully evolved. This has the effect of viewing the true time-domain response through a rectangular window, which translates mathematically into the convolution of the true spectrum with a sinc function. This must be accounted for in the determination of the quality factor.

A third method calculates the ratio of cycle-averaged power absorbed in the boundary to the total energy in the cavity mode. This last method has the advantage of by being able to separate the radiation losses into different directions:

$$\frac{1}{Q} = \frac{\omega P}{U} = \frac{\omega(P_{\perp} + P_{\parallel})}{U} = \frac{1}{Q_{\perp}} + \frac{1}{Q_{\parallel}} \qquad [2]$$

in which, the effective vertical quality factor $Q_{\perp}$ is given by the ratio of power lost to the absorber at the top and bottom $P_{\perp}$ to the total cavity energy $U(t)$, and the effective in-plane quality factor $Q_{\parallel}$ is similarly given by the ratio of in-plane power loss $P_{\parallel}$ to the total cavity energy.

Photonic crystal lasers and resonant cavities have been demonstrated experimentally. The first demonstration of lasing in a two-dimensional photonic crystal laser occurred in a cavity formed by a single defect in a triangular lattice that was patterned into an InGaAsP membrane. The laser operated under optically pumped conditions at 143 K. Room temperature pulsed operation of photonic crystal defect lasers was reported shortly thereafter. Figure 4 shows an electron micrograph of such a resonant cavity. In the figure, the resonant cavity is formed in a suspended membrane in which a triangular lattice photonic crystal with a defect has been patterned. The fabrication of these devices involves pattern definition by electron beam lithography and pattern transfer by a series of wet and dry etching steps. The observed quality factors are, of course, sensitive to fabrication imperfections. Many other cavity designs and demonstrations have also been reported.

To operate a photonic crystal laser continuously at room temperature, it is necessary to design a high-Q resonant cavity that dissipates heat well. Poor heat dissipation is a major drawback of suspended membrane resonant cavities. One strategy for improving the heat dissipation in photonic crystal laser cavities is to form the two-dimensional photonic crystal membrane cavity on top of a low-index high-thermal-conductivity substrate. Figure 5 shows an electron micrograph of a photonic crystal cavity formed by leaving out 37 holes from a triangular lattice patterned into an InGaAsP membrane. This membrane is bonded to a sapphire substrate, which facilitates heat dissipation. This cavity is capable of room temperature continuous wave operation.

It is believed that these lasers are capable of electrically-pumped room temperature operation. This is based on the fact that large quality factors have been demonstrated in optically-pumped photonic crystal lasers and some quality factor reduction due to free carrier absorption and absorption at metal contacts can be tolerated and still lead to lasing. A reasonable assumption would be that the

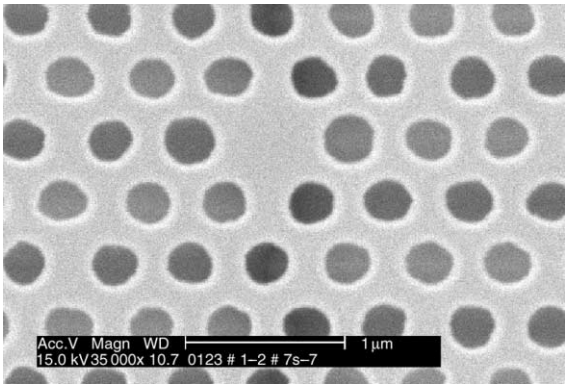


Figure 4 The scanning electron micrograph of a triangular-lattice suspended membrane photonic crystal single defect cavity.

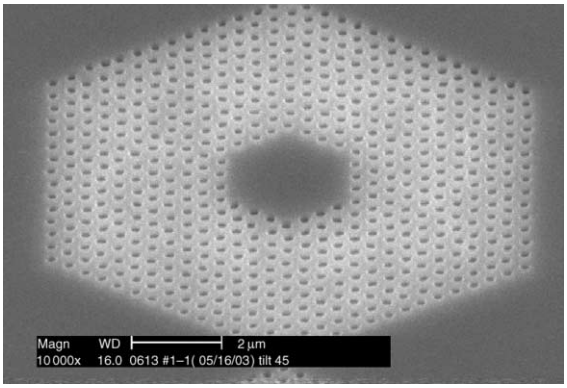


Figure 5 The scanning electron micrograph of a triangular-lattice, 37-missing-hole, photonic crystal defect cavity, bonded on sapphire which serves as lower cladding to improve heat dissipation. CW operation of the laser cavity was demonstrated.

performance of these electrically-pumped photonic crystal lasers would have much in common with the performance of VCSELs. Both VCSELs and photonic crystal cavities have small mode volumes with the cavity formed in some directions by distributed Bragg reflection. The emission direction of photonic crystal lasers can be engineered to be vertical or in-plane, but the basic resonant cavities do have important similarities. The free carrier absorption loss in photonic crystal laser cavities may be expected to be larger than in VCSELs since a photonic crystal laser mode lacks the standing wave behavior in the vertical direction that allows doping at the nodes of the standing wave to reduce the absorption loss. If free carrier absorption loss leads to larger internal losses for photonic crystal lasers than for VCSELs, then some reduction in the achievable slope efficiencies will also occur for photonic crystal lasers.

Photonic Crystal Waveguides

Another important optical element for integrated photonic circuitry is a linear waveguide. Conventional dielectric waveguides confine propagating beams by an index of refraction difference between the waveguide core and the waveguide cladding. Photonic crystal waveguides are most often formed by a linear defect, which consists of a row of modified unit cells of the lattice, patterned into a high index dielectric membrane. The guiding in this case is due to a mixture of total internal reflection at the high index membrane/low index cladding interfaces and distributed reflections from the photonic crystal in-plane. Because this confinement does not exclusively dependent on total internal reflection, the transmission loss through small bending radius waveguide branches and bends can be made very small. Predictions and experimental demonstrations exist which show that

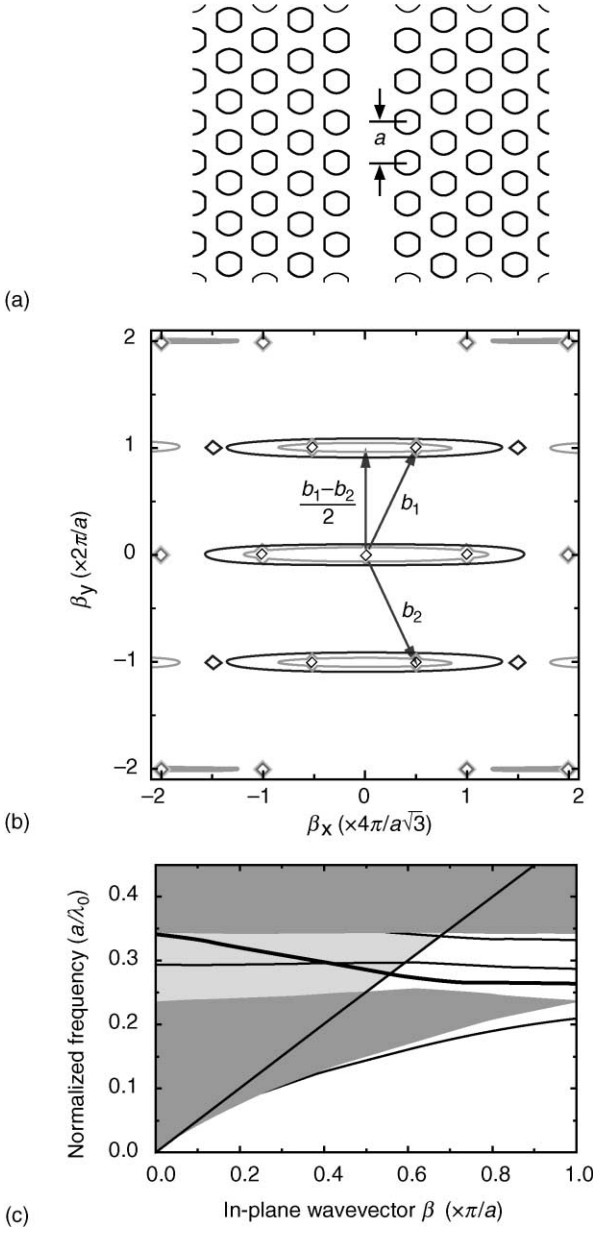


Figure 6 (a) The top view of a triangular lattice photonic crystal single-line defect waveguide with air hole radius $r/a = 0.3$. (b) The 2D spatial Fourier transform of the waveguide dielectric distribution. (c) The photonic band diagram for a suspended membrane single-line photonic crystal waveguide. The membrane has a refractive index of $n = 3.4$ and its thickness is $d/a = 0.6$. The surrounding material is air. The light gray and dark gray area indicate the regimes where field radiates through the air and photonic crystals, respectively.

photonic crystal waveguides are capable of supporting small bending radius waveguide bends with almost total transmission. This makes photonic crystal waveguides a candidate for waveguides in densely integrated photonic circuits.

Figure 6a shows an illustration of the top view of a two-dimensional photonic crystal waveguide.

This is a triangular lattice in which one row of unit cells has been modified. Shown in Figure 6b is the Fourier transform of index of refraction for the situation in which the photonic crystal consists of low index holes in a high index semiconductor. The presence of the linear defects modifies and broadens the reciprocal lattice from a simple reciprocal triangular lattice, before the creation of linear defect, to the distribution shown in the figure. This Fourier transform has components along the direction of propagation at $(\vec{b}_1 - \vec{b}_2)/2$ as shown in the figure where $\vec{b}_1$ and $\vec{b}_2$ are reciprocal lattice unit vectors. These reciprocal lattice components at $(\vec{b}_1 - \vec{b}_2)/2$ couple a planewave with wavevector β to other planewaves with wavevectors $\beta \pm (\vec{b}_1 - \vec{b}_2)/2$. The result of this coupling is that the electric field of the waveguide mode is a Bloch wave. In a photonic crystal waveguide in which the waveguide axis is the z-direction, the Bloch wavefield in the waveguide can be written in the form:

$$\vec{E}(x,y,z,t) = \sum_n c_n \vec{f}_\beta(x,z) \times \exp\left[\mathrm{i}\left(\omega t - \left[\beta + n\left(\frac{\vec{b}_1 - \vec{b}_2}{2}\right)\right]y\right)\right] \quad [3]$$

In this expression the periodic function of the Bloch wave has been explicitly expanded as a Fourier series and each term in the Fourier series is called a spatial harmonic. Note that these fields differ from that of a photonic crystal fiber because the cladding of a photonic crystal waveguide has a periodicity along the direction of a waveguide. A photonic crystal fiber does not.

Most of the realizations of photonic crystal waveguides occur in two-dimensional photonic crystals formed in high index dielectric slabs. The membrane is usually located a few microns above a high index substrate. Because the fields of the guided modes decay exponentially outside the membrane, the effect of a substrate several microns away is very nearly negligible. The photonic band structure for $r/a = 0.3$, normalized membrane thickness $d/a = 0.6$ and refractive index of $n = 3.4$ is shown in Figure 6c. Only the modes with even symmetry along the midplane of the membrane are shown. A bandgap opens for modes with odd symmetry only for large low index filling factors in the triangular lattice. Filling factors large enough to create a bandgap for the odd modes are often impractical so that the odd modes are rarely considered. The even and odd modes are orthogonal, so that plotting the dispersion relation for the even modes only is justified. In reality, even and odd modes may be weakly coupled as a result of fabrication imperfections. The horizontal axis of the dispersion relation covers the in-plane wavevector β along the propagation direction of the irreducible Brillouin zone. The vertical radiation light cone and transverse radiating region of the photonic crystal are mapped as light gray and dark gray areas, respectively. They correspond to the regimes in which light radiates through the air cladding and photonic crystal cladding, respectively. Waveguide dispersion relations can be calculated using two-dimensional approaches in which case the guiding due to the high index slab is accounted for by using an effective index of refraction for the mode. The effective index is the ratio of the propagation coefficient to the free space wavevector of the guided mode of the slab. The two-dimensional

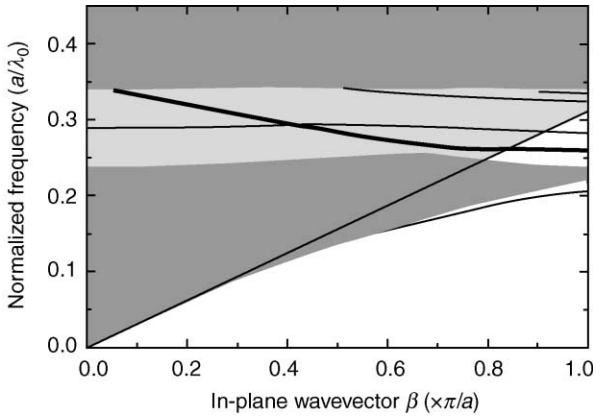


Figure 7 The photonic band diagram for a triangular-lattice sapphire-bonded single-line photonic crystal defect waveguide. The dielectric membrane has a refractive index of $n = 3.4$ and thickness of $d/a = 1.0$. The index of the sapphire bottom cladding is assumed to be 1.6.

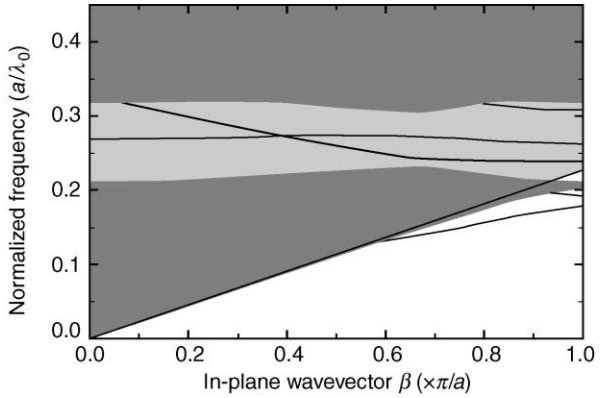


Figure 8 The photonic band diagram for a triangular-lattice deeply etched single-line photonic crystal defect waveguide. The top cladding, guiding membrane, and bottom cladding have an index of refraction of 3.0, 3.4, and 3.0, respectively. Their normalized thicknesses d/a are 1.0, 1.0, and 6.0, respectively.

calculation then uses the effective index of the mode as the high index of material in the photonic crystal. To obtain accurate results, the effective index for each mode calculated separately and the two-dimensional calculation of the photonic crystal waveguide dispersion must be repeated for each waveguide mode.

Examples of photonic crystal waveguide dispersion relations are shown in Figures 7 and 8 for waveguides formed by photonic crystals in dielectric slabs on high index lower cladding layers. Figure 7 is the dispersion relation for an oxide lower cladding layer. The area of the light gray region, which is the projection of the oxide cladding light cone onto waveguide propagation direction, increases as a result of higher index of refraction. Figure 8 shows the dispersion relation for a waveguide with an even larger lower cladding index. In this case the air holes of the two-dimensional photonic crystal are patterned all the

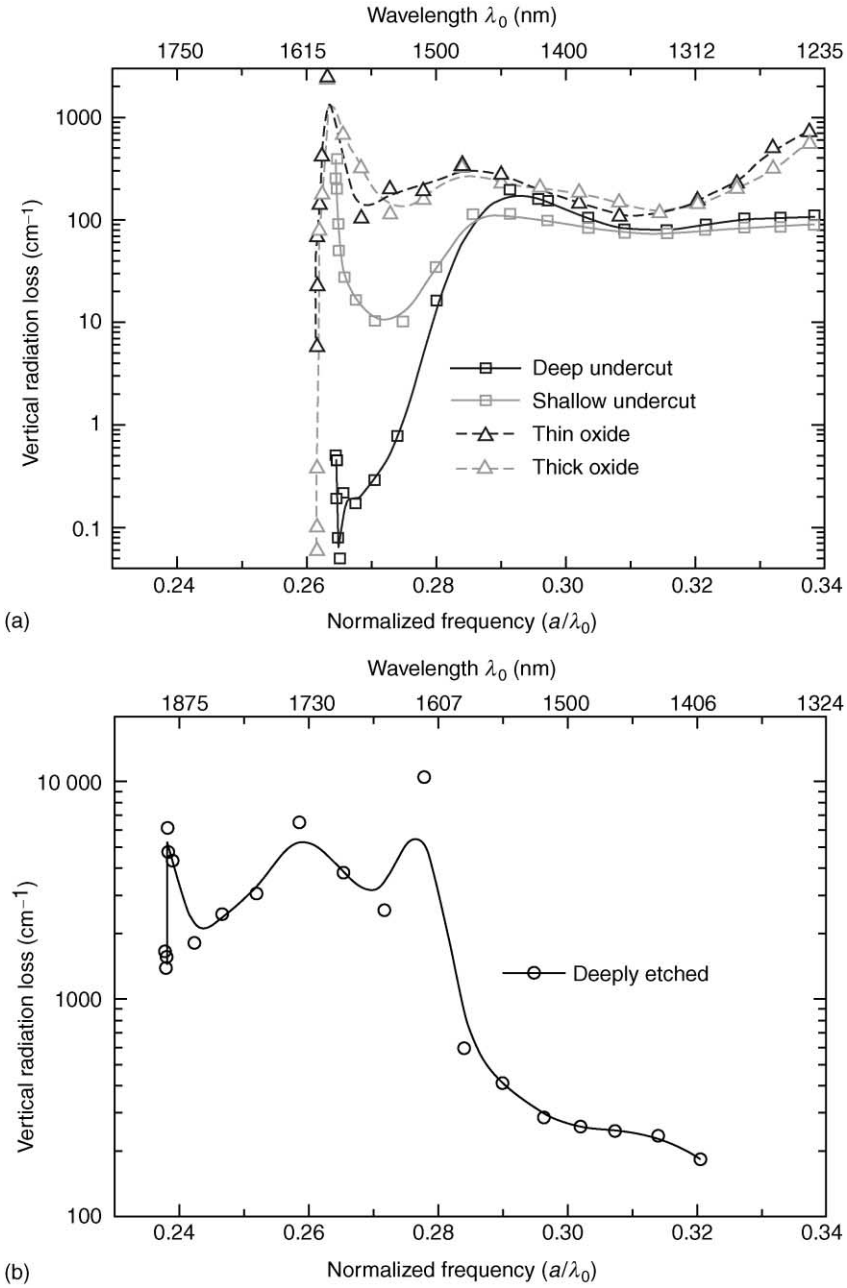


Figure 9 The out-of-plane radiation loss as a function of normalized frequency for the photonic crystal defect slab waveguides illustrated in Figures 6–8. The lattice constant $a = 450$ nm for deeply etched waveguide and 420 nm for the rest of the cases. Points are calculated values and lines are B-spline curve fits.

Table 1 Photonic crystal defect slab waveguides considered in the calculations

Waveguide structures	*Suspended membrane*		*Oxidized lower cladding*		*Deeply etched*
	Deep undercut	*Shallow undercut*	*Thin oxide*	*Thick oxide*	
Layer description: Material (normalized thickness d/a)			Air (>3.0)		
	–	–	–	–	AlGaAs (1.0)
	GaAs (0.6)	GaAs (0.6)	GaAs (0.6)	GaAs (0.6)	GaAs (1.0)
	Air (3.0)	Air (1.0)	Al_xO_y (2.0)	Al_xO_y (5.0)	AlGaAs (6.0)
			GaAs substrate (>3.0)		

Refractive index of 3.4 is assumed for GaAs, 3.0 for AlGaAs, 1.6 for Al_xO_y, and 1.0 for air.

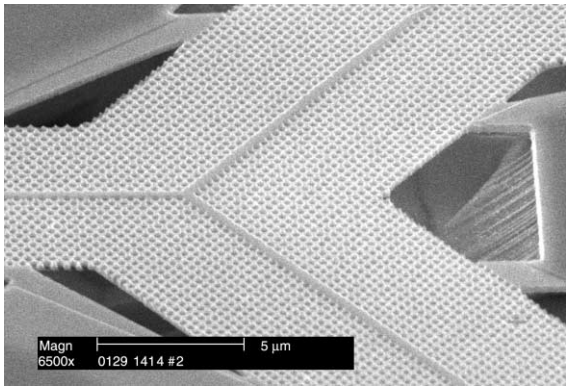


Figure 10 The scanning electron micrograph of a photonic crystal Y-branch fabricated on InP material system.

way through the lower cladding layer, with an index of 3.0, on a high index substrate. The defect modes lying within the in-plane photonic bandgap are all above the cladding light line in this case. Propagation losses are predicted to increase by over two orders of magnitude from the structure with very low index cladding layers shown in Figure 6 to the high index lower cladding structure of Figure 8.

Figure 9a and b shows the results of a fully three-dimensional calculation for the propagation loss of the three structure shown in Figures 6–8. The details of these structures are shown in Table 1. This calculated loss is due to coupling between the guided mode and the radiation modes of the high index substrate. A wavelength scale is also included using $a = 420$ nm in Figure 9a and 450 nm in Figure 9b. The suspended membrane suffers the least vertical radiation loss among all of the waveguides considered and has a minimum loss around 0.2 cm^{-1}. Low-loss waveguides formed on high index cladding layers may be possible, but care should be taken to eliminate coupling between the guided modes and the radiation modes of the cladding. In this case, the design strategy for reducing the optical loss is to reduce the magnitude of the Fourier component of the electromagnetic field inside the light cone. It should also be noted that deeply-etched waveguides formed by removing multiple rows of holes have been demonstrated with significantly lower radiation loss than is predicted for the single line defect waveguides. However, these waveguides are multi-moded at all frequencies. It is also important to remember that the nanofabricated waveguides will suffer additional losses due to fabrication imperfections that were not included in the numerical model. Figure 10 shows a nanofabricated photonic crystal waveguide. The image is an electron micrograph of a y-branch component of a Mach–Zehnder interferometer formed in a two-dimensional photonic crystal waveguide patterned into a suspended InGaAsP membrane.

Finally, it is worth noting that the group velocity of photonic crystal waveguide modes, where the group velocity is given by the gradient of the dispersion surface:

$$v_g = \nabla_{\vec{\beta}}\omega \qquad [4]$$

is very small near the zone boundary. This can be seen from the dispersion relations shown in Figures 6–8. In fact, the slope of the dispersion relations is likely to be a parameter which can be engineered by careful waveguide design. This property may allow photonic crystal waveguides to find applications in optical processing functions such as delay lines.

See also

Photonic Crystals: Atomic Physics; Electromagnetic Theory.

Further Reading

Baba T (1997) Photonic crystals and microdisk cavities based on GaInAsP–InP system. *IEEE Journal of Selected Topics in Quantum Electronics* 3: 808–830.

Joannopoulos JD, Meade RD and Winn JD (1995) *Photonic Crystals: Molding the Flow of Light*. Princeton: Princeton University Press.

Mekis A, Chen JC, Kurland I, Fan S, Villeneuve PR and Joannopalos JD (1996) High transmission through sharp bends in photonic crystal waveguides. *Physics Review Letters* 77: 3787–3790.

Painter O, Lee RK, Yariv A, *et al.* (1999) Two-dimensional photonic crystal defect laser. *Science* 284: 1819–1821.

Sakoda K (2001) *Optical Properties of Photonic Crystals*. New York: Springer.

Slater JC (1967). *Insulators, Semiconductors, and Metals*, *Vol. 3*. New York: McGraw-Hill Book Company.

Srinivasan K and Painter O (2002) Momentum space design of high-Q photonic crystal optical cavities. *Optics Express* 10: 670–684.

Villeneuve PR, Fan S and Joannopoulos JD (1996) Microcavities in photonic crystals: mode symmetry, tenability, and coupling efficiency. *Physics Review B* 54: 7837–7842.

Yariv A and Yeh P (1984) *Optical Waves in Crystals: Propagation and Control of Laser Radiation*. New York: John Wiley & Sons.

Self-Assembled and Functionalized Photonic Crystals

S Bhandarkar, Alfred University, Alfred, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Photonic crystals mimic the role of the atomic structure of a semiconductor in manipulating electrons, in this case for light. Hence, called semiconductors of light, these crystals are not generally found in nature and have to be designed and fabricated. Research and development in this field is taking place at a feverish pace to both understand the various nuances of this concept, as well as to evolve new methods to fabricate such crystals in the laboratory and ultimately on a commercial scale. The driving force behind this is the tremendous potential of these systems. They promise to revolutionize the way light can be used for future devices: from forming perfect, omnidirectional mirrors to the smallest ever lasers, and ultimately, complex optical chips with a wide range of functionality arranged in three-dimensional (3D) circuits.

This article seeks to elaborate on a specific technique to make photonic crystals called colloidal self assembly. Here, we will review the fundamental concepts underlying photonic crystals and see how the above process can lead to a versatile platform in making such systems. We will discuss the basic aspects of this method and look at the various innovative manifestations that have been used to fabricate many types of photonic bandgap materials.

Photonic Bandgap

A key attribute of a semiconductor is the formation of a bandgap, a band of energy states that the electrons cannot acquire. They may exist in the so-called valence band where their energy levels are less than that of the bandgap. Or they may be energized to states above the bandgap, i.e., to the conduction band, but the bandgap is the forbidden region. The continuous band of states that constitutes this bandgap is a result of the long-range order imposed by the crystal structure of the semiconductor, where the atomic lattice presents a periodic potential to a propagating electron. The tremendous usefulness of semiconductors comes from their ability to perform switching and logic functions, which in turn is a result of controlling the availability of electrons above and below the bandgap.

The propagation of light as an electromagnetic wave is influenced by the dielectric properties of the medium. For a waveguide aligned in the z direction, the field is the sum of the two propagating transverse modes, each with the form $E(x) = \Phi(x, y)e^{ik_z z}$, where $\Phi(x, y)$ is the field distribution and k_z is the propagation constant. The wave equation for the transverse mode reduces to

$$\frac{\lambda^2}{4\pi^2}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\Phi(x, y) + n(x, y)^2\Phi(x, y) = (\bar{n})^2\Phi(x, y) \quad [1]$$

Here, $\bar{n}$ is referred to as the effective refractive index and is given by $\bar{n} = k_z\lambda/2\pi$.

Equation [1] has the same form as the Schrödinger equation, which describes the wave functions of quantum particles such as electrons. This equivalence helps us to understand the similarity between the attractive potential well structures that bind electrons to the wells created by a local increase in the refractive index that confines light into optical modes.

The creation of a structure with a 3D periodic change in the refractive index (Figure 1) opens a gap, due to Bragg scattering off the dielectric interfaces whereby a set of frequencies of light becomes forbidden to propagate in any direction – a photonic bandgap. To understand the formation of the

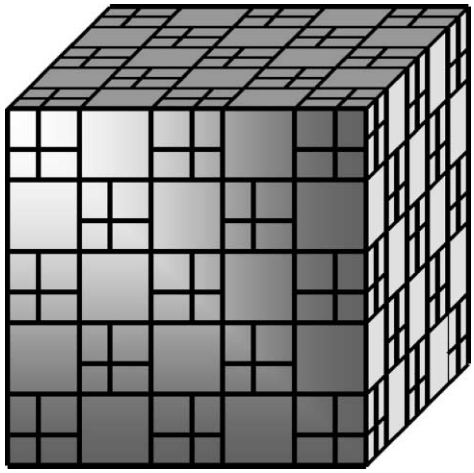


Figure 1 Schematic representation of a three-dimensional photonic crystal. The structure formed is a periodic variation in the dielectric properties of the matrix. Here, the two different dielectric regions are represented by □ and ⊞.

bandgap, one needs to consider the complex pattern of overlapping beams that results from the refractions and reflections from the numerous interfaces. This pattern is determined by the characteristics of light-waves (wavelength and direction of travel) and those of the periodic dielectric variation (pitch and magnitude of the contrast). Thus, for a select band of wavelengths, complete cancellation through destructive interference occurs in all directions.

One way to breach this restriction is by creating a defect or a disturbance in the periodicity, which would allow localized photonic states in the gap. This defect can be a point defect to form a microcavity, or have a 1D trace that behaves as a waveguide or a planar defect. Thus, a photonic crystal with a specific defect structure can be used to manipulate light in unique ways that can be used for a variety of applications, including lasers, filters, switches, waveguides, and for 3D integrated optics circuits. These structures are also interesting because they have been predicted to exhibit enhanced optical nonlinearities.

Photonic Crystals

The lattice constant for the photonic crystal must be comparable to the wavelength of light, typically in the range of 0.5 to 2 μm. This range is three orders of magnitude greater than the lattice constants found in solids and hence photonic crystals need to be synthesized. A common example of such a structure is a diffraction grating in a waveguide, which serves as a 1D photonic crystal. There are also examples of photonics crystal structures found in nature, such as the iridescent colors seen in certain butterflies and opal gemstones. The structure of opal is a 3D photonic crystal consisting of small spherical particles of hydrated silica (SiO_2). These spheres tend to be irregular in size but in rare precious opals, they are monosized. Diffraction from these spheres leads to the remarkable colors seen in opals.

Fabrication of 3D photonic crystal structures, with two materials with sufficiently different dielectric properties, has been extensively investigated over the past several years. Given the requirement of a tightly controlled periodicity, one obvious means of doing this is to use the microlithographic patterning processes that are so well developed for very large-scale integration (VLSI). Since these are planar and not volumetric techniques, several iterations of layer-by-layer deposition, lithography, and etching are necessary. A popular second method is that of colloidal self assembly.

The Self-Assembly Technique

The term 'self assembly' is usually used to describe the act of spontaneous and controlled agglomeration of some basic moieties. These basic units could be molecules, supramolecules, nanoclusters, or particles. It is similar to what is seen in living systems, where organic and inorganic materials are assembled into useful configurations mediated by various proteins. Since there is no intrinsic restriction on the number of such building blocks that can spontaneously organize into periodic structures, this process is readily scalable to large dimensions, and is the main attraction of this process.

The self-assembly technique is well suited for making periodic structures such as photonic crystals. Starting with particles of the appropriately uniform size (in the range 0.5 to 2 μm), one can conceive of organizing them into a periodic 3D array. Particles, rather than molecules or nanoclusters, are the preferred starting materials for photonic crystals because their larger size provides periodicity on the length-scale of optical wavelengths. Large block copolymer molecules have also been used instead, but they are limited in their scope.

The following sections will discuss the salient details of using self assembly to form photonic crystals. The sequence of the discussion follows the path increasing process complexity. To begin, we will examine the simplest of photonic crystals made by neatly packing uniform particles called a colloidal lattice. Given the limited range of monosized particles available, such lattices are inadequate as useful photonic crystals but can be used as templates into

which other materials can be incorporated. We shall discuss one particularly attractive incorporation technique, namely electrochemical deposition, in more detail. Finally, since the process is dominated by surface interactions, diverse classes of particles can be generated by simply functionalizing the particle surface. Looking ahead to the future, this is an important aspect that can vastly increase the capability of a colloidal assembly.

Formation of Colloidal Lattices

Colloidal crystallization has been studied for several decades but has received renewed attention due to both the availability of more uniform powders as well as new applications such as photonic crystals. Fine particles are unwieldy in their dry state due to high solid–gas interfacial energy and are typically dispersed in a liquid. Particle concentration is kept low to avoid agglomeration. Further stabilization of the particles may be achieved by either charging the particles' surface (electrostatic stabilization) or by adsorbing a neutral shielding molecule (steric stabilization). A preferred approach is to induce monosized spherical particles to settle into a tightly packed lattice. The settling rate, v, of particles of diameter d_p is given by the well-known Stokes equation:

$$v = \frac{d_p^2 a \Delta\rho}{18\mu} \qquad [2]$$

where a is the acceleration due to gravity (or a centrifugal field), $\Delta\rho$ is the density difference between the particle and the liquid, and μ is the viscosity of the liquid. Whereas settling is a popular means of self assembling, other methods such as centrifugation, filtering, pressing, or dip coating can be used (Figure 2). For charged particles, electrophoresis, where an electric field induces flow, is applicable.

The typical lattice types in these domains are hexagonal close-packed planes (hcp) as shown in Figure 3 or face-centered cubic (fcc) structures (Figure 4). In fcc crystals, the layer stacking follows the order ABCABC; in hcp it is ABAB. In both stacking sequences, hard spheres occupy a volume fraction of 0.7404 and can be shown to have little entropic differences. However, it is experimentally observed that the preferred ordering of spheres is the fcc close-packed arrangement, which is also the case in high-quality opals.

This method clearly has the advantage of being a simple, cost-effective platform to make 3D photonic crystals that can be readily scaled to various levels. There are no repetitive steps as would be the case in layer-wise lithographic pattering. However, there are still some outstanding challenges. The primary one is the difficulty in ensuring that the organization of the particles does not incur the formation of unwanted defects and domains. Also, introduction of intentional defects such as waveguides is not trivial, although this is being demonstrated in different laboratories. Another shortcoming is the

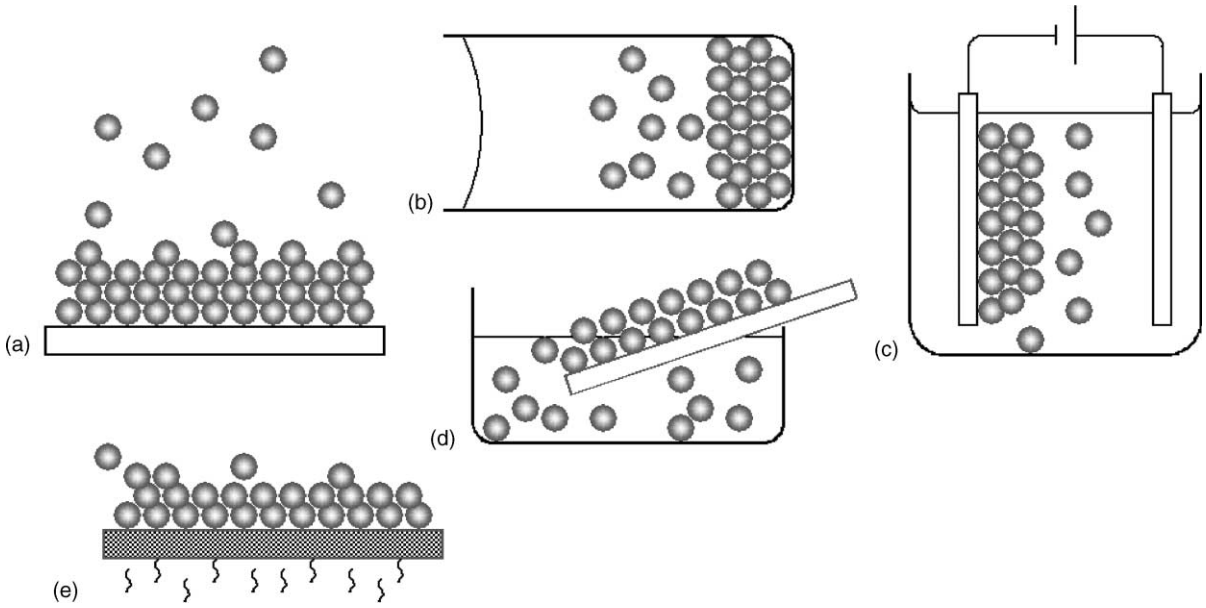


Figure 2 Diagram showing some of the techniques to use self assembly of spherical particles to make colloidal lattices. These include settling (a); centrifuging (b); electrophoresis (c); dip-coating (d); and filter-pressing (e).

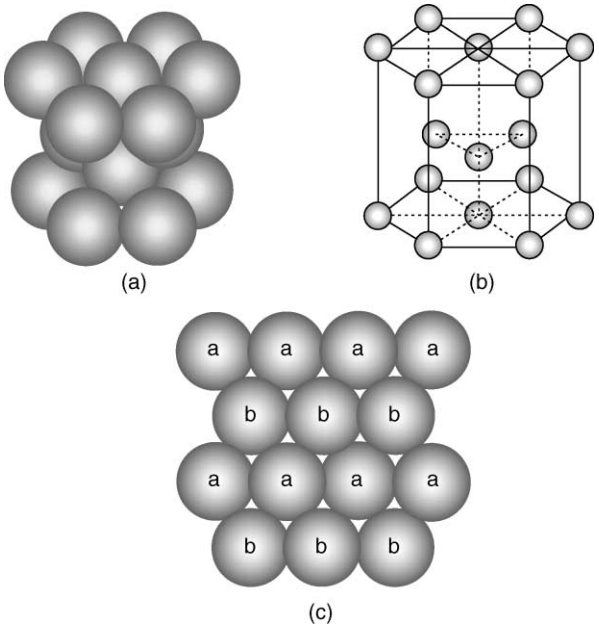


Figure 3 The arrangement of spheres in a hexagonal closed packed (hcp) arrangement (a). The unit cell is shown in (b) whereas the stacking *abab* sequence can be seen in (c).

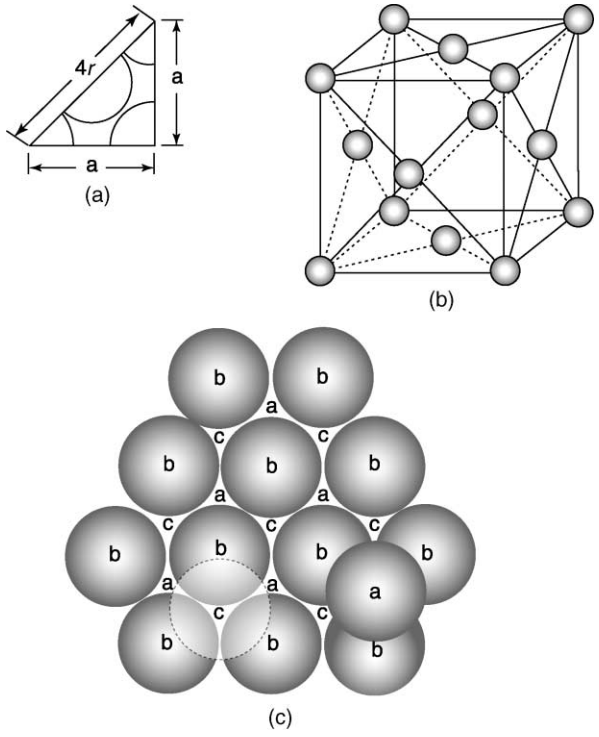


Figure 4 The arrangement of spheres in a face centered cubic (fcc) arrangement with spheres of radius *r* (a). The unit cell is shown in (b). The stacking follows an *abcabc* order. The *a* and *c* sites relative to spheres populating the *b* sites are seen in (c).

unavailability of uniformly sized particles of various different chemical compositions.

The formation of monosized spherical particles is a complex problem in its own right. Usually, such particles can be made from specific polymeric materials such as polystyrene or latex, whereas SiO_2 is the best-known inorganic system. These particles are not perfectly uniform but the typical deviations from the mean size of a few percent are acceptable for most applications. Given the attractiveness of photonic crystals for a wide array of applications in optics, research in this field is burgeoning and more breakthroughs are bound to come. Already first-generation photonic bandgap devices, such as fiber gratings or photonic crystal optical fiber, are commercially available.

Colloidal Lattices as Photonic Crystals

Due to the limited range of composition of the particles available, the particle-to-air refractive index contrast provided by straightforward colloidal lattices is small. This results in low diffraction efficiencies and hence colloidal lattices do not have photonic bandgap in all directions. However, these structures can then be used as templates for various additional modifications, as seen later.

On the other hand, a simple colloidal lattice can be more easily modeled. An analytical solution can be achieved for this system using dynamical diffraction theory for a low diffraction efficiency approximation. For a close-packed stacking of spheres in a fcc array, there exists a minimum in the transmission band at a wavelength λ_c which is given by the Bragg–Snell equation:

$$\lambda_c = 2d_s\sqrt{n_{\text{eff}}^2 - \sin^2\theta} \qquad [3]$$

where d_s is the d-spacing between the planes in the $\langle 111 \rangle$ direction, n_{eff} is the effective refractive index of the assembly and θ is the angle between the incident beam and normal to the (111) plane. This minimum in the transmission is referred to as the primary stop-band.

Photonic crystals are characterized using a Michelson or Mach Zehnder interferometer to obtain the phase and amplitude of the transmission (or reflection) as a function of the wavelength of light (Figures 5 and 6). Ideally, the whole band structure can be calculated from these data. In addition to λ_c, the close-packed structures also exhibit a transmission threshold at a wavelength λ_t, as seen in Figure 5. Both λ_c and λ_t depend on the particle diameter and increase with particle size. In a close-packed structure,

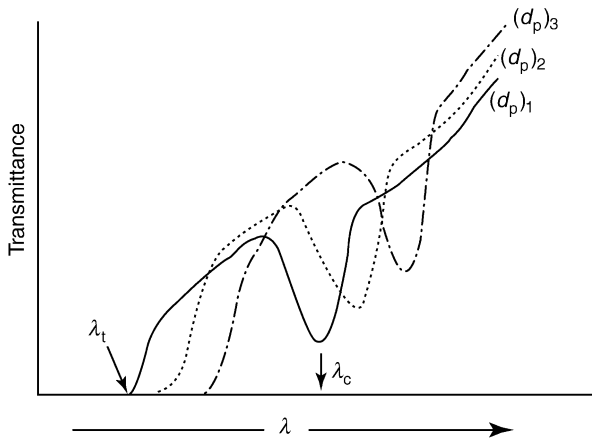


Figure 5 The typical profiles observed from transmission measurements on a photonic crystal. The stopband wavelength (λ_c) increases with d_p, where $(d_p)_1 > (d_p)_2 > (d_p)_3$. The polarization of light is fixed at *s* or *p* polarized.

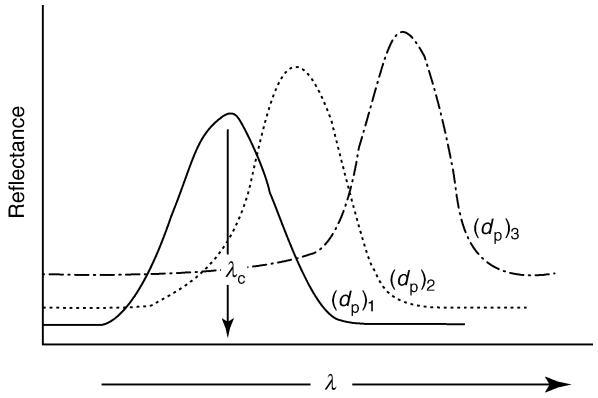


Figure 6 The typical profiles of the reflectance peaks corresponding to the transmission plot seen in Figure 5.

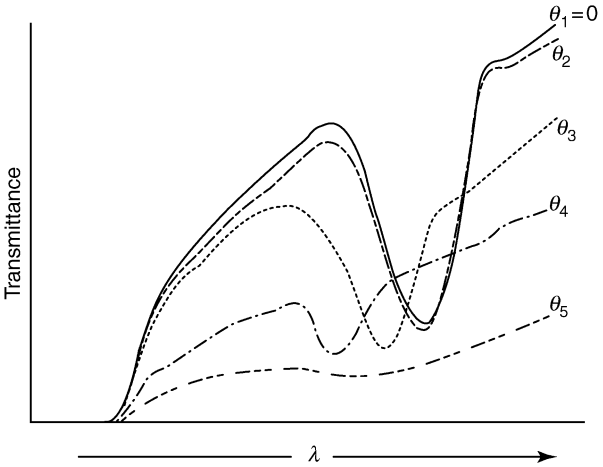


Figure 7 The transmittance profiles as a function of the angle of incidence of light of a fixed polarization. Typical values of θ would be $\theta_1 = 0$, $\theta_2 = 10°$, $\theta_3 = 30°$, $\theta_4 = 50°$, $\theta_5 = 60°$.

the lattice *d*-spacing d_s is related to the particle diameter d_p as $d_s = d_p\sqrt{2/3}$. So when the incident light is orthogonal to the structure, eqn [3] can be simplified to:

$$\lambda_c = 2\sqrt{2/3}(d_p n_{eff}) \qquad [4]$$

Thus, n_{eff} can be obtained from the slope of a plot of the stop-band wavelength λ_c as a function of the particle diameter d_p. The angular dependence of the reflectance spectra is illustrated in Figure 7. These results depend upon the polarization state of the incident light and hence need to be obtained separately for *s* and *p* polarized light.

More Complex Colloidal Lattices

While the use of SiO_2 and polymer microspheres is ubiquitous, they are optically passive materials. New methods are being developed for making uniform spherical particles from other more optically active systems. The zinc oxide (ZnO) system is one such example. The higher refractive index (~2.2) and the light emission ability of this system make it particularly interesting. Monosized particles in the size range of 100 to 600 nm can be made using a two-stage reaction method, which allows extra control over the particle formation. Other optically active systems include rare-earth doped SiO_2, dye doped PMMA spheres, or composite spheres comprised of core–shell configuration with SiO_2 or zinc sulfide (ZnS) cores and gold (Au) or SiO_2 shells, respectively.

Whereas the above discussion focuses on monosized particles, binary crystals can be made using spheres of two different sizes. The strategy here is to vary the particle size during the deposition of every alternate layer. For instance, a first layer of closely packed spheres is followed by the deposition of layers of smaller particles. The resultant crystal structure depends primarily on the concentration of the smaller particles in the dispersion. As seen in Figure 8, the second layer can vary from having three to six particles around every underlying larger particle. Thus the large-small (LS) stoichiometry of these binary crystals can be varied from LS to LS_2 to LS_3 by increasing the concentration of the smaller particles in the suspension phase. Such crystals open up interesting new avenues for complex photonic crystals.

Colloidal Lattices as Templates

By using the colloidal lattice as porous host for a second material, a greater range of photonic crystals can be synthesized. Here, the role of the colloidal array

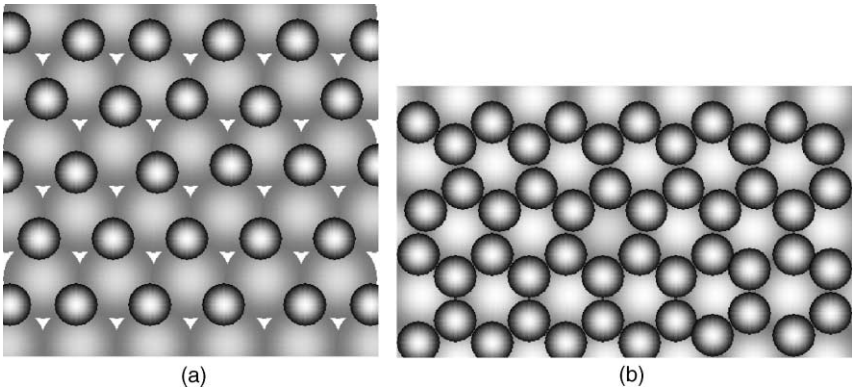


Figure 8 Binary crystals formed using spheres of two different sizes. The large-small (LS) ratio in (a) is 1:1 whereas in (b) it is 1:2.

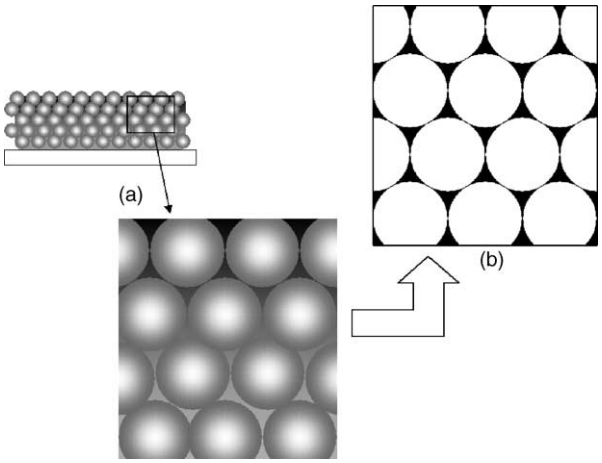


Figure 9 Schematic representation of the templating process using infiltration (a). Upon removal of the particles, a macroporous photonic crystal is formed (b). This is the inverted opal configuration.

is simply as a scaffolding structure. The second material to be infiltrated may be either a liquid (in the form of a dispersion, a solution, or a molten liquid) or a vapor. Both eventually deposit a solid material inside the template. Subsequent selective removal of the template yields a 3D photonic crystal in the form of an inverted opal (Figure 9). This is one of the best and most versatile means of achieving high index contrast, which is required for true photonic bandgap behavior.

It must be noted that to use self-assembled particle arrays as a template, the whole structure has to be often heat treated to confer greater mechanical strength. As assembled, the particles are normally attached by short-range Van der Waals forces and hydrogen bonding, which in the case of SiO_2 is silanol (Si–OH) mediated. Both of these are relatively weak bonds. By heating the assembly, a greater number of siloxane (Si–O) bonds can be formed which additionally strengthens and stabilizes the particulate network. This is important to retain the periodic structure of the assembly during the exposure to high stresses during the infiltration step, for instance. At temperatures around 750 °C, there is significant generation of siloxane bonds but minimal sintering so that the lattice does not change dimensions and retains its original open porosity. With latex particles, the annealing is done just above the glass transition temperature T_g of the polymer for no more than a critical period of time. Longer treatments can destroy the ordering of the spheres.

A template thus strengthened by heat treatment is now ready for infiltration. For example, a template comprised of polystyrene spheres can be infiltrated with a hydrolyzed sol of titanium alkoxide to ultimately form titania in the interstitial regions. Upon standing, the alkoxide sol gels in the interstitial regions leading to a polymeric network called a xerogel. This entire composite can then be further heated to high temperatures (> 300 °C). All organics are removed by decomposition and combustion so as to leave behind a titania matrix with periodic voids (in place of the polystyrene spheres). The templating spheres may also be dissolved, rather than combusted, in solvents such as toluene. The refractive index of titania depends on the crystalline phase that results after heat treatment – anatase is ~2.5 while rutile is ~2.9. High indices can also be achieved with semiconducting materials such as selenium (Se) with an index of ~2.5 or cadmium selenide (CdSe), ~2.9. These chalcogen-derived materials have low absorption in the visible and near infrared, so they are suitable candidates.

When SiO_2 particles are used for templating, they are eventually removed by etching, typically with dilute, buffered hydrofluoric acid at low temperatures. Given the thermal and chemical stability of the SiO_2 particles, molten Se (or other low melting

metals) can be used to rapidly fill the interstitial space. Vapor phase deposition has also been used to fill the porous templates. Examples include the backfilling with silicon (Si) using chemical vapor deposition (CVD), or indium phosphide (InP) using metal-organic chemical vapor deposition (MOCVD).

A variety of structured, macroporous carbon matrices has been made with remarkable optical properties. The templates are artificial opals assembled using SiO_2 microspheres, which are filled with carbon by either: (a) infiltrating with phenolic resin followed by anaerobic pyrolysis of the resin to glassy carbon; (b) CVD of graphitic carbon; and (c) growing diamond-seeded CVD graphite from a plasma. The resultant carbonaceous 'inverse opals' are highly conductive, show intense opalescence due to the ordered arrays of holes, and have a photonic bandgap in the infrared region.

Electrochemical Deposition

One particularly effective backfilling technique is electrodeposition, which has been successfully used to deposit most of the different classes of materials – semiconductors, metals, and polymers. Here, a conducting substrate serves as the appropriate electrode where deposition occurs. By placing a colloidal lattice template on this electrode, the electrochemical deposit is used to fill the interstitial voids in the template (Figure 10). This method has been successfully applied for a host of different metals such as gold (Au), nickel (Ni), platinum (Pt), and alloys such as tin–cobalt (Sn–Co). Likewise, conducting polymers can be efficiently deposited by this method. Although these polymers suffer from inherently low indices, they have favorable electro-optic properties that can be tuned by controlling the composition. In the case of II–VI semiconductors, galvanostatic or potentiostatic deposition can be used: for instance, interstitial CdSe has been deposited (Figure 10) using an acidic solution of electrolytes such as cadmium sulfate ($CdSO_4$) and selenium dioxide (SeO_2).

Electrochemically deposited structures have minimal porosity leading to densities close to bulk density. Hence, they have a higher refractive index and are generally more robust than those made by simple infiltration. Also, they exhibit little change in dimensions and a lower tendency toward cracking during the various physicochemical processes such as drying and sintering that occur en route to forming the photonic crystal. Hence, it is a preferred technique in many cases.

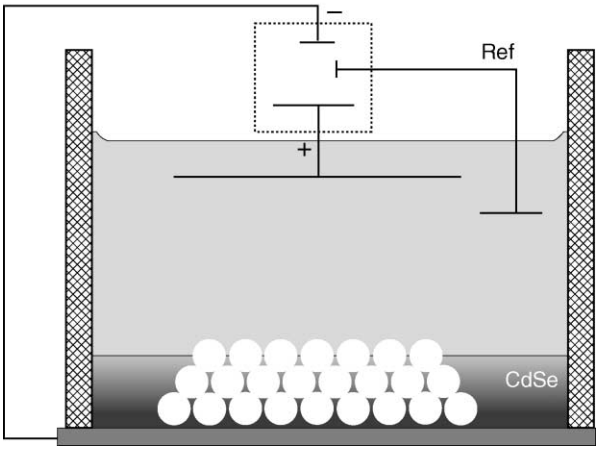


Figure 10 An electrochemical cell for deposition of CdSe from an electrolytic solution on a conducting electrode. The solution contains sulfuric acid (H_2SO_4), $CdSO_4$, and SeO_2. A potentiostat is used to generate the electric fields.

Functionalization

The self-assembly process, whether used directly or as a template to synthesize a photonic crystal, is dominated by surface interactions between particles starting from the organization process through to the thermal stabilization step. These interactions can be tailored by specially treating the surface of the particles with functional groups. Functionalized colloidal templates can enhance the infiltration of the high index material and its properties such as smoothness, density, or optical nonlinearity. Latex spheres can be surface functionalized with sulfate and carboxylic acid groups to increase their surface charging capability, which can lead to more robust self-organization. These spheres can also be coated with a layer of SiOH, which gives them a SiO_2-like appearance on the surface and allows them to be used in conjunction with SiO_2 microspheres. The infiltration of metal (such as Au) nanoparticles into a template crystal of SiO_2 microspheres can be significantly enhanced by functionalizing the SiO_2 surface. This can be achieved using long-chain organic molecules with polar groups on one end that bond with silanols on the SiO_2 surface and thiol groups on the other that strongly latch onto the gold nanoparticles. Simulations of light interaction with ordered three-dimensional arrays of these metallodielectric spheres, show that they are promising starting points for the construction of full photonic bandgap materials in the visible part of the optical spectrum. Also, in these systems, the plasmon resonance of the Au shell particles can be controlled over the visible and infrared region by varying the thickness of the Au shell.

The site-selective attachment ability of biological moieties can also be used effectively here by attaching them onto particle surfaces. DNA oligomers are well known to form conjugate specific bonds – adenine to thymine and guanine to cytosine. Thus particles functionalized with these oligomers can be made to organize in specific sequences that are differentiated, depending on the surface group. These particles can have variable sizes or compositions but their surface properties are controlled largely by the attached biomolecules. Likewise, the antigen–antibody attachment ability can be exploited to improve surface bonding.

Positioning Defects in Photonic Crystals

Much of the work on photonic crystals involves the making of a long-range periodic microstructure with high dielectric contrast because the underlying materials engineering aspects are very challenging. However, the usefulness of photonic crystals depends on the successful placement of defects in precise locations within the crystal. The placement of a point defect leads to a photonic microcavity and that of a line defect, a waveguide for propagation of a light-wave (Figure 11). However, there are certain unique advantages to this style of propagation, such as the possibility of bending light around sharp corners. Line defects in the form of air cores have been introduced in a photonic crystal using a combination of lithography and self assembly. First, ridges are defined in a photoresist on top of the substrate by optical lithography (Figure 12). Upon forming a colloidal lattice on top, the ridges are converted to linear voids by removing the photoresist using combustion or etching. Another technique is to infiltrate a photopolymerizable monomer into a colloidal crystal and preferentially polymerize the liquid by focusing a light beam inside the crystal. Thus the locus of the focused beam defines the waveguiding region that is formed upon removal of the unpolymerized liquid.

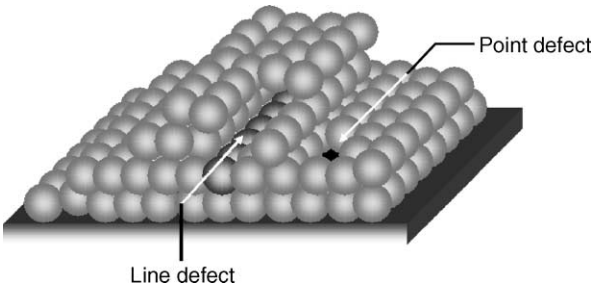


Figure 11 Two common defects in colloidal lattices.

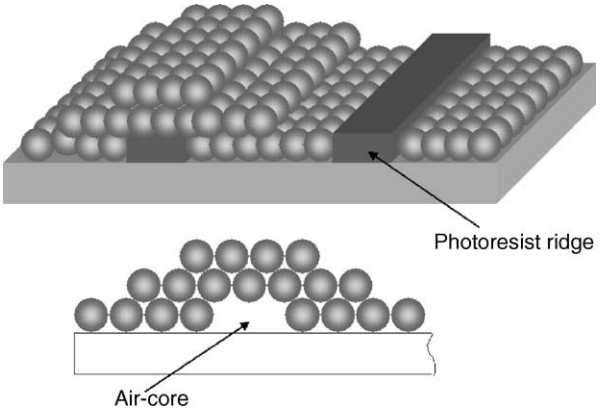


Figure 12 Formation of intentional defect using a lithographic process to define an air-core. This structure can also be used as a template for backfilling a high dielectric material.

Tunable Photonic Crystals

The holy grail in many photonic devices is tunability, and photonic crystals are no exception. Given the wide spectral range that constitutes the domain of light, the ability of devices to switch or translate their performance to another wavelength through an external stimulus is indeed very attractive. The two parameters that can be externally tuned are the dielectric contrast, which would result in a change in the n_{eff} of the crystal and the periodicity of the grating. One technique to tune the former is to form an interference pattern that organizes liquid crystals dispersed in a polymer matrix into droplets to form the photonic crystal. Applying an electric field that changes the refractive index of the droplets can therefore alter the transmission spectrum of the photonic crystal.

To control the lattice parameter, the size of or the spacing between the spheres need to be varied. This concept has also been demonstrated using self-assembled hydrogel-based nanoparticles, which are made of a thermo-responsive material such as poly-*N*-isopropylacrylamide. These particles are soft and conformable and their sizes can be tuned depending on the water content of the system. This ordered matrix is in the form of a gel that can be made to swell and compress upon changing the temperature as water is exchanged between the gel and the ambient. Wavelength tunability of one nanometer has been achieved this way. Tunability has also been sought using methods such as straining the lattice. Calculations show that a 2 or 3% shear strain ought to create a 52–73% distortion of the bandgap. In one novel approach, SiO_2 or polystyrene microspheres were strongly surface charged and ordered with a less than dense packing within a poly-acrylamide

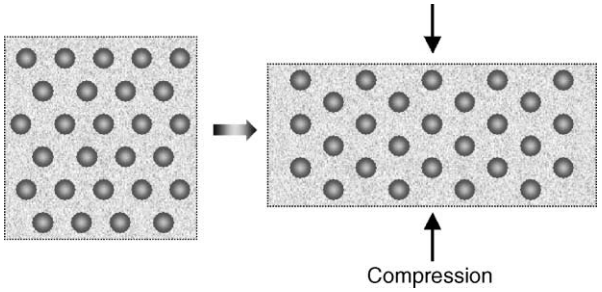


Figure 13 Tunable photonic crystal. Spherical particles are assembled in a compressible media like a hydrogel, which deforms upon applying a stress. The deformation can be used to tune the lattice parameter and hence the optical properties of the photonic crystal.

hydrogel housing (Figure 13). Due to the strong charging the arrangement tends to be body centered cubic (bcc) with the lattice spacing that depends on the concentration of the particles. When the hydrogel is deformed by applying mechanical stress, the diffraction peaks were shown to be tunable over almost the whole visible light wavelength spectrum.

Photonic crystals made from self-organized hydrogels have been used as sensors. Here, they are functionalized with chemical recognition agents that cause the gels to swell upon binding to the analyte. The resultant change in the lattice parameters causes a measurable change in the color. Specific examples include the detection of the toxic organophosphate, parathion, using acetylcholinesterase or of glucose using boronic acid and polyethylene glycol. Utilizing the same principle, even metal ions such as copper can be detected using a hydroxyquinoline-based hydrogel. These serve as a good illustration of the multifaceted potential of photonic crystals.

Summary

Photonic crystals have aroused tremendous interest with their promise to lift the field of photonics to a new level of capability. These crystals are conceptually analogous to electronic semiconductors in their function and hence are called photonic bandgap materials. The big challenge lies in synthesizing them artificially without defects. One of the few methods to synthesize these crystals is by self assembly of micron-sized building units. Self assembly leads to formation of a uniform array of monodispersed particles, typically silica or latex. This process can be further embellished by permeating a second, high-index material using the particulate assembly as a template. Useful applications of photonic crystal require them to have specifically positioned defects and tunability. These aspects, which currently constitute the state-of-the-art research in this field, promise to deliver unique devices and applications.

List of Units and Nomenclature

a	Acceleration (typically due to gravitational or applied centrifugal force)
d_p	Diameter of the particle
d_s	d-spacing or distance between lattice planes
E	Electric field
k_z	Propagation constant for travel in z direction
n	Refractive index of the medium
$\bar{n}$	Modal effective refractive index
n_{eff}	Effective refractive index of the colloidal assembly
v	Settling velocity
$\Delta\rho$	Density difference between the particle and the fluid in which it is suspended
θ	Angle between the incident beam and normal to the lattice plane in question
λ	Wavelength of the lightwave
λ_c	Stopband wavelength
λ_t	Threshold wavelength for the photonic crystal
μ	Viscosity of the suspending fluid
Φ	Electric field distribution

See also

Photonic Crystals: Atomic Physics; Electromagnetic Theory; Photonic Crystal Lasers, Cavities and Waveguide.

Further Reading

Born M and Wolf E (1981) *Principles of Optics: Electromagnetic Theory of Propagation Interference and Diffraction of Light*. Oxford, UK: Pergamon Press.

Braun PV and Wiltzius P (2002) Macroporous materials – electrochemically grown photonic crystals. *Current Opinion in Colloid and Interface Science* 7: 116–123.

Graf C and Van Blaaderen A (2002) Metallodielectric colloidal core-shell particles for photonic applications. *Langmuir* 18(2): 524–534.

Iwayama Y, Yamanaka J, Takiguchi Y, *et al.* (2003) Optically tunable gelled photonic crystal covering almost the entire visible light wavelength region. *Langmuir* 19(4): 977–980.

Joannopoulos JD, Meade RD and Winn JN (1995) *Photonic Crystals*. Princeton, NJ: Princeton University Press.

Johnson SG and Joannopoulos JD (2002) *Photonic Crystals: The Road from Theory to Practice*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kazuaki S (2001) *Optical Properties of Photonic Crystals*. New York: Springer-Verlag.

Koenderink AF, Johnson PM, Galisteo López JF and Vos WL (2002) Three-dimensional photonic crystals as a cage for light. *Comptes Rendus Physique* 3: 67–77.

Soukoulis CM (ed.) (2001) *Photonic Crystals and Light Localization in the 21st Century*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Stein A (2001) Sphere templating methods for periodic porous solids. *Macroporous and Mesoporous Materials* 44–45: 227–239.

Velev OD and Lenhoff AM (2000) Colloidal crystals as templates for porous materials. *Current Opinion in Colloid and Interface Science* 5: 56–63.

Xia Y, Gates B, Yin Y and Lu Y (2000) Monodispersed colloidal spheres: old materials with new applications. *Advanced Materials* 12(10): 693–713.

Yablonovitch E (1987) Inhibited spontaneous emission in solid-state physics and electronics. *Physics Review Letters* 58(20): 2059–2062.

PHYSICAL APPLICATIONS OF LASERS

Contents

Free-Electron Lasers in Physics
Industrial Applications
Sum-Frequency Generation at Surfaces

Free-Electron Lasers in Physics

T Dekorsy, Forschungszentrum Rossendorf, Dresden, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The topic of this article is a survey on the main applications of free-electron lasers (FELs) in physical sciences. The merits of FELs are discussed with respect to other available laser sources. For the wavelength ranges, where FELs contributed significantly to advances in physical sciences, specific examples of the research with FELs are given.

The basic physics of FELs, their historic development, and the state of the art of these laser sources are described in detail in this Encyclopedia (*see* **Lasers:** Free Electron Lasers). The present article covers the applications of FELs in physical sciences, but excludes all phenomena involved in the physics of FELs themselves. From the application point of view, one should regard an FEL as any other laser, as a source of coherent electromagnetic radiation with high power, narrow bandwidth, and a diffraction limited beam. Thus, the radiation produced by an FEL significantly differs from that generated by synchrotrons (*see* **Incoherent Sources:** Synchrotrons). The specifications of the FEL, such as wavelength, temporal structure of the output intensity, and peak power, determine the possible applications of the laser. One might now ask, why it is worthwhile to write – and read – an article on its own devoted to the applications of FELs. The answer to this question is the uniqueness of the radiation produced by FELs, which enables certain classes of experiments and applications in specific wavelength ranges not covered by other laser sources, such as gas, solid state, or semiconductor lasers.

Since the construction and operation costs of an FEL facility are usually orders of magnitude higher than those for the equipment of a conventional laser spectroscopy laboratory, FEL facilities have been built, and are planned for the future, in different places all over the world as dedicated-user facilities. These user facilities allow researchers from universities and industry to exploit the unique properties of FEL radiation for their research and provide the basic infrastructure and support for setting up the experiments. Usually potential users apply using a standard procedure for beam time at an FEL by submitting a short proposal to the facility. The proposals are judged by a panel which distributes the available beam time to the different users. So researchers active in the field of optics should keep in mind that these facilities exist and are available to anyone who has specific needs which can be fulfilled by an FEL facility. A list of all major FEL facilities is given in Table 1.

Table 1 FEL user facilities. Facilities marked with an asterisk are under construction at the time of publication

Facility name	*Wavelength ranges*
iFEL (Osaka, Japan)	230 nm–1.2 μm 1–6 μm, 5–22 μm 20–60 μm, 50–100 μm
Duke University Free Electron Laser Laboratory (US)	193–400 nm 2–9 μm
Vanderbilt University Free Electron Laser Center (US)	2.1–9.8 μm
FELIX–FOM (Rijnhuizen, Netherlands)	4.5–250 μm
Stanford Picosecond FEL Center (US)	3–15 μm 15–65 μm
CLIO–LURE (Orsay, France)	3–90 μm
Jefferson Lab (US)	3–6.2 μm
FEL–SUT–Science University of Tokyo (Japan)	5–16 μm
UCSB Center for Terahertz Science and Technology (US)	63–338 μm 338 μm–2.5 mm
ENEA Compact FEL (Frascati, Italy)	2–3.5 mm
FELBE (Dresden, Germany)*	5–150 μm
SCSS SPring-8 Compact SASE Source (Japan)*	X-ray
Institute for Plasma Research (India)*	sub-mm
BESSY FEL (Germany)*	deep UV–X-ray
Brookhaven–NSLS SDL DUV–FEL (US)*	deep UV
University of Maryland–MIRFEL (US)*	far-infrared
University of Hawaii (US)*	UV–far-infrared
SLAC SSRL X-ray FEL–LCLS (US)*	X-ray
TESLA X-ray FEL, DESY (Germany)*	X-ray
Daresbury 4GLS (UK)*	IR, X-ray
MIT Bates Lab X-ray FEL (US)*	X-ray
INFN SPARX X-FEL (Italy)*	X-ray
Tel Aviv University (Israel)*	far-infrared
Center for Advanced Technology (India)*	far-infrared

The basic principles of an FEL are briefly summarized here, while for an in-depth description the reader is referred to the article on free electron lasers (*see* **Lasers:** Free Electron Lasers). As with any other laser, the FEL consists of a laser cavity and a gain medium. The gain media are free electrons, which travel with a velocity close to the speed of light in vacuum through a periodic array of magnets called an undulator or wiggler. The magnets are arranged in a way that the electrons feel an alternating magnetic field. Due to the Lorentz force, the electrons perform an undulating or wiggling motion perpendicular to the magnetic field, which gives rise to spontaneous electromagnetic radiation into the forward direction. The radiation, emitted in due course of the passage of the electrons through the undulator, superposes constructively. The FEL may operate in such a way that the gain from one passage of an electron bunch is high enough to produce a coherent light pulse. This regime is known as self-amplified spontaneous emission (SASE) regime which is especially desirable for short wavelength operation in the UV and X-ray region, where the manufacturing of adequate resonator optics is difficult. At longer wavelengths, a laser cavity provides the feedback for the build-up of the coherent light field. The laser starts lasing when the amplification exceeds the losses of the cavity, including the light intensity coupled out of the cavity.

One of the distinct properties of FELs, in comparison to other laser sources, is their continuous tunability. The emission wavelength of an FEL is given by the following relation (*see* **Lasers:** Free Electron Lasers):

$$\lambda = \frac{d}{2\gamma^2}\left(1 + \frac{K^2}{2}\right) \qquad [1]$$

with d the period of the magnet array, the γ-parameter is the relativistic factor related to the electron velocity u via $\gamma = 1/(1 - u^2/c^2)^{1/2}$, and the K-parameter is given by d, and the magnetic field B as $K = 0.934 \times B[\mathrm{T}] \times d[\mathrm{cm}]$. From eqn [1] it becomes obvious that the emission wavelength of the FEL can be continuously tuned via the electron energy and the magnetic field. This is in distinct contrast to other laser sources, where the emission wavelength is fixed by the energy separation of electronic levels in atoms or by the bandgap energy of semiconductors.

The main advantage of FELs concerns the high power, both average and peak, which can be obtained. In addition, an FEL provides short optical pulses due to the fact that the electron beam passing through the undulator consists of short electron bunches. The temporal width of these electron bunches can be below one picosecond (10^{-12} s), so that optical pulses in the sub-picosecond to picosecond range are obtained. These pulses are extremely adjuvant for performing nonlinear optical experiments (*see* **Spectroscopy:** Nonlinear Laser Spectroscopy) and pump-probe experiments with high temporal resolution (*see* **Chemical Applications of Lasers:** Pump and Probe Studies of Femtosecond Kinetics). Typically the optical pulses are Fourier transform limited, i.e., the product of the spectral bandwidth $\Delta\nu$ times the pulse duration $\Delta\tau$ equals a constant, which depends on the specific shape of the pulse. For a pulse of Gaussian shape in frequency and time, this so called time-bandwidth product is given as $\Delta\nu\,\Delta\tau \geq 0.441$. From this equation the minimum spectral width (temporal duration) can be calculated for a given temporal duration (spectral width) of the laser pulse. When the electron beam driving the FEL is produced by a superconducting radio frequency linear electron accelerator, the repetition rate of the pulses is in the

range of some tens of MHz. The energy per optical pulse is obtained by dividing the average power of the FEL through the repetition rate. The peak power within an optical pulse is obtained by dividing the energy per pulse through the pulse duration, which can be in the picosecond to sub-picosecond range. Typical values for FELs are energies per pulse in the μJ to mJ range, pulse lengths from some 100 femtoseconds to a few picoseconds, and accordingly peak powers in the range of kW to GW. Depending on the type of electron accelerator employed, the pulse structure of an FEL may have further peculiarities. Superconducting radio-frequency accelerators provide a quasi continuous-wave (CW) train of electron pulses, thus enabling the FEL to emit a continuous optical pulse train. Electrostatic accelerators provide macropulses of electrons at some Hz repetition rates with a duration of several μs. These macropulses consist of micropulses with MHz repetition rates and picosecond or sub-picosecond duration. Hence the FEL output has the same temporal characteristics with the FEL turning on at the beginning of each macropulse.

Another important development is the synchronization of table-top short-pulse laser systems to FELs. This is accomplished by stabilizing the repetition rate of the table-top laser system via a movable intracavity mirror accurately to the repetition rate (or a higher harmonic of the FEL) of the FEL. Using a high-frequency electronic phase-locked loop, a timing jitter, down to 500 femtoseconds between the FEL pulses and those from a mode-locked titanium-sapphire laser, have been obtained. Nonlinear frequency conversion of the near-infrared pulses from the synchronized titanium-sapphire allows one to perform multi-color pump-probe measurements in a broad spectral range in combination with the FEL.

In the following, FELs will be discussed with respect to the wavelength range which they cover. For each wavelength range alternative laser sources are given and compared to FELs. It should be noted that these alternative sources evolve very rapidly and thus pose challenges on the justification of FELs in certain wavelength and power ranges. For each wavelength region selected scientific achievements obtained from experiments performed with FELs are given.

FELs in the Far Infrared Region

The far infrared region is roughly defined as the wavelength (energy) range from 20 μm to 300 μm (62 meV to 4 meV). At the long wavelength end (wavelengths larger than 220 μm) radiation can be obtained from pure electronic devices, so-called backward wave oscillators (BWO). Furthermore, optically pumped gas lasers operate at certain discrete wavelengths in this range where there exists one fundamental problem for realizing a semiconductor laser based on transitions between the conduction and the valence bands. Auger recombination, where an electron and hole combine by transferring another electron high into the conduction band, becomes increasingly important for smaller bandgaps. Recent advancements in this field are based on quantum cascade lasers, which use inter-sub-band transitions of quantum structures for achieving lasing in the wavelength range from 105 μm to 3.4 μm. Another possibility to generate coherent radiation in this wavelength range is optical rectification or difference frequency mixing of femtosecond laser pulses at near-infrared wavelengths. However, none of these sources outperform FELs in terms of peak or average power, making FELs the most important tunable laser source in this wavelength range.

In the far-infrared, several scientific opportunities exist. In condensed matter many important elementary excitations exist in this frequency range, e.g., phonon-polaritons, plasmons, polarons, and magnons. The quantum confinement in semiconductor heterostructures leads to the formation of discrete electronic levels with energy separation typically in the far-infrared. The associated radiative transitions, so-called inter-sub-band transitions, can be arbitrarily tuned by varying the materials and the dimensions of the confining potential. In surface science, intramolecular vibration of molecules adsorbed to surfaces are of large interest. They can be studied by resonant nonlinear optical experiments such as sum-difference frequency generation or second harmonic generation, giving information about the energy transfer between the molecules and the surface. In biophysics, low-frequency modes of large bio-molecules lie in the far-infrared, whose dynamics can also be studied.

Important scientific results have been obtained with FELs in the following fields, which exclude the important fields of molecular science, chemistry, biology, and medical research.

Scanning Near-Field Optical Microscopy (SNOM)

Increasing the spatial resolution is an important task in spectroscopy. Free-space focusing of a laser beam is diffraction-limited to a spot with the dimensions of the wavelength. Especially in the far-infrared region this would allow to resolve objects larger than 10 μm only. However, the imaging of elementary excitations in solid state or of collective excitation in biomolecules on a sub-μm scale, would be highly desirable. SNOM allows to circumvent the diffraction limit. Several different SNOM techniques have been

introduced in recent years, most of them working with apertures smaller than the wavelength. Imaging is performed by scanning the aperture over the sample or vice versa. An evanescent electromagnetic field is present at the aperture which can couple to the specimen brought into the proximity of the aperture. Since the aperture is smaller than the wavelength, the light intensity drops by order of magnitudes. It is here that the high power available from FELs becomes advantageous. Usually, the end of a fiber is tapered and metalized, leaving a small aperture at the tip of the fiber. In the far-infrared region the problem exists in the absorption of the fiber material. Chalcogenide fibers are capable of transmitting in the 2–12 μm range, while hollow glass capillaries coated with metals inside work also at longer wavelengths. Recent approaches try to work without apertures, e.g., by coupling the light locally via a nanometer metal tip to the sample. Another interesting approach is to use the focus of visible laser light to create a small highly reflective spot on a material. By creating free-carriers with the visible laser light in the material, the reflectivity of the far-infrared light of the FEL can be strongly increased, thus a reflecting spot is created which is diffraction-limited by the wavelength of the visible light rather than that of the infrared light. The techniques described here are, of course, not limited to the far-infrared region but may also be applied successfully, for example, in the mid-infrared region.

Rabi Oscillations

Rabi oscillations are a fundamental phenomenon in a quantum mechanical two-level system driven by a light field. The light field promotes electrons from the lower to the upper state and back, depending on the duration and intensity of the light field, i.e., the population oscillates between the lower and upper state of the two-level system with the oscillation period depending on the intensity of the light field. This phenomenon has been studied extensively in atoms. In semiconductors, scattering processes strongly aggravate the observation of Rabi oscillations. By using the Santa Barbara FEL Rabi oscillations between the ground and excited state of a hydrogenic, impurities in GaAs could be observed. From the excited state of the impurity, electrons are transferred to the conduction band, thus changing the photoconductivity of the crystal, which is measured as a function of time delay. Overdamped oscillations with frequencies changing with the laser intensity could be clearly resolved.

Optical Nonlinearities in Semiconductors

An important concept of nonlinear optics is based on the nonlinear susceptibility $\chi^{(n)}$ of a material, where n denotes the order of the nonlinearity. The nonlinear susceptibility has tensorial character and determines, for example, the efficiency of second harmonic generation and difference frequency mixing ($\chi^{(2)}$), third harmonic generation, and the Kerr nonlinearity ($\chi^{(3)}$), etc. Hence the knowledge about the dispersion and the absolute value of $\chi^{(n)}$s is both of fundamental and practical importance. Due to the lack of tunable laser sources, these values are not well known in the far-infrared. Especially close to fundamental lattice resonances in the far-infrared, a strong dispersion is theoretically expected. Here FELs play an important role in the determination of these values. For example, in GaAs, the dispersion of the $\chi^{(2)}$ has been recently determined from second harmonic generation experiments below the fundamental optical phonon resonance at 8 THz, i.e., wavelengths larger than 37 μm, where a resonance enhancement at half the phonon frequency could be observed.

Relaxation Kinetics in Semiconductor Quantum Structures

The investigations of inter-sub-band transitions in semiconductor heterostructures have opened the pathway for the realization of the quantum cascade laser. In these devices, the radiative and nonradiative transition rates between different electronic quantum levels are of prime importance for a working device. The relevant time-scales have been obtained from pump-probe experiments with FELs. In semiconductor quantum dots, the strong coupling of phonons to inter-sublevel transitions could be observed and the polaron lifetime could be determined.

Bloch Oscillations in Semiconductor Superlattices

Bloch oscillations are an intriguing concept of solid state physics. When an electron experiences the force of an electric field in a periodic potential, the resulting dynamics is oscillatory in the absence of scattering events. However, since in real solid state the scattering rates are much higher than the expected oscillation period, such oscillations cannot be observed and the application of an electric field leads to a net electric current. However, in artificial semiconductor superlattices, the Bloch oscillation period can be tailored to be smaller than the scattering rate, leading to the observation of Bloch oscillations and the associated THz emission. The inverse effect of a Bloch oscillator, i.e., the generation of a photocurrent by tuning an FEL to the Bloch oscillation frequency, could be observed.

Cluster Physics

FELs can be an important tool for studying the dynamics of clusters and molecules in the gas phase by

infrared resonance enhanced multiphoton ionization (IR-REMPI). In IR-REMPI, the species of interest are irradiated by pulsed, tunable IR radiation from a free electron laser. When the radiation is resonant with a vibrational mode of the molecule or cluster, the absorption of many – up to several hundred – photons can take place. When the internal energy is high enough, the thermal emission of an electron can take place. The resulting ions can be detected mass selectively in a time-of-flight mass spectrometer. Monitoring the mass specific ion yield as a function of IR wavelength then yields the IR spectrum of that species. This technique has given important information on the IR absorption spectra of interstellar dust.

FELs in the Mid-Infrared Region

This wavelength (energy) range extends from 20 μm to 2 μm (62 meV to 0.62 eV). Alternative sources in this wavelength range are nonlinear optical parametric generators and amplifiers (OPGs/OPAs) which compete against FELs in terms of peak power. However, there still exists a large number of interesting phenomena where FELs are better suited, e.g., due to their higher repetition rate in the MHz range as opposed to the kHz repetition rate of OPGs/OPAs and their higher average power exceeding several W up to kW as compared to mW for OPGs/OPAs. These high-power levels find applications in material processing and recently in surgery. Since these applications cover such a wide variety of different materials, we refer the reader to publications from users of the Jefferson Laboratories FEL and the Vanderbilt University FEL (see Table 1). The advantage of FELs in this context is that their wavelength can be tuned exactly to a resonance of the material under preparation. The following significant spectroscopic application was performed with FELs.

Vibrational Relaxation of Defects in Semiconductors

The heat dissipation in submicron electronic devices is of prime importance for the failure behavior of such devices. As the device size shrinks to the nanometer scale, atomic defects play an important role. In silicon, hydrogen passivation of defects is a standard process step in the production of metal-oxide-semiconductor electronic devices. Hence detailed knowledge about the stability and heat dissipation of hydrogen defects is very important. By employing nonlinear optical methods such as four-wave mixing or a transient bleaching technique, the vibrational lifetime of hydrogen vibrational modes in silicon (around a wavelength of 5 μm to 6 μm) could be determined and important information on the dependence of the local atomic arrangement and the vibrational lifetime could be obtained.

FELs in the Near-Infrared to the UV Region

The wavelength (energy) range from 2 μm to 200 nm (0.62 eV to 6.2 eV) is the realm of solid state lasers. Especially diode pumped solid state lasers have made tremendous improvements in terms of power and stability in the past years. Nonlinear frequency conversion of ultrashort optical pulses generated from titanium-sapphire lasers and amplifiers cover almost all of this wavelength range. Hence there is no specific need for operating FELs in this wavelength range for applications in physical sciences.

FELs in the Extended UV to X-ray Region

In the wavelength (energy) range shorter than 200 nm (larger than 6.2 eV) the availability of laser sources is very scarce. In this wavelength range, solid state lasers have the problem of the reabsorption of the laser light by the crystals' hosts. Also nonlinear frequency conversion processes are hampered by the absorption of the nonlinear crystals in this wavelength range. The generation of ultrashort pulses at the wavelength of the Cu K_α line has been obtained on the basis of amplified femtosecond titanium-sapphire pulses, thus enabling the study of the lattice vibrations in the time domain via a modulation of the X-ray diffraction. Since these sources are limited in their wavelength and output power, the development and operation of FELs in this wavelength range is truly important.

Only recently the first experimental verification of the feasibility of a SASE FEL at short wavelengths was obtained at the DESY facility. The lasing wavelength was 110 nm. Since several of these sources are presently under construction worldwide, there are as yet no experimental results published. However, the future development of short-wavelength FELs will enable new types of time-resolved experiments in atoms, solid state physics, chemistry, and biology. One prominent example is protein crystallography with X-ray pulses. The high intensity of FELs in this wavelength range should allow one to obtain a diffraction pattern of a single protein or other important bio-molecules in a single shot on a subpicosecond time-scale, i.e., before the protein is fragmented. The short-pulse duration opens the pathway to study the dynamical behavior of such

structures in pump-probe experiments with solid-state lasers synchronized to the FEL. In solid state physics, such experiments could yield important insights into the dynamics of phase transitions.

Other Developments

Beyond the use of an FEL as a light source in the above described wavelength ranges, there exists the possibility of using FEL facilities in a broader way. One example is the development of a high-intensity γ-ray source which can be accomplished by Compton backscattering of FEL photons from the relativistic electrons driving the FEL. This work has been performed at a storage ring FEL (OK-4/Duke University), where 12.2 MeV γ rays were obtained by backscattering 379.4 nm free-electron laser photons from 500 MeV electrons circulating in the storage ring. The advantage of these γ rays is that they are highly monochromatic and linearly polarized, thus making them an ideal source for nuclear spectroscopy and investigations in cancer therapy.

See also

Incoherent Sources: Synchrotrons. **Lasers:** Free Electron Lasers. **Spectroscopy:** Nonlinear Laser Spectroscopy.

Further Reading

Andruszkow J, Aune B, Ayvazyan V, *et al.* (2000) First observation of self-amplified spontaneous emission in a free-electron laser at 109 nm wavelength. *Physical Review Letters* 85: 3825–3828.

Benson SV (2003) Free electron lasers push into new frontiers. *Optics and Photonics News* 14: 20–25.

Budde M, Lupke G, Chen E, *et al.* (2001) Lifetimes of hydrogen and deuterium related vibrational modes in silicon. *Physical Review Letters* 87: 145501–145503.

Cole BE, Williams JB, King BT, Sherwin MS and Stanley CR (2001) Coherent manipulation of semiconductor quantum bits with terahertz radiation. *Nature* 410: 60–62.

Colson WB, Johnson ED, Kelley MJ and Schwettman HA (2002) Putting free-electron lasers to work. *Physics Today* 55: 35–41.

Dekorsy T, Yokovlev VA, Seidel W, Helm M and Keilmann F (2003) Infrared-phonon-polariton resonance of the nonlinear susceptibility in GaAs. *Physical Review Letters* 90: 055508.

Helm M (ed.) (2000) *Long Wavelength Infrared Emitters Based on Quantum Wells and Superlattices*. Amsterdam: Gordon and Breach.

Knippels GMH, van de Pol MJ, Pellemans HPM, *et al.* (1998) Two-color facility based on a broadly tunable infrared free-electron laser and a subpicosecond-synchronized 10-fs-Ti sapphire laser. *Optical Letters* 23: 1754–1756.

Litvinenko VN, Burnham B, Emamian M, *et al.* (1997) Gamma-ray production in a storage ring free-electron laser. *Physical Review Letters* 78: 4569–4572.

Murdin BN, Knippels GMH, Vandermeer AFG, *et al.* (1994) Excite-probe determination of the intersubband lifetime in wide GaAs/AlGaAs quantum wells using a far-infrared free-electron laser. *Sem. Sci. Techn.* 9: 1554–1557.

Palanker DV, Knippels GMH, Smith TI and Schwettman HA (1998) IR microscopy with a transient photo-induced near-field probe. *Opt. Comm.* 148: 215–217.

Sauvage S, Boucaud P, Lobo RPSM, *et al.* (2002) Long polaron lifetime in InAs/GaAs self-assembled quantum dots. *Physical Review Letters* 88: 177402.

Sorokina IT and Vodopyanov KL (eds) (2003) *Solid-State Mid-Infrared Laser Sources*. Berlin: Springer.

Unterrainer K, Keay BJ, Wanke MC, *et al.* (1996) Inverse Bloch oscillator: strong terahertz-photocurrent resonances at the Bloch frequency. *Physical Review Letters* 76: 2973–2976.

von Helden G, Tielens ACGM, van Heijnsbergen D, *et al.* (2000) Titanium carbide nanocrystals in circumstellar environments. *Science* 288: 313–316.

World Wide Web Virtual Library (2004) *Free Electron Laser Research and Applications*. Available online at http://sbfel3.edu/www/UL_fel.html

Industrial Applications

I P Mercer, Imperial College of Science, Technology and Medicine, London, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Lasers are seeing increasing application in industry, driven by emerging market demand and continuing enabling advances in laser technology. Core laser technology advances are based on robustness, performance, and cost. A combination of gas and solid-state lasers currently see application across the spectrum, from the extreme ultraviolet (EUV) to the infrared (IR). Industrial applications continue to emerge across the spectrum and are being extended to the extreme ends of the spectrum at terahertz (THz) and X-ray frequencies (see Figure 1).

Where a step in level of performance is offered by a laser based tool, that is robust enough and cost-effective for industrial application, it can be anticipated that new applications and markets will emerge. For this reason, many established laser manufacturers and new entries in the 1980s and 1990s were driven in part by the exploitation of laser technology advances, as opposed to market application demand. As markets mature for lasers applied to materials processing and diagnostics, so the driver across industry is beginning to shift from being predominantly set by the capability of emerging laser technology, to a combination of laser technology drive and application demand.

Figure 2 shows the current and emerging picture of laser technologies across a core range of wavelengths and pulse durations. Current technologies are shown along with emerging technologies expected to see industrial application in this decade, opening up new applications and markets. Currently established technologies include: CO_2 gas lasers at 10 μm; excimer and KrF gas lasers emitting in the ultraviolet at 248 and 193 nm, respectively; direct diode lasers emitting in the near infrared and visible; and solid state lasers emitting in the near infrared and visible. Current technology developments are centered on extending the range of performance of both gas and solid-state lasers: gas lasers are being extended to the ultraviolet with gases such as F_2; diode lasers are being extended to the ultraviolet with materials such as nitride-based semiconductors and to the infrared with the incorporation of quantum cascade technology; and solid-state diode pumped lasers are being extended to the infrared and ultraviolet via nonlinear conversion. In essence, achievable laser parameter space is seeing continued expansion, not only in terms of pulse duration and wavelength, but also in terms of average power, pulse energy, beam quality, robustness, and cost. Laser technology has now arrived as a major driver for industrial and social change through this decade and into the next.

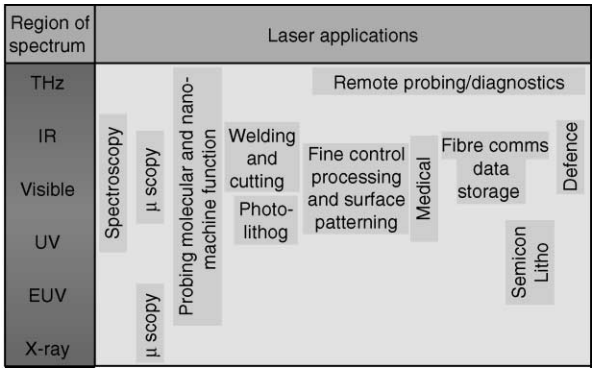


Figure 1 Industrial laser applications across the spectrum.

Infrared wavelengths directly stimulate nuclear motion, coupling to heat without necessary recourse to electronic excitation. Visible wavelengths couple to electronic transitions and at high laser intensities can ionize via multiphoton absorption. Ultraviolet wavelengths couple to electronic transitions and for wavelengths in the region of or below 200 nm, ultraviolet radiation can directly result in bond breaking. For materials processing applications, ultraviolet radiation can have advantages in reducing the laser threshold required for ablation, translating to reduced thermal load imparted to the material.

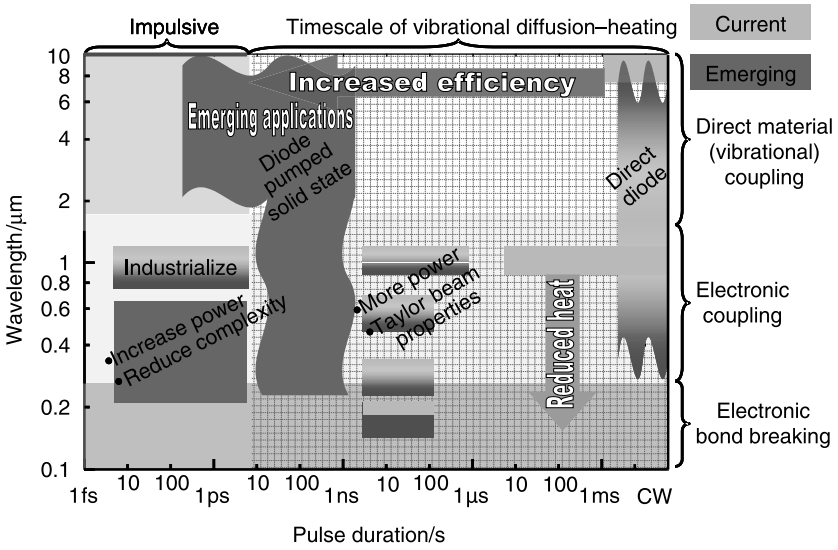


Figure 2 Current and emerging laser technologies across the spectrum as a function of laser pulse duration. Currently established technologies (light gray) are shown with technologies expected to emerge during this decade (dark gray).

Nuclear motions in materials cover a range of timescales of dynamics from 100 ps ($0.3\,cm^{-1}$) to 10 fs ($3000\,cm^{-1}$). As the laser pulse duration is reduced below 1 μs, the effects of thermal diffusion are mitigated in solids and as the pulse duration is further reduced below 100 ps, the interaction between radiation and matter is seen to adopt impulsive behavior. Reduction in pulse duration below the microsecond regime can lead to increased efficiency for material ablation combined with improved quality.

Materials Processing – Photochemical

Lithography in Semiconductor Manufacture

Laser technology's role in semiconductor and microelectronics fabrication is growing as manufacturers move towards producing smaller, more powerful devices. The overriding commercial driver for a generation has been to increase the number of integrated components per area of silicon. This has been achieved by reducing the feature size of a transistor on silicon and this trend is anticipated to remain in place through to 2010.

Features on silicon are chemically etched following lithographic exposure. To meet the increasing demand in resolution and throughput, lasers have been applied with a trend in development towards shorter (ultraviolet) wavelengths, with increasing power and reliability.

The semiconductor industry is currently in transition from 248 nm (XeCl gas) to 193 nm (ArF gas) ultraviolet (UV) pulsed laser technology to expose silicon wafers of diameter 30 cm. A typical laser power available at this time at 193 nm, is 40 W at a repetition rate of 4 kHz. To continue the reduction of feature sizes by a factor of two every four to five years, the industry is considering moving to 157 nm (F_2 gas) pulsed laser technology. As wavelengths reach the deep ultraviolet (DUV), transmission through gas is significantly reduced, requiring a move to vacuum-based systems. Associated technologies relating to translation (stepper), mask, and image relay optics also require significant advancement. A technology also being considered at this time is to image the mask to silicon through a medium with refractive index greater than unity, thereby reducing the minimum feature size afforded by diffraction theory, whilst maintaining the same laser wavelength and numerical aperture.

It is estimated that at a feature size of approximately 80 nm, the above technologies will be reaching their resolution limit. This will necessitate a switch to electron beam projection lithography or extreme ultraviolet (EUV, 13 nm) technology. Electron beam projection carries a disadvantage in terms of projected rates of wafer throughput and as such, emphasis has been placed by the industry on the current development of EUV technology. EUV radiation can be produced by laser interaction with a gaseous volume.

Computer to Plate Printing

A common advantage where lasers are replacing current technologies is to reduce the number of associated processes and thereby reduce the cost of process ownership. Computer to plate (CTP) printing involves the laser exposure of a printing plate without the use of a film intermediate. This technology is finding fast traction within the printing industry and is applicable to drum or flat bed platforms. CTP first gained serious attention in the 1990s and has since substantially advanced, with a wide range of technology suppliers (Figure 3).

Traditional methods have involved offset printing plates, imaged used traditional film, and high-power

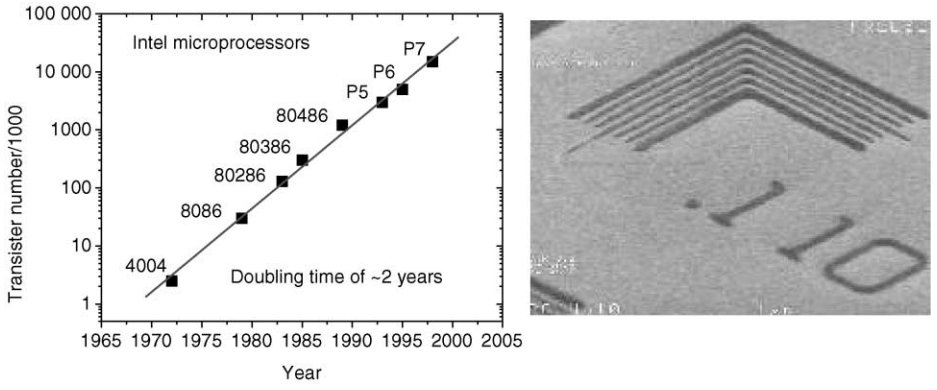


Figure 3 Moore's Law: continuing growth in computing power, showing a doubling in the number of transistors per unit area of silicon every 2 to 2.5 years. Also shown are 110 nm dense lines on silicon (state of the art in 2002). Source: International Sematech.

UV lamps. In contrast, CTP exposes the plate directly with laser radiation. The major saving reported by users of CTP is a faster turnaround time. For color printers, CTP gives almost instant color register on the press and there is a reduced requirement to stop the press to delete dust marks and scratches, as was the case with film. As the industry sees transition to reduced press run lengths, so the importance of CTP is increasing.

Digital plates generally use a high-quality, grained, and anodized aluminum substrate, as for analog plates (see Figure 4). High tolerance electrochemical graining and anodizing ensure stable ink/water balance and good-quality printed results and translates to predictable press behavior. Multiple coatings are applied for CTP, in contrast to a single coating for offset printing, and these take the form of a photopolymer or silver halide and a thermally responsive layer.

On the press, ink and water rollers are continually rolled over the plate. The grained and anodized aluminum substrate serves as a hydrophyllic surface volume to attract water. Ink is oil based and is repelled from areas of water absorbed into the substrate, but is attracted to the hydrophobic dot material.

Silver halide is rated for runs up to 350 000, proving a popular choice for commercial and newspaper applications. It is straightforward to sensitize for a range of writing wavelengths and is the fastest to write.

Thermal plates offer high resolution, equivalent to silver halide and can be baked to enable the longest application lifetimes. Typically the plates are exposed to 830 nm radiation on the external surface of a drum. This technology has served the industry well, but has struggled to find full application in the newspaper business, due to its disadvantage in writing speed.

The most common CTP technology used for newspapers uses a photopolymer coating, which is fast to write and durable. Resolution has traditionally been limited, with Agfa, for example, rating their N91 photopolymer plate to a maximum of 175 lines per inch (LPI), whilst thermal and silver plates are able to image at 200 LPI and above. However, recent advances in shorter wavelength violet diode lasers and violet sensitive polymers have improved resolution. Launched in 2000, violet technology is seeing widespread takeup. This technology is a spinoff from the DVD industry and with laser lifetimes estimated at 10 to 20 years, cost of ownership is one of the drivers. Both silver halide and photopolymers can be sensitized to violet. Current typical laser parameters are for: silver halide, 2–3 $\mu J/cm^2$ at 405 nm; photopolymer, 0.2 mJ/cm^2 at 488 or 532 nm; and thermal, 0.1 J/cm^2 at 830 nm.

The chemical composition of the CTP surface layers is balanced to determine the response to different laser wavelengths, as well as performance characteristics in different print applications. The plate coatings undergo a photochemical change during laser exposure. This is a critical stage since it is important that the plate surface is not under- or over-exposed.

Chemical post processing after exposure is generally required for CTP to date, with the exception of ablative CTP, to remove unexposed material. After imaging, thermal plates go through a pre-bake step, during which the system heats the plate to 130 °C (266 °F) for about 30 seconds. The energy from this

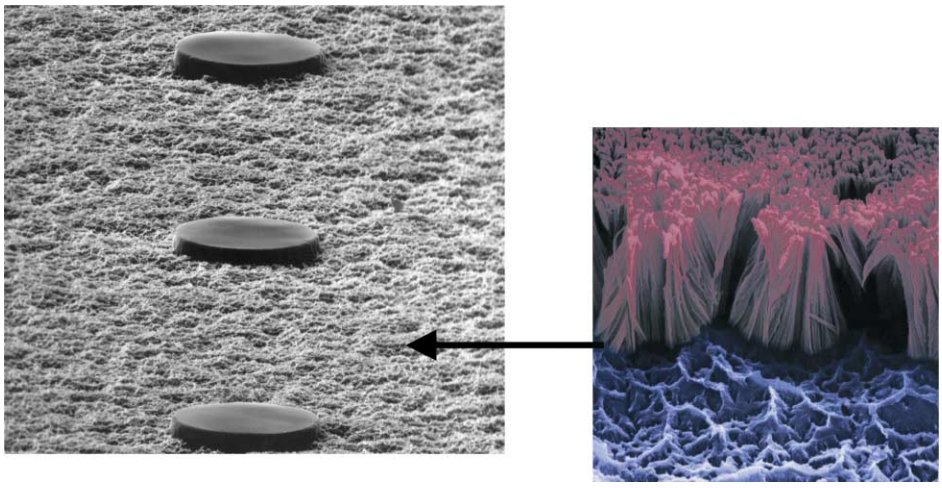


Figure 4 Three typical polymer dots of a few microns depth on an anodized aluminum substrate. Also shown is a further magnified electron microscope image of the substrate material, with the anodized layer cut away for presentation. The porous structure increases water retention. Oil-based ink is repelled from areas other than the dots. Figure courtesy of Agfa (Tony King, tony.king1@agfa.com).

heating stage completes cross-linking of polymer chains in the exposed portions of the emulsion. Chemical processing consists of processing the plate through an alkaline developer solution where the polymer that is not cross-linked washes away. Following the wash is a rinse and application of a coating. After chemical processing, some plates require baking to 200 °C (392 °F) for 35 minutes. Afterwards, the plate continues on to rinsing.

The machine tools required for these processing steps measure 30 feet long and take up to 7 minutes to process a plate after imaging. For both polymer and thermal plate processing, the temperatures are carefully controlled to prevent the aluminum substrate from deforming during the baking stages.

The effluent from thermal and other photopolymer plate processors is fairly benign and is typically directly disposed without significant environmental concern. Spent photopolymer chemistry can, however, pose problems with waste flow, it being a plastic sludge that can solidify and over time, clog plumbing systems.

Silver diffusion plate processors are relatively compact and tend to process plates in less than a minute. Environmental precautions are required with the disposal of accumulated silver particulates. Relatively simple systems are available, which neutralize the pH of the effluent and filter out silver until concentration levels are below 3 parts per million. Under these circumstances, most jurisdictions allow drain disposal.

Ablative plates do not require post processing, other than a rinse to remove residues. This is an emerging technology that has not taken hold at this time in the printing industry and requires more powerful lasers to ablate a layer from the plate. These plates use layers of ink-repellent silicon on top of an aluminum substrate to provide the non-inking portion of the plate. Variations of this technology use a laser to ablate the silicon from the substrate in the image areas of the plate, or to transfer material from a donor layer to the aluminum substrate. A weakness of this technology has been the limited durability of the plates in press. There is continuing development to produce ablative plates with other surface layers that are more durable on press.

Materials Processing – Heating

One way in which lasers can be applied to materials processing is as a directed source of heating. With material quickly raised in temperature to the point of melting and vaporization, applications lie in drilling, welding, cutting, and scintering.

The exploited characteristics of a laser source are its directionality and discrete color. Directionality enables the formation of a restricted area for heating as well as an extended working distance, whilst a discrete color enables absorption to be optimized for the material being worked. Laser technologies typical in application to materials processing through heating, range from diodes at 808 and 940 nm to flashlamp or diode pumped solid state lasers at 1 μm, and CO_2 gas lasers at 10 μm. A recent entry technology is in diode pumped fiber lasers, reaching output powers in the kilowatt regime at a wavelength of 1 μm and typically being continuous in emission or pulsed with a pulse duration in the millisecond regime.

Direct diode lasers (see Figure 5) have advantages in simplicity and robustness and with the minimum of additional assembled optical components, costs can be expected to continue to reduce, fed by Moore's law for the cost of semiconductor manufacture. Typical powers range from 1 to 10 kW and light can be fiber delivered to the work piece. A disadvantage of direct diode emission is the reduced directionality in comparison with alternative laser sources. For example, a current typical numerical aperture of emission from a diode bar array emitting a power of 60 W, is 10 and 60 degrees in directions parallel and perpendicular, respectively, to the array. These divergences can be reduced to a numerical aperture of approximately one degree with the addition of micro-optics, at a device cost increment of approximately 20%. For an array of diode bars with emitting powers in the kilowatt regime and with emitting dimensions sided 2 cm, a typical optical quality is one thousand time diffraction limited, translating to working distances are on the order of centimeters for a laser spot size at a work piece that is on the order of 1 mm.

CO_2 gas lasers and solid state lasers offer the advantage of reduced divergence and corresponding greater working distance, combined with a smaller laser spot size at the work piece for finer processing.



Figure 5 Diode bar array, with an output power exceeding 500 W. Courtesy of Rofin.

Both laser technologies can operate in the kilowatt regime, for application to sheet metal cutting and welding.

With the interaction process involving high temperatures and the ejection of material, a high-pressure gas bathes the area of illumination to remove smelt and either reduce exothermic chemical reactions, or where process acceleration is required, to utilize an assist gas to increase heating through exothermic reaction. Oxygen is typically used as an assist gas for the processing of low-alloy steels, with the process being typically known as laser flame cutting. The gas is, in general, injected coaxially with the laser beam. Laser flame cutting is used in industry for material up to 4 cm thick, albeit the width of the cut becomes significant at this thickness. Both CO_2 and Nd:YAG lasers are applied for this application, the choice of which is dependant on the cut geometry, the cycle time, and the material. Typical cutting speeds are, for example, in the region of 10 meters per minute for 1 mm, 3 meters per minute for 6 mm, and 1 meter per minute for 15 mm thick material.

In the case of high-alloy steels and aluminum, in particular, the material is bathed with an inert gas such as nitrogen or argon. The material is heated solely by the energy in the laser beam, and is typically referred to as laser fusion cutting (see **Figure 6**). The laser power required is higher than for laser flame cutting. Laser fusion cutting reduces oxidization to the cut edges, which is particularly important where welding is the next process step after cutting.

Laser fusion cutting is used in industry for material up to 25 mm thick, albeit the width of the cut becomes significant at this thickness.



Figure 6 Laser fusion cutting. Courtesy of Rofin.

As with laser flame cutting, both CO_2 and Nd:YAG lasers are suitable for this application, the choice of which is influenced by the geometry of the cut, the cycle time, and the material. Typical cutting speeds in steel are, for example, in the region of ~8 meters per minute for 1 mm, 4.5 meters per minute for 3 mm and 1.5 meters per minute for 8 mm thick material.

Materials Processing – Localized Heating

As a material is heated in a localized volume to the point of vaporization, so also is the surrounding material, which can melt and experience cracking or changes in phase. Such heating is mainly considered deleterious, reducing process tolerances and reducing the tensile strength of the surrounding material.

As the pulse duration is reduced for a localized laser heating source, so are the dimensions that experience heating around the interaction volume (the heat affected zone), translating to improved part quality, with the length for thermal diffusion, l_{th} given by:

$$l_{th} = \sqrt{Dt} \qquad [1]$$

where D is the thermal diffusivity and t, the time of diffusion. As the pulse duration is reduced, the pulse energy required to ablate material is also reduced, with the laser pulse energy required to raise the material temperature to the point of vaporization being proportional to the volume heated. For an example such as silver, a pulse duration of 200 fs is associated with a diffusion length of 6 nm. Extrapolating, a pulse duration of 100 ps, 100 ns, and 100 μs are associated with diffusion lengths of 0.1, 4, and 130 μm, respectively. For typical metals and dielectrics, the process quality and pulse energy threshold for ablation become decoupled from variation in pulse duration for pulse durations less than a few picoseconds. In this regime, the volume for heating is dictated more by the depth for optical absorption and further, the physical mechanism for absorption changes to that of material breakdown and multiphoton ionization (**Figure 7**).

Additional advantages of operating in the short-pulse regime are a sharpening of threshold to ablation, enabling the processing of features of dimension less than that of the laser spot size and an increase in the number of materials that can be processed by the same color laser source.

Given the advantages of the short-pulse regime, it should be noted that this is currently juxtaposed with the complexity and cost of the required laser sources. Solid-state laser technology is, however, advancing, significantly reducing the number of optical components required to deliver short pulses at multi-kHz

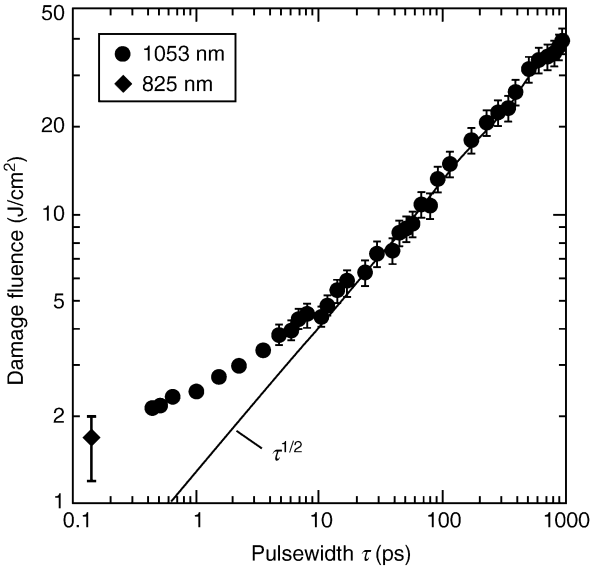


Figure 7 Pulse duration dependence of threshold damage fluence for fused silica. Reproduced with permission from Stuart BC, Feit MD, Herman S, Rubenchik AM, Shore BW and Perry MD (1996) Nanosecond-to-femtosecond laser-induced breakdown in dielectrics. *Physics Review B* 53: 1749. Copyright (1996) by The American Physical Society.

repetition rates and multi-mJ pulse energies. At this time of evolution in laser technology, solid-state, diode-pumped high repetition rate (>10 kHz), high pulse energy (>40 mJ) nanosecond (1–200 ns) lasers at a wavelength of 1 μm are seeing increasing application. Such laser sources offer a combination of robustness and cost-effectiveness, combined with a process quality improvement over sources with longer pulse duration.

In the nanosecond regime, the nature of light-matter interaction remains dependent on the wavelength of the laser source and the material properties. In particular, wavelengths in the UV exhibit reduced pulse energy threshold and can offer improved process quality. UV wavelengths less than 300 nm are currently produced by excimer and ArF gas lasers, emitting at 248 and 193 nm, respectively, or solid-state sources that are wavelength converted with nonlinear optical parametric materials to emit at 354 or 256 nm. Significant laser development effort is underway to improve optical parametric materials for nonlinear conversion, to enable greater robustness for solid-state UV sources that then offer reduced overall laser tool cost and complexity.

Semiconductor Lithography

The advantage of nanosecond and shorter pulsed lasers for materials processing extends from localization of heating. With heating being localized, a greater temperature can be achieved. This behavior is fundamental to the production of EUV radiation for application to next-generation lithographic processing.

EUV lithography is considered as an attractive candidate to succeed deep DUV-based optical lithography within this decade. EUV wavelengths around 13 nm will allow the reduction of structure sizes to less than 80 nm in semiconductor devices.

For such a high-power light source that emits at a wavelength of 13 nm, solutions are currently being pursued by multiple industry players in electrical plasma discharge and laser produced plasmas (LPP) (**Figure 8**). Electron beam and synchrotron sources could, in principle, satisfy the application requirements, but suffer from restriction relating to cost, size, and throughput. The choices of technologies remain to be finalized. It is, however, accepted across the industry that Moore's Law should remain intact and that a solution is required that is scalable to high throughput rates for wafer exposure.

For a laser-produced plasma, the peak wavelength of the light emitted from a plasma is determined by its temperature, T, where according to the blackbody formula:

$$T \sim E_{\mathrm{ph}}/3 \qquad [2]$$

where E_{ph} is the photon energy associated with the wavelength of maximum emission. Liner photopolymerization (LPP) schemes use pulsed laser radiation to heat a target to the optimum temperature required for efficient EUV production. For plasma-emitted light which peaks at a wavelength of 13.5 nm, the optimum temperature is approximately 3.5×10^5 Kelvin (30 eV).

A laser is focused onto a target, transferring its energy into the target in such a way that a plasma is formed and heated. A range of targets have been investigated for use in LPP schemes ranging from gaseous and cluster targets, to liquid targets and macroscopic solid targets. The interaction begins with the production of free electrons via multiphoton ionization, which is strongly dependent on the peak intensity of the laser. As heating progresses, there is further ionization from inelastic collisions between heated electrons and ions to reach a pseudo-equilibrium state, where cooling competes with heating on nanosecond timescales. Given this time-scale for cooling and requirements on source size and temperature, the optimum laser pulse duration lies in the regime of nanoseconds and the optimum laser pulse energy is hundreds of millijoules. Laser technology

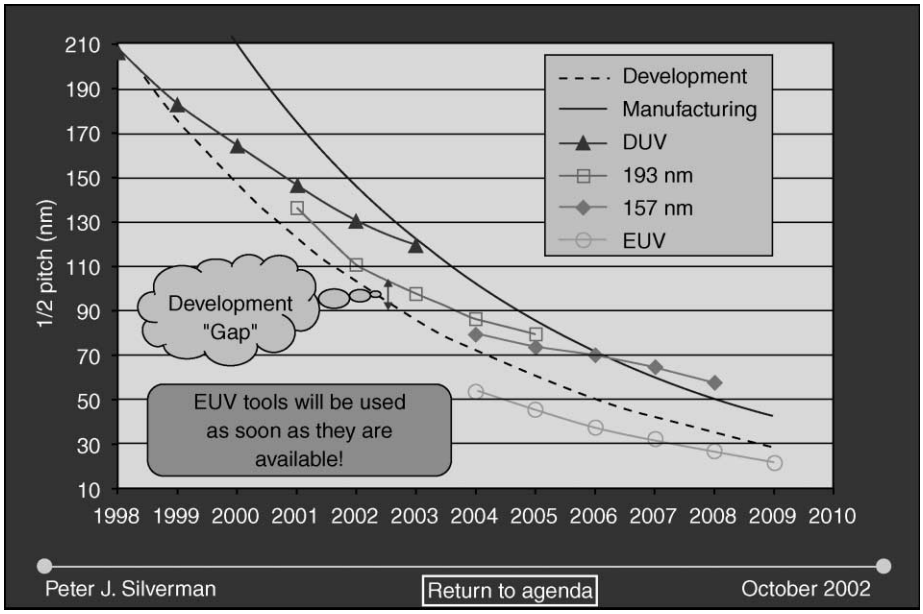


Figure 8 Light source roadmap to 2010. Courtesy of Intel Corporation.

suited to this application is diode-pumped Nd:YAG, emitting at a wavelength of 1 μm and pulsed at multikilohertz repetition rates.

Fine Control Material Removal

Localized heating afforded with the application of lasers with pulse durations in the nanosecond regime and shorter, is seeing increasing takeup to replace processes ranging from wet etching and electro-discharge machining (EDM) in applications such as scribing, drilling, and differential layer removal. Laser methods are seeing advantage in direct ablation of material, for feature sizes on the order of microns to hundreds of microns. Large areas on the order of 2 m^2 are now becoming cost-effective to process with direct laser ablation, with the advent of high average power laser sources of hundreds of watts of average power, combined with nanosecond regime or shorter pulse durations. Typical sources are gas excimer or diode-pumped solid-state.

The finest processing is advantaged by the application of pulse durations in the picosecond regime. A restriction in application of this technology has been robustness and cost of the laser sources to date; however, such technology is continuing to see rapid improvement and may be expected to see wider takeup in industry by 2010.

Also seeing application, are wavelengths in the UV regime combined with nanosecond pulse durations. UV radiation in general offers a reduced threshold of ablation, translating to reduced heating and improved process quality. High absorption at UV wavelengths in most materials, combined with photon energies that can exceed bond energies of the material, contribute to the removal of material with the minimum of heating. Excimer, KrF, and ArF gas lasers are applied here (Figure 9), but it is expected that solid-state lasers will emerge by 2010 to capture a large part of the UV market. CO_2 gas lasers are used for fine processing where the material being ablated is of low thermal diffusivity, which is typical for dielectrics or ceramics. Materials of low thermal diffusivity are associated with a low heat affected zone (see eqn [1]).

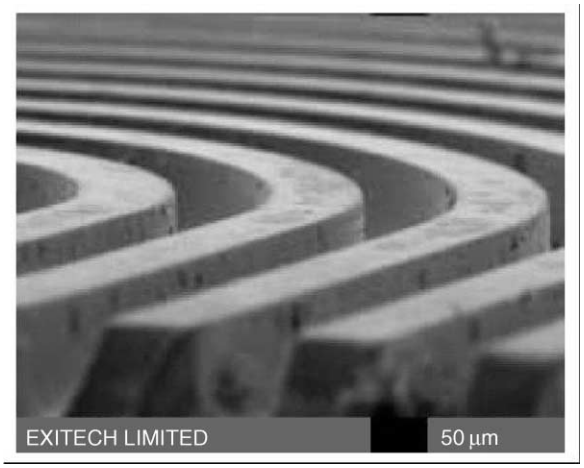


Figure 9 KrF laser micromachined microfluidic channels in polyester for application to lab-on-a-chip technologies or sensing devices. Courtesy of Exitech Ltd.

Scribing is required for dicing of wafers such as silicon and other low-*k* materials, both for computing device manufacture and for solar cell manufacture. Nanosecond and shorter pulsed lasers have seen recent application in this area, enabling higher aspect ratio scribing and a reduced material affected zone. This application has seen laser wavelengths applied from the UV to near-infrared. Short pulse laser scribing can see advantages of high throughput yet reduced or eliminated chipping, delamination, and film peeling.

Fine drilling is required for a wide variety of applications and markets ranging from constructing vias for through-layer interconnects in printed circuit board (PCB) fabrication, to ink jet nozzles (Figure 10) and high aspect ratio holes in hard materials, such as steel or alloys for application to combustion engines.

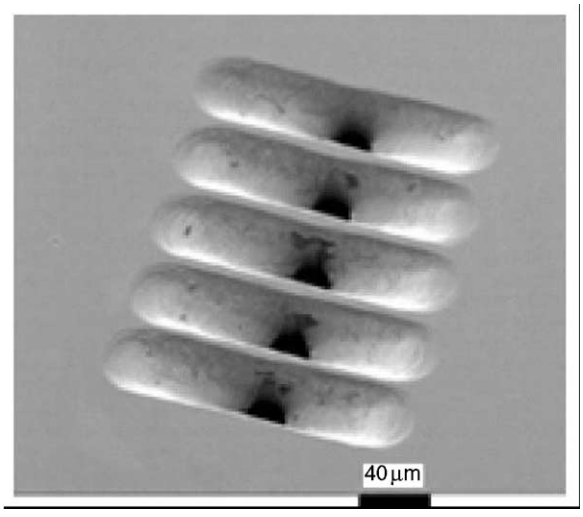


Figure 10 A laser drilled hole in steel and an array of shaped holes in polyamide for inject nozzle application. Courtesy of Exitech Ltd.

Excimer and CO_2 laser drilling has seen industrial application in the drilling of via holes for electronics interconnection packages, since early investigations in the 1980s. More recently, high repetition rate diode-pumped solid-state lasers, nonlinearly converted to emit in the UV, are seeing application. In parallel with the increasing density of transistors in silicon processors, the density of packaging of multichip modules (MCM) and PCB components has also been increasing, requiring the drilling of blind vias on the order of 100 microns in diameter to enable interconnects between multiple conducting layers (Figure 11). Also, as processor power increases, so does the number of connections required to devices. Microvias have enabled the application of two-dimensional grid array connections to the underside of devices, reducing overall package size, including processor electrical connections, by a factor of five.

Laser drilling of microvias in PCBs is expected to increase as packing density increases. PCBs remain cheaper than MCMs and traditionally have been mechanically drilled. As via sizes are reduced below 100 μm, however, it is expected that pulsed CO_2 and Nd:YAG lasers will be applied.

Thin layer ablation is required for applications ranging from surface cleaning to display manufacture and for polymers, dielectrics, and metals. For example, manufacture of plasma displays currently incorporates lithography combined with wet etching, to pattern layers of thin conducting oxide onto a substrate. Multiple process steps can be replaced by application of direct laser ablation and increased laser average power combined with pulse durations in the nanosecond regime are allowing millions of cells to be processed per minute, with the accommodation of process areas currently greater than 2.5 m^2 (see Figure 12).

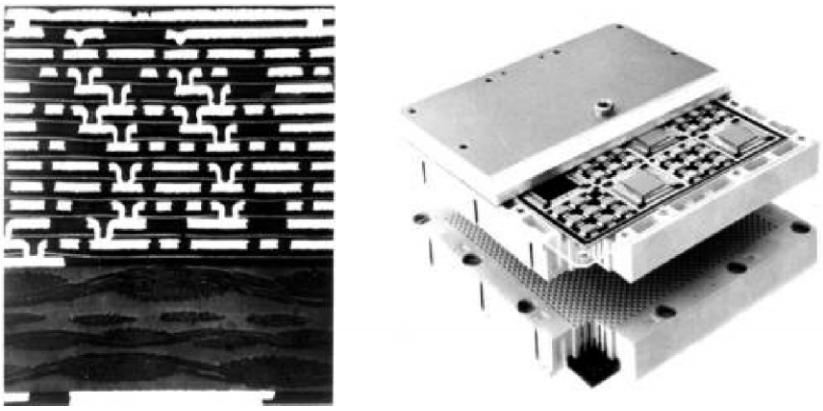


Figure 11 Cross-section of 14-layer copper plated MCM board showing 80 μm diameter microvias drilled by a KrF laser in acrylic resin-polyimide and, fully assembled MCM package. Courtesy of Siemens AG.

Laser-Based Diagnostics

With a continuing increase in reliability, performance, and range of specifications, lasers are seeing growing industrial application for process monitoring and diagnostics, ranging from gas phase detection of molecular specific concentrations to *in vitro* microscopy in the life sciences. Markets are emerging, driven in a large part by advances in laser technology. Chemical sensing with application of lasers is a vast field, which is difficult to cover here completely. Following, are some applications seeing current industrial application and growth.

Microscopy

Confocal microscopy has seen large uptake, in particular, for the application to imaging of cell and membrane function, with more than fifteen manufacturers now supplying confocal microscope tools including Zeiss, Biorad, Leica, Olympus, and Nikon. Confocal microscopy enables increases in resolution and contrast compared to standard microscopy. Techniques such as fluorescence resonant energy transfer (FRET), fluorescence recovery after photobleaching (FRAP), fluorescence loss in photobleaching (FLIP), fluorescence lifetime imaging (FLIM), or fluorescence correlation spectroscopy (FCS), where married with confocal microscopy and, in particular, where relevant molecular structures are known in advance, can offer the additional advantages of specific molecular and structural specificity through the incorporation of tracer fluorophores. This is seen as color coding of the resultant image, with sophisticated microscopes currently offering differential imaging stimulated by three laser colors and with six selectable colors for detection.

The confocal microscope operates with advantage over a standard microscope, by illuminating only a small area in the sample with dimensions of the resolution limit of the objective and re-imaging emission from that area through a pinhole spatial filter (**Figures 13** and **14**). Structures outside of the illuminated area or at a depth outside that of the focus, are suppressed from detection. Fluorophores (tracers) can be inserted into protein sequences at predetermined locations and are stimulated by either single or two-photon absorption of the stimulating laser light.

Two photon absorption has the additional advantages of increased resolution and detected contrast over single photon absorption. Laser sources applied are generally modelocked oscillators operating near 800 nm, with a pulse duration on the order of 100 fs. This technology is continuing to improve in robustness, with fiber and direct diode pumped solid-state sources available from multiple vendors.

FRAP and FLIP monitor diffusion or transport of fluorescent labeled molecules into or out of the laser focal volume, respectively. FCS monitors the fluctuations in detected signal from emissions within the laser focal volume, to determine the rate of diffusion with sensitivity achievable to the single molecule level (**Figure 15**).

FRET is seeing application for the probing of protein interactions in cells or at cell membranes. FRET involves the nonradiative transfer of excitation energy from one fluorophore to another, which subsequently emits at a different wavelength. The fluorophores are inserted into molecules, the

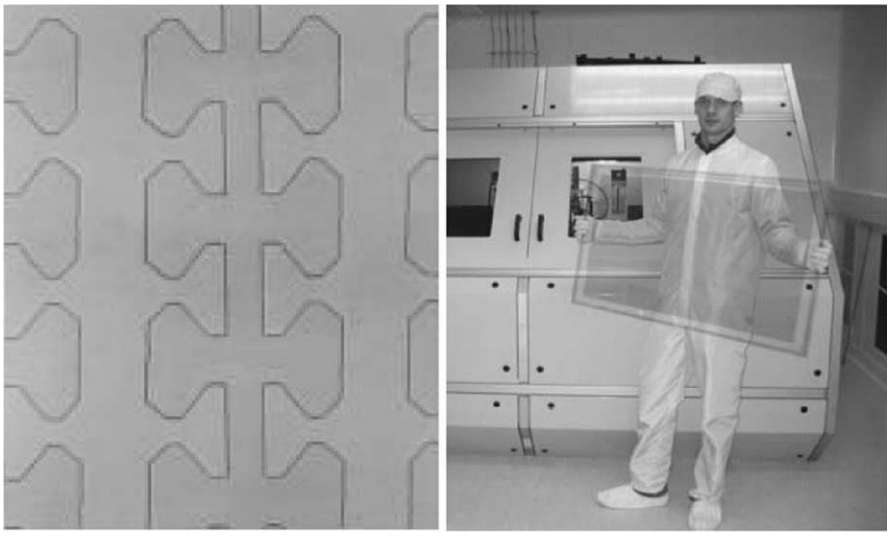


Figure 12 Plasma display manufactured by direct laser ablation of thin-film indium tin oxide (ITO). Courtesy of Exitech Ltd.

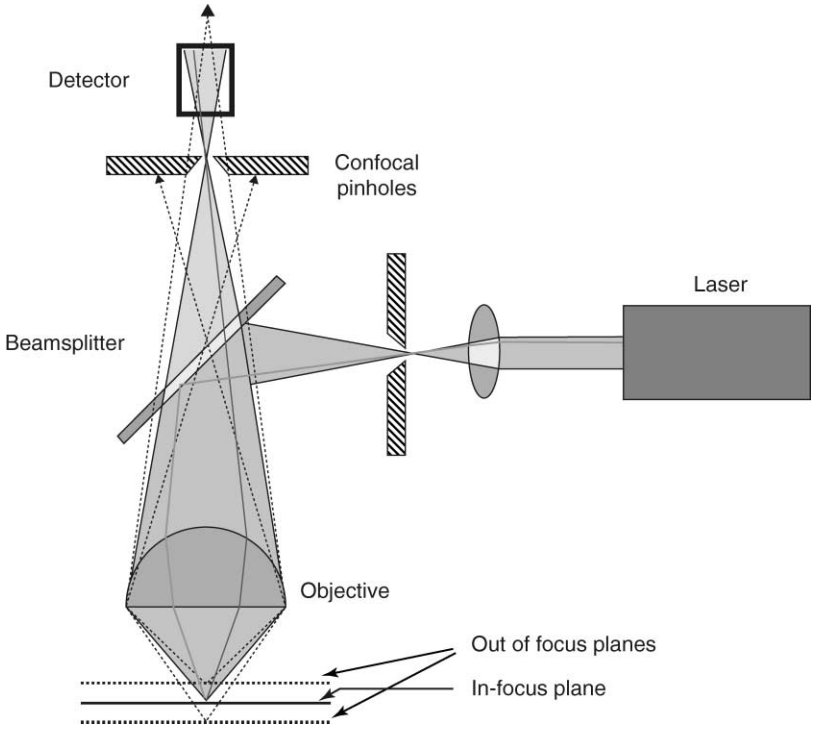


Figure 13 Confocal microscope principle of operation. Courtesy of Leica Microsystems Heidelberg GmbH.

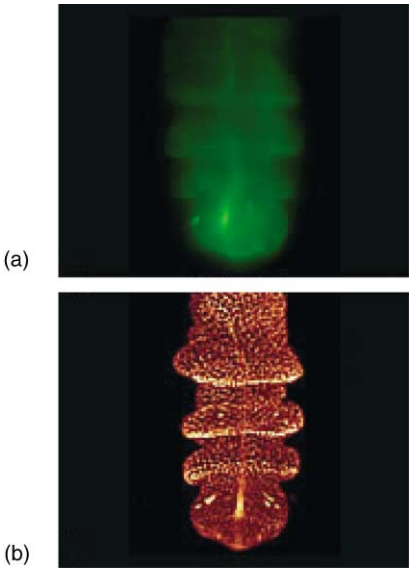


Figure 14 Image (a): *Drosophila melanogaster*, fluorescence nonconfocal configuration and Image; (b): *Drosophila melanogaster*, fluorescence confocal 3D projection. Courtesy of Leica Microsystems Heidelberg GmbH.

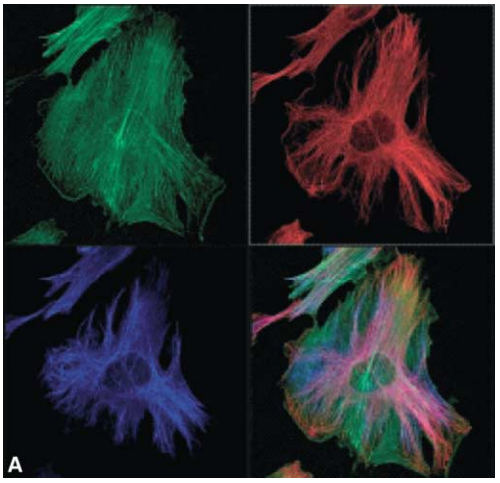


Figure 15 Spectral selective imaging with three fluorescence channels showing triple labeled mouse fibroblast: green: actin fibers, red: microtubules, blue: vimentin. Courtesy of Leica Microsystems Heidelberg GmbH.

interaction of which is being investigated. FRET effectively enables sensitivity to relative position at fluorophore separations less than 10 nm; a significant advance over the 200 nm resolution provided by a standard optical microscope. Concentrations of binding events can be monitored over time (Figure 16).

Raman Methods and Microscopy

Raman spectroscopy sees broad application in chemistry in the analysis of chemical composition and molecular structure, being used, for example, in pharmaceuticals, polymer materials science, bioclinical science, and forensic science.

The history of Raman scattering stems from theoretical predictions in the early 1920s,

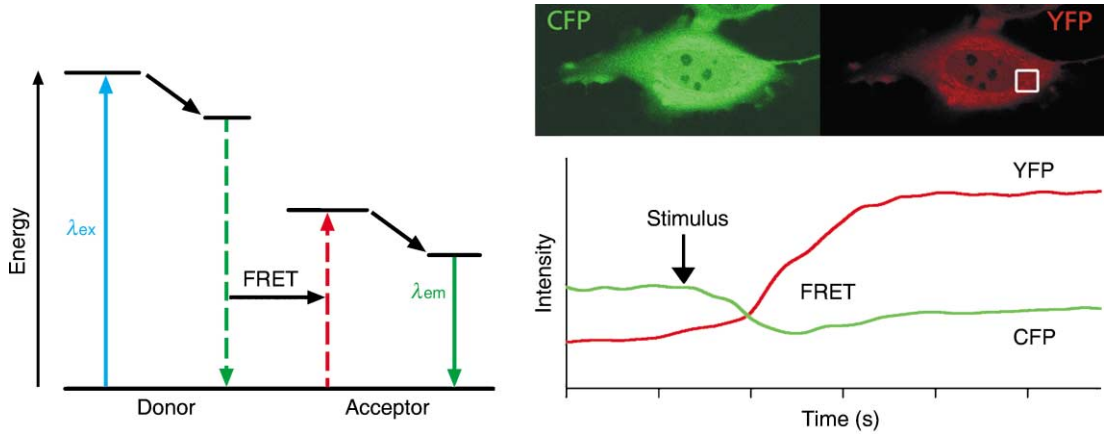


Figure 16 Fluorescence resonant energy transfer mechanism and example relating to cytoplasmic Ca^{2+} concentration. Cyan Fluorescent Protein (CFP) and Yellow Fluorescent Protein (YFP) are acceptor and donor, respectively, for the FRET. Courtesy of Carl Zeiss, Germany.

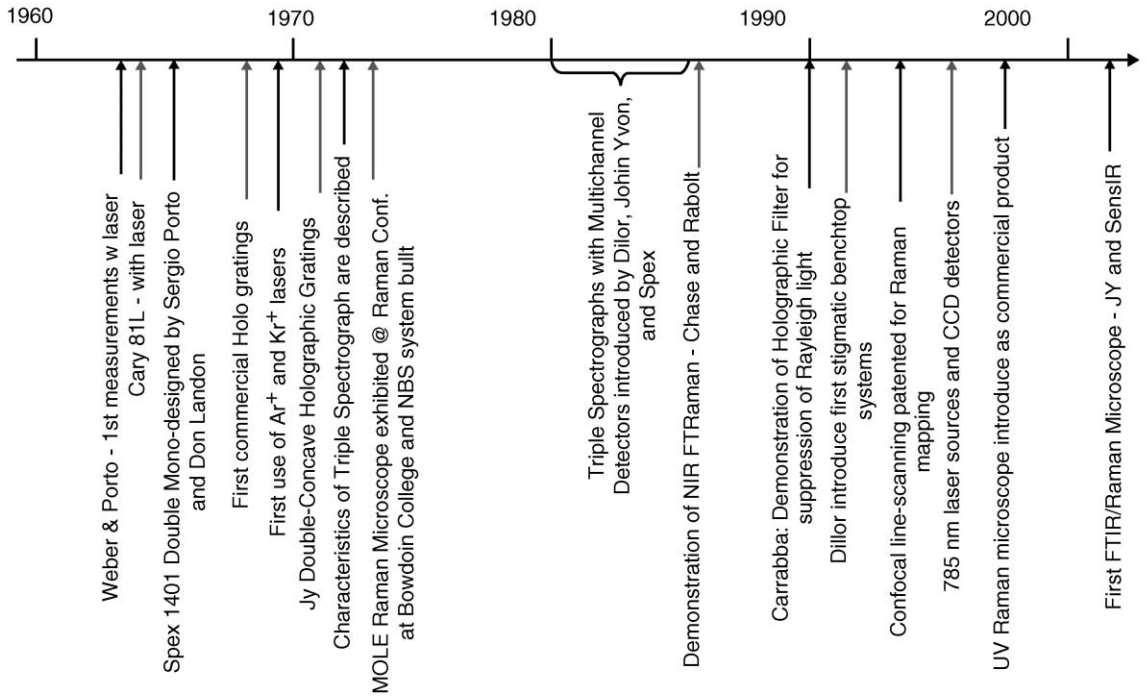


Figure 17 Timeline of Raman spectroscopy from the first incorporation of a laser source. Courtesy of Jobin Yvon.

accompanied also in the 1920s by early measurements, notable being those of C.V. Raman for which he was awarded the Nobel prize, soon followed in the 1930s by hundreds of chemistry papers describing this technique. Raman used for his early experiments filtered sunlight, a prism spectrometer, and visual observation. This was followed by application of filtered Hg emission and photographic plates. Lasers first saw application to Raman scattering in the 1960s, shortly after their innovation (see Figure 17).

In the simplest form of application of the Raman effect, UV or visible radiation illuminates a sample, coupling to excite electronic transitions near or on resonance. Spontaneous emission is shifted to a longer wavelength than that of the excitation (Stokes emission) and this is filtered and detected. In general, a broad spectrum of fluorescence is superimposed by weaker, narrow features that originate from vibration-electronic coupling (Raman emission). Inclusion of laser excitation increases the rate of

data collection and also enables nonresonant interaction of the stimulating radiation with the sample whilst maintaining data collection rates. Nonresonant interaction preferentially reduces electronic emission of fluorescence relative to that of the narrow featured Raman emission. Lasers applied to spontaneous Raman spectrometers in general emit in the visible or near-infrared running nonpulsed, such as Ar-ion, Nd:YAG or He:Ne.

Further enhancements are made to deliver spatial as well as vibration-specific information, through tight focusing of the stimulating radiation on to the sample in a confocal microscope arrangement, delivering spatial resolution down to 200 nm (Figure 18).

A limitation of standard Raman spectrometers and microscopes is the rate of signal collection, requiring concentrated samples for interrogation. The restricted signal level derives from a typical cross-section for the spontaneous emission of Raman radiation being $10^{-30}\,\mathrm{cm}^2$; 10^{10} to 10^{14} times less than that typical for spontaneous electronic fluorescence. Enhancements in the level of signal and the ratio of signal to background noise may be made through coherent laser stimulation of anti-Stokes Raman emission, known as coherent anti-stokes Raman scattering (CARS) (Figures 19 and 20).

It can be anticipated that bond-specific imaging will see wide industrial application in future, employing Raman as well as resonant infrared techniques. Application of CARS to microscopy involves two laser colors. A vibration transition is exited at the beat frequency of the two laser frequencies and the shorter laser wavelength is subsequently scattered from the sample to a yet shorter wavelength of emission. Such a coherent process has the advantage of signal scaling proportional to the square of emitters at the sample, thereby increasing the signal over that of spontaneous Raman scattering. Typical laser pump and Stokes powers range from less than 1 mW to 10s of mW, with laser repetition rates ranging from multi-kHz to greater than 100 MHz and typical pulse durations from 10s of picoseconds to 100 fs.

Infrared Methods

Infrared absorption spectroscopy provides a wealth of information on bond-specific sample concentration

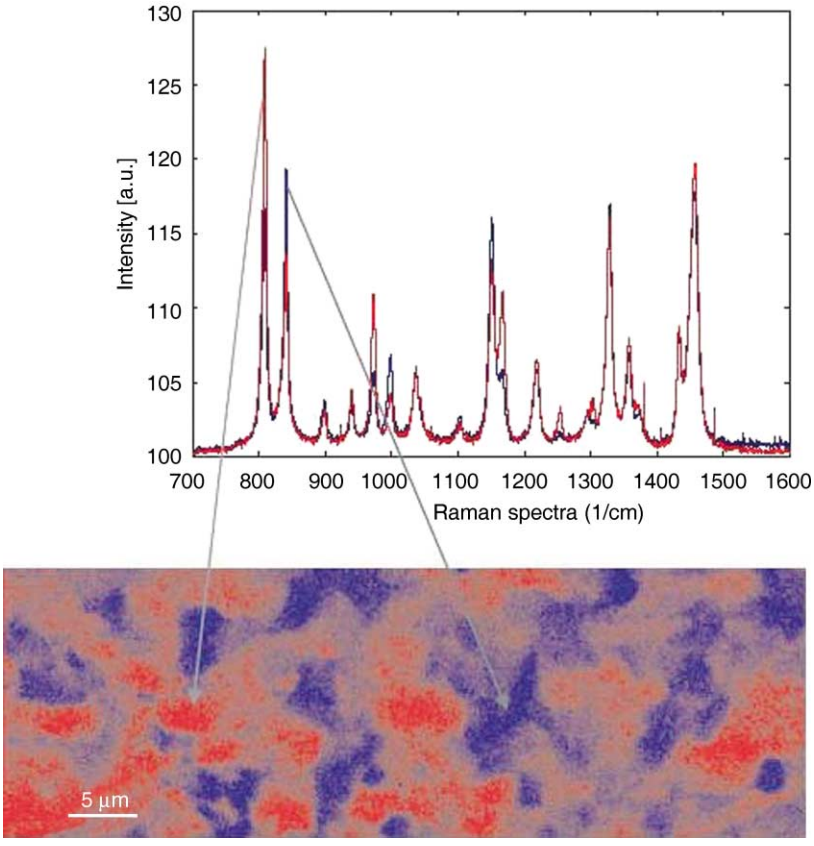


Figure 18 Confocal Raman image of semicrystalline polypropylene film. Two typical Raman spectra are shown. The ratio of the Raman lines at 809 and 841 cm increases with the degree of crystalline order, delivering a two-dimensional image (Hendra *et al.* 1995). Courtesy of Witec GmbH.

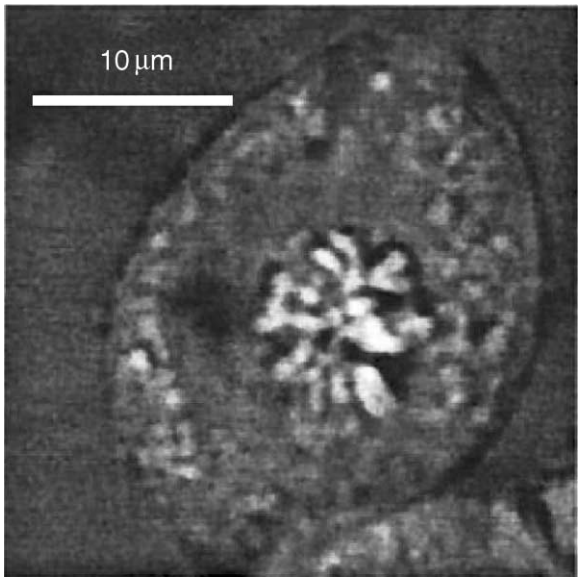


Figure 19 CARS image of a NIH 3T3 cell in metaphase, where Raman vibration stimulation is tuned to the PO_2^- symmetric stretching vibrational frequency at 1090 cm^{-1}. Reproduced from Cheng J-X, Kevin JYK, Zheng G and Xie X-S (2002) Laser-scanning coherent anti-Stokes Raman scattering microscopy and applications to cell biology. *Biophysical Journal* 83: 502–509, with permission from The Biophysical Society.

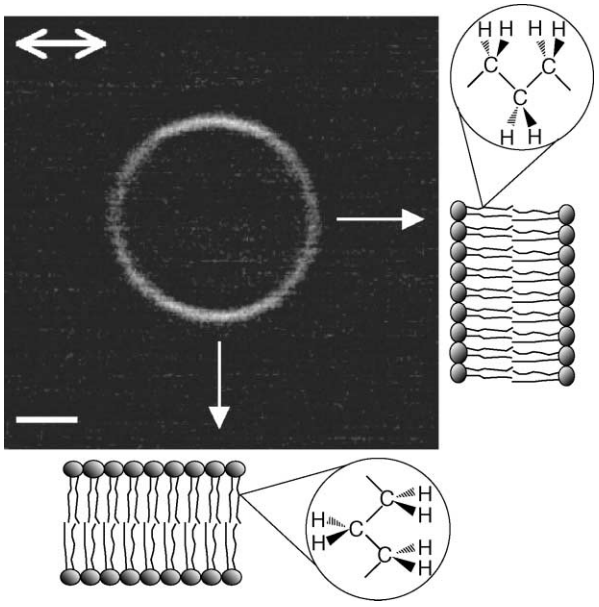


Figure 20 CARS microscope image of an erythrocyte ghost, consisting of a single lipid bilayer, with a Raman shift of 2845 cm^{-1}. The scale bar is of length 2 μm. Reproduced from Potma EO and Xie X-S (2003) Detection of single lipid bilayers with coherent anti-Stokes Raman scattering (CARS) microscopy. *Journal of Raman Spectroscopy* 36: 642–650, with permission from Wiley InterScience.

and has become a standard tool for the chemistry laboratory. Interaction between the radiation and sample is via resonant vibrational absorption and is otherwise analogous to the operation of a standard optical absorption spectrometer. Applied to a microscope arrangement, resolution on the order of 10 μm is standard. One restriction associated with infrared absorption spectroscopy applied to solution, is that absorption lineshapes are broadened and the spectra lose detail at room temperature, hindering identification of the contributing species.

A technique showing early promise for application in bond-specific molecular detection in solution, is coherent resonant infrared wavemixing. Here, cross coupling between vibrational modes are detected, promising information more detailed than that of infrared absorption. Figure 21 shows an example of triply vibrationally enhanced four-wave mixing, where two infrared laser beams mix to deliver cross peaks representing vibration mode coupling. Here, the 2D spectrum contains two sets of cross-peaks corresponding to the couplings of the two strong IR modes at 1943 and 2002 cm^{-1} of $Ni(CO)_2(PPh_3)_2$ and 1996 and 2068 cm^{-1} of $Ni(CO)_3PPh_3$. Concentrations in solution are on the order of 10 mM. Such methods provide a means sensitive to chemical structure without a requirement for fluorophores.

Current application requires complex laser equipment delivering microjoule pulse energies, tunable between 3 and 10 microns, with pulse durations on the order of 1 ps. However, following the example of other laser diagnostic techniques in application

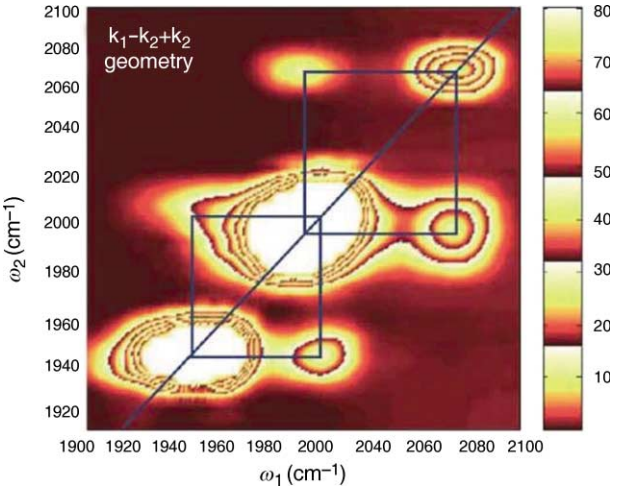


Figure 21 Triply vibrationally enhanced four-wave mixing of $Ni(CO)_2(PPh_3)_2$ and $Ni(CO)_3PPh_3$. Reproduced from Meyer KA, Besemann DM and Wright JC (2003) Coherent two dimensional spectroscopy with triply vibrationally enhanced infrared four-wave mixing. *Chemical Physics Letters* 381: 642–649, with permission of Elsevier.

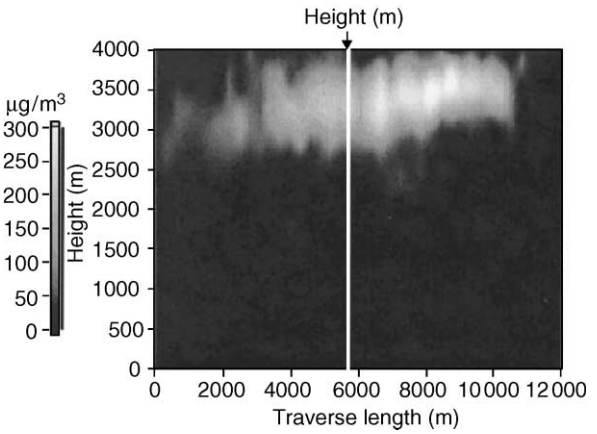


Figure 22 DIAL trace of sulfur dioxide distribution from Mount Etna. Reproduced with permission from Weibring P, Edner H and Svanberg S (2003) Versatile mobile lidar system for environmental monitoring. *Applied Optics* 42: 3583–3594. © 2003 Optical Society of America.

across industry today, as laser technology advances over this decade, we can expect such methods of probing chemical systems to see significant uptake, in particular in the bio-industries.

Laser Diagnostics and Treatment of Tumors

For clinical medicine, fluorescent detection is applied in the visualization of malignant tumors. Important natural chromophores emitting fluorescence in the blue-green region are elastin, collagen, NADH, and NAD^+, whilst administered agents such as porphyrins and phthalocyanines yield sharp peaks in the dark red.

Excitation is via UV or violet light. Where emission is monitored in both the blue-green and red regions of the spectrum and the ratio taken, an enhanced tumor differentiation is possible from that of normal tissue whilst simultaneously reducing sensitivity to geometry, illumination, and detection efficiency. Fiber optic probes are generally applied for delivery and collection of radiation.

Such detection can be combined with techniques of laser photodynamic therapy (PDT) to preferentially remove cancerous tissues. Typical laser wavelengths for photodynamic therapy are in the visible region of 600 to 700 nm with a pulse fluence of tens of joules/cm^2, a pulse duration in the nanosecond regime and a repetition rate from single shot to 10 Hz. Where combined with diagnostic detection, the PDT can be carried out by exciting the same administered sensitizer, with subsequent transfer of ground-state triplet oxygen to a toxic singlet state.

Remote Chemical Sensing

The atmosphere can be monitored by absorption and laser-induced fluorescence (LIF). Three-dimensional mapping of chemical distribution can be achieved with light detection and ranging (LIDAR), using a pulsed laser as a transmitter and an optical telescope as a receiver. Information collated includes pollution concentration, temperature, humidity, and wind velocity. Strong quenching prohibits LIF at atmospheric pressures; however, it does become practical to monitor for example, Li, Na, K, and Ca layers in the mesosphere. Pollution monitoring in the troposphere is most frequently performed with differential absorption LIDAR (DIAL). Range resolved optical transients are taken as a ratio of two laser wavelengths; on and just off resonance for the species of interest. Typical ranges for detection of SO_2, O_3, NO_2, NO, and Hg are 500 m to 5 km, making this technique suitable for urban and industrial monitoring. Sulfur dioxide is the most common pollutant emitted when fossil fuels are burned or sulfide ores are roasted and this technique sees application for the monitoring of industrial emissions. Figure 22 shows a DIAL trace recorded from Mount Etna, from which emissions of SO_2 are estimated at 75 tonnes/h. Laser wavelengths used are 300.02 and 299.30 nm for on and off resonance for absorption, respectively.

Important to the widespread application of LIDAR and LIF for atmospheric chemical monitoring, has been the advance in robustness and flexibility of the associated laser technologies. The laser technology applied for the above measurement is based on solid-sate Nd:YAG, nonlinearly converted via optical parametric amplification to wavelengths tunable between 220 nm and 4 μm, delivering up to 20 mJ in the mid-infrared at a repetition rate of 20 Hz and with bandwidths of less than 0.2 cm^{-1}, in a pulse duration of 4 ns.

See also

Imaging: Infrared Imaging. **Lasers:** Carbon Dioxide Lasers. **Microscopy:** Confocal Microscopy. **Scattering:** Raman Scattering. **Spectroscopy:** Raman Spectroscopy.

Further Reading

Allen SD, Bass M and Teisniger ML, Comparison of pulsed Nd:YAG and pulsed CO2 lasers for hole drilling in printed circuit board materials. CLEO 1982 Conference Summary.

Cheng J-X, Kevin Jia YK, Zheng G and Xie X-S (2002) Laser-scanning coherent anti-Stokes Raman scattering microscopy and applications to cell biology. *Biophysical Journal* 83: 502–509.

Cymer Inc., http://www.cymer.com

Cymer Inc., ASML, TRW, Xtreme Technologies GmbH, Fraunhoffer Institut fur Lasertechnik, Philips Extreme UV GmbH, Royal Institute of Technology Sweden, PLEX LLC.

Disco Corporation, http://www.disco.co.jp/

Exitech Ltd., http://www.exitech.co.uk/

Lamda Physik, http://www.lambdaphysik.de/

Meyer KA, Besemann DM and Wright JC (2003) Coherent two dimensional spectroscopy with triply vibrationally enhanced infrared four-wave mixing. *Chemical Physics Letters* 381: 642–649.

Perkin Elmer, www.perkinelmer.com

Phillips D (1997) Chemical mechanisms in photodynamic therapy with phthalocyanines. *Progress in Reaction Kinetics* 22(3–4): 175–300.

Potma EO and Xie X-S (2003) Detection of single lipid bilayers with coherent anti-Stokes Raman scattering (CARS) microscopy. *Journal of Raman Spectroscopy* 34: 642–650.

Powerlase Ltd., http://www.powerlase.com

Presstek Inc., Hudson, NH, http://www.presstek.com

Pronko PP, Dutta DK, Squier J, Rudd JV, Du D and Mourou G (1995) Machining of sub-micron holes using a femtosecond laser at 800 nm. *Optics Communications* 114: 106–110.

Reiser A (1989) *Photoreactive Polymers – The Science and Technology of Resists*. John Wiley & Sons.

Silverman PJ (2002) Intel Corporation. Presented at 1st Intl EUV Lithography Symposium, 15–17 October 2002, Dallas, USA.

Stuart BC, Feit MD, Herman S, Rubenchik AM, Shore BW and Perry MD (1996) Nanosecond-to-femtosecond laser-induced breakdown in dielectrics. *Physics Review B* 53(4): 1749.

Svanberg S (1996) Chemical sensing with laser spectroscopy. *Sensors and Actuators B* 33: 1–4.

Wang I, Bendsoe N, Klinteberg CAF, *et al*. (2001) Photodynamic therapy vs. cryosurgery of basal cell carcinomas: results of a phase III clinical trial. *British Journal of Dermatology* 144(4): 832–840.

Watson MN (1984) Laser drilling of printed circuit boards. *Circuit World* 11: 13.

Waynant RW (ed.) (2002) *Lasers in Medicine*. CRC Press.

Weibring P, Edner H and Svanberg S (2003) Versatile mobile lidar system for environmental monitoring. *Applied Optics* 42(18): 3583–3594.

WITEC, WITec Wissenschaftliche, Instrumente und Technologie GmbH, Hörvelsinger Weg 6, D-89081 Ulm, Germany, www.WITec.de

Yvon J, www.jobinyvon.com

Sum-Frequency Generation at Surfaces

M B Raschke and Y R Shen, University of California, Berkeley, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Under sufficiently intense illumination the optical properties of a material system depend nonlinearly on the strength of the electromagnetic field. Related to this nonlinear optical response is a large number of phenomena and fundamental processes which are discussed in various articles of this encyclopedia. This article will focus on second-harmonic generation (SHG) and sum-frequency generation (SFG) for the investigation of surfaces and interfaces. The intrinsic surface sensitivity of these techniques allows for investigations of surface properties not readily accessible by other spectroscopies. Here, the basic principles of these optical processes as well as their experimental implementation are discussed, and a summary of the applications to different material systems is given.

Among the various techniques employed for the characterization of surfaces and interfaces, those using light are particularly attractive. They are applicable *in situ* to all interfaces accessible by light, are nondestructive, and offer unprecedented time resolution. However, the penetration depth of optical radiation in matter is generally of the order of a wavelength, which makes isolation of the surface or interface contribution to the optical response from the bulk contribution difficult. In contrast, for reasons of symmetry, SHG and SFG are intrinsically surface sensitive in media with inversion symmetry, and hence the signal generated mainly originates from the topmost surface layer where the inversion symmetry is broken. By means of electronic or vibrational SHG or SFG spectroscopy, information on surface structure, chemical composition and bond or molecular orientation at solid and liquid interfaces can be deduced. To date, SFG and SHG have been well established as important tools for the investigation of surfaces and interfaces of solids ranging from metals and semiconductors to insulators and magnetic materials, liquids, polymers, biological membranes, and other systems. The studies are motivated by fundamental interests as well as applications in many areas such as heterogeneous catalysis, electrochemistry, device fabrication, epitaxial growth, and environmental science.

Theory

The nonlinear optical response of a material, in the electric-dipole approximation, can often be described by an induced polarization in the form of a power series,

$$\begin{aligned}\boldsymbol{P}(\omega) = \epsilon_0[&\chi^{(1)}(\omega)\boldsymbol{E}(\omega) \\ &+\chi^{(2)}(\omega=\omega_1+\omega_2)\boldsymbol{E}(\omega_1)\boldsymbol{E}(\omega_2) \\ &+\chi^{(3)}(\omega=\omega_1+\omega_2+\omega_3)\boldsymbol{E}(\omega_1)\boldsymbol{E}(\omega_2)\boldsymbol{E}(\omega_3)+\ldots]\end{aligned} \quad [1]$$

where $\boldsymbol{E}(\omega_i)$ is the optical field with frequency ω_i and $\chi^{(n)}$ is the nth-order linear ($n=1$) or nonlinear ($n>1$) susceptibility tensor. The second term, being the lowest-order nonlinear optical response, is responsible for sum-frequency generation and second-harmonic generation (with $\omega_1=\omega_2$). In the electric-dipole approximation, all even-order terms ($\chi^2, \chi^{(4)}$, etc.) are forbidden in media with inversion symmetry. At the surface or interface, however, the inversion symmetry is necessarily broken and hence $\chi^{(2)} \neq 0$. This makes SHG and SFG surface-sensitive and specific. While electric-quadrupole and magnetic-dipole contributions from the bulk may not be totally negligible, in many cases it has been shown that the surface contribution clearly dominates.

Surface SFG and SHG is best described by radiation from a surface polarization $\boldsymbol{P}_s^{(2)}(\omega)$ induced in a thin sheet with dielectric constant ε' sandwiched between two linear media (1) and (2) as shown schematically in Figure 1. This surface polarization has an in-plane wavevector component equal to the sum of the in-plane wavevector components of the incident fields:

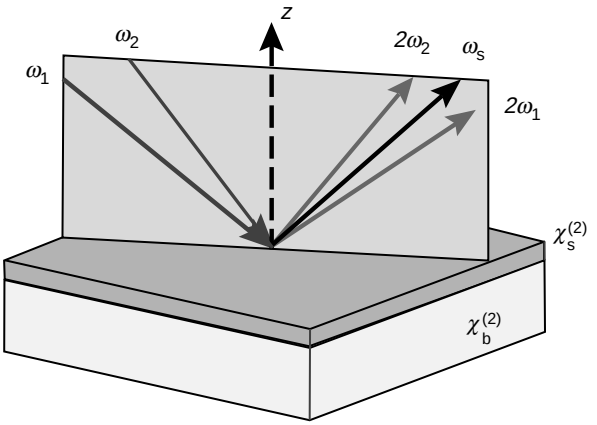


Figure 1 Schematic of SHG and SFG geometry in reflection and transmission from an interface.

$\boldsymbol{k}_\parallel(\omega) = \boldsymbol{k}_\parallel(\omega_1) + \boldsymbol{k}_\parallel(\omega_2)$. The ratio of $\boldsymbol{k}_\parallel(\omega)$ to $\boldsymbol{k}(\omega)$ then determines the direction of the sum-frequency output that appears in both transmission and reflection. The reflection geometry is often used in order to minimize the bulk contribution.

With $\boldsymbol{P}_s^{(2)}$ as the source term, the sum-frequency output can be obtained from the solution of the wave equation. With the proper boundary conditions the sum frequency (SF) intensity is given by:

$$\begin{aligned}I(\omega) = &\frac{8\pi^3\omega_s^2\sec^2\theta_\omega}{\hbar c^3\sqrt{\epsilon_1(\omega)\epsilon_1(\omega_1)\epsilon_1(\omega_2)}} \\ &\times |\boldsymbol{e}^\dagger(\omega)\cdot\boldsymbol{\chi}_s^{(2)} : \boldsymbol{e}(\omega_1)\boldsymbol{e}(\omega_2)|^2 I_1(\omega_1)I_2(\omega_2)\end{aligned} \quad [2]$$

In this expression, $\chi_s^{(2)}$ is the surface nonlinear susceptibility defined by $\boldsymbol{P}_s^{(2)} = \varepsilon_0\boldsymbol{\chi}_s^{(2)} : \boldsymbol{E}(\omega_1)\boldsymbol{E}(\omega_2)$; the quantities $\boldsymbol{e}(\omega_i)$ correspond to the unit polarization vectors after the appropriate Fresnel correction, $\boldsymbol{e}(\omega_i) \equiv \boldsymbol{F}(\omega_i)\hat{\boldsymbol{e}}(\omega_i)$, with $\boldsymbol{F}(\omega_i)$ being the transmission Fresnel factor and $\hat{\boldsymbol{e}}(\omega_i)$ the unit polarization vector of $\boldsymbol{E}(\omega_i)$; θ_ω denotes the SF output angle with respect to the surface normal, $\varepsilon(\omega_i)$ is the dielectric constant at frequency ω_i, and I_i is the input pump intensity at ω_i.

The surface nonlinear susceptibility $\chi_s^{(2)}$ is a third-rank tensor. In Cartesian coordinates it has 27 tensor elements ($\chi_{ijk}^{(2)}$), many of which could vanish or depend on others due to the surface structural symmetry. As an example, $\chi_{ijk}^{(2)}(\omega=\omega_1+\omega_2)$ for an isotropic surface, with the z-direction defined by the surface normal, has only four independent nonvanishing elements: $\chi_{xxz}^{(2)} = \chi_{yyz}^{(2)}, \chi_{xzx}^{(2)} = \chi_{yzy}^{(2)}, \chi_{zxx}^{(2)} = \chi_{zyy}^{(2)}$, and $\chi_{zzz}^{(2)}$. Different combinations of input and output beam polarizations in SFG measurements are often used to deduce values for the non-vanishing elements. Such measurements then allow the determination of surface symmetry or surface molecular orientation. Being a third-rank tensor, $\chi_s^{(2)}$ can reflect up to three-fold rotational symmetry. $\chi_{s,ijk}^{(2)}$ is related to the molecular nonlinear polarizability or hyperpolarizability $\alpha_{\xi\eta\zeta}^{(2)}$, where $\hat{\xi}$, $\hat{\eta}$, and $\hat{\zeta}$ define the molecular coordinates, through the coordinate transformation

$$\chi_{s,ijk}^{(2)} = N_s\left\langle(\hat{\boldsymbol{i}}\cdot\hat{\boldsymbol{\xi}})(\hat{\boldsymbol{j}}\cdot\hat{\boldsymbol{\eta}})(\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{\xi}})\right\rangle\alpha_{\xi\eta\zeta}^{(2)} \quad [3]$$

The angular brackets denote an average over the molecular orientational distribution, and N_s is the surface molecular density. For simplicity, microscopic local-field correction is neglected in eqn [3]. Knowing $\chi_{s,ijk}^{(2)}$ and $\alpha_{\xi\eta\zeta}^{(2)}$ thus permits deduction of information on the orientational distribution.

An explicit expression for the nonlinear optical polarizability is obtained from a second-order quantum mechanical perturbation calculation:

$$\alpha^{(2)}_{s,ijk}(\omega_s) = -\frac{e^3}{\hbar^2}\sum_{g,n,n'}\left[\frac{\langle g|r_i|n\rangle}{\omega_s-\omega_{ng}+i\Gamma_{ng}} \times\left(\frac{\langle n|r_j|n'\rangle\langle n'|r_k|g\rangle}{\omega_2-\omega_{n'g}+i\Gamma_{n'g}} + \frac{\langle n|r_k|n'\rangle\langle n'|r_j|g\rangle}{\omega_1-\omega_{n'g}+i\Gamma_{n'g}}\right)+\cdots\right]\rho^{(0)}_g \quad [4]$$

This expression shows how the nonlinear polarizability or susceptibility depends on material parameters such as the dipole transition moments $\langle n|r_i|g\rangle$ and energy levels. It contains a sum over eight resonant terms. The quantities ω_{ng} and Γ_{ng} are the frequencies and half-widths for the transitions between quantum states $|n\rangle$ and $|g\rangle$, and $\rho^{(0)}_g$ denotes the population in $|g\rangle$. It can be seen that $\alpha^{(2)}$, and hence the SF output, should be resonantly enhanced when the pump frequency ω_1 or ω_2 and/or the sum frequency $\omega_1+\omega_2$ approach resonance. The resonant enhancement provides spectral information and makes SHG and SFG effective spectroscopic tools. The resonances could be electronic or vibrational, or more generally, any surface characteristic transition.

Experimental

Here we deal with basic considerations concerning the optical setup for SHG or SFG experiments in reflection from a surface as is shown schematically in Figure 2. In general, the pump radiation from pulsed laser sources is directed onto the surface. High-power tunable light pulses can be obtained from optical parametric generation and amplification (OPG/OPA) together with harmonic, sum- or difference-frequency generation, preferably pumped by picosecond or femtosecond lasers with high repetition rates. By this means radiation tunable from near UV at ~200 nm to mid IR at ~18 μm can be generated. For SFG the input pulses with frequencies ω_1 and ω_2 are directed to overlap spatially and temporally on the sample. The sample could be the surface of a liquid or solid in air, a single crystalline surface in ultrahigh vacuum, buried interfaces, etc. Polarizers and half-wave plates allow for tuning of incident power and polarizations at the sample. Together with the polarizer in the output path, different input/output polarization combinations can be chosen. In the detection system the signal has to be discriminated against reflected and scattered pump light. This is achieved by spatial and spectral filtering with apertures, dielectric mirrors, and interference and color glass filters. An optional monochromator can also be used for additional stray-light suppression. The signal is then detected by a photomultiplier tube and gated electronics are used for signal integration or photon counting.

From eqn [2] the expected signal strength for SHG/SFG can be estimated. With a typical value of $\chi^{(2)}_s$ of 10^{-21} m^2 V^{-1}, a single 1-μm pump pulse, incident at $\theta = 45°$ having pulse energy $E = I\,A\tau = 100$ μJ, beam cross-section $A = 1$ mm^2, and pulse duration $\tau = 10$ ps, can generate about 10^3 photons per pulse of SHG. By means of photon counting a minimal count rate of 10^{-3} photons/pulse

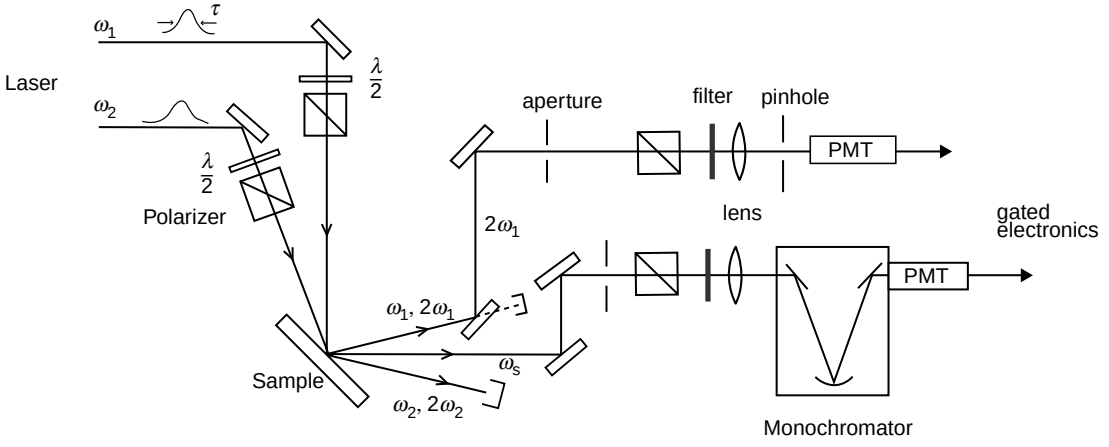


Figure 2 Schematic of experimental setup for second-harmonic and sum-frequency generation.

can be achieved which allows the detection of $\chi_s^{(2)}$ as small as 10^{-24} $m^2\,V^{-1}$. The spectral resolution is limited by the spectral width of the pump pulses. In order to obtain accurate spectral data over a wide tuning range or to obtain absolute values for $\chi_{s,ijk}^{(2)}$ of the material investigated, a reference material such as *z*-cut quartz with known values of χ is to be measured for comparison.

The simplest kind of experiments are SHG at a fixed pump frequency. From the signal variation in response to surface modification one can probe kinetics and dynamics of adsorption, desorption, diffusion, surface melting, phase transitions, etc. By tuning the pump wavelength, second-harmonic spectroscopy can give information on surface electronic states. For surface vibrational spectroscopy, however, IR-vis SFG must be employed. In the latter case, tunable IR input is mixed with visible input to yield a SF signal in the visible region.

Surface Specificity

In many cases, the surface contribution to SFG or SHG from a centrosymmetric medium clearly dominates over the bulk contribution. This generally occurs when the surface or interface is composed of a polar oriented molecular layer. An example is shown in Figure 3, where the vibrational spectra in the C–H stretch region of three buried solid–liquid interfaces are presented. The spectrum for hexadecane ($C_{16}H_{34}$)/fused silica shows no C–H peaks. This is in spite of the fact that liquid hexadecane has strong infrared and Raman activity in this spectral range. The result indicates that the bulk signal contribution is negligible. In contrast, strong resonances are observed from the hexadecane/OTS/silica system, where OTS is a monolayer of octadecyltrichlorosilane ($CH_3(CH_2)_{17}SiCl_3$) chemisorbed on the fused silica plate. A similar spectrum is obtained for the CCl_4/OTS/silica interface. This clearly indicates that it is the contribution from the OTS monolayer that dominates the spectra.

Another example is described in Figure 4 where the SFG spectrum in reflection from a free liquid water surface is shown. The resonant features between 3000 and 3600 cm^{-1} are due to the O–H stretches of hydrogen-bonded OH groups. The sharp peak at 3700 cm^{-1} can be identified with the free OH bonds. Since no dangling OH bonds can exist below the surface its presence indicates that the spectrum must originate from the topmost layer of water molecules. Furthermore it can be shown that the surface water molecules are oriented with one of the OH bonds directed out of the liquid. In addition to the above examples, the surface sensitivity and specificity has been demonstrated and successfully applied to a large number of systems as will be shown below.

A note on the definition of surface or interface is in order. In SHG and SFG, the surface or interface layer refers to a thin layer between two adjacent bulk media that has a different structure from the bulk media and lacks inversion symmetry. If molecules in the surface layer are polar-ordered, then SHG and

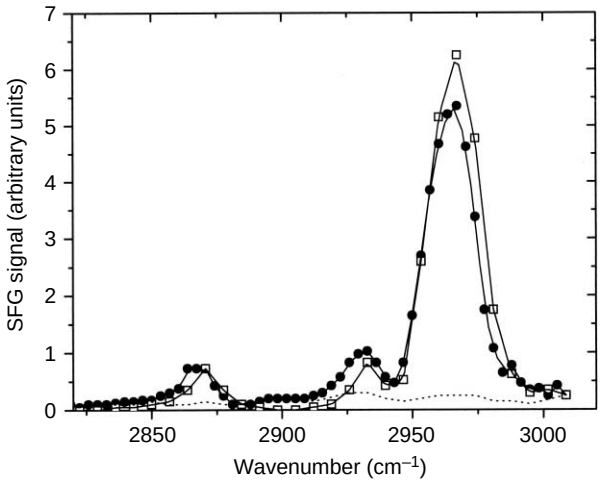


Figure 3 SFG spectra of different interfaces obtained with *ssp* polarization combination: hexadecane/silica (dashed), CCl_4/ OTS/silica (solid circles), and hexadecane/OTS/silica (open squares). OTS refers to a monolayer of octadecyltrichlorosilane.

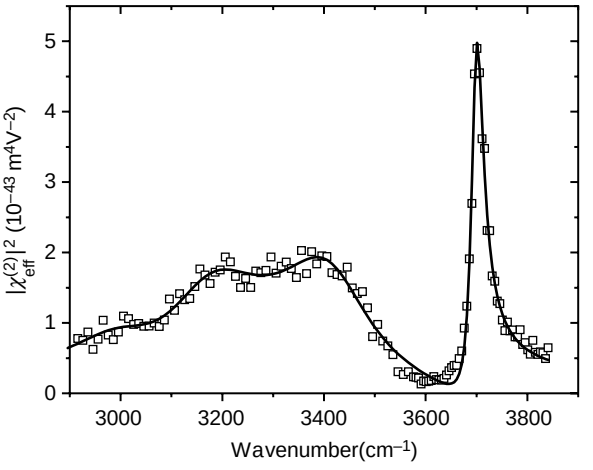


Figure 4 SFG spectrum of the liquid–vapor interface of pure water (*ssp* polarization combination). The mode at 3700 cm^{-1} is due to the dangling OH bonds at the surface.

SFG are often dominated by contributions from the surface layer. However, the bulk contribution is not strictly zero even if the media have inversion symmetry, as $\chi^{(2)}$ can arise from electric-quadrupole and magnetic-dipole contributions. Separation of surface and bulk contributions in SHG and SFG is a subtle problem in general. Experimentally, how surface modification affects the signal can be used to evaluate whether or not the surface contribution dominates.

A special case concerns materials that lack bulk inversion symmetry, for example, III–V and II–VI semiconductors such as GaAs and ZnSe. In this case, the bulk contribution to SHG and SFG is electric-dipole allowed and may be very significant. With an appropriate choice of polarization combinations and surface orientation, it can nevertheless be well suppressed, leaving the surface contribution dominant.

Applications

Both SFG and SHG have been applied to a wide variety of surface and interfacial systems. To illustrate their capabilities and versatility as surface spectroscopic tools, the focus of this summary will be on examples where the information deduced cannot readily be obtained by other techniques. (For details, readers are referred to the review articles cited in Further Reading and the original papers cited therein.)

SFG and SHG have been used for the study of adsorbates at surfaces even under ambient conditions. The adsorbate alters electronic or structural properties of the substrate and the signal is thus related to the surface density of the adsorbate and its orientation.

Elementary surface reactions, in particular adsorption, dissociation, or desorption of simple molecules such as CO, hydrogen, oxygen, or water on metal and semiconductor single-crystal surfaces under ultra-high-vacuum conditions, have been investigated. Sensitivity into the sub-percent monolayer regime is achieved and allows the identifictation of competing reaction channels and adsorbate dynamics on surfaces. Such studies can help to understand epitaxial growth or heterogeneous catalysis because the techniques are applicable at any gas pressure and surface temperature. Related to this is the study of surface diffusion, which can be probed following the temporal decay of a diffracted SHG signal off a submonolayer adsorbate grating formed by laser-induced desorption.

Making use of the spectroscopic capabilities of SHG, the surface electronic structure of selected metals and semiconductors, in particular silicon, were studied with an emphasis on the effects of temperature, crystallographic orientation and different adsorbates. Among buried interfaces, the Si/SiO_2 interface has been the subject of intense investigations and the sensitivity of SHG to static electric fields at the interface and to inhomogeneous strain was also demonstrated.

Different in-plane surface symmetries associated with different crystallographic orientations or surface reconstructions yield different SHG and SFG responses. This allowed for the investigation of surface reconstructions (e.g., $Si(111)2\times1 \rightarrow 7\times7$), order–disorder transitions (e.g., $Si(111)7\times7 \rightarrow 1\times1$, Au/Si(111)) and surface melting (e.g., single-crystal ice, Si(111), Ge(111)), providing information on latent heat, superheating, and its dynamics using pump-probe techniques. Magnetization-induced SHG was also used for probing surface and interface magnetism of ferromagnetic metals and bimetallic systems.

Many unique applications of spectroscopic SFG have been developed based on its capability to study large molecules, in particular under ambient conditions. Surface vibrational spectra with different input/output polarization combinations and sample geometries provide information on orientation, conformation and alignment of surface molecules and the composition and structure of the surface layer. The study of self-assembled monolayers (e.g., OTS on quartz) to deduce the orientation and conformation of the alkyl chains in the monolayer is an example. The effect of molecular density and temperature on the spectrum gives information on intermolecular interactions and phase transitions. Another example concerns the adsorption of liquid crystal molecules on nanostructured polymer substrates. Both the chain orientation at the polymer surface and its effect on the alignment of adsorbed liquid crystal molecules can be addressed.

Surface structures of liquids, especially those of pure liquids, are of great interest in science and technology, and SFG is a unique spectroscopic tool for probing liquid surfaces and interfaces. For example, the SFG vibrational spectrum for the water–vapor interface in **Figure 4** gives information on the density and orientation of the surface water molecules at the water–vapor interface. The surface layer was found to be more ordered compared to the bulk as was also observed for various other organic liquids studied with SFG. In other cases soluble and insoluble organic molecules at a liquid–vapor,

liquid–liquid, or liquid–solid interface have also been studied. They appear as adsorbed monolayer films and their structure and phase behavior are relevant to many applications. Identification of the adsorbed molecular species at solid–liquid interfaces is also important for understanding electrochemical processes. Various studies with SFG have addressed bonding of, e.g., CO, CN^-, and SCN^- at electrodes and its variation with electrode potential.

SFG spectroscopy also finds unique applications in studies of polymer surfaces and interfaces most relevant to modern science and technology. The observed surface vibrational spectra of polymers provide information on surface composition, molecular orientation, and conformation of neat and blended polymers. The effect of environment and surface treatment on the surface structure can also be studied. Investigation of the biological functions of complex molecular systems has attracted increasing interest, and SHG and SFG have been applied successfully to selected systems ranging from the isomerization of retinal – a molecule involved in the vision process – to functional aspects of model membranes.

Both SHG and SFG can be combined with microscopy techniques for surface microscopic studies. Microscopic imaging can be achieved by rastering the probe beam or, preferably, the sample position. Near-field SFG/SHG spectroscopy has recently been developed.

Performing surface SHG and SFG experiments in the time domain can provide information on surface dynamics. Pump-probe experiments allow for measurements of energy relaxation and phase coherence of excitations such as surface electronic states and surface phonons and vibrations.

Outlook

Although SHG and SFG were established as surface analytical tools more than 15 years ago, it is still an active field of research with much potential that has not yet been explored. As an example, it has only recently been demonstrated that doubly resonant infrared–visible SFG, as a two-dimensional spectroscopy, can give access to couplings between vibrational modes and surface electronic transitions. With the commercial availability of suitable laser sources, the techniques can be extended to a much wider range of applications. This includes possible investigations of ultrafast surface dynamics, nanostructures, and biological systems.

Further Reading

Bennemann KH (ed.) (1998) *Nonlinear Optics in Metals*. Oxford, UK: Clarendon Press.

Bloembergen N (1977) *Nonlinear Optics*. New York: Benjamin.

Boyd RW (1993) *Nonlinear Optics*, 2nd edn. San Diego, CA: Academic Press.

Dumas P, Weldon MK, Chabal YJ and Williams GP (1999) Molecules at surfaces and interfaces studied using vibrational spectroscopies and related techniques. *Surface Review & Letters* 6(2): 225–255.

Guyot-Sionnest P and Harris AL (1995) Surface vibrational dynamics probed by sum frequency generation. In: Ho W and Dai HL (eds) *Laser Spectroscopy and Photochemistry of Metal Surfaces*, pp. 405–458. Singapore: World Scientific.

Guyot-Sionnest P, Superfine R, Hunt JH and Shen YR (1988) Vibrational spectroscopy of a silane monolayer at air/solid and liquid/solid interfaces using sum-frequency generation. *Chemical Physics Letter* 144: 1–5.

Heinz TF (1991) Second-order nonlinear optical effects at surfaces and interfaces. In: Stegeman GI and Ponath H-E (eds), *Nonlinear Surface Electromagnetic Phenomena*, pp. 353–416. Amsterdam, The Netherlands: North-Holland.

Matthias E and Träger F (eds) (1999) *Applied Physics B, Special Issue: Nonlinear Optics at Interfaces*, Volume 210.

McGilp JF (1999) Optical second-harmonic generation as a semiconductor surface and interface probe. *Physica Status Solidi (a)* 175: 153–167.

Miranda PB and Shen YR (1999) Liquid interfaces: A study by sum-frequency vibrational spectroscopy. *Journal of Physical Chemistry B* 103: 3292–3307.

Reider GA and Heinz TF (1995) Second-order nonlinear optical effects at surfaces and interfaces: recent advances. In: Halevi P (ed.) *Photonic Probes of Interfaces*, 413. Amsterdam, The Netherlands: Elsevier.

Shen YR (1984) The Principles of Nonlinear Optics. New York: Wiley.

Shen YR (1994) Nonlinear optical studies of polymer interfaces. *International Journal of Nonlinear Optical Physics* 3: 459–468.

Shen YR (1994) Surface spectroscopy by nonlinear optics. In: Ḧansch TW and Inguscio M (eds) *Frontiers in Laser Spectroscopy*, pp. 139–165. Amsterdam, The Netherlands: North-Holland.

Shen YR (2001) Vibrational spectroscopy of water at interfaces. In: Morra M (ed.) *Water in Biomaterials Surface Science*, Chapter 8, pp. 215–244. Chichester, UK: Wiley.

Wei X and Shen YR (2001) Motional effect in surface sum-frequency vibrational spectroscopy. *Physical Review Letter* 86: 4799–4802.

POLARIZATION

Contents

Introduction
Matrix Analysis

Introduction

J M Bennett, Michelson Laboratory, China Lake, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Polarization is one property of light waves and will be defined after a brief overview of the properties of light. Light is part of the electromagnetic spectrum of waves that have both particle-like and wave-like properties. Light waves carry energy in the form of photons which act like particles; the photon energy increases in proportion to the frequency of the wave. The particle-like properties of light and other electromagnetic waves are described by quantum mechanics. Light also acts like transverse waves that travel in straight lines in air and vacuum and can be described by classical electromagnetic theory. Polarization deals with the wave-like properties of light, and is described mathematically by Maxwell's equations, the key relations in electromagnetic theory. Polarization effects occur when light interacts with matter. In order to understand polarization, there will be a brief introduction to the subject of electromagnetic waves.

For a more in-depth treatment of the material in this article, the reader is directed to the Further Reading list at the end of this article and in particular to the two chapters on 'Polarization' and 'Polarizers' written by the author in Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chapter 5 and vol. II, chapter 3. New York: McGraw-Hill Inc.

Electromagnetic waves have both electric and magnetic fields associated with them. These are vibrations in directions perpendicular to the direction the wave is traveling, i.e., the direction of propagation. $\vec{E}$ represents the vector of the electric field and $\vec{H}$, perpendicular to $\vec{E}$, represents the magnetic field vector. Both these vectors are complex and have real and imaginary parts. Polarization effects are always associated with the $\vec{E}$ vector. Specifically, the plane of polarization is defined as the plane in which the $\vec{E}$ vector is vibrating. Waves having different amplitudes, phases, or angular orientations (azimuths) of their electric or magnetic vectors can be combined by conventional vector addition methods. Also the $\vec{E}$ vector of a particular vibration can be resolved into two components in mutually perpendicular directions that are vibrating in phase.

If a light source such as the sun, a candle flame, or an electric light bulb is considered on a microscopic scale, each vibrating atom or molecule emits linearly polarized light (see the definition below). But the individual atoms or molecules do not act together, so their vibrations have no fixed phase relationships to each other and they cannot be added into a single linearly polarized beam. Thus, we call light from these sources unpolarized. In an unpolarized light beam, the $\vec{E}$ vector vibrates in all directions perpendicular to the direction of propagation. If a snapshot is taken at a particular instant of time, different parts of the beam will have $\vec{E}$ vectors vibrating with different amplitudes and phases at different angles to each other, but all in a plane perpendicular to the direction of propagation. In the most common convention used in optics, the wave travels in the $+z$ direction in a right hand coordinate system and the $\vec{E}$ vectors are all vibrating at various angles in the $x-y$ plane. The angle of vibration is measured from the positive x axis in a counterclockwise direction when the observer is looking against the direction of propagation of the light beam.

For linearly polarized or plane polarized light, if a snapshot of a light beam is taken at a particular instant, the $\vec{E}$ vector will be vibrating at a certain angle in the $x-y$ plane. As time (or position on the traveling wave) varies, the amplitude of the $\vec{E}$ vector will vary in a sinusoidal manner, but the vibration will remain at the same angle in the $x-y$ plane. As an example, the real part of the electric vector of a linearly polarized beam that is vibrating in the $+x$ and $-x$ directions and traveling in the $+z$ direction is

given by the relation:

$$\vec{E}_x(z,t) = \hat{i}E_0 \cos(kz - \omega t) \quad [1]$$

where t is the time, $\hat{i}$ is a unit vector in the $+x$ direction, E_0 is the amplitude of the vibration, $\vec{k}$ is the propagation vector in the direction the wave is traveling, i.e., the z direction (its magnitude is $2\pi/\lambda$), and $\omega = 2\pi f$ where f is the vibration frequency of the wave. In free space, $\omega = 2\pi c/\lambda$ where c is the velocity of light and λ is the wavelength of the vibration.

Light can also be circularly or elliptically polarized. Circularly polarized light is produced by adding the electric vectors of two waves, each having the same amplitude but which are 90° out of phase; the resulting vibration sweeps out a circle in the x–y plane. When the first wave is vibrating in the $+x$ and $-x$ directions (eqn [1]), the second wave is vibrating in the $+y$ and $-y$ directions:

$$\vec{E}_y(z,t) = \hat{j}E_0 \sin(kz - \omega t) \quad [2]$$

and is added to the first wave. The resultant wave is the sum of eqns [1] and [2]:

$$\vec{E} = E_0[\hat{i}\cos(kz - \omega t) + \hat{j}\sin(kz - \omega t)] \quad [3]$$

The direction of vibration of this wave will rotate in a circle as either the time increases or as the distance along the wave increases, but the amplitude of the vibration will not change. As time increases, the vibration will make a circle in the clockwise direction (to negative angles). This light is defined to be right circularly polarized. If the + sign between the two parts of eqn [3] is changed to a minus sign, as time increases, the vibration will make a circle in the counterclockwise direction (to positive angles) and the light is now defined to be left circularly polarized. These definitions are for conventional (traditional) optics. However, the opposite definitions with right (left) circularly polarized light defining circles in the counterclockwise (clockwise) direction are also in use. In modern physics, there is still another convention that defines right (left) circularly polarized light as having negative (positive) helicity.

Right and left elliptically polarized light beams have the same angle conventions as for circularly polarized beams but are produced by adding two electric vectors that have different amplitudes. In eqns [1] and [2], the amplitude terms will be, for example, E_1 and E_2 instead of E_0 and eqn [3] can no longer be used. There is now a major axis and a minor axis of the ellipse. If $E_1 > E_2$, the major axis will be along the x axis; for $E_1 < E_2$, the major axis will coincide with the y axis.

The preceding discussion has dealt entirely with the electric vector of the electromagnetic field. However, one cannot observe the $\vec{E}$ vector. The irradiance (energy per unit area per unit time), $\vec{E}\vec{E}^*$, is what can be observed visually and measured by electronic detectors. Thus, measurements of the polarization are irradiance measurements. Light transmitted by a polarizer is called the transmittance of the polarizer; similarly, light reflected from a polarizer is called the reflectance of the polarizer.

One of the most important parts of the subject of polarization is how to produce linearly polarized light starting with unpolarized light. This is done by using polarizers, as discussed in the section on polarizers below. Certain materials have special properties that make them able to polarize light. Depending on the application, different kinds of polarizers are preferred. Optics textbooks by Hecht and Guenther discuss the most important polarizers and two articles by Bennett describe many kinds of polarizers including special ones (see the Further Reading at the end of this article for full references). Only the basic principles will be described here.

Sunlight scattered by air molecules in the atmosphere (Rayleigh scattering) is also partially linearly polarized. The air molecules act like tiny dipoles and vibrate when they absorb sunlight. They emit radiation that is polarized in certain directions relative to the vibration direction. When the viewer is at a 90° angle with respect to the sun, the polarization of the skylight is a maximum. Rayleigh scattering is the subject of several books and will not be further discussed here.

Retarders are used to change linearly polarized light into circularly or elliptically polarized light and compensators, which are a form of retarders, can analyze an unknown type of polarized light and determine its composition. They are discussed below.

Polarimetry and ellipsometry are closely related techniques that are used to determine the optical properties of a material by shining a known type of polarized light on the material at non-normal incidence and analyzing the polarization properties of the light after it has been reflected from the material. These subjects are treated in other articles in this encyclopedia and in several references at the end of this article.

Changes can be produced in the optical properties of some materials by applying an electric field, a magnetic field, an acoustic field, or another form of variable pressure. The materials changed in these ways are said to be electro-optic, magneto-optic, or piezo-optic. The changes in the optical properties modify some parameter of a light wave passing through a material or reflecting from it.

The amplitude, phase, frequency, polarization, or direction of the light wave can be modified. In this article, we are only concerned with modifications that produce changes in the polarization; these are discussed towards the end of this article.

Matrix methods have been developed to handle problems involving polarization and there is also a visual representation of the matrix algebra that is based on the Poincaré sphere. Both topics are covered at the end of this article.

Polarizers

Basic Relations for Polarizers

A linear polarizer is anything which, when placed in an incident unpolarized beam, produces a beam of light whose electric vector is vibrating primarily in one direction with only a small component vibrating in the direction perpendicular to it. The transmittance T of the linear polarizer is

$$T = \tfrac{1}{2}(T_1 + T_2) \qquad [4]$$

where T_1, the principal transmittance of the polarizer, is $\gg T_2$, the transmittance of the polarizer at 90° to the principal transmittance. Thus, a perfect polarizer would transmit only 50% of an incident unpolarized beam.

If a linear polarizer is placed in a linearly polarized beam and is rotated about an axis parallel to the beam direction, the transmittance will vary between a maximum value T_1 and a minimum value T_2 according to the law:

$$T = (T_1 - T_2)\cos^2\theta + T_2 \qquad [5]$$

where θ is the angle between the plane of vibration of the principal transmittance and the plane of vibration of the electric vector in the incident beam.

The extinction ratio ρ_P of a polarizer is defined as

$$\rho_P = \frac{T_2}{T_1} \qquad [6]$$

and the degree of polarization of a polarization P is

$$P = \frac{T_1 - T_2}{T_1 + T_2} \qquad [7]$$

When two identical polarizers are placed in an unpolarized beam, and the directions of their principal transmittances, T_A and T_B, are inclined at an angle θ to each other, the transmittance of the pair will be

$$T_\theta = \tfrac{1}{2}(T_1^2 + T_2^2)\cos^2\theta + T_1 T_2 \sin^2\theta \qquad [8]$$

Thus, when the directions of the principal transmittances are aligned, $T_\parallel = \frac{1}{2}(T_1^2 + T_2^2)$, and when they are perpendicular, $T_\perp = T_1 T_2$.

Birefringent Materials (Calcite)

The majority of high-quality polarizers are made from calcite. This is a birefringent (doubly refracting) crystalline material that is uniaxial (i.e., there is one preferred direction in the crystal). A birefringent material acts differently for light going in different directions through the crystal. For example, if an unpolarized light beam passes through the crystal in a certain direction, it will be split into two spatially separated beams that are parallel but are linearly polarized at right angles to each other. A uniaxial crystal has an optic axis (i.e., a certain direction in the crystal). When light rays travel parallel to the optic axis, they travel at the same velocity and there is no difference between them. When light passes through the crystal in other directions, the ray whose vibration direction is perpendicular to the optic axis is governed by the ordinary laws of geometrical optics (the same as for isotropic materials) and is called the ordinary ray. The ray whose vibration direction is parallel to the optic axis does not follow the normal geometrical optics laws and is called the extraordinary ray. One ray travels faster than the other, so there is a phase retardation for one ray relative to the other. This is the principle of a retarder or retardation plate (see the next section).

Calcite is a negative uniaxial crystal which means that the refractive index for the ordinary ray is larger than the refractive index for the extraordinary ray. When the ordinary ray enters a block of calcite from air at non-normal incidence, it is bent more than the extraordinary ray.

Calcite can be easily cleaved along three distinct planes, making it possible to produce rhombs of the form shown in Figure 1. The optic axis, going in the HI direction, makes equal angles with all three faces at point H. Any plane, such as DBFH, which contains the optic axis and is perpendicular to the two opposite faces of the rhomb ABCD and EFGH is called a principal section. If the plane of incidence of light on the rhomb coincides with a principal section, the light entering the crystal will be split into two components polarized at right angles to each other which travel in slightly different directions and leave the crystal as two beams slightly displaced but parallel to each other.

The large birefringence of calcite and its excellent transmission through the visible spectral region and into the ultraviolet and infrared regions has made it possible to make excellent high extinction ratio

polarizers with different designs. Some of these are shown in Figure 2. There are two main types: Glan types with rectangular shapes and Nicol types with rhombohedral shapes. The polarizers are made of two pieces of calcite cemented together. Glan types have their optic axes in the plane of the entrance face. In the Nicol types, the principal section is perpendicular to the entrance face, but the optic axis is neither parallel nor perpendicular to the face. The two halves of conventional polarizing prisms are cemented together with cement that has a refractive index intermediate between the ordinary and extraordinary refractive indices. This enables one ray (generally the extraordinary, or *e* ray) to be transmitted and the other to be reflected at the cut, so that only one ray exits from the prism in the direction of the incoming ray. The Glan types are also used without cement with only an air space between the two halves. They can be used at shorter wavelengths in the ultraviolet where the cement absorbs. The extinction ratio can be very high for air-spaced prisms. For example, for a Glan–Foucault prism (an air-spaced Glan–Thompson prism, Figure 2), the extinction ratio can be better than 1×10^{-5} and the prism is usable from about 0.214 μm in the ultraviolet to 2.3 μm in the infrared. However, air-spaced polarizers have very small field angles, so they are mainly used with laser sources where the beam is parallel.

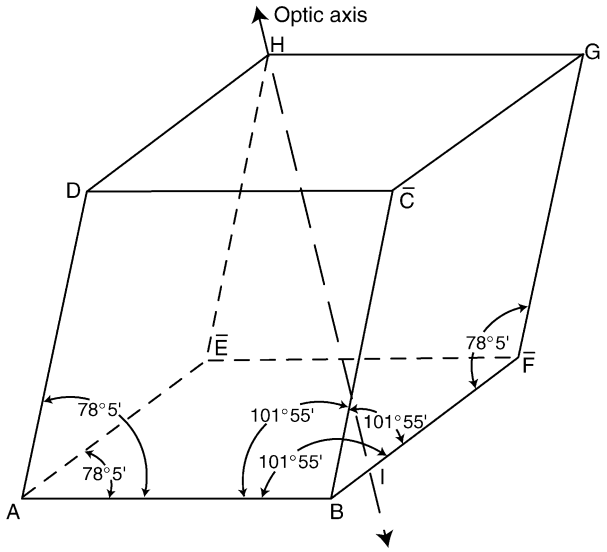


Figure 1 Schematic representation of a rhombohedral calcite crystal showing the angles between faces. The optic axis passes through corner H and point I on side BF. (Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

The extreme paths of light through a cemented Glan–Thompson prism are shown in Figure 3. This prism has a length that is three times the width of the entrance aperture (i.e., an *L*/*A* ratio of 3). The optic axis is perpendicular to the plane of incidence which is in the plane of the paper. In the first half of the polarizer, the paths of the ordinary and extraordinary rays, both of which follow the conventional law of refraction (Snell's Law, eqn [8] above) nearly coincide. Ray A is incident on the entrance face of the polarizer at an angle such that the angle of incidence on the cut is the smallest angle for which the *o* ray is totally internally reflected. Ray B is incident at an angle such that the angle of refraction in the first half of the prism is essentially equal to the cut angle, S, so that the *e* ray just passes through the cut. The field angle of the polarizer is twice the smaller of angles θ_A and θ_B. Thus, all rays having angles of incidence between rays A and B will be polarized

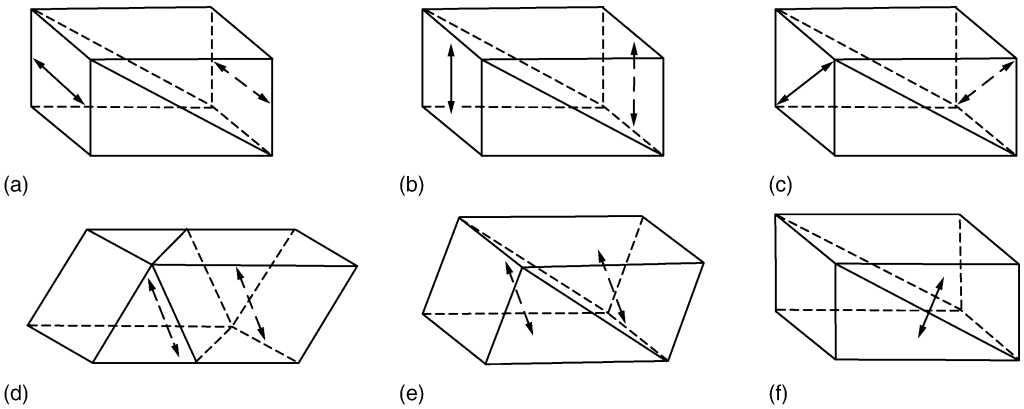


Figure 2 Types of polarizing prisms. Glan types: (a) Glan-Thompson, (b) Lippich, and (c) Frank-Ritter; Nicol types: (d) conventional Nicol, (e) Nicol, Halle form, and (f) Hartnack-Prazmowsky. The optic axes are indicated by the double-pointed arrows. (Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

when exiting the prism. The paths of similar rays can be traced through the other prism designs.

In addition to conventional polarizing prisms, there are also polarizing beamsplitter prisms and Feussner prisms. In polarizing beamsplitter prisms, the two beams that are polarized at right angles to each other both emerge from the prism but are separated spatially. Figure 4 shows a diagram of several types of polarizing beamsplitter prisms and Figure 5 shows the paths of rays through side views of these prisms.

A Feussner prism is made of isotropic material but the film separating the two halves is birefringent. These prisms have the advantage that much less birefringent material is required but they have a more limited wavelength range when calcite or sodium nitrite (another birefringent material) is used because the transmitted ordinary ray has a shorter transmission range than the extraordinary ray which is transmitted by the conventional and air-spaced polarizing prisms.

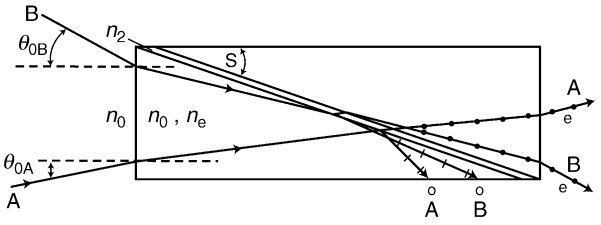


Figure 3 Extreme rays (A and B) that can pass through a cemented Glan-Thompson prism. Both rays change to ordinary and extraordinary rays in the calcite. Both ordinary rays are reflected at the boundary between the two halves of the prism and the extraordinary rays exit in the directions indicated. Rays entering the prism between rays A and B would be transmitted by the prism. The field angle of the prism is twice the smaller of the two angles of incidence (ray A).

Dichroic Absorbers

A dichroic material is one that absorbs light polarized in one direction more strongly than light polarized at right angles to that direction. Dichroic materials are different from birefringent materials because the latter usually have negligible absorption coefficients for both the ordinary and extraordinary rays. Stretched polyvinyl alcohol sheets treated with absorbing dyes or polymeric iodine are the most common type of dichroic absorbers and are primarily sold under the tradename Polaroid. These polarizers do not have as good an extinction ratio as the calcite prism polarizers (see above), but they are inexpensive, come in large sizes, are easily rotated, and produce negligible beam deviation. Also, they are thin, lightweight, and can be made in any desired shape. One of their main advantages is that they are insensitive to the degree of collimation of the beam and can be used in strongly convergent or divergent light.

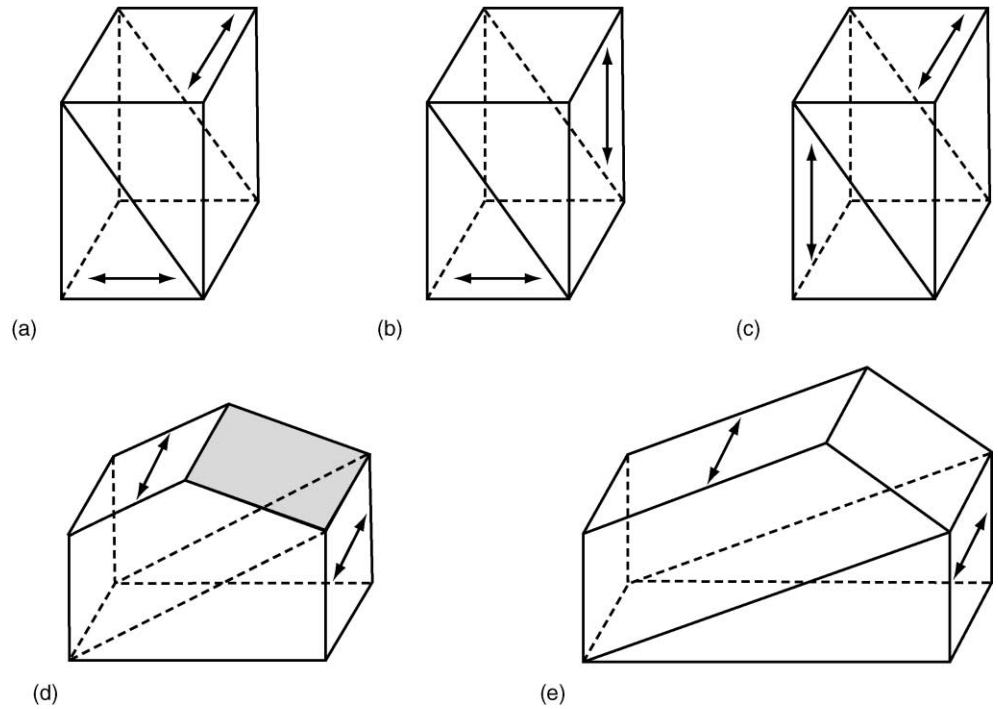


Figure 4 Three-dimensional views of various types of polarizing beamsplitter prisms: (a) Rochon; (b) Sénarmont; (c) Wollaston; (d) Foster (shaded face is silvered); and (e) beamsplitting Glan-Thompson. (Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

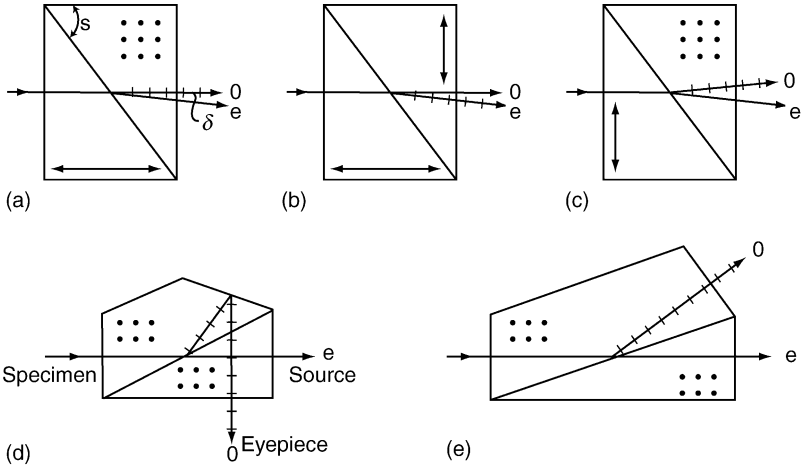


Figure 5 Side views of the polarizing beamsplitter prisms in **Figure 4**. The directions of the optic axes are indicated by the dots and the heavy double-pointed arrows. When the Foster prism is used as a microscope illuminator, the source, specimen, and eyepiece are in the positions indicated. (Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

Depending on the density and type of absorbing dye used to make the polarizer, the transmission can be maximized for the visible or near infrared spectrum. The extinction ratio of the Polaroid HN-22 sheet compares favorably with that of the Glan–Thompson prisms throughout the visible spectral region, but the transmission of the Glan–Thompson prism is superior. As the dichroic polarizer transmission increases, the extinction ratio becomes worse. Transmission and extinction ratio curves for various types of dichroic polarizers are shown in Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

A wire grid polarizer is another kind of dichroic absorber. It is made of a series of equally spaced conducting bars or wires that are either free standing or deposited onto a transparent substrate (backing plate). Energy that is polarized parallel to the length of the bars is absorbed out of the incoming wave by inducing oscillations in the electrons in the metal. Thus, only light polarized perpendicular to the bars will be transmitted. In order to have an appreciable degree of polarization, the wavelength must be at least twice the spacing between grids in the polarizer. Because of the difficulty of making wire grids with small enough spacings, these polarizers are limited to the mid- and long-wavelength infrared spectral region, beyond about 5 μm. The polarizer will have the best extinction ratio if the substrate has a low refractive index; if a high refractive index substrate such as silicon or germanium is used, it must be covered with an antireflection coating before the grid is deposited.

Most of the wire grid polarizers have been used with microwaves, so the theory is in the form for transmission lines. Similar transmission line theory has been applied to infrared polarizers in the form of bars and wires.

Reflection and Transmission

Light can be polarized by reflecting it from the flat surface of a material inclined at non-normal incidence to the light beam or by transmitting it through a transparent plate at non-normal incidence. These polarizers work because light has different reflectances when the electric vector is linearly polarized parallel and perpendicular to the plane of incidence. The plane of incidence contains both the incoming light beam (incident beam) and the reflected light beam. The angles for both the incident and reflected light beams are measured relative to an axis perpendicular to the surface (the surface normal). The reflection coefficients are given by the Fresnel equations and, for nonabsorbing materials, are:

$$R_s = r_s^2 = \frac{\vec{E}_s^2(\text{reflected})}{\vec{E}_s^2(\text{incident})} = \frac{\sin^2(\theta_0 - \theta_1)}{\sin^2(\theta_0 + \theta_1)} \qquad [9]$$

$$R_p = r_p^2 = \frac{\vec{E}_p^2(\text{reflected})}{\vec{E}_p^2(\text{incident})} = \frac{\tan^2(\theta_0 - \theta_1)}{\tan^2(\theta_0 + \theta_1)} \qquad [10]$$

where R_s and R_p are called the reflectances (previously called the intensity reflection coefficients), and r_s and r_p are the amplitude reflection coefficients. $\vec{E}_s$ ($\vec{E}_p$) is the incident or reflected electromagnetic

wave vibrating perpendicular (parallel) to the plane of incidence and θ_0 and θ_1 are the angles of incidence and refraction, respectively. The angle of refraction is the angle of the light beam in the material, measured from the surface normal. For a nonabsorbing material, it can be obtained in terms of the refractive index n_1 of the material from Snell's Law:

$$\frac{\sin\theta_0}{\sin\theta_1} = \frac{n_1}{n_0} \qquad [11]$$

where n_0 and n_1 are the refractive indexes of the incident medium (usually air with $n_0 = 1$) and the material. Sometimes the refractive index of the material is expressed as a ratio measured relative to the refractive index of the incident medium: $n = n_1/n_0$. The reflectances of absorbing materials are similar to eqns [9]–[11] but involve complex refractive indexes and angles of refraction.

Figure 6 shows curves for R_s and R_p as a function of angle of incidence for four nonabsorbing materials that have different refractive indexes. In all cases the reflectance is higher for the *s* component than for the *p* component except at 0° and 90° angles of incidence where they are the same. At the so-called Brewster angle, θ_B, $R_p = 0$, $\tan\theta_B = n_1/n_0$, and incident unpolarized light is now completely linearly polarized perpendicular to the plane of incidence (s-polarized light). Note that high refractive index materials produce more intense polarized beams (i.e., have higher reflectances) than low refractive index materials.

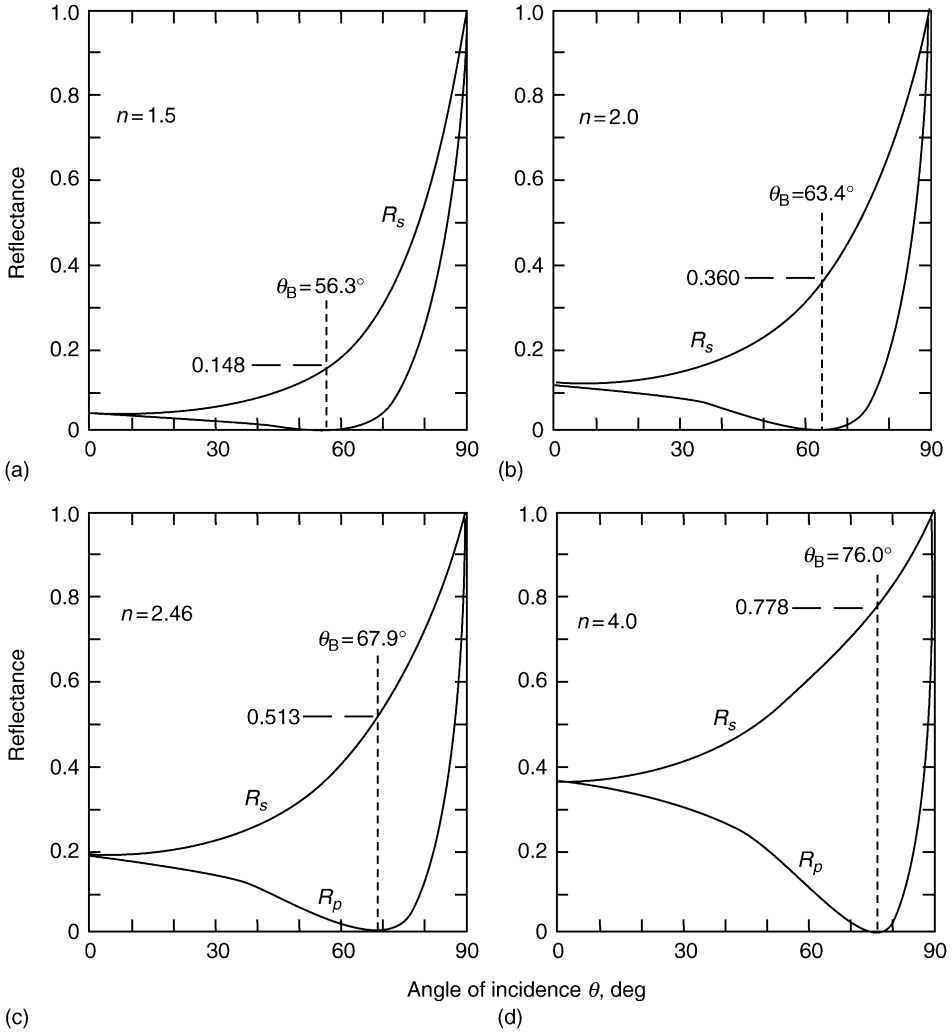


Figure 6 Reflectance of light polarized parallel R_p and perpendicular R_s to the plane of incidence from materials of different refractive index *n* as a function of angle of incidence: (a) $n = 1.5$ (alkali halides in the ultraviolet, glass (approximately) in the visible, and sheet plastics in the infrared), (b) $n = 2.0$ (AgCl in the infrared), (c) $n = 2.46$ (Se in the infrared), and (d) $n = 4.0$ (Ge in the infrared). The Brewster angle θ_B (at which R_p goes to 0) and the magnitude of R_s at θ_B are also indicated. (Reproduced with permission from Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

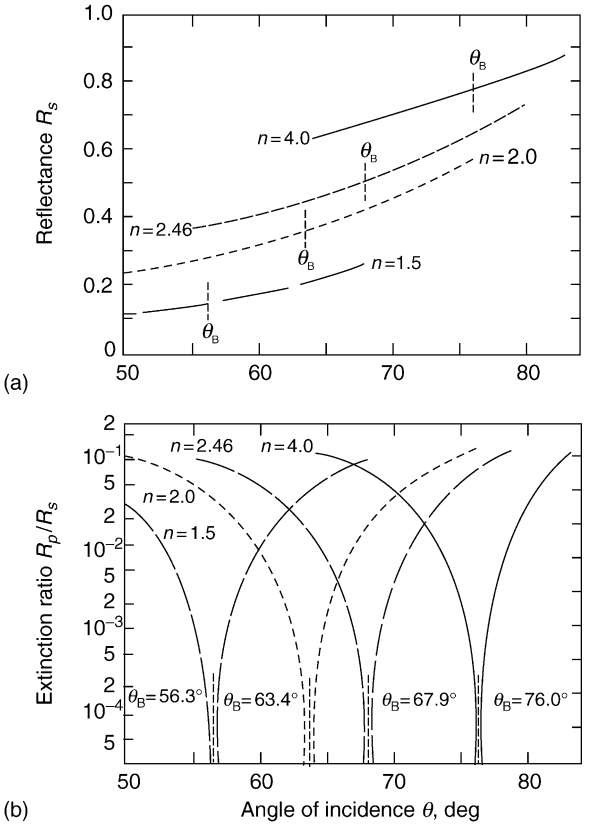


Figure 7 (a) Reflectance R_s and (b) extinction ratio R_p/R_s for materials of different refractive index at angles near the Brewster angle θ_B. A single surface of the material is assumed. (Reproduced with permission from Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

Figure 7 shows the reflectances and extinction ratios for materials having different refractive indexes. Each curve is for a single reflection. Since plates have two surfaces, there will be two reflections from an actual reflection polarizer so the reflectance and extinction ratios will increase. If the plates are thick, the reflection from the back surface of the plate will be displaced from the reflection from the front surface of the plate. A reflection polarizer is normally used close to the Brewster angle because a high degree of polarization (i.e., a very small extinction ratio) is desired. Because the extinction ratio changes rapidly for angles of incidence close to the Brewster angle, the incident beam must be well collimated to obtain a high degree of polarization. A main disadvantage of reflection polarizers is that the reflected beam is no longer parallel to the incident beam and two additional reflections are required to align the beam.

Light transmitted through a plate will only be partially linearly polarized at any angle of incidence including the Brewster angle because both s- and p-components are partially transmitted. Transmission polarizers are thus made of several plates to increase the degree of polarization. Figure 8 shows the transmittance and extinction ratio for a stack of four plane parallel plates assuming multiple incoherent reflections within each plate and no reflections between plates. At the Brewster angle the p-transmittance is theoretically 1 (assuming that there is no absorption within the material) but the extinction ratio greatly depends on the refractive index of the material. A stack of low refractive index plates ($n = 1.5$) has a poor extinction ratio, while a pile of high refractive index plates ($n = 4.0$) has a much better extinction ratio. The plates are often inclined at small angles to each other so the light multiply reflected between plates (which decreases the polarization) is removed from the transmitted beam. The sides of each plate are plane parallel to increase the polarization, but the plates are too thick for the amplitudes of the multiple internally reflected beams to add or subtract. If the amplitudes of the beams could be added, there would be interference effects and the transmittance would vary with the thickness of each plate and with the wavelength. These are so-called interference polarizers and are mentioned below.

Transmission polarizers do not have the angle deviation problem of the reflection polarizers, but the transmitted beam may be slightly displaced parallel to the incident beam if the plates are thick. The main problem with transmission polarizers is that the light is not completely polarized even with several plates in a stack. If the plates are slightly absorbing, the transmission of the polarizer is reduced.

Miscellaneous Types

There are numerous other types of polarizers that mostly use thin films. Interference effects in thin films can increase the polarization efficiency at certain wavelengths depending on the thicknesses of the films and the design of the multilayer coating. Many of the applications involve non-normal incidence designs. For example, a single high refractive index film or a multilayer coating evaporated onto a low refractive index substrate increases the degree of polarization of reflection and transmission polarizers. Polarizing beamsplitters also use multilayer dielectric coatings. Free-standing films of different thicknesses were formerly used for infrared polarizers before wire grid polarizers became available.

Dichroic absorbers have been made from two-phase lamelar eutectics of thin needles of a conducting material embedded in a transparent matrix. Thin sheets of pyrolytic graphite material also act

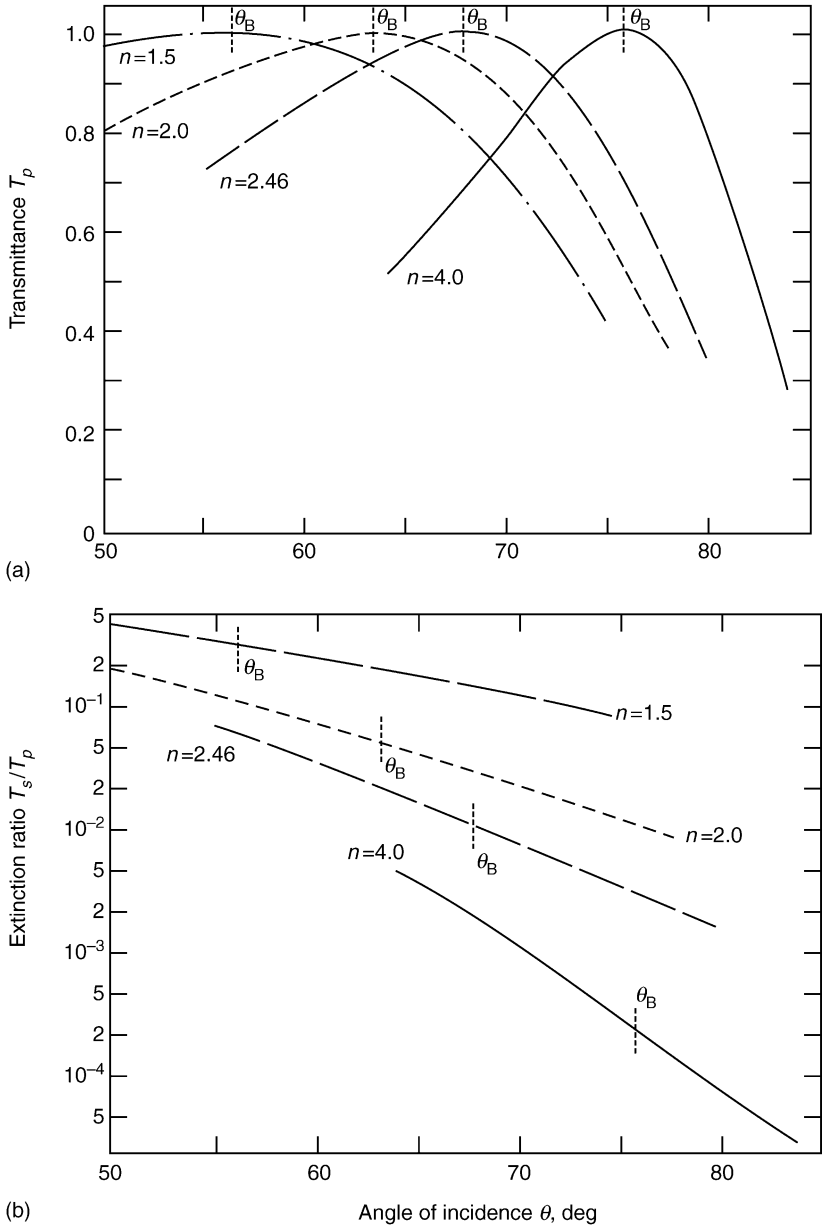


Figure 8 (a) Transmittance and (b) extinction ratio of four plane-parallel plates of refractive index n as a function of angle of incidence, for angles near the Brewster angle. Assumptions are multiple reflections but no interference within each plate and no reflections between plates. (Reproduced with permission from Bennett JM (1995) Polarization. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

as dichroic absorbers. There are also numerous miniature polarizer designs for fiber optics and nanotechnology applications.

Retarders or Retardation Plates

Retarders, or retardation plates, are devices that can change linearly polarized light into circularly or elliptically polarized light. They can also rotate the plane of polarization of linearly polarized light. A retardation plate is made from a uniaxial crystal that has an optic axis (as discussed above); the light travels in a direction perpendicular to the optic axis, as described below and shown in Figure 9. The ordinary and extraordinary rays travel at different velocities through the crystal depending on their refractive indexes: $n = c/v$, where c is the velocity of light in a vacuum and v is the velocity of light in the material. The larger the refractive index, the slower is the velocity of light in the material. The ordinary ray

with a refractive index n_o and the extraordinary ray with a refractive index n_e have mutually perpendicular vibration directions. If the ray has a vibration direction at another azimuth in the crystal, the refractive index is intermediate between n_o and n_e. For a positive uniaxial crystal $n_e > n_o$ so the extraordinary ray travels slower than the ordinary ray through the crystal. Thus, the designation 'fast axis' is often used for the ordinary ray and 'slow axis' is used for the extraordinary ray, as shown in Figure 9. In a negative uniaxial crystal, $n_e < n_o$, i.e., the velocities along the two axes are reversed.

Because there is a velocity difference between the ordinary and extraordinary rays traveling through a uniaxial crystal, they get out of phase. Depending on their velocities and the optical thickness of the material (the refractive index times the physical thickness), the addition of their amplitudes results in a wave whose vibration direction is either: (a) perpendicular to the original vibration direction; (b) rotates in a circle (right or left circularly polarized light); or (c) rotates in an ellipse (right or left elliptically polarized light). The path difference N between two rays in the crystal, measured in terms of the wavelength of the light, is $N\lambda = \pm d(n_e - n_o)$, where d is the physical thickness of the material. The corresponding phase difference δ between the two rays is $\delta = 2\pi N = \pm(2\pi d/\lambda)(n_e - n_o)$. The quantity N is important because if beams of light vibrating along the fast and slow axes of a crystal get out of step by one quarter of the wavelength of the light, the device is called a quarter wave (or $\lambda/4$) retardation plate, or simply a quarter-wave plate. There are also half-wave ($\lambda/2$) plates that rotate the plane of polarization, and full-wave (λ) plates that rotate the plane of polarization through 180°. If light passes through a full-wave plate, theoretically it is indistinguishable from the original beam. However, materials are normally temperature sensitive so that as the temperature changes, the retardation is no longer exactly one wave, but may be slightly less than or greater than one wavelength. Other thicknesses of retardation plates produce elliptically polarized light.

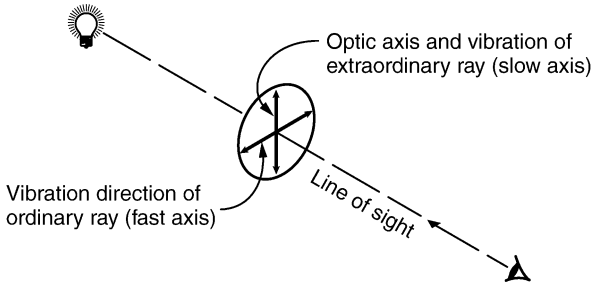


Figure 9 Light incident normally on the front surface of a retardation plate showing the vibration directions of the ordinary and extraordinary rays. In a positive uniaxial crystal, the fast and slow axes are as indicated in parentheses; in a negative uniaxial crystal, the two axes are interchanged. (Adapted with permission from Bennett JM (1995) Polarization. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

Figure 10 shows what happens to a beam of linearly polarized light when the axis of vibration is at 45° to the fast and slow axes of a positive

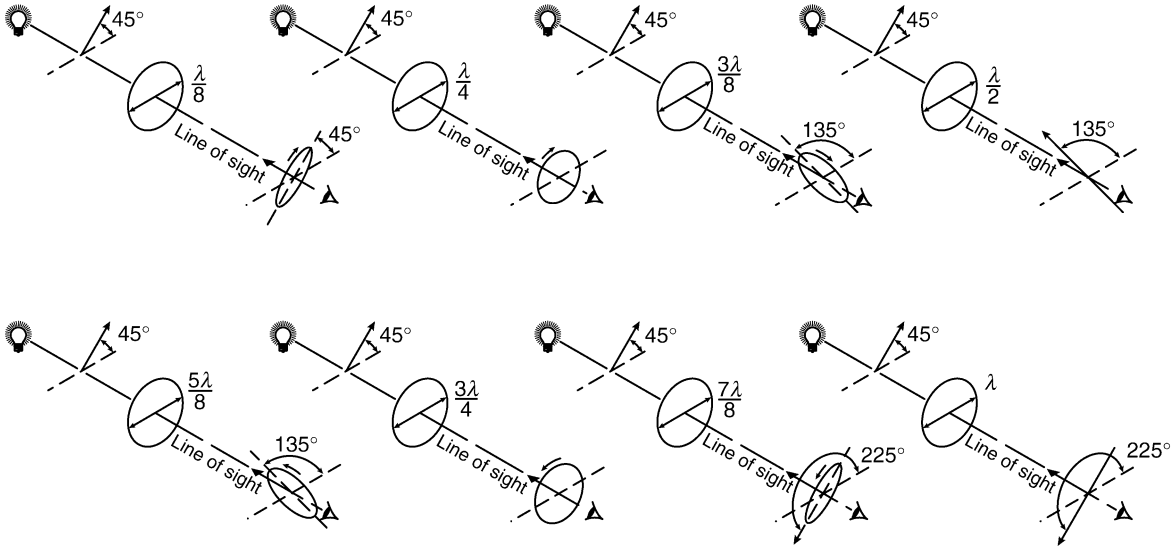


Figure 10 State of polarization of a light wave after passing through a crystal plate whose retardation is indicated in fractions of a wavelength (phase retardation $2\pi/\lambda$ times these values) and whose fast axis is indicated by the double arrow. In all cases the incident light is linearly polarized at an azimuth of 45° to the direction of the fast axis. (Adapted with permission from Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

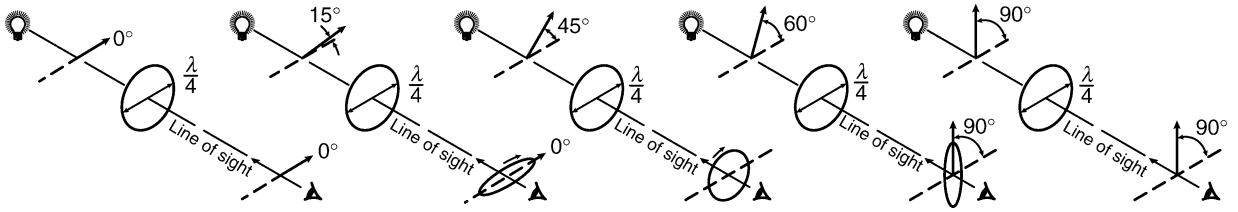


Figure 11 State of polarization of a light wave after passing through a $\lambda/4$ plate (whose fast axis is indicated by the double arrow) for different azimuths of the incident linearly polarized beam. (Adapted with permission from Bennett JM (1995) Polarization. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

uniaxial crystal. The fast axis is in the horizontal direction, the same as in Figure 9.

The main purpose of a quarter-wave plate is to change linearly polarized light into circularly polarized light. However, as Figure 11 shows, the state of polarization of the light will be quite different depending on the orientation of the plane of polarization of the incident beam relative to the fast axis of the quarter-wave plate.

A half-wave plate is used to rotate the plane of polarization of a linearly polarized beam. The plane of polarization is always rotated through an angle that is twice the angle the initial plane of polarization makes with the fast axis of the uniaxial crystal. A linearly polarized beam always remains linearly polarized.

Retardation plates are often made of mica, stretched polyvinyl alcohol, or crystal quartz, although they can also be made of other stretched plastics, sapphire, magnesium fluoride, and other materials.

Variable Retardation Plates and Compensators

Variable retardation plates are devices whose retardation can be varied in a variety of ways. They can be used to modulate or vary the phase of a beam of linearly polarized light or to analyze a beam of unknown polarization (often elliptically polarized light) such as might be produced by transmission through a birefringent material or by reflection from a metal or film-covered surface. The term compensator is often applied to a variable retardation plate since it can be used to compensate for the phase retardation produced by a material. Common types of compensators include the Babinet and Soleil compensators, in which the total thickness of a birefringent material in the light path is changed, the Sénarmont compensator which consists of a fixed quarter-wave plate and rotatable analyzer to compensate for varying amounts of ellipticity in a light beam, and tilting-plate compensators, which change the thickness of birefringent material in the light beam by changing the angle of incidence. Electro-optic and piezo-optic modulators can also be used as high-frequency variable retardation plates since their birefringence can be changed by varying the electric field or pressure (see next section).

The Babinet compensator, shown schematically in Figure 12 consists of two crystal quartz wedges, each with its axis in the plane of the face but with the axes at right angles to each other. One wedge is stationary, and the other can be moved in the direction indicated by the arrow, so that the total amount of quartz through which the light passes can be varied. The total retardation is proportional to the difference in thickness between the two wedges. This type of compensator was used extensively when the light source was a vertical slit and visual measurements were made of the state of polarization of a beam. However the bands representing different phase retardations were too narrow to be used effectively with a photoelectric detector, so the Babinet compensator was replaced by a Soleil compensator (Figure 13). This device, sometimes called a Babinet–Soleil compensator, is similar to the Babinet compensator except that the field of view has a uniform tint if the compensator is constructed correctly. This is because the ratio of the thicknesses of the two crystal quartz blocks (a movable wedge and a fixed wedge attached to a plate with the two axes perpendicular to each other) is the same over the entire field. The Soleil compensator will produce light of varying ellipticity depending on the position of the movable wedge. It is used in the same way as a Babinet compensator with the uniformly dark field of the Soleil compensator corresponding to the black zero-retardation band in the Babinet compensator.

It is sometimes necessary to accurately measure the azimuth of a beam of linearly polarized light using a photoelectric detector with a rotatable analyzer (i.e., a polarizer) directly in front it. The obvious method is to rotate the analyzer until the detector signal is a minimum and then read the analyzer angle, which equals the extinction angle (perpendicular to the azimuth of the linearly polarized beam). However, the

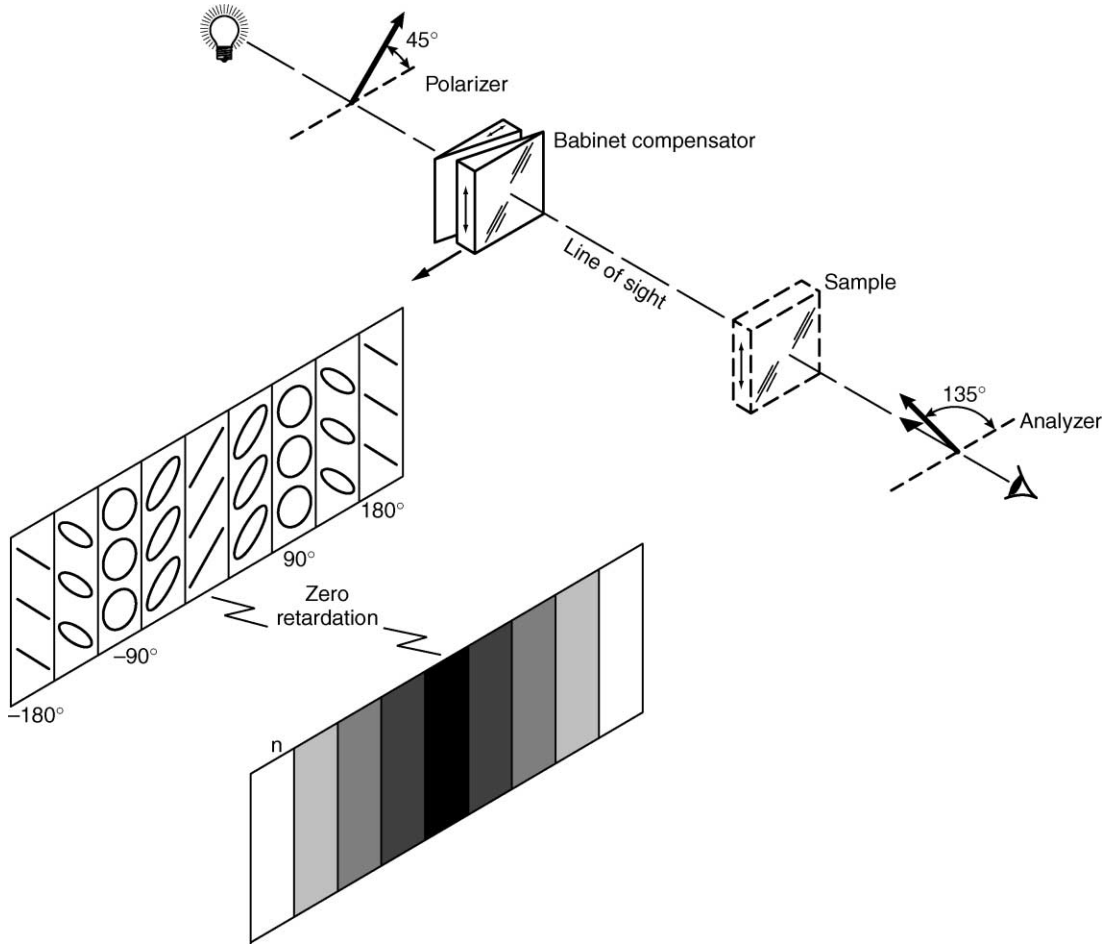


Figure 12 Arrangement of a Babinet compensator, polarizer, and analyzer for measuring the retardation of a sample. The appearance of the field after the light has passed through the compensator is shown to the left of the sample position. Retardations are indicated for alternate regions. After the beam passes through the analyzer, the field is crossed by a series of dark bands, one of which is shown to the left of the analyzer. (Adapted with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

angle can be determined more precisely if the analyzer is offset by a small angle from the extinction angle and the transmittance noted. Then the analyzer is offset by a small angle on the other side of the extinction angle at the angle where the transmittance is the same. One half the difference between these two azimuthal angles gives a more accurate value of the extinction angle than can be obtained by measuring it directly.

Before the days of photoelectric detectors, half-shade devices were extensively used to determine the azimuth of a linearly polarized beam. The device consisted of two polarizers with their axes inclined at a small angle to each other. As the device was rotated, one part of the field became darker and the other part became lighter. At the match position, both parts of the field appeared equally bright. There were a variety of these devices as well as ellipticity half-shade devices that could detect very small amounts of ellipticity in a nominally linearly polarized beam and hence could verify when a compensator had completely converted elliptically polarized light into linearly polarized light.

Electro-Optic, Magneto-Optic, and Piezo-Optic Devices

The state of polarization of light can be rapidly altered by passing it through a material that has electro-optic, magneto-optic or piezo-optic properties. The voltage, magnetic field, or pressure are varied to make the material birefringent (see above). Some materials are isotropic without the applied field, i.e., the refractive index is the same in all directions. Other materials have an optic axis

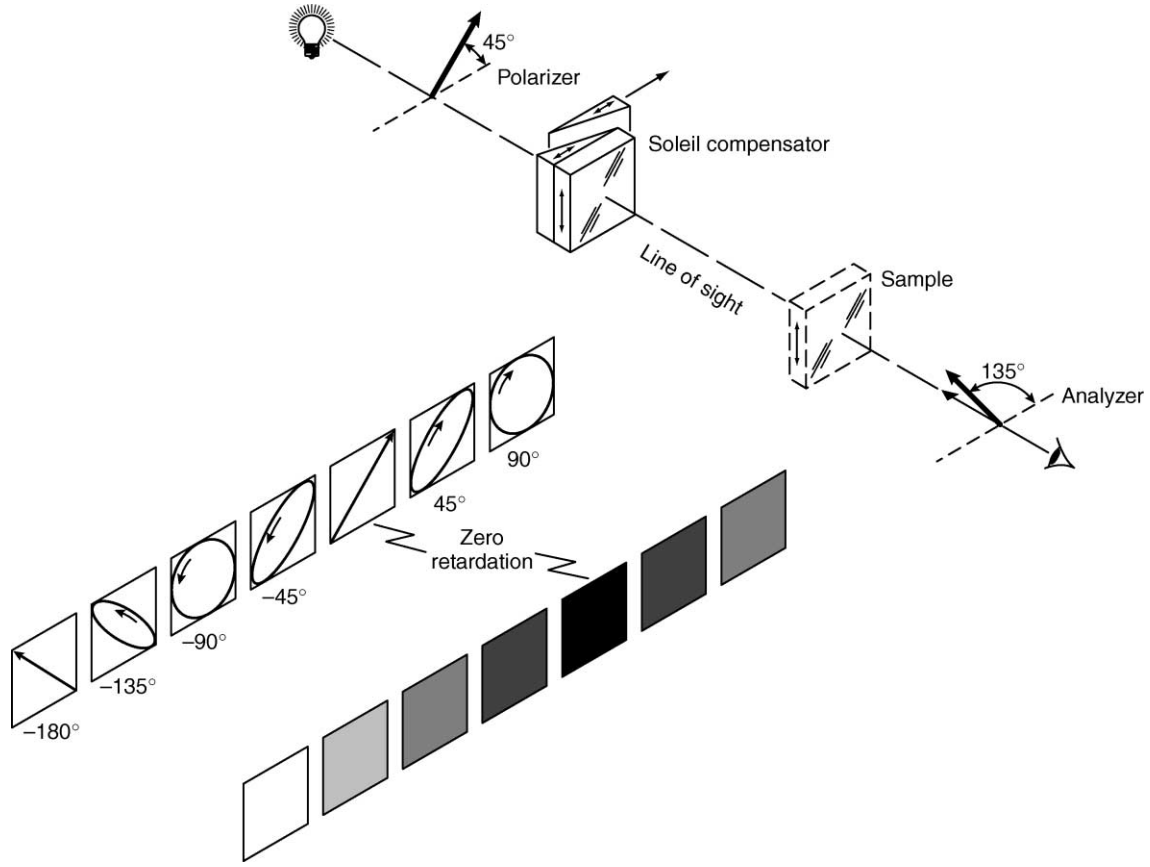


Figure 13 Arrangement of a Soleil compensator, polarizer, and analyzer for measuring the retardation of a sample. The appearance of the field after the light has passed through the compensator is shown to the left of the sample position. After the beam passes through the analyzer, the field appears as one of the shades of gray shown to the left of the analyzer. (Adapted with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

(uniaxial materials) or two optic axes (biaxial materials). In these cases, the applied field creates further asymmetries in the material.

The most common electro-optic, magneto-optic and piezo-optic effects are the Pockel's effect (depending on the electric field), the Kerr effect (depending on the square of the electric field), the Faraday effect (depending on the magnetic field), the Cotton–Mouton and Voigt effects (depending on the square of the magnetic field), and stress birefringence or the photoelastic effect (depending on pressure changes). These effects are shown in Table 1 along with order of magnitude strengths needed to produce changes in the refractive index, and materials that most strongly exhibit the effects. The mathematical descriptions of the effects involve tensors and are too involved to present here. However, simple physical descriptions will be given.

If a varying electric field acts on an electro-optic material, electrons, ions, or permanent dipoles in the material are made to reorient, inducing an electric polarization. The induced polarization creates birefringence that modifies the optical polarization of a light beam passing through the material. The electric field strength determines how the polarization is changed (see Figure 10).

Linearly polarized light passing through a magneto-optic material will be rotated in a direction parallel to the direction of the applied magnetic field. This phenomenon, called the Faraday effect, or Faraday rotation, is similar to what happens in optically active materials (see the next paragraph) except that the Faraday effect depends on the direction of the magnetic field, but not on the direction the light is traveling.

There are naturally optically active materials such as camphor, nicotine, sugar solutions, and crystal quartz that can rotate the plane of linearly polarized light passing through the material. For example, the Si–O bonds in crystal quartz form a helical path around the optic axis. This crystal structure interacts with an incoming linearly polarized beam traveling parallel to the optic axis and rotates it in a clockwise direction. A 1 mm-thick piece of quartz will rotate a

Table 1 Electro-optic, magneto-optic, and piezo-optic effects

Effect	*Device*	*Effect proportional to*	*Strength (Δn)*	*Materials*
Pockels (electro-optic)	Pockels cell	E	$10^{-12}E$	$BaTiO_2$, $LiNBO_3$
Kerr	Kerr cell	E^2	$10^{-19}E^2$	Nitrobenzene, Benzonitrile
Faraday (magneto-optic)	Faraday rotation	H	$10^{-14}H$	Zns, GaAs, CS_2
Cotton–Mounton	[a]	H^2	$10^{-25}H^2$	Chloroform, acetone
Photoelastic effect; stress birefringence	[b]	Pressure	–	Fused silica, polystyrene, Ge, KDP, ruby

[a]Too small to be of technological importance.
[b]Depends on the mounting or the processing of the material.
Adapted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley and Sons.

linearly polarized beam by 21.7°. Other materials have asymmetric structures that have mirror images. A mirror image structure cannot be obtained by simply rotating the group of atoms in space. These materials are also optically active. The amount of rotation produced by an optically active material is proportional to the thickness (optical density) of the material that the light passes through. Dextrose (a form of glucose) and levulose (also known as fructose) rotate the plane of linearly polarized light in clockwise and counterclockwise directions, respectively. Many other organic molecules have D- (right handed, dextro-rotary) and L- (left handed, levo-rotary) mirror image forms.

An isotropic material can become anisotropic when stress or an induced strain is applied because of an elasto-optic interaction known as stress birefringence, or the photoelastic effect. This effect is usually bad because it reduces the performance of optical components and devices by introducing phase distortions caused by improper mounting or unequal thermal expansion between the mounts and components. Another source of distortion is strain that has been frozen into optical components during manufacture. French curves are excellent examples that show colored strain birefringence patterns when viewed between crossed polarizers. In one application, strain birefringence has been used constructively to produce a variable phase retarder made from crystal quartz and fused silica. A block of electro-optic crystal quartz, is cemented to a block of isotropic fused silica. When a variable electric field is applied to the crystal quartz, its length changes at the resonant frequency of the block. This produces strain in the fused silica block which in turn produces a variable retardation of light passing through the strained fused silica. Depending on the magnitude of the varying electric field, the device can act like a variable quarter-wave plate, a variable half-wave plate, or have other applications.

Matrix Methods for Computing Polarization

An optical system containing various polarizing and retarding devices can be modeled using matrices. In general, there is an incident beam (in matrix form), that has some state of polarization. It interacts with a device called an instrument (also in matrix form) that alters the state of polarization, so that the exiting beam (a third matrix, the product of the first and second matrices) has another state of polarization. There can be more than one instrument matrix acting on the same incident matrix to produce the final matrix. The two most common matrix methods are the Mueller calculus and the Jones calculus.

There is also a visual representation of this process, a Poincaré sphere, on which vectors represent different states of polarization. The polarization instrument moves the vector representing the incident polarization around the sphere in a prescribed manner. The vectors are called Stokes vectors and are used in the Mueller calculus. The various states of polarization are represented on the Poincaré sphere as follows: The equator represents various forms of linear polarization, the poles represent right- and left-circular polarization, and other points on the sphere represent elliptically polarized light. Every point on the sphere corresponds to a different polarization form. The radius of the sphere represents the incident irradiance of the light beam (which is usually assumed to be unity). The effects of various polarization instruments are determined by displacements on the sphere. Partially polarized light or absorption may be dealt with approximately by ignoring the irradiance factor, since the state of polarization is generally the quantity desired. The Poincaré sphere is most useful for visualizing problems involving nonabsorbing materials, various polarization instruments including polarizers, retarders, compensators, half-shade devices, and depolarizers, and has also been applied

to ellipsometric problems and stress-optical measurements.

The Poincaré sphere is used with Stokes vectors, which are often designated S_0, S_1, S_2, and S_3. S_0 is the incident irradiance of the light beam, corresponding to the radius of the Poincaré sphere. S_1 is the difference in irradiances between the horizontal and vertical polarization components of the beam; for example, when S_1 is positive, the preference is for horizontal polarization. S_2 indicates preference for $+45°$ or $-45°$ polarization depending on whether it is positive or negative, and S_3 gives the preference for right or left circular polarization. Thus, the Stokes vectors S_1, S_2, and S_3 are simply the three Cartesian coordinates of a point on the Poincaré sphere. S_1 and S_2 are perpendicular to each other in the equatorial plane, and S_3 points toward the north pole of the sphere. Any state of polarization of a light beam can be specified by these three vectors. The irradiance vector S_0 is related to the other three by the relation $S_0^2 = S_1^2 + S_2^2 + S_3^2$ when the beam is completely polarized. If the beam is partially polarized, $S_0^2 > S_1^2 + S_2^2 + S_3^2$.

In the Mueller calculus, the incident beam is represented by the four-component Stokes vector, written as a column vector. This vector has all real elements and gives information about irradiance properties of the beam. Thus it is not able to handle problems involving phase changes or combinations of two beams that are coherent. The instrument matrix is a 4×4 matrix with all real elements.

In the Jones calculus, the Jones vector representing the incident beam is a two-component column vector that generally has complex elements. It contains information about the *amplitude* properties of the beam and hence is well suited for handling problems involving the phases of light waves. However, it cannot handle problems involving depolarization, as the Mueller calculus can. The Jones instrument matrix is a 2×2 matrix whose elements are generally complex. The Jones calculus is well suited to problems involving a large number of similar devices arranged in series in a regular manner and makes it possible to obtain a result expressed explicitly in terms of the number of such devices. The Jones instrument matrix of a train of transparent or absorbing nondepolarizing polarizers and retarders contains no redundant information. The matrix contains four elements, each of which has two parts, so that there are a total of eight constants, none of which is a function of any other. The Mueller instrument matrix of such a train contains much redundancy; there are 16 constants but only 7 of them are independent. More information about the Poincaré sphere and the Mueller and Jones calculus is available in other references.

See also

Electromagnetically Induced Transparency. **Magneto-optics:** Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping; Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering.

Further Reading

Auton JP (1967) Infrared transmission polarizers by photolithography. *Applied Optics* 6: 1023–1027.

Azzam, Rasheed MA and Bashara NM (1977) *Ellipsometry and Polarized Light*, chaps. 1 and 2. Amsterdam: North-Holland Publishing Co.

Azzam, Rasheed MA (1995) Ellipsometry. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 27. New York: McGraw-Hill.

Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill.

Bennett M (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill.

Bohren CF and Huffman DR (1983) *Absorption and Scattering of Light by Small Particles*. New York: John Wiley.

Born M and Wolf E (1999) *Principles of Optics*, 7th edn, chap. 1. Cambridge: Cambridge University Press.

Chipman RA (1995) Polarimetry. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 22. New York: McGraw-Hill.

Clark JR (1956) New calculus for the treatment of optical systems. VIII. Electromagnetic theory. *Journal of the Optical Society of America* 46: 126–131.

Guenther RD (1990) *Modern Optics*. chaps. 2, 13, and 14. New York: John Wiley and Sons.

Hass M and O'Hara M (1965) Sheet infrared transmission polarizers. *Applied Optics* 4: 1027–1031.

Hecht E and Zajac A (1997) *Optics*, 3rd edn, chap. 8. Reading, MA: Addison-Wesley Publishing Co.

Jackson JD (1999) *Classical Electrodynamics*, 3rd edn, chap. 7. New York: John Wiley and Sons, Inc.

Jasperson SN and Schnatterly SE (1969) An improved method for high reflectivity ellipsometry based on a new polarization modulation technique. *Reviews of Scientific Instruments* 40: 761–767.

Kerker M (1969) *The Scattering of Light and Other Electromagnetic Radiation*. New York: Academic Press.

Kliger DS, Lewis JW and Randall CE (1990) *Polarized Light in Optics and Spectroscopy*, chaps. 1–5. San Diego: Academic Press.

Maldonado TA (1995) Electro-optic modulators. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 13. New York: McGraw-Hill, Inc.

Markuvitz N (ed.) (1951) *Waveguide Handbook*, pp. 218–229, 285–289. New York: McGraw-Hill Book Co.

Shurcliff WA (1962) *Polarized Light, Production and Use*. Cambridge, MA: Harvard University Press.

van de Hulst HC (1957) *Light Scattering by Small Particles*. New York: John Wiley.

Matrix Analysis

B D Guenther, Duke University, Durham, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Hooke postulated, in the seventeenth century, that light waves must be transverse but his idea was forgotten. Young and Fresnel put forward the same idea in the nineteenth century and accompanied their postulation with a theoretical description of light based on transverse waves. Forty years later, Maxwell proved that light must be a transverse wave and that **E** and **H**, for a plane wave in an isotropic medium with no free charge and no currents, are mutually perpendicular and lie in a plane normal to the direction of propagation, **k**.

Convention requires that we use the electric vector to label the direction of the electromagnetic wave's polarization. The direction of the displacement vector is called the direction of polarization and the plane containing the direction of polarization and the propagation vector is called the plane of polarization. The selection of the electric field is not completely arbitrary, except in relativistic situations, when $v \approx c$, the interaction of the electromagnetic wave with matter will be dominated by the electric field.

Assume that a plane wave is propagating in the z-direction. In complex notation, the plane wave is given by

$$\tilde{\mathbf{E}} = \mathbf{E}_0\, e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi)} = \mathbf{E}_0\, e^{i(\omega t - kz + \phi)} \qquad [1]$$

The procedure used to decompose an arbitrary polarization into components parallel to two axes of a Cartesian coordinate system, is a technique used extensively in vector algebra to simplify mathematical calculations. According to the mathematical formalism associated with this technique, the polarization is described in terms of a set of basis vectors, $\mathbf{e}_i$. An arbitrary polarization would be expressed as

$$\mathbf{E} = \sum_{i=1}^{2} a_i \mathbf{e}_i \qquad [2]$$

The set of basis vectors, $\mathbf{e}_i$, are orthonormal, i.e.:

$$\mathbf{e}_i \mathbf{e}_j^* = \delta_{ij} = \begin{Bmatrix} 1 & i = j \\ 0 & i \neq j \end{Bmatrix} \qquad [3]$$

where we have assumed that the basis vectors could be complex. We mention this mathematical formalism because an identical formalism is encountered in a description of spin.

In a Cartesian coordinate system, the $\mathbf{e}_i$'s are the unit vectors **i**, **j**, **k**. The summation in eqn [2] extends over only two terms because the electromagnetic wave is transverse, confining **E** to a plane normal to the direction of propagation (according to the coordinate convention we have selected, the **E** field is in the x, y plane).

Polarization Ellipse

Following the formalism of eqn [2], a polarized wave can be written in terms of the x and y components of $\mathbf{E}_0$

$$\tilde{\mathbf{E}} = \mathbf{E}_{0x}\, e^{i(\omega t - kz + \phi_1)}\hat{\mathbf{i}} + \mathbf{E}_{0y}\, e^{i(\omega t - kz + \phi_2)}\hat{\mathbf{j}} \qquad [4]$$

We will use only the real part of **E** for manipulation, to prevent errors. We divide each component of the electric field by its maximum value, so that the problem is reduced to one of the following two sinusoidal varying unit vectors:

$$\begin{aligned} \frac{E_x}{E_{0x}} &= \cos(\omega t - kz + \phi_1) \\ &= \cos(\omega t - kz)\cos\phi_1 - \sin(\omega t - kz)\sin\phi_1 \\ \frac{E_y}{E_{0y}} &= \cos(\omega t - kz)\cos\phi_2 - \sin(\omega t - kz)\sin\phi_2 \end{aligned} \qquad [5]$$

When these unit vectors are added together, the resulting equation will describe the path taken by the tip of the resultant vector. The path will create a Lissajous figure.

To obtain the equation for the Lissajous figure, we eliminate the dependence of the unit vectors on $(\omega t - kz)$. First, multiply the equations by $\sin\phi_2$ and $\sin\phi_1$, respectively and then subtract the resulting equations. Second, multiply the two equations by $\cos\phi_2$ and $\cos\phi_1$, respectively and then subtract the new equations. These two operations yield the following pair of equations:

$$\begin{aligned} &\frac{E_x}{E_{0x}}\sin\phi_2 - \frac{E_y}{E_{0y}}\sin\phi_1 \\ &\quad = \cos(\omega t - kz)[\cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2] \end{aligned} \qquad [6]$$

$$\frac{E_x}{E_{0x}}\cos\phi_2 - \frac{E_y}{E_{0y}}\cos\phi_1 = \sin(\omega t - kz)[\cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2] \quad [7]$$

The term in brackets can be simplified using the trig identity

$$\sin\delta = \sin(\phi_2 - \phi_1) = \cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2 \quad [8]$$

After replacing the term in brackets by sin δ, the two equations are squared and added yielding the equation for the Lissajous figure:

$$\left(\frac{E_x}{E_{0x}}\right)^2 + \left(\frac{E_y}{E_{0y}}\right)^2 - \left(\frac{2E_xE_y}{E_{0x}E_{0y}}\right)\cos\delta = \sin^2\delta \quad [9]$$

The trig identity

$$\cos\delta = \cos(\phi_2 - \phi_1) = \cos\phi_1 \cos\phi_2 + \sin\phi_1 \sin\phi_2$$

was also used to further simplify eqn [9].

Equation [9] has the same form as the equation of a conic

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

Geometry defines the conic as an ellipse because, from eqn [9]

$$B^2 - 4AC = \frac{4}{E_{0x}^2E_{0y}^2}(\cos^2\delta - 1) < 0$$

This ellipse is called the polarization ellipse. The orientation of the ellipse, with respect to the x-axis, is

$$\tan 2\theta = \frac{B}{A - C} = \frac{2E_{0x}E_{0y}\cos\delta}{E_{0x}^2 - E_{0y}^2} \quad [10]$$

The tip of the resultant electric field vector obtained from eqn [4] traces out the polarization ellipse in the plane normal to **k**, as predicted by eqn [9]. The ratio of the length of the minor to the major axis of the ellipse is equal to the ellipticity, φ, i.e., the amount of deviation of the ellipse from a circle

$$\tan\varphi = \pm\left(\frac{E_m}{E_M}\right) = \frac{E_{0x}\sin\phi_1 \sin\theta - E_{0y}\sin\phi_2\cos\theta}{E_{0x}\cos\phi_1\cos\theta + E_{0y}\cos\phi_2\sin\theta} \quad [11]$$

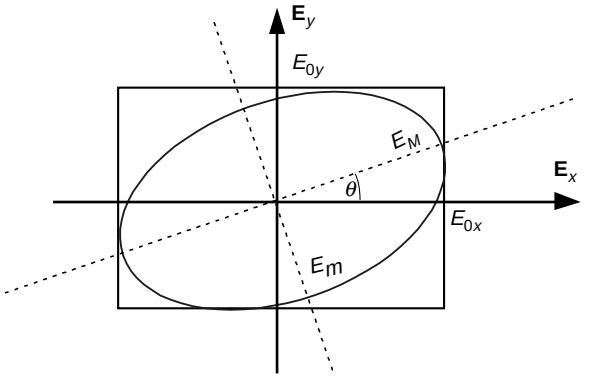


Figure 1 General form of the ellipse described by eqn [9]. Reproduced with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley and Sons.

In particle physics, the light would be said to have a negative helicity if it rotated in a clockwise direction. If we look at the source, the electric vector seems to follow the threads of a left-handed screw, agreeing with the nomenclature that left-handed quantities are negative. However, in optics the light that rotates clockwise, as we view it traveling toward us from the source, is said to be right-circularly polarized. The counterclockwise rotating light is left-circularly polarized. The association of right-circularly polarized light, with 'right-handedness' in optics, came about by looking at the path of the electric vector in space at a fixed time, then $\tan\psi = \tan(\phi - kz)$ (Figure 1).

Stokes Parameters

In the formalism associated with eqn [2], the expansion coefficients, a_i, can be used to form a 2–2 matrix which, in statistical mechanics, is called the density matrix and in optics, the coherency matrix (Table 1). The elements of the matrix are formed by the rule

$$\rho_{ij} = \mathbf{a}_i\mathbf{a}_j^* \quad [12]$$

The matrix is Hermitian, so that $\rho_{ij} = \rho_{ji}^*$. We will not develop the theory of polarization using the coherency matrix, but simply use the coherency matrix to justify the need for four independent measurements to characterize polarization. There is no unique set of measurements required by the theory, but normally measurements made are of the Stoke's parameters which are directly related to the parameters of the polarization ellipse of eqn [9].

The Stokes parameters of a light wave are measurable quantities, defined as:

$s_0 \Rightarrow$ Total flux density
$s_1 \Rightarrow$ Difference between flux density transmitted by a

Table 1 Typical Stokes vectors

Horizontal polarization	$\begin{bmatrix}1\\1\\0\\0\end{bmatrix}$	Vertical polarization	$\begin{bmatrix}1\\-1\\0\\0\end{bmatrix}$
$+45°$ polarization	$\begin{bmatrix}1\\0\\1\\0\end{bmatrix}$	$-45°$ polarization	$\begin{bmatrix}1\\0\\-1\\0\end{bmatrix}$
Right circular polarization	$\begin{bmatrix}1\\0\\0\\1\end{bmatrix}$	Left circular polarization	$\begin{bmatrix}1\\0\\0\\-1\end{bmatrix}$

linear polarizer oriented parallel to the x-axis and one oriented parallel to the y-axis. The x and y-axes are usually selected to be parallel to the horizontal and vertical directions in the laboratory.

$s_2 \Rightarrow$ Difference between flux density transmitted by a linear polarizer oriented at 45° to the x-axis and one oriented at 135°.

$s_3 \Rightarrow$ Difference between flux density transmitted by a right-circular polarizer and a left-circular polarizer.

If the Stokes parameters are to characterize the polarization of a wave, they must be related to the parameters of the polarization ellipse. It is therefore important to establish that the Stokes parameters are variables of the polarization ellipse (eqn [9]).

In its current form, eqn [9] contains no measurable quantities and thus must be modified if it is to be associated with the Stokes parameters. The time average of the Poynting vector is the quantity observed when measurements are made of light waves. We must, therefore, find the time average of eqn [9] if we wish to relate its parameters to observable quantities.

The time average eqn [9] can now be written as

$$\frac{\langle E_x^2\rangle}{E_{0x}^2} + \frac{\langle E_y^2\rangle}{E_{0y}^2} - 2\frac{\langle E_x E_y\rangle}{E_{0x}E_{0y}}\cos\delta = \sin^2\delta \qquad [13]$$

where time average is denoted by $\langle\ \rangle$

$$\langle E_x^2\rangle = \frac{1}{T}\int_{t_0}^{t_0+T} E_{0x}^2[\cos(\omega t - kz)\cos\phi_1 - \sin(\omega t - kz)\sin\phi_1]^2 dt \qquad [14]$$

With the time averages, eqn [13] can be written as

$$(E_{0x}^2 + E_{0y}^2)^2 - (E_{0x}^2 - E_{0y}^2)^2 - (2E_{0x}E_{0y}\cos\delta)^2 = (2E_{0x}E_{0y}\sin\delta)^2 \qquad [15]$$

Each term in this equation can be identified with a Stokes parameter:

$$\begin{aligned} s_0 &= \langle E_{0x}^2\rangle + \langle E_{0y}^2\rangle \qquad s_1 = \langle E_{0x}^2\rangle - \langle E_{0y}^2\rangle \\ s_2 &= \langle 2E_{0x}E_{0y}\cos\delta\rangle \quad s_3 = \langle 2E_{0x}E_{0y}\sin\delta\rangle \end{aligned} \qquad [16]$$

Equation [15] can now be written as

$$s_0^2 - s_1^2 - s_2^2 = s_3^2 \qquad [17]$$

For a polarized wave, only three of the Stokes parameters are independent. This agrees with the requirement placed upon elements of the Hermitian, coherency matrix, introduced above.

With this demonstration of the connection between the Stokes parameters and the polarization ellipse, the Stokes parameters can be written in terms of the parameters of the polarization ellipse.

$$\begin{aligned} s_1 &= s_0\cos 2\varphi\cos 2\theta \\ s_2 &= s_0\cos 2\varphi\sin 2\theta \\ s_3 &= s_0\sin 2\varphi \end{aligned} \qquad [18]$$

Mueller Calculus

Mueller pointed out that the Stokes parameters can be thought of as elements of a column matrix or a 4-vector (Table 2)

$$\mathcal{S} = \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{pmatrix} \qquad [19]$$

Each element of the Stokes vector represents a measurable intensity. The vector can represent not only polarized 1ight, but unpolarized or partially polarized light.

In 1929, Soleillet discovered that an optical device performed a linear transformation on the input wave and, in 1942, Perrin put this fact into a matrix formalism involving the Stokes vectors:

$$\mathcal{S}_1 = \mathcal{M}\cdot\mathcal{S}_0 \qquad [20]$$

Mueller used experimentally derived 4×4 matrices, the $\mathcal{M}$ in eqn [20], to describe the effect of an optical device on a light wave's polarization. The $\mathcal{S}$'s in eqn [20] are column matrices whose elements

Table 2 Mueller matrices for polarizers

Polarizer	*Transmission axis*	*Mueller matrix*
Linear	Horizontal	$\frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
	Vertical	$\frac{1}{2}\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
	$+45°$	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
	$-45°$	$\frac{1}{2}\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
Circular	Right	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$
	Left	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}$

Table 3 Mueller matrices for retarders

Retarder	*Transmission axis*	*Mueller matrix*
Quarter-wave plate	Horizontal	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}$
	Vertical	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$
	$+45°$	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$
	$-45°$	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}$
Half-wave plate	0° or 90°	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$
	$\pm 45°$	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$

are the Stokes parameters (eqn [16]). The matrices are based upon an assumed linear relationship between the incident and transmitted beams (Table 3).

The analysis of the effect of a number of polarizers and retarders is made easier by the use of the Mueller–Stokes matrix calculus, coupled with the use of the Stokes vectors. To determine the Mueller matrices, one must measure the effect a device has on unpolarized, horizontally polarized, linearly polarized at $+45°$, and right-circularly polarized light. It is then an algebraic exercise to generate the elements of the matrix. A few optical devices and their Mueller matrix are listed in Tables 2 and 3.

The Mueller matrix contains 16 parameters but there are only seven independent parameters. The matrix contains no information about absolute phase but it handles partially polarized and unpolarized light without modification (Table 4).

Jones Vector

There is one other representation of polarized light, complementary to the Stokes vector, developed by Clark Jones in 1941 and called the Jones vector. It is superior to the Stokes vector in that it handles light of a known phase and amplitude with a reduced number of parameters. However, it is inferior to the Stokes vector in that, unlike the Stokes representation, which is experimentally determined, the Jones representation cannot handle unpolarized or partially polarized light. The Jones vector is a theoretical

construct that can only describe light with a well-defined phase and frequency. The density matrix formalism can be used to correct the shortcomings of the Jones vector, but then the simplicity of the Jones representation is lost.

Assuming the coordinate system is such that the electromagnetic wave is propagating along the z-axis, it was shown earlier that any polarization could be decomposed into two orthogonal $\mathbf{E}$ vectors, i.e., parallel to the x and y directions. The Jones vector is defined as a two-row, column matrix consisting of the complex components in the x and y direction:

$$\mathbf{E} = \begin{bmatrix} E_{0x}\, e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi_1)} \\ E_{0y}\, e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi_2)} \end{bmatrix} \qquad [21]$$

Table 4 Jones vectors

Horizontal polarization	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	Vertical polarization	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$
$+45°$ polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$-45°$ polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$
Right circular polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ i \end{bmatrix}$	Left circular polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -i \end{bmatrix}$

If absolute phase is not an issue, then we may normalize the vector by dividing by that number (real or complex) that simplifies the components but keeps the sum of the square of the components equal to one. For example, assume that $E_{0x} = E_{0y}$, then

$$\mathbf{E} = E_{0x}\, e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi_1)} \begin{bmatrix} 1 \\ e^{i\delta} \end{bmatrix} \qquad [22]$$

The normalized vector would be the terms contained within the bracket, each divided by $\frac{1}{\sqrt{2}}$.

The general form of the Jones vector is

$$\mathbf{E} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \qquad \mathbf{E}^* = [\mathbf{A}^*\mathbf{B}^*] \qquad [23]$$

Some examples of Jones vectors are shown in Table 4.

Jones Calculus

The Jones calculus is complementary to the Mueller calculus and operates on the Jones vector, (eqn [23]), similar to the way the Mueller matrix operates on the Stokes vector (Table 5).

The Jones matrix contains eight independent parameters with no redundancy, making it simpler than the Mueller calculus. However, the Jones calculus only applies to polarized light. The Jones calculus can be extended, using the density matrix formalism to allow manipulation of unpolarized

Table 5 Jones matrices for retarders

Retarder	*Transmission axis*	*Jones matrix*
	Horizontal	$\begin{bmatrix} e^{i\pi/4} & 0 \\ 0 & e^{-i\pi/4} \end{bmatrix}$
	Vertical	$\begin{bmatrix} e^{-i\pi/4} & 0 \\ 0 & e^{i\pi/4} \end{bmatrix}$
	$+45°$	$\frac{1}{2}\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$
Quarter-wave plate	$-45°$	$\frac{1}{2}\begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}$
Half-wave plate	$0°$ or $90°$	$\begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}$
	$\pm 45°$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Table 6 Jones matrices for polarizers

Polarizer	*Transmission axis*	*Jones matrix*
Linear	Horizontal	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$
	Vertical	$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$
	$+45°$	$\frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$
	$-45°$	$\frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$
Circular	Right	$\frac{1}{2}\begin{bmatrix} 1 & -i \\ i & 1 \end{bmatrix}$
	Left	$\frac{1}{2}\begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix}$

light, but with a loss of simplicity. The matrix equation for Jones calculus is

$$\mathbf{E}_{\text{out}} = \mathcal{W}\mathbf{E}_{\text{in}} \qquad [24]$$

A few optical devices and their Jones matrices, $\mathcal{W}$, are listed in Tables 5 and 6.

For every matrix in Jones calculus, there is a matrix in Mueller calculus, but the converse is not true. For example, it is possible to construct a depolarizer, by using a thick piece of opal glass in the visible or by using gold covered sandpaper in the infrared. Such a device can be described in Mueller calculus, but there is no matrix for such a device in Jones calculus (Tables 5 and 6).

See also

Polarization: Introduction.

Further Reading

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.

Clark Jones R (1941) A new calculus for the treatment of optical systems. *Journal of the Optical Society of America* 32: 488–503.

Clark Jones R (1942) A new calculus for the treatment of optical systems. *Journal of the Optical Society of America* 31: 486–493.

Guenther RD (1990) *Modern Optics*. New York: Wiley.

Hecht E (1998) *Optics*, 3rd edn. Reading, MA: Addison-Wesley.

Klein MV (1970) *Optics*. New York: Wiley.

Kliger DS, Lewis JW and Randall CE (1990) *Polarized Light in Optics and Spectroscopy*. Boston, MA: Academic Press.

McMaster WH (1954) Polarization and the Stokes parameters. *American Journal of Physics* 22: 351–362.

Mueller H (1948) The foundation of optics. *Journal of the Optical Society of America* 38: 661.

Q

QUANTUM ELECTRODYNAMICS

Contents

Quantum Theory of the Electromagnetic Field
Cavity QED
Cavity QED in Semiconductors

Quantum Theory of the Electromagnetic Field

I Bialynicki-Birula and Z Bialynicka-Birula,
Polish Academy of Sciences, Warsaw, Poland

© 2005, Elsevier Ltd. All Rights Reserved.

The quantization of the free electromagnetic field is explained by analogy with the quantization of a collection of harmonic oscillators.

Brief History

The first step in the formulation of the quantum theory of the electromagnetic field (EMF) was the 'quantum hypothesis' put forward in 1900 by Planck. According to this hypothesis the exchange of energy between the EMF and the cavity walls occurs in portions $h\nu$, where h – the quantum of action – is a new fundamental constant (the Planck constant) and ν is the frequency of the material oscillator in the wall. With the help of this hypothesis Planck explained the spectrum of the electromagnetic radiation in thermodynamic equilibrium with the cavity walls. The second step was made in 1905 by Einstein, who postulated that the quantum hypothesis of Planck should be applied directly to the EMF. The assumption that the energy of the electromagnetic radiation is carried by light quanta (named photons by Lewis in 1926), each of energy $h\nu$, enabled Einstein to explain the puzzling properties of the photoelectric effect, namely, the gross disparity between the tiny classical energy flux delivered to an electron and the amount of energy needed to eject the photo-electron from the metal. After the formulation of quantum mechanics, it became possible to complete the quantum theory of the EMF (Dirac 1927) by applying the general rules of quantization to the classical Maxwell theory. However, this theory ran into difficulties when it was applied to the interaction of the EMF with charged matter.

The lightest carriers of the electric charge are the electrons and their dynamics determines to a very large extent the electromagnetic properties of ordinary matter. The main principles of quantum electrodynamics (QED) – the theory unifying the quantum theory of the EMF and the relativistic theory of electrons (Dirac 1928) – have been set forth by Heisenberg and Pauli already in 1929 but the proper tool (renormalization theory) to perform highly accurate calculations was developed by Feynman and Schwinger in 1949. In 1967 QED was unified by Weinberg and Salam with the theory of weak interactions. Later, after the unification with the theory of strong interactions, QED became part of the so called 'standard model' – the present day theory of elementary ingredients of matter: quarks, leptons, gluons, intermediate bosons, and photons. Therefore, the modern quantum theory of the EMF is not a self-contained theory but part of a much more encompassing theory. Often it requires some input from a more general theory or from phenomenology. However, the quantum theory of the EMF with only a very

crude description of its coupling to the rest of the world (for example, in the form of boundary conditions or constitutive relations) is still not only a good starting point to a more elaborate theory but also a source of important information about the behavior of photons under realistic conditions. The main reason why treating photons as free particles gives a reasonably good approximation is the smallness of the Sommerfeld fine structure constant $\alpha = e^2/4\pi\varepsilon_0\hbar c \approx 1/137$ that measures the strength of the electromagnetic interactions between photons and charged particles.

Electromagnetic Field as a Collection of Harmonic Oscillators

In order to apply the same quantization procedure as in quantum mechanics, it is useful to formulate classical electrodynamics in a way that exhibits its close relationship with mechanical systems. Many properties of the electromagnetic field may be described and understood in the simplest case of the free field. Therefore, in this article we restrict ourselves to the quantum description of the free EMF.

It was discovered by Rayleigh and Jeans that the electromagnetic field may be viewed as a collection of harmonic oscillators. This analogy enables one to apply well-established methods of simple quantum mechanics to describe the much more complicated problem of the quantum field. The dynamics of a single, one-dimensional harmonic oscillator is governed by the Hamiltonian:

$$H(p,q) = \frac{p^2}{2m} + \frac{m\omega^2 q^2}{2}$$

The equations of motion for the position and momentum of a harmonic oscillator have the same form in classical and in quantum theory:

$$\frac{\mathrm{d}q(t)}{\mathrm{d}t} = \frac{p(t)}{m}, \qquad \frac{\mathrm{d}p(t)}{\mathrm{d}t} = -m\omega^2 q(t)$$

The solutions of these equations describe harmonic oscillations with the frequency ω

$$q(t) = q_0\cos(\omega t) + (p_0/m\omega)\sin(\omega t)$$

$$p(t) = -m\omega q_0\sin(\omega t) + p_0\cos(\omega t)$$

A complete description of a multidimensional harmonic oscillator may be obtained in terms of normal modes. The transformation to normal modes leads to a decomposition of the Hamiltonian of the system into the sum, one term for each mode,

$$H = \sum_k \left(\frac{p_k^2}{2m} + \frac{m\omega_k^2 q_k^2}{2}\right)$$

Thus, each mode may be viewed as a separate system – a one-dimensional harmonic oscillator with characteristic frequency ω_k – noninteracting with the remaining oscillators.

The Maxwell equations for the free EMF read

$$\partial_t \mathbf{B}(\mathbf{r},t) = -\nabla\times\mathbf{D}(\mathbf{r},t)/\varepsilon,\ \nabla\cdot\mathbf{B}(\mathbf{r},t) = 0$$

$$\partial_t \mathbf{D}(\mathbf{r},t) = \nabla\times\mathbf{B}(\mathbf{r},t)/\mu,\ \nabla\cdot\mathbf{D}(\mathbf{r},t) = 0$$

In the vacuum, the constants ε and μ must be replaced by ε_0 and μ_0. The Maxwell equations lead to the wave equation for the vector $\mathbf{D}(\mathbf{r},t)$

$$\left(\frac{1}{c^2}\partial_t^2 - \Delta\right)\mathbf{D}(\mathbf{r},t) = 0$$

and to the same equation for the vector $\mathbf{B}(\mathbf{r},t)$. The solution of this equation obtained by the separation of variables has the form

$$\mathbf{D}(\mathbf{r},t) = a(t)\mathbf{u}(\mathbf{r})$$

where $a(t)$ satisfies the harmonic equation

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}a(t) + \omega^2 a(t) = 0$$

and $\mathbf{u}(\mathbf{r})$ satisfies the Helmholtz equation

$$(\Delta + (\omega/c)^2)\mathbf{u}(\mathbf{r}) = 0$$

Without going into mathematical details we shall assume that there exist a complete set of vector functions $\mathbf{u}_k(\mathbf{r})$, called the mode functions, satisfying the Helmholtz equations

$$(\Delta + (\omega_k/c)^2)\mathbf{u}_k(\mathbf{r}) = 0$$

the divergence condition $\nabla\cdot\mathbf{u}_k(\mathbf{r}) = 0$, the orthonormality conditions

$$\int \mathrm{d}^3 r\,\mathbf{u}_k(\mathbf{r})\cdot\mathbf{u}_l(\mathbf{r}) = \delta_{kl}$$

and also the appropriate boundary conditions. In most cases of physical interest such a set does exist.

Given the set of mode functions $\mathbf{u}_k(\mathbf{r})$ one may expand the vector field $\mathbf{D}(\mathbf{r},t)$ at each t into the series

$$\mathbf{D}(\mathbf{r},t) = -\sum_k p_k(t)\mathbf{u}_k(\mathbf{r})$$

The minus sign is purely conventional; it makes the correspondence with the harmonic oscillator a closer one. The same procedure may be repeated for the

vector $\mathbf{B}$ with a different complete set of functions $\mathbf{v}_k(\mathbf{r})$, resulting in the decomposition

$$\mathbf{B}(\mathbf{r},t)=\sum_k q_k(t)\mathbf{v}_k(\mathbf{r})$$

The Maxwell equations are satisfied if the coefficient functions $p_k(t)$ and $q_k(t)$ obey the equations of motion for harmonic oscillators

$$\frac{\mathrm{d}q_k(t)}{\mathrm{d}t}=\frac{p_k(t)}{\varepsilon},\quad \frac{\mathrm{d}p_k(t)}{\mathrm{d}t}=-\varepsilon\omega^2 q_k(t) \qquad [1]$$

and the functions $\mathbf{v}_k(\mathbf{r})$ are related to $\mathbf{u}_k(\mathbf{r})$ by the formulas

$$\mathbf{v}_k(\mathbf{r})=\nabla\times\mathbf{u}_k(\mathbf{r})$$

Note that the electric constant ε plays the same role as the mass for a mechanical oscillator. The calculation of the Hamiltonian, the total field energy,

$$H=\int \mathrm{d}^3r\left(\frac{\mathbf{D}^2}{2\varepsilon}+\frac{\mathbf{B}^2}{2\mu}\right)$$

completes the reduction of the electromagnetic field to a collection of harmonic oscillators. Upon inserting the representation of the field vectors in terms of the complete set of functions $\mathbf{u}_k(\mathbf{r})$ into the integral, one obtains the following formula

$$H=\sum_k\left(\frac{p_k^2}{2\varepsilon}+\frac{\varepsilon\omega_k^2q_k^2}{2}\right)$$

which differs from its mechanical counterpart only by the replacement $m\to\varepsilon$. The time dependence has been omitted here because the Hamiltonian is a constant of motion.

Mode Functions

The simplest example of a complete set of mode functions is that obtained for a rectangular metallic cavity. In this case the functions $\mathbf{u}_k(\mathbf{r})$ may be expressed in terms of trigonometric functions. In order to write them down, it is convenient to replace one label k by a collection of three numbers $\boldsymbol{\kappa}=\{\kappa_x,\kappa_y,\kappa_z\}$ and an additional index λ taking on the values 1 and 2. The mode functions $\mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r})$ satisfying the proper boundary conditions have the following components

$$\mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r})=N\begin{pmatrix}a_x^{(\lambda)}\cos(\kappa_x x)\sin(\kappa_y y)\sin(\kappa_z z)\\ a_y^{(\lambda)}\sin(\kappa_x x)\cos(\kappa_y y)\sin(\kappa_z z)\\ a_z^{(\lambda)}\sin(\kappa_x x)\sin(\kappa_y y)\cos(\kappa_z z)\end{pmatrix}$$

where $N=\sqrt{8/(L_xL_yL_z)}$, the coefficients $\{\kappa_x,\kappa_y,\kappa_z\}$ have the form $\{\kappa_x,\kappa_y,\kappa_z\}=\{\pi n_x/L_x,\pi n_y/L_y,\pi n_z/L_z\}$ and all the n's are natural numbers. The components of both vectors $\mathbf{a}^{(\lambda)}$ are subjected to the conditions $\boldsymbol{\kappa}\cdot\mathbf{a}^{(\lambda)}=0$ (vanishing divergence) and orthonormalization $\mathbf{a}^{(\lambda)}\cdot\mathbf{a}^{(\lambda')}=\delta_{\lambda\lambda'}$. For each set $\{\kappa_x,\kappa_y,\kappa_z\}$ one must choose two independent solutions of these conditions $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$. These two vectors define two states of linear polarization. The mode function $\mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r})$ is a solution of the Helmholtz equation for the value of ω/c equal to the length of the vector $\boldsymbol{\kappa}$. The expansion of the EMF vectors into the mode functions in a rectangular cavity has the form

$$\mathbf{D}(\mathbf{r},t)=-\sum_{\boldsymbol{\kappa}\lambda}p_{\boldsymbol{\kappa}\lambda}(t)\mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r})$$

$$\mathbf{B}(\mathbf{r},t)=\sum_{\boldsymbol{\kappa}\lambda}q_{\boldsymbol{\kappa}\lambda}(t)\nabla\times\mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r})$$

where the summation extends over all allowed values of $\boldsymbol{\kappa}$ and two values of λ. One may explicitly write down mode functions also for a cylindrical and for a spherical cavity. They involve Bessel functions and are obtained by solving the Helmholtz equation in cylindrical and spherical coordinates, respectively.

One often uses another set of mode functions that does not correspond to any physical boundary conditions but is very helpful in the theoretical analysis. It is obtained by imposing the so-called periodic boundary conditions that were first introduced by Born in the study of crystal lattices. The periodic boundary conditions require that the electromagnetic field is periodic in space; it does not change when the coordinates are displaced by the 'period of the lattice' in all three directions. Assuming, for simplicity, that this period is the same, equal to L, in all directions, one obtains the following conditions for the mode functions

$$\mathbf{u}_k(\mathbf{r})=\mathbf{u}_k(\mathbf{r}+L\mathbf{i}_x)=\mathbf{u}_k(\mathbf{r}+L\mathbf{i}_y)=\mathbf{u}_k(\mathbf{r}+L\mathbf{i}_z)$$

where $\mathbf{i}_i$ are the unit vectors in the directions of the coordinate axes. The complete set of solutions of the Helmholtz equation satisfying the periodic boundary conditions is labeled by the wavevector $\mathbf{k}=\{2\pi n_x/L,2\pi n_y/L,2\pi n_z/L\}$, where all the n's are arbitrary integers, and by the polarization index λ. The value of ω/c in the Helmholtz equation is equal in this case to the length of the vector k. Each member of the set $\mathbf{u}_{\mathbf{k}\lambda}$ has the form

$$\mathbf{u}_{\mathbf{k}\lambda}(\mathbf{r})=\frac{1}{\sqrt{V}}\mathbf{e}_{\mathbf{k}\lambda}\exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r}) \qquad [2]$$

where $V = L^3$. The polarization vectors $\mathbf{e}_{\mathbf{k}\lambda}$ must be orthogonal to the wave vector $\mathbf{k}$. It is convenient to impose the unitary orthonormality conditions $\mathbf{e}_{\mathbf{k}\lambda} \cdot \mathbf{e}^*_{\mathbf{k}\lambda'} = \delta_{\lambda\lambda'}$ and also to assume that $\mathbf{e}^*_{\mathbf{k}\lambda} = \mathbf{e}_{-\mathbf{k}\lambda}$. The expansion of the EMF vectors into the mode functions obeying the periodic boundary conditions has the form

$$\mathbf{D}(\mathbf{r},t) = -\frac{1}{\sqrt{V}} \sum_{\mathbf{k}\lambda} P_{\mathbf{k}\lambda}(t)\mathbf{e}_{\mathbf{k}\lambda} e^{i\mathbf{k}\cdot\mathbf{r}}$$

$$\mathbf{B}(\mathbf{r},t) = \frac{1}{\sqrt{V}} \sum_{\mathbf{k}\lambda} Q_{\mathbf{k}\lambda}(t) i\mathbf{k} \times \mathbf{e}_{\mathbf{k}\lambda} e^{i\mathbf{k}\cdot\mathbf{r}} \qquad [3]$$

where the sum extends again over all allowed values of the vectors $\mathbf{k}$ and over two values of λ. The expansion coefficients $P_{\mathbf{k}\lambda}$ and $Q_{\mathbf{k}\lambda}$ are now complex. In order to guarantee that the field vectors are real, they must obey the conditions

$$P^*_{\mathbf{k}\lambda} = P_{-\mathbf{k}\lambda},\ Q^*_{\mathbf{k}\lambda} = Q_{-\mathbf{k}\lambda}$$

These conditions are satisfied if $P_{\mathbf{k}\lambda}$ and $Q_{\mathbf{k}\lambda}$ are parametrized by the real (but otherwise arbitrary) variables $p_{\mathbf{k}\lambda}$ and $q_{\mathbf{k}\lambda}$

$$P_{\mathbf{k}\lambda} = \frac{p_{\mathbf{k}\lambda} + p_{-\mathbf{k}\lambda}}{2} + \frac{\varepsilon\omega_{\mathbf{k}}(q_{\mathbf{k}\lambda} - q_{-\mathbf{k}\lambda})}{2i}$$

$$Q_{\mathbf{k}\lambda} = \frac{q_{\mathbf{k}\lambda} + q_{-\mathbf{k}\lambda}}{2} + \frac{i(p_{\mathbf{k}\lambda} - p_{-\mathbf{k}\lambda})}{2\varepsilon\omega_{\mathbf{k}}}$$

The energy of the EMF splits again into a sum of independent contributions from all the modes

$$H = \sum_{\mathbf{k}\lambda} \omega_{\mathbf{k}} \left(\frac{p_{\mathbf{k}\lambda} p_{\mathbf{k}\lambda}}{2\varepsilon\omega_{\mathbf{k}}} + \frac{\varepsilon\omega_{\mathbf{k}} q_{\mathbf{k}\lambda} q_{\mathbf{k}\lambda}}{2} \right)$$

An important property of the periodic mode functions [2] is that not only the Hamiltonian but also the total momentum of the field splits into a sum of independent contributions. The total momentum $\mathbf{P}$ of the EMF is

$$\mathbf{P} = \int d^3r\, \mathbf{D} \times \mathbf{B}$$

Upon substituting into this expression the expansions [3] of the field vectors, one obtains, with the use of the orthonormality conditions,

$$\mathbf{P} = \sum_{\mathbf{k}\lambda} \mathbf{k} \left(\frac{p_{\mathbf{k}\lambda} p_{\mathbf{k}\lambda}}{2\varepsilon\omega_{\mathbf{k}}} + \frac{\varepsilon\omega_{\mathbf{k}} q_{\mathbf{k}\lambda} q_{\mathbf{k}\lambda}}{2} \right)$$

Canonical Quantization

The canonical quantization in quantum mechanics consists in replacing the c-numbers representing the classical canonical variables q_k and p_k by the operators $\hat{q}_k$ and $\hat{p}_k$. These operators are assumed to obey the canonical commutation relations

$$[\hat{q}_k, \hat{p}_l] = i\hbar\delta_{kl} \qquad [4]$$

The canonical quantization of the EMF may proceed along the same lines as in quantum mechanics. All the variables q_k and p_k appearing in the decomposition into normal modes are replaced by the operators. When this is done, the field vectors $\mathbf{B}$ and $\mathbf{D}$ become the field operators $\hat{\mathbf{B}}$ and $\hat{\mathbf{D}}$ and the expression for the energy of the field becomes the Hamiltonian operator $\hat{H}$

$$\hat{H} = \sum_k \frac{\hat{p}_k^2}{2\varepsilon} + \frac{\varepsilon\omega_k^2 \hat{q}_k^2}{2}$$

The only difference between quantum mechanics and quantum field theory is that in the second case the number of canonical variables is infinite but this may cause complications only when one wants to give the quantum theory of the EMF a firm mathematical foundation. In physical applications one may think of the number of field oscillators as being very large but finite.

Time evolution of a quantum system is described by associating the time dependence either with the operators (the Heisenberg picture) or with the state vectors (the Schrödinger picture). In both pictures the time dependence is governed by the system Hamiltonian. In the Heisenberg picture the time dependence of the operators is given by the commutators with $\hat{H}$

$$\frac{d\hat{q}_k(t)}{dt} = \frac{1}{i\hbar}[\hat{q}_k(t), \hat{H}] = \frac{\hat{p}_k(t)}{\varepsilon}$$

$$\frac{d\hat{p}_k(t)}{dt} = \frac{1}{i\hbar}[\hat{p}_k(t), \hat{H}] = -\varepsilon\omega_k^2 \hat{q}_k(t)$$

These equations have the same form as in the classical theory (cf. eqn [1]). Therefore, in the Heisenberg picture the field operators will also obey the same equations as in the classical theory – the Maxwell equations:

$$\partial_t \hat{\mathbf{B}}(\mathbf{r},t) = -\nabla \times \hat{\mathbf{D}}(\mathbf{r},t)/\varepsilon,\ \ \nabla \cdot \hat{\mathbf{B}}(\mathbf{r},t) = 0$$

$$\partial_t \hat{\mathbf{D}}(\mathbf{r},t) = \nabla \times \hat{\mathbf{B}}(\mathbf{r},t)/\mu,\ \ \nabla \cdot \hat{\mathbf{D}}(\mathbf{r},t) = 0$$

Of course, the average values of the field operators evaluated in both pictures are the same and for every quantum state they obey the classical Maxwell equations.

The commutation relations [4] for the canonical operators and the completeness of the mode functions imply the following equal-time commutation relations for the field operators

$$[\hat{B}_i(\mathbf{r},t),\hat{D}_j(\mathbf{r}',t)] = \mathrm{i}\hbar\sum_k \varepsilon_{ijk}\partial_k\delta^{(3)}(\mathbf{r}-\mathbf{r}')$$

These relations do not depend on the choice of the mode functions. They summarize in a universal manner the canonical properties of the electromagnetic field operators.

Annihilation and Creation Operators

The canonical variables q_k and p_k are convenient for bringing home the analogy between mechanical oscillators and wave fields. However, there exists also a complementary, corpuscular aspect of the quantum theory of the EMF and that is most easily expressed in terms of photons. The best way to describe photons in quantum field theory is in terms of annihilation and creation operators. These operators were introduced by Schrödinger in the context of the quantum-mechanical harmonic oscillator but they have become really indispensable in quantum field theory. Formally, the annihilation $\hat{a}_k$ and the creation $\hat{a}_k^\dagger$ operators are defined as the following complex combinations of canonical operators

$$\begin{aligned}\hat{a}_k &= \sqrt{\frac{\varepsilon\omega}{2\hbar}}\left(\hat{q}_k + \mathrm{i}\frac{\hat{p}_k}{\varepsilon\omega}\right)\\ \hat{a}_k^\dagger &= \sqrt{\frac{\varepsilon\omega}{2\hbar}}\left(\hat{q}_k - \mathrm{i}\frac{\hat{p}_k}{\varepsilon\omega}\right)\end{aligned} \qquad [5]$$

Annihilation and creation operators are dimensionless, one is the Hermitian adjoint of the other, and they obey very simple commutation relations:

$$[\hat{a}_k,\hat{a}_l^\dagger] = \delta_{kl} \qquad [6]$$

The Hamiltonian expressed in terms of annihilation and creation operators reads

$$\hat{H} = \sum_k \hbar\omega_k\hat{a}_k^\dagger\hat{a}_k + \frac{1}{2}\sum_k \hbar\omega_k \qquad [7]$$

This Hamiltonian determines the following time dependence of the annihilation and creation operators

$$\hat{a}_k(t) = \hat{a}_k\exp(-\mathrm{i}\omega_k t),\ \hat{a}_k^\dagger(t) = \hat{a}_k^\dagger\exp(\mathrm{i}\omega_k t)$$

The field operators expressed in terms of annihilation and creation operators corresponding to the periodic boundary conditions are

$$\hat{\mathbf{D}}(\mathbf{r},t) = \mathrm{i}\sum_{\mathbf{k}\lambda}\sqrt{\frac{\hbar\varepsilon\omega_\mathbf{k}}{2V}}\hat{a}_{\mathbf{k}\lambda}\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_\mathbf{k}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}} + \mathrm{H.c.}$$

$$\hat{\mathbf{B}}(\mathbf{r},t) = \mathrm{i}\sum_{\mathbf{k}\lambda}\sqrt{\frac{\hbar}{2V\varepsilon\omega_\mathbf{k}}}\hat{a}_{\mathbf{k}\lambda}\mathbf{k}\times\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_\mathbf{k}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}} + \mathrm{H.c.}$$

where H.c. stands for the Hermitian conjugate of the first term. In the limiting case for infinite empty space, when $V\to\infty$, the Fourier sum becomes the Fourier integral:

$$\begin{aligned}\hat{\mathbf{D}}(\mathbf{r},t) = \mathrm{i}\sum_\lambda\int \mathrm{d}^3k\sqrt{\frac{\hbar\varepsilon\omega_\mathbf{k}}{2(2\pi)^3}}(&\hat{a}_\lambda(\mathbf{k})\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_\mathbf{k}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\\ &-\hat{a}_\lambda^\dagger(\mathbf{k})\mathbf{e}_{\mathbf{k}\lambda}^*\mathrm{e}^{\mathrm{i}\omega_\mathbf{k}t-\mathrm{i}\mathbf{k}\cdot\mathbf{r}})\end{aligned}$$

$$\begin{aligned}\hat{\mathbf{B}}(\mathbf{r},t) = \mathrm{i}\sum_\lambda\int \mathrm{d}^3k\sqrt{\frac{\hbar}{2\varepsilon\omega_\mathbf{k}(2\pi)^3}}(&\hat{a}_\lambda(\mathbf{k})\mathbf{k}\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_\mathbf{k}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\\ &-\hat{a}_\lambda^\dagger(\mathbf{k})\mathbf{k}\times\mathbf{e}_{\mathbf{k}\lambda}^*\mathrm{e}^{\mathrm{i}\omega_\mathbf{k}t-\mathrm{i}\mathbf{k}\cdot\mathbf{r}})\end{aligned}$$

In this case, the commutation relations involve the (three-dimensional) Dirac delta function

$$[\hat{a}_\lambda(\mathbf{k}),\hat{a}_{\lambda'}^\dagger(\mathbf{k}')] = \delta_{\lambda\lambda'}\delta^{(3)}(\mathbf{k}-\mathbf{k}') \qquad [8]$$

The field operators may also be expressed in terms of the vector potential operator $\hat{\mathbf{A}}(\mathbf{r},t)$

$$\hat{\mathbf{D}}(\mathbf{r},t) = -\varepsilon\partial_t\hat{\mathbf{A}}(\mathbf{r},t),\ \hat{\mathbf{B}}(\mathbf{r},t) = \nabla\times\hat{\mathbf{A}}(\mathbf{r},t)$$

$$\begin{aligned}\hat{\mathbf{A}}(\mathbf{r},t) = \sum_\lambda\int \mathrm{d}^3k\sqrt{\frac{\hbar}{2\varepsilon\omega_\mathbf{k}(2\pi)^3}}(&\hat{a}_\lambda(\mathbf{k})\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_\mathbf{k}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\\ &+\hat{a}_\lambda^\dagger(\mathbf{k})\mathbf{e}_{\mathbf{k}\lambda}^*\mathrm{e}^{\mathrm{i}\omega_\mathbf{k}t-\mathrm{i}\mathbf{k}\cdot\mathbf{r}})\end{aligned}$$

The vector potential is needed for the description of the interaction with charged particles.

Space of Quantum States

The standard basis in the space of quantum states consists of the state vectors of stationary states – the eigenvectors of the Hamiltonian.

Vacuum State

The lowest energy state of the EMF – the ground state of the system – is often called the vacuum state. The state vector describing the vacuum state is denoted by $|0\rangle$. The vacuum state is the common ground state of all harmonic oscillators making up the EMF. Since the ground state of a single harmonic oscillator has the energy $\hbar\omega/2$, the energy of the ground state for an infinite number of oscillators is infinite (last term in eqn [7]). This infinity has no physical consequences because only the energy differences are observable. In the presence of obstacles (mirrors, resonators, beamsplitters, etc.), owing to the change in the mode structure, there will be a change in the energy of the new ground state relative to the energy in empty space. The energy differences between various ground states do have physical significance.

Fock Basis

It follows from the commutation relations [6] and from the form [7] of the Hamiltonian that the operator $\hat{a}_k$ decreases the energy of the EMF by the Planck quantum of energy $\hbar\omega_k$. Therefore, all operators $\hat{a}_k$ must annihilate the ground state vector, $\hat{a}_k|0\rangle = 0$, because there is no state with a lower energy. Each creation operator $\hat{a}_k^\dagger$ acting on any state vector adds one excitation with the energy $\hbar\omega_k$. These excitations of the electromagnetic field with characteristic energies $\hbar\omega_k$ are interpreted as photons. The state vectors of the form $\hat{a}_k^\dagger|0\rangle$ describe one-photon states, the state vectors of the form $\hat{a}_k^\dagger\hat{a}_l^\dagger|0\rangle$ describe two-photon states, etc. All the vectors generated from the vacuum state vector by acting with the creation operators form a basis, called the Fock basis. The general state vector of this basis is characterized by a sequence of natural numbers $\{n_1, n_2, \ldots\}$ and is generated from the vacuum state according to the formula

$$|n_1, n_2, \ldots\rangle = (n_1!n_2!\cdots)^{-1/2}(\hat{a}_1^\dagger)^{n_1}(\hat{a}_2^\dagger)^{n_2}\ldots|0\rangle \qquad [9]$$

where the prefactor is needed for normalization. The number n_k tells us how many photons of type k are present in this state. The spatial structure of a field excitation created by $\hat{a}_k^\dagger$ is described by the mode function $\mathbf{u}_k(\mathbf{r})$. Thus, this mode function plays the same role as the wavefunction $\psi(\mathbf{r})$ of a massive particle and the set of indices labeling the mode functions is the analog of quantum numbers for $\psi(\mathbf{r})$. In particular, for the mode function [2], the photons have a definite momentum $\hbar\mathbf{k}$. In a cylindrical waveguide the photons may have a definite momentum only along the axis of the cylinder. In a spherical cavity the photons may have the definite value of square of the total angular momentum and its projection on a given axis. The Fock states are highly nonclassical; all average values of the EMF in these states vanish.

All states of the EMF are either pure or mixed. Pure states are described by all linear superpositions of the basis states. In particular, by superposing photon states with a range of energies, one may form nonmonochromatic photon wavepackets. All superpositions are allowed since photons do not have any absolutely conserved quantum numbers (like, e.g., charge). Mixed states are described by the density operators. Among the mixed states the prominent role is played by the thermal state describing the EMF in equilibrium with a reservoir at temperature T.

Coherent States

The one-mode coherent states $|\alpha\rangle$ are obtained from the vacuum state by the action of the unitary displacement operator $\hat{D}(\alpha, \alpha^*)$

$$|\alpha\rangle = D(\alpha, \alpha^*)|0\rangle = \mathrm{e}^{\alpha\hat{a}^\dagger - \alpha^*\hat{a}}|0\rangle = \mathrm{e}^{-|\alpha|^2/2}\mathrm{e}^{\alpha\hat{a}^\dagger}|0\rangle$$

where $\hat{a}$ and $\hat{a}^\dagger$ are the annihilation and creation operators corresponding to a given mode. Coherent states are superpositions of an infinite number of Fock states. The multimode coherent state is generated by the product of displacement operators

$$|\alpha_1, \alpha_2, \ldots\rangle = \mathrm{e}^{\sum_k (\alpha_k\hat{a}_k^\dagger - \alpha_k^*\hat{a}_k)}|0\rangle$$

In reality, the coherent states are generated by the action of an external (strong) time-dependent electric current. In that case the coefficients α_k are the projections of the current on the mode functions. The average values of the field operators $\langle\hat{B}_i(\mathbf{r}, t)\rangle$ and $\langle\hat{D}_j(\mathbf{r}, t)\rangle$ in coherent states are the solutions of the Maxwell equations in the presence of an external electric current.

Entangled States

The states of the EMF describing photons in different modes may exhibit correlations. The simplest such state is a two-photon state of the form $|\phi\rangle = 2^{-1/2}(\hat{a}_1^\dagger\hat{a}_2^\dagger + \hat{a}_3^\dagger\hat{a}_4^\dagger)|0\rangle$. In this case, when one photon is found, say, in the state 1, the other will be found with the probability 1 in the state 2. The states that exhibit correlations between photons belonging to different modes are called the entangled states.

Correlations occur also in the classical theory, but only for statistical mixtures when there is a loss of information. In quantum theory the correlations may arise even without a loss of information, i.e., for pure states. A striking feature of entangled states is that the decomposition of such a state into a sum of Fock states is not unique. For example, the state vector $|\phi\rangle$ may also be written in the form $|\phi\rangle = 2^{-1/2}(\hat{a}^\dagger_{1'}\hat{a}^\dagger_{2'} + \hat{a}^\dagger_{3'}\hat{a}^\dagger_{4'})|0\rangle$, where $\hat{a}^\dagger_{1'} = 2^{-1/2}(\hat{a}^\dagger_1 - \hat{a}^\dagger_3)$, $\hat{a}^\dagger_{2'} = 2^{-1/2}(\hat{a}^\dagger_2 - \hat{a}^\dagger_4)$, $\hat{a}^\dagger_{3'} = 2^{-1/2}(\hat{a}^\dagger_1 + \hat{a}^\dagger_3)$, $\hat{a}^\dagger_{4'} = 2^{-1/2}(\hat{a}^\dagger_2 + \hat{a}^\dagger_4)$. Entangled states play a crucial role in quantum information theory, especially in quantum cryptography.

Quantum Fluctuations and the Uncertainty Relations

It follows from the basic principles of quantum theory that there are no states of the system for which both canonical variables q and p have sharply defined values. The limits on the quantum fluctuations (variances) of these values are determined by the uncertainty relation. The most stringent form of this relation obtained from the commutation relations [4] (for a single mode) reads

$$(\Delta q)^2(\Delta p)^2 \geq \frac{\hbar^2}{4} + \frac{\langle \delta\hat{q}\delta\hat{p} + \delta\hat{p}\delta\hat{q}\rangle^2}{4} \qquad [10]$$

where $\delta\hat{q} = \hat{q} - \langle\hat{q}\rangle$, $\delta\hat{p} = \hat{p} - \langle\hat{p}\rangle$, $(\Delta q)^2 = \langle(\delta\hat{q})^2\rangle$, and $(\Delta p)^2 = \langle(\delta\hat{p})^2\rangle$. The states for which this inequality is saturated play a distinguished role in quantum theory. Since the uncertainty relation expresses the limitations on the quantum fluctuations of q and p, the best one can do in approaching the classical regime is to achieve equality in the uncertainty relation. That is why the quantum states that saturate [10] are viewed as the counterparts of the classical states. In the quantum theory of the EMF, the uncertainty relations express the complementarity between the magnetic and electric fields. The better one knows the magnetic field, the more uncertain is the electric field, and vice versa. However, the projections of the field operators $\hat{\mathbf{B}}$ and $\hat{\mathbf{D}}$ on the same direction commute (cf. eqn [5]). Therefore, there are no limitations on the magnetic and electric field values in a given direction and only the values in different directions are subjected to the quantum uncertainty relation.

In the vacuum state the variances of q and p for each mode are

$$(\Delta q)^2 = \frac{\hbar}{2\varepsilon\omega}, (\Delta p)^2 = \frac{\hbar\varepsilon\omega}{2} \qquad [11]$$

and they saturate the uncertainty relation. These fluctuations are called the shot noise. Very often one uses dimensionless quadrature operators $\hat{X}_1 = \hat{q}\sqrt{2\varepsilon\omega/\hbar}$ and $\hat{X}_2 = \hat{p}\sqrt{2/\varepsilon\omega\hbar}$, whose fluctuations in the vacuum state are equal to 1. All coherent states have the same fluctuations as the vacuum state. They may be viewed in the phase space as the vacuum state displaced from the origin (corresponding to the vacuum state) by the mean values of q and p.

Coherent states are not the only states that saturate the uncertainty relations. The other states with this property are called the squeezed states. The fluctuations of X_1 and X_2 for squeezed states are different from 1 but their product is still equal to 1, $(\Delta X_1)^2(\Delta X_2)^2 = 1$.

Fock states have well-defined energies and numbers of photons since they are the eigenstates of the Hamiltonian. However, the phases of the EMF are maximally uncertain; they are evenly spread over the whole range from 0 to 2π. That is why the average values of the field vectors vanish in all Fock states.

See also

Photon Picture of Light. **Quantum Electrodynamics:** Cavity QED.

Further Reading

Bialynicki-Birula I (1996) Photon wave function. In: Wolf E (ed.) *Progress in Optics*, vol. 36. Amsterdam, The Netherlands: Elsevier.

Bialynicki-Birula I and Bialynicka-Birula Z (1975) *Quantum Electrodynamics*. London: Pergamon.

Loudon R (1973) *The Quantum Theory of Light*. Oxford, UK: Clarendon Press.

Louisell W (1990) *Quantum Statistical Properties of Radiation*. New York: Wiley.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge, UK: Cambridge University Press.

Power EA (1964) *Introductory Quantum Electrodynamics*. London: Longman.

Scully MO and Zubairy MS (1997) *Quantum Optics*. Cambridge, UK: Cambridge University Press.

Cavity QED

H Walther, University of Munich and
Max-Planck-Institute for Quantum Optics, Garching, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Modification of the Spontaneous Transition Rate in Confined Space

In order to understand the modification of the spontaneous emission rate in an external cavity-like structure, it must be remembered that in quantum electrodynamics this rate is proportional to the density of modes of the electromagnetic field, i.e., the vacuum field fluctuations above the atomic transition frequency ω_0. As a consequence the spontaneous emission rate is increased if the atom is surrounded by a cavity tuned to the transition frequency. Conversely, the decay rate decreases when the cavity is mistuned. This fact was already recognized in the pioneering days of nuclear magnetic resonance by Purcell.

The spontaneous decay rate of the atom in the cavity, γ_c, will then be enhanced in relation to that in free space, γ_f, by a factor given by the ratio of the corresponding mode densities

$$\frac{\gamma_c}{\gamma_f} = \frac{\rho_c(\omega_0)}{\rho_f(\omega_0)} = \frac{2\pi Q}{V_c \omega_0^3} = \frac{Q\lambda_0^3}{4\pi^2 V_c}$$

where V_c is the volume of the cavity and Q is the quality factor of the cavity which expresses the 'sharpness' of the mode. For low-order cavities $V_c \approx \lambda_0^3$ this means that the spontaneous emission rate is increased by roughly a factor of Q. As mentioned already, when the cavity is detuned, the decay rate will decrease. In this case the atom cannot emit a photon, since the cavity is not able to accept it, and the energy will stay with the atom, i.e., its decay is inhibited.

To change the decay rate of an atom, in principle no resonator has to be present; any conducting surface near the radiator affects the mode density and, therefore, the spontaneous radiation rate. Parallel conducting planes can alter the emission rate somewhat but can only reduce the rate by a factor of 2 owing to the existence of modes, independent of the plate separation with an electric field perpendicular to the planes.

In order to demonstrate experimentally the modification of the spontaneous decay rate, it is not necessary to go to single-atom densities. The experiments where the spontaneous emission is inhibited can also be performed with large atom numbers. However, in the opposite case, when the increase of the spontaneous rate is observed, a large number of excited atoms may disturb the experiment by induced transitions. The first experimental work on inhibited spontaneous emission was done by Drexhage, Kuhn and Schäfer in 1974. The fluorescence of a thin dye film near a mirror was investigated. A reduction of the fluorescence decay by up to 25% was observed. Later experiments with microcavities filled with dye solutions were performed by de Martini *et al*. These experiments demonstrated that thresholdless lasing is possible in such systems.

Inhibited spontaneous emission was observed by Gabrielse and Dehmelt. In these neat experiments with a single electron stored in a Penning trap they observed that cyclotron orbits show lifetimes which are up to 10 times longer than that calculated for free space. The electrodes of the trap form a cavity which decouples the cyclotron motion from the vacuum field leading to the longer lifetime.

A new stage in experiments on cavity QED was reached when it was recognized that highly excited alkaline atoms are very suitable for these experiments. The main quantum numbers are in the range $n = 20-60$. Those so-called Rydberg atoms couple very strongly to the radiation field as will be discussed later. The transitions to neighboring states are in the microwave region. The cavities can therefore be built as low-order cavities with dimensions on the order of the wavelength of the transitions being in the mm or cm regions. The spontaneous rate of these transitions in free space is small owing to the low value of their transition frequency, therefore an enhancement is possible. Experiments with Rydberg atoms on the inhibition of spontaneous emission have been performed by Kleppner and coworkers and by Haroche and coworkers. In the latter experiment a 3.4 μm transition of the Cs atom was suppressed.

The first observation of enhanced atomic spontaneous emission in a resonant cavity was published by Haroche *et al*. This experiment was performed with Rydberg atoms of Na excited in the 23s state in a niobium superconducting cavity resonant at 340 GHz. Cavity tuning-dependent shortening of the lifetime was observed. The cooling of the cavity had the advantage of totally suppressing the black-body field. The latter effect is completely absent if optical transitions are observed, however, in this case it is more difficult to obtain low-order cavities. The first experiments on optical transitions were performed by Feld and collaborators. They succeeded

in demonstrating the enhancement of spontaneous transitions even in higher-order optical cavities.

In modern semiconductor devices both electronic and optical properties can be tailored with a high degree of precision. Therefore, electron–hole systems producing recombination radiation analogous to radiating atoms can be localized in cavity-like structures, e.g., in quantum wells. Thus optical microcavities of half or full wavelength size are obtained. Both suppression and enhancement of spontaneous emission in semiconductor microcavities were demonstrated in experiments by Yamamoto and collaborators. Yablonovitch *et al.* proposed the use of photonic band gap structures in semiconductors. In those systems spontaneous emission is forbidden in certain frequency regions owing to the special material properties.

Energy Shift in Confined Space

In the previous section we focused on changes of radiation rates of atoms near conducting walls or in cavity-like structures. Next we will discuss the more subtle phenomenon of energy shifts. While radiation rates are modified by the influence of the vacuum field on the real emission of photons the energy shifts are caused by the modification of a virtual photon interaction.

Resonant and nonresonant phenomena have to be distinguished. The resonant self-energy shift of a decaying atomic dipole in the vicinity of a conducting wall can be determined from the average polarization energy produced by the image dipole field. For distances comparable to the wavelength the near-field condition is satisfied, resulting in the z^{-3} dependence of the static dipole–dipole interaction characteristic for the van der Waals energy; under far-field conditions the distance dependence is given by z^{-1}. The polarization of the atom by the nonresonant parts of the broadband electromagnetic field causes energy shifts, of which the Lamb shift is the most prominent. In the sense of the nonrelativistic treatment by Bethe the major contribution of that shift can be described as being a result of the emission and reabsorption of virtual photons. It is plausible that just as the real emission of a photon is modified in confined space, so also is the virtual process. The latter 'real' radiation energy shift is thus a consequence of vacuum fluctuations only. It is identical with the energy shift predicted by Casimir and Polder and is analogous to the better-known result of Casimir on the force between two plane neutral conducting plates.

The question of modification of atomic energies in confined space has recently found considerable interest and many calculations of the phenomenon have been performed. Direct application to the energy shift of Rydberg atoms, which are of special interest for experimental studies, was performed by Barton in 1987. He calculated the direct electrostatic interaction with a conducting wall and the radiation induced (retarded) effects. The result is that in the case of two parallel plates the electrostatic effect is dominant when the distance L between the conducting plates is small, $L < n^3 a_0/\alpha$ (n is the principal quantum number of the Rydberg atom, a_0 the Bohr radius, and α the fine-structure constant), and the radiative effect plays the major role when large distances are used, $L > n^3 a_0 \alpha$.

Experiments to measure the van der Waals force of conducting planes have been performed by Hinds and coworkers. They could clearly demonstrate the z^{-3} dependence of the van der Waals force. The retarded effects were not detectable at large distances but at small distances from the wall a deviation from the z^{-3} dependence was found which was attributed to the retarded QED potential.

The energy shift of rubidium Rydberg atoms in confined space has been measured by Marrocco *et al.* in an experiment with extreme high-resolution using the Ramsey experiment two-field method. The atoms are excited in s-Rydberg states ($n \approx 30$) by two-photon transitions using the light of an ultrastable dye laser, the linewidth of which was less than 10 Hz. The laser intensity was enhanced in a folded cavity locked to the laser frequency. Both interaction zones necessary for the Ramsey method were enclosed in this cavity. Between the two interaction regions the atoms pass through a pair of conducting plates, the distance between which can be changed. The shift of the Ramsey interference was measured as a function of the plate distance. Using the Ramsey method has the advantage that the shift can be determined without a direct probing of the atoms in the space between the plates. A level shift on the order of about 150 Hz could be measured, in very good agreement with theory.

Maser Operation

If the rate of atoms crossing a cavity exceeds the cavity damping rate ω/Q, the photon released by each atom is stored long enough to interact with the next atom. Here ω stands for the cavity frequency and Q for the quality factor of the cavity. The atom–field coupling becomes stronger and stronger as the field builds up and evolves to a steady state. Using Rydberg atoms with a large field–atom coupling constant leads to a new kind of maser which operates with exceedingly small numbers of atoms and photons.

The photons corresponding to transitions between neighboring Rydberg atoms are in the microwave region at about 20–100 GHz. Atomic fluxes as small as a few atoms per second have generated maser action, as could be demonstrated by Walther *et al*. of the University of Munich in 1985 using a superconducting cavity. For such a low flux there is never more than a single atom in the resonator – in fact, most of the time the cavity is empty of atoms. It should also be mentioned that in the case of such a setup a single resonant photon is sufficient to saturate the maser transition.

The scheme for this one-atom maser or micromaser is shown in **Figure 1**. The setup represents the simplest system in radiation–atom interaction: a single atom is interacting with a single mode of the radiation field. This device was used to verify the complex dynamics of a single atom in a quantized field predicted by the Jaynes–Cummings model. All of the features are explicitly a consequence of the quantum nature of the electromagnetic field: the statistical and discrete nature of the photon field leads to new characteristic dynamics such as collapse and revivals in the exchange of a photon between the atom and the cavity mode. The frequency of this exchange is usually called the Rabi flopping frequency. The field in the cavity is measured through the number of atoms in the lower state of the maser transition. Due to the strong coupling between the atom and maser field both are entangled. The strong coupling can also be used to entangle subsequent atoms.

The steady-state field of the micromaser shows sub-Poisson statistics. This is in contrast to regular masers and lasers where coherent fields (Poisson statistics) are observed. The reason for nonclassical radiation being produced is that a fixed interaction time of the atoms is chosen, leading to careful control of the atom–field interaction dynamics.

Under steady-state conditions, the photon statistics $P(n)$ of the field of the micromaser are essentially determined by the pump parameter $\Theta = N_{\mathrm{ex}}^{1/2}\Omega t_{\mathrm{int}}$, denoting the angular rotation of the pseudospin vector of the interacting atom. Here, N_{ex} is the average number of atoms that enter the cavity during the decay time of the cavity τ_{cav}, Ω the vacuum Rabi flopping frequency, and t_{int} is the atom–cavity interaction time. The normalized photon number of the maser $\langle \nu \rangle = \langle n \rangle / N_{\mathrm{ex}}$ shows the following generic behavior

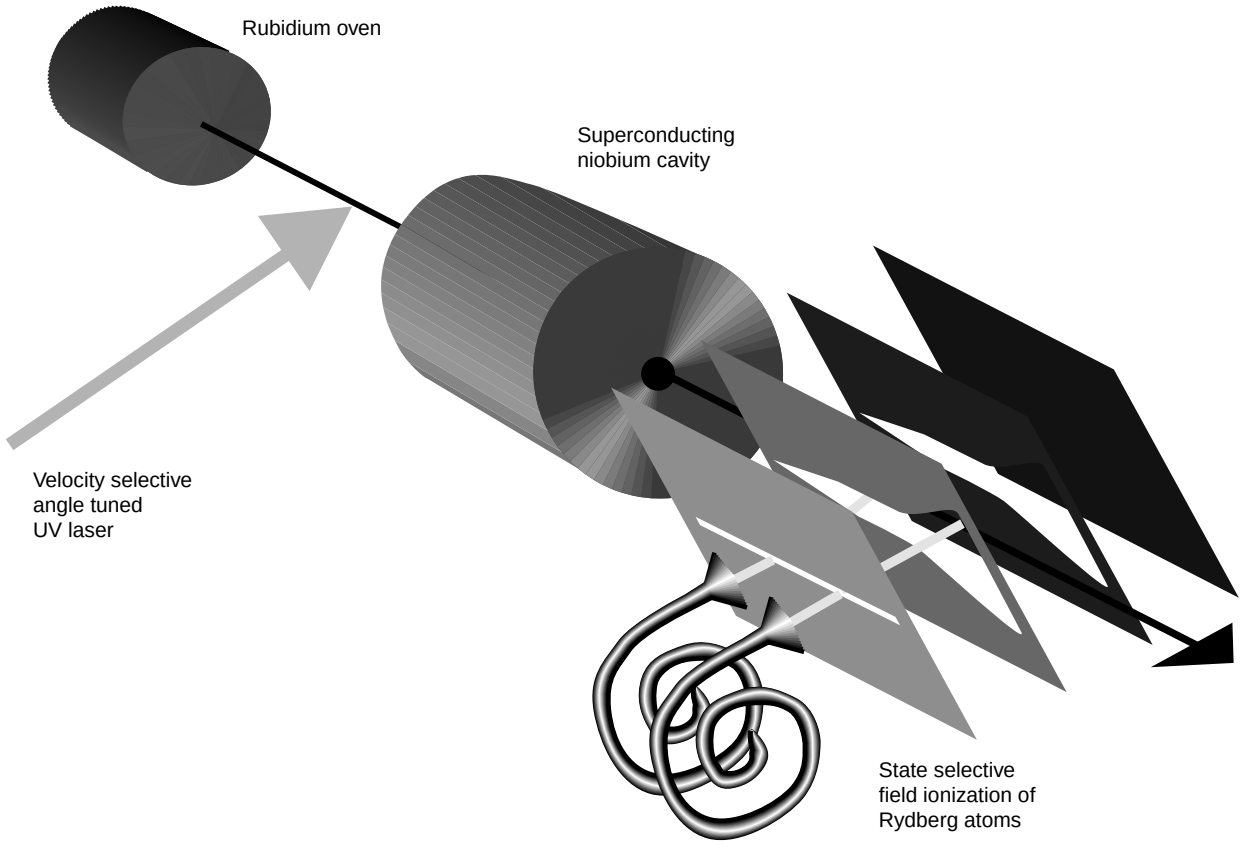


Figure 1 Micromaser setup of rubidium Rydberg atoms in the 63p state. The velocity of the atoms is controlled by exciting a velocity subgroup of atoms in the atomic beam. The atoms in the upper and lower maser levels are selectively detected by field ionization. The number of photons deposited in the cavity is determined through the number of atoms in the lower state (61d).

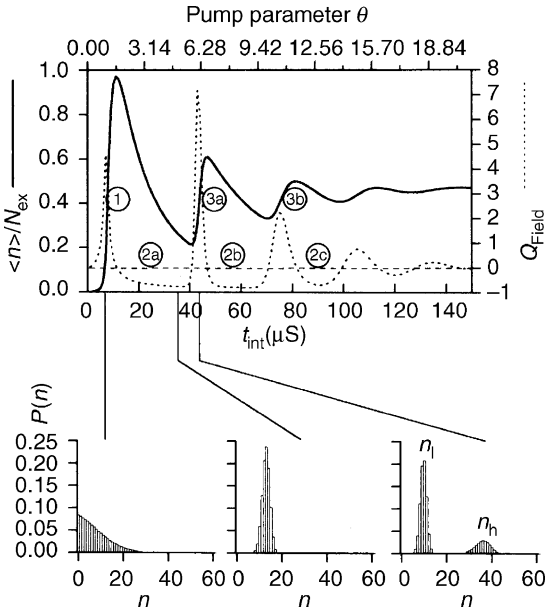


Figure 2 One-atom maser or micromaser characteristics. The upper part of the figure shows the average photon number versus interaction time (solid curve) and the photon number fluctuations represented by the Q-factor (dotted line). Both curves are determined by the photon exchange dynamics between atom and field. The pump parameter gives the angular rotation of the pseudospin vector of the interacting atom (e.g., 2π corresponds to a full rotation, i.e., the atom is again in the upper level). In the lower part of the figure the steady-state photon number distribution $P(n)$ is shown for three values of t_{int}. The distribution on the left side corresponds to the maser threshold, that on the right gives an example of the double-peaked distribution associated with the quantum jump behavior. In this situation, the atom is back in the excited state and can again emit, leading to a higher steady-state photon number n_{h}. With increasing t_{int} this part will grow and n_{l} will decrease and disappear. A new jump occurs in the region 3b.

(see Figure 2). It suddenly increases at the maser threshold value $\Theta = 1$ and reaches a maximum for $\Theta \approx 2$ (denoted by 1 in Figure 2). The behavior at threshold corresponds to the characteristics of a continuous phase transition. As Θ increases further, the normalized averaged photon number $\langle \nu \rangle / N_{\text{ex}}$ decreases to a minimum slightly below $\Theta \approx 2\pi$, and then abruptly increases to a second maximum (3a in Figure 2). This general type of behavior recurs roughly at integer multiples of 2π, but becomes less pronounced with increasing Θ. The reason for the periodic maxima of the average photon number is that for integer multiples of $\Theta = 2\pi$ the pump atoms perform an integer number of full Rabi flopping cycles, and start to flip over at a slightly larger value of Θ, thus leading to enhanced photon emission. The periodic maxima for $\Theta = 2\pi, 4\pi$, and so on may be interpreted as first-order phase transitions.

The photon statistics of the maser radiation is usually characterized by the Q-parameter introduced by Mandel:

$$Q_{\text{field}} = \left\{ \left(\langle n^2 \rangle - \langle n \rangle^2 \right) / \langle n \rangle \right\} - 1$$

It can be seen that $Q_{\text{field}} = 0$ corresponds to a Poissonian photon distribution. Q_{field} for the micromaser is plotted as a dotted line in Figure 2. The value drops below zero in the region 2a, 2b, etc. This shows the highly sub-Poissonian character of the one-atom-maser field being present over a large range of parameters.

The reason for the sub-Poissonian atomic statistics is the following. A changing flux of atoms changes the Rabi frequency via the stored photon number in the cavity. Adjusting the interaction time allows the phase of the Rabi nutation cycle to be chosen such that the probability of the atoms leaving the cavity in the upper maser level increases when the flux, and hence the photon number in the cavity, is raised. This leads to sub-Poissonian atomic statistics since the number of atoms in the lower state decreases with increasing flux and photon number in the cavity. This feedback mechanism can be neatly demonstrated when the anticorrelation of atoms leaving the cavity in the lower state is investigated. Measurements of this anticorrelation phenomenon could be made.

The fact that anticorrelation is observed shows that the atoms in the lower state are more equally spaced than expected for a Poissonian distribution of the atoms in the beam. This means, for example, that when two atoms enter the cavity close to each other, the second one performs a transition to the lower state with reduced probability.

The interaction with the cavity field thus leads to an atomic beam with atoms in the lower maser level showing number fluctuations which are up to 40% below those of a Poissonian distribution found in usual atomic beams. This is interesting because atoms in the lower level have emitted a photon to compensate for cavity losses inevitably present. Although this process is induced by dissipation giving rise to fluctuations, the atoms still obey sub-Poissonian statistics.

Generation of Number States (Fock States) of the Radiation Field

When the micromaser is operated at low temperatures, a very interesting feature is observed in the inversion of the population of the two maser levels called trapping states. They are a steady-state feature of the maser field; they occur in the micromaser as a direct consequence of field quantization in a cavity. At low cavity temperatures the number of black-body

photons, n_{th}, in the cavity mode is reduced and trapping states begin to appear; at higher temperatures they are washed out by the thermal photons. The trapping states show up when the atom–field coupling Ω and the interaction time t_{int} are chosen such that in a cavity field with n_q photons the atoms undergo an integer number k of Rabi cycles. This is summarized by the condition

$$\Omega t_{int}\sqrt{n_q + 1} = k\pi \qquad [1]$$

for the trapping state, denoted by (n_q, k). When eqn [1] is satisfied, the cavity photon number is left unchanged after interaction of an atom and hence the photon number is 'trapped'. This will occur regardless of the atomic pump rate N_{ex}. The trapping state is therefore characterized by the upper-bound photon number n_q and the number of integer multiples of full Rabi cycles k. In this situation the field is stabilized. Whenever a photon disappears, e.g., due to dissipation, the next incoming atom experiences a changed Rabi nutation frequency and emits a photon with high probability. At the trapping condition a quantum nondemolition situation is present. Through the dynamics of the Rabi nutation the field is measured, without any net change of the field.

The build-up of the cavity field can be seen in Figure 3, where the emerging atom inversion is plotted against the interaction time and pump rate. At low atomic pump rates (low N_{ex}) the maser field cannot build up and the maser exhibits Rabi oscillations due to the interaction with the vacuum field. At the positions of the trapping states, the field

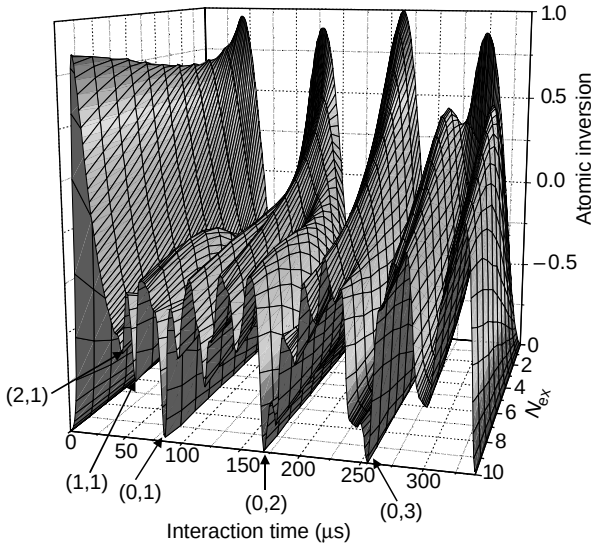


Figure 3 Behavior of the micromaser at low temperature showing the trapping states, manifested as valleys along the N_{ex} axis. For the designation of the trapping states see text.

increases until it reaches the trapping-state condition. This is manifested as a reduced emission probability and hence as a dip in the atomic inversion. Once in a trapping state, the maser will remain there regardless of the pump rate.

Under ideal conditions the micromaser field in a trapping state is a Fock state, but when the micromaser is operated in a continuous wave mode, the field state is very fragile and highly sensitive to external influences and experimental parameters. However, in contrast to continuous-wave operation, in pulsed operation trapping states are more stable and more practical and can be used over a much broader parameter range than in continuous-wave operation.

To demonstrate the principle, Figure 4 shows a simulation of a sequence of 20 pulses of the pumping atoms in which an average of seven excited atoms per pulse are present. Under the trapping condition only a single emission event occurs, producing a single lower state atom which leaves a single photon in the cavity. Since the atom–cavity system is in the trapping condition, the emission probability is reduced to zero and the photon number is stabilized. Consequently, excited state atoms following the emitting atom stay in the upper maser level. The variation of the time when an emission event occurs during the atom pulses in Figure 4 is due to the Poissonian spacing of upper-state atoms entering the cavity and the stochasticity of the quantum process.

The lower part of Figure 4 shows the photon number distribution resulting from this process. It was demonstrated experimentally by Brattke *et al.* that a single photon number state could be generated with a success rate of 85%. By improving the experimental parameters one can expect to prepare single-photon Fock states in 98% of the pulses by this method.

Other Cavity Experiments

Besides the micromaser, there are also other experiments using Rydberg atoms in cavities and the strong interaction of these atoms with cavities which lead to interesting applications. One example is an experiment by S. Haroche, J. M. Raimond *et al.*, who succeeded in realizing a Schrödinger cat state in a cavity. Studies on the decoherence of this state could be conducted. The experiments are quite important in exploring the boundaries between the quantum and classical worlds. Another example is the quantum nondemolition detection of a single photon in a cavity by S. Haroche employing the dispersion level shift of probing atoms.

There is an interesting equivalence between an atom interacting with a single-mode field and a quantum

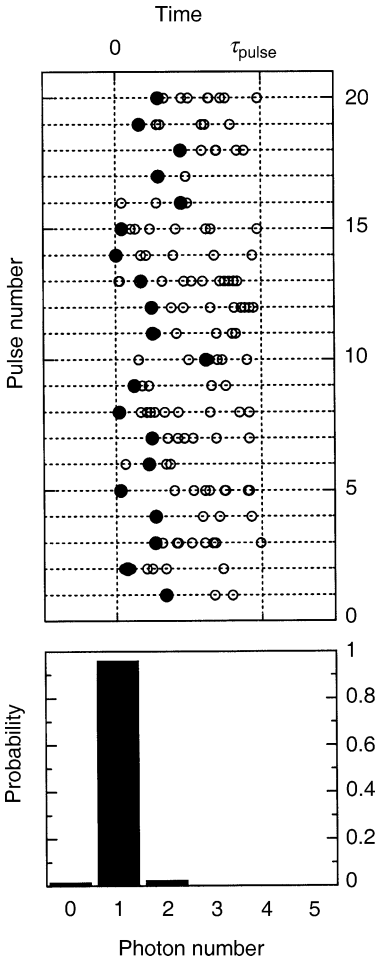


Figure 4 A simulation of a subset of 20 subsequent atom bunches generated by pulsed laser excitation with the associated probability distribution for photons in the cavity or lower-state atom production (filled circles represent lower-state atoms and open circles represent excited-state atoms). The start and end of each pulse is indicated by the vertical dotted lines marked 0 and τ_{pulse}, respectively. For details, see Battke S, Varcoe BTH and Walther H (2001) Generation of photon number states on demand via cavity, quantum electrodynamics. *Physical Review Letters* 86: 3534–3537.

particle in a harmonic potential, as was first pointed out by D. F. Walls and H. Risken; this connection results from the fact that the radiation field is quantized on the basis of the harmonic oscillator. The Jaynes–Cummings dynamics can therefore also be observed with trapped ions, as recently demonstrated in a series of beautiful experiments by D. Wineland *et al*. They also produced a Schrödinger cat state by preparing a single trapped ion in a superposition of two spatially separated wavepackets which are formed by coupling different vibrational quantum states in the excitation process.

Besides the experiments in the microwave region, a single-atom laser emitting in the visible range has also been realized by M. Feld *et al*. Furthermore, cavity quantum electrodynamic effects have been studied in the optical spectral region by J. L. Kimble *et al*.

Today's technology in microfabrication of semiconductor diode structures allows the realization of low-order cavity structures for diode lasers. In these systems the spontaneous emission is controlled in the same way as in the micromaser. Since spontaneous decay is a source of strong losses, control of this phenomenon leads to highly efficient laser systems. Quantum control of spontaneous decay thus has important consequences for technical applications. This topic will be briefly described in the next section.

Microlasers

The simplest approach to fabricating an optical microcavity is to shrink the spacing between the mirrors of a Fabry–Perot resonator to λ/n (where n stands for the refractive index) while reducing the lateral dimensions to a range of the same order of magnitude. This structure provides a single dominant longitudinal field mode that radiates into a narrow range of angles around the cavity axis.

The first optical microcavity experiments used dye molecules between high-reflectivity dielectric mirrors in the Fabry–Perot configuration. Because spontaneous emission is a major source of energy loss, speed limitations, and noise in lasers, the capability to control spontaneous emission is expected to improve laser performance. If the fraction of spontaneous emission coupled into the lasing mode is made close to one, the 'thresholdless' laser is obtained, in which the light output increases almost linearly with the pump power instead of exhibiting a sharp turn-on at the pump threshold.

Semiconductor microcavities provide high-Q Fabry–Perot cavities for both basic studies and potential applications. Molecular beam epitaxy or organometallic chemical vapor deposition techniques are used to deposit high-reflectivity mirrors consisting of alternating quarter-wavelength layers of lattice-matched semiconductors. For example, 15–20 pairs of quarter-wave layers of $Al_{0.2}Ga_{0.8}As$ and AlAs result in a reflectivity greater than 99% and Q values greater than 500. The optically active layer in such a microcavity is typically a GaAs quantum well located at the midplane of the cavity, where the field strength is maximum.

Since the microcavities have an extremely low threshold, their efficiency will also be high. Low-cost, high-density light source arrays and photonic circuits are made possible by the small size and low power consumption of such resonators. It will be possible to produce entire wafers containing millions

of microlasers with a multipole arrangement instead of having to cleave each individual semiconductor laser as at present. This improvement will lead to higher yields and lower cost per element. Another advantage of surface-emitting microcavity sources is the efficient optical coupling of their stable symmetric mode patterns into optical fibers or waveguides. These lasers are thus certain to spark a revolution in optical communication in the future.

Conclusion

This article reviewed experiments in cavity quantum electrodynamics. The experiments shed new light on our understanding of the radiation interaction of atoms. The achievable strong coupling between atoms and radiation leads to the possibility to generate entanglement between the generated radiation field and the atoms, presenting a basis for interesting applications in connection with, e.g., quantum computing and quantum information processing. Furthermore Fock states of the radiation field can be generated. The control of spontaneous decay finally leads to interesting new laser systems offering high efficiency.

See also

Quantum Electrodynamics: Cavity QED in Semiconductors; Quantum Theory of the Electromagnetic Field. **Quantum Optics:** Quantum Computing with Atoms.

Further Reading

Berman PR (1994) Cavity quantum electrodynamics. In: Bederson B and Bates D (eds). *Advances in Atomic, Molecular, and Optical Physics*, Supplement 2. Boston, MA: Academic Press.

Englert B-G, Löffler M, Benson O, *et al.* (1998) Entangled atoms in micromaser physics. *Fortschritte der Physik* 46: 897–926.

Haroche S (1992) Cavity quantum electrodynamics. In: Dalibard J, Raimond J-M and Zinn-Justin J (eds) *Fundamental Systems in Quantum Optics*. pp. 767–935. Amsterdam: North-Holland.

Haroche S and Kleppner D (1989) Cavity quantum electrodynamics. *Physics Today* 42: 24–30.

Kimble HJ, Carnal O, Georgiades N, *et al.* (1995) Quantum optics with strong coupling. In: Wineland DJ, Wieman CE and Smith SJ (eds) *Atomic Physics*, vol. 14, pp. 314–335. New York: American Institute of Physics.

Marrocco M, Weidinger M, Sang RT and Walther H (1998) Quantum electrodynamic shifts of Rydberg energy levels between parallel metal plates. *Physical Review Letters* 81: 5784–5787.

Meschede D (1992) Radiating atoms in confined space: from spontaneous emission to micromasers. *Physics Reports* 211: 201–250.

Meystre P (1992) Cavity quantum optics and the quantum measurement process. In: Wolf E (ed.) *Progress in Optics*, vol. 30. Amsterdam: North-Holland.

Walther H (1992) Experiments on cavity quantum electrodynamics. *Physics Reports* 219: 263–281.

Cavity QED in Semiconductors

M Kira, W Hoyer and S W Koch,
Philipps-University, Marburg, Germany

G Khitrova and H M Gibbs,
University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

According to classical electromagnetism, propagating light fields obey Maxwell's equations under which electric and magnetic fields exchange their strength in an oscillatory manner. If a propagating field is followed at a single reference point, it is convenient to represent the electric field as a complex number $|E(t)|\exp(-i\omega t)$ where $|E(t)|$ is the magnitude of the field and the phase of the field $\exp(-i\omega t)$ oscillates in time t with the optical frequency ω. Figure 1 illustrates how the classical field is determined by one single point in the complex plane; consequently, the phase and amplitude of the field are known exactly. According to quantum mechanics, this simple picture has to be altered because the Heisenberg uncertainty relation states that complementary quantities like phase and amplitude cannot be determined precisely at the same time. As a result, the real and imaginary parts of the field can be determined only with accuracy $\Delta\mathrm{Re}[E]$ and $\Delta\mathrm{Im}[E]$, respectively, and the Heisenberg uncertainty principle determines the best possible accuracy $\Delta\mathrm{Re}[E]\,\Delta\mathrm{Im}[E] \geq 1$. To incorporate the fundamental inaccuracy, the electric field has to be defined by using complex-valued distributions or equivalently wavefunctions as shown in Figure 1. This quantization procedure leads to the Schrödinger equation for light analogous to that for particles. The field called quantum optics investigates the quantum electrodynamics (QED) features of light.

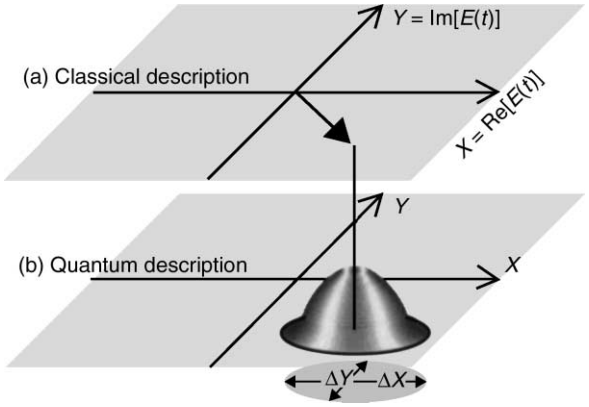


Figure 1 (a) The classically described light field $E(t)$ is a single point in the complex XY-plane; the corresponding (X, Y) vector has a definite phase and angle. (b) Quantum mechanically described light is a distribution with fluctuations ΔX and ΔY described by the shaded circle.

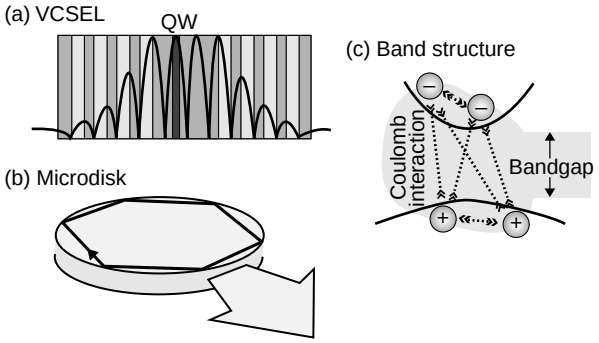


Figure 2 (a) A typical vertical-cavity surface-emitting laser (VCSEL) structure; a stack of quarter-wavelength layers with different refractive index provides a mirror such that the light intensity (solid line) is strongly confined inside the cavity. A quantum well (QW) is positioned at the field maximum. (b) In a microdisk laser, light encounters ideal total reflection when the edge of the disk is reached. The so-called whispering gallery mode emits (leaks) light from the edge. (c) A schematic illustration of semiconductor band structure, bandgap, and Coulomb interaction for an electro–hole system.

In general, light does not propagate freely in space, but it interacts strongly with the surrounding matter. For example, light can be absorbed, and its energy can be converted into excitation of the matter. As a result, the light may be slowed down, reflected, scattered, or diffracted. In order to explain the implications of the light–matter interaction in detail, one obviously has to apply a quantum mechanical description also for the matter. This approach determines the so-called eigenstates of the matter, so that the matter may occupy only certain discrete states with discrete energies. The Rydberg series of a hydrogen atom is a typical example. Similar discrete energy levels are also found for the quantized light; these levels are often referred to as photons which heuristically describe the particle properties of the light. The lowest-order light–matter interaction consists of processes where one photon is absorbed (emitted) while the matter is simultaneously excited (de-excited) from one eigenstate to another. This generic type of interaction leads to a microscopic description of the optical effects mentioned above. The magnitude of these effects depends on the strength of the interaction. By implementing different cavity configurations, the reflected light can be forced to travel across the same matter many times. As a result, the light–matter interaction is enhanced by a factor proportional to the multiple passes of the light. Thus, a cavity can be efficiently used to amplify optical phenomena. Figure 2 shows cavity configurations commonly used for semiconductors. Semiconductor cavity QED investigates quantum optical effects in semiconductor systems by enhancing them with cavity mirrors. In general, quantum optical features produce classically unexpected effects which typically stem from: (a) the discrete nature of eigenstates of light and matter; (b) the superposition principle of quantum mechanics; and (c) the Heisenberg uncertainty principle. In this article, we review both theoretical and experimental advances made so far to predict, observe, and control quantum optical effects in semiconductors with respect to effects (a)–(c).

During the last few decades, atomic quantum optics has already developed toward QED investigations, and semiconductor optics is developing rapidly into the same regime. Advanced atomic QED theories have successfully explained and guided intriguing experimental developments like laser cooling, atom condensation, and photon teleportation. However, the atomic approach can mostly be applied to describe dilute and only weakly interacting atomic gases since relatively simple models of few-level systems are used to describe the material. The elementary electronic excitations in semiconductors consist of electrons and holes (missing electrons) lifted into the conduction and valence bands, respectively. The corresponding eigenstates form continuous energy bands with the band-gap energy separation indicated in Figure 2. Since electrons and holes have opposite charges, they experience Coulomb attraction whereas bare electrons or holes repel each other. Under favorable conditions, the attractive Coulomb interaction prevails and atom-like bound electron–hole pairs (excitons) may be formed. However, since the Coulomb force has an infinite range and an electron–hole system is typically dense, excitons cannot be treated as weakly interacting quasiparticles. As a result, the atomic QED approaches cannot be used to describe quantum optics of semiconductors

in general. Thus, this article concentrates mostly on semiconductor QED theory which fully includes the interaction of charge carriers, i.e., electrons and holes, as well as quantum features of light.

The development work on quantum devices and processes is vital for advancements in several key technologies such as computers and telecommunications. The continuous decrease in component dimensions leads to a microscopic structure size (less than 1 μm); at the same time, increase in the device performance is accompanied by ultrafast operation time (faster than 1 ps). Due to these development trends, the properties of components and processes are becoming more and more quantum mechanical. Although the description of the quantum processes is complicated, the microscopic behavior offers entirely new operational functionality such as massively parallel quantum computers. These possibilities are based on the controlled manipulation of the quantum mechanical state of the light and matter. Due to rapid recent developments, optically coupled semiconductor devices have a great potential to become technologically successful and commercially viable quantum components.

Quantized States: Observation via Strong Light–Matter Interaction

In order to enhance the light–matter interaction, we choose a system where the semiconductor is placed inside a cavity in the position where the field intensity is maximum. For an empty cavity, the field intensity has a strong dependence upon frequency with a strong resonance at ω_R when the cavity length is an integer number of light wavelengths in the material. The width of the resonance is directly related to the quality of the cavity as determined by the number of back-and-forth reflections of light inside the cavity. If one chooses a vertical-cavity surface-emitting laser (VCSEL) structure, the most efficient coupling is obtained by placing a thin planar semiconductor structure at the field maximum as shown in Figure 2. If the semiconductor is planar and thin, the structure is called a quantum well because electrons and holes are confined in one direction in analogy to the standard particle-in-a-box problem of fundamental quantum mechanics. As a result, the energy levels of a quantum well are discrete in the confinement direction. If the structure is narrow enough, the system dynamics is confined to the lowest quantum-well level such that the carriers are quasi-two-dimensional due to the free in-plane motion. The optical response of such a system to a weak classical probe beam has been successfully analyzed with the so-called semiconductor Bloch equations. For the quantum-well system alone, the absorption spectrum may have a sharp resonance; this is often referred to as an excitonic resonance since it is located below the fundamental bandgap energy at a position corresponding to the exciton binding energy. Obviously, the light–matter coupling becomes large when the exciton and cavity resonances coincide. However, since these resonances are coupled, the optical response is altered; we observe a splitting into two absorption peaks instead of the original degenerate resonances. This phenomenon is commonly referred to as normal-mode coupling; in general, it is a quite common feature in quantum mechanics that degenerate states split due to an additional interaction. Figure 3 shows the first experimental observation of normal-mode coupling in semiconductors.

Since a semiconductor is a strongly interacting many-body system, several effects alter the specific character of an excitonic resonance when electrons and holes are excited. The attractive interaction between electrons and holes becomes weaker for

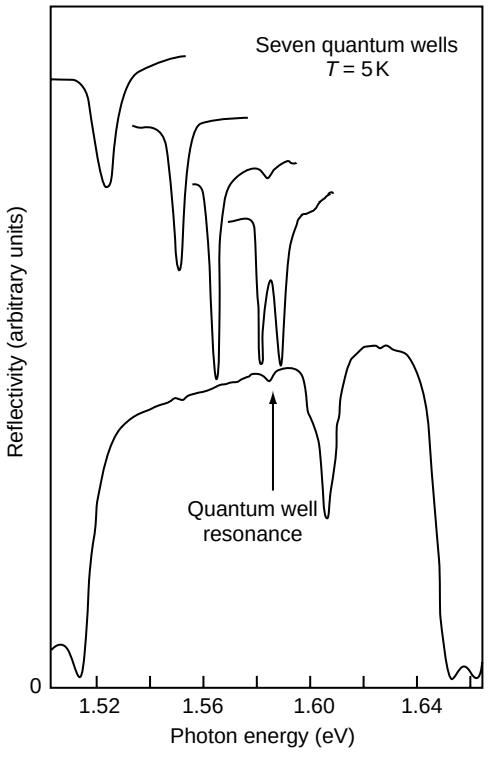


Figure 3 Probe reflectivity of a seven-quantum-well microcavity structure. The various curves correspond to the energy difference between exciton and cavity resonances. When these resonances become degenerate, the reflectivity shows two resonances corresponding to the normal-mode splitting. From Weisbuch C, Nishioka M, Ishikawa A and Arakawa Y (1992) Observation of the coupled exciton-photon mode splitting in a semiconductor quantum microcavity. *Physical Review Letters* 69: 3314. Copyright (2004) by the American Physical Society.

increased carrier density since the surrounding charges screen the bare Coulomb interaction. A particular exciton may also experience scattering from nearby electrons, holes, or other excitons via the Coulomb interaction. Furthermore, electrons and holes are Fermions, which requires that any given carrier state can only be occupied once. This poses fundamental limits on how many carriers can coexist in a specific volume; above a certain limit, additional occupation is prevented, and so-called Pauli blocking is observed. Since an exciton consists fundamentally of Fermionic constituents, eventually Pauli blocking effects become important even for excitons. Due to a combination of these many-body effects, the excitonic resonance is weakened by an increasing carrier density. Figure 4 shows a comparison between theory and experiment of the excitonic absorption spectrum for different carrier densities. For elevated densities, the exciton resonance is broadened and its height reduced. When the same investigations are repeated in a microcavity, we observe that the normal-mode peaks decrease in height but their separation is roughly unchanged for moderate densities. In general, the normal-mode spitting is proportional to the oscillator strength of the excitonic absorption, i.e., the area under the resonance. Thus, the observed constant splitting suggests an unchanged oscillator strength for moderate densities. The invariant oscillator strength was unexpected and can only be explained by using a theory which includes the Coulomb interaction of carriers microscopically. When the carrier density is increased even further, the exciton resonance is completely bleached, and the microcavity transmission has only a single peak, at the cavity mode energy; this is commonly referred to as the weak-coupling regime, in contrast to the nonperturbative strong coupling with two transmission peaks. The strong coupling regime provides several intriguing phenomena; for example, parametric amplification of emission has been demonstrated by applying simultaneously multiple light beams to the sample.

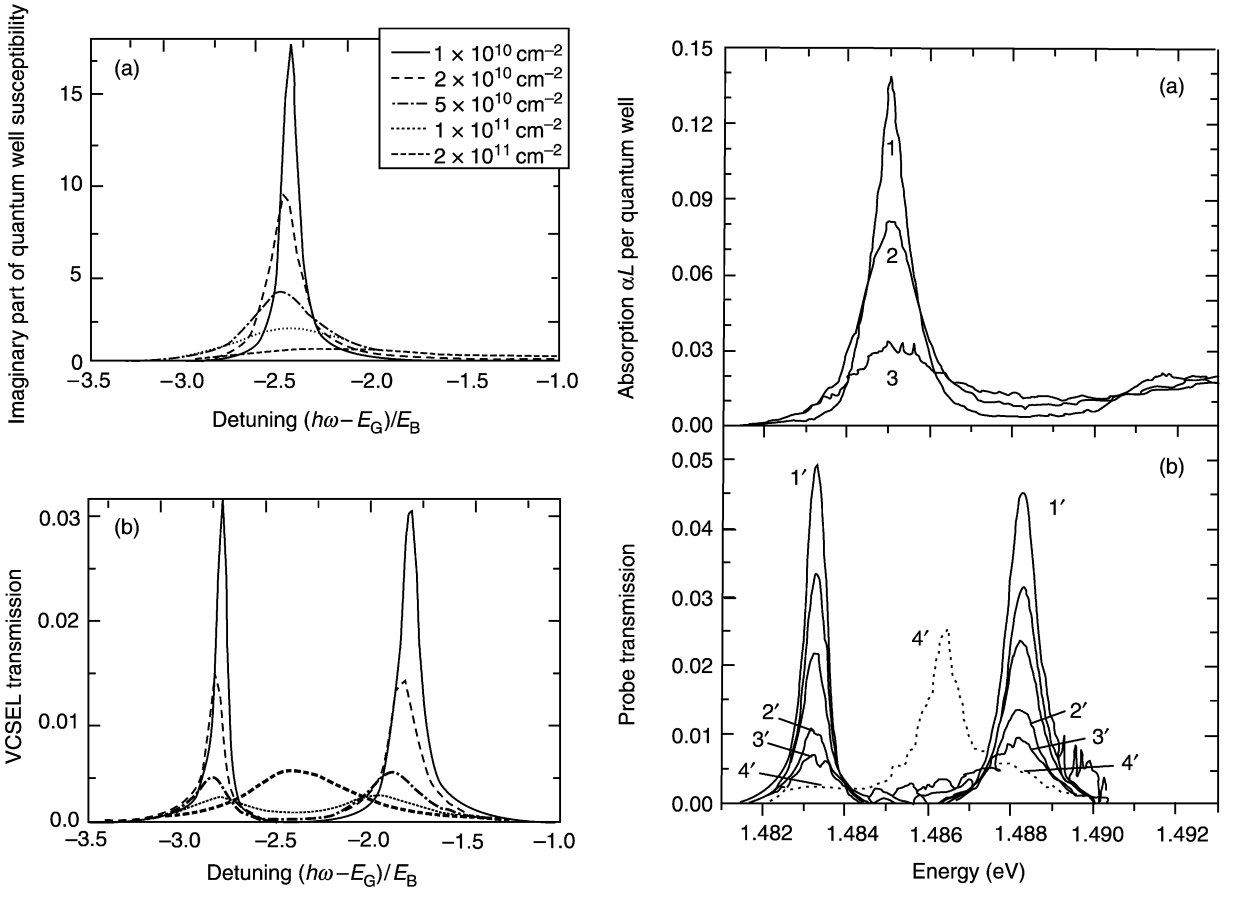


Figure 4 Left column: (a) microscopically computed bare quantum-well absorption spectrum, i.e., the imaginary part of the susceptibility, as a function of carrier density. (b) Calculated transmission of the quantum-well microcavity for corresponding densities. The right column shows experimentally determined (a) absorption and (b) microcavity transmission for the same conditions as in the calculations. From Jahnke F, Kira M, Koch SW, *et al.* (1996) Excitonic nonlinearities of semiconductor microcavities in the nonperturbative regime. *Physical Review Letters* 77: 5257. Copyright (2004) by the American Physical Society.

The ultimate cavity QED limit of normal-mode coupling follows when the light–matter interaction is so strong that even a single photon leads to splitting (genuine strong coupling). We briefly outline some aspects of the quantum statistical limit known from atomic physics. This situation can be analyzed with the so-called Jaynes–Cummings model where only two states of the atom are included together with a single light mode. In this case, the interaction between a single photon and an atom leads to a normal-mode splitting g_0. If two photons interact with one atom, the splitting increases to $2^{1/2}g_0$; more generally coupling between n photons and an atom leads to a splitting of $n^{1/2}g_0$. In the QED limit, the matter response is very nonlinear because one can detect the quantized nature of light directly from the energy splitting. This limit has already been reached in atomic cavity QED, whereas this regime is hard to approach with semiconductors because they behave more like multi-atom systems. If one has N atoms in a cavity, the first photon excites one atom, but there are N different ways of doing this, so the splitting is $N^{1/2}g_0$. If a second photon arrives, it interacts with the remaining unexcited atoms rather than the excited one, so that the splitting remains $N^{1/2}g_0$ as long as the number of photons is much smaller than N. Thus, in an N-atom system, or equivalently in semiconductors, the quantized light effects are much harder to observe than for a single atom. Currently, the normal-mode splitting in a semiconductor microcavity is explainable by the classical features of light, so that only the quantized nature of the matter is observed. In order to approach the ultimate cavity QED limit, one obviously has to reduce N and increase the light–matter coupling. In the future, this objective might be possible in quantum-dot/nanocavity systems where the semiconductor is confined in all spatial directions.

When an excited semiconductor is not under any influence of external classical light fields, it can still emit light via spontaneous emission resulting from the recombination of electron–hole pairs. The resulting light emission is called photoluminescence. Since the emission process takes place in a strongly interacting many-body system, one has to systematically include the Coulomb interaction and Fermionic features. The emission properties can be consistently described by the so-called semiconductor luminescence equations. When the microcavity photoluminescence spectrum is investigated, one observes a similar normal-mode splitting as for the transmission studies. However, this normal-mode coupling is not in the ultimate cavity QED limit even though the quantum fluctuations of the light field trigger the spontaneous emission. Nevertheless,

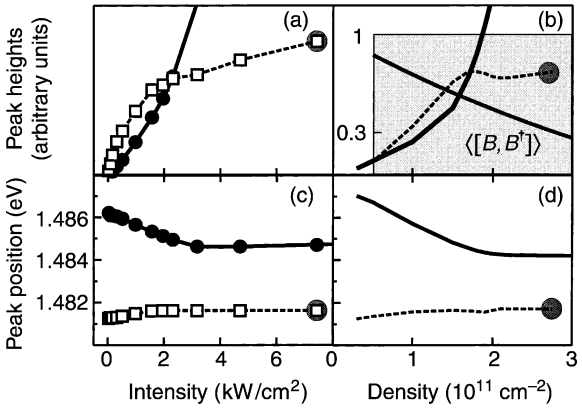


Figure 5 Measured microcavity luminescence intensities (a) and peak energies (c) versus excitation intensity for the high-energy peak (solid line) and the low-energy peak (dashed line). The results of the microscopic theory are shown in (b) and (d). From Kira M, Jahnke F, Koch SW, *et al.* (1997) Quantum theory of nonlinear semiconductor microcavity luminescence explaining "boser" experiments. *Physical Review Letters* 79: 5170. Copyright (2004) by the American Physical Society.

the luminescence shows an interesting new transition as a function of the excitation level. Figure 5 contains a comparison between theory and experiment of normal-mode peak heights and positions as a function of carrier density. For low densities, the high-energy peak is lower in height but it overtakes the low-energy peak for moderate carrier densities that are still below lasing threshold. This threshold-like overtaking was attributed to Boser action, involving exciton formation, final-state stimulation, and Bose condensation. The inset shows an estimate of just how Bosonic the excitons are, i.e., how much they actually behave as independent atoms; the value unity corresponds to the fully Bosonic situation. The nonlinear luminescence transition takes place at a density regime where the underlying Fermionic electron and hole contributions become important (commutator is roughly 0.5). A more detailed analysis with the semiconductor luminescence equations shows that Fermionic carrier nonlinearities in a detuned cavity are responsible for the experimental observations. This example of a misinterpretation based on a Bosonic analysis stresses how important it is to include the Coulomb interaction and Fermion character of carriers when analyzing the properties of semiconductors.

Superposition Principle: Observation of Quantum Optical Correlations

The above normal-mode-coupling investigations have shown interesting nonlinear effects, but they also revealed that the ultimate QED limit – demonstrating

discrete states of light directly – has not yet been reached. Alternatively, a QED investigation may concentrate on other fundamental features of quantum mechanics. One possibility is to study the implications of the superposition principle which states that the joint properties of a light–matter system can always be expressed as a superposition that combines light and matter wavefunctions. The most dramatic consequence can be observed in so-called entangled wavefunctions which cannot be factorized into light and matter parts. In a factorized wavefunction, the light and matter parts are independent whereas entangled wavefunctions conditionally connect light and matter wavefunctions. To elaborate the subtle details of entanglement, we first analyze a simple example. Assume that light can be in two different polarization states, $|\sigma^+\rangle$ or $|\sigma^-\rangle$, and the matter part is either excited $|\text{up}\rangle$ or de-excited $|\text{down}\rangle$. A wavefunction, $[|\text{up}\rangle + |\text{down}\rangle]$ $[|\sigma^+\rangle + |\sigma^-\rangle]$, is clearly a superposition of the fundamental states, and at the same time the light and matter parts are completely factorized. However, the wavefunction $[|\text{up}\rangle\ |\sigma^+\rangle + |\text{down}\rangle\,|\sigma^-\rangle]$ is entangled, since the light and matter parts cannot be separated. In the entangled state, measurement on the light state will conditionally determine a definite state of the matter, whereas the factorized state has no such conditionality. The entanglement has far-reaching consequences which cannot be understood with a classical analysis. For example, the principles of teleportation and quantum computing follow directly from the controlled manipulation of different parts of the entangled wavefunction. For atomic systems, wavefunction entanglement has been demonstrated and utilized in several experiments. For semiconductors, the direct manipulation and detection of the wavefunction seems difficult due to the overwhelming complexity of the many-body wavefunctions. Once again, the entanglement and the wavefunction are simplest for low-dimensional structures; direct entanglement effects have been demonstrated recently in a single quantum dot and between a pair of dots. Also for more complicated semiconductor systems, such as a quantum well in a microcavity, the entanglement can be observed as correlations between light and matter. In this case, the existence of QED correlations basically means that light and matter properties depend conditionally on each other. In general, the direct entanglement and QED correlation investigations have a large development potential for semiconductors.

When a semiconductor is excited with an external light pulse, entanglement-related correlations couple the classical and luminescence emission dynamics. The resulting dynamic interplay between the semiconductor Bloch and luminescence equations is mainly mediated by the correlation between photons and electron–hole densities. In the following, we analyze a microcavity configuration where a strong pump pulse generates large QED correlations. The effect of the correlations is then measured by the response of a weak probe beam. Figure 6 shows a comparison of theory and experiment of probe reflection in a configuration where the pump and probe do not have any spectral overlap. The measured probe reflection displays long-living oscillations as a function of time delay, i.e., phase difference, between the pump and the probe. Only by including the QED correlation in the theory can the oscillatory probe reflection be explained. When the QED contributions are omitted from the theory, the phase difference does not have any effect on the probe reflection. Thus, this experiment–theory example represents a direct observation of the cavity QED effects in semiconductors. Figure 6 also shows more pronounced oscillations when two phase-locked pump pulses provide the excitation; again the full QED theory explains the enhanced oscillation features.

The QED features can also alter the normal-mode coupling characteristic of a weak probe beam. Figure 7 shows a comparison of theory and experiment for a situation where the pump spectrum is

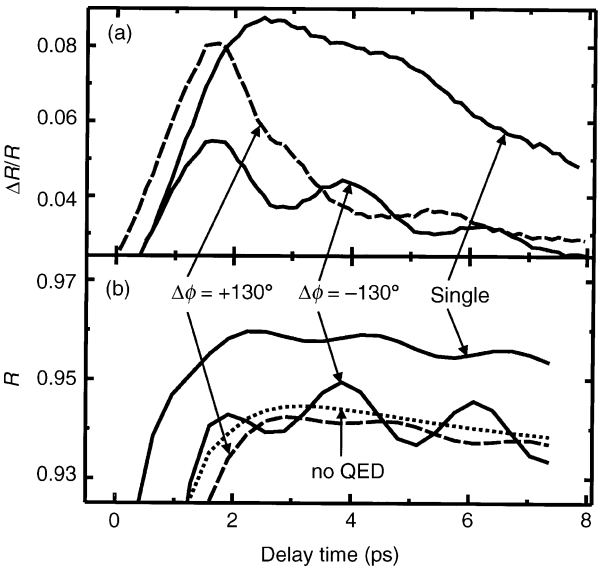


Figure 6 (a) Measured differential reflectivity and (b) computed reflection probability of the probe pulse as function of probe delay with respect to the excitation pulse. For two-pulse excitations, ±130° relative phase shifts are used. The dotted line is computed without the QED correlations. From Lee Y-S, Norris TB, Kira M, *et al.* (1999) Quantum correlations and intraband coherences in semiconductor cavity QED. *Physical Review Letters* 83: 5338. Copyright (2004) by the American Physical Society.

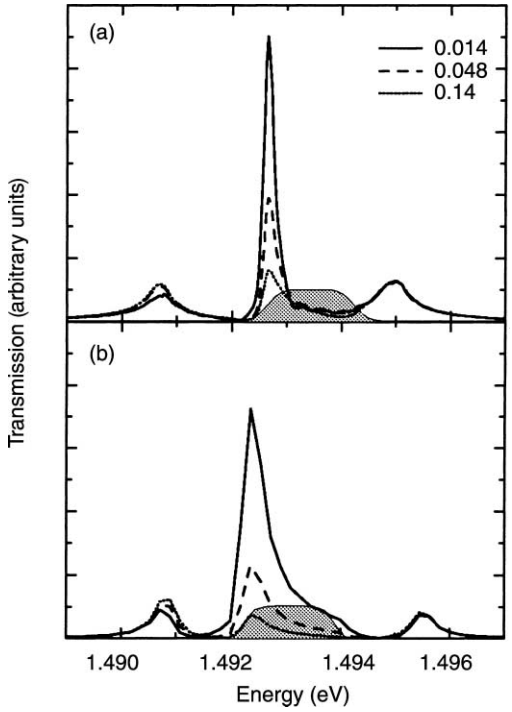


Figure 7 Dependence of probe transmission on probe intensity (nJ cm^{-2}) at a constant pump intensity of 330 nJ cm^{-2} for the microcavity. (a) Experiment, and (b) theory. The shaded regions indicate the pump spectrum. From Ell C, Brick P, Hübner M, *et al.* (2000) Quantum correlations in the nonperturbative regime of semiconductor microcavities. *Physical Review Letters* 85: 5392. Copyright (2004) by the American Physical Society.

located between the two normal-mode coupling resonances. Both the theory and experiment show a new third resonance which follows the energetic position of the pump. The third peak is a direct consequence of the QED correlations. Since the third peak is generated from the QED field–carrier correlations generated by the pump pulse, its magnitude follows the intensity of the pump. As the probe intensity is decreased the relative effect of the QED correlations is increased so that the third peak grows in the probe transmission.

In these QED investigations, the cavity plays an important role since the mirror feedback leads to a significant amplification of the QED effects through the enhanced light–matter coupling. From a practical point of view, normal-mode coupling also provides well-separated spectral features which can be used as classical-emission reference points.

Heisenberg Uncertainty Principle: Squeezing of Light Emission

Quantized light effects can also be investigated by measuring the light fluctuations of the field directly. To maximize the quantum effects, we analyze the field fluctuations in the emission directions where the classical field vanishes. This situation can be realized in planar quantum-well structures which are nearly free of disorder. In such systems, a light pulse propagating perpendicular to the structure leads to transmission and reflection of classical light only along the excitation axis. If the detection is performed at an angle, the so-called secondary emission is purely quantum mechanical. However, the classical and quantum emissions are coupled in the same way as for the QED correlation study. The resulting fluctuations of secondary emission obey the Heisenberg uncertainty principle; in the following, the special quantum features of these fluctuations are investigated.

In the full analysis, we determine the variance ΔX and ΔY of the emission as shown in Figure 8 (see also Figure 1). The Heisenberg uncertainty principle requires that the quadrature fluctuations obey $\Delta X \Delta Y \geq 1$. For a specific quadrature, the minimum uncertainty limit is usually defined to be $\Delta X = 1$ or $\Delta Y = 1$. If ΔX and ΔY are different, the observed light field is squeezed; and if the variance in one quadrature is less than one, the field is squeezed below the minimum uncertainty limit. In both cases, the field has a strong quantum nature; in particular, squeezing below unity suggests that measurements can be more accurate than the standard quantum limit for that quadrature.

To illustrate the behavior of the quantum fluctuations, we excite the quantum well resonantly with a relatively strong pulse. The emission is detected at an angle of 45° away from the excitation axis. Since the excitation pulse is relatively strong, the carrier density starts to oscillate during the pulse because Fermionic carriers can be excited only once; thus, further excitation actually leads to de-excitation. In other words, when these states become almost fully excited, the pulse can no longer excite the system, and we observe periodic de-excitation and excitation analogous to the Rabi-flopping of a strongly excited two-level system. Figure 8 shows the exciting pulse and corresponding quadrature fluctuations during the excitation process. As long as the pulse is present, the squeezing is at the 4% level and oscillates with the Rabi-flopping frequency. The quadrature fluctuations clearly show squeezing below the minimum uncertainty limit. Hence, the field has obvious quantum properties.

Similar squeezing and quantum characteristics statistics have been predicted and observed in the photon statistics of resonance fluorescence from two-level atoms subjected to an intense coherent light beam. The quantum properties of the scattered light

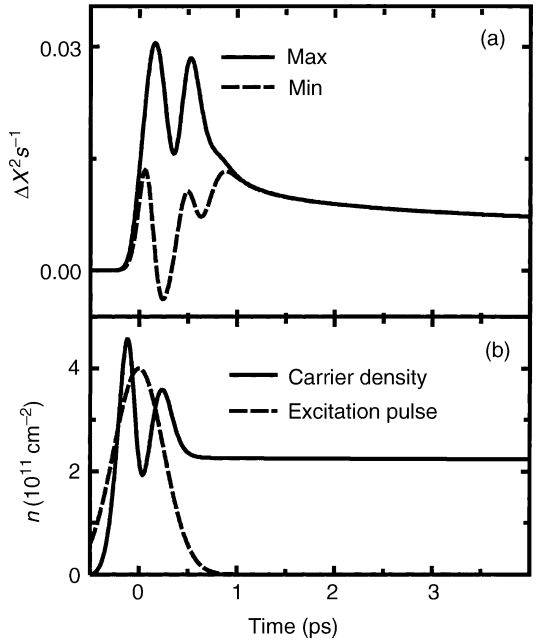


Figure 8 (a) Time-resolved maximum (solid line) and minimum (dashed line) deviations of the mode quadratures from the vacuum value of unity. (b) Corresponding time-resolved carrier density (solid line) and excitation pulse (dashed line). From Kira M, Jahnke F and Koch SW (1999) Quantum theory of secondary emission in optically excited semiconductor quantum wells. *Physical Review Letters* 82: 3544. Copyright (2004) by the American Physical Society.

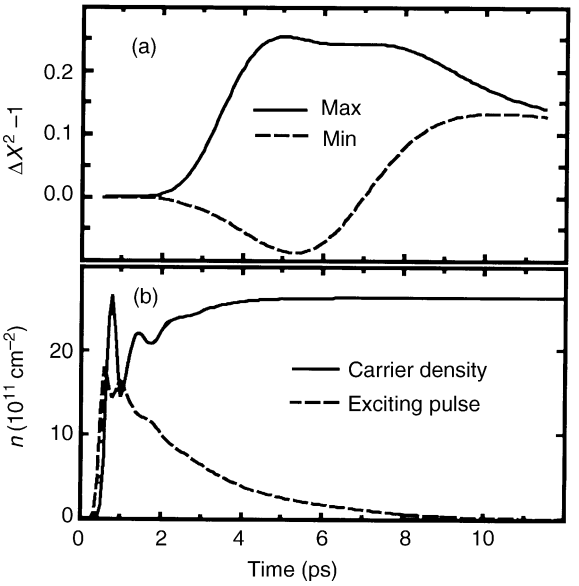


Figure 9 Squeezed light from a quantum well inside a microcavity: (a) the time evolution of maximum and minimum of the quadrature fluctuations; (b) the corresponding density and light field intensity at the quantum well.

in both quantum-well and atomic emission are enhanced when the driving field forces the system to oscillate between the excited and de-excited state. Just as described above, when the system becomes de-excited, it can no longer emit an additional photon; this inhibition manifests itself as sub-Poissonian photon statistics and squeezing in the mode quadratures. When the excitation pulse re-excites the system, such restrictions are no longer imposed. Thus, the field properties show quantum features oscillating with the Rabi-flopping frequency. Squeezing effects can also be observed with excitation by an electric current; for example, amplitude squeezing of diode laser emission has been demonstrated by controlling the electron statistics of the current flow.

The squeezing investigations demonstrate that some QED effects are expected to be seen without a cavity. Figure 9 shows the quadrature fluctuations when the cavity is added to the squeezing analysis. We observe qualitatively similar squeezing, but now the level of squeezing is enhanced up to 30% compared to the minimum uncertainty. This is once again a demonstration how QED effects can be amplified by using an increased light–matter coupling provided by the cavity.

Summary

This article overviews some of the most intriguing features of cavity quantum electrodynamics in semiconductors. Even though the related research has started only recently, several important quantum phenomena have already been predicted and measured: (a) the discrete quantum mechanical states of light and/or matter can be measured with strong coupling; (b) the consequences of the superposition principle have been detected as light–matter entanglement; and (c) the Heisenberg uncertainty principle has been tested via the squeezing of light. In all cases, the cavity enhances light–matter coupling leading to more pronounced quantum effects. Compared to atomic systems, the semiconductor has strong Coulomb correlation effects due to the relatively high density, which makes the analysis challenging. However, the semiconductor can also provide new mechanisms like the photon/carrier-density correlations which trigger new unexpected quantum phenomena.

The field of semiconductor quantum optics is very active and is developing rapidly. It benefits both from advances in computer capabilities and in semiconductor crystal growth. The increasing computer capacity allows more profound, accurate, and realistic modeling of semiconductor structures. At the same time, advances in crystal growth technology provide us with improved samples which are almost disorderless. In particular, the growth of quantum

wells with narrower exciton linewidths and quantum dots with larger dipole moments and reduced dephasing rates may be achieved, which is certain to make quantum features increasingly apparent and unavoidable.

All these research efforts eventually focus on producing devices utilizing quantum mechanical principles. One of the main objectives endeavors to develop quantum logic components for building blocks of quantum computers. Similar expanding possibilities can be expected, e.g., for accuracy of detection, device efficiency, and component design in general. Considering all of this, we are almost guaranteed to see new astonishing advancements; in this bright future, semiconductor cavity QED research will most likely be a pre-eminent element.

See also

Quantum Electrodynamics: Cavity QED.

Further Reading

Allen L and Eberly JH (1987) Optical Resonance and Two-Level Atoms, 1st edn. New York: Dover.

Banyai L and Koch SW (1993) Semiconductor Quantum Dots, 1st edn. Singapore: World Scientific.

Berman PR (ed.) (1994) Several articles in *Cavity Quantum Electrodynamics*, 1st edn. Boston, MA: Academic Press.

Cohen-Tannoudji C, Dupont-Roc J and Grynberg G (1989) Photons and Atoms, 3rd edn. New York: Wiley.

Haug H and Koch SW (2004) Quantum Theory of the Optical and Electronic Properties of Semiconductors, 4th edn. Singapore: World Scientific.

Khitrova G, Gibbs HM, Jahnke F, Kira M and Koch SW (1999) Nonlinear optics of normal-mode-coupling semiconductor microcavities. *Reviews of Modern Physics* 71: 1591–1639.

Kira M, Jahnke F, Hoyer W and Koch SW (1999) Quantum theory of spontaneous emission and coherent effects in semiconductor microstructures. *Progress in Quantum Electronics* 23: 189–279.

Scully MO and Zubairy MS (1999) Quantum Optics, 1st edn. Cambridge, UK: Cambridge University Press.

Walls DF and Milburn GJ (1994) Quantum Optics, 1st edn. New York: Springer-Verlag.

QUANTUM OPTICS

Contents

Atom Optics
Atomic Coherence Effects
Entanglement and Quantum Information
Laser Cooling of Ions
Quantum Computing with Atoms
Squeezed Phonons in Solids

Atom Optics

A D Cronin and D E Pritchard, Massachussetts Institute of Technology, Cambridge, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

In 1923 Louis de Broglie suggested that particles of matter propagate as waves with a wavelength

$$\lambda_{dB} = h/p \qquad [1]$$

where h is Planck's constant and p is the particle's momentum. This breakthrough idea motivated Schrödinger's equation, which is the equation of motion of quantum mechanics.

Schrödinger's equation is a wave propagation equation, and therefore implies the existence of a whole new type of optics – matter wave optics – in which electron waves, neutron waves, atom waves, and entire molecule waves can be manipulated coherently. Such optics enable new measurement technologies and devices that depend on matter wave interference.

Guided by a knowledge of light waves, one can understand that the de Broglie wavelength of a particle sets the scale for focusing, diffraction, and interference. However, there are differences between

matter and light. For example, the vacuum is tremendously dispersive for matter waves. In other words, the group velocity for a particle depends on its de Broglie wavelength, while for light waves the group velocity is independent of wavelength.

To illuminate the mathematical connection with light optics, we note the wave equation for light is

$$\nabla^2 \vec{E} = \mu\varepsilon \frac{\partial^2 \vec{E}}{\partial t^2} \qquad [2]$$

where $\vec{E} = \vec{E}(\vec{x}, t)$ is the optical electric field, and the parameters μ and ε may vary throughout space, describing an index of refraction. The wave equation for matter waves is Schrödinger's equation:

$$\left(-\frac{\hbar^2 \nabla^2}{2m} + V \right) \psi = i\hbar \frac{\partial}{\partial t} \psi \qquad [3]$$

where $\psi = \psi(\vec{x}, t)$ is the probability amplitude, or wavefunction, for a particle of mass m. Here, the potential energy V can vary throughout space, and thus control the phase and amplitude of propagating atoms. Just as density for photons is proportional to $\vec{E}(\vec{x})^2$, the probability density for particles is given by $|\psi(\vec{x})|^2$.

In a vacuum $V = 0$ and $(\mu_0 \varepsilon_0)^{-1/2} = c$, and if we assume that both ψ and E have a time dependence described by $e^{-ik\omega}$, the wave equations become

$$\left[\nabla^2 + \frac{\omega^2}{c^2} \right] \vec{E} = 0 \qquad [4]$$

and

$$\left[\nabla^2 + \frac{2m\omega}{\hbar} \right] \psi = 0 \qquad [5]$$

which look similar, but have a critical difference stemming from the single time derivative in Schrödinger's equation. To see how this makes the vacuum dispersive for matter waves, but not for light, consider a propagating plane wave of either ψ or $\vec{E}$ described by $e^{i(kx - \omega t)}$ (for either kind of wave, $k \equiv 2\pi(\lambda)^{-1}$). Using the wave equation for light waves, the dispersion relation is

$$\omega = ck \qquad [6]$$

where c is the phase (and group) velocity of light, which is independent of k. The same plane wave substituted into Schrödinger's equation for matter waves yields the dispersion relation

$$\omega = \frac{\hbar k^2}{2m} \qquad [7]$$

This is quadratic for matter waves, which means the group velocity $v_g = d\omega/dk$ is twice the phase velocity $v_p = \omega/k$, and both still depend on k.

From an engineering perspective, matter wave optics consists of lenses, mirrors, beam-splitters, and other components of an optics toolkit. The fact that electrons and neutrons can be transmitted through solid material, and that electrons can be reflected or refracted by static electromagnetic fields, has enabled optical elements to be created for these kinds of matter waves. For atoms, however, finding suitable materials for lenses or mirrors is a daunting task, since with rare exceptions (e.g., cold H on liquid He) atoms thermalize on surfaces rather than elastically bouncing off or passing through them. This was the major obstacle which delayed the blossoming of atom optics until the mid-1980s.

This obstacle has now been circumvented, but only with difficulty. Electromagnetic fields (whose spatial configurations are severely constrained by Maxwell's equations) can change the potential energy of neutral atoms, but since an atom's net charge is zero, provide only a weak and highly dispersive index of refraction based on the atom's polarizability or magnetic moment. However, these effects are tremendously enhanced if the electromagnetic fields oscillate near a transition frequency of the atom. Hence, propagating waves and standing waves of single-frequency laser light have become standard tools in atom optics. Standing light waves are often referred to as a light crystal, because the periodic potential can diffract atom beams.

An alternative source of atom optical components is based on nanofabrication. Thin sheets of material with patterned holes serve as absorptive atom optics by transmitting atom waves only through the nanometer scale openings. A variety of techniques have been used to make physical diffraction gratings, zone plates, and holograms for atom waves. Methods such as electron beam lithography and ultraviolet photolithography both rely on sophistocated etching procedures to make the final free-standing structures. Thus, in either approach – near-resonant light, or miniature physical structures – current technologies such as the tunable laser or nanofabrication techniques have been required for the development of useful atom optics.

While coherent manipulation of atom waves is often the goal, incoherent manipulations are possible also, e.g., if photons are spontaneously scattered. Dissipative processes offer an additional class of possibilities such as increasing the brightness of an atom source, and are therefore an important part of atom optics. This article excludes atom slowing and trapping, and focuses instead on atom diffraction – a coherent

process which can be used to make beam splitters for atom interferometers.

Nanofabricated Atom Optics

Because atoms stick to surfaces, mechanical structures are absorptive atom optics. Yet the de Broglie waves transmitted through slots of a mask are nearly unaffected, provided the mask is thin enough. This is because the van der Waals potential energy of interaction with nearby surfaces falls off rapidly for neutral atoms.

Consider, for example, an atom wave packet being diffracted by a mechanical grating made with physical bars and slots as shown in Figure 1. Initially, let the atom wave packet have a slowly varying spatial envelope, $f = f(x, y, z)$, and momentum only in the $\hat{x}$-direction, $\vec{p} = \hbar\vec{k}_0 = h(\lambda_{dB})^{-1}\hat{x}$, so the wavefunction can be written as

$$\psi = fe^{i(k_0 x - wt)} \qquad [8]$$

Immediately after the grating, only the portions of the wavefunction passing through slots remain nonzero, as depicted in Figure 2. Here, the wavefunction is now periodic both due to the grating lattice vector $k_g\hat{y}$ and the initial wavevector $k_0\hat{x}$. According to the Schrödinger equation, a periodicity in any direction implies a momentum in that direction. Thus, variation of ψ in the $\hat{y}$ variation gives rise to a new momentum distribution.

The square wave modulating the envelope of the wave packet can be described by a Fourier sum times the incident wavefunction

$$\psi = fe^{i(k_0 x - \omega t)} \sum_{n=-\infty}^{\infty} a_n \cos(nk_g y) \qquad [9]$$

where $k_g = 2\pi(d_g)^{-1}$ is the grating lattice vector. The only other physical parameter of the grating is the open fraction, γ, and this determines the Fourier amplitudes a_n

$$a_n = \frac{\sin(n\pi\gamma)}{n\pi} \qquad [10]$$

From the principle of superposition (which is applicable here because the Schrödinger equation is linear) the wavefunction can be written as a sum of traveling waves with wavevectors $\vec{k}_n$:

$$\psi = f\sum_{n=-\infty}^{\infty} a_n e^{(i\vec{k}_n\cdot\vec{r} - wt)} \qquad [11]$$

$$\vec{k}_n = \frac{k_0\hat{x} + nk_g\hat{y}}{N} \qquad [12]$$

where, just as in standard optics, the factor N serves to keep $|\vec{k}_n|$ on a spherical shell, i.e., the grating does not add energy to the atom.

Inserting this expression for the wavefunction into Schrödinger's equation we get the dispersion relation

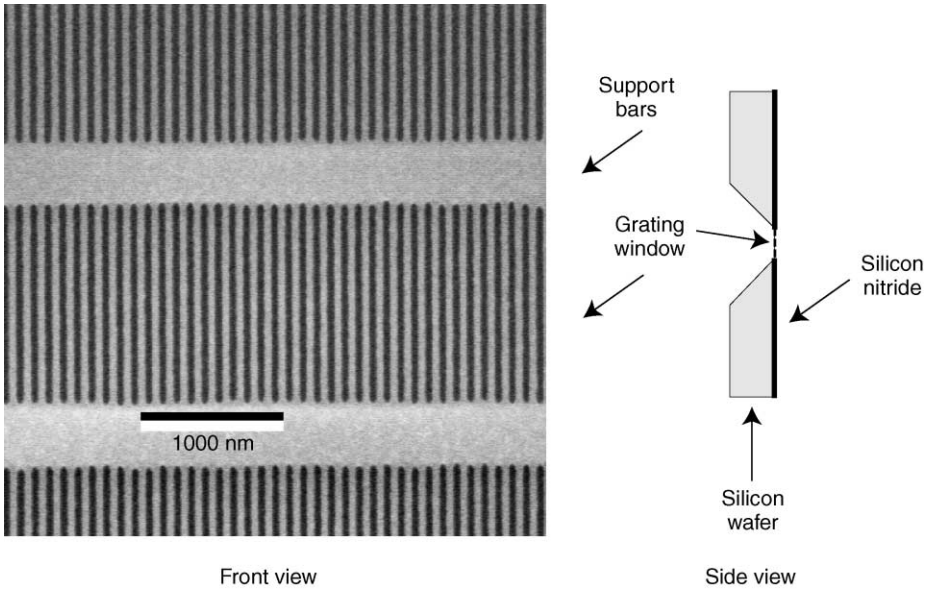


Figure 1 Scanning electron microscope image of a silicon nitride diffraction grating for atom waves. The bars (light) absorb, and the slots (dark) transmit atom waves. The grating period is 100 nm, the open fraction is 0.67, and the free-standing bars are supported every 5 μm with a thicker strip of silicon nitride material. The grating was fabricated by Tim Savas and Henry I Smith at MIT NanoStructures laboratory.

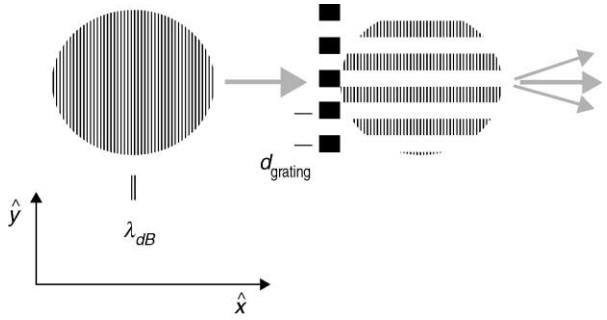


Figure 2 Schematic diagram of an atom wavepacket passing through an absorbing grating, as seen from above. The atom wavepacket is depicted before and after passing through the grating from left to right.

for matter waves and a different group velocity, v_n, for each k_n

$$\hbar\omega = \frac{\hbar^2 k_n^2}{2m} \qquad [13]$$

$$v_n = \frac{d\omega}{dk_n} = \frac{\hbar k_n}{m} = p_n/m \qquad [14]$$

Thus, diffraction through a grating puts an atom into a coherent superposition of different momenta, each separated by $\sim\hbar k_g\hat{y}$. Some time later, the atom will be in a coherent superposition of multiple locations, and the far-field probability density will be peaked at integer multiples of the diffraction angle

$$\Theta_D = \frac{k_g}{k_0} = \frac{\lambda_{dB}}{d_{\text{grating}}} \qquad [15]$$

where d_{grating} is the lattice spacing of the grating and the small-angle approximation for $\sin(\theta)$ has been used. Exactly as in light optics, these diffraction gratings are good momentum spectrometers, i.e., the diffraction angle is a measure of longitudinal momentum. (Contrary to light optics, group velocity can also be used to measure matter wave momentum.) Coherence between the diffraction orders remains until a perturbation occurs which would (even in principle) be sufficient to determine which path the atom took. This assertion has been tested with atoms in an interferometer (discussed below).

Atom flux in the nth diffraction order is given by

$$I_n = \left(\frac{\sin(n\pi\gamma)}{n\pi}\right)^2 \qquad [16]$$

To resolve the diffracted orders the initial transverse momentum distribution of the atoms must be smaller

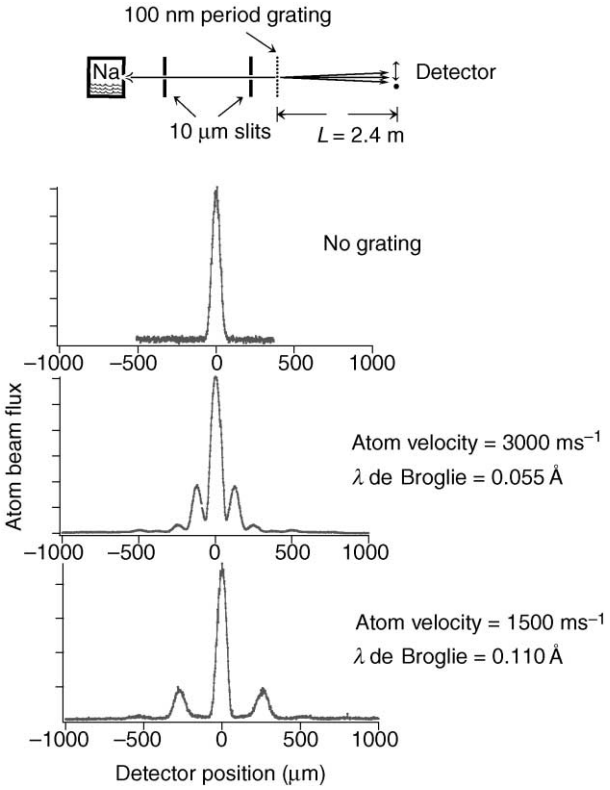


Figure 3 Diffraction data from a beam of sodium formed by passing atoms through two collimating slits followed by a nanofabricated diffraction grating. The momentum of the atoms determines their de Broglie wavelength, and thus their diffraction angle $\sin(\Theta_D) = \lambda_{dB}/d_{\text{grating}}$. These data were obtained at MIT by the authors.

than k_g, and the time until the far-field observation must be sufficiently large given the beam width. At intermediate times, the atom waves are in the Fresnel diffraction region where the Talbot effect can be observed. This effect causes the atom waves to form a replica of the grating in free space at integer multiples of the Talbot length, $Z_T = 2d^2_{\text{grating}}\lambda_{dB}^{-1}$.

Figure 3 shows the far-field diffraction pattern from a beam of sodium (^{23}Na) atoms using a material grating with a period of 100 nm. The atom beam velocity peaked at 1500 ms^{-1} corresponds to a de Broglie wavelength of 0.11 Å; thus, $\Theta_D \approx 10^{-4}$ rad. (The beam of 3000 ms^{-1} atoms diffracts at half this angle.) The spread in longitudinal velocity of $\sigma_v/v \approx 6\%$ can be deduced from the data because it slightly increases the width for higher order diffraction peaks. The open fraction of the grating, $\gamma \approx 60\%$, can also be determined from the diffraction data, because it determines the relative intensity of the orders. Effects of the atoms' finite size and interaction with the grating bars make small corrections to this simple theory.

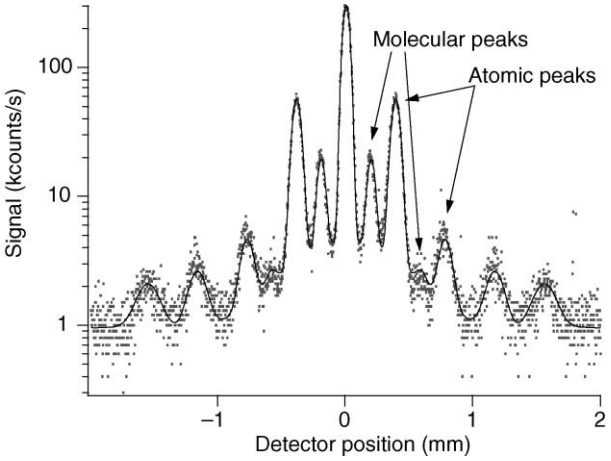


Figure 4 Diffraction data using a mixed beam of sodium atoms (Na) and sodium dimer molecules (Na_2). From this source molecules have the same speed as atoms, thus twice the momentum. Thus, molecule waves are diffracted at half the diffraction angle compared to atom waves. These data were obtained at MIT by the authors.

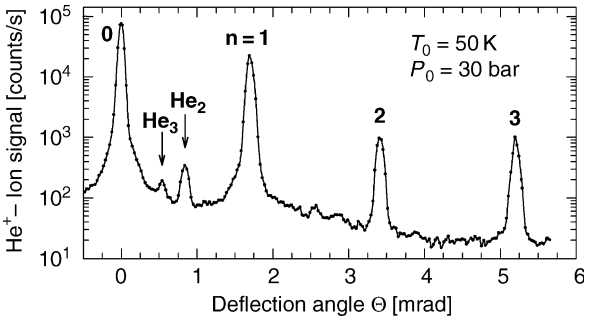


Figure 5 Diffraction of helium atoms and helium molecules through a nanofabricated grating. These data, reproduced with permission from Wieland Schoellkopf, were obtained at the Max-Planck-Institute in Göttingen.

Nanofabricated atom optics differ from those based on light forces in a number of ways: they are amplitude structures (with corresponding loss of transmission intensity), they are species-independent, their scale size can be several times smaller than attainable with light, and they can be arbitrarily patterned since they are fabricated by electron beam lithography. Insensitivity to species is demonstrated in Figure 4, where a beam of Na_2 molecules and Na atoms is diffracted and separated by a nanofabricated diffraction grating. The flexibility of electron beam lithography has allowed fabrication not only of diffraction gratings, but also of spherical and cylindrical zone plates, as well as a combination of lens and a hologram that generates a focused atom image. One stunning application laid to rest a long-standing argument concerning whether a stable bound state of the $^4He_2$ dimer exists. For this a diffraction grating was used to separate and resolve $^4He_2$ and more massive clusters (Figure 5). Subsequently a nano-sieve was used to estimate the size of the $^4He_2$ dimers to be 62 ± 10 Å.

Standing Waves of Light

In contrast to nanofabricated structures, standing waves of near-resonant light are phase gratings for atoms. The optical electric field dynamically polarizes the atoms and causes a shift to the ground state energy (the ac-Stark shift). For optical electric fields oscillating above (below) the atomic resonance frequency, the ac-Stark shift increases (decreases) potential energy for atoms and in either case perturbs the de Broglie wave phase.

In many respects, the interaction of an atom with a standing light wave is richer than the more familiar topic of light–atom interactions in a traveling light wave. Part of this richness reflects two ways in which a standing wave can be considered, either as two counter-propagating traveling waves or as a single, stationary standing wave. The standing wave–atom interaction is capable of transferring momentum in well-determined quantities, coherently splitting atomic wave packets, and generating forces much larger than possible with spontaneously scattered light. Not surprisingly, this interaction has many distinct facets that only appear in different regimes of the interaction parameters (intensity, detuning, and pulse duration) and atomic parameters (mass, initial momentum with respect to the standing wave, and excited state natural lifetime).

We begin with the Bragg scattering of atoms from a standing wave light grating. Although it can be difficult to realize the physical conditions that assure its occurrence, pure Bragg scattering is simpler than intermediate cases involving spontaneous decay (from light too close to resonance) or Kapitza–Dirac diffraction (from shorter interaction times).

Consider a standing wave light grating formed by two counter-propagating plane waves (traveling parallel to the z-axis) of equal amplitude, E_0, wavevector, k, frequency, ω, polarization vector, $\hat{e}$, and temporal envelope function, $f(t)$:

$$\vec{E}(z,t) = E_0 f(t)\sin(kz - \omega t)\hat{e} + E_0 f(t)\sin(kz + \omega t)\hat{e} \quad [17]$$

$$= 2E_0 f(t)\sin(kz)\cos(\omega t)\hat{e} \quad [18]$$

We would like to work with momentum states as our atomic basis, thus it is easiest to consider the

description of the electric field driving the transitions in terms of two counter-propagating traveling waves of definite momentum (eqn [17]), as opposed to the single standing wave they jointly form (eqn [18]).

Momentum is transferred by paired stimulated absorption and emission processes, resulting in a transfer of photons between the traveling waves. An N_B^{th} order diffraction process transfers N_B photons from one traveling wave to the counter-propagating traveling wave and changes the atomic momentum by $2N_B\hbar k\hat{z}$. Furthermore, atomic population is transferred only between $|g, -N_B\hbar k\rangle$ and $|g, +N_B\hbar k\rangle$, where $|g(e), \pm N\hbar k\rangle$ denotes a two-level atom in its ground (excited) state with momentum $\pm N\hbar k$ parallel to the standing wave axis. The excited state remains nominally unpopulated so long as the temporal envelope function, $f(t)$, does not have strong frequency components near the laser detuning

$$\delta = \omega - \omega_0 \qquad [19]$$

where ω_0 is the unperturbed frequency of the atomic transition. Furthermore, for a given initial state ($|g, -N_B\hbar k\rangle$) the uniqueness of the final state ($|g, +N_B\hbar k\rangle$) occurs because of the fundamental assumption that makes Bragg scattering so simple to describe; the uncertainty in the photon energy driving the transitions is small compared to the energy separation between neighboring momentum states. A quantitative discussion of the validity of this assumption will be given later.

We now calculate the probability, $P_1^B(\tau)$, of the first order ($N_B = 1$) Bragg process. Scattering transfers an atom from $|g, -\hbar k\rangle$ to $|g, +\hbar k\rangle$ when the atoms interact with a constant light intensity for a time τ (i.e., in eqns [17] and [18] $f(t)$ is a square wave of unit amplitude and duration τ). The geometry for Bragg diffraction from a standing wave of light is shown in Figure 6, and the energy/momentum states involved in the transition are depicted in Figure 7.

Considering the electronic degree of freedom, we rewrite the Schrödinger equation as

$$(\mathcal{H}_0 + \mathcal{H}_{\text{int}})|\psi\rangle = i\hbar\frac{\partial}{\partial t}|\psi\rangle \qquad [20]$$

where the total Hamiltonian $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_{\text{int}}$ is a matrix. The wavefunction is a vector of coefficients c_i in the basis $\{|g, -\hbar k\rangle, |e, 0\rangle, |g, +\hbar k\rangle\}$, with normalized population amplitudes for each

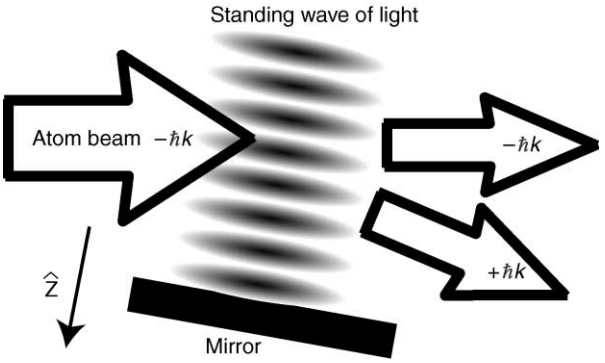


Figure 6 Two counter-propagating running waves of light superimpose to form a standing wave of light. This 'light crystal' acts as a diffraction grating for atoms, shown here in the configuration for Bragg diffraction.

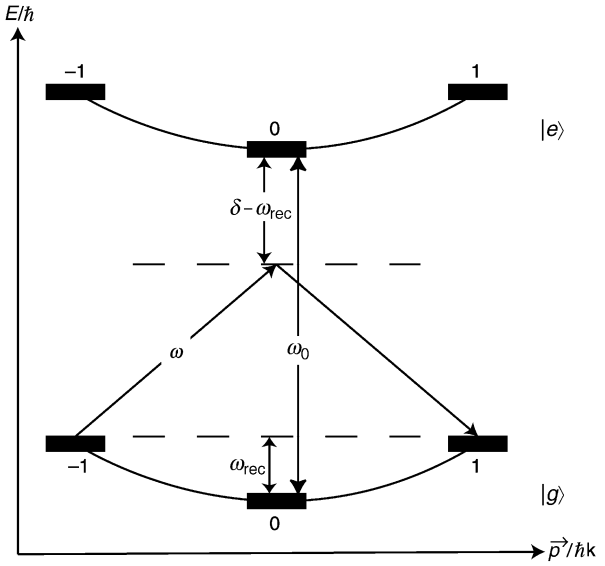


Figure 7 Energy vs. momentum diagram describing atomic states involved in Bragg diffraction. The total energy is due to the electronic state ($|g\rangle$, $|e\rangle$) plus the kinetic energy associated with the recoil from photon emission (or absorption). In Bragg diffraction, atoms make a coherent transition from $|g, -\hbar k\rangle \leftrightarrow |g, +\hbar k\rangle$. Reprinted from Gupta S, *et al.* (2001) Coherent manipulation of atoms with standing light waves. *Comptes Rendus de L'Academie des Sciences – Series IV – Physics-Astrophysics* 2:479–495, copyright (2001) with permission from Elsevier.

component:

$$|\psi\rangle = \begin{pmatrix} c_{-1}(t) \\ c_0(t) \\ c_{+1}(t) \end{pmatrix} \qquad [21]$$

In the electric dipole approximation, the interaction Hamiltonian is $\mathcal{H}_{\text{int}}(t) = -\vec{\mu}\cdot\vec{E}(t)$, where $\mu = \langle e|q\vec{r}|g\rangle\cdot\hat{e}$ is the electric dipole matrix element connecting the ground ($|g\rangle$) and excited ($|e\rangle$) states

of the atom. $\hat{e}$ is the polarization vector of the light and q is the charge of the electron. By momentum conservation, only the plane wave traveling in the $\pm\hat{z}$ direction couples the $|e,0\rangle \leftrightarrow |g,\mp\hbar k\rangle$ transitions. By using this argument, we are effectively viewing the electric field as a quantum mechanical operator. Expanding the sinusoidal variation of the electric field in terms of complex exponentials and treating the spatially dependent complex exponential terms as quantum mechanical momentum translation operators yields the interaction Hamiltonian:

$$\mathcal{H}_{\text{int}} = \frac{\hbar\omega_R}{2}\begin{pmatrix} 0 & ie^{i\omega t} & 0 \\ -ie^{-i\omega t} & 0 & ie^{-i\omega t} \\ 0 & -ie^{i\omega t} & 0 \end{pmatrix} \qquad [22]$$

For example, $e^{\pm ik\hat{z}}|g,n\hbar k\rangle = |e,(n\pm 1)\hbar k\rangle$ describes absorption, and for stimulated emission, $e^{\mp ik\hat{z}}|e,n\hbar k\rangle = |g,(n\pm 1)\hbar k\rangle$. This procedure includes the rotating wave approximation. In addition, we have neglected any frequency components associated with the sudden switch on of the fields and the finite duration of the light–atom interaction. In formulating $\mathcal{H}_{\text{int}}$, the strength of the light–atom interaction is parameterized by the single-photon Rabi frequency:

$$\omega_R = \frac{\mu E_0}{\hbar} \qquad [23]$$

Without loss of generality, we will take μ and hence ω_R to be positive, real-valued quantities.

The total Hamiltonian follows simply by including the electronic and kinetic energy terms:

$$\mathcal{H}_0 = \hbar\begin{pmatrix} \omega_{\text{rec}} & 0 & 0 \\ 0 & \omega_0 & 0 \\ 0 & 0 & \omega_{\text{rec}} \end{pmatrix} \qquad [24]$$

where the single-photon recoil energy, E_{rec}, of an atom of mass m is given by

$$E_{\text{rec}} = \hbar\omega_{\text{rec}} = \frac{\hbar^2k^2}{2m} \qquad [25]$$

Making the *ansatz* for the solution wavefunction as

$$|\psi\rangle = \begin{pmatrix} c_{-1}(t)e^{-i\omega_{\text{rec}}t} \\ c_0(t)e^{-i\omega_0 t} \\ c_{+1}(t)e^{-i\omega_{\text{rec}}t} \end{pmatrix} \qquad [26]$$

and substituting into the Schrödinger equation yields three coupled first-order differential equations:

$$\dot{c}_{\pm1}(t) = \mp\frac{\omega_R}{2}e^{i\Delta t}c_0(t), \qquad [27]$$

$$\dot{c}_0(t) = \frac{\omega_R}{2}e^{-i\Delta t}\big(c_{+1}(t) - c_{-1}(t)\big), \qquad [28]$$

where $\Delta = \delta + \omega_{\text{rec}} = (\omega - \omega_0) + \omega_{\text{rec}}$. Differentiating eqn [27] and substituting eqn [28] into the result yields two coupled second-order differential equations:

$$\ddot{c}_{\pm1}(t) - i\Delta\dot{c}_{\pm1}(t) + \frac{\omega_R^2}{4}\big(c_{\pm1}(t) - c_{\mp1}(t)\big) = 0 \qquad [29]$$

With the initial conditions:

$$c_{-1}(0) = 1 \qquad [30]$$

$$c_0(0) = 0 \Rightarrow \dot{c}_{\pm1}(0) = 0 \qquad [31]$$

$$c_{+1}(0) = 0 \qquad [32]$$

and the assumption $\Delta^2 \gg \omega_R^2$, the solutions to eqn [29] are

$$c_{-1}(t) = e^{-i\frac{\omega_R^{(2)}}{2}t}\cos\left(\frac{\omega_R^{(2)}}{2}t\right) \qquad [33]$$

$$c_{+1}(t) = ie^{-i\frac{\omega_R^{(2)}}{2}t}\sin\left(\frac{\omega_R^{(2)}}{2}t\right) \qquad [34]$$

where the two-photon Rabi frequency is

$$\omega_R^{(2)} = \frac{\omega_R^2}{2\Delta} \rightarrow \frac{\omega_R^2}{2\delta}, \quad |\delta| \gg \omega_{\text{rec}} \qquad [35]$$

with both transitions driven at equal single-photon Rabi frequencies, ω_R.

Substituting the solution of eqn [33] or [34] into eqn [27] yields an expression for the excited state amplitude:

$$c_0(t) = -i\frac{\omega_R}{2\Delta}e^{-i\Delta t}e^{-i\omega_R^{(2)}t} \qquad [36]$$

This will be important in calculating the rate of spontaneous emission events later.

The solutions for $c_{-1}(t)$ and $c_{+1}(t)$ oscillate with the interaction duration, τ, yielding the result for the $|g,-\hbar k\rangle \rightarrow |g,+\hbar k\rangle$ transition probability:

$$P_1^B(\tau) = |c_{+1}(\tau)|^2 = \sin^2\left(\frac{\omega_R^{(2)}}{2}\tau\right) \qquad [37]$$

Thus, the system oscillates between the two momentum states $|g,-\hbar k\rangle$ and $|g,+\hbar k\rangle$ in a manner analogous to the Rabi oscillation of atomic population between two resonantly coupled states. This solution with

oscillatory probabilities for the two Bragg coupled states is known as the Pendellösung and has been observed for atoms, neutrons, and X-rays (Figure 8). The nice feature here is that the strength of the grating can be actually controlled by the intensity of the light.

Viewing Bragg scattering as a two-photon transition from the initial ground state to the final ground state with opposite momentum, illuminates the close connection with a Raman transition between two internal sub-states of the ground state manifold, each with its own external momentum state. The formalism describing the Raman transition is basically the same as that presented here, except the two transitions can be driven at different single-photon Rabi frequencies, ω_{R1} and ω_{R2}, so that the generic two-photon Rabi frequency is given by $\omega_R^{(2)} = \omega_{R1}\omega_{R2}/2\Delta$, where Δ is the detuning from the intermediate state.

An N_B^{th} order Bragg process (similar to a $2N_B$-photon Raman process) is a coherent succession of N_B two-photon transitions from $|g, -N_B\hbar k\rangle$ to $|g, +N_B\hbar k\rangle$ with $2N_B - 1$ intermediate states of the form:

$$|e, (-N_B + 1)\hbar k\rangle, |g, (-N_B + 2)\hbar k\rangle \ldots$$

$$\ldots |g, (N_B - 2)\hbar k\rangle, |e, (N_B - 1)\hbar k\rangle$$

Such a process is characterized by a $2N_B$-photon Rabi frequency given by

$$\omega_R^{(2N_B)} = \frac{[\omega_R]^{2N_B}}{2^{2N_B-1}\Delta_1\Delta_2\cdots\Delta_{2N_B-1}} \qquad [38]$$

where Δ_n is the detuning from the n^{th} intermediate state. Figure 9 shows what this process would look like for an N_B^{th} order Bragg transition where the intermediate state detunings are given by

$$\Delta_n = \left\{ \begin{array}{ll} \delta + (2N_B n - n^2)\omega_{\text{rec}} : & n \text{ odd} \\ (2N_B n - n^2)\omega_{\text{rec}} : & n \text{ even} \end{array} \right\} \qquad [39]$$

Substituting these detunings into eqn [38] yields the N_B^{th} order Bragg transition $2N_B$-photon Rabi frequency, $\omega_R^{(2N_B)}$:

$$\omega_R^{(2N_B)} = \frac{[\omega_R]^{2N_B}}{2^{4N_B-3}[(N_B - 1)!]^2\delta^{N_B}\omega_{\text{rec}}^{N_B-1}} \qquad [40]$$

where we have assumed $|\delta| \gg N_B^2\omega_{\text{rec}}$.

To ensure that the system truly undergoes Bragg scattering, and validate the assumption that only states of equal kinetic energy and opposite

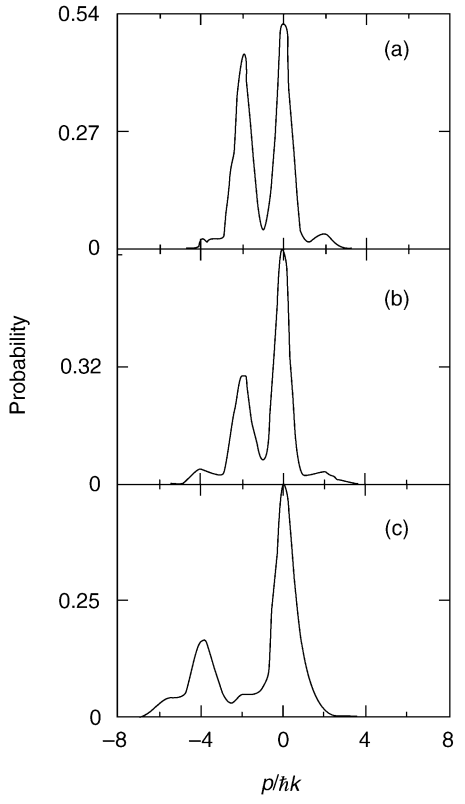


Figure 8 Bragg scattering of an atom beam from an optical standing wave. First-order Bragg scattering at (a) lower power, (b) higher power of the laser beams (Pendellösung has set in, increasing the amplitude of the undeflected peak), (c) second-order Bragg scattering. Reprinted with permission from *Physical Review Letters* 60(6): 515–518. Copyright (1998) by the American Physical Society.

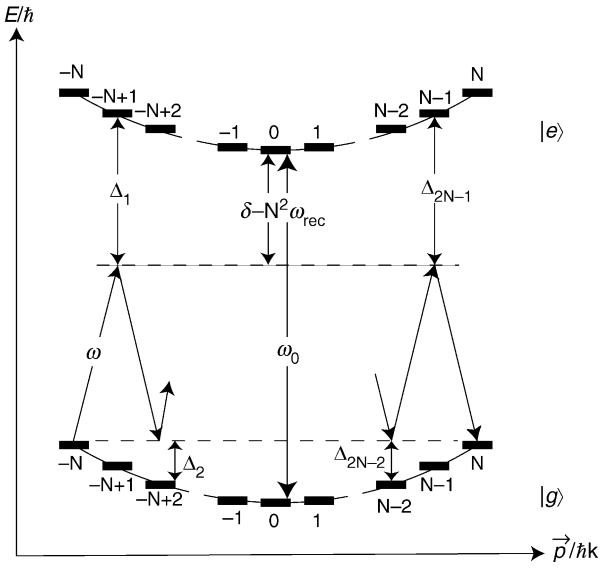


Figure 9 Energy vs. momentum diagram describing atomic states involved in higher-order Bragg diffraction. Reprinted from Gupta S, *et al.* (2001) Coherent manipulation of atoms with standing light waves. *Comptes Rendus de L'Academie des Sciences – Series IV – Physics-Astrophysics* 2:479–495, copyright (2001) with permission from Elsevier.

momentum are coupled, the overall exposure time, τ, of the atoms to the fields must be limited both from below and above.

The lower bound on τ is necessary to resolve the final momentum state ($|g, N_B\hbar k\rangle$) from neighboring momentum states ($|g, (N_B \pm 2)\hbar k\rangle$) two photon recoil momenta away. Avoiding population transfer into these states requires that the energy separation be resolvable, or

$$\tau(\delta E) \gg h \qquad [41]$$

For first-order Bragg scattering processes ($N_B = 1$) the nearest lying momentum states that may be mistakenly populated are $|g, \pm 3\hbar k\rangle$, for which

$$\delta E = \frac{p^2_{n=3}}{2m} - \frac{p^2_{n=1}}{2m} = 8\hbar\omega_{\text{rec}} \qquad [42]$$

Therefore

$$\tau \gg \frac{h}{8\hbar\omega_{\text{rec}}} = \frac{\pi}{4\omega_{\text{rec}}} \qquad [43]$$

For sodium atoms (^{23}Na) in 590 nm light, $\tau \gg 6\ \mu$s is the lower bound required for first-order Bragg scattering.

For all higher-order ($N_B > 1$) Bragg scattering processes, the nearest momentum states are $|g, \pm(N_B - 2)\hbar k\rangle$, which limits the interaction time to

$$\tau \gg \frac{\pi}{2(N_B - 1)\omega_{\text{rec}}} \qquad [44]$$

which for $N_B \gg 1$ reduces to

$$\tau \gg \frac{\pi}{2N_B\omega_{\text{rec}}} \qquad [45]$$

The upper bound on the interaction duration is necessary to avoid spontaneous emission. The interaction duration must be short enough so that the expected number, N_s, of spontaneous emission events per atom during the time τ is negligible. N_s is simply given by the product of the excited state fraction (eqn [36]) and the probability of spontaneous decay given that the atom is in the excited state:

$$N_s = |c_0(t)|^2\Gamma\tau = \frac{\omega_R^2}{4\Delta^2}\Gamma\tau \qquad [46]$$

where Γ is the natural decay rate of the excited state. Avoiding spontaneous emission ($N_s \ll 1$) while still having a significant probability for transitions ($\omega_R^{(2N_B)}\tau \simeq \pi$) forces one to work in the regime where $\Delta \gg \Gamma$, which is a practical requirement to avoid spontaneous emission and maintain coherence. For sodium atoms on the 590 nm transition, $\Gamma = 6.3 \times 10^7\ \text{s}^{-1}$, and a typical choice in Bragg diffraction experiments is $\Delta = 50\Gamma$, or $\Delta/2\pi = 500$ MHz.

Briefly, to relate Rabi frequency to optical intensity, the transition rate W on resonance with a Lorentzian spectral line is given by Fermi's golden rule:

$$W = \frac{\omega_R^2}{\Gamma} = \frac{\lambda^2_{\text{photon}}}{2\pi}\frac{I}{\hbar\omega} \qquad [47]$$

where the last term can be regarded as the cross section for absorption multiplied by intensity in units of photon flux. Hence

$$\omega_R^2 = \Gamma\frac{\lambda^2_{\text{photon}}}{2\pi}\frac{I}{\hbar\omega} \qquad [48]$$

For Na atoms in 12 mW(cm)$^{-2}$ resonant light, $\omega_R = \Gamma = 6.3 \times 10^7\ \text{s}^{-1}$. Both the probability of Bragg scattering and the spontaneous emission increase linearly with $\omega_R^2\tau\Gamma^{-1}$. However, spontaneous emission decreases as Δ^2 while the Bragg scattering probability decreases as Δ. Thus, as shown in **Figure 10**, at sufficient detuning and laser power, Bragg scattering occurs with no spontaneous emission. For reference, 12 mW(cm)$^{-2} \times 10\ \mu$s corresponds to $\omega_R^2\tau\Gamma^{-1} = 630$ for sodium atoms.

Bragg scattering of atoms from a standing light wave was first observed at MIT in 1988. A supersonic atomic beam was diffracted from a standing wave of near-resonant laser light as depicted in **Figure 6**. The angle between the atomic beam (of thermal wavelength λ_{dB}) and the light grating (of periodicity $\lambda_L/2$) was tuned to the appropriate Bragg angle, θ_B, where:

$$\lambda_{dB} = \lambda_L \sin(\theta_B) \qquad [49]$$

Population transfer corresponding to both first- and second-order Bragg scattering was observed (**Figure 7**). The experiment required a sub-recoil transverse momentum spread of the atomic beam in order to resolve the different momentum states in the far field and limit the final state to only one

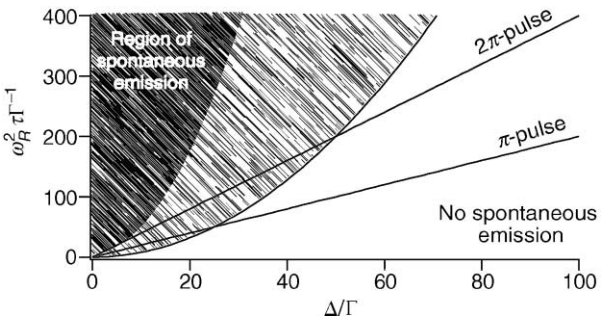


Figure 10 The conditions for which $N_s > 0.02(0.10)$ are hashed in light(dark) gray. ω_R^2/Γ is proportional to optical intensity. The line labeled π-pulse indicates the parameters which cause complete first-order Bragg diffraction. The line labeled 2π-pulse indicates conditions where population is completely transferred back to the initial state (Pendellösung oscillation).

diffracted order. The Pendellösung was observed as an oscillation in population transfer as a function of standing wave intensity, $I \propto \omega_R^{(2)}$, for a fixed interaction time, τ.

Atomic beam diffraction from an optical standing wave is a continuous-wave (CW) experiment in which the selectivity needed for the Bragg process is imposed by good angular resolution of the particle beam and a high degree of parallelism between the light crystal planes. This ensures that of the various final Bragg orders allowed by momentum conservation, only one conserves energy (energy conservation is exact in a CW experiment). For atoms scattering from a light crystal, parallelism of the crystal planes requires highly parallel photon momentum that implies a minimum width of the standing wave (the diffraction limit for the collimated photons). The transit time, τ, of the atoms across this width then exceeds the lower bound given in eqn [44].

The excellent collimation required of the atomic beam to ensure resolution of the Bragg scattered atoms reduces the intensity of the source by many orders of magnitude. A Bose–Einstein Condensate (BEC) is an attractive alternative source of atoms because its momentum spread is typically an order of magnitude below a single-photon recoil momentum. To Bragg diffract atoms initially in a stationary BEC, it is easier to move the light crystal than to accelerate the condensate. This is done by simply frequency shifting one of the traveling waves so that the resultant standing wave formed by its interference with the unshifted traveling wave moves with the proper velocity ($N_B\hbar k/m$) relative to the stationary atoms to impart the necessary momentum. The Bragg scattered atoms will then have momentum $2N_B\hbar k$ in the laboratory frame. The resonance condition thus becomes a condition on relative detuning, δ_{N_B}, between the two laser beams forming the diffraction grating. For N_B^{th}-order Bragg diffraction, the relative detuning is given by

$$\delta_{N_B} = \frac{2N_B\hbar k^2}{m} = 4N_B\omega_{\text{rec}} \qquad [50]$$

The first demonstration of Bragg scattering in a BEC was at NIST in 1999. They used Bragg scattering mainly as a tool to manipulate the momentum of the BEC, observing up to sixth-order processes. At MIT the interaction time was lengthened (to ≈100 times the lower bound of eqn [44] for first-order Bragg scattering), creating a new type of spectroscopy called Bragg Spectroscopy. It is a spectroscopic measurement of the shift of the Bragg scattering frequency due to the Doppler shift ($k \cdot v$) from the atoms' motion together with any mean field interaction. (These effects can be separated by going to higher-order Bragg scattering to enhance the Doppler shift.)

Bragg Spectroscopy was used to observe the momentum distribution of a BEC in a magnetic trap. The width of the Bragg resonance curve was primarily due to a 2 kHz Doppler-broadening that yielded the momentum distribution of the condensate. The spread in the corresponding velocity distribution was very small ($\approx 0.5\ \text{mm s}^{-1}$), even smaller than allowed by the Heisenberg uncertainty limit for a particle confined in the ground state of a harmonic trap. This reflects the increase in size of the condensate due to the mean field repulsion (^{23}Na, used in the MIT experiment, has a positive scattering length). The distribution was Heisenberg uncertainty limited at the observed size of the BEC, establishing for the first time that the coherence length of the condensate was equal to its size. In addition, the narrow Bragg resonance was shifted by the repulsive interactions within the condensate, resulting in a spectroscopic measurement of the mean-field energy.

Kapitza–Dirac Scattering

Diffraction of neutral atoms from a standing wave of near-resonant light with a short interaction time ($\tau_{KD} \ll 1/\omega_{\text{rec}}$) has come to be called Kapitza–Dirac scattering, in honor of their pioneering suggestion. In 1933, Kapitza and Dirac predicted that an electron beam propagating in a standing light wave would undergo stimulated Compton scattering and be reflected. This process has a tiny cross-section, equal to the classical electron radius squared: $\sigma_{\text{Compton}} = 8\pi/3(e^2/mc^2)^2 \approx 6 \times 10^{-25}\ \text{cm}^2$, and has only recently been observed using extremely high laser intensities. If the electrons are replaced by atoms, however, the scattering cross-section for resonant light $\sigma_{\text{atom}} = (1/2\pi)\lambda_{\text{photon}}^2 \approx 4 \times 10^{-10}\ \text{cm}^2$ is 15 orders of magnitude larger. This large cross-section, together with the ready availability of tunable lasers, has allowed stimulated scattering, both in the Kapitza–Dirac and Bragg regimes, to become the primary tool for the coherent manipulation of atoms.

In Kapitza–Dirac scattering, atomic motion during the interaction time is small compared to the characteristic dimensions of the interaction potential. This is equivalent to the eikonal approximation for scattering or the thin-lens approximation in optics. The idea is that the phase of the incident particle changes along each classical trajectory, but not the amplitude. Thus, there may be momentum transfer perpendicular to the trajectory, but the trajectory is not significantly displaced (until after the interaction is over). Mathematically, this regime

can be treated by neglecting the atomic kinetic energy term in the Hamiltonian (the Raman–Nath approximation).

The standing wave interaction may be treated by considering the standing wave (ac-Stark shift) potential resulting from the applied fields given in eqn [18]:

$$V(z,t) = \frac{\hbar\omega_R^2}{\delta} f^2(t)\sin^2(kz) \qquad [51]$$

where we have assumed $\delta \gg \Gamma$, because the constraint $N_s \ll 1$ (eqn [46]) holds in this regime as well, where N_s is the expected number of spontaneous emission events per atom during the time τ. Although we neglect the kinetic energy term in the Hamiltonian, we continue to will treat z in eqn [51] as an operator.

Given the initial atomic wavefunction in momentum space, $\psi(p_0)$, the atomic wavefunction immediately after the interaction is given by

$$\psi(p) = e^{-\frac{i}{\hbar}\int^t dt' U(z,t')}\psi(p_0) \qquad [52]$$

$$= e^{-i\frac{\omega_R^2}{2\delta}\tau} e^{i\frac{\omega_R^2}{2\delta}\tau\cos(2kz)}\psi(p_0) \qquad [53]$$

where $\tau = \int dt' f^2(t')$ and the integral is over the interaction duration. With the use of the identity for Bessel functions of the first kind, $e^{i\alpha\cos(\beta)} = \sum_{n=-\infty}^{\infty} i^n J_n(\alpha) e^{in\beta}$, the atomic wavefunction can be written as

$$\psi(p) = e^{-i\frac{\omega_R^2}{2\delta}\tau} \sum_{n=-\infty}^{\infty} i^n J_n\left(\frac{\omega_R^2}{2\delta}\tau\right) e^{i2nkz}\psi(p_0) \qquad [54]$$

$$= e^{-i\frac{\omega_R^2}{2\delta}\tau} \sum_{n=-\infty}^{\infty} i^n J_n\left(\frac{\omega_R^2}{2\delta}\tau\right) \psi(p_0 + 2n\hbar k) \qquad [55]$$

States with $2N\hbar k$ of momentum are populated with the probability:

$$P_N = J_N^2(\theta), \quad N = 0, \pm 1, \pm 2, \ldots \qquad [56]$$

where

$$\theta = \frac{\omega_R^2}{2\delta}\tau = \omega_R^{(2)}\tau \qquad [57]$$

is the pulse area. This leads to a transverse rms momentum of the diffracted atoms that is linearly proportional to the pulse area:

$$p_{\text{rms}} = \sum_{n=-\infty}^{\infty} (n\hbar k)^2 P_n = 2^{1/2}\theta\hbar k \qquad [58]$$

The maximum and minimum bounds on the interaction time, τ, and the amount of momentum transfer to the atoms are discussed in the next section.

Kapitza–Dirac diffraction of atoms was first observed at MIT in 1986. Diffraction of a well-collimated (subrecoil) supersonic atomic beam was observed after passage through the tightly focused waist of a near-resonant standing wave (Figure 11). Significant diffraction into momentum states $|g, \pm 10\hbar k\rangle$ was observed (Figure 12).

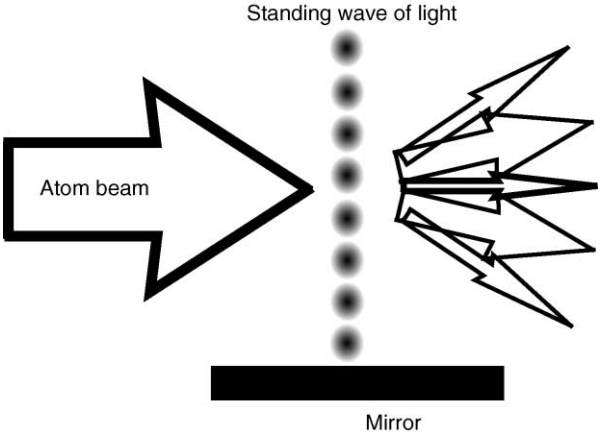


Figure 11 Two counter-propagating running waves of light superimpose to form a standing wave of light. The more narrow beam shown here represents the configuration for Kapitza–Dirac diffraction.

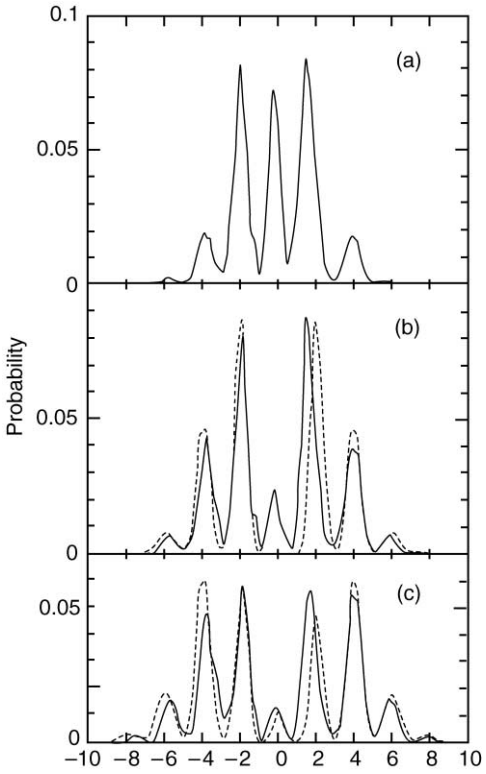


Figure 12 Kapitza–Dirac diffraction of an atomic beam from a standing wavelight grating. The solid lines are experimental data, the dashed lines are theoretical curves. (a) $\theta = 1.69$, (b) $\theta = 2.33$, (c) $\theta = 2.84$. Reprinted with permission from *Physical Review Letters* 56(8): 827–830. Copyright (1986) by the American Physical Society.

Even higher diffracted orders should be observable in the future using laser beams directed at small BECs for somewhat longer times.

Comparison of Atom-Standing Wave Interactions

The atomic motion induced by interaction with a standing light wave varies qualitatively with the interaction parameters. Figure 13 summarizes the various interaction regimes of an atom with a standing wave in terms of the two most basic parameters: the two-photon Rabi frequency, $\omega_R^{(2)} = \omega_R^2/2\delta$, and the duration of the interaction, τ. To render the plot independent of atomic species, $\omega_R^{(2)}$ is given in units of the single-photon recoil frequency, ω_{rec}, and τ is given in units of the inverse single-photon recoil frequency, ω_{rec}^{-1}. Thus, for ^{23}Na the point (1,1) on Figure 13 corresponds to (6.4 μs, $2\pi \times 25$ kHz). The product of the two-photon Rabi frequency and the duration of the interaction is the pulse area, $\theta = \omega_R^{(2)}\tau$ (eqn [57]), which is likewise the product of the pair of coordinates forming a point on the plot.

The lines KD1(KD10) on Figure 13 show where the first maxima of $P_{1,10} = J_{1,10}^2(\theta)$ (eqn [56]) occur, corresponding to the maximum possible population transfer into the first and tenth Kapitza–Dirac diffracted order respectively. Since the momentum distribution of diffracted atoms depends only on the pulse area (θ), all Kapitza–Dirac orders are parallel to each other and further offset from the origin in increasing order number.

The Kapitza–Dirac regime ends at large interaction times where the Raman–Nath approximation fails due to motion of the atoms down the slope of the standing wave potential. We show these lines dashed as the interaction time approaches the beginning of the classical oscillation regime, and terminate these lines where the oscillation produces its first focus, at the first focus line. This line corresponds to the atoms completing a quarter period of oscillation in the standing wave potential, $\tau = \tau_{\text{osc}}/4$, where the oscillation time is derived from the approximation that the potential $\langle V(z,t)\rangle_t$ (eqn [51]) is harmonic

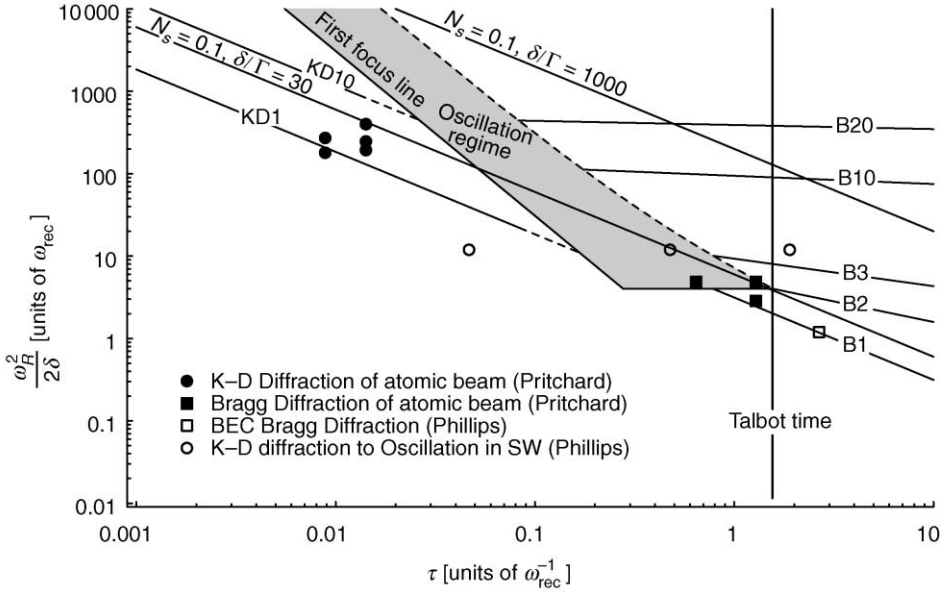


Figure 13 Atomic diffraction from a standing light wave. The vertical axis is the two-photon Rabi frequency, $\omega_R^2/2\delta$, in units of the single-photon recoil frequency, ω_{rec}, and the horizontal axis is the pulse duration, τ, in units of the inverse single-photon recoil frequency, ω_{rec}^{-1}. The scaling is chosen to eliminate the atomic species dependence of the plot. All coherent momentum transfer processes are destroyed by spontaneous decay, which occurs with probability $N_s = 0.1$ along the lines labeled accordingly and parameterized by the given ratio δ/Γ. Lines KD1 and KD10 show conditions for the maximum transfer into the first and tenth Kapitza–Dirac diffracted order respectively. As the interaction time is increased the Raman–Nath approximation is violated (termination of Kapitza–Dirac curves) and the atoms enter the oscillatory regime, executing at least a quarter period of oscillation above the first focus line (shaded area). Cures B1, B2, B3, B10, and B20 correspond to conditions that generate complete Bragg reflection in the first-, second-, third-, tenth-, and twentieth-order respectively. Experimental conditions are shown as points. Filled circles: Kapitza–Dirac diffraction of an atomic beam (Pritchard 1986). Filled squares: Bragg diffraction of an atomic beam (Pritchard 1988). Open squares: Bragg diffraction of a BEC (Phillips 1999). Open circles: transition from Kapitza-Dirac diffraction to oscillation of a BEC in a standing wave light pulse (Phillips 1999). Reprinted from Gupta S, *et al.* (2001) Coherent manipulation of atoms with standing light waves. *Comptes Rendus de L'Academie des Sciences – Series IV – Physics-Astrophysics* 2:479–495, copyright (2001) with permission from Elsevier.

about the minimum:

$$\tau_{\text{osc}} = \frac{\pi}{\omega_R}\left(\frac{|\delta|}{\omega_{\text{rec}}}\right)^{1/2} \qquad [59]$$

In the shaded region of Figure 13, the atoms will oscillate classically about the potential minimum, causing a periodic focusing of the atoms alternately in position and momentum space. However, anharmonicities of the potential away from its minimum will degrade the quality of the focusing effects.

The maximum diffracted order that can be significantly populated by the light–atom interaction is limited by energy conservation. Classically, the maximum momentum transfer to the atoms due to the sudden switch on of the standing wave is delivered to atoms that convert the full height of the standing wave potential ($\hbar\omega_R^2/\delta$), into kinetic energy. This gives

$$\frac{p_{\text{max}}^2}{2m} \simeq \frac{\hbar\omega_R^2}{\delta} \Rightarrow p_{\text{max}} \simeq \omega_R\left(\frac{2\hbar m}{\delta}\right)^{1/2} \qquad [60]$$

Equating the maximum absorbed momentum, p_{max}, to an integer number of two-photon recoil momenta yields the maximum expected diffracted order, N_{max}:

$$p_{\text{max}} = 2N_{\text{max}}\hbar k \Rightarrow N_{\text{max}} \simeq \frac{\omega_R}{2}(\omega_{\text{rec}}\delta)^{-1/2} \qquad [61]$$

This is just half of the square-root of the standing wave potential height measured in units of the single-photon recoil energy:

$$N_{\text{max}} = \frac{1}{2}\left(\frac{\omega_R^2}{\frac{\delta}{\omega_{\text{rec}}}}\right)^{1/2} = \frac{1}{2}\left(\frac{\omega_R^{(2)}}{\omega_{\text{rec}}}\right)^{1/2} \qquad [62]$$

As the two-photon Rabi frequency is reduced, N_{max} falls below unity. As a result, there is no longer time for the higher-order multi-photon processes to generate significant amplitudes in diffracted orders with $N > 1$ before dephasing due to the kinetic energy term (eqn [24]) becomes significant and the higher-order processes become negligible. Only a small population is ever transferred to the $N = \pm 1$ orders and it does not oscillate at ω_{rec}. Therefore the classical oscillation regime does not extend below $N_{\text{max}} \simeq 1$ where it is terminated on Figure 13.

Classical oscillation would result in atoms with momentum from zero up to the maximum allowed by energy conservation. Therefore, the classical focusing of atoms must be manifest as an atomic population distribution over many of the quantized momentum states (separated by $2\hbar k$) allowed by energy conservation (eqn [61]). However, as the interaction time lengthens and extends into the Bragg regime, the populated momentum states become restricted by energy conservation until only the oscillatory Pendellösung into and back from only one final state remains. Therefore, we have ended the classical oscillation regime where the Bragg condition (eqn [44]) is satisfied. The Bragg regime presupposes a smooth light pulse shape. For a pulse with sharp edges, classically oscillatory behavior can still be observed at longer times than included in the shaded region of Figure 13, which is why we show the large τ boundary of the classical oscillation regime as dashed.

In the Bragg regime, transfer of population is restricted to (and back from) only one final momentum state. The allowable final states are restricted by limiting the frequency bandwidth (i.e., energy uncertainty) of the light fields in the atomic rest frame. This is accomplished by lengthening the interaction time and smoothing the rise and fall of the electromagnetic fields. The parameters for a first-order Bragg transition (Table 1) are typically $\delta/\Gamma \simeq 50$, $\Gamma\tau \simeq 1$, and $\omega_R/\Gamma \simeq 10$ giving $N_s \simeq 10^{-2}$ and $\omega_R^{(2)}\tau \simeq 1$. Obtaining significant population transfer with higher-order Bragg processes requires larger intensities. Various orders (1, 2, 3, 10, and 20) of Bragg diffraction are shown as lines on Figure 13 corresponding to $\omega_R^{(2N_B)}\tau = \pi$, where $\omega_R^{(2N_B)}$ is given in eqn [40]. The lines are terminated at the appropriate interaction time determined from the final momentum state resolution condition (eqn [44]). The Bragg regime extends indefinitely to larger interaction times, which might be termed the region of Bragg spectroscopy. In experiments in this regime with ^{23}Na, atomic velocity resolution below 1 mm s^{-1} was obtained at $\tau = 80\omega_{\text{rec}}^{-1}$. The study of adiabatically expanded BECs would require larger interaction times to resolve their smaller velocity spread and weaker mean field shifts.

The Kaptiza–Dirac and Bragg regimes assume a different initial atomic momentum (in the rest frame of the standing wave) parallel to the standing

Table 1 Natural parameters and typical experimental parameters involved in standing wave diffraction. The relevant frequencies together with the corresponding times ($1/\omega$) are tabulated. The parameters for ^{23}Na have been used for system dependent quantities

ω	$\omega/2\pi$	$1/\omega$
ω_0	500 THz	0.3 fs
δ (Bragg)	500 MHz	0.3 ns
ω_R (Bragg)	100 MHz	1.6 ns
ω_R (K–D)	100 MHz	1.6 ns
Γ	10 MHz	16 ns
ω_{rec}	25 kHz	6.4 μs
ω_{osc}	1 MHz	160 ns

wave axis. Kapitza–Dirac scattering assumes no component of the initial atomic momentum along this axis. The efficiency of Kapitza–Dirac diffraction falls rapidly as the initial momentum in units of the photon momentum approaches $1/\omega_{\text{rec}}\tau$. To observe N_B^{th}-order Bragg scattering, the initial atomic momentum along the standing wave axis must have (nonzero) magnitude $N_B\hbar k$. Without adhering to this constraint, no final momentum state will be energetically degenerate with the initial state and the atomic sample will not respond to the presence of the standing wave, even if the interaction parameters are appropriate ($\omega_R^{(2N_B)}\tau \simeq 1$) for N_B^{th}-order Bragg scattering.

The condition $N_s = 0.1$ (eqn [46]), where N_s is the expected number of spontaneous emission events per atom during the light–atom interaction, is drawn on Figure 13 for two ratios (30 and 1000) of standing wave detuning, δ, to the atomic excited state natural lifetime, Γ. For fixed δ/Γ, we see that $N_s \propto \theta$, thus conditions with an increased pulse area (above and to the right of the $N_s = 0.1$ lines) result in a proportionally increased N_s. These $N_s = 0.1$ lines are extended across all regimes of Figure 13 because restricting spontaneous emission is required for both Kapitza–Dirac and Bragg scattering. As N_s approaches unity, the correct optical potential describing the light–atom interaction no longer is eqn [51], and scattering of atoms becomes an incoherent process which can lead to new possibilities such as a complex valued optical potential. Thus, while the standing wave fields in pure Bragg or Kapitza–Dirac scattering behaved as phase gratings for atoms, light gratings on resonance can effectively become amplitude diffraction gratings.

Atom Interferometry

In the 19th century, Fizeau (1853), Michelson (1881), Rayleigh (1881), and Fabry and Perot (Fabry 1899) exploited the interference properties of light waves to create the light interferometer which has since resulted in many beautiful experiments and precise measurements. Using technologies invented since the Second World War, the initial ideas of de Broglie and Schrödinger have evolved into construction of interferometers for neutrons, electrons, and atoms. These new interferometers are proving to be valuable tools for probing fundamental physics, for studying quantum mechanical phenomena, and for making inertial measurements.

The scientific value of interferometry with atoms and molecules has long been recognized. In fact, the concept of an atom interferometer was patented in 1973 and it has been extensively discussed since. Atom interferometry offers great richness stemming from the varied internal structure of atoms, the wide range of properties possessed by different atoms (e.g., mass, magnetic moment, absorption frequencies, and polarizability), and the great variety of interactions between atoms and their environment (e.g., static E–M fields, radiation, and other atoms).

Light interferometers are generally based either on achromatic beamsplitters such as half-silvered mirrors or on other semi-transparent membranes whose structure is small compared to the wavelength of the wave they are splitting. Lacking material structures that are either transparent to atoms or smaller than their de Broglie wavelength, diffraction gratings have been pressed into service both as beamsplitters and mirrors for atom waves. This means that atom interferometers are constrained to designs which somehow compensate for the dependence of diffraction angle on the wavelength of the individual atoms. In spite of this challenge, a surprising variety of atom and molecule interferometers have been built since 1991. A majority have used the three-grating configuration in which the first grating splits the incident beam, the second reverses the differential momenta given by the first, and the third recombines the two beams at the location where they overlap. Both material gratings and standing waves of light have been used in the Raman–Nath, Bragg, and adiabatic regimes to obtain interference fringes. Some designs render the interference fringes in position space, others in internal state space.

Figure 14 shows the MIT setup for a three-grating atom interferometer which features sufficient separation between interfering paths to accommodate an interposed metal foil. This allows different fields or media to act on atom waves only in one of the two

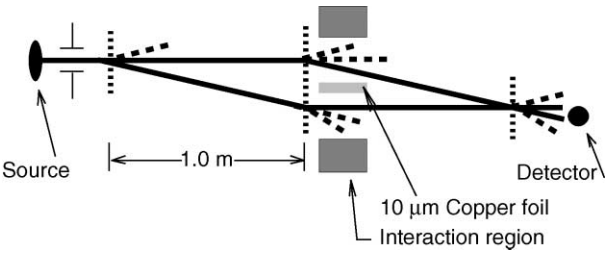


Figure 14 An interferometer for atoms built with three nanofabricated gratings. Each atom propagates in a superposition of two paths, shown in bold. The beam of atoms is well enough collimated, and the diffraction angle is large enough, that a metal barrier can be inserted between the two paths inside the interferometer. The interference fringes are observed by translating one grating transverse to the atom beam. See for example: Berman PR (ed.) (1997) *Atom Interferometry*. San Diego, CA: Academic Press.

arms of the interferometer. Such an interaction region has been used to measure phase shifts due to an electric field applied to only one of the interfering paths. The phase difference between the two paths is then given by the Feynman path integral:

$$\phi = \frac{-i}{\hbar}\int V(x)dt \qquad [63]$$

where

$$V = -\frac{1}{2}\alpha E^2 \qquad [64]$$

is the potential energy due to the electric filed E, and α is the electric polarizability of the atoms. The resulting phase shift of the interference fringes (**Figure 15**) has been used to measure the polarizability of sodium atoms to unprecedented accuracy.

Phase shifts have also been measured due a dilute target gas in one arm of the interferometer. In this case there is a complex index of refraction for the forward scattered atoms (some attenuation, and some phase shift) which depends on matter wave wavelength. Recent measurements reveal glory oscillations in the matter wave index of refraction due to passing through a dilute gas, and are a sensitive probe of interatomic potentials.

In the domain of inertial sensing, atom interferometers have been used to the local gravitational acceleration, $\vec{g}$, Newton's constant, G, gravity gradients, $\nabla\vec{g}$, and rotations Ω, each with sensitivity rivaling if not exceeding any other method. The Sagnac phase shift, due to rotations, for a two-path interferometer is

$$\phi_S = \frac{4\pi}{\lambda_{dB}v}\vec{\Omega}\cdot\vec{A} \qquad [65]$$

where v is the velocity of atoms, $\vec{A}$ is the enclosed area of the interferometer, and $\vec{\Omega}$ is the rotation rate. The two factors in the denominator, λ_{dB} and v, are both $\sim 10^5$ times larger for light than for atoms, which is why atom interferometers have the potential to be 10^{11} times more accurate than optical interferometers. In practice, atom interferometers only outperform laser gyros by a small amount because they have dramatically smaller enclosed areas and particle flux than optical interferometers.

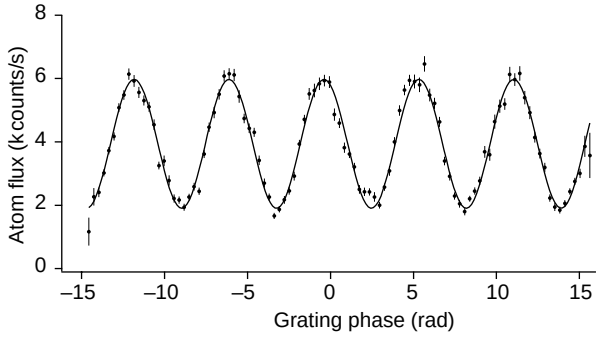


Figure 15 Interference fringes from atoms. 2π of grating phase corresponds to translating the grating one grating period, or 100 nm. These data were obtained at MIT.

Since the first atom interferometers for Na and He^* began working in 1991, others have been made for Li, Ar^*, Ca, Cs, K, Mg, Ne^*, Rb, and Li_2, Na_2, I_2, C_{60} and C_{70} molecules, interferometers starting with trapped atoms have been made for Cs, Ca, He^*, Mg, and Rb. Interferometers using Bose–Einstein condensates have been demonstrated with Na and Rb. These lists are still growing. Diffraction from a single grating has been observed for molecules as large as C_{70} buckyballs, which may someday be used in a three-grating interferometer.

Conclusion

The manipulation of atoms using light forces from standing light waves is a rich subject. The seminal suggestion of Kapitza and Dirac laid dormant for 50 years due to lack of experimental technology. However, in the 1980s this suggestion was realized with atomic sources and considerably extended both experimentally and theoretically. In the early 1990s, coherent standing wave manipulation and nonofabricated atom optics became the two major routes to making atom interferometers. As the new century begins, these techniques are being refined further with improved sources, new detection schemes, and better atom optics. Replacing thermal atomic beams by Bose–Einstein condensates (BECs), as sources of atoms, will revolutionize the field of atom optics just as lasers did in light optics.

Conversely, stimulated light scattering and nanostructures provide new ways to study atomic and molecular properties, such as Bragg spectroscopy for BEC characterization. As progress continues, exciting developments should be forthcoming, such as definitive measurements of the fine structure constant, α, atom gyroscopes that are far superior to the best laser gyroscope technology, and improved studies of BECs. Another area that is on the brink of spectacular development is the confinement of coherent matter waves in atom waveguides and the development of the scientific and technological opportunities that these represent, in analogy to fiber optic waveguides for light.

While we have described the application of nanofabrication techniques to atom optics, it is possible to imagine technology transfer in the other direction. The fundamental problem of fabricating

ever smaller structures might be tremendously advanced by atom optics because the de Broglie wavelength of atoms is so much smaller than that of light, permitting them to be focused to directly deposit much smaller features than possible with photolithography.

See also

Interferometry: Overview. **Scattering:** Scattering Theory.

Further Reading

Cohen-Tannoudji C, Diu B and Laloe F (1977) *Quantum Mechanics*. New York: Wiley and Sons.

Hecht E (1987) *Optics*. Reading, MA: Addison-Wesley.

Berman PR (ed.) (1997) *Atom Interferometry*. San Diego, CA: Academic Press.

Mandel L and Wolf EM (1995) *Optical Coherence and Quantum Optics*. London: Cambridge University Press.

Balykin VI and Letokhov VS (1995) *Atom Optics with Laser Light*. Chur, Switzerland: Harwood Academic Publishers.

Meystre P and Sargent M III (1999) *Elements of Quantum Optics*. Heidelberg, Germany: Springer Verlag.

Rauch H (2000) *Neutron Interferometry*. Oxford, UK: Oxford University Press.

Meystre P (2001) *Atom Optics*. New York: Springer-Verlag.

Campargue R (ed.) (2001) *Atomic and Molecular Beams, the State of the Art 2000*. Berlin: Springer.

Atomic Coherence Effects

A Belyanin, G R Welch and M O Scully,
Texas A&M University, College Station, TX, USA

© 2005, Elsevier Ltd. All Rights Reserved.

History and Perspective

Quantum optics and laser physics have entered a new era. This started with the invention of the laser and the associated theoretical developments such as the theory of photon correlations and statistics, and the quantum theory of the laser. Later, new and novel lasers such as the free electron laser and tunable dye laser fueled many new research directions. With the development of the micromaser, quantum optics became a beautiful new test bed for the study of nonequilibrium statistical physics of photons interacting with matter in a controlled environment. The realization that a deeper understanding was needed to understand the measurement process led to the utilization of the laser as a probe of gravity, and to insight into nonclassical processes such as quantum nondemolition measurements and the production of squeezed states of light.

Effects of atomic coherence in quantum electronics have a long history. One of the first examples was the Hanle effect: excitation by polarized light creates atomic coherence that can evolve in a magnetic field, causing the change of polarization of the resulting fluorescence. Another important example is the interference of decay processes, first suggested by Fano. Modern schemes utilizing quantum coherence in atomic vapors or solid media usually involve strong coupling laser fields (laser drives). In each case, the effect of quantum coherence can be understood as interference of multiple (usually two) transition paths between quantum states of a system.

In the past decade, the concepts of atomic coherence and interference have led to many new and surprising innovations in laser physics such as electromagnetically induced transparency (EIT), refractive index enhancement, lasers without population inversion, ultra-slow light group velocity, storage of the quantum states of light, teleportation of quantum states, and nonlinear optics at an ultra-low intensity level.

Electromagnetically Induced Transparency

One of the most vivid manifestations of quantum coherence effects is the reduction of absorption in a three-level system that is driven by a coherent laser field. This effect is known as electromagnetically induced transparency, or EIT. We consider in more detail the case when the probe and drive fields are in a so-called Λ-configuration (Figure 1a) named after the appearance of the level scheme. In the conventional usage, the two lower levels are coupled to the upper level with two lasers. One of the lasers may be strong and is called the driving or coupling laser, and the other is weak and is called the probe laser.

Consider a quasi-monochromatic laser field E_d (called the 'driving' field, or 'coupling' field) and probe field E_p interacting with a three-level medium of Figure 1a. This interaction is described by a density matrix ρ_{ij}, where i, j take the values a, b, c. Within the

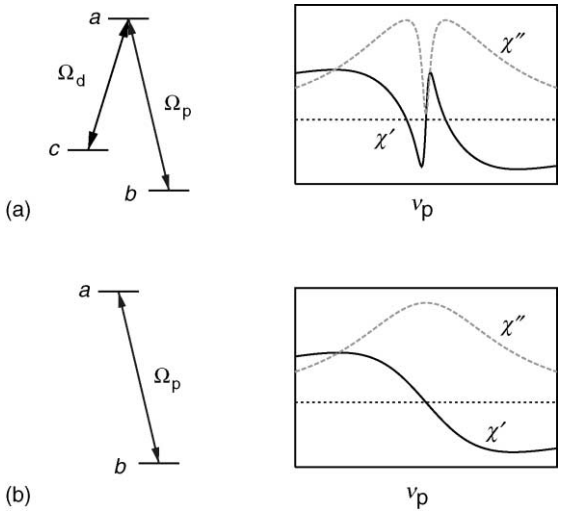


Figure 1 Real and imaginary parts of the susceptibility as functions of the frequency of the probe laser for (a) three-level Λ–type atoms, and (b) two-level atoms.

density matrix approach, quantum coherence in the system has the precise meaning of an off-diagonal element of the density matrix at the two-photon transition c–b. To have strong quantum coherence in a Λ-system, one must excite a large value of ρ_{cb}.

It is convenient to simplify the equations from the very beginning by using the rotating-wave approximation. We introduce slowly varying complex amplitudes of fields and polarizations and neglect their second derivatives. For example, for the probe field applied at the $a \rightarrow b$ transition we write

$$E_{\mathrm{p}}(z,t) = \frac{1}{2}\varepsilon_{\mathrm{p}}(z,t)\exp(-\mathrm{i}\omega_{\mathrm{p}}t + \mathrm{i}k_{\mathrm{p}}z) + \mathrm{c.c.} \qquad [1]$$

After introducing the complex Rabi frequency $\Omega_{\mathrm{p}}(z,t) = d_{\mathrm{p}}\varepsilon_{\mathrm{p}}(z,t)/2\hbar$, the wave equation for waves propagating in the z-direction can be written as

$$\frac{\partial\Omega_{\mathrm{p}}}{\partial z} + \frac{n_{\mathrm{p}}}{c}\frac{\partial\Omega_{\mathrm{p}}}{\partial t} + \kappa_{\mathrm{p}}\Omega_{\mathrm{p}} = \frac{2\pi\mathrm{i}\nu_{\mathrm{p}}d_{\mathrm{p}}^2N}{\hbar c n_{\mathrm{p}}^2}\sigma_{ab} \qquad [2]$$

Here d_{p} is the dipole moment of the probe transition a–b, N the total volume density of atoms in the active region, κ_{p} represents cavity losses, and $n_{\mathrm{p}} = k_{\mathrm{p}}c/\omega_{\mathrm{p}}$ is the refraction index. The variable σ_{ab} is the slowly varying amplitude of the corresponding off-diagonal element $\rho_{ab} = \sigma_{ab}\exp(-\mathrm{i}\nu_{\mathrm{p}}t)$ of the density matrix. The same representation is assumed for the drive field E_{d}, with parameters n, d, ν, κ having index 'd'.

The functions σ_{ab}, σ_{ac}, and σ_{cb}, which determine the polarization response of the medium at the frequency of the external field, should be found from the density matrix equations

$$\frac{\mathrm{d}\sigma_{cb}}{\mathrm{d}t} + \Gamma_{cb}\sigma_{cb} = -\mathrm{i}\Omega_{\mathrm{p}}\sigma_{ac}^* + \mathrm{i}\Omega_{\mathrm{d}}^*\sigma_{ab} \qquad [3]$$

$$\frac{\mathrm{d}\sigma_{ab}}{\mathrm{d}t} + \Gamma_{ab}\sigma_{ab} = \mathrm{i}\Omega_{\mathrm{p}}n_{ba} + \mathrm{i}\Omega_{\mathrm{d}}\sigma_{cb} \qquad [4]$$

$$\frac{\mathrm{d}\sigma_{ac}}{\mathrm{d}t} + \Gamma_{ac}\sigma_{ac} = \mathrm{i}\Omega_{\mathrm{d}}n_{ca} + \mathrm{i}\Omega_{\mathrm{p}}\sigma_{cb}^* \qquad [5]$$

where

$$\begin{aligned}\Gamma_{cb} &= \gamma_{cb} + \mathrm{i}(\nu_{cb} - \nu_{\mathrm{p}} + \nu_{\mathrm{d}}),\\ \Gamma_{ab} &= \gamma_{ab} + \mathrm{i}(\nu_{ab} - \nu_{\mathrm{p}}),\\ \Gamma_{ac} &= \gamma_{ac} + \mathrm{i}(\nu_{ac} - \nu_{\mathrm{d}})\end{aligned} \qquad [6]$$

Here the γ's are relaxation rates of the polarizations at the corresponding transitions (the so-called homogeneous broadening). The population differences $n_{ik} = \rho_{ii} - \rho_{kk}$ are defined from eqns [2]–[6] together with the three equations for $\rho_{ii}, i = a, b, c$, with phenomenological rates of population relaxation and pumping:

$$\begin{aligned}\frac{\mathrm{d}\rho_{bb}}{\mathrm{d}t} &= -2\mathrm{Im}[\Omega_{\mathrm{p}}^*\sigma_{ab}] + r_{cb}\rho_{cc} + r_{ab}\rho_{aa} - r_{ba}\rho_{bb}\\ \frac{\mathrm{d}\rho_{cc}}{\mathrm{d}t} &= -2\mathrm{Im}[\Omega_{\mathrm{d}}^*\sigma_{ac}] + r_{ac}\rho_{aa} - r_{cb}\rho_{cc}\\ \frac{\mathrm{d}\rho_{aa}}{\mathrm{d}t} &= 2\mathrm{Im}[\Omega_{\mathrm{p}}^*\sigma_{ab}] + 2\mathrm{Im}[\Omega_{\mathrm{d}}^*\sigma_{ac}] + r_{ba}\rho_{bb}\\ &\quad - (r_{ab} + r_{ac})\rho_{aa}\end{aligned} \qquad [7]$$

Here we have included only the pumping term r_{ba} between levels b and a for simplicity. Later, we will consider the effect of incoherent pumping on gain, but in this section we put $r_{ba} = 0$. The sum of populations is evidently conserved in this model. Analysis of open systems with, e.g., flow of atoms through a cavity, leads to qualitatively similar conclusions.

Suppose that the probe field is weak, $|\Omega_{\mathrm{p}}|^2 \ll \gamma_{cb}\gamma_{ac}$, and the drive-field intensity is constant in a sample. The complex susceptibility at the probe-field frequency, which is proportional to $\sigma_{ab}/\varepsilon_{\mathrm{p}}$, is given by

$$\chi = \frac{\mathrm{i}d_{\mathrm{p}}^2N}{\hbar c}\left\{\frac{1}{\Gamma_{ab} + |\Omega_{\mathrm{d}}|^2/\Gamma_{cb}}\left(n_{ba} - \frac{|\Omega_{\mathrm{d}}|^2 n_{ca}}{\Gamma_{cb}\Gamma_{ac}^*}\right)\right\} \qquad [8]$$

This expression can be further simplified if there is no pumping to the states $|c\rangle$ and $|a\rangle$, so that we can

assume that $n_{ca} = 0$ and drop the second term in the brackets. The behavior of χ as a function of detuning from resonance $\nu_p = \nu_{ab}$ is shown in Figure 1a. In the absence of the drive field Ω_d, we retrieve the usual two-level susceptibility, shown in Figure 1b.

There are two remarkable features in the susceptibility of Figure 1a. The first one is the transparency window at the point of two-photon resonance. The absorption coefficient at resonance is

$$K'' = \frac{2\pi\nu_p d_p^2 N}{\hbar c(\gamma_{ab} + |\Omega_d|^2/\gamma_{cb})} \qquad [9]$$

Clearly, absorption can be arbitrarily small when $\gamma_{cb} \ll \gamma_{ab}$ and $|\Omega_d|^2 \gg \gamma_{cb}\gamma_{ab}$, since it is proportional to $\gamma_{cb}/|\Omega_d|^2$. For any value of $|\Omega_d|^2$, the absorption is much smaller than that resulting from simple Stark splitting of two Lorentzians. Moreover, in the most interesting case

$$\gamma_{cb}\gamma_{ab} \ll |\Omega_d|^2 \ll \gamma_{ab}^2$$

the Stark splitting is negligible while the deep and narrow EIT feature remains, with width $\gamma_{cb} + |\Omega_d|^2/\gamma_{ab}$.

Note that the intensity of the drive field required for the coherent bleaching of the medium is a factor $r_{ab}/\gamma_{cb} \gg 1$ smaller than the intensity needed for usual saturation bleaching.

EIT is closely related to the phenomenon of coherent population trapping (CPT). This can be illustrated by transforming from states $|a\rangle$, $|b\rangle$, $|c\rangle$ to a new set of states $|a\rangle$, $|B\rangle$, $|D\rangle$, where

$$|D\rangle = (\Omega_d|b\rangle - \Omega_p|c\rangle)/\sqrt{|\Omega_d|^2 + |\Omega_p|^2}$$

$$|B\rangle = (\Omega_d|b\rangle + \Omega_p|c\rangle)/\sqrt{|\Omega_d|^2 + |\Omega_p|^2}$$

It can be easily found that under the condition of two-photon resonance $\nu_p - \nu_d = \nu_{cb}$ the dipole matrix element from $|D\rangle$ to $|a\rangle$ goes to zero. This means that the population in state $|D\rangle$ does not interact with the coherent radiation field and does not contribute to absorption. That is why this state is called a non-absorbing or dark state. Under the condition $|\Omega_d|^2/\gamma_{cb}\gamma_{ab} \gg 1$ practically all the population of states $|b\rangle$ and $|c\rangle$ is optically pumped into the dark state, and only a small fraction $\propto \gamma_{cb}$ remains in the absorbing state $|A\rangle$ and contributes to absorption.

The second remarkable feature of a driven Λ-system is sharp variation of the real part of χ near the resonance, which can be used for manipulation of the group velocity of light, as discussed in the next section.

Slow, Ultra-Slow, Stored, and Frozen Light

Introduction

There are five useful definitions for the velocity of a traveling wave:

- the phase velocity, which is the speed at which the zero crossings of the carrier wave move;
- the group velocity, at which the peak of a wave packet moves;
- the energy velocity, at which energy is transported by the wave;
- the signal velocity, at which the half-maximum wave amplitude moves;
- the front velocity, at which the first appearance of a discontinuity moves.

Although these can all be different, in most cases (linear, passive, dispersive media) the last four coincide, and are usually less than the phase velocity. By using the very steep frequency dispersion in the vicinity of the narrow resonance associated with electromagnetically induced transparency (EIT), it has been shown that the group velocity of light can be reduced by approximately 100 million compared with its velocity in vacuum.

In order to understand the concept of group velocity, it is useful to consider the superposition of two waves of the same amplitude E_1 and E_2 where $E_i = E_0\cos(k_i z - \nu_i t)$ with $i = 1, 2$. Addition of these waves gives rise to the modulation shown in Figure 2:

$$E = E_0(\cos(k_1 z - \nu_1 t) + \cos(k_2 z - \nu_2 t))$$

$$= 2E_0\cos(\Delta k z - \Delta\nu t)\cos(kz - \nu t)$$

where $\Delta k = (k_1 - k_2)/2$, $\Delta\nu = (\nu_1 - \nu_2)/2$, $k = (k_1 + k_2)/2$, and $\nu = (\nu_1 + \nu_2)/2$. This superposition

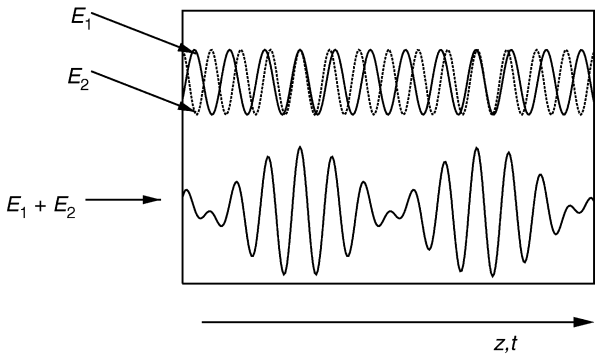


Figure 2 Interference of two monochromatic waves with different frequencies results in a wave modulated in time and space.

displays an interference pattern consisting of a rapid oscillation propagating with the so-called phase velocity,

$$v_{\text{phase}} = \frac{\nu}{k} \qquad [10]$$

and a slowly varying envelope propagating at the group velocity,

$$v_{\text{g}} = \frac{\Delta\nu}{\Delta k} \qquad [11]$$

Usually, the group velocity of wave packets consisting of more than two harmonics is given by transforming the ratio $\Delta\nu/\Delta k$ into $\mathrm{d}\nu/\mathrm{d}k$.

Slow and Ultra-Slow Light

In ordinary absorption and refraction of light, it is sufficient to consider only two atomic energy levels interacting with the light. However, in multilevel systems, coherent excitation of more than one transition can dramatically alter the dispersion and hence the group velocity. The first experiments of this effect demonstrated a group velocity of $c/13$. However, this was progressively lowered in later experiments as the steepness of the dispersion was increased, culminating in a group velocity as low as $c/300\,000\,000$ (1 m/s) in a Bose–Einstein condensate of ultra-cold sodium atoms. Slow light has been observed in a wide range of systems, from the Bose–Einstein condensates just mentioned, to hot optically dense gases of atoms and molecules, and in solids ranging from rare-earth doped crystals to solid hydrogen.

Most of these experiments occurred in a Λ system. The susceptibility of this medium has very steep dispersion without absorption, as shown in Figure 1.

The minimum possible group velocity slowed by the steep dispersion in EIT is given by

$$v_{\text{g}} = \frac{8\pi\Omega^2}{3\lambda^2 N\gamma} > 0 \qquad [12]$$

where $\gamma = \gamma_{ab} = \gamma_{ac}$. The driving field Ω should be strong enough to meet the condition $\Omega^2 \gg \gamma\gamma_{bc}$ for coherent effects. It can be shown that the group velocity approaches its minimum $v_{\text{g min}} = 8\pi\gamma_{bc}/(3\lambda^2 N)$ when $\Omega^2 \approx \gamma\gamma_{bc}$.

Bringing Light to a Halt: Frozen Light

Consider a laser pulse propagating through a medium with ultra-slow group velocity while the medium itself is moving with some velocity v in the opposite direction to the direction of laser pulse propagation. Let us denote the group velocity of the light in the frame co-moving with the atoms as $\tilde{v}_{\text{g}}$. In this frame, atoms are at rest, and hence, there is no spatial dispersion. The Galilean transformation to the laboratory frame, $k = \tilde{k}$, $\nu = \tilde{\nu} - \tilde{k}v$, where v is the atomic velocity, yields the group velocity $v_{\text{g}} = Re(\mathrm{d}\nu/\mathrm{d}k) = \tilde{v}_{\text{g}} - v$. This simple transformation shows that the light pulse is 'dragged' by the moving atoms. If the velocity of the atoms with respect to the laboratory frame is exactly equal and opposite to the group velocity of light in the frame where atoms are at rest, then we find that the group velocity in the laboratory frame is zero. We refer to this as 'freezing' of light. We may even make the group velocity negative so that the propagation direction is opposite to the wavevector (that is, that $v_{\text{g}} < 0$). The simplest example of spatial dispersion is the so-called drift dispersion corresponding to a mono-velocity atomic beam or moving sample with velocity v.

Slow light experiments in Doppler broadened systems such as warm gases allow group velocities that are much lower than the mean thermal speed (v_{T}) of the atoms in the thermal sample (Figure 3). This means that freezing of light can occur without actually moving the atomic sample. When the experimentally obtained group velocity is less than or about equal to the mean thermal speed of the atoms, a pulse of light can be stopped in a stationary cell. In this case the delay time for a pulse passing through the medium would tend to infinity (contrary to the case of a finite atomic beam) in the sense that this pulse would never leave the cell.

This is accomplished by using a single velocity group from the Maxwellian thermal distribution of atomic velocities, and adjusting the frequency of the driving field to be resonant with that velocity group (see Figure 4). In this case it is mainly the atoms in this single velocity group that support the ultra-slow group velocity as a slow EIT polariton, and the atoms

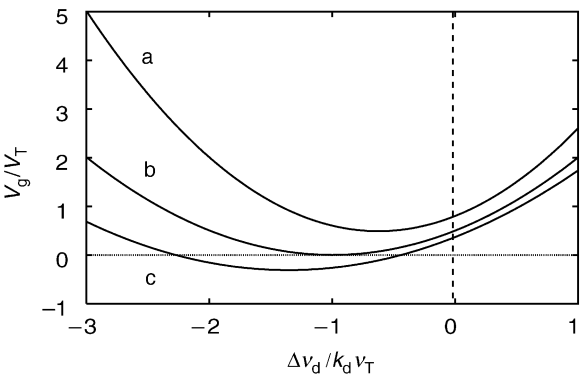


Figure 3 Ultra-slow and negative group velocity of EIT polariton versus detuning of drive laser; $\Omega = 0.25\gamma$, $k_{\text{d}}\nu_{\text{T}} = 100\gamma$, $\gamma_{cb} = 0.001\gamma$, (a) $N = 0.6\,N_{\text{cr}}$; (b) $N = N_{\text{cr}}$; (c) $N = 1.5\,N_{\text{cr}}$.

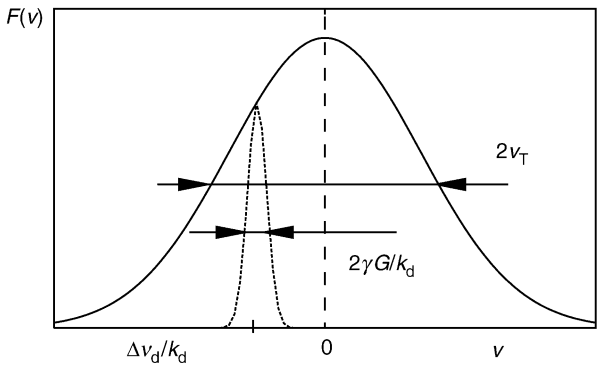


Figure 4 The velocity distribution of atoms in a cell (solid line). Effective drifting beam (dotted) selected by drive laser.

then act as an effective atomic beam. If the driving field is on resonance with the velocity group of atoms moving into the light with speed equal to $\tilde{v}$, we will have $v_g = \tilde{v}'_g - v_d$, where $\tilde{v}'_g$ is the group velocity of the light defined by the effective atom beam if those atoms were at rest. This requires that the intensity of the drive is strong enough to provide EIT for the resonant velocity group of atoms, i.e., $\Omega^2 > \gamma_{bc}\gamma$ (see Figure 4) but at the same time weak enough to avoid interaction with off-resonant atoms moving with the 'wrong' velocities (i.e. $\Omega < kv_T\sqrt{\gamma_{bc}/\gamma}$).

Storing and Retrieving Quantum Information

Photons are potentially the best particles for use in quantum computing and quantum information processing. They are the fastest, and perhaps the simplest and most robust carriers of quantum information, but their storage and manipulation is difficult without destroying their quantum state. However, for application to quantum cryptography, quantum teleportation, and quantum computation it is necessary to achieve a strong coupling between electromagnetic waves and a large nonlinear susceptibility with small losses. For this task, it may be resonant coherent media that offer the promise of success.

Recently, new coherent and reversible methods based on EIT have been found for the transfer and manipulation of photons and their quantum information. Quantum states of photons are trapped in coherently driven atomic media in which the group velocity is reduced adiabatically toward zero. The methods allow for an ideal transfer of quantum correlations between light fields and metastable states of matter. Currently, interesting applications such as quantum state memories and quantum information processing including transporting and time reversing the state of light are being investigated.

Although experiments have not yet been conducted with single quanta, proof-of-principle experiments using classical laser fields have demonstrated the technique. In these experiments control and signal light pulses propagate in a gas of three-level Λ-type atoms and excite a spatial profile of a long-lived coherence of ground spin states. This spin coherence profile contains quantum information about the 'writing' pulses. When the control field is switched back on, the signal pulse is restored via Raman scattering on the atomic coherence. Ideally, the scattered pulse is in the same quantum state as the input signal pulse, preserving entanglement, quantum statistics, etc.

Similar experiments have also shown that different reading pulses (rather than switching back on the same control pulse) can be used to obtain new tools for quantum information storage and processing. Time reversing, teleportation, color switch and multiplexing of the signal pulse of light have all been demonstrated. These new operational possibilities arise as a result of the reading of the stored information with a light pulse that is spatially separated from, has different frequency from, and propagates in the opposite direction to the signal beam.

To demonstrate these effects, one pair of control and signal pulses (the 'writing' fields) prepares a spin coherence of the lower level (Figure 5). A second control pulse (which can be thought of as a 'reading' field) scatters from this coherence resulting in the generation of a new, or 'recovered', field. This recovered field acquires some properties of the reading field because it is a Raman scattered component of the reading pulse. If the reading field pulse is centered about a frequency other than that of the writing fields, and propagates in the opposite direction to the writing fields, then the recovered field propagates in the same direction as the reading field and has a different frequency from the incident writing fields. Even if the writing and reading pulses are separated in space as well as time the scattering effect will persist. This allows the transportation of the coherence grating to another point in the atomic cell via the atomic center-of-mass motion.

In the ideal case the signal pulse profile and quantum statistics are exactly stored by the atomic coherence grating. Thus, reading by a backward propagating pulse results in a generated pulse that is also backward propagating. That is it represents an exactly time-reversed copy of the signal pulse. Similarly, atomic motion in the direction transverse to the light propagation moves the coherence grating to another spatial position. Reading the light by a spatially shifted laser results in the equivalent of

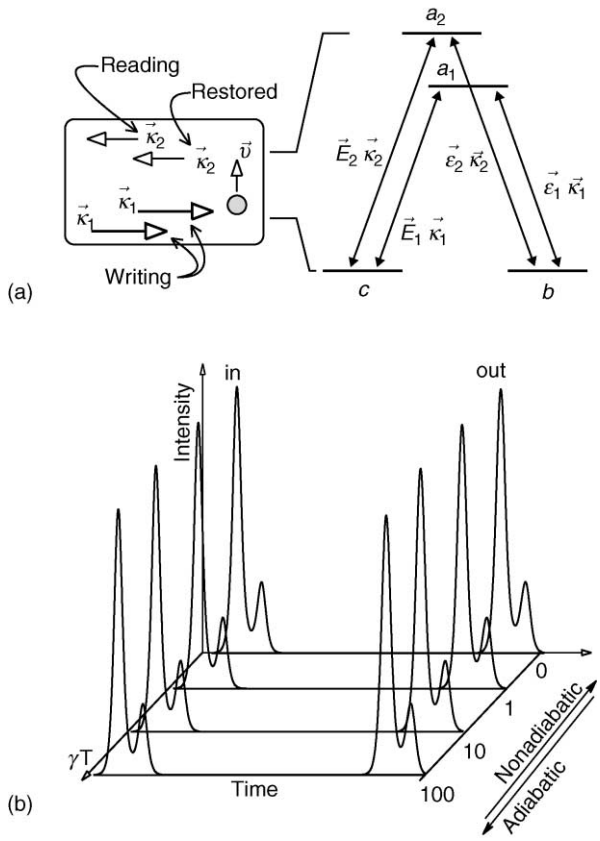


Figure 5 (a) Energy level scheme used in experiments for multiplexing, transporting, and time-reversing stored light. (b) 'Storage of light': intensity of the signal field versus time for different switching times (T) of the reading and writing fields.

'teleportation' of light by atoms in the same spirit as the teleportation of the quantum state of light by atoms that has been discussed recently. It is also possible for the frequency of the 'restored' pulse to be easily shifted by switching the frequency of the reading pulse. This means that the use of a frequency comb of reading pulses propagating in different spatial modes should result in multiplexing of the signal light.

Novel Laser Sources Based on Quantum Coherence

Introduction

In the previous sections we have seen how quantum coherence can be used to overcome resonant absorption on a transition of interest, leading to EIT. A natural question arises whether it is possible to achieve gain without population inversion at this transition, i.e. lasing without inversion (LWI). This subject has been intensively investigated since

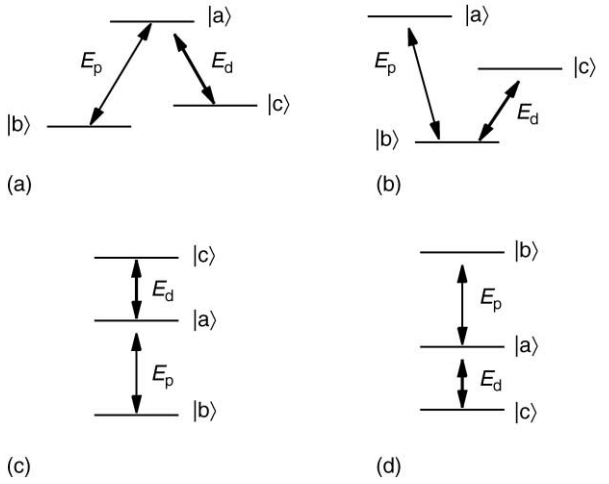


Figure 6 LWI schemes in a three-level system coherently driven by a laser field.

the early 1990s, and the first proof-of-principle experiments have been successful.

Besides the fundamental physical beauty of this phenomenon, the research effort has been motivated by potentially important applications. It is clear that LWI could be particularly useful in situations where a population inversion is difficult to maintain due to, e.g., very fast relaxation or inefficient pumping. The obvious examples are high-frequency UV, X-ray and gamma-ray lasers, or semiconductor mid/far-infrared lasers based on intersub-band transitions. Another promising application of LWI is efficient frequency up- or down-conversion. Furthermore, if resonant absorption can be eliminated, resonant nonlinear optical interactions become a reality. Since all nonlinear coefficients are maximized at resonance, highly efficient nonlinear optical generators of light can be created.

In this section we discuss the simplest possible schemes of inversionless lasing which allow one to clarify the physical origin of LWI and to see the shortcomings which have so far prevented LWI-based schemes from reaching the level of practical laser devices.

Consider a three-level system, coherently driven by a strong quasi-monochromatic field nearly resonant with one transition, and probed by a field that is nearly resonant with an adjacent transition. Some possible configurations are shown in Figure 6. We consider in detail the case when the two fields are in a Λ-configuration (Figure 6a). This is probably the most popular and well-studied scheme since it can provide probe-field gain for a very low pumping rate, when most of the population is at the lowest energy level. However, we will see that the gain

condition for the Λ-scheme is actually rather restrictive and does not allow one to realize large frequency up-conversion. For the latter purpose, ladder-type schemes (Figure 6c,d) turn out to be more favorable.

Weak-field gain in the Λ-scheme

From the general formalism presented earlier, it is straightforward to get the amplification condition

$$K'' = \mathrm{Re}\left\{\eta_\mathrm{p}\frac{1}{\Gamma_{ab}+|\Omega_\mathrm{d}|^2/\Gamma_{cb}}\left(n_{ba}-\frac{|\Omega_\mathrm{d}|^2 n_{ca}}{\Gamma_{cb}\Gamma_{ac}^*}\right)\right\}+\kappa_\mathrm{p} < 0 \qquad [13]$$

where

$$\eta_\mathrm{p} = \frac{2\pi\omega_\mathrm{p} d_\mathrm{p}^2 N}{\hbar c n_\mathrm{p}^2}$$

The physical meaning of the two terms in brackets in eqn [13], which originate from the two terms on the right-hand side of eqn [4] for σ_{ab}, can be easily understood. The first term is just the resonant one-photon absorption proportional to the population difference n_{ba}. It can be strongly reduced in the presence of a drive field Ω_d, as explained earlier. However, it can change sign and give rise to the field amplification only if $n_{ba} < 0$, meaning there is a population inversion!

The second term, originating from the product $\Omega_\mathrm{d}\sigma_{cb}$ in eqn [4], is due to the mixing of the drive field Ω_d and polarization σ_{cb} where the latter is parametrically excited by a two-photon process: $\sigma_{cb} \propto e_\mathrm{p}\sigma_{ac}^* \propto e_\mathrm{p}\Omega_\mathrm{d}^* n_{ca}$. It is this term that can overcome one-photon absorption to provide amplification without inversion in the Λ-scheme. As we see, the possibility of LWI depends crucially on our ability to excite a sufficiently large polarization σ_{cb} at the transition resonant with the beat frequency of the probe and drive fields. This conclusion is universal for all LWI schemes with a coherent drive. We see that the sometimes loosely defined term 'quantum coherence' has a precise meaning here as the polarization σ_{cb} at a two-photon transition $c-b$. It is clear that the effects of quantum coherence are maximized when γ_{cb} is small, i.e., the transition $c-b$ is long-lived.

We are of course interested in the case of weak incoherent pumping. The typical dependence of gain on detuning from two-photon resonance is shown in Figure 7. At resonance, where the gain is maximized, the amplification condition $K'' < 0$ yields (in the absence of nonresonant losses):

$$r_{ac} > 2\gamma_{cb}(r_{ab}/r_{ba} - 1) \geq r_{ab} - r_{ba} \simeq r_{ab} \qquad [14]$$

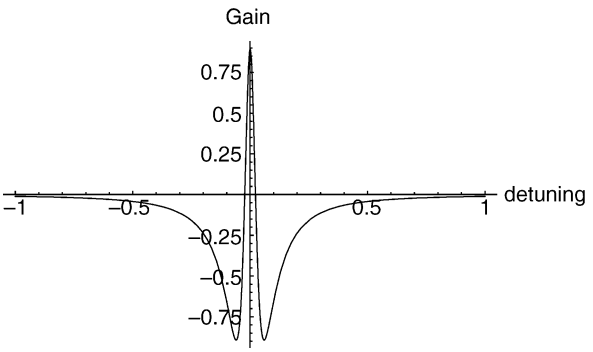


Figure 7 Gain without inversion in Λ-scheme.

because $\gamma_{cb} \geq r_{ba}/2$. Here we assumed that the incoherent pumping is weak, $r_{ba} \ll r_{ab}$, so that it does not produce population inversion at the probe transition.

Equation [14] is a simple and very important result, which is valid for all three-level laser schemes in the absence of inversion at both probe and drive transitions: the relaxation rate at the drive transition should exceed that at the probe transition. This result is an immediate consequence of the fact that the long-lived coherence σ_{cb} is inevitably destroyed by incoherent pumping, and γ_{cb} is now bounded from below. One can follow how the EIT regime is transformed into lasing as follows. With increasing incoherent pumping rate r_{ba}, EIT is actually destroyed due to the growth of $\gamma_{cb} \simeq r_{ba}/2$. In terms of CPT states, this means increase of population of the absorbing state $|B\rangle$. At the same time, the resonant Raman term in [13] also grows since n_{ac} is proportional to r_{ba}. With increase in the incoherent pumping rate, the contribution to gain overcomes the contribution to absorption if $r_{ac} > r_{ab}$.

Propagation effects

Inequality [14] has important implications for the problem of co-propagation of probe and drive fields. Indeed, to provide a significant amplification of the probe field, the absorption length $L_\mathrm{d} = \gamma_{ac}/(\eta_\mathrm{d} n_{ca})$ of the drive field should be larger than the amplification length $L_\mathrm{p} \simeq \gamma_{ac}/(\eta_\mathrm{p} n_{ca})$ of the probe. This requirement gives simply

$$\eta_\mathrm{p}/\eta_\mathrm{d} > 1 \qquad [15]$$

which in the absence of nonradiative relaxation can be rewritten as

$$\frac{r_{ab}}{r_{ac}}\frac{\nu_d^2}{\nu_p^2} > 1$$

In view of the gain condition given in eqn [14], the above requirement can be satisfied only if $\nu_d > \nu_p$, i.e., up-conversion is impossible.

An interesting solution to this problem is to realize population inversion at the driving transition. Then up-conversion can be realized with a lower relaxation rate at the drive transition as compared with probe transition. In this case in the Λ and V-schemes the two-photon term contributes to absorption at resonance, and gain is only possible far off-resonance. At the same time, in the ladder schemes shown in Figure 6c and d the gain is achieved at resonance, as discussed below in more detail.

It should also be noted that to excite laser oscillations in a high-Q cavity, large one-pass gain is not required at all, so that a Λ-scheme could still be employed for up-conversion.

Now, consider a probe field of arbitrary intensity. An important question is where the energy for amplification of the probe field comes from. It is known that for parametric processes in a transparent medium the Manley–Rowe relations are satisfied. Therefore, it might seem that the drive field is also the ultimate energy source for LWI. However, this is the case only for very large detunings of the drive and probe fields from the corresponding transition frequencies. In that case, we can neglect one-photon absorption, and the wave equations take the form typical for stimulated Raman scattering. They possess the first integral

$$\frac{|\varepsilon_d|^2}{\nu_d} + \frac{|\varepsilon_p|^2}{\nu_p} = \text{const} \qquad [16]$$

which expresses the conservation of the total number of quanta in both waves (Manley–Rowe conditions).

However, in the case of exact resonance, one can show that one-photon absorption of the drive field is always important, and the Manley–Rowe-type conservation law does not exist at all. The optimal condition for the probe field amplification is $\eta_p/\eta_d \gg 1$. In this case the stage of exponential amplification of the probe is maintained over a long distance, and the probe field intensity can grow to much larger values than the sum of initial intensities of the drive and probe fields. Apparently, the energy for the probe field amplification comes mainly from the medium, namely from the upper-level population ρ_{aa} supported by incoherent pumping.

V-Scheme

The V-scheme, shown in Figure 6b, is qualitatively similar to the Λ-scheme. An important difference is that there is now no long-lived coherence at the mixing transition c–a. Indeed, even if the transition $a \to c$ is strongly forbidden, the coherence γ_{ac} is destroyed by fast relaxations at the c–b and a–b transitions. We can immediately suggest that for the V-scheme inversionless lasing takes place at large driving field intensities exceeding the saturation value. However, one must be careful: although it appears that the reduction of absorption in the absence of pumping cannot be greater than a factor of 2, this turns out to be incorrect. Rather, almost complete transparency can be achieved, as confirmed by calculations and recent experiments. Analysis of the laser gain condition shows that the V-scheme shares with Λ-scheme the same negative feature: relaxation at the driving transition should be much faster than at the probe transition.

Schemes with Population Inversion at the Driving Transition

The above analysis illustrates the common problem of LWI schemes: the requirement of rapid relaxation at the driving transition and the resulting strong absorption of the drive field. This difficulty can be overcome in schemes with population inversion at the driving transition. Then, of course, all this makes practical sense only if the relaxation at the drive transition is much slower than that on the probe transition, so that it is easy to provide population inversion for the drive transition as compared with the probe transition. Luckily, such schemes do exist in principle, and seem to have many potential applications. The simplest possibility is the so-called ladder (or cascade) schemes shown in Figure 6c,d. LWI in ladder schemes is usually studied in the regime when there is no inversion at both probe and drive transitions. In this case gain is achieved at large detunings from two-photon resonance and under the same unfortunate condition of fast relaxation at the drive transition as compared with the probe transition. In the opposite case, population inversion at the driving transition is required for LWI at the probe transition. In ladder schemes the maximum gain of the probe field is achieved when both fields are at resonance with atomic transitions, while in V and Λ-schemes the resonance point corresponds to maximum absorption, and the gain is achieved in sidebands.

Note that population inversion at the driving transition means that the drive field can be amplified. This immediately puts forward the possibility to use a drive field which is self-generated in the same active medium. In other words, the system supports lasing with inversion at the drive frequency and LWI at the probe frequency. This situation is especially beneficial for dense gases or solids where the problems of

external drive absorption, reflection, and inhomogeneity may become otherwise insurmountable. One specific example considered in the literature is LWI on rapidly decaying UV and soft X-ray transitions in gas lasers in the presence of lasing with inversion on long-lived adjacent transitions in the visible range (as in Figure 6c). Another example is LWI on rapidly decaying mid/far-infrared intersub-band transitions in semiconductor quantum wells in the presence of lasing with inversion on the interband transition, as in Figure 6d. These interband transitions have 10^4 times longer lifetime than the intersub-band transitions.

Experiments

Experimentally, amplification without inversion has been observed in rarefied atomic vapors in the early and mid-1990s. In particular, amplification of a probe beam in the Λ-configuration was realized in atomic samarium, sodium, and cadmium, while the V-scheme was successfully implemented in atomic rubidium. Inversionless laser oscillation was observed in a V-scheme in rubidium vapor and a Λ-scheme in a sodium atomic beam. In all of these proof-of-principle experiments, the frequencies of the probe and drive fields were very close to each other, and the coherence was excited between two hyperfine states or Zeeman sublevels. One-pass gain of a few per cent was observed.

Nonlinear Optical Sources Based on Quantum Coherence

Excitation of a strong coherence has a profound impact on nonlinear optical interactions. Strong scattering on the polarization wave excited at the beat frequency of two or more laser fields allows one to achieve huge nonlinear coefficients and conversion efficiencies. A particularly impressive example is ultra-broadband collinear Raman generation in a resonant Λ-scheme, that has been observed experimentally. A pair of strong laser pulses propagating in molecular deuterium excites a strong polarization wave at a beat frequency close to a vibronic transition in deuterium. As a result, a comb of Raman sidebands from the infrared to the deep-UV region of the spectrum has been observed.

Another example is an extremely efficient simultaneous blue light and mid-infrared generation in a four-wave mixing experiment in rubidium vapor. Here, all four fields are resonant with corresponding transitions.

A final example illustrates the use of self-generated drive fields for nonlinear generation. It has been suggested that one can use two intracavity generated laser modes in a semiconductor quantum-well laser for the difference frequency, sum frequency, and second-harmonic generation in fully resonant three-wave mixing processes. This idea has already been successfully realized in two-wavelength quantum-cascade lasers, and experiments with other types of semiconductor lasers are under way.

The whole field of resonant nonlinear optics, stemming from original ideas of LWI in systems with quantum coherence, is now in explosive development.

Summary

Atomic coherence effects have already found many important applications in metrology and low-intensity nonlinear optics. Novel applications in quantum nonlinear optics, laser physics, and quantum information processing seem to be feasible. The coherence allows the storage of information about the probe light, its transportion in space, time reversing, and multiplexing the light. Moreover, the coherence allows an increase of the coupling between light fields such that it becomes possible to study the interaction between single photons. New coherent radiation sources based on the ideas of LWI and coherent nonlinear optics are at the stage of active research. There are surely many other applications ahead.

See also

Electromagnetically Induced Transparency. **Non-Classical Light**. **Nonlinear Optics, Applications:** Raman Lasers. **Nonlinear Optics, Basics:** Cascading; Four-Wave Mixing. **Quantum Optics:** Entanglement and Quantum Information.

Further Reading

Belyanin AA, Bentley C, Capasso F, *et al.* (2001) Inversionless lasing with self-generated driving field. *Physical Review A* 64: 013814.

Harris SE (1997) Electromagnetically induced transparency. *Physics Today* 50(9): 36–42.

Kocharovskaya O (1992) Amplification and lasing without inversion. *Physics Reports* 219: 175–190.

Mompart J and Corbalán R (2000) Lasing without inversion. *Journal of Optics B: Quantum and Semiclassical Optics* 2: R7–R24.

Rostovtsev Y, Kocharovskaya O, Welch GR and Scully MO (2002) Slow, ultra-slow, stored and frozen light. *Optics Photonics News*, June.

Scully M and Zubairy S (1997) Quantum Optics. Cambridge, UK: Cambridge University Press.

Entanglement and Quantum Information

P G Kwiat, University of Illinois at Urbana-Champaign, Urbana, IL, USA

D F V James, Los Alamos National Laboratory, Los Alamos, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Quantum information is an emerging field of technology that encompasses the application of fundamental quantum mechanical phenomena, such as entanglement, to tasks in information processing and communications. The importance of optical technology in quantum information is not surprising, given the success of quantum optics in the investigation of these fundamental phenomena and the pre-eminent role of optics in modern communications. This article describes some of the important advances in photonic quantum information and discusses the prospects for the future.

Basic Notions: Qubits and Entanglement

Qubits

Quantum information is usually confined to two-level quantum systems, which are referred to as quantum bits, or qubits. The qubits form the register of a quantum computer or carry the information in a quantum communications channel. Physical examples of qubits under active study include the spin degrees of freedom of an electron or spin-1/2 nucleus, two well-isolated energy levels of an atom or trapped ion, and the polarization degrees of freedom of a photon; here we concentrate on the latter.[1] The two states of each qubit are generically denoted by $|0\rangle$ and $|1\rangle$ (although, because of our concentration on optical polarization, we shall use the horizontal and vertical states $|H\rangle$ and $|V\rangle$ interchangeably with $|0\rangle$ and $|1\rangle$ when describing experiments). An arbitrary state of a single qubit, $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, is specified by the complex probability amplitudes α and β associated with the two possible states, constrained by $|\alpha|^2 + |\beta|^2 = 1$. States for which α and β are deterministic quantities are referred to as pure states; when they are stochastic, the state is mixed, and must be described by a density matrix, e.g., $\rho = \sum_i p_i|\psi_i\rangle\langle\psi_i|$, where p_i is the probability of the pure state $|\psi_i\rangle$. Pure states are a useful but unattainable ideal, just like a perfectly coherent laser beam; in reality even the best-prepared quantum states have some level of mixedness, which can be quantified by the entropy: $S \equiv -Tr(\rho \ln \rho) = -\sum_i \lambda_i \ln \lambda_i$, where λ_i are the eigenvalues of the matrix ρ.

Multiqubit States and Entanglement

For two qubits, the possible pure states of the system are specified by four probability amplitudes:[2] $|\psi_2\rangle = \alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle$, with $|\alpha|^2 + |\beta|^2 + |\gamma|^2 + |\delta|^2 = 1$, and, e.g., $|00\rangle \equiv |0\rangle \otimes |0\rangle$ denoting the state in which the first and the second qubits are both in state $|0\rangle$. In general, 2^n amplitudes are needed to characterize a pure state of n qubits. Immediately we can see one advantage of quantum systems for information storage: if each distinct probability amplitude is regarded as a data register, the size of the memory grows exponentially with the number of qubits. Further, if one were to flip just one qubit, for example the second, the state would be transformed into $|\psi'_2\rangle = \beta|00\rangle + \alpha|01\rangle + \delta|10\rangle + \gamma|11\rangle$, where all of the amplitudes have been affected. This simple example demonstrates the notion of quantum parallelism, one of the most powerful properties of quantum information processors.

If the two-particle state can be written as a product of two single-particle states, i.e., if $\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle = (A|0\rangle + B|1\rangle) \otimes (C|0\rangle + D|1\rangle)$, then these two particles can be considered as separate, unconnected entities: the state is said to be separable.[3] When the state cannot be written in this form, it is called an entangled state. Entanglement is one of the most fascinating fundamental properties of quantum systems: Erwin Schrödinger described it as '*the* characteristic trait of quantum mechanics … that enforces its entire departure from classical lines of thought.' Consider performing a measurement on the second of the two qubits, to establish which of its two states the particle was in, leaving the first qubit

[1] Recently some attention has been devoted to quantum information based on continuous variables, such as those arising in connection with 'squeezed light'. However, here we will focus on discrete systems, i.e., qubits.

[2] Qubits are assumed to be distinguishable particles, and so the wavefunctions describing their states need not be symmetrized under exchange of particles.

[3] When mixture is also present, the state is separable if it can be written as $\rho = \sum_i p_i \rho_i^A \otimes \rho_i^B$, i.e., as a sum of product states for systems A and B.

untouched. The result would collapse the state of the first qubit into one of two possibilities: either $|\phi_0\rangle = (\alpha|0\rangle + \gamma|1\rangle)/\sqrt{P_0}$ if the outcome of the measurement were 0 (which occurs with probability $P_0 = |\alpha|^2 + |\gamma|^2$) or $|\phi_1\rangle = (\beta|0\rangle + \delta|1\rangle)/\sqrt{P_1}$ if it were 1 (with probability $P_1 = |\beta|^2 + |\delta|^2$). Thus the state of the first qubit is instantaneously projected into a specific state by performing a measurement on the second qubit, which can be in a completely different location. This example demonstrates the nonlocality of quantum mechanics, which has been confirmed experimentally in some depth by a number of elegant quantum optics experiments over the last 30 years.

The fidelity $F = |\langle\Phi|\Psi\rangle|^2$ characterizes the overlap of two states.[4] The fidelity of $|\phi_0\rangle$ and $|\phi_1\rangle$ is a way to quantify the amount of entanglement in the initial two-qubit state. If the final two states are identical, the fidelity will be unity, and the initial state is separable; conversely if the two final states are orthogonal and occur with equal probability, then in a sense the initial state is maximally entangled, since the largest possible nonlocal influence would occur due to the measurement. Mathematically, the fidelity of the two final states is $|\langle\phi_0|\phi_1\rangle|^2 = 1 - C^2/4P_0P_1$, where $C = 2|\alpha\delta - \beta\gamma|$ is called the concurrence of the state. If $C = 0$, the state is separable; if $C = 1$, the state is maximally entangled.[5]

As discussed below, entanglement forms the heart of a number of quantum information protocols, such as dense coding, teleportation, and one type of quantum key distribution (also known as quantum cryptography), and large-scale entangled states of many qubits seem to be a requirement for the more ambitious goal of practical quantum computing.

Creating Entangled States Experimentally

Entangled states can currently be created in a controlled manner using technologies such as ion traps, cavity quantum electrodynamics, and optical spontaneous parametric down-conversion (SPDC). Down-conversion is a nonlinear optical process by which an incident 'pump' photon can be split (or down-converted) into a pair of longer-wavelength daughter photons (historically called 'signal' and 'idler') in a crystal possessing a $\chi^{(2)}$ nonlinearity, such as beta-barium borate (BBO). Mathematically, the process is described using the creation and annihilation operators of the field modes $(\hat{a}_\lambda^\dagger, \hat{a}_\lambda)$, thus:

$$|\Phi_{\text{out}}\rangle = \exp\left[-i \sum_{p,s,i} c_{p,s,i} (\hat{a}_p^\dagger \hat{a}_s \hat{a}_i + \hat{a}_i^\dagger \hat{a}_s^\dagger \hat{a}_p) \right] |\Phi_{in}\rangle$$

where the triple sum is over all field modes (subscripts p, s, and i refer to pump, signal, and idler modes, respectively) and $c_{p,s,i}$ is a coefficient linearly dependent on the second-order nonlinear susceptibility $\chi^{(2)}_{ijk}$ and also on the birefringent properties of the crystal. Further, $c_{p,s,i}$ will be negligibly small unless both total photon energy and momentum are conserved (i.e., $\omega_p = \omega_s + \omega_i$ and $\vec{\kappa}_p = \vec{\kappa}_s + \vec{\kappa}_i$, where $\vec{\kappa}$ is the intracrystal momentum).

One particularly efficient method for using this phenomenon to create entangled states is as follows. For a specific geometry (type-I phase matching), the daughter photons emerge from the crystal with identical polarizations (perpendicular to the parent polarization and the crystal optic axis) on opposite sides of a cone that is centered about the pump beam (Figure 1a). Because each photon is in a definite state of polarization, the two photons are not in an entangled state. Two crystals, aligned with their axes of symmetry oriented at 90° to each other, as shown in Figure 1b, can then be used to create an entangled state.[6] With crossed crystals, two processes are possible: the parent photon can down-convert in the first crystal to yield two vertically polarized photons, or it can down-convert in the second crystal to yield two horizontally polarized photons. Because it is impossible to distinguish which of these processes has occurred, the state of the daughter photons is a coherent quantum-mechanical superposition of the states that would arise from each crystal alone; the output of the crossed crystals is photons in the maximally entangled state $|\Phi_+\rangle \equiv \frac{1}{\sqrt{2}}(|HH\rangle + |VV\rangle)$.[7]

Figure 2 shows how this basic source can be adapted to produce any pure quantum state of two photons by placing rotatable half- and quarter-wave plates (which can be used to transform the

[4] The generalized fidelity between mixed states ρ_A and ρ_B is $F = |Tr\sqrt{\sqrt{\rho_A}\rho_B\sqrt{\rho_A}}|^2$.

[5] The concurrence can also be generalized to mixed states, although it has a much more complicated expression. The quantification of entanglement for mixed states with more than two subsystems is presently an active area of research.

[6] Another widely used method employs type-II phase matching in a single crystal. The down-conversion photons are emitted with perpendicular polarizations along a pair of cones. Pairs emitted along particular directions are in the maximally polarization-entangled state $|\psi_-\rangle = \frac{1}{\sqrt{2}}(|HV\rangle - |HV\rangle)$. Such a source was used in the first demonstrations of superdense coding and quantum teleportation.

[7] Note that in an arbitrary basis $|\theta\rangle \equiv \cos\theta|H\rangle + \sin\theta|V\rangle$ and $|\theta^\perp\rangle \equiv -\sin\theta|H\rangle + \cos\theta|V\rangle$, this maximally entangled state has the form $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|\theta\theta\rangle + |\theta^\perp\theta^\perp\rangle)$, demonstrating that the nonlocal correlations are present regardless of the bases used to represent the state.

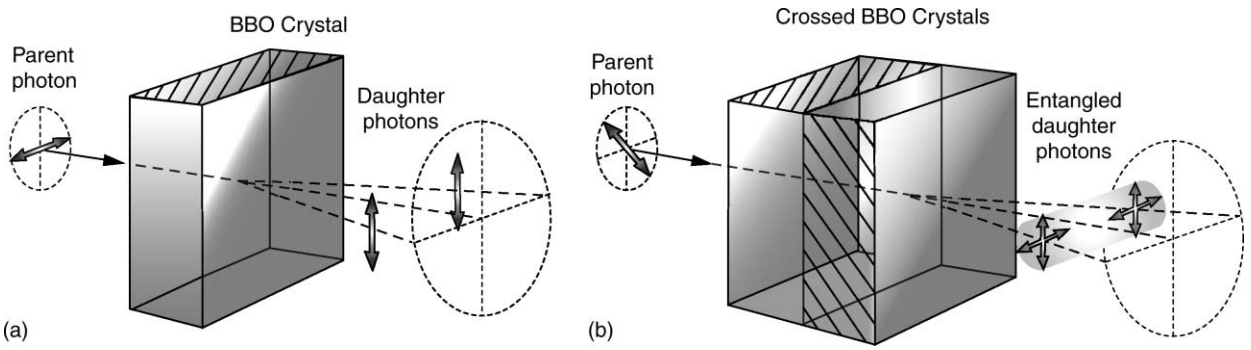


Figure 1 An entangled-photon source. (a) For a given orientation of the nonlinear crystal, a horizontally polarized parent photon produces a pair of vertically polarized daughters. The daughters emerge on opposite sides of an imaginary cone. The cone's axis is parallel to the original direction taken by the parent photon. The two daughter photons are not in an entangled state. Reorienting the BBO crystal by 90° will produce a pair of horizontally polarized daughters if a vertically polarized pump beam is used. (b) Passing a photon polarized at 45° through two crossed BBO crystals can produce two photons in an entangled state. Because of the Heisenberg uncertainty principle, there is no way to tell in which crystal the parent photon 'gave birth,' and so a coherent superposition of two possible outcomes results: a pair of vertically polarized photons or a pair of horizontally polarized photons. (Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.)

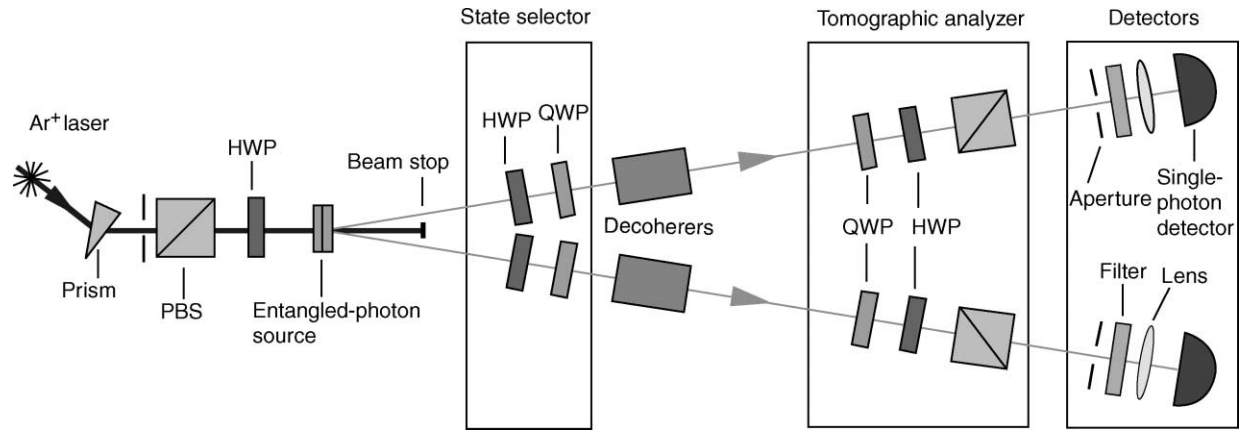


Figure 2 Creating and measuring two-photon entangled states. The 'pump' photons are created, for example, in an argon ion laser and are linearly polarized with a polarizing beamsplitter (PBS). The half-wave plate (HWP) rotates the polarization state before the photon enters the pair of nonlinear crystals that constitute the entangled-photon source; the initial angle of pump polarization controls the entanglement of the pair produced. Each photon's polarization state can be altered at will by the subsequent HWP and quarter-wave plate (QWP). The decoherers following the state selection allow the production of mixed photon states. The optical elements (QWP, HWP, and PBS) in the tomographic analyzer allow the measurement of each photon in an arbitrary basis. Coincidence measurements of photons allow the quantum state to be determined. (Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.)

polarization state of a single photon) before the crystal pair and in the paths of the two daughter photons. To create mixed states, a long birefringent crystal can be used to delay one polarization component with respect to the other. If the delay is longer than the coherence time of the photons, the horizontal and vertical components are effectively decohered; that is, the phase relationship between the different states is destroyed.

The figure also shows schematically the apparatus required for measuring the quantum state. In classical optics, four Stokes parameters are required to specify the polarization of a single beam (i.e., an ensemble of uncorrelated photons); for a pair of photons, 16 projective measurements, each with different wave plate settings, is required. From these 16 measurements, all of the elements of the 4×4 density matrix describing the (in general, mixed) state of the photon pairs can be deduced. This is an example of quantum state tomography, a technique that has found application to a number of quantum optical systems.

Quantum Key Distribution

Two parties, historically known as Alice and Bob, want to have a secret conversation. A generic, classical encryption protocol would begin when Alice and Bob convert their messages to separate binary streams of 0's and 1's, which are then encrypted and decrypted with a set of secret 'keys' known only to the two. Each key is a random string of 0's and 1's that is as long as the binary string comprising each message. To encrypt, Alice (the sender) sequentially adds each bit of the key to each bit of her message, using addition modulo 2. She then sends the encrypted message over a public channel to Bob, who decrypts it by simply repeating the addition modulo 2 of the key to the message. This type of encryption, known as a one-time pad, is currently the only provably secure encryption protocol. But the one-time pad is effective only if Alice and Bob never reuse the key, and more obviously, if the key remains secret. A potential eavesdropper, Eve, cannot be allowed to glean any part of the bit stream that makes up the key. Therein lies a central problem of cryptography – how can secret keys be created and then securely distributed?

Quantum key distribution (QKD) exploits the fundamentally indivisible nature of photons to perform this task. There are a variety of QKD protocols; here we describe one that employs entangled photon pairs (see Figure 3). Alice and Bob use a source such as the one described above to produce maximally entangled photons in the state $|\Phi_+\rangle \equiv \frac{1}{\sqrt{2}}(|HH\rangle + |VV\rangle) = \frac{1}{\sqrt{2}}(|45°,45°\rangle + |-45°,-45°\rangle)$. One photon goes to Alice and the other to Bob. For each pair, Alice and Bob randomly and independently analyze their respective photons (using a polarizing beamsplitter) in the H/V or the 45°/−45° basis. They record a bit value of 0 for all H or 45° results, and a 1 for all V or −45° results. After a sufficient number of measurements (dictated by the length of the key), Alice and Bob have a public discussion, e.g., over the internet. For each detected photon, they announce which basis they used for the measurement, but not the actual measurement result. Whenever they made the same basis choice (50% of the time), the correlations of the entangled state ensure their measured bit values agree. By contrast, they discard the results when they used different bases, because their measurements are completely uncorrelated (see Table 1).

An eavesdropper (Eve) cannot tap the line, as she might with conventional communications, because of the indivisibility of individual photons and the fact

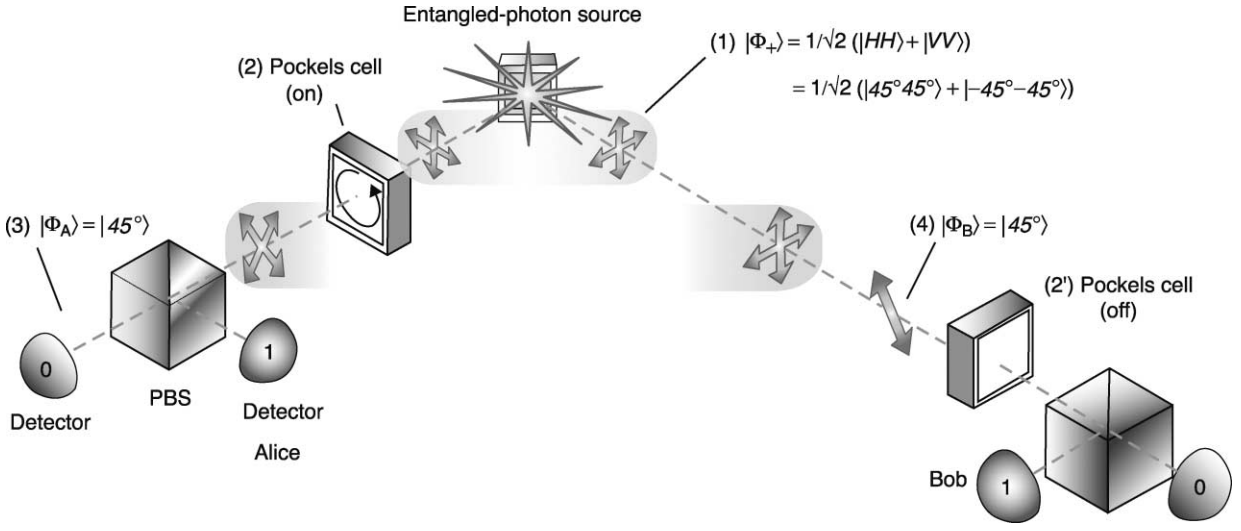


Figure 3 Quantum cryptography using entangled photons. Entangled photons can be used to create a pair of identical cryptographic keys. One member of an entangled pair (1) is sent to Alice, and the other to Bob. Alice and Bob each randomly and independently analyze their respective photon in one of two linear polarization bases with a polarizing beamsplitter (PBS); the basis can be actively chosen using a Pockels cell before the PBS, as shown, to rotate the polarization (alternatively, the 'choice' could be made by directing the photon onto a nonpolarizing 50/50 beamsplitter; in the transmitted path, one analysis basis is used, in the reflected path, the other is used). In the example shown, Alice used the 45°/−45° basis (2), and measured 45° polarization (3), thus projecting Bob's photon into the identical state (4). Since he chose the H/V analysis basis (2′), he is equally likely to detect a 0 or a 1. By subsequent public discussions Alice and Bob determine the events for which they used the same analysis basis (and discard the other events). For these events Alice and Bob will have obtained identical measurement results, which they may interpret as raw key material. After classical error correction and privacy amplification techniques are applied, the remaining string is the sought-after shared secret key. (Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.)

Table 1 Polarization entanglement-based quantum cryptography protocol

Alice's analysis basis	H/V	H/V	45°/−45°	H/V	45°/−45°	45°/−45°	H/V	45°/−45°
Alice's measurement result[a]	*H*	–	*−45°*	*V*	*45°*	*45°*	*V*	*−45°*
Alice's bit value	0	–	1	1	0	0	1	1
Bob's analysis basis	H/V	45°/−45°	45°/−45°	45°/−45°	H/V	45°/−45°	H/V	HV
Bob's measurement result	*H*	*−45°*	*−45°*	*45°*	*H*	–	*V*	*H*
Bob's bit value	0	1	1	0	0	–	1	0
Public discussion: Both photons detected and same basis used?	yes	no	yes	no	no	no	yes	no
Remaining secret key	0		1				1	

[a]In some cases Alice or Bob may not detect a photon due to loss or detector inefficiency. These events simply do not contribute to the key data.

that arbitrary quantum systems cannot be accurately cloned.[8] If Eve steals Bob's photon (a 'denial-of-service' attack), the photon's information never becomes part of the key. Thus, although a wiretap would reduce the rate of the transmission, it would not jeopardize the security of the key. Eve can try to intercept the photon, measure it, and send another one to Bob. But any measurement Eve would make to determine the photon's polarization state would perturb the photon and collapse the entangled state. The photon she sends to Bob would therefore only be 'classically' correlated with Alice's photon. Consequently, Eve's intervention necessarily induces additional errors into Bob's key, which Alice and Bob can detect by publicly revealing a small subset of their actual key. Unfortunately, even with no eavesdropper, the encryption keys created by any real-world quantum cryptography system typically possess a few-percent errors. To make sure their key is secure, Alice and Bob ascribe all errors to Eve and then estimate the maximum amount of information available to the eavesdropper. They then use a classical privacy amplification protocol to reduce Eve's knowledge of the secret key to less than one bit by reducing the length of the key. It has been proven that if the initial error probability per bit is greater than ~15%, no secret bits will remain after error detection and privacy amplification.

Researchers are working to make entanglement-based quantum cryptography more practical. Also, a number of longer distance quantum cryptography demonstrations using weak pulses have been carried out, over tens of kilometers in fibers and in free space. In fact, the first commercially available systems have recently been announced (in Europe and the United States).

[8] The no-cloning proof is very simple. If we have a copying operation such that $|0\rangle|c\rangle \rightarrow |0\rangle|0\rangle$ and $|1\rangle|c\rangle \rightarrow |1\rangle|1\rangle$ ($|c\rangle$ is the initial state of the copier), then by linearity, a superposition becomes an entangled state, instead of two copies: $\frac{(|0\rangle+|1\rangle)}{\sqrt{2}}|c\rangle \rightarrow \frac{(|0\rangle|0\rangle+|1\rangle|1\rangle)}{\sqrt{2}} \neq \frac{(|0\rangle+|1\rangle)}{\sqrt{2}}\frac{(|0\rangle+|1\rangle)}{\sqrt{2}}$.

Table 2 Method to convert the initial state $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ into any of the four Bell states

Alice's operation	*Polarization transformation*	*Resultant two-qubit state*
$\hat{I}$, the identity	H → H; V → V	$\|\Phi_+\rangle = \frac{1}{\sqrt{2}}(\|00\rangle + \|11\rangle)$
$\hat{\sigma}_x$	H → V; V → H	$\|\Psi_+\rangle = \frac{1}{\sqrt{2}}(\|01\rangle + \|10\rangle)$
$i\hat{\sigma}_y$	H → V; V → −H	$\|\Psi_-\rangle = \frac{1}{\sqrt{2}}(\|01\rangle - \|10\rangle)$
$\hat{\sigma}_z$	H → H; V → −V	$\|\Phi_-\rangle = \frac{1}{\sqrt{2}}(\|00\rangle - \|11\rangle)$

Quantum Superdense Coding

It is possible for Alice to send Bob two bits of classical information using a single qubit in the quantum superdense coding protocol. Suppose that Alice and Bob share one qubit each of an entangled pair in the maximally entangled state $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, where we have returned to the generic labeling scheme for the qubit states. Alice has two classical bits of information, which is equivalent to one of four choices. She can encode this information by applying one of four possible transformations on her qubit; a suitable set of transformations are the three Pauli matrices and the identity (i.e., do nothing). This set of operations can be performed experimentally on photon qubits quite easily, e.g., using waveplates, as indicated in Table 2. The four resultant two-qubit states are all maximally entangled and are orthonormal. They form a special basis for the two-qubit states, the Bell basis, which is of particular use in analyzing entanglement.

Alice now sends her qubit to Bob, e.g., through an optical fiber. Bob can then perform a Bell state analysis[9] on his qubit pair, i.e., a projective measurement of the two-qubit state in the Bell basis. The result

[9] Complete discrimination of all four Bell states is presently an unsolved technical problem: for photons one needs either a nonlinear interaction (which is typically very weak) or to exploit so-called 'hyper-entanglement' involving other entangled degrees of freedom of the photons.

immediately reveals the choice of operation Alice made, and in effect, two bits of classical information have been encoded on a single qubit.

Quantum Teleportation

Another application of entanglement is quantum state teleportation (Figure 4), in which the infinite amount of information contained in an arbitrary qubit state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ is transferred by communication of two bits of classical information between Alice and Bob, if they also share two qubits in the maximally entangled state $|\Psi_-\rangle$. The three-photon initial state (i.e., the input photon plus the two entangled photons) is

$$|\psi_0\rangle = (\alpha|H\rangle + \beta|V\rangle) \otimes \tfrac{1}{\sqrt{2}}(|HV\rangle - |VH\rangle),$$

which can be rewritten with the first two photons (the input plus the first half of the entangled pair) represented by the Bell state basis:

$$\begin{aligned}|\psi_0\rangle = \tfrac{1}{2}\{&|\Phi_+\rangle \otimes (-\beta|H\rangle + \alpha|V\rangle)\\ &+ |\Phi_-\rangle \otimes (\beta|H\rangle + \alpha|V\rangle)\\ &+ |\Psi_+\rangle \otimes (-\alpha|H\rangle + \beta|V\rangle)\\ &+ |\Psi_-\rangle \otimes (-\alpha|H\rangle - \beta|V\rangle)\}.\end{aligned}$$

Now suppose that a Bell state measurement is performed on the first two photons. The third photon

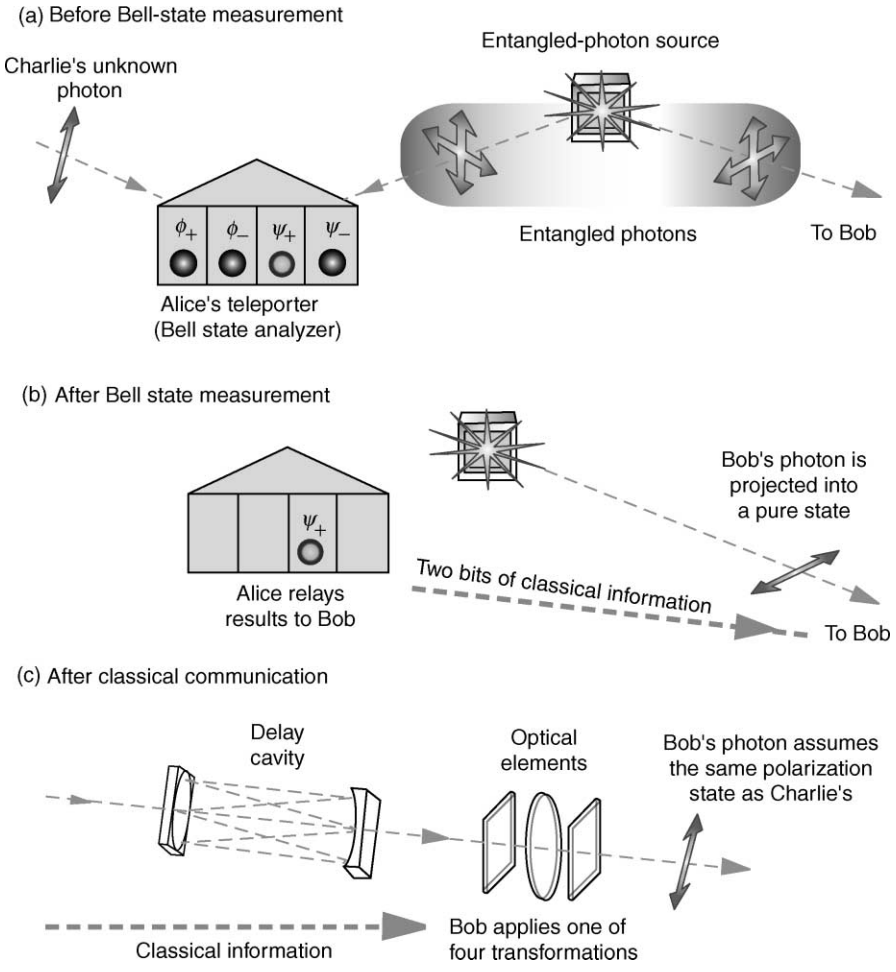


Figure 4 Quantum state teleportation. (a) Teleportation requires an entangled photon source, a Bell state analyzer and a classical communications channel. One entangled photon goes to Bob and the other to Alice, who also receives a photon of unknown polarization state. (b) Alice performs a joint polarization measurement of the two photons and relays the result to Bob using two classical bits of information. The photon going to Bob is projected into a pure state as a result of Alice's measurement. (c) Upon receiving Alice's classical information, Bob performs a simple transformation on his photon (which he has been storing), such as a rotation of the polarization vector. The resulting state of his photon is then identical to the unknown state Alice wished to teleport. (Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.)

is immediately projected into one of four possible states, which can be transformed back into the state of the original input photon by a simple operation, e.g., with a waveplate. For example, if the Bell state measurement produced the result $|\Psi_+\rangle$, the third photon is immediately 'collapsed' into the pure state $|\psi_3\rangle = -\alpha|H\rangle + \beta|V\rangle$. By applying a π-phase shift to the horizontal polarization (relative to the vertical), $|\psi_3\rangle$ can be transformed into the original input state: $|\psi'_3\rangle = \alpha|H\rangle + \beta|V\rangle$. Alice (who does the Bell state measurement) need only communicate to Bob (who wants to receive the teleported photon) which of the four Bell states she measured. Note that during the entire teleportation procedure, neither Alice nor Bob can obtain any idea of the values of the parameters α and β, which specify the state. Also, because Alice's measurement collapses the unknown state, there is only a single copy at the end of the protocol.

Quantum state teleportation was first demonstrated experimentally by Zeilinger and coworkers (University of Innsbruck, Austria). The group was able to determine two of the four Bell states unambiguously (the other two states gave the same experimental signature) and proved for those cases that the state of the input photon could indeed be transferred to Bob. Other experiments have realized modified forms of quantum teleportation in different systems. For example, Kimble's group (California Institute of Technology) teleported the coherent state of an optical mode using squeezed (rather than polarization-entangled) light. Researchers have recently suggested how teleportation might form the basis of a distributed network of quantum communication channels, and how it might enable quantum computing in all-optical systems.

Other Application of Optical Entanglement

Quantum Computing

The most challenging and powerful application of quantum information is large-scale quantum computing. Two features that make quantum information processing potentially powerful are the exponentially large Hilbert space, which gives quantum registers very large capacities, and quantum parallelism, which means that data processing tasks can be performed very efficiently. One fundamental drawback, which severely constrains useful applications of quantum computers, is that the final measurement can only produce a number of classical bits equal to the number of qubits. Thus, quantum computers are limited to performing tasks in which a small amount of information is meant to be gleaned from processing a large amount of data; examples include searching an unstructured database for a specific entry (Grover's algorithm) or finding the periodicity of a function (the quantum Fourier transform). This second task is central to Shor's factor-finding algorithm, the most famous quantum computing algorithm to date.

A practicable quantum computer technology must have at least the following features, first identified by DiVincenzo:

- A set of well-characterized, distinguishable qubits to form a quantum data storage register.
- The ability to initialize the qubits of the register in a simple fiducial state.
- Decoherence times that are much longer than the time needed to perform logical operations.
- The ability to perform any single qubit operation on any qubit in the register, and the ability to perform two-qubit conditional logic gates (such as the CNOT gate: $\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle \rightarrow \alpha|00\rangle + \beta|01\rangle + \gamma|11\rangle + \delta|10\rangle$). Together, these operations constitute a universal set of quantum gates, from which all other gates can be synthesized.
- The ability to measure each qubit.

All-optical schemes have been used to implement small quantum algorithms. However, most of the approaches are not scalable, due to the limitations of non-linear optics to perform a CNOT gate at the single-photon level. Recent proposals have suggested that very high efficiency single-photon detectors, along with sources of single photons 'on-demand,' allow scalable quantum computing with only linear optics; preliminary two-qubit gates have been experimentally demonstrated. A number of other promising candidate technologies that meet DiVincenzo's five requirements are being pursued vigorously. For example, the ability to create multi-qubit entanglement and perform reliable measurements on trapped ions cooled and manipulated by lasers has recently been demonstrated. Solid state systems offer the possibility of scalability, and a number of schemes, such as quantum dots, isolated impurities with nuclear spins, and superconducting quantum interference devices (SQUIDs), are being investigated. It is possible that the final quantum computing technology may take the form of a hybrid between various present approaches.

Lithography

Lithography, in which a pattern is optically imaged onto some photoresistive material, is the primary method of manufacturing microscale or nanoscale

electronic devices. An inherent limitation of this process is that details smaller than a wavelength of light cannot be written reliably. However, quantum state entanglement might circumvent this limitation. Under the right circumstances, the interference pattern formed by beams of entangled photons can have half the classical fringe spacing. Quantum lithography requires two beams of photons, a coherent superposition consisting of the state in which two photons are in beam A while none are in B, and the state in which no photon is in beam A while two photons are in B. Such number-entangled states can be made in the laboratory, and the predictions about fringe spacings have been verified. However, other obstacles must be overcome to surpass current classical-lithography techniques.

Two-Photon Imaging and Microscopy

At present, two-photon microscopy is widely used to produce high-resolution images, often of biological systems. However, the classical light sources (lasers) used for the imaging have random spreads in the temporal and spatial distributions of the photons, and the light intensity must be very high if two photons are to intersect within a small enough volume to cause a detectable excitation. Such high intensity can damage the system under investigation. Because the temporal and spatial correlations may be much stronger between members of an entangled photon pair, much weaker light sources could be used, which would be much less damaging to the systems being observed. The development of such systems is currently an active area of research.

List of Units and Nomenclature

Entanglement: a property of quantum systems consisting of two or more distinct subsystems, often separate particles. When transformations on one system do not affect the other, the state is said to be separable; when this is not the case, and the quantum state of the overall system cannot be resolved into separate states of the individual pieces, the system is entangled.

Pure and mixed states: when the probability amplitudes specifying a quantum state are deterministic, the state is said to be pure; when the amplitudes are random quantities, the state is mixed. The distinction is similar to that of coherent and partially coherent optical fields.

Quantum computer: an information processing device that exploits quantum mechanical phenomena to greatly enhance computational power for certain problems. Some mathematical problems thought to be intractable on conventional computers can, in theory, be performed efficiently on a quantum computer.

Quantum crytography (also known as quantum key distribution): a quantum information protocol by which two parties acquire a shared series of random numbers (a cryptographic key) by exploiting the quantum nature of light. Absolute security can be proven by virtue of the indivisibility and uncopy-ability of individual quanta.

Quantum state tomography: a means by which quantum states can be determined experimentally by performing a series of appropriate measurements on multiple identically prepared systems. From such measurements, the elements of the density matrix, which fully specifies the state, may be inferred.

Quantum superdense coding: A quantum information protocol by which two bits of classical information may be communicated by a single quantum bit, initially part of an appropriate entangled quantum state.

Quantum teleportation: a quantum information protocol by which the unknown quantum state of one particle can be transferred to another distant particle, using a pair of entangled particles, a projective measurement, and exchange of two bits of classical information.

Qubit (quantum bit): a two-level quantum mechanical system, which constitutes the building blocks of quantum information processing devices. Examples include a spin-1/2 particle, an atom with two well-isolated levels, or the polarization of a photon.

See also

Nonlinear Optics, Applications: Three-Dimensional Microfabrication. **Quantum Optics:** Quantum Computing with Atoms.

Further Reading

Berman GP, Doolen GD, Mainieri R and Tsifrinovitch V (1998) *Introduction to Quantum Computers*. Singapore: World Scientific.

Bouwmeester D, Ekert AK and Zeilinger A (eds) (2000) *The Physics of Quantum Information*. Berlin: Springer-Verlag.

Braunstein SL and Lo HK (eds) (2001) *Scalable Quantum Computers: Paving the Way to Realization*. Berlin: Wiley-VCH.

DiVincenzo DP (2000) The physical implementation of quantum computation. *Fortschrcitte der Physik* 48: 771–783.

Gisin N, Ribordy G, Tittel W and Zbinden H (2002) Quantum cryptography. *Review of Modern Physics* 74: 145–195.

Kok P, Braunstein SL and Dowling JP (2002) Quantum lithography. *Optics and Photonics News* 13(9): 24–27.

Lo HK, Popescu S and Spiller T (eds) (1998) *Introduction to Quantum Computation and Information*. Singapore: World Scientific.

Macchiavello C, Palma GM and Zeilinger A (eds) (2000) *Quantum Computation and Quantum Information Theory*. Singapore: World Scientific.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge, UK: Cambridge University Press.

Nielsen MA and Chuang IL (2000) *Quantum Computation and Quantum Information*. Cambridge, UK: Cambridge University Press.

Williams CP and Clearwater SH (1998) *Explorations in Quantum Computing*. Santa Clara: Telos.

Yariv A (1989) *Quantum Electronics*. New York: Wiley.

Laser Cooling of Ions

H Walther, University of Munich and Max-Planck-Institute for Quantum Optics, Garching, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The fact that light can exert a mechanical force on atoms by the transfer of linear momentum was recognized as long ago as the 1930s, when Frisch demonstrated the deflection of an atomic beam using resonant light. With the advent of tunable laser light in the early 1970s, a much larger excitation rate for atoms was achieved, leading to a much stronger deflection. A little earlier it was also observed that in addition to the force resulting from photon recoil, a dipole force is also produced in a focused light beam so that even small solid particles can be trapped and manipulated. The original proposal for laser cooling of atomic gases was put forward in 1975 and later laser cooling of trapped ions was proposed. The principle of laser cooling is very simple. The ion is excited by laser light below a resonance frequency; in order that a transition can be made, the residual energy has to be taken from the kinetic energy of the particle, leading to cooling of the particle. In this article, we will discuss this cooling process in more detail.

Laser Cooling of Ions in a Radiofrequency Trap

Of the various proposed and realized methods for laser cooling of atoms or trapped ions, sideband cooling is conceptually the simplest and most efficient. It allows a trapped particle to be cooled to the quantum-mechanical ground state of a harmonic trap potential. The process can be regarded as anti-Stokes Raman scattering of the pseudomolecule formed by the ion and the trap. If the linewidth of an optical transition in this atom is smaller than the vibrational frequency of the trapped ion, both electronic levels of the transition split into a ladder of resolved vibrational levels. The absorption spectrum consists of a carrier at the resonance of the free atom and sidebands at multiples of the vibrational frequency. With a laser tuned to the m-th lower frequency sideband, each absorption of a photon will reduce the vibrational energy of the atom by m quanta. Since the subsequent spontaneous re-emission will, on average, not change the vibrational excitation, repeated photon scattering processes lead to efficient cooling of the atom. This cooling will continue until the absorption probability for the low-frequency sidebands decreases when approaching the vibrational ground state of the ion in the trap. The final temperature is limited by the heating process caused by photon recoil. However, after the ion is cooled to the vibrational ground state this heating process disappears, when the vibrational energy of the ion is larger than the recoil energy of a photon. In this case, the recoil is taken up by the entire trap setup, as it is the case in the Mössbauer effect.

Despite these obvious advantages of sideband cooling, relatively few experiments have been made using this method. This is because the regime, where the oscillation frequency ω is larger than the natural linewidth Γ of an optical transition (the so-called 'strong binding regime'), is not easily accessible. Vibrational frequencies of ions in standard electromagnetic traps are of the order of 1 MHz, whereas linewidths of electric dipole transitions in positive ions are in the range 20–50 MHz. In this 'weak binding regime' with $\Gamma > \omega$, laser cooling of trapped

ions is essentially analogous to Doppler cooling of free atoms. The minimum temperature is given by the Doppler limit $k_B T = \hbar\Gamma/2$, and the mean vibrational quantum number is $\langle n\rangle = \Gamma/2\omega > 1$. The first experiments with cooling on resolved sidebands, therefore, used a strongly forbidden optical transition or stimulated Raman transitions. In these cases, however, sideband cooling is inevitably slow and it was used only to extract a few vibrational quanta from Doppler-precooled ions. In both experiments the vibrational ground state was reached to a very good approximation (mean quantum number $\langle n\rangle \approx 0.05$). Raman sideband cooling has also recently been applied to neutral atoms in optical traps.

Laser cooling of trapped ions was motivated by their possible use for optical frequency standards where the relativistic Doppler shift, as well as the Stark shift due to the trap potential, are eliminated at low temperatures. Cooling to the ground state also opens up the possibility of studying a well-controlled mechanical quantum oscillator experimentally. This led to the realization of nonclassical states of motion, investigation of quantum decoherence phenomena, and the demonstration of a quantum logic gate for quantum computing.

Sideband Cooling

Most theoretical treatments of sideband cooling consider a static harmonic trap, whereby the time dependence of the potential of the radio-frequency (rf) ion trap is neglected. In such a model a running laser wave $\propto\exp(ikr)$ of wave number k induces transitions between the eigenstates $|n\rangle$ and $|n'\rangle$ of a harmonic oscillator. The relevant transition matrix elements are

$$U_{n'n} = \langle n'|\exp(ikr)|n\rangle \qquad [1]$$

$$U_{n'n} = \exp(-\eta^2/2)\sqrt{\frac{n_<!}{(n_< + \Delta n)!}}(i\eta)^{\Delta n} L_{n_<}^{\Delta n}(\eta^2) \qquad [2]$$

Here $\eta = kx_0 = k\sqrt{\hbar/2m\omega}$ is the Lamb–Dicke parameter for the spatial extension x_0 of the ground state, $n_<$ the smaller of the two numbers n and n', $\Delta n = |n - n'|$, and L_n^m denotes the generalized Laguerre polynomial. In the case of high vibrational excitation $n, n' \gg 1$, $U_{n',n}$ can be approximated as

$$U_{n+\Delta n,n} \approx i^{\Delta n} J_{\Delta n}(2\eta\sqrt{n}) \qquad [3]$$

with the Bessel function of n-th order J_n. The argument of the Bessel function is the product of the wave number and the classical vibration amplitude of the highly excited state. This is the result one would expect for the frequency-modulated spectrum of a classical oscillator. For arguments β among the Bessel functions $J_n(\beta)$, those of order $n \approx \beta$ have the largest modulus. Consequently, during the initial stages of cooling it is most efficient to tune the laser so that it induces transitions with $\Delta n \approx 2\eta\sqrt{n} \gg 1$.

For the final stages of cooling one can usually assume that the ion is in the Lamb–Dicke regime, i.e., that it is localized within dimensions of the optical wavelength. A detailed study of the limits of sideband cooling in a static trap was carried out in 1984. The final distribution of population over the vibrational states is thermal and can be characterized by a mean occupation number $\langle n\rangle$ or a temperature $T = \hbar\omega/k_B \ln(1 + 1/\langle n\rangle)$. The number $\langle n\rangle$ is a rational function of the four frequencies involved, i.e., the natural linewidth Γ, the oscillation frequency in the trap ω, the laser detuning δ, and the Rabi frequency ω_R. For certain limiting cases, simple expressions for the mean vibrational excitation $\langle n\rangle$ can be derived from the general result. The lowest vibrational excitation is achieved in the case of well-resolved sidebands ($\omega \gg \Gamma$) and under weak laser excitation ($\omega_R \ll \Gamma$), tuned to the first sideband ($\delta = -\omega$):

$$\langle n\rangle = \left(\alpha + \frac{1}{4}\right)\frac{\Gamma^2}{4\omega^2} \qquad [4]$$

The constant α depends on the geometry of excitation and photon re-emission and is of the order of 1. In the case of strong saturation ($\omega_R \gg \Gamma$) of the first sideband ($\delta = -\omega$), the mean vibrational quantum number is proportional to the laser intensity:

$$\langle n\rangle = \frac{\omega_R^2}{8\omega^2} \qquad [5]$$

This case is of practical relevance if one wants to detect a fluorescence signal on the cooling transition, for example, to measure the internal state of the ion in a double-resonance experiment, while keeping the ion cooled. For the choice of parameters $\omega_R = \omega = -\delta$, the ion is still predominantly in the vibrational ground state ($\langle n\rangle \approx 1/8$), but photons are scattered from the wing of the saturation-broadened carrier at a rate $P = \Gamma\omega_R^2/(\Gamma^2 + 2\omega_R^2 + 4\delta^2) \approx \Gamma/6$, which is already one-third of the maximum scattering rate. Finally, in the case of large laser detuning ($\delta \gg \Gamma$, ω, ω_R), the energy is proportional to the detuning because the lower vibrational levels are no longer depleted by the laser:

$$\langle n\rangle \approx \frac{\alpha + 1}{4}\frac{\delta}{\omega} \qquad [6]$$

If the Lamb–Dicke parameter satisfies $\eta \ll 1$, the strength of the sidebands can be calculated from a power series expansion of the matrix elements of eqn [2] and the thermal distribution of population over the oscillator levels. The $U_{n',n}$ are evaluated to first-order in η only. In this approximation, the height of the first lower-frequency sideband in the absorption spectrum relative to the carrier is $\eta^2\langle n\rangle$ and that of the first higher-frequency sideband $\eta^2(\langle n\rangle + 1)$. This information can be used to determine experimentally the vibrational quantum number from the absorption spectrum. The strengths of the higher-order sidebands contain higher powers of η^2 and therefore decrease rapidly.

Let us now briefly review some facts about the motional spectrum of an ion in a Paul trap, which is not just a static system but possesses an explicitly time-dependent potential. The equation of motion for a classical particle of charge e and mass m in the quadrupole potential of the trap $\phi(r,z) = (V_0 - U_0 \cos \Omega t)(r^2 - 2z^2)/2r_0^2$ is given by the Mathieu differential equation:

$$\frac{\delta^2 r}{\delta \tau^2} + (a_r - 2q_r \cos 2\tau)r = 0 \qquad [7]$$

with dimensionless parameters:

$$\tau = \frac{\Omega}{2}t, \qquad a_r = \frac{4eV_0}{m\Omega^2 r_0^2}, \qquad q_r = \frac{2eU_0}{m\Omega^2 r_0^2} \qquad [8]$$

Here a static voltage V_0 and an alternating voltage with amplitude U_0 and frequency Ω are applied to a Paul trap with the radius of the ring electrode r_0. The equation of motion for the z coordinate is obtained by multiplying the parameters a and q by the factor -2. The general solution of the Mathieu equation in the first stability region of the (a,q) parameter space can be written by using the Floquet ansatz:

$$r(\tau) = A\sum_{j=-\infty}^{\infty} c_j \cos(2j+\beta_r)\tau + B\sum_{j=-\infty}^{\infty} c_j \sin(2j+\beta_r)\tau \qquad [9]$$

where A and B are constants determined by the initial conditions, and the c_j and β_r depend on a and q. In the first stability region, where most of the experiments are carried out, the value of β_r is between 0 and 1. According to eqn [9] the motional spectrum of the ion has resonances at

$$\omega_{r,j} = \left(j \pm \frac{\beta_r}{2}\right)\Omega, \qquad j = 0, 1, 2\ldots \qquad [10]$$

In the adiabatic approximation, which is valid for $a, q^2 \ll 1$, the motion of the ion can be separated into an oscillation at the driving frequency Ω, the so-called micromotion, and a slower oscillation, called the secular motion, describing the oscillation in the time-averaged pseudopotential. This pseudopotential describes the effect of the ponderomotive force of the oscillating trapping field that drives the ion to the field minimum at the center of the trap. Since, in the quadrupole trap the pseudopotential is harmonic, the solution to the equation of motion can be approximated as

$$r(\tau) = r_0 \cos\beta_r\tau\left(1 + \frac{q_r}{2}\cos 2\tau\right) \qquad [11]$$

The approximate expression for β_r is $\beta_r = \sqrt{q_r^2/2 + a_r}$ and the secular frequencies are given by

$$\omega_r = \frac{\Omega}{\sqrt{8}}\sqrt{q_r^2 + 2a_r}, \qquad \omega_z = \frac{\Omega}{\sqrt{2}}\sqrt{q_r^2 - 2a_r} \qquad [12]$$

which are the lowest-order resonances $\omega_{r,z} = \beta_{r,z}\Omega/2$ in the Floquet ansatz, eqn [9]. From eqn [11] it is seen that the amplitude of motion at the resonances $\Omega \pm \omega$ is a factor $q_r/4$ smaller than those at the resonance ω. It is noteworthy that the frequency Ω of the driving field does not appear directly in the motional spectrum, but only in combination with the secular frequency ω. This changes, however, if there is an additional static force acting on the ion, displacing it from the center of the quadrupole. The approximate solution to the equation of motion then becomes:

$$r(\tau) = r_1 + r_0 \cos\beta\tau\left(1 + \frac{q_r}{2}\cos 2\tau\right) \qquad [13]$$

and a resonance at Ω appears. The force that displaces the ion by r_1 may be due to static electric stray fields or the presence of a second ion in the trap. These effects can have a strong influence on sideband laser cooling of the trapped ions. A number of theoretical studies have been published on the quantum-mechanical treatment of the Paul trap. The explicitly time-dependent potential does not allow stationary states with time-independent energy eigenvalues. One can, however, as in the case of the pseudopotential model, distinguish between two time-scales and construct wave functions of the harmonic secular oscillator that show some additional breathing motion at the trap frequency. A quantum-mechanical theory of sideband cooling in the time-dependent potential of the Paul trap thus far exists only for special cases; in 1994 there was a case where the ion was localized in the Lamb–Dicke regime at the node of a standing laser wave and time

averaged kinetic energies were calculated. Minima of the energy that agree with the results of the theory for the static harmonic trap to within a factor of 2 were found. The strongest modifications to the predictions of the cooling theory for the static trap arise in the dependence of the energy on the laser detuning. The influence of the micromotion gives rise to heating for certain regions of negative laser detuning (on the high-frequency side of a low-frequency micromotion sideband) and to cooling for some positive detunings (on the low-frequency side of a high-frequency micromotion sideband).

Trapped LASER-cooled ions have a variety of applications in spectroscopy and other fields.

In the following, a phenomenon will be discussed where observations during the cooling process of ions in traps is interpreted as phase transitions between chaotic and ordered ion structures during the cooling process.

Generation of Chaos During the Cooling Process and Phase Transitions to Ordered Configurations

Using a Paul trap, it is possible to study the physics of a few-body phase transition of ions. The ordered state of the trapped ions can be observed directly by monitoring their resonantly scattered light. The transition to an ordered or 'crystalline' configuration is expected when the Coulomb coupling constant which corresponds to the ratio of Coulomb energy to the kinetic energy of a particle, is larger than 1. The ordered structures represent configurations of maximum density for the ions, they are determined by the trap potential pushing the ions towards the trap center, and by the Coulomb repulsion between the ions acting in the opposite direction, so that minimum energy configurations are finally adopted. The system corresponds to an ordered one-component plasma or a strongly coupled plasma.

In the following, computer simulations of the phase transition in the Paul trap will be reviewed. In the trap the ions are subjected to five different forces. These are the force $\mathbf{F}^{(\mathrm{trap})}$, arising from the dynamical rf trapping field, the Coulomb interaction between the ions $\mathbf{F}^{(\mathrm{Coul})}$, the laser cooling force $\mathbf{F}^{(\mathrm{laser})}$, a random force $\mathbf{F}^{(\mathrm{rand})}$ arising from the recoil of the spontaneously emitted photons, and a parasitary force $\mathbf{F}^{(\mathrm{cont})}$, arising from a contact potential, which may be generated unintentionally by locally coating the electrodes with the trapped atoms.

The potential generated by the rf voltage applied to the hyperbolic electrodes of the trap is given by

$$\phi(\mathbf{r},t) = \varphi(t)[x^2 + y^2 - 2z^2]$$
$$\varphi(t) = \frac{U_0 + V_0\cos(\Omega t)}{r_0^2 + 2z_0^2} \qquad [14]$$

where U_0 and V_0 are the dc and ac components of the voltages applied to the trap. From eqn [1] we obtain the trapping force and the resulting equation of motion of a single ion in the trap:

$$\mathbf{F}^{(\mathrm{trap})} = m\ddot{\mathbf{r}} = \mathbf{e}\nabla\phi(\mathbf{r},t) = -2\mathbf{e}\varphi(t)[\mathbf{r} - 3z\mathbf{e}_z] \qquad [15]$$

where m, $\mathbf{r}$, and $z\mathbf{e}_z$ denote the mass of the ion, the position, and the $z\mathbf{e}_z$-component of the position vector, respectively.

The Coulomb interaction between the particles results in the force:

$$\mathbf{F}_i^{(\mathrm{Coul})} = \frac{e^2}{4\pi\varepsilon_0}, \sum_{m\neq 1}\frac{\mathbf{r}_i - \mathbf{r}_m}{[\mathbf{r}_i - \mathbf{r}_m]^3} \qquad [16]$$

In the process of laser cooling, every scattered photon changes the momentum of an ion on the average by an amount $\hbar k$.

The ensemble average of the resulting laser cooling force acting on the i-th ion is thus:

$$\mathbf{F}_i^{(\mathrm{laser})} = \hbar k R[\mathbf{r}_i(t);\ \dot{\mathbf{r}}_i(t)] \qquad [17]$$

The number $N_i(l)$ of spontaneously emitted photons from the i-th ion during the l-th cycle of period $T = 2\pi/\Omega$ is

$$N_i(l)\int_{(l-1)T}^{lT} R[\mathbf{r}_i(t);\dot{\mathbf{r}}_i(t)]\mathrm{d}t \qquad [18]$$

In the simulations, it is assumed that these N_i spontaneous photons are emitted one by one, at the end of each rf cycle, rather than emitting them one by one, according to the appropriate photon-statistics during the cycle. The orientation of the unit vector $\hat{\mathbf{u}}_j$, which points in the direction of the j-th emitted photon, is chosen at random, but weighted with a $\cos^2$ distribution (dipole characteristic) with respect to the laser polarization axis. This way, only the linearly polarized spontaneously emitted photons have been simulated in the calculations. The circularly polarized ones, which occur with a $\sin^2$ distribution, have been neglected. The random force $\mathbf{F}_i^{(\mathrm{rand})}$, is thus:

$$\mathbf{F}_i^{(\mathrm{rand})}(t) = \hbar k\sum_{l=-\infty}^{\infty}\delta(t - lT)\sum_{j=1}^{N_i}\hat{\mathbf{u}}_j \qquad [19]$$

In some of the calculations, a contact potential has been taken into account which is generated experimentally by a coating of the quadrupole electrodes by the trapped particles. Close to the center of the trap, the contact potential gives rise to an approximately homogeneous electric field. Altogether, the motion of the i-th particle is governed by the sum of the above forces, that is:

$$m\ddot{\mathbf{r}}_i = \mathbf{F}_i^{(\mathrm{trap})} + \mathbf{F}_i^{(\mathrm{Coul})} + \mathbf{F}_i^{(\mathrm{laser})} + \mathbf{F}_i^{(\mathrm{rand})} + \mathbf{F}_i^{(\mathrm{cont})} \qquad [20]$$

Given appropriate initial conditions, the eqns of motion [20] are integrated forward in time to obtain the trajectories of the n-ion system. These trajectories provide all the information necessary to find the quantities of interest, such as the occupation probability density of the i-th ion in phase space (i.e., the probability density of finding the i-th ion at a given position, and with a given velocity), the (total) kinetic energy, and the fluorescence intensity.

Some of the results of the described simulations will be summarized below. Figure 1 shows the excitation spectrum of the ions as a function of laser detuning. The experiment confirms the results of the simulations.

Owing to the Coulomb interaction of the ions at particular phases of the cooling process, chaos in the ion cloud is observed leading to a heating process of the ions by the applied field of the trap. This heating mechanism was investigated in the simulations in detail and the summary of the results is given in Figure 2.

For zero laser power and large r, there is no net heating of the ions. This is confirmed by the experiments, in which, even in the absence of a cooling laser, large clouds of ions can be stored in a Paul trap over several hours without being heated out of the trap. The ions are far apart, the Coulomb force is small, and on short time-scales the ions behave essentially like independent singly stored ions. For this reason, this part of the heating diagram is called the 'Mathieu regime' as it is well described by the Mathieu equation [7]. Turning on a small laser power, the rms radius r reduces drastically due to laser

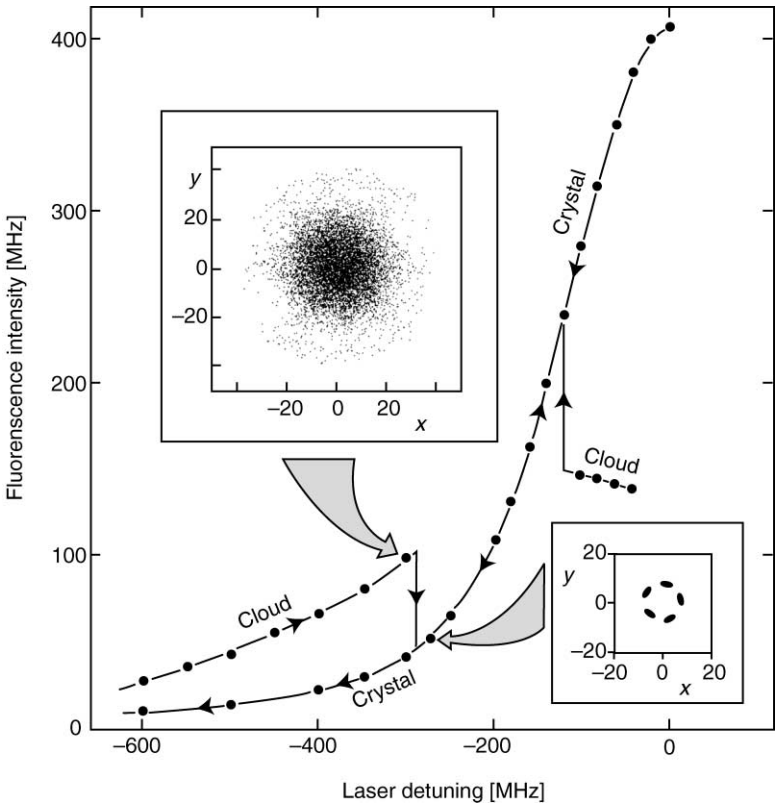


Figure 1 Excitation spectrum of five ions as a function of laser detuning. When the laser detuning is changed the cooling rate of the ion is varying. Two phase transition points (and bistability) are apparent. Insets show the results of the molecular dynamics calculations shortly before and after the jump, which demonstrates directly the existence of a transition from a cloud phase to a crystalline state. The units given at the insets are μm.

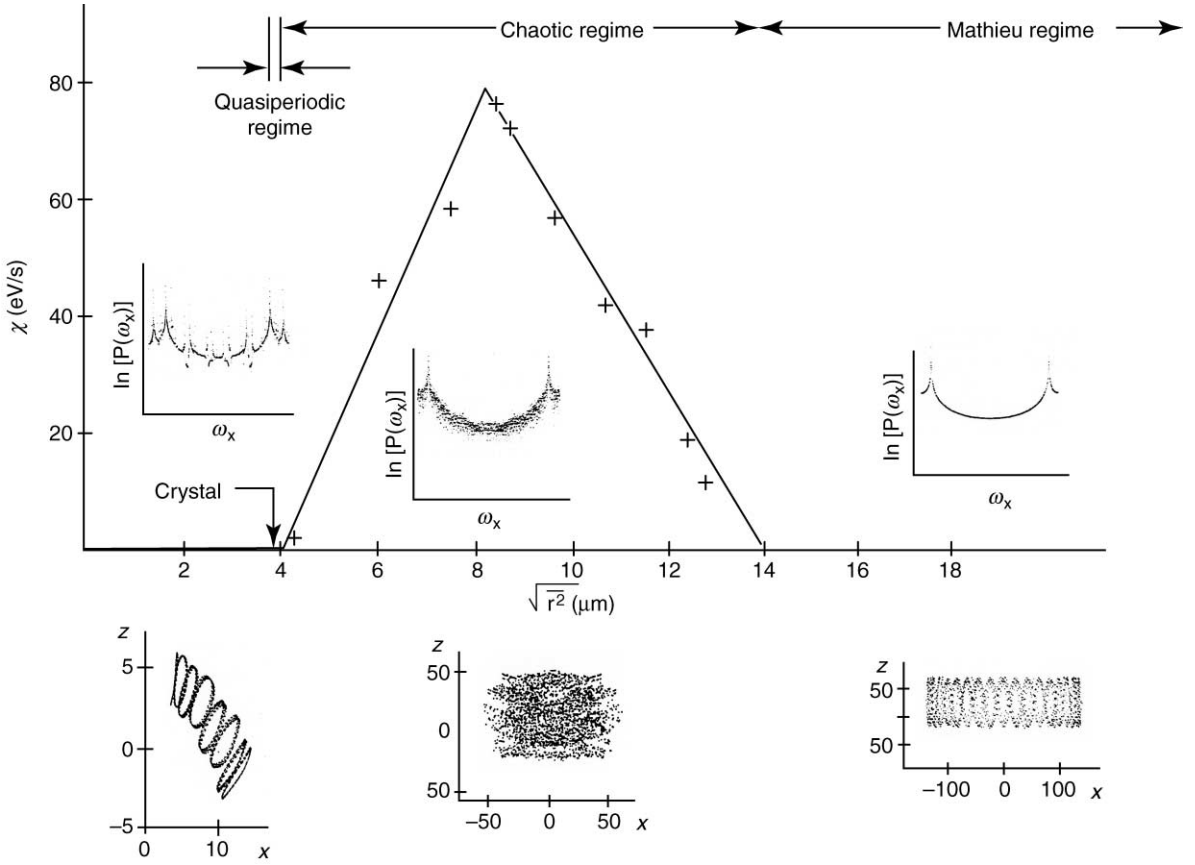


Figure 2 Average heating rate of five ions in a Paul trap versus the rms radius of the ion configuration. The insets show the power spectrum and the corresponding stroboscopic Poincaré sections in the x–z plane of relative separation for two ions in three characteristic domains: the quasiperiodic regime, the chaotic regime and the Mathieu regime. All length scales are in units of micrometers. Units gives at the x–z axis of the insets are in μm.

cooling, but comes to a halt at about 14 μm where chaotic rf heating sets in and balances the cooling power of the laser. Increasing the laser power results in an even smaller cloud. The smaller cloud produces more chaotic rf heating, as seen clearly by the negative slope of the heating curve (Figure 2) in the range $8 < r < 14$ μm.

In the range 4 μm $< r <$ 8 μm, there is still chaotic heating, but the slope of the heating curve is positive. As a consequence of the resulting triangular shape of the heating curve, at about 8 μm, the chaotic heating power can no longer balance the cooling power of the laser, and the cloud collapses into the 'crystalline' state located at $r \approx 3.8$ μm; a 'phase transition' has occurred. Due to this collapse, the behavior of the heating rate in the range 3.8 μm $< r <$ 8 μm cannot be studied by equilibrating laser cooling and rf heating. In this case, we start out from the crystal state and slightly displace the ions to explore the 'vicinity' of the crystal. We observe no heating for 3.8 μm $< r <$ 4 μm, but quasiperiodic motion, and thus name this regime the 'quasi-periodic' regime. We call the upper edge of the quasiperiodic regime ($r \approx 4$ μm) the 'chaos threshold'. An initial condition beyond the chaos threshold, i.e., satisfying $r > 4$ μm, leads to heating and expansion of the ion configuration and numerical data relevant for the shape of the heating curve can be taken during this 'explosion phase'. The laser power P is set to zero for this type of experiment. We conjecture that, apart from the trivial case of a single stored ion, the heating curve is 'universal', i.e., its qualitative shape, including the existence of the chaotic regime, does not depend on the number of simultaneously trapped ions.

For the quasiperiodic, the chaotic and the Mathieu-regime, respectively, the corresponding type of power spectrum are shown in the insets above the abscissa of Figure 2. The data were actually taken from the case of two ions, but would look similar in the five-ion case. A discrete spectrum in the quasi-periodic regime and a complicated

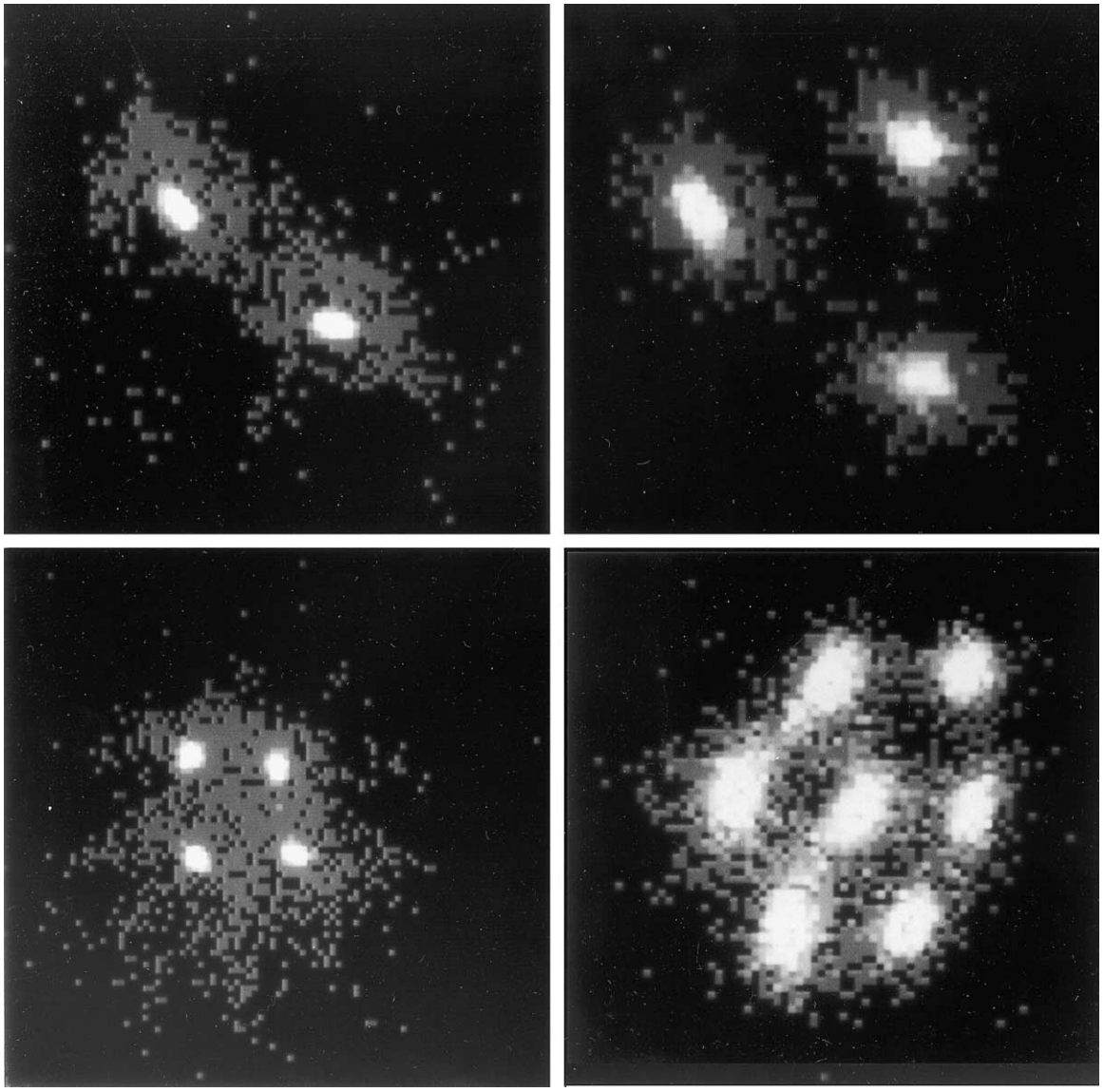


Figure 3 Ordered ion configuration of two, three, four, and seven laser-cooled Mg^+ ions confined in a Paul trap and observed using resonantly scattered light. The average ion separation is 20 μm. The structures of four and seven ions are slightly deformed owing to contact potentials in the trap.

noisy spectrum in the chaotic regime is obtained. The spectrum in the Mathieu regime is again quite simple and dominated by the secular motion frequency. Stroboscopic pictures of the locations of two ions in the x–y plane of the trap characterizing the three regions, are shown on the insets below the abscissa of Figure 2.

Some of the ordered ion structures observed experimentally are shown on Figures 3 and 4. The structures in Figure 3 were obtained in a Paul trap; the others are measured in a storage ring leading to quasi-linear structures of the ion configurations, depending on the symmetry of the trap. Ordered ion structures have also been observed in Penning traps. Due to the symmetry of this trap rotationally symmetric structures are observed.

Conclusion

Laser cooling of ions in traps provides an excellent tool for numerous experiments, varying from tests of basic physics to important and new applications in quantum information processing. The linear ion structure shown on Figure 4 is presently investigated in many laboratories in order to realize gates for

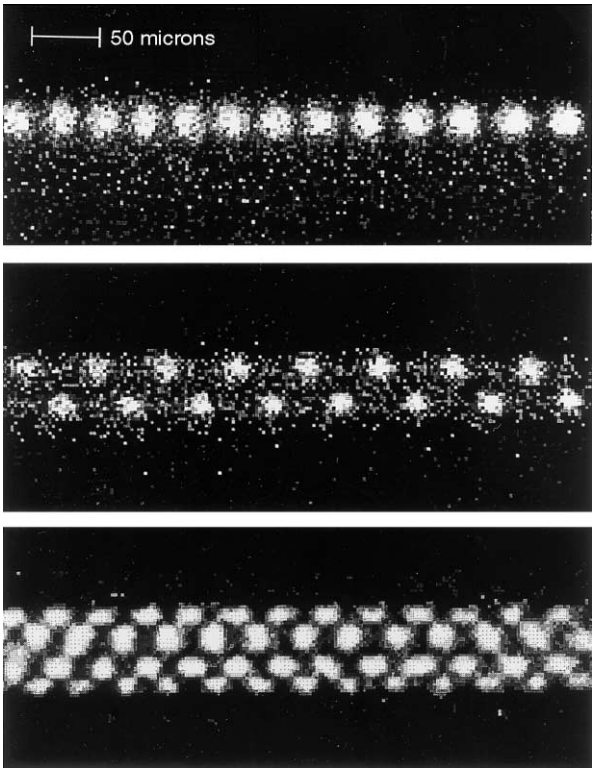


Figure 4 Crystalline structures of laser-cooled $^{24}Mg^{+}$ ions in the quadrupole storage ring. At a low ion density the ions form a string along the field axis (upper). Increasing the ion density transforms the configuration to a zigzag (middle). At still higher ion densities, the ions form ordered helical structures on the surface of a cylinder, e.g., three interwoven helices (lower). As the fluorescent light is projected onto the plane of observation in this case the inner spots are each created by two ions seated on opposite sides of the cylindrical surface, resulting in a single, bright area. (Reproduced by permission of *Nature* from Birkl G, Kassner S and Walther H (1992) Multiple shell structures of laser cooled $^{24}Mg^{+}$ ions in a quadrupole storage ring. *Nature* 357: 310–313.)

quantum computing. Single trapped ions in cavities may be used as deterministic light sources for secure single photon communication links. Furthermore, in many laboratories single laser-cooled ions are investigated as the basis for new time and frequency standards, being several orders of magnitude more accurate than our present atomic clock.

See also

Chaos in Nonlinear Optics. **Quantum Optics:** Entanglement and Quantum Information; Quantum Computing with Atoms.

Further Reading

Arimondo E, Philips WD and Strumia F (eds) (1992) Laser manipulation of atoms and ions. *Proceedings of the International School of Physics Enrico Fermi, Course CXVIII*. Amsterdam, Oxford, New York, Tokyo: North Holland.

Ashkin A (1970) Atomic-beam deflection by resonance-radiation pressure. *Physical Review Letters* 25: 1321–1324.

Birkl G, Kassner S and Walther H (1992) Multiple shell structures of laser cooled $^{24}Mg^{+}$ ions in a quadrupole storage ring. *Nature* 357: 310–313.

Blümel R, Chen JM, Peik E, Quint W, Schleich W, Shen YR and Walther H (1988) Phase transitions of stored laser-cooled ions. *Nature* 334: 309–313.

Blümel R, Kappler C, Quint W and Walther H (1989) Chaos and order of laser-cooled ions in a Paul trap. *Physical Review A* 40: 808–823.

Bollinger JJ, Kriesel JM, Mitchell TB, King LB, Jensen MJ, Itano WM and Dubin DHE (2003) Laser-cooled ion plasmas in Penning traps. *Journal of Physics B: Atomic Molecular Optical Physical* 36: 499–510.

Cirac JI, Garay LJ, Blatt R, Parkins AS and Zoller P (1994) Laser cooling of trapped ions: the influence of micro-motion. *Physical Review A* 49: 421–432.

Dehmelt H (1967 and 1969) In: Bates DR and Estermann I (eds) *Advances in Atomic and Molecular Physics*, vols. 3 and 5. New York: Academic Press.

Dehmelt HG (1983) Stored ion spectroscopy – three lectures. In: Arecchi FT, Strumia F and Walther H (eds) *Advances in Laser Spectroscopy*. New York: Plenum.

Diedrich F, Peik E, Chen JM, Quint W and Walther H (1987) Observation of a phase transition of stored laser-cooled ions. *Physical Review Letters* 59: 2931–2934.

Glauber R (1991) In: Arimondo E and Phillips W (eds) *Proceedings of the International School of Physics 'Enrico Fermi,'* vol. 118. Amsterdam: North Holland.

Hänsch TW and Schawlow AL (1975) Cooling of gases by laser radiation. *Optics Communications* 13: 68–69.

Javanainen J, Lindberg M and Stenholm S (1984) Laser cooling of trapped ions: dynamics of the final stages. *Journal of Optical Society of America B* 1: 111–115.

Leibfried D, Blatt R, Monroe C and Wineland D (2004) Quantum dynamics of single trapped ions. *Review of Modern Physics* 75: 281–324.

Peik E, Abel A, Becker Th, von Zanthier J and Walther H (1999) Sideband cooling of ions in radio-frequency traps. *Physical Review A* 60: 439–449.

Schieder R, Walther H and Wöste L (1972) Atomic beam detection by the light of a tunable dye laser. *Optics Communications* 5: 337–340.

Wineland D and Dehmelt H (1975) Proposed 10–14 laser fluorescence spectroscopy on Tl^{+} mono-ion oscillator. *Bulletin of the American Physical Society* 20: 637.

Quantum Computing with Atoms

S F Huelga, University of Hertfordshire, Hatfield, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

It is a well-known fact that the speed of commercial microprocessors is doubling, roughly, every 18 months (Moore's law). This speedup is achieved mainly at the expense of reducing the size of the microchip. If this tendency is to continue, it seems unavoidable that quantum mechanical effects will eventually become important.

Remarkably, far from being a technological nuisance, the fact that information carriers are truly quantum mechanical systems offers an entirely new paradigm for computation. As opposed to classical bits, with logical values 0 and 1, quantum bits (qubits) can be prepared in arbitrary, normalized superpositions of the logic values. In addition, quantum states of multiqubit systems can be entangled, i.e., cannot be written as tensor products of states of the individual components. The theoretical description of entanglement requires a huge number of parameters which, in fact, grows exponentially with the number of particles that are involved. This exponential number of new degrees of freedom is responsible for the computational potential that quantum systems offer. Recently it has been demonstrated theoretically how we can harness this potential in the form of quantum algorithms, including Shor's factorization algorithm and Grover's protocol for data base search.

From a theoretical point of view, the implementation of quantum computation is greatly simplified by the fact that any unitary transformation involving an arbitrarily large number of qubits can be decomposed into two building blocks: single qubit operations and two-qubit controlled-NOT gates (CNOT-gates). We only need to identify two-level systems whose interaction can be mediated in a way that allows for conditional dynamics, that is, the second system undergoes certain operation (a NOT operation in our case), conditional on the state of the first qubit. However, it is easy to convince ourselves that from the experimental point of view, the challenge is going to be formidable. The following (DiVincenzo)-checklist summarizes the necessary requirements for the potential physical realization of a quantum computer:

- The system needs to provide a well-defined qubit, with two robust levels to represent the logical values 0 and 1. Moreover, the system should be scalable.
- Both the initial preparation and the final readout of the system has to be done with sufficient accuracy.
- The system has to allow the implementation of a universal set of gates (single-qubit rotations and CNOTs gates) with sufficient high fidelity. Implicit in this requirement is the ability to address individual qubits.

From the above list it is clear that the practical realization of a quantum computer relies on the ability to achieve a delicate equilibrium: the system needs to interact strongly with selected external agents, typically laser light, both during the gate operations and the preparation and readout steps, but should be shielded from any unwanted interaction if the desired dynamics is to be preserved. Perfect, selective, shielding from the environment is not possible in practice and unwanted interactions with the surroundings will unavoidably result in some form of decoherence. In this article we will review the fundamentals of the implementation of a universal set of quantum gates using (i) atomic system coupled to a resonant field sustained in a microwave cavity and (ii) trapped ions confined in an electromagnetic trap. Atomic and ionic systems provide a good means for storing quantum information in long-lived electronic states. In addition, both neutral atoms and ions can be selectively detected with high efficiency. While none of the presented schemes is expected to be scalable to a very large number of qubits, they constitute the first prototypes for quantum registers and are expected to play an important role in applications where only a moderate number of qubits are required. These applications range from the development of new frequency standards to quantum communication protocols. We will discuss in the following sections how a set of universal gates can build within these systems and will present the state of the art of the experiments performed to date.

Quantum Computation: Basic Definitions

As stated before, the basic unit of quantum information is a qubit, an arbitrary quantum mechanical two-level system. We will associate logical values 0 and 1 with quantum states $|0\rangle$ and $|1\rangle$, the

eigenvectors of an observable such as the energy, and refer to such states as the computational basis. Any unitary operation on a qubit will be called a single qubit gate. Joint operations on more than one qubit will be denoted as multi-qubit gates. Using the fact that any two-level system is isomorphic to a spin-1/2 particle (e.g., an electron), we can always think of a qubit in terms of a certain pseudo-spin whose dynamics can be represented as rotation of the associated Bloch vector within the Bloch sphere.

For instance, as depicted in Figure 1, a qubit initially prepared in state $|1\rangle$, after experiencing a single qubit gate will be represented by certain Bloch vector $\vec{S}$, whose location on the Bloch sphere is a function of the gate parameters. Note that a classical system can only assume two positions (upwards and downwards), while a quantum system can explore the whole surface of the sphere.

It is also quite standard in quantum computation to use a diagramatic representation in terms of quantum circuits. Quantum circuits are built up of wires to denote the quantum bits and boxes to represent the gates, or quantum operations. Figure 2 shows two basics elements of a quantum circuit. The element on the left is the so-called Hadamard transformation, a particular example of single-qubit gate which transform each element of the computational basis as follows:

$$|0\rangle \rightarrow \frac{1}{\sqrt{2}}\left(|0\rangle + |1\rangle\right)$$
$$|1\rangle \rightarrow \frac{1}{\sqrt{2}}\left(|0\rangle - |1\rangle\right) \qquad [1]$$

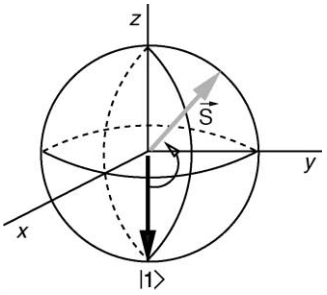


Figure 1 Geometrical representation of qubit systems using the Bloch sphere formalism. See the text for details.

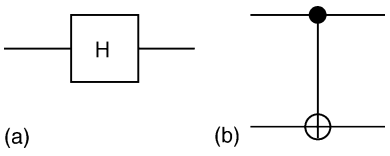


Figure 2 Examples of single qubit gates (a) Hadamard transformation and two qubit gates (b) CNOT gate as customarily represented in quantum circuits diagrams.

Part (b) of Figure 2 depicts a CNOT-gate. This is now an operation involving two qubits. The first one, the control qubit, remains unaltered after the gate operation while the second qubit, referred to as the target, undergoes a spin flip (logical NOT operation) only when the control is in the state $|1\rangle$. Specifically, the action of a CNOT-gate is defined by the following transformation rule:

$$\begin{aligned} &\text{CNOT} \\ &|00\rangle \rightarrow |00\rangle \\ &|01\rangle \rightarrow |01\rangle \\ &|10\rangle \rightarrow |11\rangle \\ &|11\rangle \rightarrow |10\rangle \end{aligned} \qquad [2]$$

We have already mentioned that single qubit rotations and CNOT gates provide a universal set of gates. Our aim now is to show how such operations can be built when the so far generic qubit are realized in terms of real atomic systems. In the two experimental scenarios we are going to analyze here, the CNOT operation will be achieved by means of a different type of gate, a controlled-phase gate supplemented by two one-qubit Hadamard gates on the target qubit. The arrangement is showed in Figure 3. The controlled-phase gate is again a two-qubit operation defined by the following transformation rule:

$$\begin{aligned} &\text{CNOT} \\ &|00\rangle \rightarrow |00\rangle \\ &|01\rangle \rightarrow |01\rangle \\ &|10\rangle \rightarrow |10\rangle \\ &|11\rangle \rightarrow -|11\rangle, \end{aligned} \qquad [3]$$

i.e., a global phase equal to π is acquired if, and only if, the two basis states are in the logical state 1. It is easy to check that when the target bit is rotated by a Hadamard gate prior and subsequently to the action

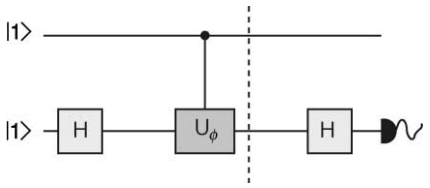


Figure 3 Controlled-phase gate with both inputs, control, and target qubits, in state $|1\rangle$. The first Hadamard gate rotates the target to state $(|0\rangle - |1\rangle)/\sqrt{2}$. The action of the controlled-phase gate with $\phi = \pi$ yields the joint state $|1\rangle + (|0\rangle - |1\rangle)/\sqrt{2}$ and a final Hadamard rotation on the target leaves the system in the product state $|1\rangle|0\rangle$.

of the joint phase gate, the resulting state is tantamount to the action of a single CNOT gate. Therefore, to prove that our atomic system allows for the implementation of a universal set of gates, it suffices to show that both single qubit rotations and controlled phase gates can be implemented with sufficient accuracy.

The next section analyzes a typical cavity QED setup and presents the basic physical ideas to understand how single and two-qubit gates can be implemented within atomic systems interacting with radiation fields. Subsequent sections will rely heavily in this analysis.

Quantum Computing with Atoms I: Cavity QED

A simplified cavity-QED arrangement for performing both single- and two-qubit gate operations is depicted in Figure 4. This configuration is well-known to quantum opticians and it is often referred to as a Ramsey interferometric setup. For concreteness, we will focus here in the experimental set up currently in operation at the ENS in Paris. There, the atomic system is a highly excited Rydberg atom initially prepared in a long-lived circular state. The relevant atomic states are represented in Figure 5, with the qubit being stored in levels $|i\rangle$ and $|g\rangle$ which will play the role of the general basis states $|0\rangle$ and $|1\rangle$.

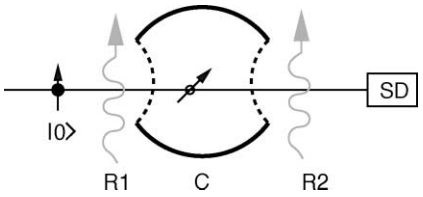


Figure 4 Cavity QED setup for single-qubit and two-qubit operations. Rotations of the atomic pseudo spin are performed via interation with classical radiation (Zone R), while the interaction with a resonant cavity field sustained in C allows the realization of a phase gate where the cavity photon acts as the control qubit.

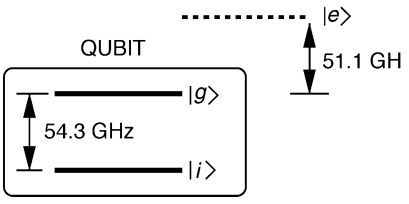


Figure 5 Circular Rydberg levels involved in the ENS experimental setup. The qubit is stored in levels $|i\rangle$ (logical value 0) and $|g\rangle$ (logical value 1). The radiative lifetime of these levels lies in the millisecond range. Dissipation can then be neglected along the 20 cm path inside the apparatus.

Regions denoted by R represent the interaction of the atomic system with classical microwave radiation. As detailed below, this interaction will allow arbitrary rotations of the atomic pseudo spin. Region C represents the interaction of the atomic system with a single mode cavity field resonant the atomic transition $g - e$. The highest Q value reported to date is of the order of 10^8, which corresponds to a photon storage time of a few milliseconds. Therefore, the relaxation time of the cavity field is much longer than the atom-cavity interaction time, which defines the two-qubit gate operation time and is of the order of a few tens of μs. The cavity field can be prepared in either of the Fock states $|0\rangle_\gamma$ or $|1\rangle_\gamma$ and will play the role of the control qubit for the realization of a phase gate, as detailed below. This set up is completed with a detection scheme (SD in the figure) that allows the selective detection of the atomic electronic state.

We will start our discussion showing how to perform an arbitrary single-qubit gate by means of irradiating the atomic system with classical radiation. We will keep the analysis in very general terms with the aim of using it later in the description of the implementation of quantum logic with trapped ions.

Single-Qubit Rotations

Let our qubit be a two-level atomic system with frequency separation ω_0. We will show that an arbitrary rotation can be achieved by means of irradiating the system with classical light, e.g., microwave radiation or a laser field, with tunable frequency ω_L. The radiation field will be specified by two controllable parameters, its Rabi frequency Ω and its phase ϕ. The atom-field Hamiltonian can be written as

$$H = \frac{\hbar\omega_0}{2}\left(|1\rangle\langle 1| - |0\rangle\langle 0|\right) + \hbar\left(g e^{i\omega_L t}|0\rangle\langle 1| + g^* e^{-i\omega_L t}|1\rangle\langle 0|\right) \qquad [4]$$

where the first term corresponds to the atomic internal energy and the second describes the coupling with the external field. The coupling constant g can be written in terms of the field parameters as

$$g = \frac{i\Omega}{2} e^{-i\phi} \qquad [5]$$

In order to write down a time-independent Hamiltonian, we will move to an interaction picture with

respect to $H_0 = \hbar\omega_L/2(|1\rangle\langle 1| - |0\rangle\langle 0|)$. In this frame,

$$H_{\text{int}} = \frac{\hbar\delta}{2}\Big(|1\rangle\langle 1| - |0\rangle\langle 0|\Big) + \hbar\Big(g|0\rangle\langle 1| + g^*|1\rangle\langle 0|\Big) \quad [6]$$

The interaction Hamiltonian can be readily expressed in terms of the Pauli spin operators

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_y = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix} \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad [7]$$

as

$$H_{\text{int}} = \frac{\hbar\delta}{2}\sigma_z + \frac{\hbar\Omega}{2}(\sin\phi\,\sigma_x - \cos\phi\,\sigma_y) \quad [8]$$

where $\delta = \omega_L - \omega_0$ denotes the detuning from the atomic transition and we have taken into account the explicit form of g as given by eqn [5]. We can now easily derive the time evolution operator by noting that H_{int} can be written in a compact form as $H_{\text{int}} = \vec{\sigma}\cdot\vec{n}$, where the vector $\vec{n}$ has cartesian coordinates $\vec{n} = (\Omega/2\sin\phi, -\Omega/2\cos\phi, \delta/2)$. Note that $|\vec{n}| \equiv \Omega_g = 1/2\sqrt{\Omega^2 + \delta^2}$. The time evolution operator can then be written as follows:

$$U = e^{-iH_{\text{int}}t} = \cos(|\vec{n}|t)1\!1 - i\sin(|\vec{n}|t)\vec{\sigma}\cdot\vec{n}/|\vec{n}| \quad [9]$$

In matrix form:

$$U = \begin{pmatrix} \cos\left(\frac{\Omega_g t}{2}\right) - i\frac{\delta}{\Omega_g}\sin\left(\frac{\Omega_g t}{2}\right) & \frac{\Omega}{\Omega_g}\sin\left(\frac{\Omega_g t}{2}\right)e^{-i\phi} \\ -\frac{\Omega}{\Omega_g}\sin\left(\frac{\Omega_g t}{2}\right)e^{i\phi} & \cos\left(\frac{\Omega_g t}{2}\right) + i\frac{\delta}{\Omega_g}\sin\left(\frac{\Omega_g t}{2}\right) \end{pmatrix} \quad [10]$$

In the limit of large Rabi frequency, where $\Omega \gg \delta$, we can write the following approximate expressions for the time evolution of the computational basis states

$$\begin{aligned} U|0\rangle &= \cos\left(\frac{\Omega_g t}{2}\right)|0\rangle - \sin\left(\frac{\Omega_g t}{2}\right)e^{i\phi}|1\rangle \\ U|1\rangle &= \sin\left(\frac{\Omega_g t}{2}\right)e^{-i\phi}|0\rangle + \cos\left(\frac{\Omega_g t}{2}\right)|1\rangle \end{aligned} \quad [11]$$

The transformed states by U are eigenstates of the spin operator along the direction $\vec{n}$ which is a function of the external field parameters. Therefore, by adjusting the Rabi frequency and the field's phase, it is possible to transform the basis states into spin eigenstates along any arbitrary direction, and to perform an arbitrary rotation of our qubit, we simply have to tailor appropriately the parameters of the driving field. For example, when the Rabi frequency and the interaction time are chosen such that $\Omega t/2 = \pi/4$, and the field's phase is taken equal to π, we get

$$\begin{aligned} U|0\rangle &= \frac{1}{2}\Big(|0\rangle + |1\rangle\Big) \\ U|1\rangle &= \frac{1}{2}\Big(|0\rangle - |1\rangle\Big) \end{aligned} \quad [12]$$

This relationship defines the Hadamard transformation H introduced before. Alternatively, using quantum optics jargon, we have performed a $\pi/2$-pulse. Note that using the Bloch sphere representation, we can visualize the action of a Hadamard gate on the computational basis states as a rotations that brings them to the equatorial plane of the Bloch sphere, the transformed states by H now lying in opposite directions along the x-axis.

The next step is to show that our system also allows the implementation of conditional dynamics and, in particular, allows to build a controlled-phase gate.

Conditional Dynamics: Controlled-Phase Gate

The interaction of the atomic system with the cavity field sustained by the microwave cavity is described by the Jaynes–Cummings (JC) Hamiltonian. We can see this Hamilton operator as a generalization of the operator defined by eqn [4], where the radiation field is now quantized and the terms $e^{\pm i\omega_L t}$ are replaced by the operators a and $a^\dagger$ corresponding to the annihilation and creation of a cavity photon of frequency ω. The interaction with the quantized cavity field couples levels $|e\rangle$ and $|g\rangle$, while the far detuned level $|i\rangle$ will now play the role of a passive spectator. Remember that our atomic qubit is stored in levels $|i\rangle$ and $|g\rangle$. To avoid confusion with the notation, we will write the Hamilton operator in terms of the atomic states e and g. Explicitly:

$$\begin{aligned} H_{\text{JC}} = &\frac{\hbar\omega_0}{2}\Big(|e\rangle\langle e| - |g\rangle\langle g|\Big) + \hbar\omega a^\dagger a \\ &- i\frac{\hbar\Omega}{2}\Big(a|g\rangle\langle e| - a^\dagger|g\rangle\langle e|\Big) \end{aligned} \quad [13]$$

where the coupling constant g has been taken to be $g = -i(\hbar\Omega/2)$, proportional to the Rabi frequency averaged over the mode structure, Ω. Using the same techniques employed in the previous section, we can derive the corresponding JC time evolution operator,

which yields, for exact resonance, the following transformation rules:

$$\begin{aligned} |g\rangle|0\rangle_\gamma &\rightarrow |g\rangle|0\rangle_\gamma \\ |g\rangle|1\rangle_\gamma &\rightarrow \cos\frac{\Omega t}{2}|g\rangle|1\rangle_\gamma - \sin\frac{\Omega t}{2}|e\rangle|0\rangle_\gamma \\ |e\rangle|0\rangle_\gamma &\rightarrow \cos\frac{\Omega t}{2}|e\rangle|0\rangle_\gamma + \sin\frac{\Omega t}{2}|g\rangle|1\rangle_\gamma \end{aligned} \quad [14]$$

where states $|\rangle_\gamma$ refer to the cavity field. Note that:

- The cavity field will play the role of our control qubit in the performance of a phase gate, the target qubit being the atomic system with computational states defined as before. The computational basis for the photonic qubit is provided by the Fock states $|0\rangle_\gamma$ and $|1\rangle_\gamma$.
- Strictly speaking, the time t entering in the equations above corresponds to an effective interaction time which accounts for the spatial variation of the coupling along the field mode.
- When the atom enters the cavity in state $|i\rangle$, it will acquire a global phase proportional to the ratio Ω^2/δ_i, where δ_i denotes the detuning between the atomic transition $e \leftrightarrow i$ and the mode frequency ω. In conditions of resonance, or quasi-resonance, this factor can be safely neglected, which justifies the term *passive spectator* used before.

We now have all the ingredients to understand how we can perform a controlled phase gate. Consider the situation where $\Omega t = 2\pi$, so the system experiences a full Rabi oscillation. According to the previous discussion, the basis states of two-qubit system atom-cavity field will transform as follows:

$$\begin{aligned} |i\rangle|0\rangle_\gamma &\rightarrow |i\rangle|0\rangle_\gamma \\ |i\rangle|1\rangle_\gamma &\rightarrow |g\rangle|0\rangle_\gamma \\ |g\rangle|0\rangle_\gamma &\rightarrow |g\rangle|0\rangle_\gamma \\ |g\rangle|1\rangle_\gamma &\rightarrow -|g\rangle|1\rangle_\gamma \end{aligned} \quad [15]$$

That is, whenever the atomic system enters the cavity in state $|i\rangle$, it will be left unaffected, irrespective of the photon state. However, if the atom enters the cavity in state $|g\rangle$, the system will undergo a phase shift equal to π conditional to the photon number. And that is precisely the dynamics required for a controlled phase gate as defined earlier.

A detailed account of the experiments in cavity QED, showing coherent gate operation, can be found in Further Reading. These experiments prove the working principle of a universal set of quantum logic gates with neutral atoms and provide the basic elements for the development of a cavity QED quantum register in the near future.

Quantum Computing with Atoms II: Ion Traps

Charged atomic particles, ions, provided the first atomic system where conditional dynamics was demonstrated experimentally. The theoretical proposal in 1995, by Cirac and Zoller, for performing a two-qubit controlled NOT gate between trapped ions cooled to their motional ground state, was followed a few months later by a pioneering experiment at NIST, in Colorado. While the physics involved in the description of laser cooled and trapped ions is conceptually more involved than the dynamics ruled by a purely JC Hamiltonian, we can review the rudiments of the Cirac–Zoller scheme on the basis of the techniques developed in the previous sections. Nevertheless, as discussed briefly at the end of this section, we should point out that the exact procedure followed in the actual experiments departs from the analysis presented here.

Figure 6 depicts a linear ion trap where a string of ions is confined in an essentially 1-D structure. Quantum bits are stored in two long-lived hyperfine states which can be driven with laser light of controllable Rabi frequency Ω and phase ϕ. Ideally, the separation d between ions exceeds the wavelength of the external field so that ions can be addressed individually. Detection of the atomic state relies on the well mastered quantum jump technique which provides an efficiency close to 100%. From our previous discussion, it is clear that we already have all the ingredients for performing arbitrary single qubits gates, but we still need to provide a procedure to achieve two-qubit gate operations. As detailed in the next section, this will be achieved using the motional degrees of freedom of the ions as a bus for information transfer between an arbitrary pair of ions.

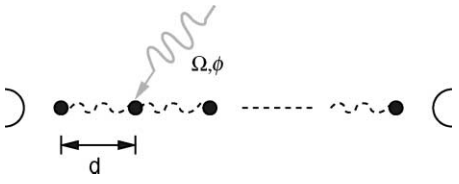


Figure 6 Schematic representation of a linear ion trap. Ions are individually addressed by external laser fields to perform single qubit operations, while two-qubit gates are mediated by the coupling to the motional degrees of freedom, depicted here as mechanical strings linking the chain of ions in the trap. The exact duration of a two-qubit gate operation and the characteristic decoherence time depend very much on the trap design, but orientative values are in the range of 50 μs and 100 − 1000 μs respectively.

The Cirac and Zoller Quantum Computer

At sufficiently low temperatures, the process of absorption and emission of photons produces a non-negligible recoil of the atomic system and the ion's motion can no longer be treated classically. In this regime, we can think of the string of ions as a system of coupled quantum harmonic oscillators whose normal mode configurations correspond to the collective motional states of the ions. The lowest energy solution corresponds to an oscillation where all ions are in phase (center of mass (CM) mode) and will be the one considered in our description. Figure 7 depicts the level scheme for ions initially cooled down to the motional ground state, illustrating both internal levels, $|e\rangle$ and $|g\rangle$, and motional (vibrational) states, denoted by $0, 1, \ldots$ Laser radiation can be tuned to be resonant with the $|g\rangle|0\rangle_{cm} \leftrightarrow |e\rangle|0\rangle_{cm}$ transition (carrier transition), but may also be tuned to drive transitions with different vibrational states (sideband transition), as represented by the external arrows in the figure. The realization of single qubit gates simply involves tuning the laser towards the carrier transition. The interaction is ruled by essentially the same Hamiltonian discussed above. Therefore, choosing appropriately the laser parameters, it is possible to induce an arbitrary rotation of the ion pseudo-spin.

To prove that the system allows the implementation of a universal set of gates, we will now show that it is possible to operate a controlled phase gate between two arbitrary ions. For that we need to couple internal and motional degrees of freedom which will be done by means of tuning the laser light to a sideband transition. The corresponding Hamiltonian will not be derived here in detail but it can be understood as describing a Jaynes–Cummings

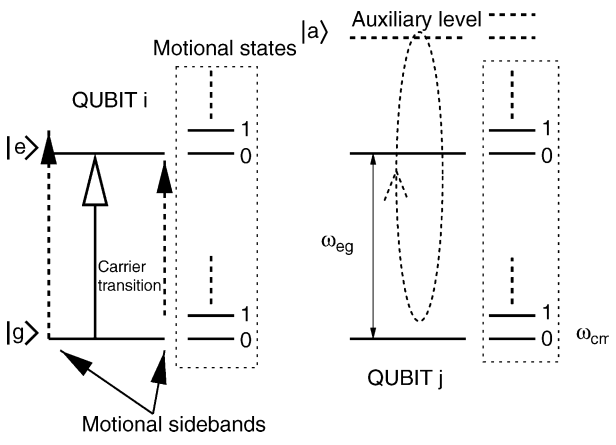


Figure 7 Relevant internal and motional atomic levels for the realization of a controlled NOT gate between two arbitrary ions within a chain. See the text for details.

form of dynamics where the quantized electromagnetic field within a cavity is replaced by the quantum mechanical oscillator describing the vibrational states of the ion. Explicitly:

$$H = \frac{\eta\Omega}{2}\left(c|e\rangle\langle g| + c^{\dagger}|g\rangle\langle e|\right) \qquad [16]$$

where c and $c^{\dagger}$ denote the annihilation and creation operators of vibrational quanta and η is the so-called Lamb–Dicke parameter, a measure of the recoil energy. The gate operation involves the following sequence:

- Mapping the internal state of the control ion into the collective motional state of the chain,
- performing an operation on the target ion conditional to the excitation number of the vibrational state and
- reverse the initial operation restoring the vibrational state to its ground state.

Imagine ion i, which will play the role of the control, and j, the target, initially prepared in arbitrary superpositions of the corresponding internal states, with the whole system cooled down to their motional ground state $|0\rangle_{cm}$. The joint state of the two particles can be written as a product state of the form:

$$\left(\alpha_i|g\rangle_i + \beta_i|e\rangle_i\right) \otimes |0\rangle_{cm} \otimes \left(\alpha_j|g\rangle_j + \beta_j|e\rangle_j\right) \qquad [17]$$

When ion i is irradiated by a laser tuned to the first motional side band $|g\rangle|1\rangle_{cm} \leftrightarrow |e\rangle|0\rangle_{cm}$, and the timing is such that the system undergoes a π-Rabi rotation, the internal state of the ion i is mapped onto the vibrational state (Note the transformation rules given in eqn [14]. As a result, the joint state of the system is still factorizable as

$$\left(|g\rangle_i \otimes \alpha_i|0\rangle_{cm} + \beta_i|1\rangle_{cm}\right) \otimes \left(\alpha_j|g\rangle_j + \beta_j|e\rangle_j\right) \qquad [18]$$

The second step of the protocol consists of applying an operation on ion j conditional on the vibrational state of the joint system. For that we apply a laser pulse resonant with the transition $|e\rangle_j|1\rangle_{cm} \leftrightarrow |a\rangle_j|0\rangle_{cm}$ with an auxiliary level $|a\rangle_j$ so that ion j undergoes a full Rabi oscillation. This results in acquiring a phase shift equal to π only if the system is state $|e\rangle_j|1\rangle_{cm}$ (see eqn [15]). This step entangles the internal and vibrational degrees of freedom of ion j, leading to a joint state of

the form:

$$|g\rangle_i \otimes \Big(\alpha_i|0\rangle_{cm}(\alpha_j|g\rangle_j + \beta_j|e\rangle_j) + \beta_i|1\rangle_{cm}(\alpha_j|g\rangle_j - \beta_j|e\rangle_j)\Big) \quad [19]$$

Applying a final laser pulse on ion i identical to the initial one, the state of the vibrational mode is transferred back to ion i and the atomic systems are left in the, generally, entangled state

$$\alpha_i\alpha_j|g\rangle_i|g\rangle_j + \alpha_i\beta_j|g\rangle_i|e\rangle_i + \beta_i\alpha_j|e\rangle_i|g\rangle_j - \beta_i\beta_j|e\rangle_i|e\rangle_j, \quad [20]$$

which completes the controlled-phase gate operation.

A simplified version of this scheme shortly followed the theoretical proposal by Cirac and Zoller, proving the working principle of the ions trap quantum processor. Since then, both theoretical and experimental progress have allowed the development of schemes for implementing faster gates as well as protocols which do not require cooling to the motional ground state. We will briefly revise the current state of the art and future prospects of this technology in the last section of this article.

Frequency Estimations as a Quantum Computation: Building Frequency Standards with Entangled States

Recently, it has been proposed to make use of entangled states of trapped ions for performing high-resolution spectroscopy beyond the shot noise limit. This boundary is the ultimate limit achievable when performing a frequency measurement using uncorrelated particles. Consider n atomic two-level systems with frequency separation ω_0 initially prepared in their ground state $|0\rangle$. The aim is to lock a certain oscillator of frequency ω_L to the atomic frequency. To do that, the currently employed protocols use the so-called Ramsey method of separate oscillating fields, where each atomic system undergoes two $\pi/2$-pulses (or Hadamard transformations) separated by a free time evolution of duration t, much larger than the duration of the pulse. Looking at the Hamiltonian describing the interaction between the ion and the field, eqn [1], one can easily see that the free time evolution can be understood as a controlled phase gate in the following sense. If we set $g = 0$ in eqn [1] and for simplicity re-scale the ion internal energies so that $H = \hbar\delta|1\rangle\langle 1|$, the free evolution operator can be written as $U = -\hbar\delta|1\rangle\langle 1|$, so that the atomic basis states evolve as:

$$\begin{aligned} U|0\rangle &= |0\rangle \\ U|1\rangle &= e^{-i\delta t}|1\rangle \end{aligned} \quad [21]$$

In other words, the ground state remains invariant, whereas the upper levels pick up a phase which is proportional to the detuning from the atomic transition. The state of each particle after the time evolution can then be written as

$$\frac{1}{\sqrt{2}}\Big(|0\rangle + e^{-i\delta t}|1\rangle\Big) \quad [22]$$

After the second pulse, which again mixes populations of levels $|0\rangle$ and $|1\rangle$, one measures the population of, say, level $|1\rangle$. The signal is therefore a sinusoidal function of the form

$$P_1 = \frac{1}{2}\Big(1 + \cos(\delta t)\Big) \quad [23]$$

which is symmetrical with respect to $\delta = 0$. This procedure is repeated a number of times to accumulate a sufficient data. Let us call T the total duration of the procedure, so that we accumulate $N = nT/t$ experimental data. Finally, whenever statistics in the populations are different from 50:50 at half maximum are detected, the oscillator frequency is steered to achieve exact resonance. The ultimate resolution $\Delta\omega_0$ achievable with this technique is limited simply by projection noise and can be written as

$$\Delta\omega_0 = \frac{1}{\sqrt{nTt}} \quad [24]$$

After time T, the oscillator's frequency is steered towards the resonance and new sequence starts. We will now illustrate how the use of entanglement can yield an uncertainty in the measurement beyond the shot noise limit eqn [24]. Consider the quantum network depicted in Figure 8. The n particles are now initially prepared in a maximally entangled state. This is achieved by means of subjecting the first ion to

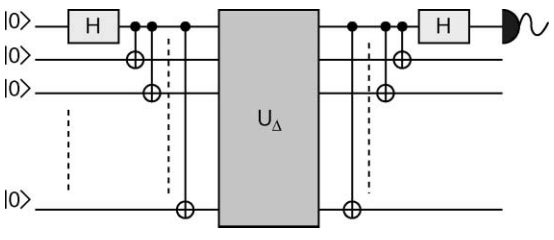


Figure 8 Quantum circuit representation for an experimental setup for a frequency estimation using maximally entangled states.

a Hadamard transformation followed by a cascade of CNOTs between the first ion, which acts as a control, and each of the remaining $(n-1)$ particles.

The resulting state is of the form

$$\frac{1}{\sqrt{2}}\Big(|0\ldots0\rangle + |1\ldots1\rangle\Big) \qquad [25]$$

The systems now evolve freely for a time t and, in light of the discussion above, we can represent this evolution in terms of a quantum gate which will leave state $|0\rangle$ invariant while introducing a phase factor proportional to the detuning if the system is in state $|1\rangle$. As a result, the joint state of the n particle after the free time evolution is given by

$$\frac{1}{\sqrt{2}}\Big(|0\ldots0\rangle + e^{-in\delta t}|1\ldots1\rangle\Big) \qquad [26]$$

with a phase factor amplified by a factor n as compared with the single particle state given by eqn [22]. The second part of the network disentangles the n particles, leaving the first ion in the state $(1/\sqrt{2})(|0\rangle + e^{-in\delta t}|1\rangle)$. If we now subject it to an additional Hadamard gate, the probability for the ion to be excited is

$$P_{1,ent} = \frac{1}{2}\Big(1 + \cos(n\delta t)\Big) \qquad [27]$$

which oscillates n times faster than the signal eqn [23]. The corresponding uncertainty for the estimation of the atomic frequency is now

$$\Delta\omega_0, ent = \frac{1}{n\sqrt{Tt}} \qquad [28]$$

This corresponds to an increase in the resolution of $\sqrt{n}$ beyond the shot noise bound. In practice, the achievable resolution using entangles states is limited by the decoherence time of the system t_{dec}, which sets the condition $T < t_{\mathrm{dec}}/2$ for the procedure to be valid. Remarkably, entanglement is still advantageous in a situation where T is chosen such that it exceeds the decoherence time. However, maximally entangled states should now be replaced by certain partially entangled ones to overcome the precision achievable with uncorrelated particles.

The application of entangled states in precision spectroscopy is a nice illustration of the potential of the ideas developed in quantum information beyond the ultimate task of building a large-scale quantum computer. The first experimental results showing the enhanced performance of entangled states have been reported by the NIST group in Boulder.

Conclusion and Perspectives

We have revised the fundamental concepts for the practical realization of quantum computation using atomic systems. We have showed how single qubit and two-qubit controlled-phase gates can be implemented using atoms interacting with the cavity field sustained by a microwave resonator and ion traps. We have illustrated how a very simple quantum computation can be used for the development of more precise frequency standards, emphasizing the potential of the ideas developed in quantum information beyond the implementation of large scale quantum computation.

Cavity QED technology has so far allowed the generation of entangled states of two and three particles and has demonstrated coherent gate operation. Ion trap technology has pushed controllable multi-particle entanglement to four particles and has showed the possibility of implementing forms of error correction. The combination of these two technologies offers interesting prospects for the near future. Surrounding traps by optical cavities and networking these devices via optical fibers is expected to allow the construction of the first prototypes for a quantum communication network.

See also

Quantum Optics: Entanglement and Quantum Information.

Further Reading

Cirac JI and Zoller P (1995) Quantum computations with cold trapped ions. *Physical Review Letters* 74: 4091.

Deutsch D, Barenco A and Ekert AK (1995) Universality in quantum computation. *Proc. Roy. Soc. London A MAT* 449: 669–677.

Grover L (1999) Quantum mechanics helps in searching for a needle in a haystack. *Physical Review Letters* 79: 325–328.

Huelga SF, *et al.* (1997) Improvement of frequency standards with quantum entanglement. *Physical Review Letters* 79: 3865.

Jaynes ET and Cummings FW (1963) Comparison of quantum and semiclassical radiation theories with application to the beam maser. *Proc. IEEE* 51: 89.

Kielpinski D, Meyer V, Rowe MA, *et al.* (2001) A decoherence-free quantum memory using trapped ions. *Science* 291: 1013–1015.

Monroe C, Meekhof DM, King BE, *et al.* (1995) Demonstration of a fundamental quantum logic gate. *Physical Review Letters* 75: 4714–4717.

Raimond JM, Brune M and Haroche S (2001) *Colloquium*: manipulating quantum entanglement with atoms and

photons in a cavity. *Review Modern Physics* 73: 565 and references therein.

Sackett CA, Kielpinski D, King BE, *et al.* (2000) Experimental entanglement of four particles. *Nature* 404: 256–259.

Shor P (1994) *Proc. of the 35th Annual Symposium on the Foundations of Computer Science, IEEE Computer Society, Los Alamitos*. Santa Fe, NM: IEEE Computer Society Press.

Sorensen A and Molmer K (1999) Quantum computation with ions in thermal motion. *Physical Review Letters* 82: 1971.

Wineland DJ, Monroe C, Itano WM, Leibfried D, King BE and Meekhof DM (1998) Experimental issues in coherent quantum-state manipulation of trapped atomic ions. *J. Res. Natl. Inst. Stand. Tech.* 103: 259 and references therein.

Squeezed Phonons in Solids

A V Bragas, Universidad de Buenos Aires, Buenos Aires, Argentina

R Merlin, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The reduction of noise is crucial for measurements requiring high accuracy and, often, for the observation of new phenomena. Among noise sources, noise inherently related to the quantum nature of an object persists even after all the classical sources of noise are removed. Quantum noise is a fundamental limitation, loosely related to the Heisenberg uncertainty principle, which imposes a lower limit to the fluctuations of any pair of conjugated variables, like position and momentum. For conjugated variables A_1 and A_2 we have

$$\Delta A_1 \Delta A_2 \geq \frac{\hbar}{2} \qquad [1]$$

where

$$\Delta A_i = \sqrt{\langle A_i^2 \rangle - \langle A_i \rangle^2} \qquad i = 1, 2 \qquad [2]$$

While quantum mechanics does not forbid reducing the noise in one of the conjugated observables down to zero, the noise for the other variable will increase according to eqn [1]. Squeezing refers to the reduction of noise of a quantum variable at the expenses of its conjugated variable. The term originates from quantum optics. Nonlinear interactions of the electromagnetic field with passive and active atomic media have been successfully used to generate squeezed photons, opening new possibilities for essentially noiseless optical communications and precision measurements. The electromagnetic field is not the only quantum object that can be squeezed. By performing appropriate manipulations of the internal quantum state of a system, it is, in principle, possible to squeeze an arbitrary variable. In the past few years, vibrational states have also been experimentally squeezed in molecules and solids, by exciting the system with ultrafast optical pulses. The concept of squeezing can also be extended to an ensemble of classical objects for which the fluctuations are of thermal origin.

Squeezing applies in particular to situations where the relevant variables are the position Q and momentum P of a harmonic oscillator. Consider the Hamiltonian of a particle of mass m in a harmonic potential of frequency Ω

$$H_0 = \frac{P^2}{2m} + \frac{m}{2}\Omega^2 Q^2 \qquad [3]$$

The ground state wavefunction is

$$\varphi_0 = \frac{1}{\left[2\langle Q_0^2 \rangle \pi\right]^{1/4}} \exp\left(-\frac{Q^2}{4\langle Q_0^2 \rangle}\right) \qquad [4]$$

where $\langle Q_0^2 \rangle = \hbar/2m\Omega$. The stationary probability, $|\varphi_0|^2$, is represented in Figure 1a. This state is also a minimal uncertainty state in that $\Delta Q \Delta P \equiv \hbar/2$. $\langle Q_0^2 \rangle$ gives a measure of the quantum noise and it is known as the vacuum noise level.

Let us now discuss the behavior of a harmonic oscillator that interacts with an external source. There are two types of quantum states for which the Gaussian shape of the wavefunction is preserved after interacting with an external driving force. These are the coherent state and the squeezed state. If the interaction places the system in a state of the same form as eqn [4] but displaced with respect to the center, the wavepacket will oscillate periodically, with period $\tau = 2\pi/\Omega$, without changing its shape as shown in Figure 1b. This quantum state is known as the coherent state. It has the same uncertainty as the ground state but the expectation value $\langle Q(t) \rangle$ depends on the time following the classical trajectory in the harmonic potential. If, instead, we introduce a sudden change in the spring constant at $t = 0$, the

wavepacket will begin to 'breathe' for $t > 0$, so that its width will vary periodically with a period given by $\tau/2$. This is the squeezed state. Here the noise of one of the variables is periodically reduced below the level of the vacuum noise (Figure 1c). A state that is a combination of a coherent and a squeezed state is shown in Figure 1d. The wavepacket follows the classical trajectory and breathes at the same time.

The results in Figure 1 apply only to harmonic oscillators. While the concept of squeezing can be extended to anharmonic potentials, there are additional features to consider such as, for example, wavepacket revivals.

Dynamics of a Parametrically Excited Crystal

Harmonic oscillators describe a wide range of physical systems. In particular, the dynamics of a crystal lattice can be described in terms of a set of independent harmonic oscillators, the quantum nature of which is described in terms of phonons. In this section we discuss interactions of a harmonic oscillator with external sources that produce squeezing.

The normal modes in a crystal obey equations of the form:

$$\ddot{Q}_q + \Omega_q^2 Q_q = 0 \qquad [5]$$

where Q_q and Ω_q are the phonon coordinate and frequency, and q is the wavevector.

As briefly mentioned in the previous section, a sudden change in the spring constant produces a squeeze state. To do this, we need what is known as time-dependent parametric coupling. As in the experiments on phonon squeezing, using ultrashort optical pulses, we consider an impulsive driving force proportional to Q. The relevant equation is:

$$\ddot{Q}_q + \Omega_q^2 Q_q = \lambda_q Q_q \delta(t) \qquad [6]$$

where $\delta(t)$ is the Dirac delta function, and λ accounts for the strength of the external impulsive force. This impulsive excitation induces a sudden change in momentum, proportional to the normal coordinate Q at $t = 0$, as can be shown by integrating eqn [6]. Explicitly, $\Delta\dot{Q}_q = \Delta P_q = \lambda_q Q_q(0)$. $\Delta P_q = P_q(0^+) - P_q(0^-)$, where $0^-(0^+)$ is the time immediately before (after) the excitation. This result is illustrated in Figure 2. The main effect of the interaction is to change the width of the Gaussian packet or variance of the distribution. This leads to a time-dependence of the noise defined in eqn [2], $\Delta Q^2(t) = \langle Q^2(t)\rangle$ reflecting the 'breathing' of the wavepacket. Since P_q is the conjugated variable of Q_q, $\langle P_q^2(t)\rangle$ shows the same time-dependent behavior as $\langle Q_q^2(t)\rangle$ but in quadrature with it, so as to fulfill eqn [2] at all time. Assuming an oscillator initially in its ground state, the solution of eqn [6] is

$$\langle Q_q^2(t)\rangle = \frac{\hbar}{2m\Omega_q}\left[1 + \left(\frac{\lambda_q}{\Omega_q}\right)\sin(2\Omega_q t) + \left(\frac{\lambda_q}{\Omega_q}\right)^2 \sin^2(\Omega_q t)\right] \qquad [7]$$

It is then clear that the fluctuations of Q_q periodically fall below the vacuum quantum noise level, determined by $\langle Q_0^2\rangle$, and that the oscillations have a frequency that is twice the frequency of the oscillator.

The considerations above apply to a quantum harmonic oscillator at temperature $T = 0$. In the classical case, we need to consider an ensemble of oscillators in equilibrium with a thermal bath at temperature T. It can be shown that eqn [7] also applies to the classical case, provided $\hbar/2m\Omega$ is replaced by $k_B T/\Omega^2 m$, where k_B is the Boltzman constant. In classical squeezing, the external perturbation is applied simultaneously to the ensemble, leading to a sudden change of the momentum of all the oscillators. The jump in momentum is proportional to the position of the particle at $t = 0$. After the interaction, the noise of the ensemble will fall periodically below the thermal noise at temperature T given by $k_B T/\Omega^2 m$. A pictorial representation of classical squeezing is shown in Figure 3.

To describe a more realistic situation, it is necessary to take into account effects due to dissipation. Theoretically, this is usually accomplished by considering the coupling with a continuum of external modes simulating a thermal bath. The solution to this type of problem is nontrivial even for the simplest case of linear coupling to the bath.

The impulsive excitation of eqn [6] is a parametric excitation containing all the frequencies. A more common way to parametrically excite an oscillator is to use a monochromatic source, as described by the equation

$$\ddot{Q} + [\Omega^2 + \varepsilon\cos(\Omega_0 t)]Q = 0 \qquad [8]$$

where the parametric excitation is characterized by the amplitude ε and the modulation frequency Ω_0. Notice that the excitation induces a time-dependent frequency $\omega^2(t) = \Omega^2 + \varepsilon\cos(\Omega_0 t)$, and, hence, it

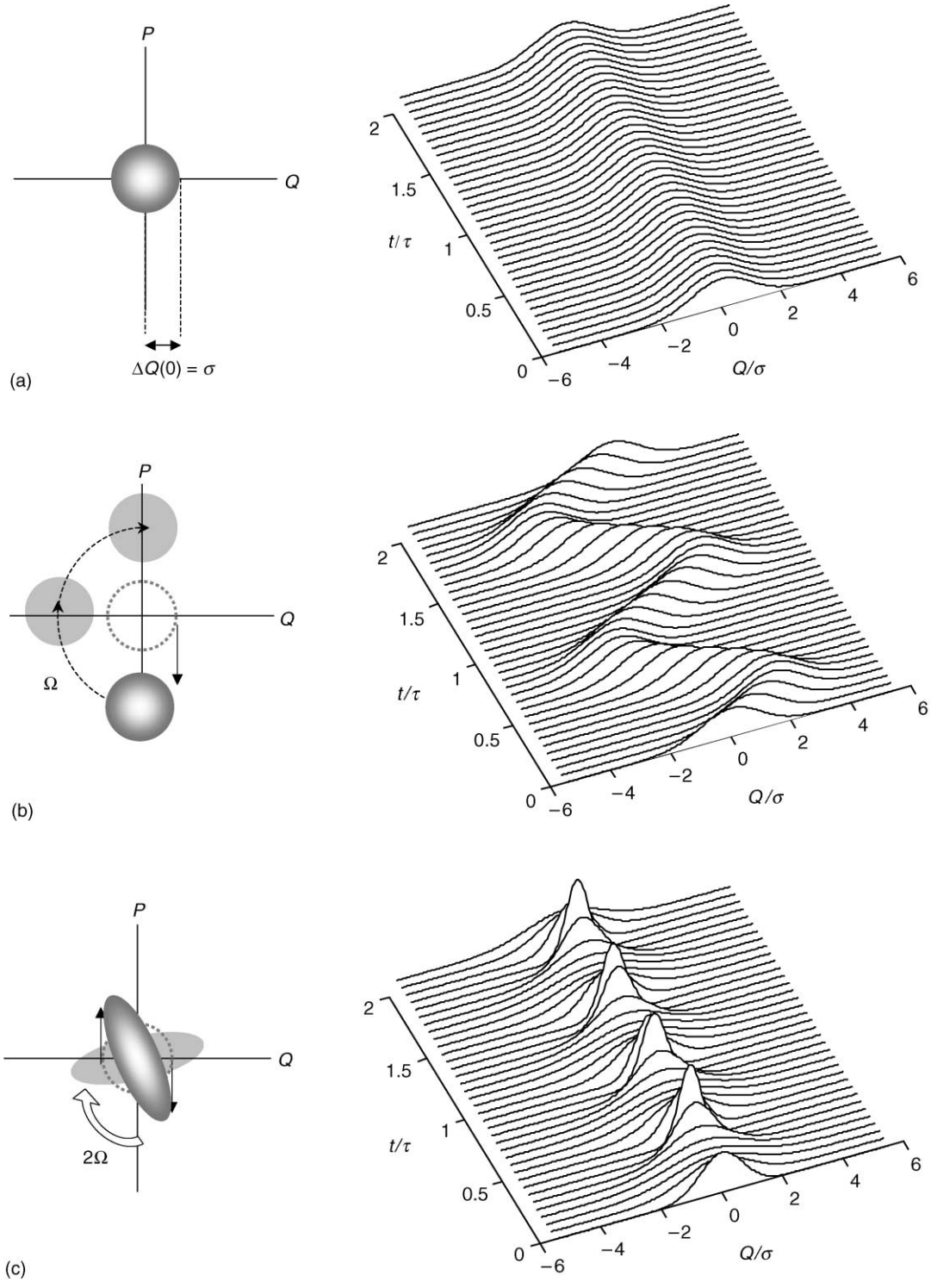


Figure 1 (a) Ground state probability density $|\varphi_0|^2$ for a quantum harmonic oscillator of frequency Ω. Q is the position, P is the momentum and σ is the standard deviation of the Gaussian wavepacket, or the square root of $\langle Q_0^2 \rangle$. The phase-space diagram on the left shows the uncertainty region for the ground state. (b) Coherent state. The uncertainty $\langle Q^2 \rangle$ does not change in time and it is always equal to $\langle Q_0^2 \rangle$, but the wavepacket oscillates with the period τ of the harmonic oscillator. The phase diagram shows the evolution of the uncertainty region, starting from the ground state (dashed line). A sudden constant change in momentum at time $t = 0$ produces a change in the equilibrium position and the wavepacket begins to oscillate. (c) Squeezed state. The wavepacket evolves by changing its width periodically with period $\pi/2$. The uncertainty, $\langle Q^2(t) \rangle$, takes values smaller than $\langle Q_0^2 \rangle$ for a fraction of the cycle. Due to the sudden change on momentum, proportional to Q, at $t = 0$, the distribution becomes ellipsoidal and rotates with frequency 2Ω. (d) Coherent-squeezed state. It is a combination of both (b) and (c). The wavepacket oscillates and breathes at the same time.

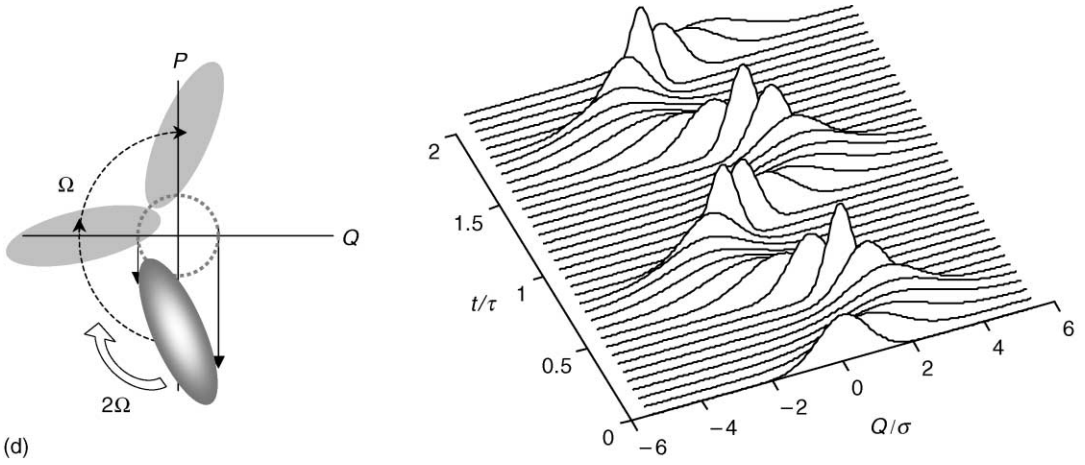


Figure 1 *Continued.*

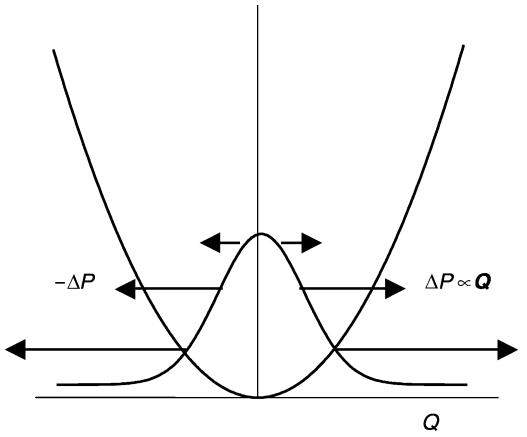


Figure 2 Schematic representation of the effect of an impulsive excitation on the ground state of the harmonic oscillator. At $t = 0$, the pulse induces a change in momentum P, proportional to the position Q. As a result, the wavepacket width in momentum space changes suddenly and starts to oscillate.

shows squeezing effects. The complete set of solutions with and without dissipation has been obtained by Zerbe and Hanggi for the quantum and classical cases.

Physical Interactions Giving Phonon Squeezing

The considerations above are rather general and specific to any physical system that behaves like a set of harmonic oscillators. Here, we consider proposals for phonon squeezing.

Equations [6] and [8] show that we need a driving force proportional to the phonon coordinate Q to achieve squeezing. One way to do this is for the external interaction to produce time-dependent correlations between two modes. Prior to the experimental achievement of phonon squeezing in 1997, using stimulated second-order Raman scattering, models based on (a) three-phonon parametric amplification and (b) polaritons had been proposed.

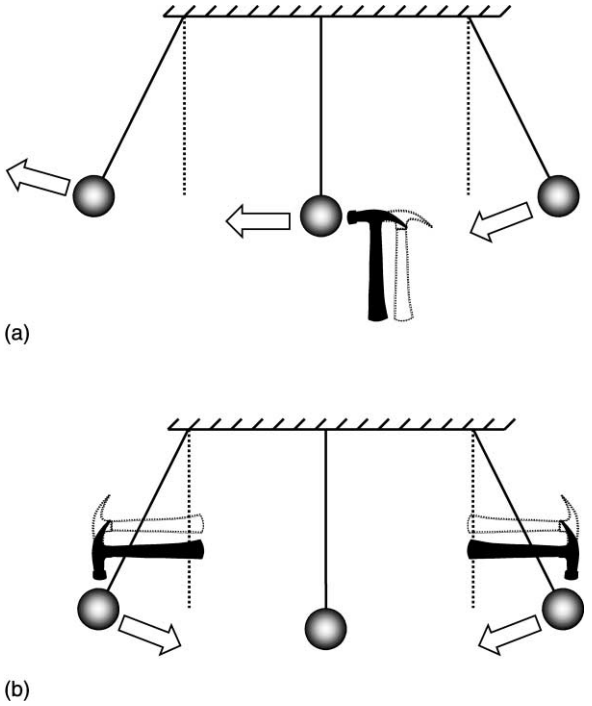


Figure 3 (a) Classical representation of a coherent phonon. An impulsive horizontal force acts on all members of a thermal distribution of pendula, displacing the group in phase space along the momentum axis. The ensemble precesses with an average position $\langle x \rangle \sim \sin(\Omega t)$. (b) Classical representation of a squeezed phonon. An impulsive vertical force produces an impulse proportional to the pendulum's displacement from equilibrium. Those pendula farthest from equilibrium receive the largest kick. The phase space distribution becomes elliptical and precesses with an ensemble averaged variance $\langle x^2 \rangle \sim \sin(2\Omega t)$.

These proposals have not yet been implemented experimentally.

The scheme proposed in (a) is based on three-phonon parametric amplification, similar to photon squeezed states that are generated by optical parametric amplification. Specifically, the proposal relies on a two-step process. First the generation by external excitation with two lasers of a 'pump' longitudinal optical phonon (LO) and, second, the decay of the optical phonon into two longitudinal acoustic phonons (LA), which will be squeezed. This is the kind of parametric oscillator described by eqn [8], for which the parametric control is driven by the coherent optical phonon field.

In case (b) it has been shown that mixed polarization modes known as phonon polaritons are intrinsically squeezed. These hybrid objects, sometimes referred to as dressed photons, are a mixture between phonons and photons that cannot be separated once they are formed.

Squeezing of vibrational states in molecules was confirmed experimentally earlier than in solids. Here, a molecule initially in the ground electronic and vibrational state is excited by an ultrafast optical pulse tuned with an electronic resonance. After the pulse arrives, the molecule is placed in a excited electronic state and the nuclei in a new (excited) potential with a new set of discrete vibrational modes. In Figure 4 we show, as an example of a diatomic molecule, the potential of a sodium dimer. For the sake of simplicity, we consider a single vibrational mode in the excited state and we disregard effects due to anharmonicity. If the optical pulsewidth is much shorter than any vibrational frequency, and the frequencies are different in the ground and excited state, the sudden change in frequency produces a vibrational squeezed state. On the other hand, even if the frequencies in the ground and excited state are equal, but the laser pulse is finite or chirped (frequency sweep on time), the final vibrational state may be squeezed. The formal demonstration of this is not trivial. Analytic solutions for the vibrational wavefunction for an arbitrary external field excitation are known for a very limited number of cases.

Squeezing of a wavepacket in an excited sodium diatomic molecule was achieved by Dunn and coworkers, measuring the time and frequency-resolved spectrum of the spontaneous emission. In this experiment, the molecule can return at any time to its electronic ground state by emitting light, the wavelength of which depends on the internuclear separation at that time. As shown in Figure 4, this is because the equilibrium position for the potential surfaces in the ground and excited states are different.

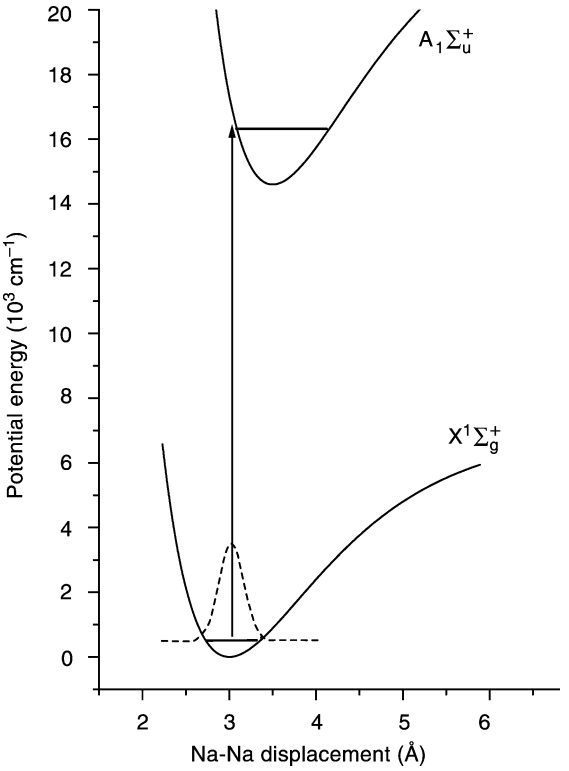


Figure 4 Potential curves for a Na_2 molecule, showing the ground and first excited states. The dashed line represents the ground state wave function. The equilibrium position of both potentials are shifted and the vibrational frequencies are different.

Thus, the time-dependent average wavelength of the emitted light oscillates as the atom vibrates, giving a signature of the dynamics of the wavepacket. Dunn observed oscillations in the signal strength that are consistent with a wavepacket expanding and contracting.

Experimental Generation and Detection of Squeezed Phonons by Second Order Raman Scattering

The first experimental demonstration of phonon squeezing was performed in 1997 by exciting a $KTaO_3$ (potassium tantalate) crystal with an ultrafast pulsed laser and coupling the light to the phonons through second order Raman Scattering. The squeezed phonon field produced by an intense optical pulse, referred to as the 'pump', is detected by a second low intensity pulse, known as the 'probe'. Phonon squeezing was also reported for $SrTiO_3$ (strontium titanate). In addition, correlations between pairs of phonons of different frequencies was reported in a closely related experiment on $KTaO_3$.

The simplest Hamiltonian describing this experiment contains the harmonic contribution from the lattice and a Raman-type interaction term. Raman scattering (RS) refers to light-crystal interaction in which the initial quantum of radiation energy, $\hbar\omega_i$, is divided into a vibrational excitation of the crystal with energy $\hbar\Omega$ and a scattered photon with energy $\hbar\omega_f$. Here, the total energy and momentum are conserved. The interaction energy is proportional to the second-order electronic susceptibility which can be expanded in terms of the atomic displacement Q. The nth term of the Taylor expansion is associated with RS of order n. Two phonons participate in the second order RS process. Although the absence of first-order RS is not required for achieving squeezing, $KTaO_3$ and SrTiO3 (above ~100 K) show no first-order Raman active modes. As a result, the leading contribution is

$$H_{\text{int}} = -\frac{1}{4}|E(t)|^2 \sum_q \chi_2(q) \qquad [9]$$

with $\chi_2(q) = (\partial^2\chi/\partial Q_q^2)Q_q^2$, where $\partial^2\chi/\partial Q_q^2$ is the second order Raman tensor, $E(t)$ is the magnitude of the light electric field, and χ is the linear electronic susceptibility. The Hamiltonian eqn [9], reflecting contributions from pairs of modes at $\pm q$, represents a change in the phonon frequency $\Delta\Omega_q(t) \approx -(1/4\Omega_q)(\partial^2\chi/\partial Q_0^2)|E(t)|^2$ that results in a time-dependent variance $\langle Q_q^2(t)\rangle$ and thus, squeezing. We note that, while individual scattering events produce a two-phonon state, the coherent ensemble of these correlated two-phonon states modulates the variance of Q. Notice that second-order RS excites phonons throughout the whole Brillouin zone instead of just the zone center at $q \approx 0$, as in first-order RS. This means that the light couples to a continuum of modes. The strength of the scattering at any particular frequency is proportional to the number of modes, i.e., the density of states, which is inversely proportional to the gradient $\nabla_q\Omega_q$ and thus, exhibits singularities at points where the gradient vanishes. These so-called critical points lead to associated singularities in the density of states referred to as van Hove singularities. $KTaO_3$ and $SrTiO_3$ show a sharp peak in the density of states at frequencies associated with certain van Hove singularities and, therefore, the continuum of modes becomes quasimonochromatic.

In the experiments, the period of the relevant phonons, $2\pi/\Omega_q$, is large compared with both the pulsewidth and the time the pulse takes for traversing the sample. Then, the time-dependent electric field can be approximated by the delta function $|E(t)|^2 \cong (4\pi I_0/nc)\delta(t)$, where I_0 is the pulse energy density, n the refractive index and c the speed of the light. Within this approximation, the dynamics of the system at $T = 0$ is described by eqn [6]. To lowest order in I_0, we have

$$\Delta Q_q = \langle Q_q^2(t)\rangle^{\frac{1}{2}} \approx \langle Q_q^2(0)\rangle^{\frac{1}{2}}\left[1 + \left(\frac{\lambda}{\Omega_q}\right)\sin(2\Omega_q t)\right] \qquad [10]$$

with $\lambda = 2\pi(I_0/nc)(\partial^2\chi/\partial Q_q^2)$.

As before, the variance drops below the quantum limit $\langle Q_q^2(0)\rangle$ for some fraction of the cycle. This zero-temperature description is only accurate for temperatures such that k_BT is very small compared with the phonon energies. For $T \neq 0$, the conditions overcoming the thermal contribution and produce quantum squeezing at low intensities are

$$n_q \ll 1 \qquad \text{and} \qquad \left(\frac{\lambda}{\Omega_q}\right) > 2n_q \qquad [11]$$

where $n_q = 1/[\exp(\hbar\Omega_q/k_BT) - 1]$ is the Bose factor.

Next we discuss the detection of the squeezed field. As mentioned earlier, the pump pulse creates the squeezed phonon state and the probe is affected by the changes in the material produced by the pump. The scattering of the probe by the squeezed phonon is also described by the nonlinear polarization of the material $P^{NL} = (2V)^{-1}E'(t)\sum_q \chi_2(q)$, where $E'(t)$ is the probe field, and V is the scattering volume. If the probe pulse is Gaussian-shaped of width τ_0 and centered at $\omega = \omega_0$, the total normalized change in the probe transmission ΔT is

$$\frac{\Delta T}{T}(\omega) \approx -\frac{l\pi\tau_0^2\omega(\omega-\omega_0)}{ncV}\sum_q \frac{\partial^2\chi}{\partial Q_q^2}|E'(t)|^2 \times \lambda\langle Q_q^2(0)\rangle\cos(2\Omega_q\Delta t)\exp(-2\Omega_q^2\tau_0^2) \qquad [12]$$

where l is the length of the sample. Equation [12] can be obtained from well-known results in coherent phonon detection. Comparing eqn [12] with eqn [10] it follows that the change in probe transmission is proportional to the derivative of the total time-dependent variance $\langle Q^2(t)\rangle = \sum_q \langle Q_q^2(t)\rangle$. Accordingly, the integral of $\Delta T/T$ probes $\langle Q^2(t)\rangle$.

$KTaO_3$

The perovskite $KTaO_3$ was the first material for which phonon squeezing was achieved. This compound has no first-order Raman active modes. The time-domain data in Figure 5a shows behavior

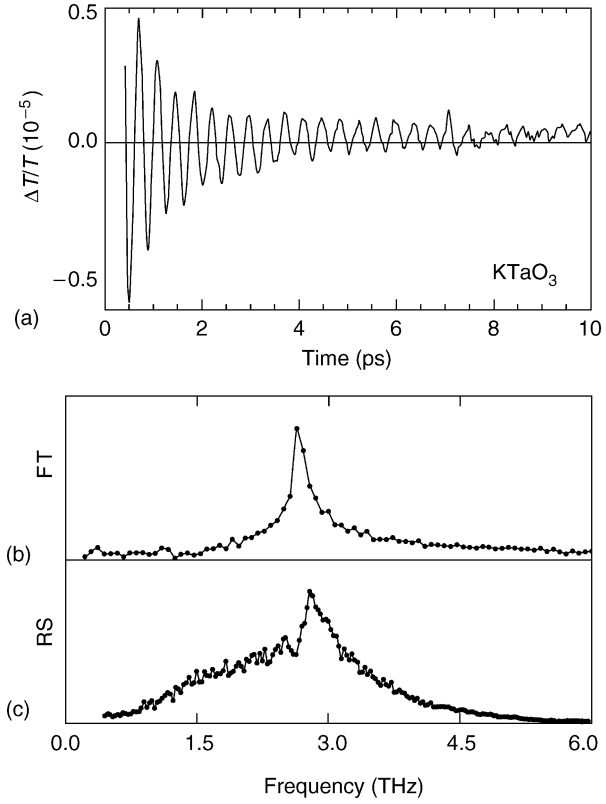


Figure 5 (a) The normalized transmitted probe intensity as a function of the probe delay for $KTaO_3$ at $T = 10$ K. Reprinted with permission from Garrett GA, Rojo AG, Sood AK, Whitaker JF and Merlin R (1997) Vacuum squeezing of solids: Macroscopic quantum states driven by light pulses. *Science*, 275(5306): 1638–1640. Copyright 1997 American Association for Advancement of Science. (b) Fourier transform of the time domain data in (a). The narrow peak appears very close to twice the frequency of the TA (transverse acoustic) mode at the X point of the Brillouin zone. (c) Second order Raman cross section.

consistent with eqn [12]. The Fourier transform $\mathcal{F}_{FT}$ in Figure 5b, is dominated by a narrow peak, strongly dependent on temperature, that appears very close to twice the frequency of the TA (transverse acoustic) mode at the X point of the Brillouin zone. The comparison with the spontaneous Raman scattering data, Figure 5c, indicates that the $\mathcal{F}_{FT}$ peak is due to the 2TA overtone. The squeezing factor defined as $S = 1 - [\langle Q^2(t)\rangle / \langle Q^2(0)\rangle]^{1/2}$, is depicted in Figure 6. Notice that, whereas the squeezing factor in a quantum optics experiment can reach values as large as 70%, phonons in a solid can be squeezed by no more than, say, 0.01%. This is because larger values may lead to collapse of the solid since, according to the Lindemann criterion, solids melt when the root-mean-square displacement of an atom is a fraction of 0.2–0.25 of the lattice parameter.

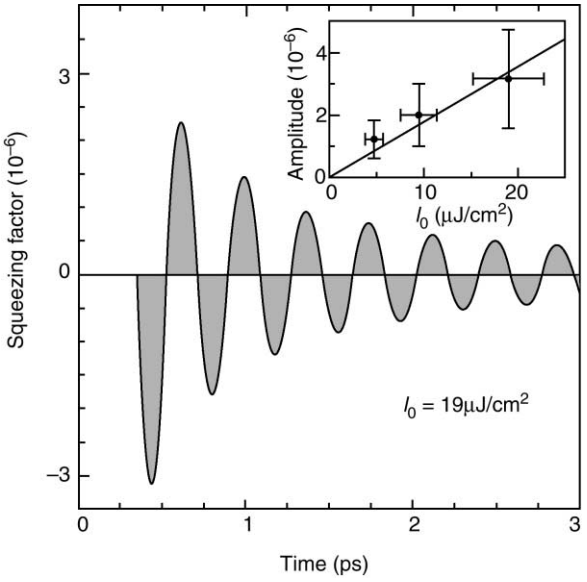


Figure 6 Experimental time dependence of the squeezing factor $S = 1 - [\langle Q^2(t)\rangle / \langle Q^2(0)\rangle]^{1/2}$. I_0 is the pulse intensity. Inset: the amplitude of S as a function of the integrated intensity of the pulse, showing the linear behavior predicted by eqn [12]. Reprinted with permission from Garrett GA, Rojo AG, Sood AK, Whitaker JF and Merlin R (1997) Vacuum squeezing of solids: Macroscopic quantum states driven by light pulses. *Science* 275(5306): 1638–1640. Copyright 1997 American Association for the Advancement of Science.

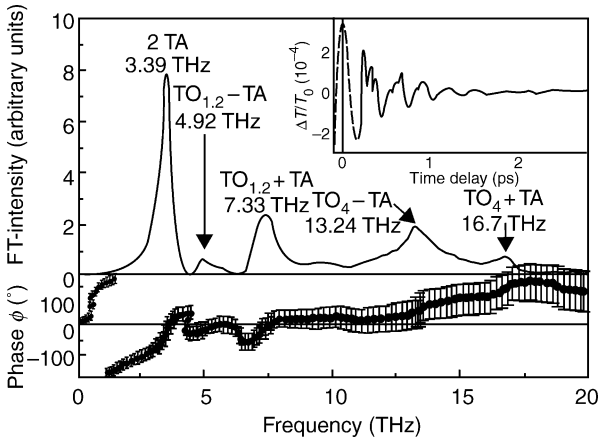


Figure 7 Fourier transform of the time domain data (inset) of the probe differential transmitivity of $KTaO_3$ at room temperature. The excited phonon combinations are indicated in the figure. Reproduced with permission from Bartels A, Dekorsy T and Kurz H (2000) Impulsive excitation of phonon-pair combination states by second-order Raman scattering. *Physical Review Letters* 84(13): 2981–2984. Copyright 2000 by the American Physical Society.

In $KTaO_3$ it was also possible to induce a time-dependent covariance between two different independent modes by using second order RS to couple the two oscillators. In Figure 7 we show

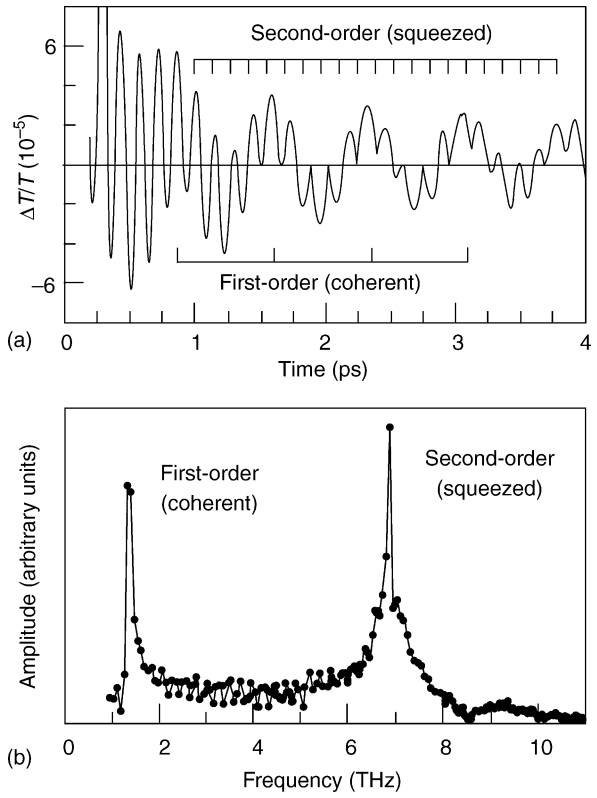


Figure 8 (a) The normalized transmitted probe intensity as a function of the probe delay for $SrTiO_3$ at $T = 7$ K. (b) Fourier transform of the time domain data. The 1.3 THz peak is the soft A_{1g}-phonon (coherent), and the 6.9 THz is the 2 TA overtone (squeezed). The second order feature is dominated by a sharp peak very close to twice the frequency of TA modes at X and M points of the Brillouin zone.

data revealing correlations between transverse optical (TO) and transverse acoustical (TA) modes.

$SrTiO_3$

This material undergoes a structural phase transition, called antiferro-distortive, at $T_C \approx 110$ K. Second-order RS is allowed at all temperatures, but first-order RS is only allowed for temperatures below T_C. Figure 8 shows data at low temperatures, revealing both first and second order features. From the comparison with spontaneous RS, the peak appearing in the $\mathcal{F}_{FT}$ at ~ 1.3 THz is due to first-order RS, while the peak at ~ 6.9 THz is due to second order RS (2TA overtone). The quantum state created by the utrafast pulse is then a mixture of a coherent and a squeezed state. Hence the phonon wavefunction will evolve qualitatively as in the picture shown in Figure 1d.

List of Units and Nomenclature

Gaussian units

I_0 energy density [Jcm^{-2}]

E electric field [statvolt cm^{-1}]

See also

Scattering: Stimulated Scattering; Raman Scattering. **Semiconductor Physics:** Light Scattering. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Dunn TJ, Sweetser JN, Walmsley IA and Radzewicz C (1995) Experimental determination of the dynamics of a molecular nuclear wave-packet via the spectra of spontaneous emission. *Physical Review Letters* 74(6): 884–887.

Garraway BM and Suominen KA (1995) Wave-packet dynamics: new physics and chemistry in femto-time. *Reports on Progress in Physics* 58(4): 365–419.

Garrett GA, Rojo AG, Sood AK, Whitaker JF and Merlin R (1997) Vacuum squeezing of solids: Macroscopic quantum states driven by light pulses. *Science* 275(5306): 1638–1640.

Hayes W and Loudon R (1978) *Scattering of Light by Crystals*. New York: John Wiley & Sons, Inc.

Lindemann F (1910) Über die Berechnung molekularer Eigenfrequenzen. *Zeitschrift Für Physik* 11: 609–612.

Walls DF and Milburn GJ (1994) Generation and Applications of Squeezed Light. *Quantum Optics*, pp. 137–174. Berlin Heidelberg: Springer Verlag.

Zerbe C and Hanggi P (1995) Brownian parametric quantum oscillator with dissipation. *Physical Review E* 52(2): 1533–1543.

R

RELATIVISTIC NONLINEAR OPTICS

D P Umstadter, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Immediately after the invention of the laser in 1960, light could be focused to sufficient intensity to cause nonlinear optical effects in atomic media. In conventional nonlinear optics, where the electric field of a light wave E is much smaller than an atomic field, $E_{at} = 3 \times 10^9$ V/cm, various nonlinear phenomena, such as self-focusing, harmonic generation, and Raman scattering, arise due to the anharmonic motion of electrons in the combined fields of atom and laser. Approximate analytical solutions can be obtained by means of perturbation expansion methods, using E/E_{at} as the expansion parameter. At higher light fields, when E approaches E_{at}, this method breaks down and the medium becomes photo-ionized, creating a plasma, as illustrated by Figure 1. Further increases in light intensity enabled nonlinear optical effects of even these free plasma electrons (see Figure 2). The nonlinearity arises, in this case, because the electrons oscillate at relativistic velocities in laser fields that exceed 10^{11} V/cm, resulting in relativistic mass changes exceeding the electron rest mass and the light's magnetic field becomes important. The work done by the electromagnetic field (E) on an electron ($eE\lambda$) over the distance of a laser wavelength (λ), then approaches the electron rest mass energy ($m_e c^2$), where e is the elementary charge of an electron, m_e is the electron rest mass, and c is the speed of light. Effects analogous to those studied with conventional nonlinear optics – self-focusing, self-modulation, harmonic generation, and so on – are all found, but based on this entirely different physical mechanism. Thus, a new field of nonlinear optics, that of relativistic electrons, has been launched, as illustrated by Figure 1.

One outcome of accessing this new optical regime is the generation of frequency-shifted light in a spectral region where there are no other compact sources. Another is the acceleration of other types of particles, such as positrons, ions, and neutrons. These novel radiation sources have properties (femtosecond duration, micron source size, MeV energy) that make them suitable for numerous applications in imaging and spectroscopy in basic research, as well as medical diagnostics, cancer therapy, energy production, and space propulsion. Rapid advancement is underway and new research tools, subfields, and commercial products are on the horizon, e.g., compact and ultrashort pulse duration laser-based electron accelerators and X-ray sources.

Another physical regime will be encountered at even higher intensities ($I\lambda^2 \simeq 10^{24}$ W/cm^2), when even protons will quiver relativistically, i.e., the work done on a proton over the distance of a laser wavelength approaches its rest mass energy. This might be called the nuclear regime of laser–plasma interactions, because of the fusion and fission reactions and the generation of pions, muons, and neutrinos that should occur as nuclei collide in such energetic plasmas.

The recent dramatic increase in light intensity was partly made possible by the development, in the last decade, of compact lasers that have the ability to amplify shorter light pulses. For instance, solid-state lasers use the technique of chirped-pulse amplification to generate femtosecond duration pulses. To accomplish this, an ultrashort, low-energy laser pulse is first stretched in time before it is amplified and then recompressed. Gas or dye lasers, using solid-state switches, have produced picosecond duration pulses. Present day advanced laser systems now have multiterawatt peak powers and, when focused to micron spotsizes with adaptive optics, can produce electromagnetic intensities $I\lambda^2 \simeq 10^{21}$ W/cm^2, as illustrated by Figure 3. An example of a modern ultra-high-power (0.9 PW) solid-state (Ti:Sapphire) laser

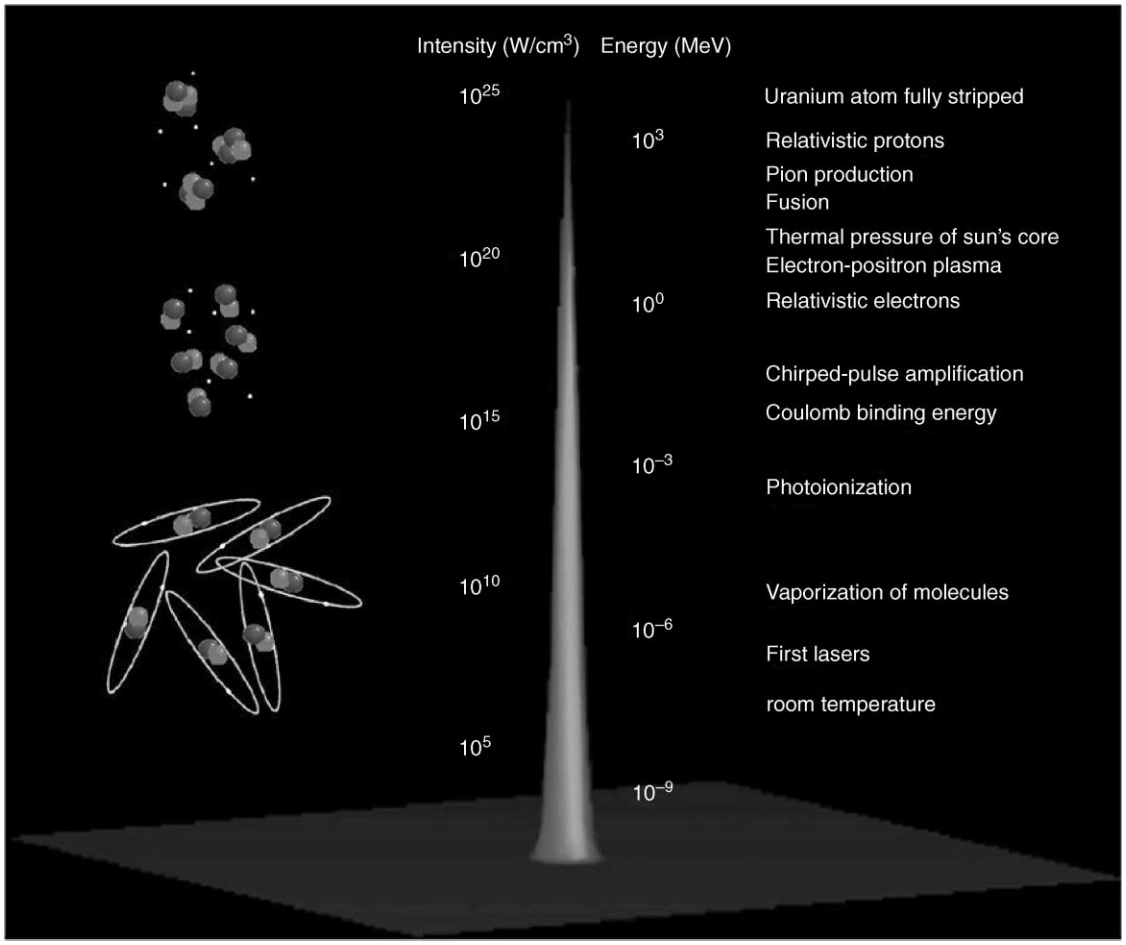


Figure 1 The various regimes of laser–matter interactions, represented by the ideal laser pulse. As the intensity of laser light increases, so does the energy of electrons accelerated in the light field and the regime of conventional nonlinear optics with electrons bound to atoms is replaced by the regime of relativistic nonlinear optics with free electrons in relativistic plasmas. At the highest intensities, even protons become relativistic, giving rise to what might be called the regime of nuclear optics, in which various nuclear processes, such as fusion, can take place. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

system, located at the Advanced Photon Research Center, Kansai Research Establishment, Japan Atomic Energy Research Institute (JAERI), can be seen in the photograph shown in Figure 4.

This paper will discuss the fundamental concepts underlying what might be referred to more broadly as the field of high-field science. For more detailed descriptions of recent progress in experiment or theory, several review papers have been published on related topics: (i) relativistic nonlinear optics; (ii) high-intensity laser development; (iii) laser accelerators; (iv) intense laser–plasma interactions; and (v) relativistic scattering.

Definitions

For an infinite plane wave for the light wave propagating in the $\hat{z}$ direction (shown in Figure 5), the fields may be described by

$$\mathbf{E} = \hat{e}_x E_0 \exp[\mathrm{i}(kz - \omega t)] = \hat{e}_x E_0 \, \mathrm{e}^{\mathrm{i}\psi} \qquad [1]$$

and

$$\mathbf{B} = \hat{e}_y B_0 \exp[\mathrm{i}(kz - \omega t)] = \hat{e}_y B_0 \, \mathrm{e}^{\mathrm{i}\psi} \qquad [2]$$

where ω and k are the frequency and wavenumber of the laser light, respectively. For the ultrashort pulse durations of the lasers typically used to produce high fields, it is assumed that the ions do not have time to move because of their greater inertia. In this case, $m(\mathrm{d}v/\mathrm{d}t) \approx -e\mathbf{E}$ and the term v/c can be used as an expansion parameter. At even higher fields, electrons can be accelerated to nearly the speed of light as they quiver in the oscillating fields of the light wave. In this case, the force due to the magnetic field can no longer be ignored and the mass of the electron

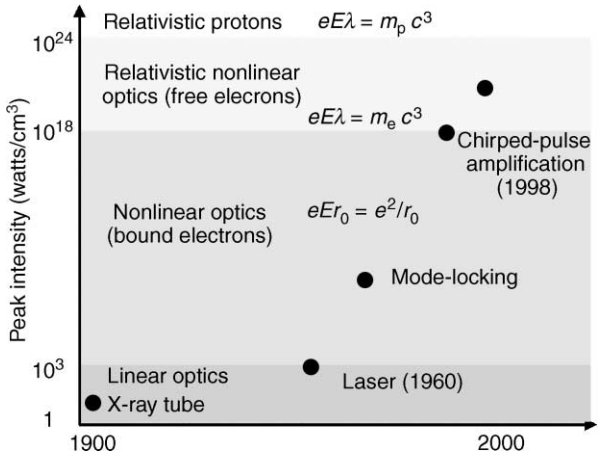


Figure 2 History of light sources over the last century. Each advance in laser power enables a new regime of optics. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

can increase as $m = \gamma m_0$, where m_0 is the electron rest mass and the relativistic gamma factor is given by

$$\gamma = \frac{1}{\sqrt{1 - \left(\frac{v}{c}\right)^2}} \qquad [3]$$

The threshold for the relativistic regime of optics is crossed when the electron momentum p exceeds the product of the electron rest mass m_0 and c.

For problems involving charged particle motion in intense electromagnetic fields, it is often convenient to introduce the vector potential: $\mathbf{A}(\mathbf{r}, t) = \mathrm{Re}\{\mathbf{A}_0\, e^{i\psi}\}$. From Maxwell's equation $\nabla \cdot B = 0$, it can be seen that without loss of generality, the magnetic field can be written as the curl of a potential $B = \nabla \times A$. Further, from Faraday's law in vacuum:

$$\nabla \times E = -\frac{1}{c}\frac{\partial B}{\partial t} = -\frac{1}{c}\frac{\partial(\nabla \times A)}{\partial t} \qquad [4]$$

it is clear that $\mathbf{E} = -(1/c)\partial\mathbf{A}/\partial t$. To include the space–charge field of the plasma, as we will do in the section on collective plasma response below, we introduce the potential Φ, and the definition:

$$\mathbf{E} = -\nabla\Phi - \frac{1}{c}\frac{\partial}{\partial t}\mathbf{A} \qquad [5]$$

Assuming the Coulomb gauge:

$$\nabla \cdot \mathbf{A} = 0 \qquad [6]$$

or $A_z = 0$ in 1D, then $\mathbf{A}_\perp$ represents the laser field.

If we neglect Φ for now, then, by Fourier analysis:

$$\mathbf{E} = \mathrm{Re}\left\{\frac{i\omega}{c}\mathbf{A}_0\, e^{i\psi}\right\} \qquad [7]$$

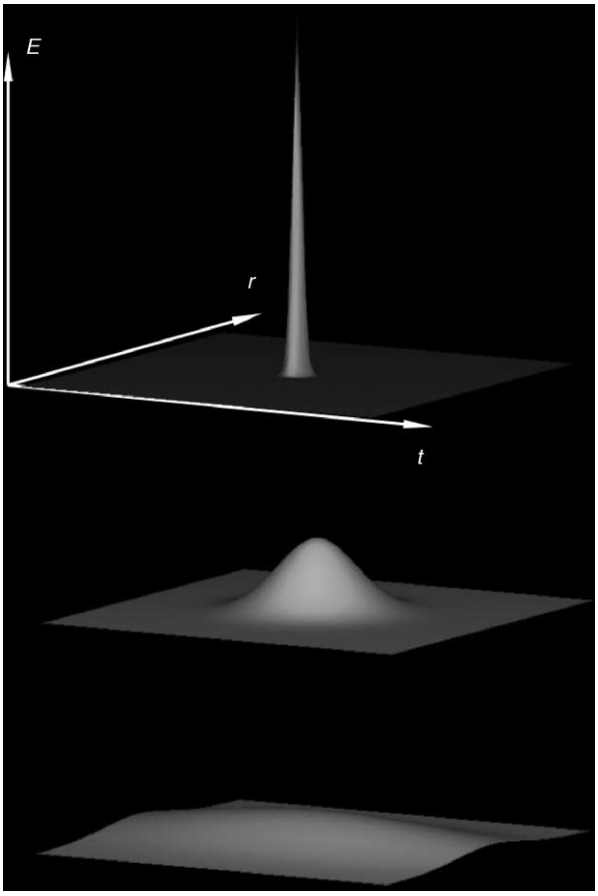


Figure 3 Laser light has become concentrated to ever-smaller regions of space (r) and time (t), dramatically increasing the peak electric field (E) at the laser focus. Prior to the development of chirped-pulse amplification (CPA), the energy of light was produced in long-duration pulses, as shown in the pulse of the bottom figure. After CPA, the pulse duration decreased dramatically as shown in the pulse in the middle. The latest improvement in laser technology has been the use of deformable mirrors, which has allowed lasers to be focused to a spatial dimension that is as small as the temporal dimension, a few laser wavelengths, as shown in the pulse on top. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

$$\mathbf{B} = \mathrm{Re}\{ik\hat{z} \times \mathbf{A}_0\, e^{i\psi}\} \qquad [8]$$

The light intensity is given by the time-averaged Poynting's vector, which can be written for linear (LP), $\mathbf{A}_0 = A_0\hat{e}_y$ and circular polarizations (CP), $\mathbf{A}_0 = A_0(\hat{e}_x \pm \hat{e}_y)$ as:

$$I = \langle|\mathbf{S}|\rangle = \left\langle\left|\frac{c\mathbf{E} \times \mathbf{B}}{8\pi}\right|\right\rangle = \frac{\omega k}{16\pi}A_0^2 \qquad [9]$$

It is often the combination $I\lambda^3$ that matters, which has units of power and for linear polarization can be



Figure 4 Photograph of the 33-fs duration 0.9-PW-peak power laser system at JAERI. (Reproduced with permission from Yamakawa K (2004) Table-top lasers create ultrahigh peak powers. *OYOBUTURI* 73: 186–193.)

written as

$$I\lambda^2 = \frac{\omega k \lambda^2}{16\pi} A_0^2 = \frac{\pi}{2} c A_0^2 = \left[1.37 \times 10^{18} \frac{\mathrm{W}}{\mathrm{cm}^2} \mu\mathrm{m}^2\right] a_0^2 \qquad [10]$$

where the normalized vector potential is defined as

$$a_0 = \frac{eA_0}{mc^2} \qquad [11]$$

Relativistic effects become significant when a_0 approaches unity, which, as can be seen from eqn [10] occurs for 1 μm wavelength light at a laser intensity of $\sim 1.37 \times 10^{18}$ W/cm^2, corresponding to a electric field strength of 6×10^9 V/cm, which is four orders of magnitude greater than E_{at}.

Single-Particle Motion

The electron orbit, subject to a linearly polarized electromagnetic wave propagating in the $+z$ direction, is governed by the Lorentz equation:

$$m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -e\left(\mathbf{E} + \frac{\mathbf{v}}{c} \times \mathbf{B}\right) \qquad [12]$$

where v is the electron velocity, and E and B are the light's electric and magnetic field.

A zeroth order solution to eqn [12] is found by setting $v/c \ll 1$, which allows the term $(v/c) \times B$ to be neglected and γ from eqn [8] is set to unity. Integrating eqn [12] in this limit once over time (t) yields for the velocity:

$$\mathbf{v} = \mathrm{Re}\left\{\frac{e\mathbf{E}}{\mathrm{i}m\omega}\right\} = \frac{a_0}{c}\begin{cases} \hat{e}_y \cos\psi, & \mathrm{LP} \\ (\hat{e}_x \cos\varphi \mp \hat{e}_y \sin\psi), & \mathrm{CP} \end{cases} \qquad [13]$$

An electron in low-intensity light oscillates with this velocity in a straight line along the polarization vector ($\hat{e}_y$), which when normalized to c, is the normalized vector potential a_0. Integrating again yields for the

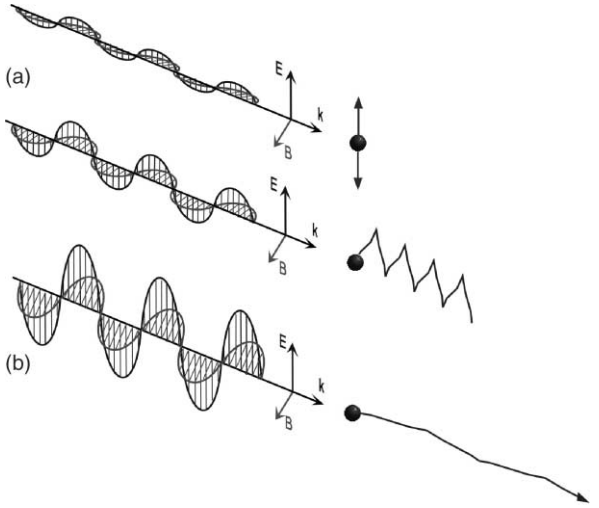


Figure 5 Classical optics versus relativistic optics. (a) In classical optics, the amplitude of the light wave is small, electrons oscillate in the direction of the electric field at the light's frequency, and there is no displacement along the light's propagation direction. Note that only the *E* field acts on the electron and the electron-oscillation velocity is very small compared with the speed of light. (b) In relativistic optics, the amplitude of the light wave is very large, the light's magnetic field becomes important and the combined action of the electric and magnetic fields push the electron forward. In this case, the electron velocity becomes close to the speed of light and its motion is a figure-eight superimposed upon a steady drift. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

transverse displacement:

$$\mathbf{x} = \mathrm{Re}\left\{\frac{e\mathbf{E}}{m\omega^2}\right\} = \frac{a_0}{\omega c}\begin{cases}\hat{e}_y \sin\psi, & \text{LP}\\ (\hat{e}_y \sin\varphi \mp \hat{e}_z \cos\psi), & \text{CP}\end{cases} \quad [14]$$

Thus, for the $a_0 \sim 1$, the electron excursion during its oscillation is approximately λ.

A first-order approximation for the electron motion can be found by substituting the zeroth order velocity eqn [13] into the $v/c \times B$ term of eqn [12]. The latter then becomes proportional to

$$E \times B \propto \frac{a_0^2}{2}[1 + \cos(2\omega_0)]\hat{e}_z \quad [15]$$

In the frame in which the electron is on the average at rest, the relativistic motion of an electron is thus described by a figure eight, oscillating twice in the $\hat{e}_z$ or $\hat{k}$ direction for every once in the polarization direction ($\hat{e}_x$); i.e., a figure of eight motion. This originates from the fact that $v \times B \propto E \times B \propto E^2\hat{k}$, which is a product of two functions that vary sinusoidally at frequency ω and thus varies itself at frequency 2ω. In the lab frame,

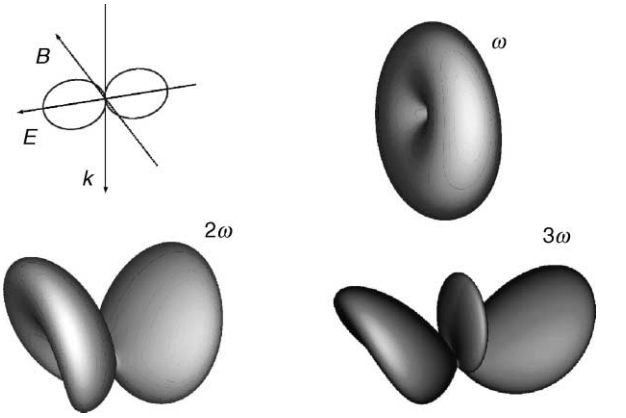


Figure 6 Harmonics driven by relativistic Thomson scattering as the electrons in high-intensity laser fields (a_0^2) undergo figure-eight motion display unique angular distributions. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

this transverse motion is superimposed upon a steady drift in the ($\hat{e}_z$) direction (shown in Figure 5), originating from the DC term in eqn [15]. The next order approximation would include the mass shift $m = \gamma m_0$. As the field strength increases ($a_0^2 \gg 1$), the longitudinal motion ($\propto a_0^2$) begins to dominate the transverse motion ($\propto a_0$), as shown in Figure 5. This is described more formally below.

In the regime $a_0 \lesssim 1$, electrons radiate photons at harmonics of a modified laser frequency ω_0, with each harmonic order having its own unique angular distribution, as shown in Figure 6. The radiation at the fundamental is the usual donut pattern, with a maximum in the direction perpendicular, and a minimum along, the polarization vector ($\hat{e}_x$). The second harmonic has two emission lobes with maxima pointing at an angle between $\hat{e}_z$ and $\hat{e}_x$. An additional lobe is added for each additional harmonic order. This is referred to as nonlinear or relativistic Thomson scattering. The unique angular distributions of the second and third harmonics emitted from nonlinear relativistic Thomson scattering were observed experimentally and are shown in Figure 7. As will be shown in detail in the section on radiation from relativistic electrons, below, the situation is more complex for $a_0 \gtrsim 1$, and the scattered light is no longer simply harmonic.

Constants of the Motion

Several constants of the motion can be found from an exact treatment for the motion. The starting point is the relativistically correct Lagrangian, which is written as

$$L(\mathbf{r}, \mathbf{v}, t) = -mc^2\sqrt{1 - \frac{v^2}{c^2}} + \frac{q}{c}\mathbf{v}\cdot\mathbf{A} + q\phi \quad [16]$$

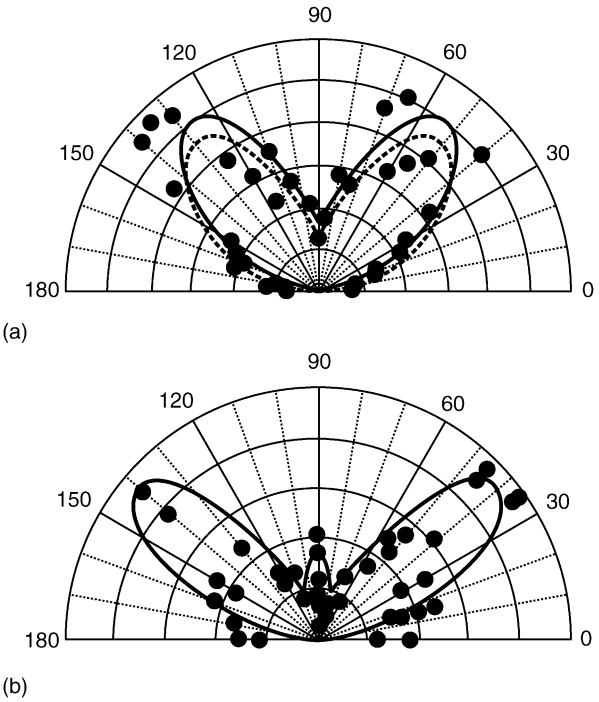


Figure 7 Angular pattern of higher-order harmonic light. Shown are polar plots of the intensity of the second-harmonic light (top) and third harmonic (bottom) as a function of azimuthal angle. Filled circles, experimental data; solid and dashed lines, theoretical results. (Reproduced with permission from Chen S-Y, Maksimchuk A and Umstadter D (1998) Experimental observation of relativistic nonlinear Thomson scattering. *Nature* 396: 653.)

where ϕ is the scalar potential associated with electrostatic fields. The Lorentz equation, eqn [12], is derived from the Euler–Lagrange equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \mathbf{v}} - \frac{\partial L}{\partial \mathbf{r}} = 0 \qquad [17]$$

For an infinite plane wave, in which the Lagrangian is independent of space in the transverse direction, eqn [17] yields the conservation of transverse canonical momentum:

$$\frac{\partial L}{\partial \mathbf{v}_\perp} = \mathbf{p}_\perp + \frac{q}{c}\mathbf{A}_\perp = \text{constant} \qquad [18]$$

The sum of the transverse momentum and field strength remain constant. The next constant follows from $\mathrm{d}H/\mathrm{d}t = -\partial L/\partial t$, yielding the relation:

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\frac{\partial L}{\partial t} = c\frac{\partial L}{\partial z} = c\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial v_z} = c\frac{\mathrm{d}p_z}{\mathrm{d}t} \qquad [19]$$

where E is the time-dependent energy of the particle, which yields our second constant of the motion:

$$E - cp_z = \text{constant} \qquad [20]$$

For a particle initially at rest, the kinetic energy E_{kin} becomes

$$E_{\mathrm{kin}} = E - mc^2 = p_z c \qquad [21]$$

which yields

$$E_{\mathrm{kin}} = \frac{p_\perp^2}{2m} = p_x c = mc^2(\gamma - 1) \qquad [22]$$

Thus, the electron scattering angle θ is related to the transverse and longitudinal momenta or the kinetic energy by

$$\tan^2\theta = \left(\frac{p_\perp}{p_z}\right)^2 = \frac{2mE_{\mathrm{kin}}}{(E_{\mathrm{kin}}/c)^2} = \frac{2}{\gamma - 1} \qquad [23]$$

The angle of electrons produced by photo-ionization with intense lasers has been shown experimentally to obey the conservation of canonical momentum, as demonstrated in experiments that studied the angular distribution of relativistic electrons emitted from barrier-suppression ionization of atoms in intense laser fields (shown in Figure 8). The transverse normalized vector potential, $\mathbf{a} = e\mathbf{A}_\perp/mc^2$, is thus equal to the transverse normalized momentum:

$$\hat{\mathbf{p}}_\perp = \frac{\mathbf{p}_\perp}{mc} = \mathbf{a} = (0, a_y, a_x) \qquad [24]$$

The normalized kinetic energy can also be written in terms of a:

$$\hat{E}_{\mathrm{kin}} = \frac{E_{\mathrm{kin}}}{mc^2} = \gamma - 1 = \hat{p}_z = \frac{\hat{p}_\perp^2}{2} = \frac{a^2}{2} \qquad [25]$$

This can be rewritten as

$$\gamma = 1 + \frac{a^2}{2} \qquad [26]$$

From eqn [25], the normalized momenta in each direction can be found:

$$\hat{p}_z = \gamma\beta_z = \frac{\gamma}{c}\frac{\mathrm{d}z}{\mathrm{d}t} = \frac{a^2}{2} \qquad [27]$$

$$\hat{p}_y = \gamma\beta_y = \frac{\gamma}{c}\frac{\mathrm{d}y}{\mathrm{d}t} = a_y \qquad [28]$$

$$\hat{p}_x = \gamma\beta_x = \frac{\gamma}{c}\frac{\mathrm{d}x}{\mathrm{d}t} = a_x \qquad [29]$$

Note that the momentum in the longitudinal ($\hat{e}_z$) direction is proportional to a^2 while in the two

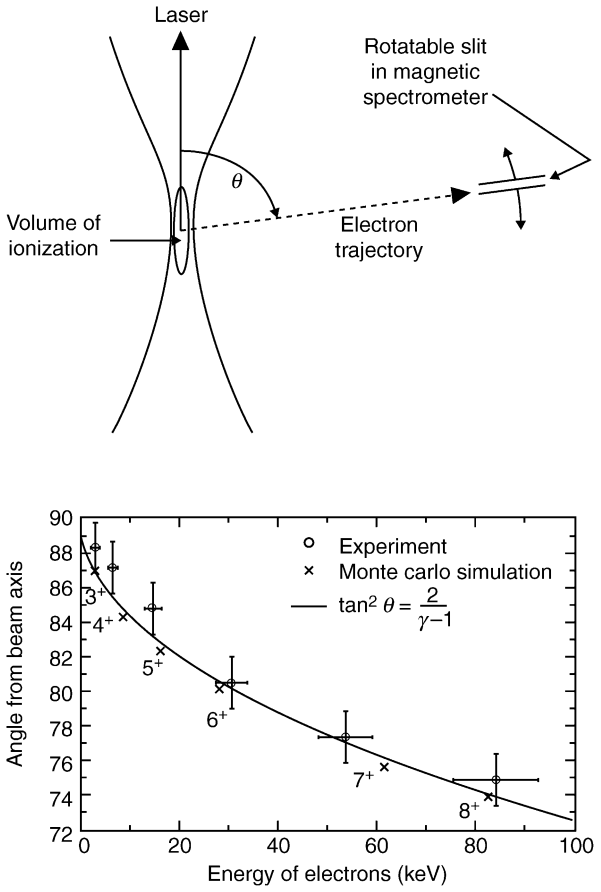


Figure 8 Top: Experiment to measure the angle of electron ejection during photo-ionization. Bottom: Kinetic energy versus angle during the production of 3^+ through 8^+ of neon. (Reproduced with permission from Moore CI, Knauer JP and Meyerhofer DD (1995) Observation of the transition from Thomson to Compton scattering in multiphoton interactions with low-energy electrons. *Physics Review Letters* 74: 2439. Copyright (1995) by the American Physical Society.)

transverse directions it is proportional to a. Equations [26]–[29] indicate that in the limit of large a, the normalized velocities are given by

$$\beta_z = \frac{a^2/2}{1+a^2/2} \to 1, \qquad \beta_y = \frac{a_y}{1+a^2/2} \to 0, \qquad \beta_x = \frac{a_x}{1+a^2/2} \to 0 \qquad [30]$$

This indicates that in the deeply relativistic regime, the electron is accelerated to near c in the longitudinal direction, while in the transverse directions its velocity goes to zero. It follows that, as a increases and the electron moves in phase with the light wave, relativistic Thomson scattering will be significantly reduced. Integrating eqn [29] in the variable $\tau = t - x(t)/c$ gives

$$\gamma\frac{\mathrm{d}}{\mathrm{d}t} = \gamma\frac{\mathrm{d}\tau}{\mathrm{d}t}\frac{\mathrm{d}}{\mathrm{d}\tau} = \gamma\left(1 - \frac{1}{c}\frac{\mathrm{d}z}{\mathrm{d}\tau}\right)\frac{\mathrm{d}}{\mathrm{d}\tau} = \left(1 + \frac{a^2}{2} - \frac{a^2}{2}\right)\frac{\mathrm{d}}{\mathrm{d}\tau} = \frac{\mathrm{d}}{\mathrm{d}\tau} \qquad [31]$$

which allows the equations for the displacement to be written simply as

$$\frac{\mathrm{d}z}{\mathrm{d}\tau} = c\frac{a^2}{2} \qquad [32]$$

$$\frac{\mathrm{d}y}{\mathrm{d}\tau} = ca_y \qquad [33]$$

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = ca_x \qquad [34]$$

For a box-shaped pulse, $a_x = a_0\cos(\omega\tau)$ for $0 < \tau < N(2\pi/\omega)$, $a_z = 0$, $a^2 = a_x^2\cos^2(\omega\tau)$, the trajectory for an electron initially at rest at $x = y = z = 0$ at $\tau = 0$ is then

$$x(\tau) = \frac{ca_0^2}{2}\int_0^\tau \cos^2(\omega\hat{\tau})\mathrm{d}\hat{\tau} = \frac{ca_0^2}{4}\left[\tau + \frac{1}{2\omega}\sin(2\omega\tau)\right] \qquad [35]$$

$$y(\tau) = ca_0\int_0^\tau \cos(\omega\hat{\tau})\mathrm{d}\hat{\tau} = \frac{ca_0}{4}\sin(\omega\tau) \qquad [36]$$

Thus, as we saw from the first-order appoximation above, eqn [15], the motion consists of a drift along $\hat{x}$:

$$x_\mathrm{d}(t) = \frac{a_0^2}{a_0^2+4}ct \qquad [37]$$

and a figure-eight motion in the drift frame of reference:

$$ky = a_0\sin(\omega\tau), \qquad k(z - z_\mathrm{d}) = \frac{a_0}{8}\sin(2\omega\tau) \qquad [38]$$

In the strongly relativistic regime, because the electron is strongly accelerated in the direction of the light wave, the harmonic emission from relativistic Thomson scattering loses its dependence on the laser frequency, ω, but instead depends only on the direction of observation, the amplitude a_0 and the initial phase of the electron, relative to the light wave. Electrons that do not begin their trajectories at zero or π phase with respect to the light wave can

drift in the transverse direction. For pulses with finite spatial width, such as a Gaussian pulse with field strength peaked on axis, the ponderomotive force of the light can also act to expel the electrons transversely from the axis.

For circular polarization:

$$\mathbf{a}(\mathbf{r}, t) = \mathrm{Re}\{a_0(\hat{e}_y \pm \mathrm{i}\hat{e}_x)\mathrm{e}^{-\mathrm{i}\omega\tau}\} \qquad [39]$$

the electron drifts as it did in the case of linear polarization and described in eqn [37] but orbits in a helical trajectory:

$$y(\tau) = \frac{ca_0}{\sqrt{2}\omega}\sin(\omega t/\gamma) \qquad [40]$$

$$x(\tau) = \mp\frac{ca_0}{\sqrt{2}\omega}\cos(\omega t/\gamma) \qquad [41]$$

Rather than oscillate in amplitude, the relativistic factor remains constant:

$$\gamma = 1 + \frac{a_0^2}{4} \qquad [42]$$

In either polarization case, the light pulse eventually overtakes the electron, leaving the latter at rest. However, net acceleration in vacuum can be obtained only under special circumstances, described by the Lawson–Woodward criterion, such as in the near field of metallic surfaces; with high-order focusing modes instead of infinite plane waves; or with mixed frequencies, or with optics to terminate the light but not the electrons. Regions of a tightly focused Gaussian pulse, where the accelerating field propagates at sub-luminal velocities also exist, but they are effective for only extremely large values of a and thus only very short acceleration lengths.

Role of Initial Phase

If we wish to understand the role of the electron's initial phase, the relativistic Lorentz equation (eqn [12]) may also be written, for an $\hat{x}$-polarized wave traveling in the $\hat{z}$ direction, as

$$\frac{\mathrm{d}}{\mathrm{d}t}(\gamma\beta) = (\hat{x} - \hat{z}\times\beta)a\cos(t - z) \qquad [43]$$

where we normalize time by $1/\omega_0$, velocity by c, and distance by c/ω_0. In eqn [43], $a = eE_0/m\omega_0 c$ is the dimensionless parameter measuring the electric field strength, $\gamma = (1 - \beta_x^2 - \beta_y^2 - \beta_z^2)^{-1/2}$ is the relativistic mass factor, and $\beta = (\beta_x, \beta_y, \beta_z)$ is the electron velocity (in units of c). The electron orbit, subject to the following general initial conditions at time $t = 0$:

$$x = 0, \qquad y = 0, \qquad z = z_{\mathrm{in}} \qquad [44]$$

$$\beta_x = \beta_{x0}, \qquad \beta_y = \beta_{y0}, \qquad \beta_x = \beta_{z0} \qquad [45]$$

$$\beta_{x0} = \beta_{y0} = 0 \qquad [46]$$

has a closed form solution when it is expressed parametrically: $t = t(\theta)$, $r = r(\theta)$, $\beta = \beta(\theta)$, where θ is the phase of the wave, defined by

$$\theta = t - z \qquad [47]$$

Note that $\beta_0 = (\beta_{x0}, \beta_{y0}, \beta_{z0})$ is the unperturbed velocity of the electron ($a = 0$ limit) and that the initial phase that the electron sees is $\theta_{\mathrm{in}} = -z_{\mathrm{in}}$ according to eqns [44] and [47]. This phase can be important in the ionization of the gas by an intense laser.

For the special case $\beta_{x0} = 0$, $\beta_{y0} = 0$, one finds $\beta_y = 0$, and the orbital equation yields the following closed form solution:

$$\gamma = \gamma_0 + \frac{a^2(\sin\theta - \sin\theta_{\mathrm{in}})^2}{2\gamma_0(1 - \beta_{z0})} \qquad [48]$$

$$\gamma\beta_z = \gamma - \gamma_0(1 - \beta_{z0}) \qquad [49]$$

$$\gamma\beta_x = \frac{1}{\gamma_0(1 - \beta_{z0})}a(\sin\theta - \sin\theta_{\mathrm{in}}) \qquad [50]$$

$$x = \frac{a[(\cos\theta_{\mathrm{in}} - \cos\theta) - (\theta - \theta_{\mathrm{in}})\sin\theta_{\mathrm{in}}]}{\gamma_0(1 - \beta_{z0})} \qquad [51]$$

$$t = \frac{(\theta - \theta_{\mathrm{in}})}{1 - \beta_{z0}}\left[1 + \frac{a^2}{2}\left(\frac{1}{2} + \sin^2\theta_{\mathrm{in}}\right)\right] + \frac{a^2(1 + \beta_{z0})}{2(1 - \beta_{z0})}\left[-\frac{\sin 2\theta}{4} + 2\cos\theta\sin\theta_{\mathrm{in}} - \frac{3\sin 2\theta_{\mathrm{in}}}{4}\right] \qquad [52]$$

In eqns [48]–[52], $\gamma_0 = (1 - \beta_{z0}^2)^{-1/2}$ and β_{z0} may either be negative (counter-propagating against the laser), or zero (Thomson scattering), or positive (co-propagating with the laser). Note that the velocity components β_x and β_z are given as explicit functions of θ according to eqns [49] and [50], upon using eqn [48]. They are periodic functions of θ of period 2π. The period, T, of this periodic (**Figure 8**) motion is thus equal to the increase in t as θ increases by 2π. Thus, we obtain from eqn [53]:

$$T = \frac{2\pi}{1 - \beta_{z0}}\left[1 + \frac{a^2}{2}\left(\frac{1}{2} + \sin^2\theta_{\mathrm{in}}\right)\right] \qquad [53]$$

The parametric solution for the z-coordinate of the electron orbit is given by $z = t - \theta$, in which t is given by eqn [52]. Over one orbital period, T, the electron undergoes a net displacement $r_0 = (x_0, 0, z_0)$ where x_0 is given by the increase in eqn [51] as θ increases by 2π, and z_0 is simply $T - 2\pi$:

$$\begin{aligned} z_0 &= T - 2\pi \\ &= \frac{2\pi}{1-\beta_{z0}}\left[\beta_{z0} + \frac{a^3}{2}\left(\frac{1}{2} + \sin^2\theta_{\text{in}}\right)\right] \end{aligned} \quad [54]$$

$$x_0 = \frac{-2\pi a \sin\theta_{\text{in}}}{\gamma_0(1-\beta_{z0})} \quad [55]$$

Note that the electron trajectory depends on a, β_{z0}, and θ_{in} in a rather complicated manner. Accordingly, the fundamental frequency ω_1 of the radiation spectrum depends on these three quantities. For backscattered radiation ($\hat{n} = -\hat{z}$), the dependence on electron beam and on the laser becomes decoupled.

Following is a useful formula relating the change of t with respect to θ:

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = 1 - \beta_z = \frac{\gamma_0(1-\beta_{z0})}{\gamma} \quad [56]$$

which may be verified from eqns [47] and [49]. Equation [56] is also valid for arbitrary values of β_{x0}, β_{y0}, β_{z0}, and θ_{in}, in which case γ_0 is the electron's relativistic mass factor in the absence of the laser.

Radiation from Relativistic Electrons

An electron with displacement $r(t)$ and velocity $v(t)$ carries a current density $J(r,t) = ev(t)\delta r - r(t)]$, whose Fourier transform $J_{\text{F}}(k,\omega)$ may easily be obtained. The total work done, W, in ergs, performed by the current $\mathbf{J}(\mathbf{r},t)$ on the electric field $\mathbf{E}(\mathbf{r},t)$ is given by

$$\begin{aligned} W &\equiv \int \mathrm{d}\omega\,\mathrm{d}\Omega \frac{\mathrm{d}^2 W}{\mathrm{d}\omega\,\mathrm{d}\Omega} \\ &= \int \mathrm{d}\omega\,\mathrm{d}\Omega \frac{\omega^2}{c^2}\left[|\mathbf{J}_{\text{F}}(\mathbf{k},\omega)|^2 - |\mathbf{n}\cdot\mathbf{J}_{\text{F}}(\mathbf{k},\omega)|^2\right] \end{aligned} \quad [57]$$

where the $\mathbf{k}$-space differential volume $\mathrm{d}^3\mathbf{k} = k^2\,\mathrm{d}k\,\mathrm{d}\Omega$ is expressed in terms of the solid angle (Ω) in the direction of the unit vector $\mathbf{n} = c\mathbf{k}/\omega$. We immediately obtain from eqn [56], under the far-field approximation:

$$\frac{\mathrm{d}^2 W}{\mathrm{d}\Omega\,\mathrm{d}\omega} = \frac{e^2\omega^2}{4\pi^2 c}|\mathbf{n}\times[\mathbf{n}\times\mathbf{F}(\omega)]|^2 \quad [58]$$

$$\mathbf{F}(\omega) = \int_{-\infty}^{\infty} \mathrm{d}t\ \beta(t)\mathrm{e}^{\mathrm{i}\omega[t-\mathbf{n}\cdot\mathbf{r}(t)/c]} \quad [59]$$

Equation [58] gives the energy radiated by the electron in the direction of the unit vector, per unit solid angle, per unit frequency ω. Radiation damping is ignored throughout.

Let us consider the simplest case where the electron orbit is strictly a periodic function of time with period T and a net displacement $\mathbf{r}_0$ per period. Thus, we have for all integers m (positive, negative, or zero):

$$\beta(t+mT) = \beta(t), \quad \mathbf{r}(t+mT) = m\mathbf{r}_0 + \mathbf{r}(t) \quad [60]$$

Equation [59] may then be written as, because of Thomson scattering the energy of the scattered photon is much less than the electron rest mass in the average rest frame of the electron, i.e.:

$$\frac{\hbar\omega_{\text{L}}}{mc^3} \ll 1 \quad [61]$$

$$\begin{aligned} \mathbf{F}(\omega) &= \sum_{m=-\infty}^{\infty}\int_{mT}^{(m+1)T} \mathrm{d}t\ \beta(t)\mathrm{e}^{\mathrm{i}\omega[t-\mathbf{n}\cdot\mathbf{r}(t)/c]} \\ &= \sum_{m=-\infty}^{\infty} f(\omega)\mathrm{e}^{\mathrm{i}m\omega[T-\mathbf{n}\cdot\mathbf{r}_0/c]} \end{aligned} \quad [62]$$

where

$$f(\omega) = \int_0^T \mathrm{d}t\ \beta(t)\mathrm{e}^{\mathrm{i}\omega[t-\mathbf{n}\cdot\mathbf{r}(t)/c]} \quad [63]$$

and we have used eqn [60]. Upon using $\sum_m \mathrm{e}^{\mathrm{i}mx} = \sum_m 2\pi\delta(x - 2m\pi)$ in the last infinite sum in eqn [64] and the property of the Dirac delta function, $\delta(ax) = (1/a)\delta(x)$, we obtain from eqns [62] and [63], the following expression for the spectrum:

$$\mathbf{F}(\omega) = \sum_{m=-\infty}^{\infty} \mathbf{F}_m \delta(\omega - m\omega_1) \quad [64]$$

$$\omega_1 = \frac{2\pi}{T - \mathbf{n}\cdot\mathbf{r}_0/c}, \quad [65]$$

$$\mathbf{F}_m = \frac{\omega_1}{2\pi}\int_0^T \mathrm{d}t\ \bar{\beta}(t)\mathrm{e}^{\mathrm{i}m\omega_1[t-\mathbf{n}\cdot\mathbf{r}(t)/c]} \quad [66]$$

Note from eqn [64] that the radiation spectrum is discrete, for strictly periodic motion of the electron. The base frequency of this spectrum, ω_1, depends on the periodicity (T) of the electron, on the electron's net displacement ($\mathbf{r}_0$) in one such period, and on the direction ($\hat{n}$) in which the radiation is observed. Thus, the radiation spectrum is, in general, not at the

harmonic frequency of the laser (nor at the harmonics of the electron orbital frequency, $2\pi/T$). It would thus be wrong to simply insert $\omega = n\omega_0$ in eqn [63] and to replace the electron's orbital period T there by the laser's optical period $2\pi/\omega_0$ and consider the resultant value of that integral to give the spectral amplitude of the radiation at the nth harmonic of the laser frequency. Erroneous conclusions regarding generation of high laser harmonics have appeared in the literature based on such an intuitive (but incorrect) substitution.

The power, p_m (in erg/s), radiated at the harmonic frequency $\omega = m\omega_1$ per unit solid angle in the direction of the unit vector $\hat{n}$ is then given by (from eqns [58] and [64]):

$$p_m = \frac{e^2 m^2 \omega_1^2}{4\pi^2 c} |\mathbf{n} \times \mathbf{F}_m|^2 \qquad [67]$$

where the dimensionless spectral amplitude $\mathbf{F}_m$ is given in eqn [66]. It is easy to show from eqns [60] and [65] that the integrand in eqn [66] is a periodic function of t of period T. Integrals of this type are readily evaluated by the Romberg method. The radiation spectrum observed *exactly* in the forward direction of the laser ($\hat{n} = \hat{z}$) always has only one discrete frequency, $\omega = \omega_1 = \omega_0$, which is easily shown from eqn [65] upon using the first equality of eqn [53]. This statement is true regardless of the velocity of the electron or the laser intensity, and may easily be deduced from eqn [66] for this case. However, for an energetic electron beam that is almost co-propagating with the laser, such as that produced by the laser itself, high harmonics at the laser frequency may be observed in the direction just slightly off the laser direction.

In the backscattering direction of the laser ($n = -z$), if we set $\theta_{\text{in}} = 0$, one obtains the following expressions for ω_1 and p_m, the backscatter power at $\omega = m\omega_1$ (from eqn [67]):

$$\frac{\omega_1}{\omega_0} = \left(\frac{2}{2+a^2}\right)\left(\frac{1-\beta_{z0}}{1+\beta_{z0}}\right) = \left(\frac{2}{2+a^2}\right)\gamma_0^2(1-\beta_{z0})^2 \qquad [68]$$

$$p_m = \frac{A}{\gamma_0^2(1-\beta_{z0})^2}\left(\frac{\omega_1}{\omega_0}\right)^4 s_m \qquad [69]$$

where $A = e^2\omega_0^2/4\pi^2 c = 0.69[\lambda(1\ \mu\text{m})]^{-2}$ erg/s, $s_m = 0$ for $m = 0, \pm 2, \pm 4, \ldots$, and for $m = \pm 1, \pm 3, \pm 5, \ldots$,

$$s_m = (a\pi)^2 m^2 \left[J_{\frac{(m-1)}{2}}(m\kappa) - J_{\frac{(m+1)}{2}}(m\kappa)\right]^2 \qquad [70]$$

$$\kappa = \frac{a^2}{2(a^2+2)} \qquad [71]$$

The $J_\nu(x)$ above are Bessel functions of the first kind of order ν. They also appear in the quantity '$[JJ]$' or '$F_m(K)$' in the FEL/synchrotron literature, where K is the undulator/wiggler parameter K. Setting $K = a$, one finds $s_m = \pi^2(1 + K^2/2)^2 F_m(K)$. It is easy to show that $s_m = s_{-m}$ for all odd integers m. Note that the relative spectral shape of s_m depends only on a, and is independent of the electron beam energy. The discrete spectrum for small a approaches a continuum for $a^2 \gg 1$.

The maximum values of s_m, occurring at $m = M$ with a value s_M, are shown in Figure 9. Note that the frequency component $\omega = M\omega_1$ contains the highest backscattered power. In terms of the laser frequency, the frequency component $\omega = N\omega_0$ would contain the highest backscatter power, where $N = M\omega_1/\omega_0$. The total backscatter power, P_{T} (in ergs/s), per unit solid angle in the $\mathbf{n} = -\mathbf{z}$ direction is then given by $P_{\text{T}} = \Sigma p_m$, where the sum is taken over all odd values of m:

$$P_{\text{T}} \approx \frac{\gamma_0^6(1-\beta_{z0})^6}{(\lambda/1\ \mu\text{m})^3} \times \begin{cases} 13.7a^3, & (a < 0.3) \\ 11.1/a, & (a \gg 1) \end{cases} \qquad [72]$$

To summarize, due to the rapid acceleration of electrons in the direction of the light wave and decrease of the oscillation frequency in the deeply relativistic regime, the photon energy from relativistic nonlinear scattering scales only linearly with laser field strength, $\sim a$. However, the relativistic motion of the electron results in a reduction in the angle of the scattered light, such that the harmonics are generated in a low-divergence-angle forward-propagating beam. There is also a relativistic Doppler shift ($\sim\gamma^2$). For instance, this mechanism will be used in high-energy physics experiments to cleanly make constituent particles in the gamma–gamma ($\gamma\gamma$) collider, in which gamma rays with energy 200 GeV will be generated by Compton scattering 1 eV photons from 250 GeV energy conventionally accelerated electron beams ($\gamma = 5 \times 10^{11}$).

Using a much smaller, laser-based accelerator (discussed in the following section), a 1 eV photon can be upshifted by this Doppler shift to an energy of 50 keV, corresponding to subatomic spatial resolution and of interest in medical diagnostics, by an electron beam with only $\gamma = 200$ (100 MeV). In this case, the maximum efficiency is obtained for laser fields a of order unity. When compared with conventional light sources based on cm-wavelength magnetostatic wigglers, the electromagnetic wigglers

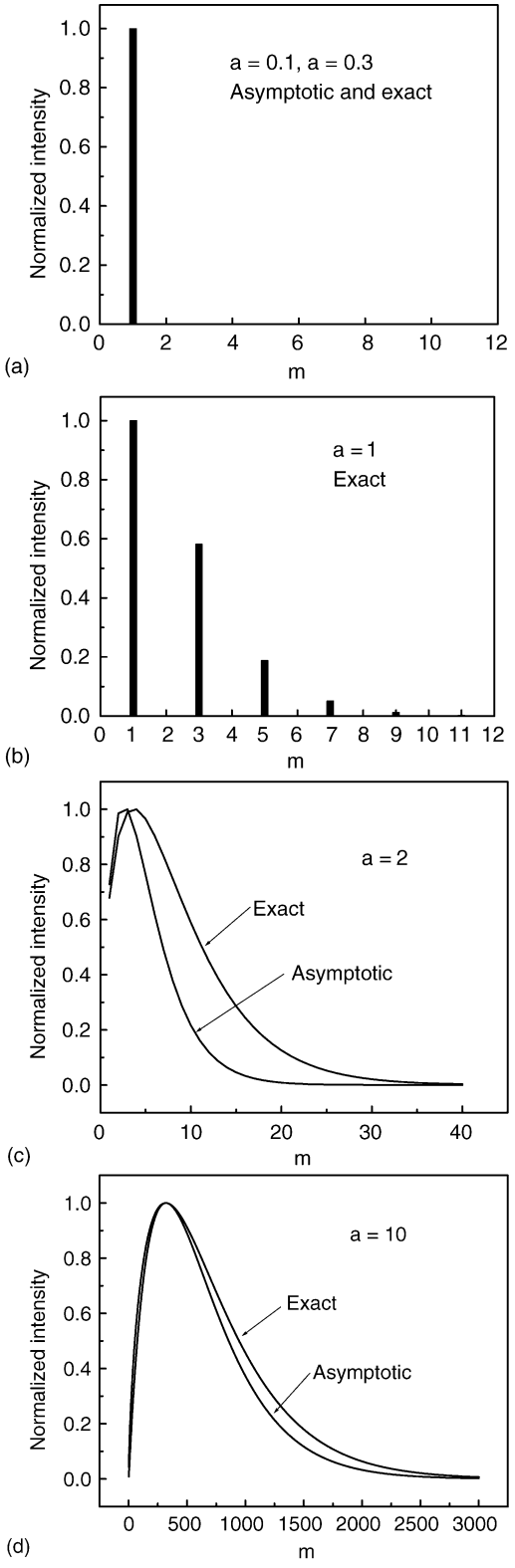


Figure 9 Normalized spectral distribution of s_m, at frequency $\omega = m\omega_1$. Here, s_m is normalized with respect to the maximum value s_M, occurring at $m = M$. (Reproduced from Lau Y, He F and Umstadter D (2003) Nonlinear Thomson scattering: A tutorial. *Physics Plasmas* 10: 2155, with permission from the American Institute of Physics.)

of such all-optical-laser-based EUV sources have ten-thousand times shorter wavelength (micron-scale). Thus, the total length of the wiggler region is correspondingly smaller (only mm in length). Another consequence of this is that the frequency upshift required to reach a given output wavelength is also ten thousand times smaller. Also, given that the required electron energy scales as the square root of the upshift, the required electron energy can be one-hundred-times lower (10–100 MeV). It follows from this, and the fact that the field gradients of these accelerators can be ten thousand times higher (1 GeV/cm) than convention RF-based accelerators, that the size of the accelerating region can, in principle, be a million times smaller (only mm in length). Besides its small size, this EUV source can produce femtosecond duration pulse and be synchronized with a relatively low jitter with another femtosecond light pulse having a different wavelength (by virtue of the possibility of deriving the two pulses from the same laser pulse); this is advantageous for the study of ultrafast pump-and-probe photo-initiated processes. The exceptionally low transverse emittance of laser-accelerated electron beams may even make it possible to generate coherent XUV radiation by means of the self-amplified spontaneous emission (SASE) free-electron lasing (FEL) mechanism.

Collective Plasma Response

The above equations apply to electron motion in vacuum. In the case of electrons in plasmas, the collective plasma electron response needs to be considered. A theory for the 1D nonlinear interaction of intense laser fields with electrons has been developed.

The variables are normalized such that

$$\mathbf{a} = e\mathbf{A}/mc^2, \qquad \phi = e\Phi/mc^2 \qquad [73]$$

$$\mathbf{u} = \mathbf{p}/mc, \qquad \beta = \mathbf{v}/c, \qquad \mathbf{u} = \gamma\beta \qquad [74]$$

$$\gamma = (1 + u^2)^{1/2} = (1 - \beta^2)^{-1/2} \qquad [75]$$

The relativistic force equation (Lorentz equation) is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u} = \nabla\phi + \frac{\partial}{\partial t}\mathbf{a} - \bar{\beta}\times(\nabla\times a) \qquad [76]$$

The energy equation ($\mathbf{u}\cdot$ force equation) is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\gamma = \beta\cdot\left(\frac{\partial}{\partial z}\phi + \frac{\partial}{\partial t}\mathbf{a}\right) \qquad [77]$$

The transverse force equation (1D) is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u}_\perp = \left(\frac{\partial}{\partial t} + \beta_z \frac{\partial}{\partial z}\right)\mathbf{a}_\perp \qquad [78]$$

Since $\mathbf{a} = \mathbf{a}(z,t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{a}_\perp = \left(\frac{\partial}{\partial t} + \beta_z \frac{\partial}{\partial z}\right)\mathbf{a}_\perp \qquad [79]$$

hence:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{u}_\perp - \mathbf{a}_\perp) = 0 \qquad [80]$$

This is just conservation of transverse canonical momentum. Assuming $\mathbf{u}_\perp = 0$ prior to the laser interaction gives

$$\mathbf{u}_\perp = \mathbf{a}_\perp \qquad [81]$$

which is the same as eqn [24].

The electron response to the normalized scaler and vector potentials of the form $\phi = \phi(z - ct)$ and $\mathbf{a} = \mathbf{a}(z - ct)$, which are a function of only $\xi = z - ct$, is completely described by the following constants of the motion:

$$\beta_\perp - a_\perp/\gamma = 0, \qquad [82]$$

$$\gamma(1 - \beta_z) - \phi = 1, \qquad [83]$$

$$n(1 - \beta_z) = n_0, \qquad [84]$$

where $\gamma = (1 - \beta^2)^{-1/2}$ is the relativistic factor and n is the electron density. This allows the various electron quantities to be specified solely in terms of the potentials:

$$\beta_\perp = \frac{2(1+\phi)a_\perp}{(1+a_\perp^2)+(1+\phi)^2}, \qquad [85]$$

$$\beta_z = \frac{(1+a_\perp^2)-(1+\phi)^2}{(1+a_\perp^2)+(1+\phi)^2}, \qquad [86]$$

$$\gamma = \frac{(1+a_\perp^2)+(1+\phi)^2}{2(1+\phi)}, \qquad [87]$$

$$\frac{n}{n_0} = \frac{(1+a_\perp^2)+(1+\phi)^2}{2(1+\phi)^2} \qquad [88]$$

In the single particle limit, $\phi \equiv 0$, eqns [26] and [30] are retrieved. For a long pulse interacting with a dense plasma, $\tau_\mathrm{L} \gg \omega_\mathrm{p}^{-1}$ (i.e., neglecting wakefield effects) it can be shown that $(1+\phi) \simeq (1+a_0^2)^{1/2}$ for circular polarization and $(1+\phi) \simeq (1+a_0^2/2)^{1/2}$ for linear polarization.

Propagation

To understand the propagation of high-intensity light in plasma, we need to understand how the dielectric properties of a plasma medium are affected by the relativistic electron mass change. A wave equation is found by taking the curl of Faraday's law:

$$\nabla \times \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} \qquad [89]$$

and using Ampere's law:

$$\nabla \times \mathbf{B} = 4\pi \mathbf{J}/c + \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} \qquad [90]$$

gives

$$\nabla \times \nabla \times \mathbf{E} = -\nabla^2 \mathbf{E} + \nabla(\nabla \cdot \mathbf{E}) = \frac{1}{c^2}\left[4\pi \frac{\partial \mathbf{J}}{\partial t} - \frac{\partial^2 \mathbf{E}}{\partial t^2}\right] \qquad [91]$$

where $\mathbf{J} = \sum n_i q_i v_i$ and the summation over the index i is done over all charged species of electrons and ions. Using Coulomb's law:

$$\nabla \cdot \mathbf{E} = 4\pi\rho \qquad [92]$$

where $\rho = \sum n_i q_i$, and assuming a uniform plasma, the second term on the left vanishes. Substituting eqn [13] for the velocity, we find

$$\left(-\frac{\partial^2}{\partial t^2} + c^2\nabla^2 - \omega_\mathrm{p}^2\right)\mathbf{E} = 0 \qquad [93]$$

where

$$\omega_\mathrm{p} = (4\pi n_i q_i^2/m_i)^{1/2} \qquad [94]$$

is the plasma frequency. Assuming again plane waves and Fourier analyzing, yields the well-known dispersion relation for electromagnetic waves in plasma:

$$\omega^2 = \omega_\mathrm{p}^2 + c^2 k^2 \qquad [95]$$

The index of refraction can be written fully relativistically as

$$\eta^2 = \frac{c^2 k^2}{\omega^2} = \frac{c^2}{v_\phi^2} = 1 - \frac{\omega_\mathrm{p}^2}{\omega^2} \qquad [96]$$

where v_ϕ is the phase velocity of the light wave. Assuming infinitely massive ions, the plasma frequency can be written as

$$\omega_\mathrm{p} = \frac{\omega_{\mathrm{p}0}}{\gamma^{1/2}} = \left(\frac{4\pi n_\mathrm{e} e^2}{\gamma m_0}\right)^{1/2} \qquad [97]$$

where ω_{p0} is the plasma frequency in a quiescent plasma, e is the electron charge, m_0 is the electron rest mass, and n_e is the plasma electron density. A change in mass changes the plasma frequency, which in turn modifies the index of refraction and the velocity of the light wave. The light's phase velocity then depends on the laser intensity. This can be seen clearly if we expand the phase velocity for small field strength ($a_0 \ll 1$):

$$v_\phi = \frac{c}{\eta} \sim c\left[1 + \frac{\omega_{p0}^2}{2\omega^2}\left(1 - \frac{\langle a_0^2\rangle}{2}\right)\right] \qquad [98]$$

An on-axis minimum of the phase velocity [i.e., $v_\phi(r) > v_\phi(0)$] can be created by a laser beam with an intensity profile peaked on axis, such as with a Gaussian beam, causing the wavefronts to curve inward and the laser beam to focus, as shown in Figure 10. When this focusing effect just balances the defocusing caused by diffraction, the laser pulse can propagate over a longer distance than it could in vacuum, while maintaining a small cross-section. This mechanism is referred to as relativistic self-guiding.

Relativistic Self-Focusing

The threshold for relativistic self-focusing can be obtained using the fully relativistic formalism with the vector potential, as in our previous discussion of electron motion in vacuum. In this case, we can rewrite eqn [91] as

$$\nabla^2\mathbf{a} - \frac{1}{c^2}\frac{\partial^2\mathbf{a}}{\partial t^2} = \frac{\omega_p^2}{c^2}\frac{n_e\mathbf{a}}{n_0\gamma} \qquad [99]$$

where $n_0 = n_i$ is the uniform background density. For a Gaussian beam focused in vacuum, where

$$\mathbf{a}(\mathbf{r},t) = \mathrm{Re}\left\{\mathbf{a}_0(\mathbf{r},t)\exp[\mathrm{i}(\mathbf{k}\cdot\mathbf{r} - \omega t)]\right\} \qquad [100]$$

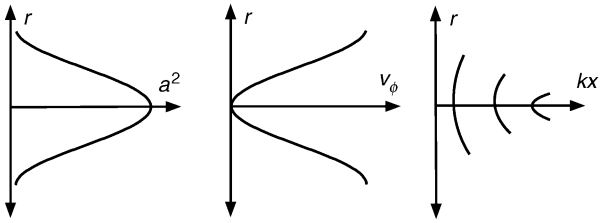


Figure 10 Relativistic self-focusing occurs when an on-axis peak in light intensity (left) produces an on-axis dip in the phase velocity (middle), which acts like a positive lens to cause the phase fronts to curve inward (right). (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

and the amplitude $\mathbf{a}(\mathbf{r},t)$ varies much less less in r and t than does the phase:

$$|\partial a_0/\partial t| \ll |\omega a_0|, \qquad |\partial a_0/\partial z| \ll |k a_0| \qquad [101]$$

then

$$a_0(r,z) = \frac{e^{-r^2/(r_0^2(1+z^2/L_r^2))}}{\sqrt{1+z^2/L_R^2}}\exp\left\{-\mathrm{i}\arctan\left(\frac{z}{L_R}\right) + \mathrm{i}\left(\frac{r}{r_0}\right)^2\frac{z/L_R}{1+z^2/L_R^2}\right\} \qquad [102]$$

satisfies the envelope equation (assuming second deriviatives are small compared to first derivatives), where L_R is the Rayleigh range and r_0 is the focal radius. In plasma, eqn [99] can then be written in this envelope approximation as

$$\left[\nabla_\perp^2 + 2\mathrm{i}k\frac{\partial}{\partial z}\right]a = -\frac{\omega_p^2}{c^2}\frac{|a|^2}{4}a \qquad [103]$$

The power at which the first and last terms of eqn [103] are in balance is

$$P = \frac{\pi R^2 I_0}{2} \qquad [104]$$

An equation for the beam radius R can thus be found:

$$\frac{\mathrm{d}^2R(z)}{\mathrm{d}z^2} = \frac{4}{k^2r_0^2}\left[1 - \frac{1}{22}\frac{\omega_p^2}{c^2}|a_0|^2 r_0^2\right] \qquad [105]$$

Self-focusing occurs when the two terms in the brackets in eqn [105] are equal, which gives for the critical power:

$$P_{crit} \simeq 17.4\cdot(\omega/\omega_p)^2\ \mathrm{GW} \qquad [106]$$

This corresponds to 1 TW for 1 μm light focused into a gas with electron density of $10^{19}/\mathrm{cm}^3$. Note that this is a power threshold, not an intensity threshold, because the tighter the focusing, the greater the diffraction (the first term in eqn [103]). Numerous recent experiments have confirmed this focusing mechanism when $P > P_{crit}$.

Any spatial variation of the laser intensity will act to push an electron to regions of lower intensity through the so-called ponderomotive force:

$$\mathbf{F} = \nabla\gamma = \nabla\sqrt{1+a^2} = (2\gamma)^{-1}\nabla a^2 \qquad [107]$$

$$\mathbf{F}_{\text{pond}} = -\frac{m_0 c^2}{8\pi} \frac{\nabla a_0^2}{\sqrt{1+\frac{a_0^2}{2}}} \cdot \left[\left(\frac{2}{a_0^2} + 1 \right) E(2\pi, \kappa) - \frac{2}{a_0^2} F(2\pi, \kappa) \right] \quad [108]$$

where

$$\kappa = \frac{a_0^2/2}{1 + a_0^2/2} \quad [109]$$

and F and E are the elliptic integrals of the first and second kind, respectively. In the low intensity limit, i.e., when $a_0^2 \ll 1$;

$$\mathbf{F}_{\text{pond}} = -\frac{m_0 c^2}{4} \nabla a_0^2 \quad [110]$$

That is, the laser ponderomotive force is roughly proportional to the gradient of laser intensity.

A Gaussian-shaped laser's ponderomotive force will tend to expel electrons radially from the region of the axis, so called 'electron cavitation.' If the ponderomotive force is high enough for long enough, the charge displacement due to expelled electrons (in either the lateral or longitudinal direction) will eventually cause the ions to move as well through the Coulomb electrostatic force, forming a density channel. Because $n_e(0) < n_e(r)$, and thus $v_\phi(0) < v_\phi(r)$, this enhances the previously-discussed relativistic self-guiding and can itself guide a laser pulse.

For plasmas created by photo-ionization of a gas by a Gaussian laser pulse, the density will be higher on the axis than off the axis. If we instead expand the phase velocity in terms of changes in density:

$$\frac{v_\phi}{c} \sim 1 + \frac{\omega_{p0}^2}{\omega^2} \frac{\Delta n_e}{n_e} \quad [111]$$

The phase velocity will thus be higher on axis, which will tend to defocus the light and increase the self-guiding threshold. In order to avoid this, gases with low atomic number and thus fewer available electrons, such as H_3, are commonly used as targets.

Raman Scattering, Plasma Wave Excitation and Electron Acceleration

The local phase velocity, described in eqns [98] and [111], can also vary longitudinally if the intensity and/or electron density does. Local variation in the index of refraction can 'accelerate' photons, i.e., shift their frequency, resulting in photon bunching, which in turn bunches the electron density through the ponderomotive force (F), and so on. When the laser pulse duration is longer than an electron plasma period, $\tau \gg \tau_p = 2\pi/\omega_p$, this photon and electron bunching grows exponentially, leading to the stimulated Raman scattering instability. Energy and momentum must be conserved when the electromagnetic wave $(\omega_0, \mathbf{k}_0)$ decays into a plasma wave $(\omega_p, \mathbf{k}_p)$ and another light wave $(\omega_0 - \omega_p, \mathbf{k}_0 - \mathbf{k}_p)$.

From an equivalent viewpoint, the process begins with a small density perturbation, Δn_e, which, when coupled with the quiver motion, eqn [13], drives a current $J = \Delta n_e e v_e$. This current then becomes the source term for the wave equation (eqn [91]), driving the scattered light wave. The ponderomotive force, due to the beating of the incident and scattered light wave, enhances the density perturbation, creating a plasma wave and the process begins anew. In three dimensions, a plasma wave can be driven when transverse self-focusing and stimulated Raman scattering occur together, a process called the self-modulated wakefield instability.

Two conditions must be satisfied for self-modulation to occur in the plasma. First, the laser pulse must be long compared to the plasma wave, $L \gg \lambda_p$. This allows the Raman instability time to grow, and it allows for feedback from the plasma to the laser pulse to occur. Second, the laser must be intense enough for relativistic self-focusing to occur, $P > P_c$, so that the laser can be locally modified by the plasma. Under these conditions, the laser can form a large plasma wave useful for accelerating electrons.

As the long laser pulse enters the plasma, it will begin to drive a small plasma wave due to either forward Raman scattering or the laser wakefield effect from the front of the laser pulse. This small plasma wave will have regions of higher and lower density with both longitudinal and radial dependence. That is, the plasma wave will be three-dimensional in nature with a modulation along the propagation direction of the laser and a decay in the radial direction to the ambient density (see **Figure 11**). The importance of this lies with how it affects the index of refraction in the plasma. In the regions of the plasma wave where the plasma density is lower, the radial change in the index of refraction is negative, $\partial n(r)/\partial r < 0$. This means that this part of the plasma acts like a positive lens and focuses the laser. Whereas regions of the plasma wave where the density is higher, $\partial n(r)/\partial r > 0$, the opposite occurs and the laser defocuses. This has the effect of breaking up the laser pulse into a series of shorter pulses of length $\lambda_p/2$ which will be separated by the plasma period. The instability occurs because of how the plasma responds to this. Where the laser is more tightly focused, the ponderomotive force will be

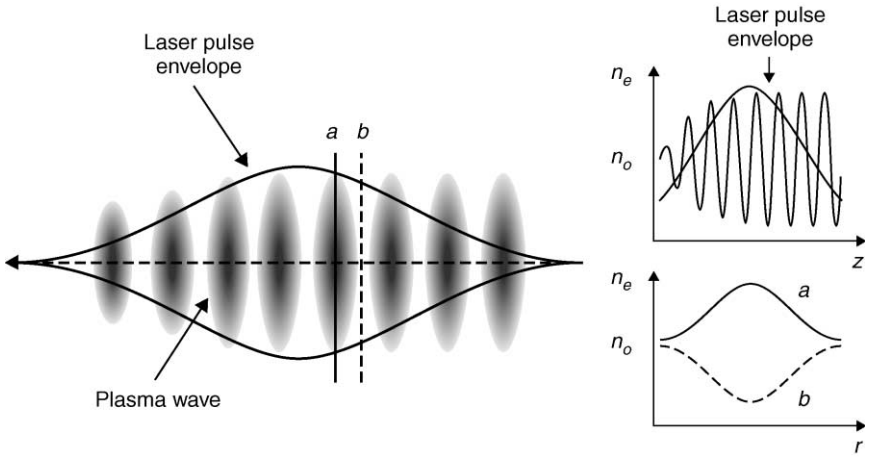


Figure 11 The plasma wave generated by a SMLWFA is three-dimensional in nature. Note that the darker regions correspond to areas of higher plasma density. The graphs to the right represent lineouts of the plasma density longitudinally and radially at the indicated points. (Reproduced with permission from Wagner R (1998) Laser–plasma electron accelerators and nonlinear relativistic optics. PhD thesis, University of Michigan.)

greater and will tend to expel more electrons. This decreases the density in these regions even further, resulting in more focusing of the laser. This feedback rapidly grows, hence the instability.

The phase velocity of the plasma wave in the case of forward scattering is equal to the group velocity of the beat wave, which for low-density plasma is close to the speed of light, as can be seen from the relation:

$$v_\phi = \omega_p/k_p = \Delta\omega/\Delta k = v_g = c\eta \sim c \qquad [112]$$

where eqn [94] and $\omega_p^2 \ll \omega^2$ were used to show that η is close to unity. Such relativistic plasma waves can also be driven by short pulses ($\tau \sim \tau_p$). In this case, the process is referred to as laser–wakefield generation, referring to the analogy with the wake driven by the bow of a boat moving through water, but the mechanism is similar (except it has the advantage that the plasma wave is driven linearly instead of as an instability).

In either case, the resulting electrostatic plasma wave can continuously accelerate relativistic electrons with enormous acceleration gradients. The gradient can be estimated from eqn [92] and the fact that because

$$\nabla \cdot \mathbf{E} \sim E\omega_p/c \propto E\sqrt{n_e} \qquad [113]$$

then

$$E \sim \sqrt{n_e}\ \text{eV/cm} \qquad [114]$$

corresponding to 1 GeV/cm for $n_e = 10^{18}\ \text{cm}^{-3}$. Because this gradient is four orders of magnitude greater than achieved by conventional accelerators (based on fields driven by radio-frequency waves pumped into metal cavities), laser-driven plasma accelerators have received considerable recent attention. They have been shown to accelerate an amount of electron charge (100 nC) comparable to that from conventional accelerators and to have superior transverse geometrical emittance (product of divergence angle and spotsize, similar to the $f/\#$ in light optics). However, their longitudinal emittance is currently much inferior, energy spreads of 100%. They have been shown to be useful for much of the same applications: radio-isotope production, radiation chemistry, as well as X-ray, proton, and neutron generation. Once the longitudinal emittance can be reduced, they may be advantageous for, among other applications, injectors (especially of short-lived unstable particles) into larger conventional accelerators for high-energy physics research and light sources, and, as discussed in the section on radiation from relativistic electrons about, as stand-alone all-optically driven ultrashort-pulse duration X-ray sources.

The SIMLAC code has been used to study wakefield generation and laser propagation in the limit $a^2 \ll 1$. It draws from nonlinear optics models and treats propagation in the group velocity frame. In this idealized model (which assumes perfect Gaussian beams), the pulse and wake are maintained over long enough propagation distances to accelerate an electron to GeV energy, as shown in Figure 12. A three-dimensional envelope equation for the laser field was derived that includes nonparaxial effects, wakefields, and relativistic nonlinearities.

The resonant wakefield has been characterized experimentally by temporal interferometry, as shown in Figure 13. However, this was done only for the tight-focusing case in which the laser spotsize is much

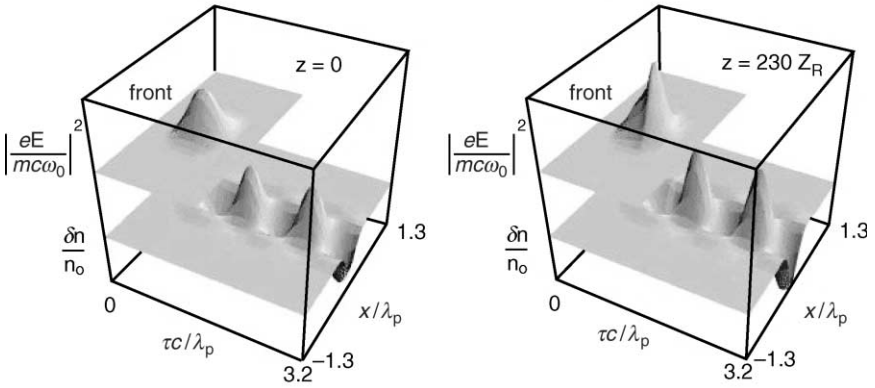


Figure 12 The 'standard' resonant wakefield simulated with SIMLAC, a code that moves at the light pulse's group velocity. (Reproduced with permission from Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774.)

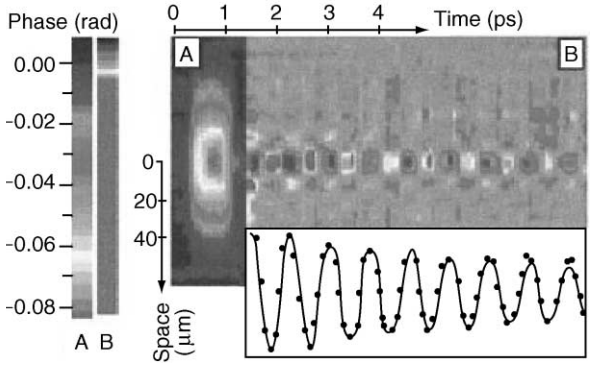


Figure 13 Typical result of a phase shift measurement to study a resonantly excited laser wakefield plasma wave by means of time-domain interferometry. Parts (A) and (B) have different color scales. The bottom graph is a line out of part (B) along the laser axis. (Reproduced with permission from Marquès JR, Dorchies F, Audebert P, Ceindre JP, Amiranoff F, Gauthier JC, Hammoniaux G, Antonetti A, Chessa P, Mora PTM and Antonson J (1997) Frequency increase and damping of nonlinear electron plasma oscillations in cylindrical symmetry. *Physics Review Letters* 78: 3463. Copyright (1995) by the American Physical Society.)

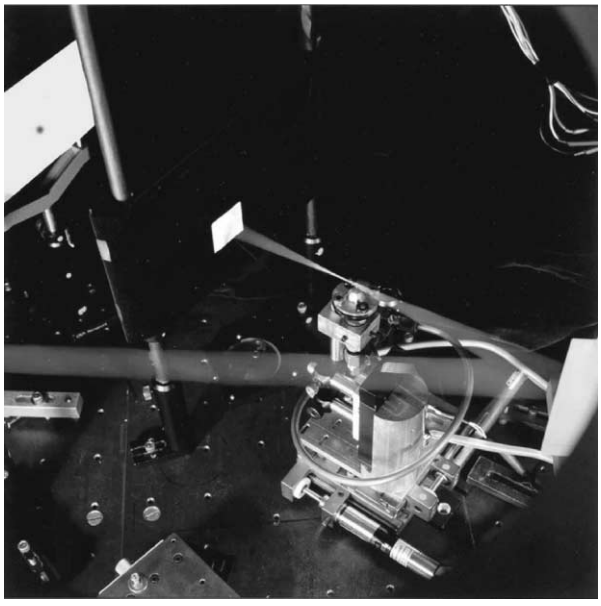


Figure 14 Artistically enhanced photograph of the acceleration of an electron beam by a laser interacting with a gas jet inside a vacuum chamber. The laser crosses the picture from left to right and is focused by a parabolic mirror (right side of the picture). The supersonic nozzle (shown in the middle of the picture) is positioned with micron accuracy with a 3-axis micropositioner. The e-beam makes a small spot on a white flourescent (LANEX) screen, shown in the upper left-hand corner of the picture. (Reproduced with permission from Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774, with permission from the American Institute of Physics.)

smaller than the plasma wave wavelength ($r_l \ll \lambda_p$) and thus the transverse wakefield was much greater than the longitudinal wakefield.

A typical experimental setup, used to study electron acceleration, is shown in Figure 14.

Dramatic reduction of the angular divergence of a laser accelerated electron beam was observed with increasing laser power above the relativistic self-focusing threshold, as shown in Figure 15.

Interactions with Solid-Density Targets

The generation of electrons by high-intensity laser light interacting with solid targets can generate energetic X-rays, accelerate other types of particles and induce nuclear reactions, as illustrated in Figure 16. For instance, high-order harmonics have been generated by the oscillation of the critical density surface, in the so-called moving mirror model. Bright X-rays, originating from Bremstrahlung caused by electron collisions with high-Z atoms in solid targets, have created isotopes by means of photofission. Laser-accelerated electron energies and

angular distributions have been inferred from analyzing (γ, n) and $(\gamma, 2n)$ reactions in composite Pb/Cu targets and in Ta/Cu targets. Positrons were created by colliding laser-accelerated electrons with a tungsten target.

When electrons are heated to high temperatures or accelerated to high energies, they can separate from plasma ions. Such charge displacement creates an electrostatic sheath, which eventually accelerates the ions. The ions are pulled by the charge of the electrons and pushed by the other ions' unshielded charges (similar to the 'Coulomb explosion' that can occur during the ionization of atoms). When the charge displacement is driven by thermal expansion, as in long-pulse (low-power) laser–plasma experiments, the maximum ion energies are limited to less than 100 keV. However, when the charge displacement is driven by direct laser heating, as in short-pulse high-power laser–plasma experiments, multi-MeV ion energies are possible. This was first shown with gas jet targets, in which case the ions were accelerated radially into 2π, and then later with thin solid-density-films, in which case the ions were accelerated into collimated beams. In the latter case, hydrocarbons and water on the surface of the film can become ionized and provide a source of protons to be accelerated.

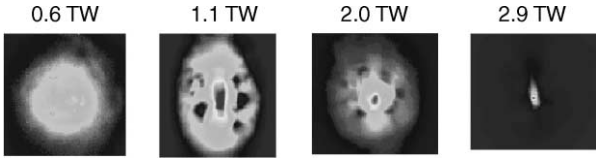


Figure 15 Images of the spatial profiles of the electron beam measured by a ccd camera imaging a LANEX screen at a distance of 15 cm from the gas jet for various laser powers. The divergence angle of the beam decreases to a value of $\Delta\theta = 1°$ at a power of 2.9 TW, corresponding to a transverse geomtrical emittance of just $\varepsilon_\perp \lesssim 0.06\pi$ mm-mrad. (Reproduced with permission from Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774, with permission from the American Institute of Physics.)

An intense laser can ponderomotively heat electrons. If the laser contrast is high, vacuum heating can occur in the following manner. When light encounters a sharp interface between vacuum and solid density, the electromagnetic field becomes evanescent in the region above the critical density. The instantaneous '$v \times B$' force can push electrons in the direction of the light's propagation vector; it also has a frequency twice that of the pump and a magnitude proportional to the square of the normalized vector potential, a_0^3. Thus, electrons can only complete half of their figure-eight orbits, on the vacuum side, gaining relativistic energies; they move through the overdense region without the electromagnetic field to pull them back. An electrostatic sheath can thus form, which will accelerate the ions left behind. Another important heating mechanism is stochastic heating, which occurs when the light that is reflected from the critical surface beats with the incoming wave to create a standing wave. The motion of electrons in such a wave can become chaotic, resulting in a large increase in electron temperature (>100 keV).

As the heated electrons propagate through a solid, they can instantaneously field-ionize the neutral atoms of the solid. This will both modify the solid's

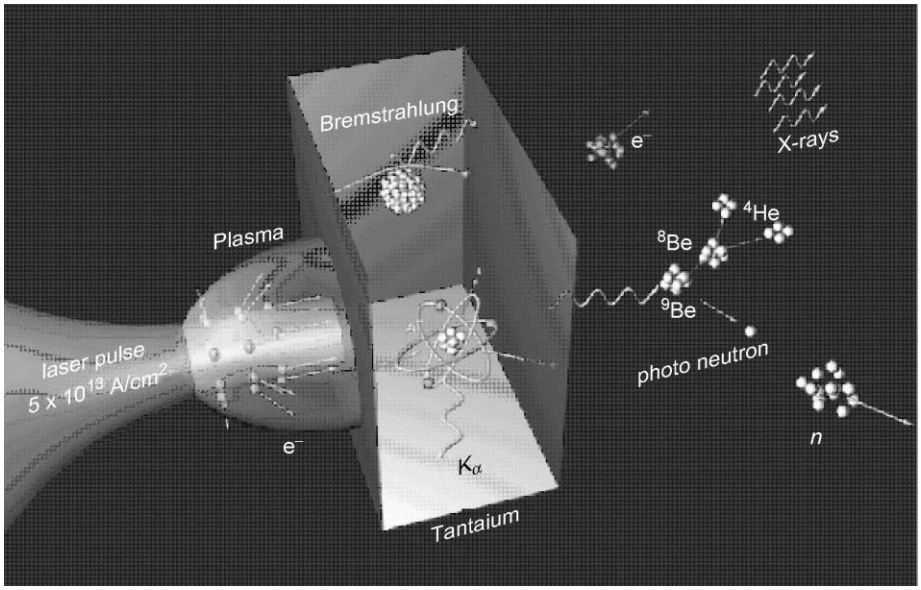


Figure 16 Illustration of the various mechanisms of particle acceleration and X-ray generation in high-intensity laser interactions with solid-density targets. (Courtesy of R. Sauerbrey.)

conductivity and provide a source of protons on the rear-side of the target. If the film is thin enough, the electrons can pass through, and create a sheath on the rear-side of the target. This latter mechanism has been dubbed the target normal sheath acceleration mechanism. The ions from thin foils have been claimed to originate from both the front and rear-side of the foil.

Several groups have reported the observation of ions originating from thin-film solid-density targets. Unlike previous long-pulse experiments, the ions were accelerated along the direction normal to the side of the target, that is opposite to that upon which the laser was incident. The ions generally originate from water or hydrocarbons on the surface of the material. The acceleration results from several different mechanisms, which may be occurring simultaneously. Charge-displacement is again common to all, with the electrons being heated ponderomotively, such as by Brunel, $J \times B$ or stochastic heating. In one case, the electrostatic sheath is formed at the backside of the ionization layer formed on the side of the target upon which the laser is incident (front side). In another case, the electrostatic sheath is formed by field ionization of the ion layer on the opposite side of the thin film target (back-side), the target normal sheath acceleration (TNSA) hypothesis. Numerical simulations show evidence for both front- and back-side acceleration.

The results of these experiments indicate that a large number of protons (10^{13} p) can be accelerated, corresponding to source current densities (10^8 A/cm^2) that are nine orders of magnitude higher than produced by cyclotrons, but with comparable, or even lower, transverse emittances ($\varepsilon_\perp \leq 1.0\pi$ mm-mrad). Proton energies up to 60 MeV have been observed in experiments at intensities exceeding 10^{20} W/cm^2 (using a petawatt power laser). The high end of the proton spectrum typically has a sharp cut-off, but is a continuum. In one experiment, protons were observed to be emitted in ring patterns, the radii of which depend on the proton energy, which was explained by self-generated magnetic fields.

A typical experimental setup used to observe the acceleration of ions is shown in Figure 17. The production of radionuclides have been used as an ion energy diagnostic. In another example of a nuclear reaction initiated by an intense laser, neutrons

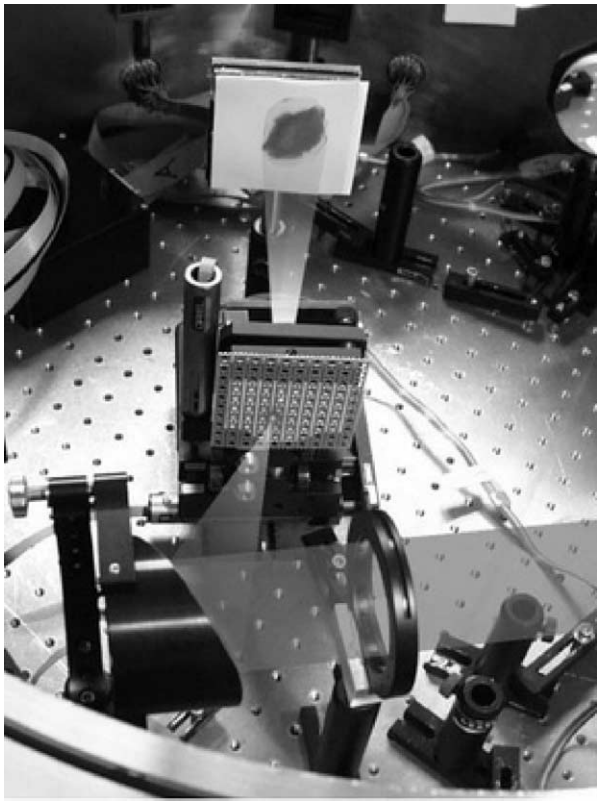


Figure 17 Artistically enhanced photograph of the typical setup used to observe the acceleration of ions. The laser (shown in the foreground) is focused with an off-axis parabola onto a thin-foil, held by a mesh that is positioned by a 3-axis micropositioner. CR-39, a nuclear track detector (shown in the background) is used to detect the ions. An actual proton-produced pattern is shown in red. (Reproduced from Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774, with permission from the American Institute of Physics).

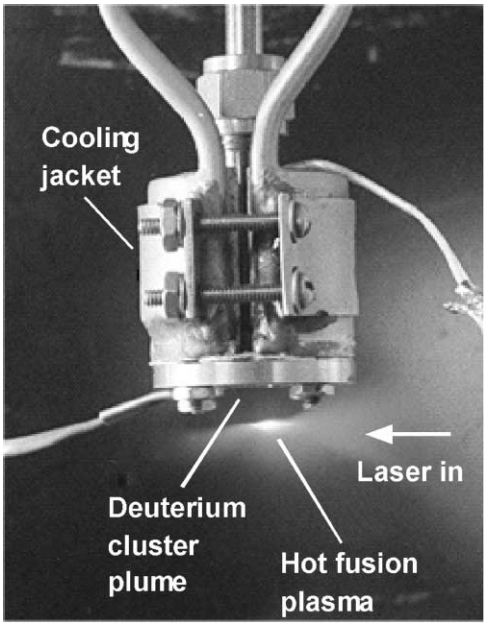


Figure 18 Photograph of a laser interacting with a deuterium cluster jet emanating from cooled gas nozzle. Fusion neutrons are produced with relatively low laser intensities. (Reproduced with permission from Ditmire T (2002) Laser fusion on a tabletop. *Optics and Photonics News* 13: 28, with permission from IEEE.)

have been produced by the He fusion reaction in the focus of 200 mJ, 160 fs Ti:sapphire laser pulses on a deuterated polyethylene target. Optimizing the fast electron and ion generation by applying a well-defined prepulse led to an average rate of 140 neutrons per shot. Neutrons have also been generated from cluster plasmas, which were produced by the cooled-nozzle depicted in Figure 18, but with significantly lower laser intensities than required with planar solid targets.

Concluding Remarks

The field of nonlinear optics, with electrons bound to atoms, has over the last forty years given rise to many scientific discoveries and technologies that are now commonplace. The relatively young field of relativistic nonlinear optics has already begun to do the same. Some of the exciting phenomena and applications that have already been identified have been discussed in this article. Others await discovery as ever higher laser intensities are reached, with no foreseeable limit.

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Scattering:** Raman Scattering.

Further Reading

Chen S-Y, Maksimchuk A and Umstadter D (1998) Experimental observation of relativistic nonlinear Thomson scattering. *Nature* 396: 653.

Cowan TE, Hunt AW, Phillips TW, Wilks SC, Perry MD, Brown C, Fountain W, Hatchett S, Johnson J, Key MH, Parnell T, Pennington DM, Snavely RA and Takahashi Y (2000) Photonuclear fission from high energy electrons from ultraintense laser–solid interactions. *Physics Review Letters* 84: 903.

Ditmire T (2002) Laser fusion on a tabletop. *Optics and Photonics News* 13: 28.

Eberly JH (1986) Fundamentals of laser interactions, vol. 229 of lecture notes in physics. *Journal Optical Society of America B* 3: 1324.

Esarey E, Sprangle P, Krall J and Ting A (1996) Overview of plasma-based accelerator concepts. *IEEE Transactions on Plasma Science* 24: 252.

Gibbon P and Förster E (1996) Short-pulse laser–plasma interactions. *Plasma Phys. Controlled Fusion* 38: 769.

Hartemann F (2001) *High-Field Electrodynamics*. Boca Raton: CRC Press.

Joshi CJ and Corkum PB (1995) Interactions of ultra-intense laser light with matter. *Physics Today* 48: 36.

Lau Y, He F and Umstadter D (2003) Nonlinear Thomson scattering: A tutorial. *Physics Plasmas* 10: 2155.

Ledingham KWD, Spencer I, McCanny T, Singhal RP, Santala MI, Clark E, Watts I, Beg FN, Zepf M, Krushelnick K, Tatarakis M, Dangor AE, Norreys PA, Allott R, Neely D, Clark RJ, Machacock AC, Wark JS, Cresswell AJ, Sanderson DCW and Magill J (2000) Photonuclear physics when a multiterawatt laser pulse interacts with solid targets. *Physics Review Letters* 84: 899.

Luther-Davies B, Gamaly EG, Wang Y, Rode AV and Tikhonchuk V (1992) Matter in ultrastrong laser fields. *Soviet Journal Quantum Electronics* 22: 289.

Marquès JR, Dorchies F, Audebert P, Ceindre JP, Amiranoff F, Gauthier JC, Hammoniaux G, Antonetti A, Chessa P, Mora PTM and Antonson J (1997) Frequency increase and damping of nonlinear electron plasma oscillations in cylindrical symmetry. *Physics Review Letters* 78: 3463.

Meyer-ter-Vehn J, Pukhov A and Sheng Z-M (2001) Relativistic laser plasma interaction. In: Batani D (ed.) *Atoms, Solids and Plasmas in Super-Intense Laser Fields*, pp. 167. New York: Kluwer Academic–Plenum.

Moore CI, Knauer JP and Meyerhofer DD (1995) Observation of the transition from Thomson to Compton scattering in multiphoton interactions with low-energy electrons. *Physics Review Letters* 74: 2439.

Mourou G and Umstadter D (1992) Development and applications of compact high-intensity lasers. *Physics Fluids B* 4: 2315.

Mourou GA and Umstadter D (2002) Extreme light. *Scientific America* 81.

Mourou GA, Barty CPJ and Perry MD (1998) Ultrahigh-intensity lasers: Physics of the extreme on a tabletop. *Physics Today* 22.

Siders CW, Blanc SPL, Fisher D, Tajima T, Downer MC, Babine A, Stepanov A and Sergeev A (1996) Laser wakefield excitation and measurement by femtosecond longitudinal interferometry. *Physics Review Letters* 76: 3570.

Sprangle P, Peñano JR, Hafizi B, Hubbard RF, Ting A, Gordon DF, Zigler A and Antonsen TM (2002) GeV acceleration in tapered plasma channels. *Physics Plasmas* 9: 2364.

Tajima T and Mourou G (2002) Zettawatt-exawatt lasers and their applications in ultrastrong-field physics. *Physics Review STAB* 5: 031301.

Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774.

Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165.

Umstadter D and Norris TB (eds) (1997) Nonlinear optics with relativistic electrons. *IEEE Journal Quantum Electronics* 33: 1878.

Wagner R (1998) Laser–plasma electron accelerators and nonlinear relativistic optics. PhD thesis, University of Michigan.

Wagner R, Chen S-Y, Maksimchuk A and Umstadter D (1997) Electron acceleration by a laser wakefield in a relativistically self-guided channel. *Physics Review Letters* 78: 3125.

Yamakawa K (2004) Table-top lasers create ultrahigh peak powers. *OYOBUTURI* 73: 186–193.

S

SCATTERING

Contents

Raman Scattering
Scattering from Surfaces and Thin Films
Scattering Phenomena in Optical Fibers
Scattering Theory
Stimulated Scattering

Raman Scattering

F Kannari, Keio University, Yokohama, Japan

© 2005, Elsevier Ltd. All Rights Reserved.

When an electromagnetic wave passes over a small elastically bound charged particle, the particle will be set into motion by the electric field. Scattering takes place for frequencies not corresponding to the natural frequencies of the particles. If the particle is bound by a force obeying Hooke's law, this vibration will have the same frequency and spatial direction as that of the electric force in the electromagnetic wave. The phase, on the other hand, of the forced vibration will differ from that of the incident wave, which causes the basis of dispersion. If the charged oscillation is bound by a force which does not obey Hooke's law, but some more complicated law, it will reradiate not only the impressed frequency, but also various combinations of this frequency with the fundamental and overtone frequencies of the oscillator.

The Raman effect was discovered by C.V. Raman (1888–1971). Ordinary Raman spectroscopy is used as a tool for studying the vibrational energy levels of molecules and of lattice optical branch vibration in crystals. A spectroscopic measurement of the scattered light reveals the existence of frequencies shifted down by increments corresponding to vibrational frequencies of the material. This down-shifted scattering is referred to as Stokes scattering. Frequencies shifted up by the vibrational frequencies are also present in the scattered radiation, which are the so-called anti-Stokes scattering.

Stimulated amplification of scattered Stokes radiation occurs through a third-order nonlinear process. To introduce stimulated Raman scattering (SRS), first we describe the steady-state stimulated scattering theory.

In general, nonlinear optical processes are described by the classical Maxwell equation with the nonlinear polarization of a material $\mathbf{P}^{\mathrm{NL}}$:

$$\nabla \times \nabla \times \mathbf{E} + \varepsilon\mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} = -\mu_0 \frac{\partial^2 \mathbf{P}^{\mathrm{NL}}}{\partial t^2} \qquad [1]$$

where ε and μ_0 are the linear dielectric constant of the material and the magnetic permeability of the vacuum. We consider a relatively intense excitation radiation $\mathbf{E}_{\mathrm{L}}$ (frequency ω_{L}) and a weak signal radiation $\mathbf{E}_{\mathrm{S}}$, having the carrier frequency ω_{s} corresponding to the scattered light frequency ($\omega_{\mathrm{s}} = \omega_{\mathrm{L}} - n\omega_0$);

$$\mathbf{E}_{\mathrm{L}} = \frac{1}{2}\{\bar{\mathbf{E}}_{\mathrm{L}}(\mathbf{r}, t) \exp \mathrm{i}(\mathbf{k}_{\mathrm{L}} \times \mathbf{r} - \omega_{\mathrm{L}} t) + \mathrm{c.c.}\} \qquad [2]$$

$$\mathbf{E}_{\mathrm{S}} = \frac{1}{2}\{\bar{\mathbf{E}}_{\mathrm{S}}(\mathbf{r}, t) \exp \mathrm{i}(\mathbf{k}_{\mathrm{S}} \times \mathbf{r} - \omega_{\mathrm{S}} t) + \mathrm{c.c.}\} \qquad [3]$$

where $\bar{\mathbf{E}}_L(\mathbf{r},t), \bar{\mathbf{E}}_S(\mathbf{r},t)$ are a slowly varying envelope of the complex electric fields. Assuming that the variation of the electric field during the time-scale shorter than the decay time of the induced oscillation in the material is negligible, nonlinear polarization induced by the excitation radiation is described as follows:

$$\mathbf{P}_S^{NL} = \frac{1}{2}\{\bar{\mathbf{P}}_S^{NL}(\mathbf{r},t)\exp i(\mathbf{k}_S\times\mathbf{r}-\omega_S t)+\text{c.c.}\} \quad [4]$$

$$\bar{\mathbf{P}}_S^{NL} = \varepsilon_0\chi^{NL}(-\omega_S,-\omega_L,\omega_L,\omega_S)\bar{\mathbf{E}}_L^*\bar{\mathbf{E}}_L\bar{\mathbf{E}}_S \quad [5]$$

where ε_0 is the vacuum dielectric constant and $\chi^{NL}(-\omega_S,-\omega_L,\omega_L,\omega_S)$ is the third-order nonlinear susceptibility, of which the real part corresponds to the variation of the dielectric constant depending on the optical radiation, whereas the imaginary part describes the stimulated scattering. In steady-state approximation, substituting eqns [3]–[5] into [1], we obtain

$$\frac{d\bar{\mathbf{E}}_S}{dz} = \frac{1}{2}\left(-a+i\frac{\omega_S}{cn_S}\chi^{NL}|\bar{\mathbf{E}}_L|^2\right)\bar{\mathbf{E}}_S \quad [6]$$

where we ignore the higher-order differential of amplitude. n_S is the refractive index at ω_S, and a is the linear absorption coefficient of the signal radiation. When ignoring the decay of the excitation radiation along the z direction in the material, for making the right-hand term of eqn [6] positive, the next condition must be satisfied:

$$I_L = \frac{1}{2}c\varepsilon_0 n_L|\bar{\mathbf{E}}_L|^2 \geq I_L^{th} = \frac{c^2\varepsilon_0 n_S n_L}{2\omega_S}\frac{a}{-\text{Im}\chi^{NL}} \quad [7]$$

The signal radiation is then amplified along the z direction at an amplification constant of g,

$$G = g(I_L - I_L^{th}) \quad [8]$$

$$g = \frac{\omega_S}{c^2\varepsilon_0 n_S n_L}(-\text{Im}\,\chi^{NL}) \quad [9]$$

Stimulated Raman Scattering

In SRS, the frequency difference $\omega_L - \omega_S$ corresponds to the resonance frequency of vibrational oscillation ω_V. Therefore, the vibrational oscillation mode is excited in the material as well as the generation of scattered radiation. In the following section, we consider the SRS by optical phonon. In general, nonlinear polarization under the optical excitation is described as follows:

$$\mathbf{P}^{NL} = N_0\alpha' q\mathbf{E}, \quad [10]$$

$$\ddot{q} = \Gamma_V\dot{q} + \omega_V^2 q = \frac{1}{2}\alpha'\mathbf{E}^2 \quad [11]$$

where q is the basic coordinate of molecular vibration, Γ_V is the decay rate of the vibrational oscillation, $\alpha' = (\partial\alpha/\partial q)_0$ is the Raman polarizability of a molecule, and N_0 is the molecule density. We can obtain the steady-state nonlinear polarization from eqns [4], [5], [10], and [11] as

$$\chi^{NL} = (-\omega_S,-\omega_L,\omega_L,\omega_S) = \frac{N_0(\alpha')^2}{4\varepsilon_0}\frac{1}{\omega_V^2-\omega^2+i\omega\Gamma_V} \quad [12]$$

where $\omega = \omega_L - \omega_S$. Thus, the gain coefficient can be obtained from eqns [9] and [12] under the assumption of $\omega_V \gg \Gamma_V$, as follows:

$$g_R(\omega_S) = \pm\frac{N_0(\alpha')^2\omega_S}{2c^2\varepsilon_0 n_S n_L\omega_V\Gamma_V} \times\frac{(\Gamma_V/2)^2}{(\omega_L-\omega_S\mp\omega_V)^2+(\Gamma_V/2)^2} \quad [13]$$

In general, Raman scattering is characterized by the scattering cross-section $d\sigma/d\Omega$, which is defined for one of the optical polarization directions in terms of the optical power density:

$$g_R(\omega_L-\omega_V) = \frac{16\pi^2c^2N_0}{\hbar n_S^2\omega_S^3\Gamma_V}\frac{d\sigma}{d\Omega}\left\{1-\exp\left(-\frac{\hbar\omega_V}{k_BT}\right)\right\} \quad [14]$$

The value of $(d\sigma/d\Omega)\Gamma_V^{-1}$ can be characterized as a peak intensity in the measured spontaneous Raman scattering spectrum.

When the conversion to the first Stokes wave is large, we must consider both pump depletion and conversion to higher-order Stokes waves (**Figure 1a**). Pump depletion in first Stokes generation can be accounted for by solving the following coupling

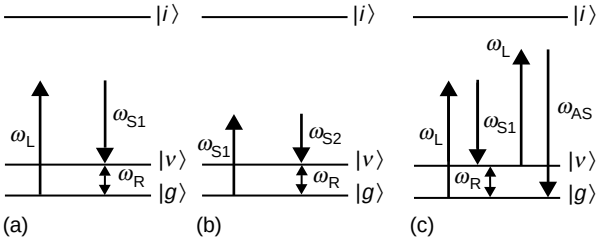


Figure 1 Energy diagrams of stimulated Raman scattering: (a) first Stokes scattering; (b) cascaded second Stokes scattering; and (c) first anti-Stokes scattering.

equations between pump power and Stokes power:

$$\frac{\partial I_L}{\partial z} + \frac{n_L}{c}\frac{\partial I_L}{\partial t} = -g_L I_L I_S \quad [15]$$

$$\frac{\partial I_S}{\partial z} + \frac{n_S}{c}\frac{\partial I_S}{\partial t} = g_R I_L I_S \quad [16]$$

where

$$g_L = \frac{\omega_L}{\omega_S} g_R$$

In SRS the power generated in the first Stokes serves as a pump for a second Stokes wave (Figure 1b). Cascade shifts of this type, in which the Stokes wave of order n serves as a pump for the generation of the Stokes wave order $n+1$, can result in the generation of radiation at several frequencies, each one shifted from the original pump frequency by a multiple of the vibrational mode frequency. In addition to the cascade higher-order Stokes generation, multiple Stokes orders can also be generated by four-wave mixing interactions (Figure 2), of the form:

$$\omega_{S,n+1} = 2\omega_{S,n} - \omega_{S,n-1} \quad [17]$$

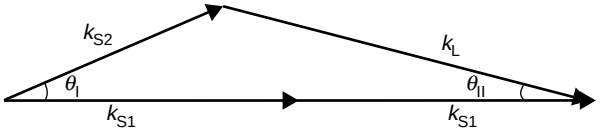


Figure 2 Vector diagram of multiple Stokes generation through four-wave mixing in positively dispersive media.

These interactions are important when the wavevector mismatch among the various waves is small. In most media exhibiting positive dispersion, the phase matching of the various interactions involves off-axis components of the generated waves. Thus, these four-wave mixing interactions are stronger for tightly focused beams than for collimated pump beams.

SRS can be realized in a variety of configurations, including single pass Raman generation, or using a variety of extra-cavity or intra-cavity Raman resonator. Variation of the Raman cross-section for the wavelength is shown in Figure 3 for various gaseous Raman media. The Raman gain coefficient simply increases with decreasing wavelength, as indicted by eqn [13]. Therefore, higher conversion efficiencies are obtainable with UV lasers such as rare-gas halide excimer lasers in a single pass configuration. Note that in some gases, such as H_2, the Raman gain becomes independent of pressure at higher pressures since the increase in linewidth, due to pressure broadening, cancels the increase in gain originated from number density. Typical SRS gains for some gas, liquid, and solid media are listed in Table 1.

Although SRS in high-pressure gases has been widely employed for frequency conversion, SRS in crystals is currently a growing area of research activity. Table 2 lists SRS frequency shifts of various crystals. Since a very high pump laser power ~1 GW/cm^2, which is already close to the laser damage threshold of many crystals, solid-state Raman lasers are

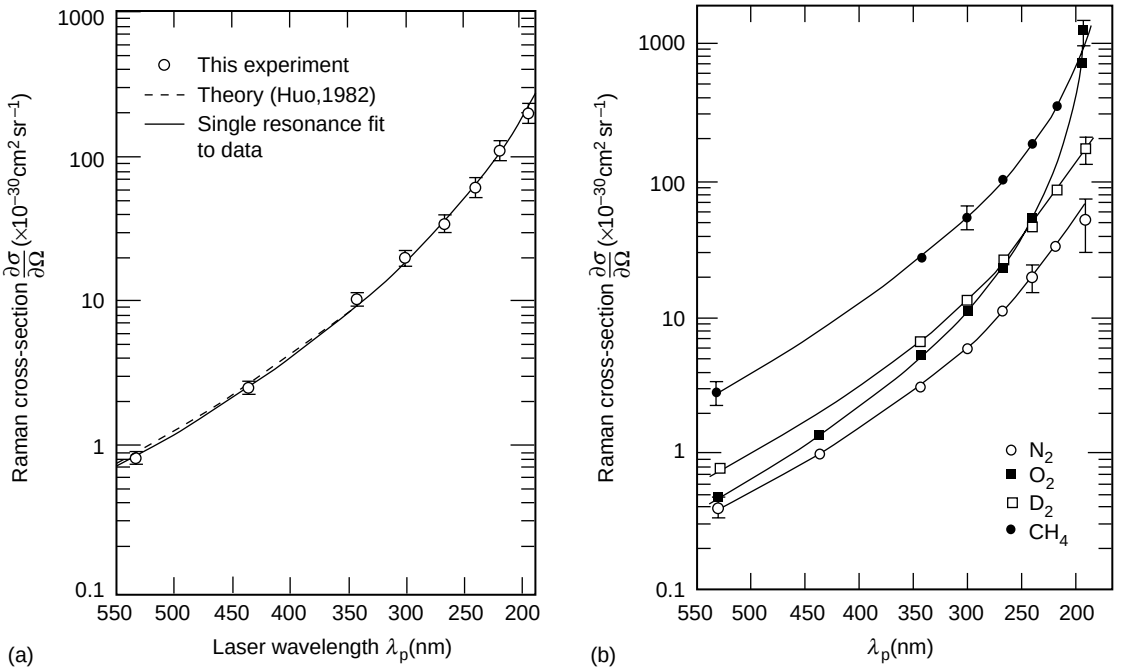


Figure 3 Variation of the Raman cross-section with pump wavelength for various gases. Source: AIP Conference Proceedings No. 100 (Excimer Laser – 1983), 181.

Table 1 Typical Raman media and their parameters

Materials	*Stokes shift (cm^{-1})*	*Raman linewidth (cm^{-1})*	$N\frac{d\sigma}{d\Omega}$ *($10^{-6}\,m^{-1}\,str^{-1}$)*	*Raman gain (10^{-5} m/MW)*
H_2 (100 atm)	4155	0.20	3.0×10^3	1.5
H_2O	3420	176	0.51	0.07
Acetone CH_3COCH_3	2921	18	5.4	1.2
Cyclohexane C_6H_{12}	2852	11	3.7	1.2
Liquid N_2	2326.5	0.067	0.29 ± 0.09	17.5 ± 5
Liquid O_2	1552	0.117	0.48 ± 0.14	14.5 ± 4
Nitrobenzene $C_6H_5NO_2$	1345	6.6	6.4	2.1
Toluene C_7H_8	1003	1.94	1.1	1.2
C_6H_5Cl	1002	1.6	1.5	1.9
C_6H5Br	1000	1.9	1.5	1.5
C_6H_6	992	2.15	3.06	2.8
CS_2	655.6	0.5	7.55	23.8
CC_{l4}	459	5	2.3	0.85
Diamond	1332	2.04	17	6.9
Calcite	1086	1.1	2.9	4.4
Si	521	0.8	305	190
Quartz	467	6.7	3.1	0.8
Li^7TaO_3	201	22	238	4.4
	215	12	167	10
Li^7NbO_3	256	23	381	8.9
	258	7	262	28.7
	637	20	231	9.4
	643	16	231	12.6
InSb ($n_e \sim 2 \times 10^{22}\,m^{-3}$)	100	2	30	16.7

designed with an intra-cavity, coupled cavities, or an extra-cavity. Efficient Raman lasers have been demonstrated using external resonator configuration when the pumping pulse is intense enough. For example, $\sim 40\%$ conversion efficiency to S1 was reported with a Q-switched frequency doubled Nd:YAG laser in a $Ba(NO_3)_2$ crystal using a pump fluence of $1.4\,J\,cm^{-1}$ and a pulse repetition rate of 30 Hz. The resonator mirror characteristics can be designed to enhance the desired Stokes-order radiation by completely confining the lower-order Stokes radiation inside the cavity. For a lower power pump source, such as cw-pumped Q-switch lasers, a Raman crystal is placed inside a pump laser resonator to reduce the effective threshold for SRS. The resonator configuration has a significant influence on the Stokes output performance. The optimum spot size in the resonator and power density in the Raman crystal is a compromise between maximizing the conversion efficiency and avoiding optical damage. It is noteworthy that the thermal lens in the Raman crystal is caused by the energy dissipation of the phonon and thus depends on the power density of the Stokes field.

A fiber Raman laser is an optical fiber with pump light focused into one end and SRS light appearing at the output. By injecting a weak frequency-shifted signal light together with the pump light, the signal light experienced substantial amplification along the length of the fiber. This fiber Raman amplifier has been actively developed to extend the wavelength band for optical fiber communication. The fiber Raman lasers have utilized fused silica (SiO_2), germanosilicate, or phosphosilicate optical fibers for which the Raman frequency shifts are ~ 440–$490\,cm^{-1}$. In contrast to Erbium-doped fiber amplifier (EDFA), since the single mode fiber itself can act as an amplifier, the Raman amplifier can be distributed over optical transmission lines in lengths of >20 km, which can keep the transmission power always nearly constant and consequently reduce the signal noise. Moreover, by preparing multipumping wavelengths, the number of signal wavelengths can increase without the limitation of an amplification gain bandwidth. High-power light sources around 1480 nm developed for EDFA can be utilized by the fiber Raman amplifier. By tailoring the power levels among multiwavelength pump lights, a flat amplification gain spectral width of ~ 100 nm is obtainable.

Transient Effects

In the steady-state SRS analysis, we assume that the pump pulse duration is much longer than the vibrational mode dephasing time ($\Delta\nu_p \ll \Gamma_V$). In a transient regime, when the pump pulse duration is shorter than the dephasing time ($\Delta\nu_p \gg \Gamma_V$), the effect of the transient response of the medium reduces the gain and delays the Stokes pulse relative to the pump pulse. With no pump depletion, the analytical expression for the Stokes intensity for large

Table 2 Typical solid-state Raman materials and parameters

Materials	*Stokes shift (cm^{-1})*	*Raman linewidth (cm^{-1})*	$\frac{d\sigma}{d\Omega}$ *(arb.u.)*
Diamond	1332.9	2.7	100
SO_2	464.5	7.0	2.2
Nitrates and calcite			
$Ba(NO_3)_2$	1048.6	0.4	21
$NaNO_3$	1069.2	1.0	23
$CaCO_3$	1086.4	1.2	6.0
Tungstates			
$CaWO_4$	910.7	4.8	47
$SrWO_4$	921.5	3	–
$BaWO_4$	926.5	2.2	–
$NaY(WO_4)_2$	918	15	–
$KGd(WO_4)_2$	901	5.4	54
$KGd(WO_4)_2$	901	5.4	43
$KGd(WO_4)_2$	768	6.4	19
$KGd(WO_4)_2$	768	6.4	65
$KY(WO_4)_2$	905.6	7	50
$KY(WO_4)_2$	905.6	7	45
$KY(WO_4)_2$	767.4	8.4	20
$KY(WO_4)_2$	767.4	8.4	64
$KYb(WO_4)_2$	908	7.4	48
$KYb(WO_4)_2$	908	7.4	48
$KYb(WO_4)_2$	757	15	25
$KYb(WO_4)_2$	757	15	70
Molybdates			
$CaMoO_4$	879.3	5.0	64
$SrMoO_4$	887.7	2.8	55
$BaMoO_4$	892.4	2.1	52
Iodate and niobates			
$LiIO_3$	821.6	5.0	54
$LiNO_3$	872	21.4	44
$LiNbO_3$	632	27	166
$LiNbO_3$	250	28	–
$LaNbO_4$	805	9	22
Phosphates			
$Ca_2(PO_4)_3F$	964.7	2.8	3.4
$Sr_5(PO_4)_3F$	950.3	2.8	3.4

amplification gain is given as

$$I_S(\ell) \propto I_S(0)\exp\left(-\frac{\tau_p}{T_V}\right) \times \exp\left\{2\left[I_L\tau_p\ell\frac{16\pi^2c^2N_0}{\hbar n_S^2\omega_S^3}\left(\frac{d\sigma}{d\Omega}\right)\right]^{1/2}\right\} \quad [18]$$

Here, T_V is the dephasing time ($T_V = \pi\Gamma_V^{-1}$), and τ_p is the pump pulse width. One can see that the Raman gain is proportional to the square root of the multiplication of the pump pulse energy $I_L \times \tau_p$, the crystal length ℓ, and the total integral Raman scattering cross-section ($d\sigma/d\Omega$), but does not depend on the Raman linewidth Γ_V.

Anti-Stokes Raman Scattering

From eqn [13] the gain coefficient is negative at $\omega = \omega_L + \omega_S$. However, anti-Stokes scattering (Figure 1c) is stimulated through a four-wave mixing interaction of the form:

$$\omega_A = 2\omega_L - \omega_S \quad [19]$$

The equations describing anti-Stokes Raman scattering in the absence of pump depletion are:

$$\frac{d\bar{E}_S}{dz} + \frac{a}{2}\bar{E}_S = i\frac{\omega_S^2}{2c^2\varepsilon_0 k_{Sz}}\bar{P}_S^{NL} \quad [20]$$

$$\frac{dE_A^*}{dz} + \frac{a}{2}\bar{E}_A^* = -i\frac{\omega_A^2}{2c^2\varepsilon_0 k_{Az}}\bar{P}_A^{NL*} \quad [21]$$

$$\bar{P}_S^{NL} = \varepsilon_0\chi^{NL}\left\{|\bar{E}_L|^2\bar{E}_S + \bar{E}_L^2\bar{E}_A^*\exp(i\Delta kz)\right\} \quad [22]$$

$$\bar{P}_A^{NL*} = \varepsilon_0\chi^{NL}\left\{\bar{E}_L^{*2}\bar{E}_S\exp(-i\Delta kz) + |\bar{E}_L|^2\bar{E}_A^*\right\} \quad [23]$$

Here, Δk is the wavevector mismatch, given by

$$\Delta k = k_S + k_A - 2k_L \quad [24]$$

Then, power gain per unit length is obtained to give

$$G \cong -a \pm 2\,\mathrm{Im}\left\{\left(\frac{\Delta k}{2}\right)^2 - \frac{\omega_S}{2cn_S}\chi^{NL}|\bar{E}_L|^2\Delta k\right\}^{1/2} \quad [25]$$

The interaction of the anti-Stokes wave with the other waves is stronger when the wavevector mismatch is small. The exact value of the phase-matching angle depends on the dispersion of the medium. At the tightly focused pumping condition, strong anti-Stokes radiation is generated so that the wavevector mismatch is minimum. Just as in Stokes scattering, multiple anti-Stokes orders can be generated (Figure 4).

So far, Stokes and anti-Stokes Raman scattering have been used to generate radiation over much of the UV and VUV, ranging from 138 to 400 nm. The shortest wavelength achieved in this manner is at 138 nm, generated by a seventh-order anti-Stokes shift in H_2 using 139 nm radiation of an ArF laser. In general, the pump intensity used in experiments with rare-gas halide excimer lasers was in the range of 100 MW/cm^2 to 5 GW/cm^2. The conversion efficiency and the distribution of power among the various Stokes and anti-Stokes

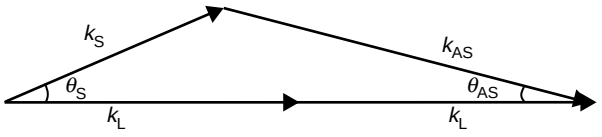


Figure 4 Vector diagram of anti-Stokes Raman scattering in positively dispersive media.

components depend on the focal geometry, and the extent of the spectrum could be varied by changing the gas pressure.

See also

Magneto-Optics: Cyclotron Resonance, Interband Magnetoabsorption, Spin Flip Raman Scattering. **Nonlinear Optics, Applications:** Raman Lasers.

Further Reading

Basiev TT, Sobol AA, Zverey PG, Ivleva LI, Osiko VV and Powell RC (1999) Raman spectroscopy of crystals for stimulated Raman scattering. *Optical Matters* 11: 307.

Bischel WK and Blach G (1983) New lines in the UV: SRS of excimer laser wavelengths. *AIP Conference Proceedings No. 100 (Excimer Lasers-1983)*, 181.

Loree TR, Sze RC, Barker DL and Scott PB (1979) *IEEE Journal of Quantum Electronics* QE-15: 337.

Reinties JF (1985) Coherent ultraviolet and vacuum ultraviolet sources. In: Laser Bass M and Stitch ML (eds) *Laser Handbook*. Amsterdam: North-Holland Physics Publishing.

Weber MJ (ed.) (2001) Raman, Brillouin and Soliton lasers. In: *Handbook of Laser*, p. 291. Boca Raton, FL: CRC Press.

Scattering from Surfaces and Thin Films

A Duparré, Fraunhofer Institut IOF, Jena, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The quality of optical components can be critically limited by light scattering. The majority of these components, such as lenses, mirrors, beamsplitters, and polarizers are covered with thin film optical coatings. To reduce scattering losses, their main sources and mechanisms must be known and, whenever possible, be controlled. Moreover, light scattering techniques have been widely recognized as a powerful tool for roughness and defect analysis as well as for general quality assessment.

Scattering can be divided into surface/interface scattering (SIS) and volume scattering (VS), attributed to the following scatter sources:

- roughness and surface defects of the substrate: SIS,
- surface/interface roughness and defects of the thin film coating: SIS,
- imperfections in the bulk substrate material: VS,
- intrinsic thin film morphology and defect structures: VS,
- surface contaminations: SIS.

Even though in particular cases and applications, VS can constitute the main scattering source, SIS largely dominates both substrate and thin film scattering losses. Studies into light scattering have widely focused on SIS effects from surface and interface micro-roughness. The theoretical outline given in the next section will also follow this concept in describing scattering from surface and interface roughness. It is emphasized, however, that volume scattering theories are accessible in the literature.

Light Scattering Models

Scattering from Rough Surfaces

A randomly rough surface can be considered as a Fourier series of sinusoidal waves with different amplitudes, periods, and phases. Following the grating equation, a single grating with spacing D causes scatter into the angle Θ according to

$$\sin\Theta = \lambda/D \qquad [1]$$

where λ is the wavelength of light. D represents one spatial wavelength in the Fourier series. Accordingly, $f = 1/D$ represents one single spatial frequency. A randomly rough surface contains many different spatial frequencies. This is quantitatively expressed by the power spectral density (PSD), giving the relative strength of each roughness component of a surface microstructure as a function of spatial frequency:

$$\mathrm{PSD}(\mathbf{f}) = \lim_{A\to\infty}\frac{1}{A}\left|\int_{\mathbf{A}}\zeta(\mathbf{r})\exp(-2\pi i\mathbf{f}\cdot\mathbf{r})\mathrm{d}\mathbf{r}\right|^2 \qquad [2]$$

where $\zeta(\mathbf{r})$ represents the height of the surface roughness profile, $\mathbf{r}$ is the position vector, and $\mathbf{f}$ is the spatial frequency vector in the x–y plane. A is the area of the measured region $\mathbf{A}$. We confine our discussion to isotropic surfaces, which represent the majority of cases in optical surface and thin film scattering studies. Thus, a PSD(f) independent of the surface direction Φ of vector $\mathbf{f}$ is obtained by averaging the two-dimensional function PSD($\mathbf{f}$) over all surface directions after transformation into polar coordinates:

$$\mathrm{PSD}(f) = \frac{1}{2\pi}\int_0^{2\pi}\mathrm{PSD}(f,\Phi)\mathrm{d}\Phi \qquad [3]$$

Well-established vector scattering theories developed, for instance, by Bousquet *et al.* or Elson provide the link between the PSD and the scattering intensity per solid angle of a surface:

$$\frac{dP}{P_0\, d\Omega} = F(\lambda, n, \Theta, \Phi)\mathrm{PSD}(f) \qquad [4]$$

This theory is valid for surfaces whose rms roughness is small compared to the wavelength. $dP/(P_0\, d\Omega)$ denotes the differential power scattered into the direction (Θ, Φ) per unit solid angle $d\Omega = \sin\Theta\, d\Theta\, d\Phi$ divided by the incident power P_0. Θ and Φ are the polar and azimuthal angles of scattering, respectively. The optical factor F contains all information on the corresponding perfect surface (without the roughness properties), i.e., the refractive index n, wavelength, and the conditions of illumination and observation. Both backscattering and forward scattering can be expressed by eqn [4], according to the illumination and observation conditions chosen. Without loss of generality, all formulas have been written here for normal incidence. The formalism, however, allows consideration of all possible cases, including oblique angles of incidence and arbitrary polarization properties. $dP/(P_0\, d\Omega)$ is called angle resolved scattering (ARS) which is related to the well-established term BRDF/BTDF (bidirectional reflectance/transmittance distribution function) by multiplication with $\cos\Theta$:

$$\frac{dP}{P_0\, d\Omega} = \mathrm{ARS} = \mathrm{BRDF\ (or\ BTDF)}\cdot\cos\Theta \qquad [5]$$

Total scattering (TS), which is defined as the power P scattered into the backward or forward hemisphere divided by the incident power P_0 (see also section on scattering measurement below), is obtained by integrating eqn [4] over the forward or backward hemisphere:

$$\mathrm{TS} = 2\pi \int_0^{\frac{\pi}{2}} \left(\frac{dP}{P_0\, d\Omega}\right) \sin\Theta\, d\Theta \qquad [6]$$

If the correlation length of surface roughness is much larger than the wavelength, scalar scattering theories like the one from Carniglia can also be employed. Moreover, in this case, the well-known simple approximate formula for total backscattering can be derived from both vector and scalar theories:

$$\mathrm{TS}_{\mathrm{back}} = R_0 \left(\frac{4\pi\sigma}{\lambda}\right)^2 \qquad [7]$$

where σ is the rms roughness and R_0 the specular reflectance.

It must be emphasized that this formula is only valid if the above-mentioned condition is met and as long as only single surfaces without coatings are considered.

Scattering from Dielectric Thin Films

For surfaces coated with a dielectric single layer or multilayer, ARS for a system of N layers is given by

$$\frac{dP}{P_0\, d\Omega} = \sum_{i=0}^{N} \sum_{j=0}^{N} F_i F_j^* \mathrm{PSD}_{ij}(f) \qquad [8]$$

where F_i is the optical factor at the i-th interface and F_j^* is the conjugate complex number of the optical factor at the j-th interface. The optical factors include both the conditions of illumination and observation and the characteristics of the ideal multilayer (refractive indices, film thickness). PSD_{ij} are the power spectral densities of the corresponding interfaces. For $i \neq j$, they describe the roughness cross-correlation between two interfaces.

From eqn [8] it becomes obvious that the scattering mechanism of coated surfaces is considerably more complex as compared to single surfaces.

The amplitudes of the scattered fields from all individual interfaces add up to the total scatter. In particular, the cross-correlation properties between the interface roughness profiles significantly determine the overall total scatter. So even for coatings having the same statistical roughness, the scatter losses can differ drastically when the cross-correlation between the interfaces is different. An example is demonstrated in a model calculation of total backscattering for $\lambda = 632.8$ nm in Figure 1.

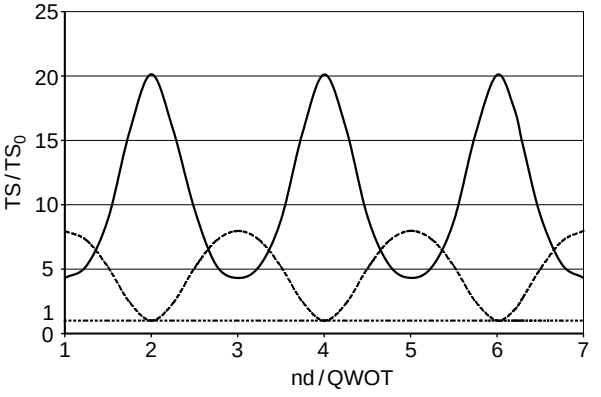


Figure 1 Calculated ratio of the total backscattering of a single high index layer ($n = 2.4$) on BK7 substrate ($n = 1.52$) to the scattering of the bare substrate as a function of optical thickness nd. (QWOT: quarter wave optical thickness). rms roughness was 1 nm, correlation length 1 μm for both the substrate–film and film–air interfaces. Full line: full interface cross-correlation. Dashed line: zero interface cross-correlation. Lower dashed line: bare substrate.

A thin film with refractive index $n = 2.4$ (typical for Titania films), and continuously increasing thickness, was modeled onto a BK 7 substrate ($n = 1.52$). Gaussian autocovariance functions with an rms roughness $\sigma = 1$ nm and a correlation length $\tau = 1$ μm were chosen for both the substrate–film and film–air interfaces. Only the cross-correlation function was varied. The two extreme cases of fully uncorrelated and fully correlated interfaces are depicted in the figure and related to the scattering of the bare substrate (TS_0). As a result of scattering interference effects, in both cases the scattering varies periodically with optical film thickness. The appearance of maxima and minima as well as the total amount of scattering crucially depends on the type of cross-correlation.

These effects have to be taken into careful deliberation whenever scattering in dielectric thin film coatings is considered. Otherwise, misinterpretation of the scatter loss origin is possible and may result in unsuitable technological attempts to minimize scatter losses.

Instrumentation for Light Scattering Measurement

Types of Measurements

Light scattering experiments on optical surfaces and thin films are, in most cases, either performed as ARS measurements using goniophotometers, or the scattering is collected over the backward and forward hemispheres yielding total integrated scattering (TIS) or TS. The latter can be measured either by using a Coblentz sphere or an integrating sphere (see schematic picture in Figure 2).

As Coblentz spheres image the scattered radiation directly onto the detector, they are much less sensitive to scattering by air particles than integrating spheres. So if super-smooth samples with scattering levels in the visible range of 1 ppm and lower shall be measured, integrating spheres require operation in vacuum or He-gas atmosphere whereas Coblentz spheres can still be used under usual clean room conditions.

Angle Resolved Scattering (ARS)

Sophisticated instruments for ARS measurements have been established in a number of laboratories. Most frequently the systems are operated at the He–Ne laser wavelength 632.8 nm, but also at 10.6 μm (CO_2 laser), 325 nm (He–Cd laser), and yet other wavelengths. These techniques are versatile and powerful, but rather unsuitable for routine measurements. For single surfaces, PSD can be calculated from such measurements. A standard procedure for ARS measurements is defined in ASTM standard E 1392 and was successfully verified in various round-robin experiments at different wavelengths. Figure 3 shows the photograph of a typical instrumentation.

The main parts of this setup are the illumination system and a double goniometer. The sample holder and the detector are each mounted on a precision goniometer. The detector head is located on the outer goniometer and can be revolved 360 degrees around the sample with an angular resolution of 0.01°. The performance of a scatterometer depends also considerably on the quality of the illumination system. Laser radiation passes several optical elements, such as a spatial filter, diaphragms, polarizers, and mirrors providing the high beam quality necessary for precision measurement.

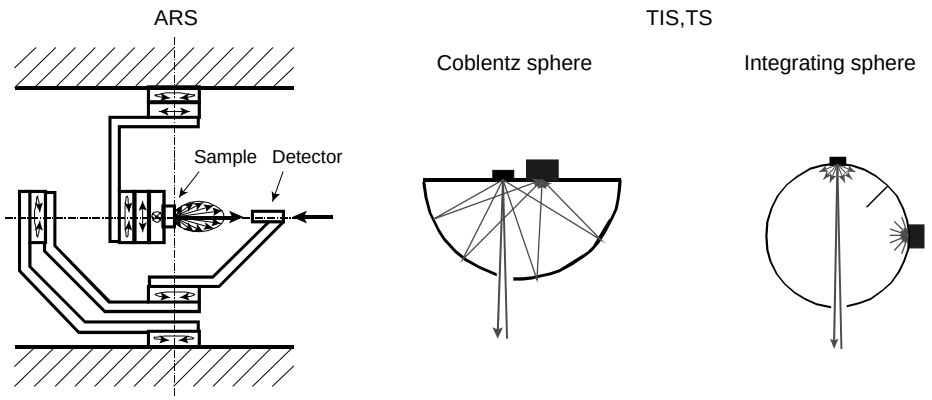


Figure 2 Schematic picture of types of scattering measurement. ARS: angle resolved scattering, TIS, TS: total integrated scattering, total scattering.



Figure 3 Instrumentation for angle resolved scattering measurement (courtesy of Fraunhofer Institute for Applied Optics and Precision Engineering, Jena, Germany).

Total Scattering (TS)

TIS measurements collect the light scattered into the backward hemisphere and are defined as the backscattered radiation divided by the total reflectance. The TIS measurement procedure is prescribed in ASTM standard F 1048 which applies to opaque reflective surfaces. Transparent or semi-transparent samples such as substrates, AR coatings, and beamsplitters cannot be measured without ambiguous additional assumption. The new international standard ISO 13696 defines TS as the backscattered and forward scattered radiation P divided by the incident radiation P_0. TS is hence equivalent to the scattering loss of the component. Opaque as well as transparent surfaces and coatings can be measured both in the backward and forward directions. TS and TIS can, however, be converted into one another, if the reflectance of the sample is known. This standard procedure was proved in an international round-robin experiment at 632.8 nm.

An example of a facility for TS measurements, in a wide range of wavelengths, is given in Figure 4. The apparatus for total backscattering and forward scattering measurements is operated from 193 nm to 10.6 μm by using several lasers. The setup is based on a Coblentz sphere imaging the light scattered into the backward or forward hemisphere within an angular range from 2° to 85° onto the detector unit. This angular range complies with the specification given in ISO 13696.

A special arrangement allows easy change from the backscatter to forward scatter operation modus. The detector unit consists of the detector (photomultiplier in the UV, VIS, NIR, and HgCdTe-element in the IR) and a small integrating sphere. The latter is used for homogeneously illuminating the detector. At all wavelengths, the light beam is modulated by a chopper and passes a beam-cleaning element. The incident radiation hits the sample surface at nearly zero degrees and the specular beam is guided back through the entrance/exit aperture of the Coblentz sphere. While performing the measurement, the sample is scanned across its surface by a positioning system, yielding one- or two-dimensional scattering diagrams. Calibration is performed with a commercial diffusing (Lambertian) standard. Filters are used for attenuating the beam during measurement of high-scatter samples and diffuse reflectance standards. Background levels smaller than 0.1 ppm at 632.8 nm were achieved without the necessity of He-gas flow or operation in vacuum.

Whereas in the past the majority of such facilities were designed for the visible, infrared, and near-ultraviolet spectral regions, the increasing demands for low-scatter optical components in the vacuum ultraviolet (VUV) region for application in photolithography has driven the development of scattering facilities for wavelengths as short as 157 nm. Because of the enhanced technical effort needed for the development of VUV scattering measurement instrumentation, only few setups exist that can be operated at 157 nm.

Applications

Fields of Applications

There exists a diversity of purposes for light-scattering investigations of surfaces and thin films, which requires proper selection of the particular method. For a number of purposes, routine measurements without detailed understanding of the scatter sources can be sufficient, while other problems need deeper insight into the scatter mechanisms and hence, require the combination of measurement with theoretical modeling. It is up to the engineer or scientist to decide in each particular case, to which level the

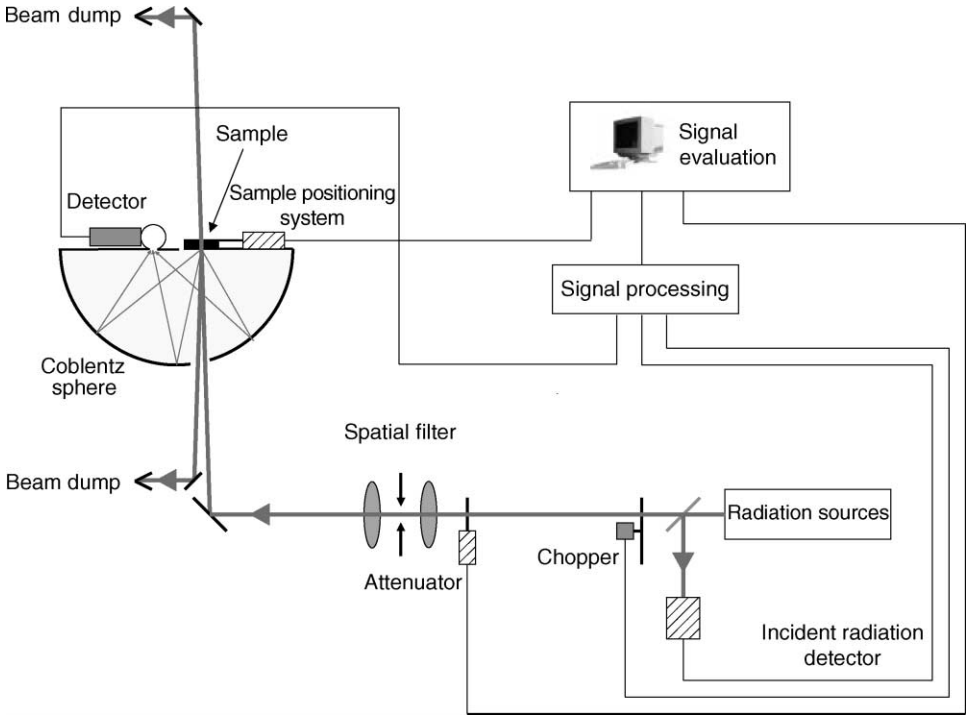


Figure 4 Schematic picture of an instrument for total scattering measurement (courtesy of Fraunhofer Institute for Applied Optics and Precision Engineering, Jena, Germany).

light-scattering study has to be extended. Examples for typical application tasks arising in industry and institutes are:

- determination of total scattering as an optical loss limiting the performance of a thin film component;
- surface finish assessment (roughness, defects) of optical substrates;
- study into the relation between scattering and the morphology and roughness of optical thin films as a precondition for optimizing deposition processes;
- assessment of cleanliness of coated components;
- detection of degradation effects in coatings and substrate materials;
- separation of interface scattering from the film and substrate;
- distinction of volume scattering from interface scattering in substrates and films.

Scattering measurements are noncontact, do not require sample preparation, enable rapid mapping of large sample areas, and are highly sensitive. This makes them usable for a large variety of applications extending from super-smooth substrates with surface roughness below 0.01 nm, thin film structures in the nanometer range, and small and large defect features up to rough engineering surfaces with micron roughness.

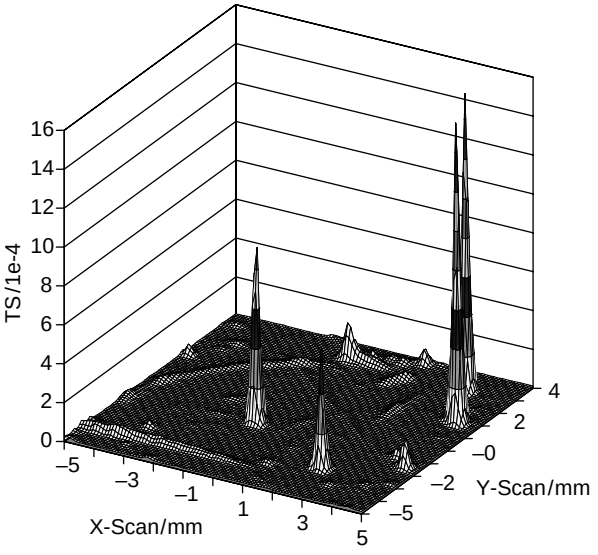


Figure 5 TS measurement (backscattering) on conventionally polished fused silica at 632.8 nm (2D-mapping) for surface quality assessment. Measurement performed with the TS instrument shown in Figure 4.

Examples of Measurements

Surface finish assessment

The result of a scattering measurement (TS, backscattering) at the He–Ne laser wavelength (632.8 nm), on conventionally polished fused silica, is shown in Figure 5. Symmetric defects, as well as

scratch-like features, can be recognized in this diagram. The minimum and averaged rms-roughness determined from the minimum (defect-free) and averaged (including defects) scattering levels are 0.51 and 0.72 nm, respectively.

Scattering from thin film components

Figure 6 shows a TS measurement (backscattering) performed on a polished CaF_2 sample. Half of the area of one surface of this sample was coated with a fluoride multilayer mirror system designed for 248 nm. The area mapping of the TS measurement at 248 nm delivers information of both the bare substrate quality and the increase in scattering after coating. According to the explanations in the first section, this increase has to be interpreted as a result of two influences: the increased reflectance (i.e., changed optical factors) and the roughness growth during deposition of the multilayer (i.e., changed PSD).

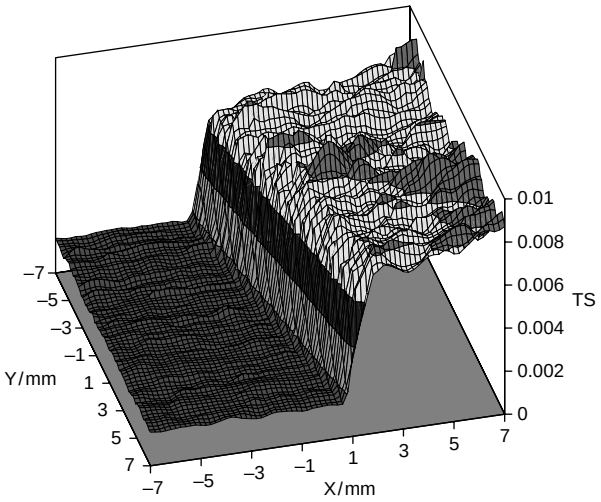


Figure 6 TS measurement (backscattering) at 248 nm on a CaF_2 substrate half of which was coated with a fluoride multilayer mirror system designed for 248 nm. Measurement performed with the TS instrument shown in Figure 4.

On the basis of such measurements, detailed interpretation can be accomplished by additional roughness analysis using scanning force microscopy, subsequent scattering modeling employing the formulas given above, and comparison of the modeled and measured results.

ARS techniques can, for example, be employed to compare the angular distribution of the scattering of a coating to that of the bare substrate. Figure 7 displays ARS measurements in the backward hemisphere at 325 nm (He–Cd laser) on a fused silica substrate before and after deposition (magnetron sputtering) of a Ta_2O_5/SiO_2 quarterwave multilayer mirror (325 nm design wavelength). As with the example discussed above, the higher scatter of the multilayer is caused by both the optical factors and increased roughness. The slight bump at large scatter angles in particular reveals high spatial frequency roughness attributed to the film morphology.

Figure 8 displays TS measurements at 157 nm. The diagram contains one-dimensional scans of forward scattering from an uncoated CaF_2 substrate and an antireflective (AR) coating on CaF_2. The background signal level is also included.

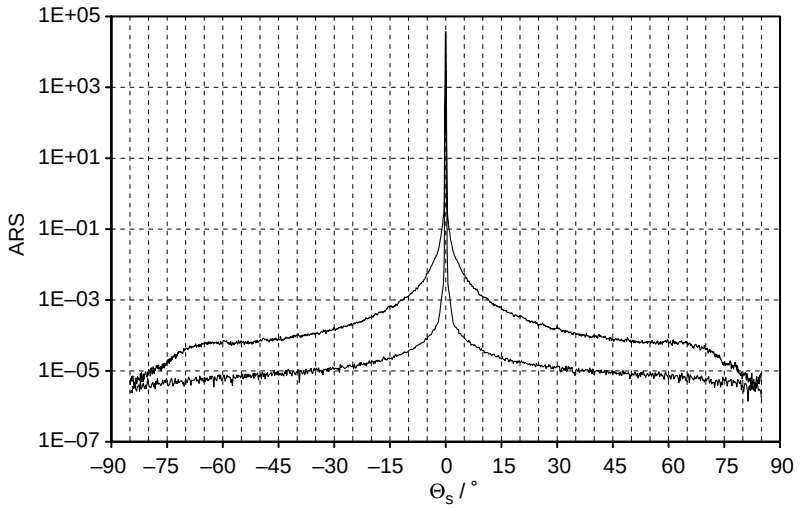


Figure 7 ARS measurement at 325 nm (backward hemisphere) on a fused silica substrate before (lower curve) and after (upper curve) deposition of a Ta_2O_5/SiO_2 quarterwave multilayer mirror with 325 nm design wavelength. Measurements performed with the ARS instrument shown in Figure 3.

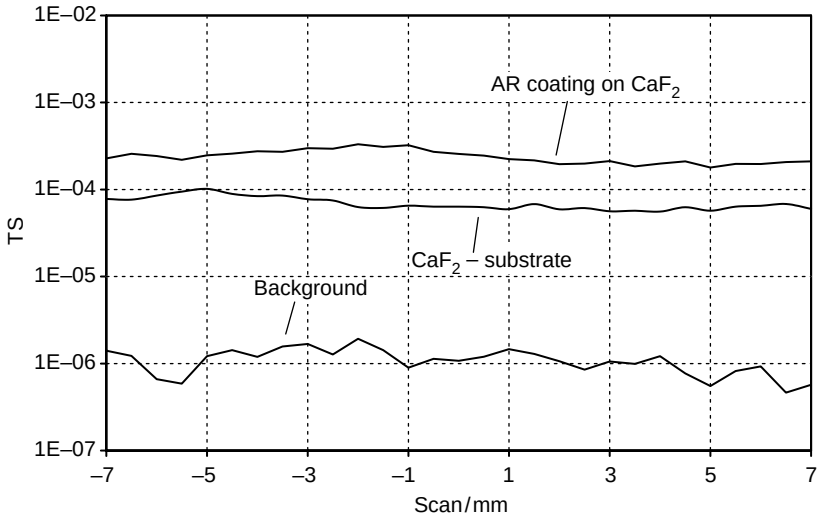


Figure 8 TS measurements at 157 nm (forward scattering): bare CaF_2 substrate, AR coating on CaF_2 and background scattering level. Measurements performed with the VUV total scattering instrumentation at the Fraunhofer Institute for Applied Optics and Precision Engineering, Jena, Germany.

List of Units and Nomenclature

Angle resolved scattering	[nondimensional]	ARS
Azimuthal scattering angle	[degree]	Φ
Correlation length	[nm]	τ
Differential scattered power	[nondimensional]	$dP/(P_0\, d\Omega)$
Grating spacing	[μm]	D
Height of surface roughness profile	[nm]	ζ
Incident power	[W]	P_0
Optical factor	[nm^{-4}]	F
Polar scattering angle	[degree]	Θ
Power spectral density	[nm^4]	PSD
Refractive index	[nondimensional]	n
Rms roughness	[nm]	σ
Spatial frequency	[μm^{-1}]	f
Specular reflectance	[nondimensional]	R_0
Total integrated scattering	[nondimensional]	TIS
Total scattering	[nondimensional]	TS
Wavelength	[nm]	λ

See also

Optical Coatings: Thin-Film Optical Coatings. **Semiconductor Physics:** Light Scattering.

Further Reading

Baeumer S, Duparré A, Herrmann T, *et al.* (1999) SLIOS – a contribution to standard procedures in stray light measurement. *Proceedings SPIE* 3739: 414–421.

Bennett JM and Mattsson L (1999) *Introduction to Surface Roughness and Scattering*, 2nd edn., pp. 28–35, 62–70. Washington, DC: Optical Society of America.

Bousquet P, Flory F and Roche P (1981) Scattering from multilayer thin films: theory and experiment. *Journal of Optical Society of America* 71: 1115–1123.

Carniglia CK (1979) Scalar scattering theory for multilayer optical coatings. *Optical Engineering* 18: 104–115.

Church L, Jenkinson HA and Zavada JM (1979) Relationship between surface scattering and microtopographic features. *Optical Engineering* 18: 125–136.

Duparré A (1995) Light scattering of thin dielectric films. In: Hummel RE and Günther KH (eds) *Thin Films for Optical Coatings*, pp. 273–303. Boca Raton, FL: CRC Press.

Elson JM (1984) Theory of light scattering from a rough surface with an inhomogeneous dielectric permittivity. *Physical Review B* 30: 5460–5480.

Elson JM and Bennett JM (1979) Vector scattering theory. *Optical Engineering* 18: 116–124.

Herman IP (1996) *Optical Diagnostics for Thin Film Processing*, pp. 522–529. London: Academic Press.

Kadkhoda P, Müller A, Ristau D, *et al.* (2000) International round-robin experiment to test the international organization for standardization total-scattering draft standard. *Applied Optics* 39: 3321–3332.

Kassam S, Duparré A, Hehl K, Bussemer P and Neubert J (1992) Light scattering from the volume of optical thin films: theory and experiment. *Applied Optics* 31: 1304–1313.

Leonhard TA and Rudolph P (1995) BRDF round robin test of ASTME 1392. *Proceedings SPIE* 1995: 285–293.

Macleod HA (2001) *Thin-Film Optical Filters*, 3rd edn., pp. 12–52. Bristol, UK and Philadelphia: Institute of Physics Publishing.

Rönnow D and Veszelei E (1994) Design review of an instrument for spectroscopic total integrated light scattering measurements in the visible wavelength region. *Review of Scientific Instruments* 65: 327–334.

Stover JC (1995) *Optical Scattering, Measurement and Analysis*, 2nd edn., pp. 133–176. Bellingham, CA: SPIE Press.

Van de Hulst HC (1957) *Light Scattering by Small Particles*, pp. 28–31, 63–84. London: Chapman and Hall Ltd.

Whitehouse DJ (1994) *Handbook of Surface Metrology*, pp. 49–60, 512–517. Bristol, UK and Philadelphia: Institute of Physics Publishing.

Scattering Phenomena in Optical Fibers

P Tchofo-Dinda and G Millot, Université de Bourgogne, Dijon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Nonlinear phenomena that result from interaction of intense light beams with dielectric media can be broadly divided into two main categories: namely, parametric effects and scattering phenomena. Parametric effects arise whenever the state of the matter is left unchanged by the interaction, whereas scattering phenomena imply transitions between vibrational or rotational energy levels in the medium. The two dominant scattering phenomena in optical fibers are stimulated Raman scattering (SRS) and stimulated Brillouin scattering (SBS), which are nonlinear processes in which a part of the energy propagating at a given frequency is converted into one or several down-shifted Stokes beams, through interaction with the vibrational modes of the materials. Although SBS and SRS phenomena are beneficial for developing all-optical fiber lasers and amplifiers for long-haul fiber transmissions, they are detrimental to transmission of ultrashort light pulses in silica fibers. In this article we describe the fundamental aspects of SBS and SRS.

General Features

Three well-known kinds of scattering phenomena in optical fibers are Rayleigh, Raman and Brillouin scattering. Although these three phenomena manifest themselves in qualitatively different ways, all of them originate fundamentally from the effects of a light wave on the atomic and molecular charges of the dielectric medium. The Raman and Brillouin scattering phenomena involve the vibrational modes of the material that are associated with the optical and acoustic branches of the dispersion curve of the material, respectively. Such vibrational modes, called phonons, are schematically represented in Figure 1 with the help of typical dispersion curves of a one-dimensional atomic lattice.

In standard silica fibers (SiO_2), the phonons associated with the optical branch correspond to intramolecular vibrations, which give rise to the Raman scattering phenomenon, whereas those associated with the acoustic branch correspond to intermolecular vibrations, which are responsible for Brillouin scattering. In either of these two scattering processes, part of the energy of the incoming light, which propagates at a given frequency ω, is converted into downshifted (Stokes) or upshifted (anti-Stokes) light waves, at frequencies $\omega - \Omega$ and $\omega + \Omega$, respectively. The frequency shift Ω is determined by the vibrational modes involved in the scattering process. Figure 2 represents schematically

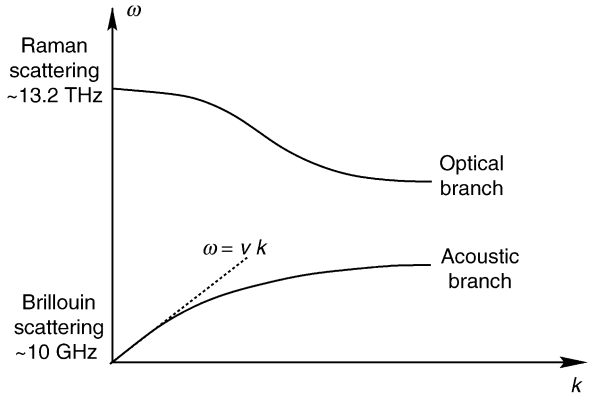


Figure 1 Schematic representation of the dispersion curves for material waves.

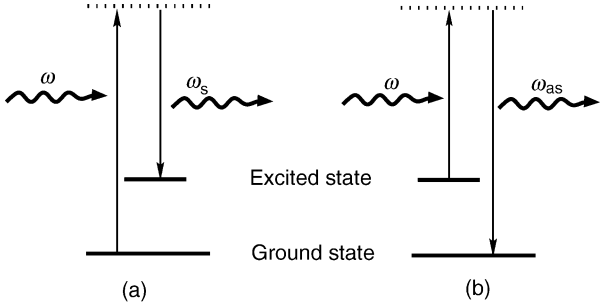


Figure 2 Schematic representation of the diagrams of energy transfer between incoming photons with frequency ω and the material, leading to (a) generation of the Stokes radiation ω_s, and (b) generation of the anti-Stokes radiation ω_{as}.

these energy conversion processes with the help of only two vibrational states, to simplify matters.

As Figure 2a shows, the material, which is initially in the ground state, absorbs a fraction $\hbar\Omega$ of the energy $\hbar\omega$ of incoming photons, to move up to the excited state. As a result, the frequency of the photons involved in this process is downshifted to $\omega_s = \omega - \Omega$. Those photons that are generated at frequency ω_s are commonly called 'Stokes radiation'. On the other hand, when a sufficiently large number of molecules is excited, they can interact with the incoming photons and return back to the ground state. As Figure 2b shows, in this process the energy is transferred from the material to the incoming photons, which thereby generates an upshifted anti-Stokes radiation at frequency $\omega_{as} = \omega + \Omega$. The Raman (Brillouin) frequency shift Ω is of the order of $\Omega_R \approx 13\,\text{THz}$ ($\Omega_B \approx 10\,\text{GHz}$).

In fact, the Raman and Brillouin effects may lead to so-called spontaneous processes, which occur in the limit of low-amplitude for all waves other than the incoming wave. For example, when a laser beam is injected in the fiber with a sufficiently low intensity, one can observe a spontaneous scattering process as schematically represented in Figure 3a.

One can clearly identify in Figure 3a the Raman and Brillouin radiation, as well as radiation scattered at the same frequency as the incoming wave, that is, Rayleigh radiation. Now, when the incoming wave propagates together with light waves whose intensity is no longer negligible with respect to that of the incoming wave, the Raman and Brillouin effects can lead to stimulated processes, in which energy is continually converted from higher to lower frequency photons. As a result, the number of Stokes photons is continually amplified at the expense of anti-Stokes photons, as schematically

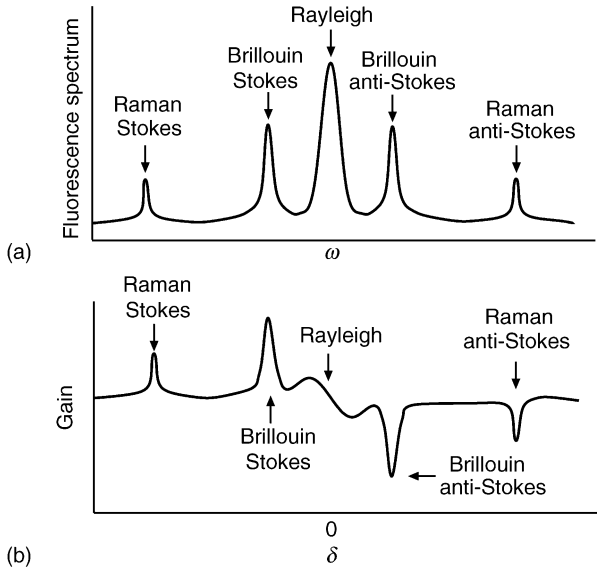


Figure 3 (a) Schematic representation of the frequency spectra for spontaneous and stimulated Raman and Brillouin scattering processes. (a) Fluorescence spectrum; (b) gain spectrum for stimulated Raman and Brillouin scattering processes. A pump and a probe waves are injected together in the fiber. δ is the frequency spacing between the pump and the probe wave.

represented in Figure 3b. The stimulated process can be obtained by either launching the incoming wave ω together with a probe wave at frequency $\omega - \Omega$ (which corresponds to the situation represented in Figure 3b), or by using a sufficiently intense incoming light beam. Indeed, above a certain power threshold (which depends strongly on the fiber length and the amount of dopants in the fiber), for which the gain corresponding to the Stokes amplification compensates for the linear fiber losses, the energy conversion from the incoming wave to the scattered waves can become sufficiently important for the scattering process to become stimulated. This process is commonly referred to as a self-stimulated process, as it does not require the use of a probe wave. Although the SRS and SBS exhibit some common general features as those mentioned above, there exist however some major differences between these two phenomena, from a fundamental point of view as well as for their practical applications to optical communication systems. We discuss separately these specific features for SRS and SBS.

Stimulated Raman Scattering

Under adiabatic following by the electrons of the optical-field-induced nuclear motions, the third-order nonlinear polarization in optical fibers may be

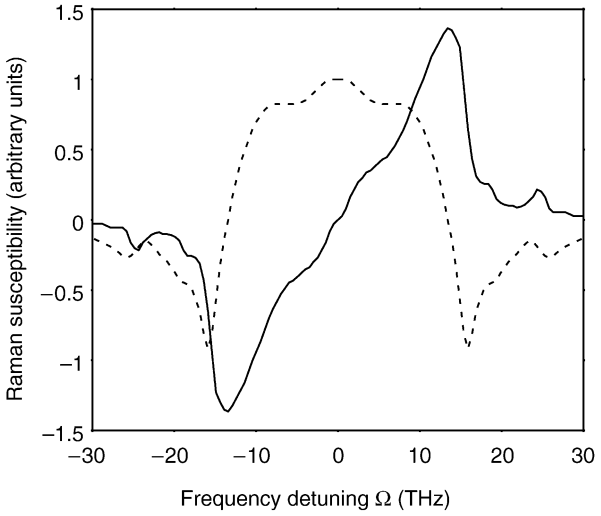


Figure 4 Schematic representation of the Raman susceptibility in an optical fiber. The solid curve represents the imaginary part of the susceptibility, that is, the Raman gain. The dashed curve represents the real part of the susceptibility.

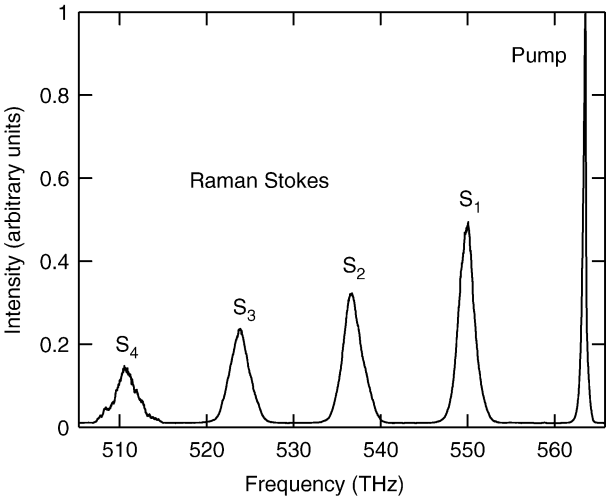


Figure 5 Self-stimulated Raman spectrum (i.e., amplified spontaneous noise), as recorded for a peak pump power $P_0 =$ 150 W and a fiber length $L = 10$ m.

written as

$$P(t) = \sigma|E(t)|^2E(t) + E(t)\int_{-\infty}^{t} \mathrm{d}s\sigma_{\mathrm{R}}(t-s)|E(s)|^2 \quad [1]$$

where σ and σ_{R} represent the instantaneous (Kerr) and noninstantaneous (Raman) parts of the fiber response, respectively. Here, we consider linearly polarized waves, for simplicity. Figure 4 shows schematically the Raman susceptibility of the fused silica, $\chi_{\mathrm{R}}(\Omega) = \sigma_{\mathrm{R}}(\Omega)/2$, which is quite close to that of standard silica fibers.

In Figure 4, the real and imaginary parts of χ_{R} are represented in solid and dashed curves, respectively. The imaginary part of the Raman susceptibility, which corresponds to the Raman gain $G_{\mathrm{R}} = \mathrm{Im}(\chi_{\mathrm{R}})$, is an experimentally measurable quantity. For example, at a pump wavelength $\lambda = 1\mu$m, the maximum value of the Raman gain in standard silica fibers is $\approx 10^{-13}\,\mathrm{mW}^{-1}$. In fact the Raman gain $G_{\mathrm{R}}(\Omega)$ depends on the type and amount of dopants used in the fabrication of the fiber. In contrast to crystalline silica where the Raman gain occurs at a well-defined frequency of 13.2 THz, the Raman gain of silica fibers extends over a large frequency range, as can be seen in Figure 4, with a broad prominent peak around 13.2 THz. In the absence of an input signal the broadband noise is amplified within the Raman Stokes components, which are centered about the frequency detuning of 13.2 THz.

If the power of the Stokes wave becomes large enough, the Stokes wave serves as a pump to generate a second-order Stokes wave. This process of cascade SRS can generate multiple Stokes waves whose number depends on the input pump power and fiber length. Figure 5 displays a typical experimental Raman spectrum generated from a Q-switched, frequency-doubled Nd:YAG laser, emitting 5 ns duration (30 GHz spectral width) pulses at 563.63 THz (532.26 nm). The laser output was injected along one birefringence axis of a 14 m length of polarization-maintaining single-mode fiber. At an estimated injected peak power of 200 W, several higher-order Raman Stokes orders S_j $(j = 1, 4)$ (separated by 13.2 THz) are generated between 510 and 550 THz (588 and 545 nm). The spectral profile of each resulting Stokes line is near-Lorentzian with a full width at half maximum ranging from about 2 THz for the first Stokes to about 3.5 THz for the fourth Stokes. The broadening of the Raman Stokes wave, as the Stokes order increases, is due to several competing nonlinear processes.

This broadband gain is due to the amorphous nature of fused silica, in which the molecular vibrational frequencies spread out into bands which overlap and create a continuum. This outstanding feature has been fully exploited to develop wide-band amplifiers for multichannel optical communications. Recent work demonstrated the achievement of a Raman amplifier with a bandwidth of 100 nm, by use of a set of height pumps and an appropriate choice of the frequency separations between the pumps. Another important application of the Raman gain in optical fibers lies in the development of fiber-Raman lasers, which can be

tuned over a wide frequency range (≈10THz). Such lasers can be obtained by inserting a piece of single-mode fiber inside a Fabry–Perot cavity formed by two partially reflecting mirrors. A rotating output mirror and an intracavity prism allow one to select the Stokes light that provides the laser wavelength. Another example is that the Raman gain can be used to assist non-phase-matched parametric interactions.

A large power-gain enhancement was recently demonstrated for nonphase-matched waves in a three-wave mixing interaction. The Raman-assisted three-wave mixing can be used to achieve wide-bandwidth frequency conversion processes without having to satisfy the strict conditions imposed by phase-matching conditions. The above discussion clearly illustrates some useful effects of the SRS in optical fibers. On the other hand, the SRS may represent a drawback for some optical communication systems, as we discuss below.

Let us consider a pump wave injected in the fiber at frequency ω_0. The total electric field in the fiber may be written as follows:

$$E = \frac{1}{2}\mathcal{A}(z,t)\exp[\mathrm{i}(k_0 z - \omega_0 t)] \qquad [2]$$

where $\mathcal{A}$ is a slowly varying field envelope. By expanding the wavevector k_0 of the electric field in a Taylor series around ω_0, the amplitude $\mathcal{A}$ of the electric field is found to satisfy the generalized nonlinear Schrödinger equation (NLSE):

$$\begin{aligned}&\frac{\partial\mathcal{A}}{\partial z}+\frac{\mathrm{i}}{2}\beta\frac{\partial^2\mathcal{A}}{\partial t^2}\\&=\mathrm{i}\gamma(1-\rho)|\mathcal{A}|^2\mathcal{A}+\mathrm{i}\gamma\rho\mathcal{A}\int_0^\infty \chi_R(s)|\mathcal{A}|^2(t-s)\mathrm{d}s\\&=\mathrm{i}\gamma|\mathcal{A}|^2\mathcal{A}+\mathrm{i}\gamma\rho C_1\mathcal{A}\frac{\partial|\mathcal{A}|^2}{\partial t}+\frac{\mathrm{i}}{2}\gamma\rho C_2\mathcal{A}\frac{\partial^2|\mathcal{A}|^2}{\partial t^2}\\&\quad+\frac{\mathrm{i}}{6}\gamma\rho C_3\mathcal{A}\frac{\partial^3|\mathcal{A}|^2}{\partial t^3}+\cdots\end{aligned} \qquad [3]$$

In eqn [3], $C_1\approx -7.9\times10^{-3}$ ps, $C_2\approx -1.28\times10^{-4}$ ps^2, $C_3\approx 9\times10^{-6}$ ps^3, and the parameter ρ, which measures the fractional contribution of Raman effects to the total nonlinearity, is ≈ 0.18 for standard silica fibers.

When the wave propagation involves only a relatively small frequency range around the pump frequency (less that 1 THz) then only the term proportional to C_1 in eqn [3], i.e., the leading term of the Raman contribution, plays a significant role. Most high-capacity transmission systems using a single channel fall into this situation, as long as picosecond pulses are used. The term proportional to C_1 induces a continuous frequency shift of the pulse frequency, as a result of a small transfer of energy from higher- to lower-frequency waves. This phenomenon, known as the soliton self-frequency shift, constitutes one of the undesirable effects of SRS in optical communications.

On the other hand, when the wave propagation involves a relatively large frequency region (of the order of a few THz), then higher-order terms that are proportional to C_2 and C_3 begin to play a significant role. In particular, the term proportional to C_3 indicates a strong transfer of energy from higher- to lower-frequency waves. In optical communications, this effect causes higher-wavelength channels of wavelength-division multiplexing systems to be amplified while depleting the lower-wavelength channels. This effect can be reduced by using filters and more amplification to compensate for the depletion at lower wavelengths (but this is achieved at the expense of a degradation of the signal-to-noise ratio).

Stimulated Brillouin Scattering

In optical fibers, the SBS occurs at input power levels that are much lower that those needed for SRS, and manifests itself through the generation of a backward-propagating Stokes wave. Although SBS is not (strictly speaking) a parametric process, the generation of the Brillouin Stokes radiation is commonly described as a parametric interaction between the pump wave, the Stokes wave and an acoustic wave (material wave). Indeed, one assumes that the annihilation of a pump photon creates simultaneously a Stokes photon and an acoustic phonon, with the following conservation laws for the energy and the momentum:

$$\Omega_\mathrm{B} = \omega_0 - \omega_\mathrm{s} \qquad [4]$$

$$K_\mathrm{B} = k_0 - k_\mathrm{s} \qquad [5]$$

where ω_0 and ω_s are the frequencies, and k_0 and k_s are the wavevectors of the incoming and Stokes waves, respectively. Then, assuming $|k_\mathrm{s}| \approx |k_0|$, the acoustic wave is found to satisfy the following dispersion relation

$$\Omega_\mathrm{B} = K_\mathrm{B}v = 2v|k_0|\sin(\theta/2) \qquad [6]$$

where v is the phase velocity of the acoustic wave, and θ is the angle between the pump and Stokes wavevectors. As eqn [6] shows, the frequency shift

Ω_B depends on the scattering angle. Ω_B is a maximum for $\theta = \pi$ (i.e., in the backward direction) and vanishes for $\theta = 0$ (forward direction). Thus, although eqn [6] predicts that Brillouin scattering should not occur in the forward direction, spontaneous Brillouin scattering can occur in the forward direction in optical fibers. But this phenomenon is quite negligible. In optical fibers, SBS occurs mainly in the backward direction with a frequency shift given by

$$\nu_B = \Omega_B/(2\pi) = 2v|k_0| = 2n_0 v/\lambda_0 \qquad [7]$$

where n_0 is the refractive index and λ_0 is the pump wavelength (incoming wave). In silica fibers, $v \approx 6\,\mathrm{km\,s^{-1}}$ and $n_0 \approx 1.45$; which leads to $\nu_B \approx 10\,\mathrm{GHz}$ at $\lambda_0 = 1.55\,\mu\mathrm{m}$. Thus, the Brillouin frequency shift is smaller by three orders of magnitude compared with the Raman frequency shift ($\approx$13 THz). The amorphous nature of fused silica leads to a Brillouin-gain coefficient $G_B(\nu)$ which extends over a frequency range of $\approx 10\,\mathrm{MHz}$, with a peak value at $\nu = \nu_B$. The Brillouin gain can be approximated by a Lorentzian profile given by

$$G_B(\nu) = \frac{(\Delta\nu_B/2)^2}{(\nu - \nu_B)^2 + (\Delta\nu_B/2)^2} G_B(\nu_B), \qquad [8]$$

$$G_B(\nu_B) = \frac{2\pi n_0^7 p_{12}^2}{c d_0 \lambda_0^2 v \Delta\nu_B}$$

where $\Delta\nu_B$ is the full width at half maximum, p_{12} is the elasto-optic coefficient and d_0 is the material density.

The Brillouin gain has been exploited to develop fiber-Brillouin lasers. Such lasers are obtained by inserting a piece of fiber inside a (ring or Fabry–Perot) cavity. Another useful application of the Brillouin gain of an optical fiber lies in the amplification of a weak signal. Fiber-Brillouin amplifiers allow one to amplify signals with a frequency shift that corresponds to the Brillouin frequency shift ν_B. Fiber-Brillouin amplifiers can provide up to 40 dB gain at pump powers of a few milliwatts. But the bandwidth of such amplifiers is relatively narrow.

Conclusion

In fact scattering phenomena can be beneficial as well as detrimental for optical communication systems. If signals are transmitted in a communication channel with a power above the threshold of a given scattering phenomenon, this phenomenon will induce a transfer of energy from the signal to the Stokes radiation. The resulting signal depletion would then require more amplification. In multichannel communications, the SBS (SRS) can induce cross-talk between two bidirectional channels if their frequency separation happens to be close to the Brillouin (Raman) shift, $\approx$10 GHz ($\approx$13 THz). However, those undesirable effects can be avoided by in-line control techniques for pulse transmission in fiber-optics communication systems. On the other hand, the Raman and Brillouin effects are useful for generating light at new frequencies as well as for amplifying weak light signals. In particular Raman amplifiers are one of the key devices for high-speed telecommunication systems.

List of Units and Nomenclature

Group velocity dispersion coefficient β_2 [$\mathrm{ps^2\,km^{-1}} = 10^{-27}\mathrm{s^2\,m^{-1}}$]
Nonlinear parameter γ [$\mathrm{W^{-1}\,m^{-1}}$]
NLS (E): nonlinear Schrödinger (equation)
Optical intensity I [$\mathrm{GW\,cm^{-2}} = 10^9\,\mathrm{W\,cm^{-2}}$]
SBS: stimulated Brillouin scattering
SRS: stimulated Raman scattering
WDM: wavelength division multiplexing

See also

Fiber and Guided Wave Optics: Light Propagation; Nonlinear Effects (Basics). **Optical Amplifiers:** Raman, Brillouin and Parametric Amplifiers. **Solitons:** Soliton Communication Systems.

Further Reading

Agrawal GP (2001) Nonlinear Fiber Optics. San Diego, CA: Academic Press.

Agrawal GP (2001) Applications of Nonlinear Fiber Optics. San Diego, CA: Academic Press.

Bloembergen N (1977) Nonlinear Optics. Reading, MA: Benjamin.

Dianov EM, Mamyshev PV, Prokhorov AM and Serkin VN (1989) Nonlinear Effects in Optical Fibres. Switzerland: Harwood Academic Publishers.

Hellwarth R (1977) Third-order optical susceptibilities of liquids and solids. *Progress in Quantum Electronics* 5: 1–68.

Pocholle JP, Papuchon M, Raffy J and Desurvire E (1990) Non linearities and optical amplification in single mode fibres. *Revue Technique Thomson-CSF* 22: 187–268.

Shen YR (1984) Principles of Nonlinear Optics. New York: Wiley.

Scattering Theory

Y A Eremin, Moscow State University, Moscow, Russia

© 2005 Published by Elsevier Ltd.

Lord Rayleigh (John William Strutt) (1842–1916)

In the nineteenth century, Lord Rayleigh offered the first explanation for the sky's blue color (on a clear day!). From Rayleigh's initial theory, representing a simplification of Maxwell's electromagnetic theory for the case of particles very small compared to the wavelength of the incident light, the theory was extended to describe the scattering of light by larger macromolecules in solution.

Scattering theory has played a central role in twentieth-century mathematical physics. From Rayleigh's explanation of why the sky is blue to modern medical application of computerized tomography, has attracted both scientists and mathematicians for over a hundred years. Scattering theory is concerned with the effect an inhomogeneous medium has on an incident wave. In particular, if the total field is viewed as sum of an incident field and a scattered field, then the direct scattering problem is to determine the scattered field from a knowledge of the incident one, obstacle properties, and the equations governing the wave motion. In contrast, the inverse scattering problem consists of reconstruction of the inhomogeneity from the knowledge of asymptotic behavior of the scattered field. An everyday example is human vision: from the measurements of scattered light that reaches our retinas, our brain reconstructs a detailed three-dimensional map of the world around us. In summary, the task of direct scattering theory is to determine the relation between the incident and scattered waves. The task of inverse scattering theory is to determine properties of the obstacle.

Light propagating can be described in terms of an electromagnetic field, which is composed of two related vector fields, the electric and the magnetic fields. This means that the vectors of electric ($\mathbf{E}$) and magnetic ($\mathbf{H}$) field have values defined at each point in space and time. The vector fields $\mathbf{E}$ and $\mathbf{H}$ interact through the material parameters: electric permittivity ε and magnetic permeability μ. The behavior of electromagnetic fields can be described by Maxwell's equations.

Maxwell Equations. James Clerk Maxwell (1831–1879)

Consider the electromagnetic wave (light) propagation in a homogeneous, isotropic medium with electric permittivity $\varepsilon > 0$, electric conductivity $\sigma \geq 0$, and magnetic permeability $\mu > 0$, containing a local obstacle. We assume that $\sigma = 0$ outside an obstacle. If J is a current density, the electric field $\vec{E}(x,t)$ and magnetic field $\vec{H}(x,t)$ satisfy the Maxwell equations:

$$\begin{aligned} \nabla \times \vec{E}(x,t) + \mu \frac{\partial \vec{H}}{\partial t}(x,t) &= 0 \\ \nabla \times \vec{H}(x,t) - \varepsilon \frac{\partial \vec{E}}{\partial t}(x,t) &= \vec{J}(x,t) \end{aligned} \qquad [1]$$

Also, in an isotropic conductor, the electric field satisfies Ohm's law:

$$\sigma \vec{E}(x,t) = \vec{J}(x,t) \qquad [2]$$

The speed of wave propagation for the electric and magnetic fields is $c = 1/\sqrt{\varepsilon\mu}$ and in free space $c = 1/\sqrt{\varepsilon_0\mu_0} = 2.99792458 \times 10^8$ m s^{-1}. Assume that the electromagnetic field is time-harmonic, i.e., of the form:

$$\begin{aligned} \vec{E}(x,t) &= \frac{1}{\sqrt{\varepsilon_0}} \mathbf{E}(x) \mathrm{e}^{-\mathrm{j}\omega t} \\ \vec{H}(x,t) &= \frac{1}{\sqrt{\mu_0}} \mathbf{H}(x) \mathrm{e}^{-\mathrm{j}\omega t} \end{aligned} \qquad [3]$$

where $\omega > 0$ is the frequency. Many phenomenon of light scattering can be described based on monochromatic wave propagation. Let us consider a local obstacle occupying a bounded domain D. Inside the obstacle we introduce relative complex permittivity:

$$\varepsilon = \frac{1}{\varepsilon_0}\left(\varepsilon + \mathrm{j}\frac{\sigma}{\omega}\right) \qquad [4]$$

Then $\mathbf{E}(x)$ and $\mathbf{H}(x)$ obey the time-harmonic Maxwell equations:

$$\nabla \times \mathbf{E}(x) = \mathrm{j}k\mathbf{H}(x), \quad \nabla \times \mathbf{H}(x) = -\mathrm{j}k\varepsilon\mathbf{E}(x) \qquad [5]$$

where $k^2 = \varepsilon_0\mu_0\omega^2$.

Helmholtz Equation. Herman Ludwig von Helmholtz (1821–1894)

A wave equation for vector $\mathbf{E}$ can be readily derived from the Maxwell equations. We obtain:

$$(\nabla^2 + k^2)\mathbf{E}(x) = 0 \qquad [6]$$

This is known as the homogeneous Helmholtz equation. A general planewave solution of the equation can be written as

$$\mathbf{E}(x) = \mathbf{E}_0 \, \mathrm{e}^{\mathrm{j}\vec{k}\cdot\vec{x}} \qquad [7]$$

here $\mathbf{E}_0$ is a constant vector independent of spatial coordinates and $\vec{k}\cdot\vec{x}$ is a scalar product of the vectors $\vec{x} = \{x_1, x_2, x_3\}$ and $\vec{k} = \{k_1, k_2, k_3\}$, the so called wavevector. Equations [6] and [7] yield the dispersion relation:

$$k^2 = \varepsilon_0 \mu_0 \omega^2 = k_1^2 + k_2^2 + k_3^2 \qquad [8]$$

For a given $\vec{k}$ vector, a constant phase front is determined by $\vec{k}\cdot\vec{x} = \text{const}$, which indicates that the front is perpendicular to the wavevector. So, the phase front is a plane and the amplitude of the electric field on the plane is a constant.

The Maxwell equations for the plane wave solution become:

$$\begin{aligned} &\vec{k}\times\mathbf{E}(x) = k\mathbf{H}(x), \quad \vec{k}\times\mathbf{H}(x) = -k\mathbf{E}(x) \\ &\Rightarrow \vec{k}\cdot\mathbf{E} = 0, \quad \vec{k}\cdot\mathbf{H} = 0 \end{aligned} \qquad [9]$$

Therefore, vectors $\mathbf{E}$ and $\mathbf{H}$ belong to the constant phase plane perpendicular to vector $\vec{k}$ and are orthogonal each other.

Poynting's Vector. John Henry Poynting (1852–1914)

For time-dependent fields, the Poynting vector is introduced as

$$\vec{S}(x,t) = \vec{E}(x,t)\times\vec{H}(x,t) \qquad [10]$$

It is interpreted as the power flow density with the dimension of watts/m^2. The time-average Poynting's vector power density is given by

$$\langle S\rangle = \frac{1}{T}\int_0^T \vec{S}(x,t)\mathrm{d}t \qquad [11]$$

For the time-dependent planewave fields:

$$\begin{aligned} \vec{E}(x,t) &= \mathbf{E}_0 \cos(\vec{k}\cdot\vec{x} - \omega t) \\ \vec{H}(x,t) &= \mathbf{H}_0 \cos(\vec{k}\cdot\vec{x} - \omega t) \end{aligned} \qquad [12]$$

The vector Poynting's power density shown in eqn [10], accepts the form:

$$\vec{S}(x,t) = |\mathbf{E}_0|^2 \cos^2(\vec{k}\cdot\vec{x} - \omega t) \qquad [13]$$

The time-average vector power density gives

$$\langle\vec{S}(x,t)\rangle = \frac{1}{2}|\mathbf{E}_0|^2 \qquad [14]$$

which shows that Poynting's power density is also propagating in the direction of the wave vector $\vec{k}$.

Boundary Conditions

Let us consider a bound obstacle D_i with a smooth closed surface S. The Maxwell eqn [5] has been written in differential form. It must be supplemented with boundary conditions enforced at the obstacle surface. The boundary conditions can be derived from the integral form of Maxwell equations. The field vectors $\mathbf{E}$, $\mathbf{H}$ are assumed to be finite but may be discontinuous across the obstacle surface S. In cases where there is no surface current density (which is so for a perfect conductor) the boundary conditions accept the following form:

$$\vec{n}\times(\mathbf{E}_\mathrm{i} - \mathbf{E}_\mathrm{e}) = 0, \quad \vec{n}\times(\mathbf{H}_\mathrm{i} - \mathbf{H}_\mathrm{e}) = 0 \qquad [15]$$

where subscripts i and e correspond to the total fields inside and outside the obstacle, respectively, and $\vec{n}$ is unit outward normal to S. Essentially the boundary conditions state that tangential components for electromagnetic fields are continuous across the boundary.

Silver–Müller Radiation Conditions

Analysis of a direct scattering problem on a local obstacle involves Maxwell eqn [10] and boundary conditions as in eqn [15] and some infinity conditions, which provide the uniqueness of the scattering problem solutions. Let us assume that a local obstacle is excited by an incident electromagnetic field $\{\mathbf{E}_0, \mathbf{H}_0\}$, given as entire solution to the Maxwell equations. Total field outside the obstacle can be represented as

$$\mathbf{E}_\mathrm{e} = \mathbf{E}_0 + \mathbf{E}_\mathrm{s}, \quad \mathbf{H}_\mathrm{e} = \mathbf{H}_0 + \mathbf{H}_\mathrm{s} \qquad [16]$$

where scattered field $\{\mathbf{E}_\mathrm{s}, \mathbf{H}_\mathrm{s}\}$ in the exterior domain $D_\mathrm{e} = \mathbb{R}^3/\bar{D}_\mathrm{i}$ satisfies the Silver–Müller radiation conditions:

$$\begin{aligned} &\frac{x}{|x|}\times\mathbf{H}_\mathrm{s} + \mathbf{E}_\mathrm{s} = \mathrm{o}\left(\frac{1}{|x|}\right) \\ &\frac{x}{|x|}\times\mathbf{E}_\mathrm{s} - \mathbf{H}_\mathrm{s} = \mathrm{o}\left(\frac{1}{|x|}\right) \end{aligned} \qquad |x|\to\infty \qquad [17]$$

uniformly for all directions $x/|x|$.

Scattering Amplitude

The following result is valid: every radiating solution $\{\mathbf{E}_s, \mathbf{H}_s\}$ to the Maxwell equations has the following asymptotic form:

$$\mathbf{E}_s(x) = \frac{e^{ik|x|}}{|x|}\left\{\mathbf{E}_\infty(\boldsymbol{\varpi}) + O\left(\frac{1}{|x|^2}\right)\right\}$$
$$\mathbf{H}_s(x) = \frac{e^{ik|x|}}{|x|}\left\{\mathbf{H}_\infty(\boldsymbol{\varpi}) + O\left(\frac{1}{|x|^2}\right)\right\} \qquad |x| \to \infty \qquad [18]$$

where the vector functions $\mathbf{E}_\infty$ and $\mathbf{H}_\infty$ defined on the unit sphere $\boldsymbol{\varpi} \in \Omega = \{0 \le \vartheta \le \pi; 0 \le \varphi \le 2\pi\}$ are known as the electric far field pattern (scattering amplitude) and magnetic far field pattern, respectively. They satisfy:

$$\mathbf{H}_\infty = \vec{\nu} \times \mathbf{E}_\infty, \quad \vec{\nu} \cdot \mathbf{E}_\infty = \vec{\nu} \cdot \mathbf{H}_\infty = 0 \qquad [19]$$

with the unit outward normal $\vec{\nu}$ on Ω. The latter means that $\mathbf{E}_\infty$ and $\mathbf{H}_\infty$ have only tangential components at the unit sphere.

Far field pattern represents real value analytic functions on a unit sphere. Reillich's lemma establishes one-to-one correspondence between radiating electromagnetic field and their far field pattern. It means that there exists only one scattered field $\{\mathbf{E}_s, \mathbf{H}_s\}$ corresponding to a given far field pattern $\mathbf{E}_\infty$.

Reciprocity Relations

Examine the set of far field pattern corresponding to the scattering problem including: Maxwell eqn [5], boundary conditions in eqn [15], and radiation conditions in eqn [17] at the infinity for the scattered fields. Let the incident electromagnetic field is given by the planewave:

$$\mathbf{E}_0(x) = \vec{e}_0\, e^{ik\vec{x}\cdot\vec{\kappa}}, \quad \mathbf{H}_0(x) = \vec{\kappa} \times \vec{e}_0\, e^{ik\vec{x}\cdot\vec{\kappa}} \qquad [20]$$

where $\vec{\kappa} \in \Omega$ is a unit vector giving the propagation direction and $\vec{e}_0 \in \mathbb{R}^3$ is a constant vector responsible for the polarization. From the previous consideration one can see that:

$$\mathbf{E}_s(x; \vec{\kappa}, \vec{e}_0) = \frac{e^{ik|x|}}{|x|}\left\{\mathbf{E}_\infty(\boldsymbol{\varpi}\vec{\kappa}, \vec{e}_0) + O\left(\frac{1}{|x|^2}\right)\right\}, \quad |x| \to \infty \qquad [21]$$

Furthermore, the following reciprocity relation for the scattering amplitude holds:

$$\vec{h}_0 \cdot \mathbf{E}_\infty(\boldsymbol{\varpi}; \vec{\kappa}, \vec{e}_0) = \vec{e}_0 \cdot \mathbf{E}_\infty(-\vec{\kappa}; -\boldsymbol{\varpi}, \vec{h}_0)$$

for all vectors $\boldsymbol{\varpi}, \vec{\kappa} \in \Omega$ and $\forall \vec{e}_0, \vec{h}_0 \in \mathbb{R}^3$.

Optical Theorem

Far field pattern enables us to evaluate scattering cross section, which is defined as

$$\sigma_s = \int_\Omega |\mathbf{E}_\infty(\boldsymbol{\varpi})|^2 d\boldsymbol{\varpi} \qquad [22]$$

For the planewave given by eqn [20] and excitation of nonabsorbing obstacle ($\sigma = 0$), the following relation holds:

$$\sigma_s = \frac{4\pi}{k} \mathrm{Im}\{\mathbf{E}_\infty(\vec{\kappa}) \cdot \vec{e}_0\} \qquad [23]$$

This is so-called optical theorem.

Huygen's Principle (Green's Theorem). Christian Huygens (1629–1695)

The Huygens' principle shows that a wavefield on the surface determines the wavefield off the surface. Let us introduce the scalar Green's function which satisfies the Helmholtz equation:

$$(\nabla^2 + k^2) g(x, y) = -\delta(x - y) \qquad [24]$$

where $\delta(x - y)$ is the Dirac delta function. For three-dimensional problems, the scalar Green's function $g(x, y)$ for isotropic media accepts the form:

$$g(x, y) = \frac{e^{jk|x-y|}}{4\pi|x - y|} \qquad [25]$$

Let electromagnetic fields $\{\mathbf{E}_s, \mathbf{H}_s\}$ be a radiating solution to the Maxwell eqn [5] in D_e. Then for any $x \in D_e$ we have the Stratton–Chu formulas:

$$\mathbf{E}_s(x) = \nabla \times \int_S g(x, y) \vec{\nu}_y \times \mathbf{E}_s(y) ds_y + \frac{j}{k} \nabla \times \nabla \times \int_S g(x, y) \vec{\nu}_y \times \mathbf{H}_s(y) ds_y$$
$$\mathbf{H}_s(x) = \nabla \times \int_S g(x, y) \vec{\nu}_y \times \mathbf{H}_s(y) ds_y - \frac{j}{k} \nabla \times \nabla \times \int_S g(x, y) \vec{\nu}_y \times \mathbf{E}_s(y) ds_y \qquad [26]$$

here S is a smooth closed surface, and $\vec{\nu}$ is the unit normal vector to S directed into D_e. Similar relations are valid for an electromagnetic field inside a bounded domain D_i. The Stratton–Chu

formulas are extensively employed to construct boundary integral equations for scattering problems investigating.

Homogeneous Obstacle

Consider a mathematical statement of the scattering problem by a bound homogeneous obstacle. Let external excitation be a planewave $\{\mathbf{E}_0, \mathbf{H}_0\}$: (eqn [20]). In this case the boundary-value scattering problem can be formulated in the form:

$$\begin{aligned} &\nabla \times \mathbf{E}_\tau(x) = \mathrm{j}k\mathbf{H}_\tau(x), \quad \nabla \times \mathbf{H}_\tau(x) = -\mathrm{j}k\varepsilon_\tau \mathbf{E}_\tau(x), \\ &\quad x \in D_\tau, \tau = \mathrm{i}, \mathrm{e} \\ &\qquad \vec{\nu} \times (\mathbf{E}_\mathrm{i} + \mathbf{E}_\mathrm{e}) = \vec{\nu} \times \mathbf{E}_0, \\ &\qquad \vec{\nu} \times (\mathbf{H}_\mathrm{i} + \mathbf{H}_\mathrm{e}) = \vec{\nu} \times \mathbf{H}_0 \ \text{ on } \ S \\ &\qquad \frac{x}{|x|} \times \mathbf{H}_\mathrm{e} + \mathbf{E}_\mathrm{e} = \mathrm{o}\left(\frac{1}{|x|}\right), \quad |x| \to \infty \end{aligned} \qquad [27]$$

where $\varepsilon_\mathrm{e} = 1$, $\varepsilon_\mathrm{i} = \varepsilon$ and $\{\mathbf{E}_\mathrm{e}, \mathbf{H}_\mathrm{e}\}$ corresponds to scattered field. For a smooth surface and the case $\mathrm{Im}\,\varepsilon \geq 0$, the problem has a unique solution.

Lippman–Swinger Integral Equation

Assume we deal with the local inhomogeneous obstacle that can be described by dielectric permittivity $\varepsilon(x)$, which is a continuous complex-valued ($\mathrm{Im}\,\varepsilon(x) \geq 0$) function in D_i. The correspondent scattering problem can be formulated as below in eqn [29]. In this case, the boundary-value scattering problem can be reduced to the solution of the Lippman–Swinger integral equation

$$\begin{aligned} \mathbf{E}_\mathrm{i}(x) = \ &\mathbf{E}_0(x) + \nabla \times \nabla \times \int_{D_\mathrm{i}} g(x, y) k^2 (\varepsilon(y) - 1) \\ &\times \mathbf{E}_\mathrm{i}(y) \mathrm{d}\nu_y, \quad x \in D_\mathrm{i} \end{aligned} \qquad [28]$$

Volume integral eqn [28] is equivalent to the corresponding scattering problem. Once the solution is found, the associated scattered field outside the inhomogeneous obstacle can be represented as

$$\begin{aligned} \mathbf{E}_\mathrm{s}(x) = \ &\nabla \times \nabla \times \int_{D_\mathrm{i}} g(x, y) k^2 (\varepsilon(y) - 1) \\ &\times \mathbf{E}_\mathrm{i}(y) \mathrm{d}\nu_y, \quad x \in D_\mathrm{e} \end{aligned} \qquad [29]$$

The Lippman–Swinger integral equation plays an important role under analysis of a wave scattering by local inhomogeneous obstacle.

Born Approximation

In the case where $k^2 |\varepsilon(x) - 1| mes D_\mathrm{i} \ll 1$, or where the second term on the right-hand part of the Lippman–Swinger integral eqn [28] is small compared to the first term, one can write $\mathbf{E}_\mathrm{i}(x) \simeq \mathbf{E}_0(x)$. Then the scattered field can be approximately evaluated as

$$\begin{aligned} \mathbf{E}_\mathrm{s}(x) = \ &\nabla \times \nabla \times \int_{D_\mathrm{i}} g(x, y) k^2 (\varepsilon(y) - 1) \\ &\times \mathbf{E}_0(y) \mathrm{d}\nu_y, \quad x \in D_\mathrm{e} \end{aligned} \qquad [30]$$

The above representation for the scattered field is known as the first-order Born approximation. It is also the first-order approximation in the Neumann series for the integral eqn [28]. The Born approximation plays an important role in diffraction tomography.

Rayleigh Scattering

Rayleigh scattering characterizes the scattering of electromagnetic waves by particles much smaller then an exciting wavelength. Consider a spherical particle with constant permittivity ε and radius a deposited at the origin of a coordinate system. Let linear polarized planewave be incident upon the particle $\mathbf{E}_0(x) = \vec{e}_x e^{\mathrm{j}kz}$, and $\vec{e}_x$ be unit vector Cartesian coordinate system. It is polarized in x-direction and propagates along z-axis. In this case, for the scattered intensity, we get:

$$I_\mathrm{s} = \left(\frac{\varepsilon - 1}{\varepsilon + 2}\right)^2 \frac{k^4 a^6}{r^2} \sin^2 \vartheta \qquad [31]$$

where r is distance from the origin and $\vartheta \in [0, \pi]$.

Corresponding scattering cross-section is calculated as

$$\sigma_\mathrm{s} = \frac{8\pi}{3}\left(\frac{\varepsilon - 1}{\varepsilon + 2}\right)^2 k^4 a^6 \qquad [32]$$

Thus the total scattered power is proportional to the fourth power of the wave number (minus fourth power of the wavelength – Rayleigh's Law). The scattered power is also proportional to the sixth power of the radius.

Mie Scattering

The Mie theory is a complete mathematical–physical theory of the scattering of electromagnetic wave by homogeneous spherical particles, developed by Gustav Mie in 1908. In contrast to Rayleigh scattering, the Mie theory embraces all possible ratios of the particle radius to wavelength. It relays

upon the exact solution to Maxwell equation for spherical homogeneous obstacle with planewave excitation. In the frame of Mie theory, the scattered field is represented as an infinite series of fields of spherical multipoles, which are orthogonal at a sphere surface. Mie solution plays an important role in light scattering by clouds of particles.

See also

Scattering: Raman Scattering; Scattering from Surfaces and Thin Films; Scattering Phenomena in Optical Fibers; Stimulated Scattering.

Further Reading

Boheren C and Hoffman D (1983) *Absorption and Scattering of Light by Small Particles*. New York: Wiley.
Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.
Chew WC (1995) *Waves and Fields in Inhomogeneous Media*. New York: IEEE Press.
Colton D and Kress R (1992) *Inverse Acoustic and Electromagnetic Scattering Problems*. New York: Springer-Verlag.
Guenther RD (1990) *Modern Optics*. New York: Wiley.
Kong JA (2000) *Electromagnetic Wave Theory*. Cambridge: EMW Publishing.
Kong JA (2002) *Maxwell Equations*. Cambridge: EMW Publishing.

Stimulated Scattering

M Bashkansky and J Reintjes, Naval Research Laboratory, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Scattering of light involves conversion of an incoming light wave in one mode to an outgoing wave in a different mode. Scattering can involve changes of wavelength, direction, and polarization, and can occur in all types of materials. Without scattering we would not be able to see objects in the world around us. Scattering processes involving particular physical interactions are often given distinctive names. Familiar examples include Rayleigh scattering from the random distribution of air molecules, which makes the sky blue, and Tyndall scattering from small suspended particles, which makes the sunset red. Other types of scattering, e.g., Raman, Brillouin and Rayleigh-wing scattering, are important spectroscopic tools that provide insight into atomic and molecular properties. In their most commonly encountered form, where the intensity of the scattered light is proportional to the intensity of the incident light, the scattering processes are termed spontaneous. With the intense radiation provided by lasers, however, the intensity of the scattered light can grow exponentially, and the processes are termed stimulated scattering.

Stimulated scattering is a subset of the larger discipline of nonlinear optics. The term was first used to describe nonlinear optical interactions (e.g., Raman, Brillouin, Rayleigh) in which the intensity of the scattered light grows exponentially, in a manner similar to the exponential growth of light intensity associated with lasers. These interactions involve exchange of energy with a second mode that is an internal mode of the scattering medium, and the scattered light wave is at a different wavelength from the incoming light. The use of the term stimulated has subsequently been expanded to include a wider class of nonlinear optical interactions, such as parametric fluorescence or certain aspects of self-focusing, in which the second mode is itself a light wave, with either a different frequency or direction of propagation.

In the wider sense, scattering of light occurs through interaction of the incoming light wave with an 'idler' mode in a material. Stimulated scattering occurs when the idler mode is itself driven by interference of the incoming and scattered light waves. Under these circumstances, exponential growth of the scattered wave results. Depending on the interaction, the idler mode may be an internal mode of the material or another light wave. Most linear scattering processes have stimulated counterparts. However, there are stimulated scattering interactions that have no spontaneous analogs.

The elements of stimulated light scattering can be summarized mathematically as follows. The scattering process is modeled as an incoming laser or pump, an outgoing scattered wave and an idler mode. The interaction of the laser with the idler mode to produce the scattered light wave is described by the equation

$$\frac{\partial}{\partial z}A_{\mathrm{S}}(z,t)+\frac{n}{c}\frac{\partial}{\partial t}A_{\mathrm{S}}(z,t)=\kappa A_{\mathrm{L}}b^{*} \qquad [1]$$

where A_{S} and A_{L} are the slowly varying amplitudes of the scattered and laser light waves whose electric fields are defined as

$$E_{\mathrm{S(L)}}=\frac{1}{2}\left\{A_{\mathrm{S(L)}}\mathrm{e}^{\mathrm{i}[\omega_{\mathrm{S(L)}}t-k_{\mathrm{S(L)}}z]}+\mathrm{cc}\right\} \qquad [2]$$

$\omega_{S(L)}$ is the frequency of the scattered (laser) light, $k_{S(L)}$ is the corresponding wavevector, κ is a coupling coefficient, cc denotes complex conjugate, and b is the amplitude of the idler mode given by

$$B = \frac{1}{2}\left\{b e^{i[\omega_b t - k_b z]} + \text{cc}\right\} \quad [3]$$

The interaction of the laser and scattered wave to drive the idler mode is described by the equation

$$\mathcal{L}(b^*) = \kappa' A_L^* A_S \quad [4]$$

where $\mathcal{L}$ is a differential operator that describes the dynamics of the idler mode, e.g., a damped oscillator or a propagating wave, and κ' is a second coupling coefficient. For the class of stimulated processes that involve excitation of an internal mode of the medium, $\mathcal{L}$ has the form

$$\mathcal{L} = \frac{\partial}{\partial t} + \frac{1}{T_2} - v\frac{\partial}{\partial z} \quad [5]$$

where T_2 is the damping time of the coherence of the internal mode. When interaction times are long compared to internal response and decay times,

$$b^* \sim \kappa' A_L^* A_S$$

and the scattered wave obeys an equation of the form

$$\frac{\partial}{\partial z} A_S(z,t) = \frac{g}{2}|A_L|^2 A_S(z,t)$$

which has an exponential-gain solution for the scattered wave intensity:

$$I_S(z,t) = I_S(0,t)e^{I_L g z}$$

When the idler mode is a light wave, as in stimulated parametric fluorescence, [4] has the same form as [1] and they can be solved as a coupled pair, leading to transcendental or hyperbolic solutions. Within the framework of nonlinear optics, interactions of these types can be expressed as arising from a nonlinear optical polarization, which forms the right-hand side of [1]. The dynamics of the nonlinear polarization are then described by [4].

Stimulated processes are often associated with pumping thresholds. When these processes are configured with cavity resonators, they possess a threshold in pumping intensity just as with lasers or other oscillators. When they are operated in single or multipass scattering cells, they have no true threshold. However, they still have a practical useful threshold when the growth from quantum noise raises the scattered signal to a level comparable to the pump. By convention, this is taken to be the incident light intensity at which the scattered intensity reaches approximately 1% of the pump intensity, typically involving gains of the order of e^{23}–e^{30}. At this level, a small increase in pump intensity raises the scattered intensity to levels at which the pump starts to be depleted.

Attenuation of the scattered light can be accounted for by replacing $I_L g$ in the above expression by $I_L g - \alpha$, where α is the attenuation coefficient. Exponential gain is now experienced when the pump intensity exceeds the level given by

$$I_L > \frac{\alpha}{g}$$

This value can serve as an estimate of threshold pumping in cavity configurations, although in free propagation situations, net gain must still approach e^{23} to exhibit threshold behavior.

A list of stimulated scattering interactions is given in Table 1.

Stimulated Raman Scattering

Stimulated Raman scattering (SRS) involves conversion of an incoming light wave at one frequency, termed the laser or pump wave, into a scattered wave of different frequency, along with excitation or de-excitation of an internal mode of the medium. The frequency difference between the pump and scattered waves is equal to the frequency of the material mode and is characteristic of the material. When the frequency of the scattered wave is lower than that of the incident wave, the scattered wave is termed a Stokes wave, and the internal mode of the medium is excited in the scattering process. When the frequency of the scattered wave is higher than that of the incident wave, the scattered wave is termed an anti-Stokes wave, and the internal mode of the medium is de-excited in the scattering process.

The internal mode of the medium that is involved in Raman scattering is a nonpropagating collective mode. Many different types of internal modes in all types of materials can be involved in Raman scattering. Examples include molecular vibrations and rotations in gases, liquids, or solids, electronic states that are of the same parity as the ground state in gases or solids, optical phonons or polaritons in solids, and Langmuir waves in plasmas. Raman frequency shifts can range from tens of wavenumbers for lattice vibrations to tens of thousands of wavenumbers for electronic transitions.

Spontaneous Raman scattering was first described theoretically by Smekal in 1923 from a quantum mechanical study of light scattering, and was

Table 1 List of stimulated scattering interactions

Stimulated scattering process	*Internal material mode*	*Light mode change*
Stimulated Raman scattering	Nonpropagating internal energy mode: molecular vibrations and rotations, same parity electronic states, optical phonons and Langmuir waves	Change in wavelength, polarization, direction
Stimulated Brillouin scattering	Propagating sound waves	Change in wavelength, polarization, and direction
Stimulated Rayleigh scattering	Temperature and density fluctuations due to light absorption	Change in direction
Stimulated Rayleigh wing scattering	Molecular distribution and orientation fluctuations	Change in wavelength and direction
Stimulated concentration scattering	Fluctuations of concentration in gas mixtures	Change in direction
Photorefractive stimulated scattering	Diffusion of photo-excited carriers	Change in direction
Stimulated Thomson/Compton scattering	Momentum changes in bunched free electrons	Change in wavelength and direction
Stimulated parametric fluorescence	Nonresonant nonlinear polarization	Change in wavelength and direction
Self-focusing	Nonresonant nonlinear polarization	Change in direction

demonstrated experimentally by Raman, and shortly after by Landsberg and Mandelstamm in 1928. Stimulated Raman scattering was first observed in 1962 by Woodbury and Ng, and Eckhardt *et al.* soon after the development of pulsed lasers.

Stimulated Raman scattering can be observed in many different configurations including cavity resonators, single or multiple-pass generator cells or amplifiers. It can be observed as forward scattering, with the scattered light propagating in the same direction as the pump light, or backward scattering, with the scattered light propagating in the opposite direction to the pump light. Its most widespread use is for generation of coherent light at frequencies different from the pump light. Other applications include phase conjugation, optical gating, beam cleanup, and pulse compression.

The equations describing SRS when depletion of the pump light and diffraction of the optical beams are neglected are:

$$2ik_{\mathrm{S}}\left(\frac{\partial A_{\mathrm{S}}}{\partial z}\pm\frac{1}{v_{\mathrm{gS}}}\frac{\partial A_{\mathrm{S}}}{\partial t}\right)=-\frac{N}{2}\frac{\omega_{\mathrm{S}}^2}{c^2}\left(\frac{\partial\alpha}{\partial Q}\right)_{\mathrm{S}}Q^*A_{\mathrm{L}} \quad [6]$$

$$\frac{\partial Q^*}{\partial t}+\Gamma Q^*=-\frac{\mathrm{i}}{4\omega_{\mathrm{o}}m}\left(\frac{\partial\alpha}{\partial Q}\right)_{\mathrm{S}}A_{\mathrm{L}}^*A_{\mathrm{S}} \quad [7]$$

Here A_{S} and A_{L} are the slowly varying optical field amplitudes of the Stokes and laser waves as given in [2], $\omega_{\mathrm{S(L)}}$ are the optical frequencies, $k_{\mathrm{S(L)}}$ are the corresponding k vectors, $v_{\mathrm{S(L)}}$ are the group velocities, and Q is the amplitude of the idler mode, in this case the normal mode coordinate of the material excitation, defined by:

$$\mathbf{Q}(x,y,z,t)=\frac{1}{2}(Q(x,y,z,t)\mathrm{e}^{-\mathrm{i}(\omega_{\mathrm{o}}t-k_{\mathrm{o}}z)}+\mathrm{cc}) \quad [8]$$

$(\partial\alpha/\partial Q)_S$ is the Stokes hyperpolarizability, $\omega_0=\omega_{\mathrm{L}}-\omega_{\mathrm{S}}$ is the Raman transition frequency, and k_0 is the k vector of the material excitation. The phase matching relation $k_{\mathrm{S}}=k_{\mathrm{L}}-k_0$ is automatically satisfied for SRS because the material mode that is involved is nonpropagating. Γ is the half width at half maximum of the Raman linewidth given by $\Gamma=1/\mathrm{T}_2$, where T_2 is the dephasing time. N is the number of molecules in the ground state and m is the effective reduced mass of the material oscillation.

When the fields are constant or vary slowly compared to the response time of the material, the Stokes intensity is given by

$$I_{\mathrm{S}}(z)=I_{\mathrm{S}}(0)\mathrm{e}^{I_{\mathrm{L}}g_{\mathrm{ss}}Z} \quad [9]$$

where the intensity I is given by

$$I_{\mathrm{S(L)}}=\frac{1}{2}(cn_{\mathrm{S(L)}}\varepsilon_0)|A_{\mathrm{S(L)}}|^2$$

and

$$g_{\mathrm{ss}}=\frac{N\omega_{\mathrm{s}}}{4\Gamma n_{\mathrm{S}}n_{\mathrm{L}}m\omega_0c^2\varepsilon_0}\left(\frac{\partial\alpha}{\partial Q}\right)^2 \quad [10]$$

where $n_{S(L)}$ is the index of refraction of the Stokes (laser) in the material and ε_0 is the dielectric permittivity of the vacuum.

The Stokes wave grows exponentially with z at a rate that depends on the pump intensity. If the Stokes wave becomes intense enough, its power can become comparable to that of the original pump wave. In that case depletion of the pump wave must be taken into account with an equation for the pump wave amplitude similar to [6]. The Stokes intensity is now given by

$$I_S(z) = \frac{I_S(0)e^{I_L(0)gz}}{1 + \dfrac{\omega_L}{\omega_S}\dfrac{I_S(0)}{I_L(0)}e^{I_L(0)gz}} \qquad [11]$$

which is valid when $I_S(0) \ll I_L(0)$.

In the limit of strong interactions all of the laser energy can be depleted. The maximum Stokes energy is given by $I_S = I_L(\omega_S/\omega_L)$, where ω_S/ω_L is termed the Manly–Rowe ratio. The energy that is removed from the laser but not converted to Stokes radiation is transferred to the material.

For each photon produced in the scattered wave, a corresponding quantum of energy is given to the material mode and a photon is removed from the laser. In principle, 100% of the laser photons can be converted to Stokes photons, but complete conversion is not achieved in practice. Photon conversion efficiencies greater than 90% have been reported in some pulsed laser experiments.

SRS frequency shifts and gains of some materials are shown in Table 2.

When pulses are used with durations comparable to or shorter than the response time of the material, ranging from picoseconds to femtoseconds depending on the material, the interaction is transient and the effects of the time derivatives in [6] and [7] must be accounted for. Under these conditions the Stokes intensity is given as a temporal integral with Bessel function dependence on the laser intensity. When the pulse duration is much shorter than the material dephasing time and the incident Stokes functional form is the same as the pump, the amplified Stokes intensity is given by the expression

$$I_S(z,\tau) = I_S(0,\tau)I_0^2\left(\sqrt{2g_{ss}\Gamma z\int_{-\infty}^{\tau} I_L(\tau')d\tau'}\right) \qquad [12]$$

where I_0 is a modified Bessel function. In this regime, the Stokes intensity depends only on the integrated pump energy. The Stokes gain is reduced from its steady state value for pulses of constant intensity, but for pulses of constant energy the highest Raman gain is obtained in the limit of transient interactions.

One of the most important applications of Raman scattering is the generation of new coherent waves at frequencies shifted from the pump. In this case, an incident Stokes wave is not supplied as the starting term in [9] or [12]. Rather, the process is initiated by the vacuum fluctuations in the Stokes and material oscillator fields. Quantum field creation and annihilation operators must replace the classical field amplitudes of [6] and [7]. The intensity of the generated Stokes wave is given by the expectation value of the normally ordered number operator and in the extreme transient regime is given by:

$$I_S(z,d\tau) = \frac{1}{2A}\hbar\omega_S\Gamma g_{ss}I_L(d\tau)z \times\left\{I_0^2\left(\sqrt{2\Gamma g_{ss}z\int_{-\infty}^{\tau} I_L(\tau')\tau'}\right) - I_1^2\left(\sqrt{2\Gamma g_{ss}z\int_{-\infty}^{\tau} I_L(\tau')\tau'}\right)\right\} \qquad [13]$$

where A is the cross-sectional area of the interaction geometry. The Stokes light generated in this fashion has all the characteristics of amplified spontaneous emission of lasers operated below threshold until the Stokes intensity becomes high enough to deplete the laser.

The stochastic nature of the SRS quantum mechanical source is manifest in the macroscopic Stokes light when the Stokes generator is operated well below pump depletion levels. For a single spatial mode the probability density distribution of

Table 2 List of frequency shifts and gains of selected materials for stimulated Raman scattering

Material, λ_L (nm)	*Shift* (cm^{-1})	*Gain (cm/GW)*
H_2 gas (20 atm), 532	4155	2.5
D_2 gas (60 atm), 532	2987	0.45
CH_4 (115 atm), 532	2913	1.26
H_2O, 694	3290	0.14
CS_2, 694	655.6	24
$LiNbO_3$, 694	637	9.4
$Ba(NO_3)_2$, 532	1047	47

Table compiled in parts from: Kaiser W and Maier M (1972) Stimulated Rayleigh, Brillouin and Raman Spectroscopy. In: Arecchi FT and Schulz-DuBois EO (eds) *Laser Handbook*, vol. 2. pp. 1077–1150. Amsterdam, The Netherlands: North Holland and Reintjes J and Bashkansky M (2001) Stimulated Raman and Brillouin Scattering. In: Bass M (ed.) *Optical Society of America Handbook of Optics*, Ch. 18, 2nd ed., vol. IV, New York: McGraw-Hill.

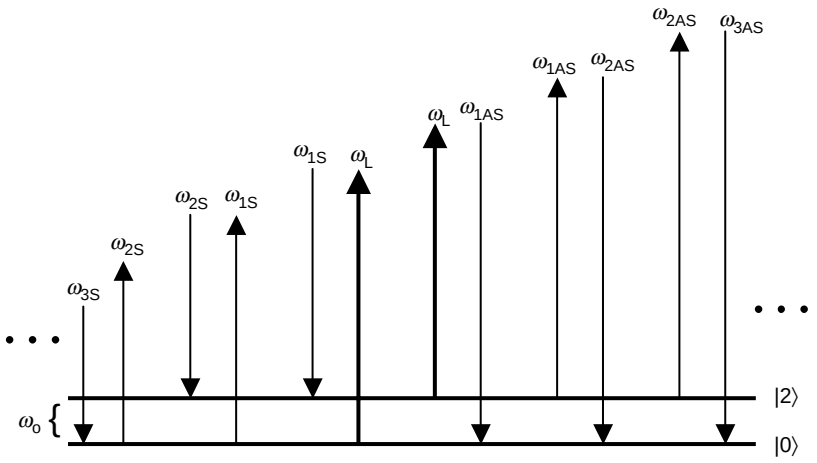


Figure 1 SRS level diagram showing generation of multiple Stokes and anti-Stokes frequencies.

the Stokes energy has the form of a negative exponential:

$$P(W) = \frac{1}{\langle W \rangle} \exp\left[- \frac{W}{\langle W \rangle} \right]$$

Statistical fluctuations in pulse energy, spatial profile, pointing, and in the spectral and temporal structure of the Stokes signal have all been observed.

If the intensity of the generated Stokes light becomes comparable to the pump intensity, it can serve as a pump wave for its own Raman interaction, producing a wave termed the second Stokes wave, shifted from the pump by frequency $2\omega_o$. If the process is driven hard enough third- and higher-order Stokes waves can be produced.

In addition to the interaction between the pump and Stokes beams described above, stimulated Raman scattering can also involve an anti-Stokes wave whose frequency is given by $\omega_{AS} = \omega_L + \omega_o$. The relationship of various Stokes and anti-Stokes waves is depicted in Figure 1. If a population inversion exists between the energy levels of the Raman process, then the anti-Stokes wave experiences exponential gain in a manner similar to that described above for the Stokes gain.

When all the population is in the ground state, which is usually the case, the anti-Stokes wave has exponential loss. The anti-Stokes wave can still grow without population inversion if the phase-matching condition

$$\Delta k = k_{AS} + k_S - 2k_L = 0$$

is approximately satisfied. Because of normal dispersion in the material this condition is usually not satisfied for collinear beams. However, if the Stokes and anti-Stokes beams propagate at small angles as shown in Figure 2, then the phase-matching condition can be met. For angles near but not exactly at phase matching the anti-Stokes wave grows exponentially as part of a mixed Stokes/anti-Stokes mode with a gain given by:

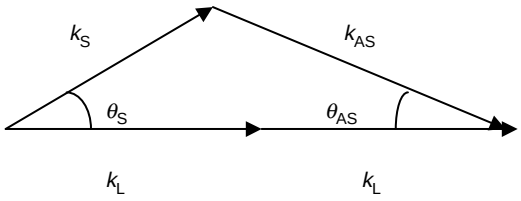


Figure 2 Phase matching diagram for Stokes–anti-Stokes coupling in a medium with positive dispersion.

$$g = \mathrm{Re}\left\{ \tfrac{1}{2}(K_3 - K_1)|A_L|^2 - \tfrac{i}{2}\sqrt{\Delta k^2 + 2i\Delta k(K_3 + K_1)|A_L|^2 - (K_1 - K_3)^2|A_L|^4} \right\} \quad [14]$$

At exact phase matching the interaction of Stokes and anti-Stokes waves serves to suppress the exponential gain of either wave.

As with Stokes waves, multiple-order anti-Stokes waves can be generated. Stokes and anti-Stokes orders up to at least 13 have been reported.

Stimulated Brillouin and Rayleigh Scattering

Light scattering can also occur through interaction with material density variations. Spontaneous

scattering through interaction with this type of internal mode produces a triplet of spectral lines. The up- and down-shifted lines are known as Brillouin anti-Stokes and Stokes lines, respectively, and are due to interaction with acoustic waves. The central peak is termed Rayleigh, or thermal Rayleigh scattering and is due to stationary density fluctuations arising from light absorption. A broad skirt around the central peak, which can be observed in anisotropic media, is due to the fluctuation of molecular distribution and orientation and is termed Rayleigh wing scattering. Each of these processes can give rise to stimulated scattering in the presence of intense radiation.

Density fluctuations are induced directly by light intensity through electrostriction, and indirectly by temperature fluctuations when optical absorption is important. A general classical model of the interaction, including both electrostriction and absorptive heating effects, is described by three equations:

The Navier-Stokes equation:

$$\rho_0\frac{\partial \mathbf{v}}{\partial t}+\frac{v^2}{\delta}\nabla(\Delta\rho)+\frac{v^2\beta\rho_0}{\delta}\nabla(\Delta T)-\eta\nabla^2\mathbf{v} = 2\gamma\varepsilon_0\nabla(\mathbf{E}_\mathrm{L}\cdot\mathbf{E}^*_\mathrm{R/B})-2\varepsilon_0\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho(\mathbf{E}_\mathrm{L}\cdot\mathbf{E}^*_\mathrm{R/B})\nabla(\Delta T) \quad [15]$$

The continuity equation:

$$\frac{\partial}{\partial t}\Delta\rho+\rho_0\nabla\cdot\mathbf{v}=0 \quad [16]$$

The energy transport equation:

$$\left(\rho_0 C_v\frac{\partial}{\partial t}-\lambda_\mathrm{T}\nabla^2\right)\Delta T-\frac{C_v(\delta-1)}{\beta}\frac{\partial}{\partial t}\Delta\rho = 4nc\alpha\varepsilon_0(\mathbf{E}_\mathrm{L}\cdot\mathbf{E}^*_\mathrm{R/B})+4\varepsilon_0\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho\left(T_0\mathbf{E}_\mathrm{L}\cdot\frac{\partial}{\partial t}\mathbf{E}^*_\mathrm{R/B}\right) \quad [17]$$

Here $\rho=\rho_0+\Delta\rho$, $T=T_0+\Delta T$, $\mathbf{v}$ is the volume element, v is the acoustic wave velocity, $\delta=C_p/C_v$ is the ratio of heat capacities at constant pressure and at constant volume, $\eta=(2\eta_\mathrm{s}+\eta_\mathrm{d})$ where η_s is the shear viscosity and η_d is the dilational viscosity, $\gamma=\rho_0(\partial\varepsilon/\partial\rho)_T$, λ_T and β are the thermal conductivity and thermal expansion coefficients, ε is the dielectric constant, and α is the linear absorption coefficient. $\mathbf{E}_\mathrm{L}$ is the pump field and $\mathbf{E}_\mathrm{R/B}$ is either the Rayleigh or Brillouin field.

Equations [15] and [16] can be combined to eliminate the $\mathbf{v}$ dependence:

$$\left[-\frac{\partial^2}{\partial t^2}+\left(\frac{v^2}{\delta}+\frac{\eta}{\rho_0}\frac{\partial}{\partial t}\right)\nabla^2\right]\Delta\rho+\frac{v^2\beta\rho_0}{\delta}\nabla^2(\Delta T) = 2\gamma\varepsilon_0\nabla^2(\mathbf{E}_\mathrm{L}\cdot\mathbf{E}^*_\mathrm{R/B})-2\varepsilon_0\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho\nabla\cdot[(\mathbf{E}_\mathrm{L}\cdot\mathbf{E}^*_\mathrm{R/B})\nabla(\Delta T)] \quad [18]$$

Together with the equations describing the effect of the medium on the optical fields [19a,b], this coupled set of equations describes simultaneously stimulated Brillouin and Rayleigh scattering:

$$\left(\frac{\partial}{\partial z}+\frac{1}{v_\mathrm{g}}\frac{\partial}{\partial t}+\frac{1}{2}\alpha\right)A_\mathrm{L} = \mathrm{i}\frac{\omega_\mathrm{L}}{4cn_\mathrm{L}}\left[\frac{\gamma}{\rho_0}\Delta\rho A_\mathrm{R/B}+\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho A_\mathrm{R/B}\Delta T\right] \quad [19a]$$

$$\left(\frac{\partial}{\partial z}-\frac{1}{v_\mathrm{g}}\frac{\partial}{\partial t}-\frac{1}{2}\alpha\right)A_\mathrm{R/B} = -\mathrm{i}\frac{\omega_\mathrm{R/B}}{4cn_\mathrm{R/B}}\left[\frac{\gamma}{\rho_0}\Delta\rho^* A_\mathrm{L}+\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho A_\mathrm{L}\Delta T^*\right] \quad [19b]$$

where the slowly varying envelope approximation [2] is used for the waves described by A_L, $A_\mathrm{R/B}$, $\Delta\rho$, and ΔT. The nonlinear polarization term is related to density and temperature wave amplitudes in the following way:

$$P^\mathrm{NL}=\varepsilon_0 E\left[\left(\frac{\partial\varepsilon}{\partial\rho}\right)_T\rho+\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho T\right] \quad [20]$$

Without the driving terms involving the optical fields, [17] and [18] describe broadband damped fluctuations in the variables $\Delta\rho$ and ΔT. When the optical driving terms are included, resonant behavior emerges. Steady state solutions can be obtained assuming strongly damped material excitations, allowing neglect of $\Delta\rho$ and ΔT derivatives. Separate contributions of the Rayleigh and Brillouin interactions to $\Delta\rho$ can be identified:

Rayleigh contribution:

$$\Delta\rho_\mathrm{R}=\left[\frac{\gamma(\Omega+\frac{1}{2}\mathrm{i}\delta\Gamma_\mathrm{R})-\mathrm{i}\gamma_\mathrm{a}\Omega_\mathrm{B}}{\Omega+\frac{1}{2}\mathrm{i}\Gamma_\mathrm{R}}\right]\frac{\varepsilon_0}{4v^2}A_\mathrm{L}A^*_\mathrm{R} \quad [21]$$

where γ_a is the absorptive coupling constant given by

$$\gamma_\mathrm{a}=\frac{2\alpha n v^2\beta}{C_p\Omega_\mathrm{B}}$$

The Brillouin contribution is:

$$\Delta\rho_B = \left[\frac{-(\gamma - i\gamma_a \nu q/\Omega)q^2\varepsilon_0}{4(\Omega^2 + i\Omega\Gamma_B - \nu^2 q^2)}\right] A_L A_B^* \qquad [22]$$

In [21] and [22] the frequencies and wavevectors satisfy the following relations:

$$\omega_L = \omega_{R/B} + \Omega \qquad [23a]$$

$$k_L = k_{R/B} + q \qquad [23b]$$

where ω_L is the incident laser frequency, $\omega_{R/B}$ is the frequency of the scattered Rayleigh or Brillouin wave, and k_L, $k_{R/B}$ are the corresponding wavevectors. Ω and q are the frequency and wavevector of the density fluctuation. Equations [23a,b] can also be described as energy and momentum conservation relations among incident and scattered photons and a phonon in the quantum mechanical picture.

The Rayleigh and Brillouin contributions exhibit resonant behavior at different frequency shifts. The Rayleigh contribution has its resonant frequency at $\Omega_R = 0$, resulting in typical diffusion behavior. The Rayleigh linewidth $\Gamma_R = 2\lambda_T q^2/\rho_0 C_p$ is the inverse of the decay time of the isobaric density fluctuations. The Brillouin contribution has its resonance at a frequency shift given by

$$\Omega_B = q\nu$$

where ν is the sound velocity. Thus at resonance the density fluctuations associated with the Brillouin effect correspond to sound waves. The Brillouin linewidth

$$\Gamma_B = \frac{q^2}{\rho_0}\left[\eta + \frac{\kappa}{C_p}\left(\frac{C_p}{C_\nu} - 1\right)\right]$$

is the inverse of the phonon lifetime τ_B.

In general Brillouin and Rayleigh interactions involve noncollinear scattering with the outgoing wave propagating at an arbitrary angle to the incident wave. In this case [23b] is a vector relation and the interaction geometry can be represented as shown in Figure 3.

The resulting relations among the k-vectors and frequencies of the various waves can place restrictions on the values of frequency shifts that are observed under various circumstances.

The coupled equations for the optical intensities in the steady state take the following form:

$$\frac{dI_L}{dz} = -g_{R/B} I_L I_{R/B} - \alpha I_L \qquad [24a]$$

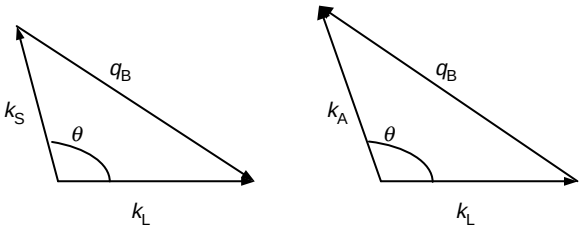


Figure 3 SBS phase matching diagram depicting relations between laser, k_L, sound wave, q_B, and Stokes and anti-Stokes, k_S, k_A k-vectors.

$$\frac{dI_{R/B}}{d\zeta} = g_{R/B} I_L I_{R/B} - \alpha I_{R/B} \qquad [24b]$$

where z is the direction of propagation of the incoming wave, ζ is the direction of propagation of the scattered wave, and $g_{R/B}$ is the gain coefficient of the Rayleigh or Brillouin effect. Equations [24a,b] describe interactions that have exponential gain. Stimulated Brillouin and Rayleigh scattering can be used in both a self-generator and an amplifier mode, but are used most often in a self-generator mode. Because the frequency shifts are small in these interactions, the starting signals for self-generators are thermal excitations of the internal modes rather than vacuum fluctuations.

Stimulated Brillouin Scattering

Stimulated Brillouin scattering (SBS) occurs through interaction with acoustic waves in solids, liquids, and gases and with ion-acoustic waves in plasmas. The Stokes or anti-Stokes frequency shifts for SBS are much smaller than for SRS, with typical values on the order of 0.1–100 GHz, depending on the excitation wavelength and interaction geometry as well as on material properties. Because the internal mode in SBS is a propagating wave, the phase matching condition shown in Figure 3 places restrictions on the frequency of the sound wave that is involved at various scattering angles. SBS is used most often for pulse compression and for correction of phase distortion in propagating light beams through phase conjugation. Although SBS is normally associated with high-power lasers in laboratory experiments, it is also a major factor in limiting the intensity of low-power lasers that can be used in optical fibers because of the long interaction lengths. SBS is also a prominent source of damage in solids and crystals caused by high-power lasers. It can also be used for generation and study of coherent hypersonic waves.

The SBS gain coefficient has both electrostrictive and absorptive contributions, $g_B = g_B^e + g_B^a$ where

$$g_B^e = \frac{\omega_B \gamma^2 \Omega_B}{2n_B n_L c^2 v^2 \rho_0 \Gamma_B} \frac{1}{1 + (2(\Omega - \Omega_B)/\Gamma_B)^2} \quad [25]$$

and

$$g_B^a = \frac{\omega_B \gamma \gamma_a \Omega_B}{4n_B n_L \rho_0 v^2 c^2 \Gamma_B} \frac{4(\Omega - \Omega_B)/\Gamma_B}{1 + (2(\Omega - \Omega_B)/\Gamma_B)^2} \quad [26]$$

The SBS gain from electrostriction has a maximum value of

$$g_B^e = \frac{\omega_B \gamma^2 \Omega_B}{2n_B n_L c^2 v^2 \rho_0 \Gamma_B} \quad [27]$$

at the resonant frequency shift

$$\Omega = \Omega_B = qv \quad [28]$$

For this condition the interaction is with a sound wave in the material. The frequency of the sound wave, and the gain, depend on the angle between the incident and scattered light. Using the vector relation

$$\mathbf{k}_B = \mathbf{k}_L - \mathbf{q}$$

depicted in Figure 3 and the approximations $\Omega_B << \omega_B, \omega_L, |k_B| \approx |k_L|$ we obtain

$$q = 2|k_L| \sin(\theta/2) \quad [29]$$

The Brillouin frequency shift is then given by

$$\Omega_B = 2n_L \omega_L (v/c) \sin(\theta/2) \quad [30]$$

and the Brillouin linewidth is given by

$$\Gamma_B = \frac{4}{\rho_0}\left[\eta + \frac{\kappa}{C_p}\left(\frac{C_p}{C_v} - 1\right)\right]\frac{\omega_L^2 n_L^2}{c^2} \sin^2(\theta/2)$$
$$= \Gamma'' \sin^2(\theta/2) \quad [31]$$

The peak Brillouin electrostrictive gain is then given by

$$g_B^e = \frac{\omega_B \gamma^2 \omega_L}{n_B c^3 v \rho_0 \Gamma_B} \sin(\theta/2) \quad [32a]$$

$$g_B^e = \frac{\omega^2 \gamma^2}{n_B c^3 v \rho_0 \Gamma'' \sin(\theta/2)} \quad [32b]$$

where we have used the approximation $\omega_B \approx \omega_L = \omega$.

The maximum electrostrictive SBS gain is independent of the laser frequency because the Brillouin linewidth is proportional to the square of the laser frequency.

As a function of angle the maximum Brillouin gain is usually observed in the backward direction. This is consistent with the form of the gain coefficient given in [32a], which indicates that the gain is proportional to $\sin(\theta/2)$. However, when the angular variation of the Brillouin linewidth is also accounted for, the gain coefficient is actually minimum in the backward direction [32b]. Typical laboratory experiments involve pencil-like geometries in which the interaction length is significant only in the backward or forward directions. In these situations, the total gain is actually maximum in the backward direction, where the interaction length is longest. Although the formulas appear to indicate that the gain goes to infinity in the forward direction, the damping time, which is proportional to Γ_B^{-1}, also goes to infinity and the interaction becomes transient. Under these conditions the gain parameter is proportional to $\sin(\theta/2)$ and the Brillouin gain goes to zero in the forward direction. Under some geometries Brillouin scattering at 90° to the laser can also be significant. It can be a source of damage to large glass components in high-power lasers.

For thermal Brillouin scattering the gain has a dispersive shape centered on the frequency shift $\Omega = \Omega_B$, with loss occurring on the low-frequency side and gain occurring on the high-frequency side, peaking at a frequency $\Omega = \Omega_B + \Gamma_B/2$. The maximum absorptive SBS gain depends on the laser frequency as ω^{-1}.

Table 3 shows representative values of the Brillouin linewidth and gain coefficients for a selection of substances.

Brillouin-enhanced four-wave mixing is a related effect that involves both a downshifted wave and an upshifted wave in which the coherent interaction is enhanced by the resonance associated with the sound wave. Gains of the order of 10^5 can be realized. Systems based on this effect are often used for phase conjugation.

Stimulated Rayleigh Scattering

Stimulated thermal Rayleigh scattering (R) can be observed in liquids while Stimulated Rayleigh wing scattering (SRWS) is observed in liquids with anisotropic molecules. Additionally, stimulated Rayleigh scattering due to fluctuations of concentration in gas mixtures can also occur. For R the internal mode obeys a diffusion equation. The R gain is given by

$$g_R = g_R^e + g_R^a$$

Table 3 Properties of stimulated Brillouin scattering for selected materials

Substance	*Laser wavelength (nm)*	*Frequency shift (Ghz)*	*$\Delta\nu$ (MHz)*	*τ_B* (ns)	*g_B^e (cm/GW)*	*g_B^a/α* (cm/GW)*
Liquid						
Acetone	532	5.93	361	0.44	12.9	22
Benzene	532	8.33	515	0.31	12.3	24
CS_2	532	7.7	120	1.9	130	20
CCl_4	532	5.72	890	0.18	8.77	13
Chloroform	532	5.75	635	0.25	11.7	
Ethanol	532	5.91	546	0.29	12*	10
Methanol	532	5.47	325	0.49	10.6	13
Water	532	7.4	607	0.26	2.94	0.8
Gas						
Xenon (7599 torr)	532	0.654 ± 0.024	98.1 ± 8.9	$0.65\lambda_p^2$P	1.38 ± 0.19	
SF_6(20 bar)	1320	0.2		35		
N_2(100 bar)	1320	0.5		30		
Solid						
BK 7	532	34.65 ± 0.039	165.0 ± 8.6		2.15 ± 0.21	
CaF_2	532	37.164 ± 1.185	45.6 ± 8.8		4.11 ± 0.65	
Plexiglas	532	15.687 ± 0.036	253.7 ± 12.6			
SiO_2	488	35.6	156		4.482	

*Absorptive maximum gain SBS values at 694 nm, with permission from Boyd RW (1992) *Nonlinear Optics*. Academic Press: New York. Reproduced from Reintjes J and Bashkansky M (2001) Stimulated Raman and Brillouin scattering. In: Bass M (ed.) *Optical Society of America Handbook of Optics*, Ch. 18, 2nd ed., vol. IV. New York: McGraw-Hill.

where g_R^e and g_R^a are the electrostrictive and absorptive contributions given by

$$g_R^e = \frac{\omega\gamma^2(\delta-1)}{4\rho_0 n_L n_R v^2 c^2}\frac{4\Omega/\Gamma_R}{1+(2\Omega/\Gamma_R)^2} \qquad [33]$$

$$g_R^a = \frac{-\omega\gamma\gamma_a\Omega_B}{2\rho_0 n^2 v^2 c^2 \Gamma_R}\frac{4\Omega/\Gamma_R}{1+(2\Omega/\Gamma_R)^2} \qquad [34]$$

Both contributions to the STRS gain have dispersive forms. The electrostrictive contribution is maximum for a down-shifted wave with $\Omega = \frac{1}{2}\Gamma_R$. Its maximum value scales linearly with laser frequency. For the absorptive process the gain depends inversely on the laser frequency, and is maximum for an upshifted wave with $\Omega = -\frac{1}{2}\Gamma_R$. Table 4 shows representative values of the Rayleigh linewidth and gain coefficients for a selection of substances.

Table 4 Properties of stimulated Rayleigh scattering for some materials at a wavelength of 694 nm

Substance	*Calculated maximum steady state gain factor*		*Linewidth $\delta\nu_R$ (MHz)*
	g_R^e(max) (cm/MW)	*g_R(max)/α (cm^2/MW)*	
CCl_4	2.6×10^{-4}	0.82	17
Methanol	8.4×10^{-4}	0.32	20
CS_2	6.0×10^{-4}	0.62	36
Benzene	2.2×10^{-4}	0.57	24
Aceton	2.0×10^{-4}	0.47	21
H_2O	0.02×10^{-4}	0.019	27.5
Ethanol		0.38	18

Reproduced from Kaiser W and Maier M (1972) Stimulated Rayleigh, Brillouin and Raman Spectroscopy. In: Arecchi FT and Schulz-DuBois EO (eds) *Laser Handbook*, vol. 2, pp. 1077–1150. Amsterdam, The Netherlands: North Holland.

Parametric Fluorescence

Some types of stimulated interactions do not involve exchange of energy with the material. These include parametric down-conversion and self-focusing. In parametric down-conversion an incident wave at frequency ω_L scatters into two longer-frequency light waves at signal and idler frequencies ω_s and ω_i. They satisfy the frequency and wavevector relations

$$\omega_L = \omega_s + \omega_i$$
$$k_L = k_s + k_i$$

The equations for the waves are:

$$\frac{\partial}{\partial z}A_s = -i\left(\frac{\omega_s}{2n_s c}\right)\chi^{(2)}A_L A_i^*$$

$$\frac{\partial}{\partial z}A_i = -i\frac{\omega_i}{2n_i c}\chi^{(2)}A_L A_s^*$$

where A_s and A_i are called the signal and idler fields. The solution for the signal field takes the

following form:

$$A_s(z) = A_s(0)\cosh\left(\sqrt{|\kappa|^2|A_L|^2}z\right) - i\sqrt{\frac{\omega_s n_i}{\omega_i n_s}}A_i^*(0)\sinh\left(\sqrt{|\kappa|^2|A_L|^2}z\right) \quad [35]$$

where

$$|\kappa|^2 = (\omega_s\omega_i/4n_s n_i c^2)|\chi^{(2)}|^2$$

The stimulated parametric interaction provides gain for input signal and idler waves. For large values of $\sqrt{|\kappa|^2|A_L|^2}\,z$ the gain becomes exponential in character, similar to the gain in SRS and SBS. When no incident signals are supplied, the source terms in [35] are the vacuum fluctuations at the signal and idler frequencies. Stimulated parametric fluorescence is the basis for parametric oscillators that provide tunable coherent radiation over a wide wavelength range in the visible and infrared.

Stimulated Photorefractive Scattering

Stimulated light scattering also takes place in photorefractive materials. These materials include crystals such as $LiNbO_3$, $BaTiO_3$, $Bi_{12}SiO_{20}$, and others. They possess a unique property of photo-induced index change with very low light intensities. Some of the potential applications include real-time holography, information recording, phase conjugation, parallel signal processing, and ultrafast optical gating. The nonlinearities leading to stimulated scattering are very large and can produce a significant gain in millimeter thick crystals. In order for amplification of scattered light to take place the nonlinearities have to have a nonlocal character due to diffusion or drift fields. Figure 4 shows a two-beam interaction effect in a photorefractive medium leading to stimulated scattering. Solid lines represent intensity maxima of interfering light beams while dashed lines represent diffused index change maxima. The fact that the two do not overlap allows amplification of the signal beam, with the maximum gain occurring at $\pi/2$ phase shift.

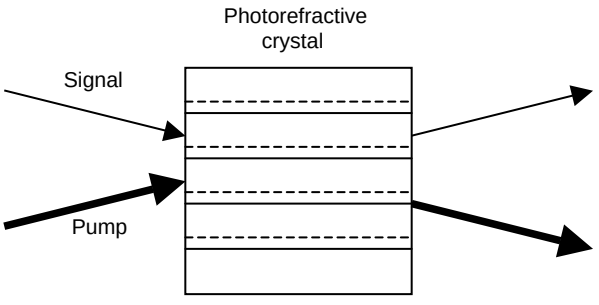


Figure 4 Stimulated photorefractive scattering diagram showing incident pump and signal beams, amplified signal beam, and an index grating formed in the photorefractive crystal.

Stimulated Compton and Thomson Scattering

Light scattering from free electrons is described by Thomson scattering in the low-energy limit, with a cross-section of $(8\pi)/(3)((e^2)/(4\pi\varepsilon_0 mc^2)^2)$, and by Compton scattering in the relativistic limit. Stimulated Thomson or Compton scattering is the source of light emitted from free electron lasers. In these interactions, the electron scatters incoming radiation in its rest frame to a backscattered wave. The ponderomotive force resulting from the interaction with the scattered waves increases the periodic electron bunching, which in turn increases the efficiency of scattering. In some cases, the incoming electromagnetic wave is provided by a microwave source, while in others it arises from a static magnetic wiggler, which appears as an incident electromagnetic wave in the electron rest frame because of a relativistic Lorentz transformation. Free electron lasers are a source of tunable coherent radiation from the microwave region to the ultraviolet.

Self-focusing

In the presence of a strong laser field with intensity I, the refractive index of many materials changes according to the relation

$$n = n_0 + n_2 I$$

where n_2 is termed the nonlinear index. This change in refractive index gives rise to several effects, depending on the temporal and spatial properties of the light. In one type of interaction, a strong forward beam couples to two weaker beams that propagate at small forward angles, providing gain for the weaker beams. One of the weak beams interferes with the strong forward beam to form a phase grating through the nonlinear index, providing gain for the second weak beam. The interference of the second weak beam with the forward beam forms a second phase grating, providing gain for the first weak beam, thereby completing the coupling

for the stimulated process. This effect is similar to Stoke/anti-Stokes Raman scattering, except that the coupling is between waves propagating in different directions rather than between waves with different frequencies.

For interactions with an instantaneous relaxation time, the gain for waves propagating at a small angle to the forward direction is given by

$$g = \frac{K_\perp}{2k}\sqrt{K_{\max}^2 - K_\perp^2}$$

where $K_\perp = \sqrt{k_x^2 + k_x^2}$ is the transverse wavevector, corresponding to the angle

$$\theta = \frac{K_\perp}{k}$$

and

$$K_{\max} = 4k^2 n^{\mathrm{L}} n_2 I$$

is the largest transverse k-vector for which gain is observed. The maximum gain is given by

$$g_{\max} = \frac{K_{\max}^2}{4k}$$

and occurs for an angle corresponding to

$$K_\perp = \frac{K_{\max}}{\sqrt{2}}$$

When the physical interaction that causes the refractive index to change with light intensity has a nonzero relaxation time, gain can be observed at angles larger than that given by $K_{\max}$.

This stimulated four-photon effect can be interpreted as the initiating stage of self-focusing. In the self-focusing effect, a laser beam with a normal mode profile (more intense in the center than at the edges) creates a positive lens in the material, focusing the beam. As the beam intensifies, the power of the lens increases, causing the beam to focus more tightly and eventually coming to a catastrophic focus if other nonlinear effects do not limit the intensity. The four-photon stimulated effect described above results in the spread of k-vectors of the forward beam, corresponding to a reduction in its diameter.

Self-focusing is responsible for breakdown and damage in many solids. Self-focusing in high-power lasers is commonly suppressed by restricting transverse k-vectors to values well below those of maximum gain.

See also

Nonlinear Optics, Applications: Self-Focusing and Related Effects (Solitons and Multiphoton Absorption). **Nonlinear Optics, Basics:** Photorefraction. **Scattering:** Scattering Theory.

Further Reading

Bloembergen N (1965) *Nonlinear Optics*. New York: Benjamin.

Bloembergen N (1967) The stimulated Raman effect. *American Journal of Physics* 35: 989–1023.

Boyd RW (1992) *Nonlinear Optics*. New York: Academic Press.

Fabellinskii IL (1968) *Molecular Scattering of Light*. New York: Plenum.

Gunter P and Huignard J-P (eds) (1989) *Photorefractive Materials and their Applications II*. Springer-Verlag Topics in Applied Physics, vol. 62. Berlin: Springer-Verlag.

Kaiser W and Maier M (1972) Stimulated Rayleigh, Brillouin and Raman spectroscopy. In: Arecchi FT and Schulz-DuBois EO (eds) *Laser Handbook*, vol. 2, pp. 1077–1150. Amsterdam, The Netherlands: North Holland.

Milanovich F (1986) Stimulated Raman scattering. In: Weber MJ (ed.) *Handbook of Laser Science and Technology. Optical Materials*, vol. III. Boca Raton, FL: CRC Press.

Penzkofer A, Laubereau A and Kaiser W (1979) High intensity Raman interactions. *Progress in Quantum Electronics* 6: 55–140.

Rabin H and Tang CL (eds) (1975) *Quantum Electronics: A Treatise*. Part A, vol. 1. New York: Academic Press.

Raymer MG and Walmsley IA (1990) The quantum coherence properties of stimulated Raman scattering. In: Wolf E (ed.) *Progress in Optics*, vol. XXVIII, p. 183. Amsterdam, The Netherlands: Elsevier.

Reintjes JF (1984) *Nonlinear Optical Parametric Processes in Liquids and Gases*. New York: Academic Press.

Reintjes JF (1995) Stimulated Raman and Brillouin scattering. Handbook of laser science and technology, supplement 2. In: Weber MJ (ed.) *Optical Materials*, pp. 334–364. Boca Raton, FL: CRC Press.

Reintjes J and Bashkansky M (2001) Stimulated Raman and Brillouin scattering. In: Bass M (ed.) *Optical Society of America Handbook of Optics*, Ch. 18, 2nd ed., vol. IV. New York: McGraw-Hill.

Shen YR (1984) *Principles of Nonlinear Optics*. New York: Wiley.

Shen YR and Bloembergen N (1965) Theory of stimulated Brillouin and Raman scattering. *Physical Review* 137: A1787.

SEMICONDUCTOR MATERIALS

Contents

Amorphous Semiconductors
Band Structure Engineering
Dilute Magnetic Semiconductors
GaAs Based Compounds
Group IV Semiconductors, Si/SiGe
III-Nitrides
Large Gap II–VI Semiconductors
Lead Salts
Mercury Cadmium Telluride
Modulation Spectroscopy of Semiconductors and Semiconductor Microstructures
Quantum Dots
Type-II Quantum Wells and Superlattices

Amorphous Semiconductors

J I B Wilson, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Amorphous semiconductors comprise several classes of materials having a range of useful properties that have made many of them commercially important. Their special characteristics are derived from the particular techniques used in their synthesis, and include the ability of some to change their structure under the influence of heat or illumination. Many are photoconductive, making them valuable for photodetectors when large area devices are required (e.g., xerography), or photovoltaic, hydrogenated amorphous silicon being widely used in thin-film solar cells.

'Amorphous' means noncrystalline, thus encompassing glassy materials and disordered covalent-bonded materials. Glasses, such as the amorphous chalcogenides (glassy sulfides, selenides, and tellurides), melt to viscous liquids when heated, whereas disordered semiconductors like amorphous silicon are metastable thin films that vaporize or crystallize when heated. Two other groups of glassy semiconductors, the covalent-bonded II–IV–V_2 compounds and ionic-bonded oxides (e.g., $V_2O_5P_2O_5$) will not be discussed here. Amorphous organic semiconductors are receiving much interest now, because of their potential use in low-cost display devices, but are more logically treated as polymer materials.

The following sections explain the main preparation methods and the structures that result, the importance of defects and the influence these have on optical and electrical behavior, and some of the most important applications.

Amorphous Structure

The common property that distinguishes all amorphous materials from crystalline ones is the absence of long-range order in their atomic structure. These disordered solids have no periodic or translational symmetry. The term is sometimes loosely applied, and it should be noted that elements and compounds that are able to form several structures might contain different phases. An example is 'hard carbon' which in thin-film form may contain tetrahedral amorphous carbon and hydrogenated amorphous carbon, as well as microcrystallites of graphite and diamond. Interestingly, the prediction in 1989 by A.Y. Liu and M.L. Cohen, that crystalline cubic-C_3N_4 would be even harder than diamond, has not been confirmed by experiment. Only a hard amorphous C_xN_y has been produced, even by nitrogen ion beam assisted sputtering of graphite, which it was thought would

have provided sufficient energy to form the close C–N bonds.

When the common crystalline semiconductors, Si and Ge, melt they become liquid metallic conductors. This is because the solid is formed by directional covalent bonds that produce tetrahedral packing of the atoms. In the liquid the bonds are weakened and a closer packed structure found in metals results. However a less ordered atomic arrangement results from condensing vaporized Si or Ge on to a cold substrate. In this case the atoms cannot move into their preferred positions for either tetrahedral or close packing arrays and instead are frozen almost wherever they land. Nonetheless, the arrangement of atoms is closer to that of the tetrahedral crystalline semiconductor than that of the metallic liquid. The radial distribution function of an amorphous semiconductor provides the numbers of atoms versus distance from an arbitrary atom. It shows that the nearest-neighbor distance is the same as in the equivalent crystalline material, although with some spread of values, but that the separation of atoms further away has a less well-defined value. Thus pure amorphous Si and Ge retain the directional covalent bonds between nearest neighbors but have distorted bond angles that make next nearest neighbors lie in different positions from the equivalent crystalline solid. Consequently there is only short-range order and the atoms form a continuous random network, a concept introduced by W.H. Zachariasen in 1932 for glasses. The deposition conditions generally lead to further disruptive effects that will incorporate small voids where the distortion is too large to allow a bond between atoms. These voids together with broken bonds give a defective network that is far from the ideal random tetrahedral 'lattice'.

Amorphous Semiconductor Energy Bands

A frequently asked question is 'Why can amorphous materials show semiconducting behavior?' since they do not have the periodic structure that leads to the allowed energy band scheme in crystalline semiconductors. However, note that metals do not become insulators when they melt, demonstrating that electrical conductivity does not require a periodic atomic structure. In fact, electron energy levels in solids may also be derived by considering chemical bonds when discrete atoms are brought closer together. The valence band in a semiconductor arises from the electron bonding orbitals and the conduction band from the antibonding electron orbitals. The forbidden energy gap (bandgap) is the separation between these two bands, which are only partially filled in semiconductors at room temperature. Thus amorphous semiconductors will have bonding and antibonding energy bands, and a more or less well-defined bandgap according to the defect content.

In fact, pure amorphous Si and Ge are extremely poor semiconductors due to an ill-defined bandgap and the large concentration of defects. Distorted bonds produce allowed electron energy levels that lie adjacent to the well-defined bands in crystals, producing tails of energy levels extending into the bandgap. Electrons that take up these 'tail state' energies are not entirely free to move through the solid but can only hop from one defective site to another. Additional defects arise from unsatisfied covalent bonds that 'dangle' in the spaces between the atoms. These produce electron energy levels within the forbidden gap (Figure 1). Indeed, their concentration is so high in amorphous silicon produced from a condensed silicon vapor, that they fix the position of the Fermi level at approximately the middle of the bandgap and the semiconductor has a high resistivity. The usual option of reducing a semiconductor's resistivity by adding small amounts of dopants from adjacent columns of the periodic table has little effect in this case. A donor atom from group V (e.g., phosphorus) may have all five valence electrons satisfied by a locally distorted arrangement of silicon atoms, instead of having only four bonded electrons and a loosely bound electron as in crystalline silicon. A similar picture exists for an acceptor atom from group III (e.g., boron).

Semiconductors containing group VI elements, S, Se, or Te, have two-fold coordination instead of the tetrahedral arrangement of the group IV elements, Si and Ge. The bonding between atoms in the solid is formed from four of the six p-electrons, which give

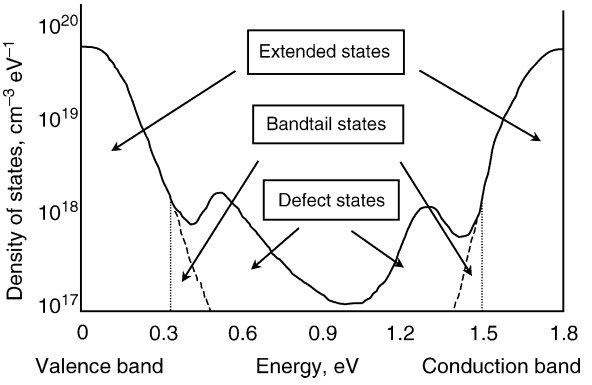


Figure 1 Schematic density of states for hydrogenated amorphous silicon.

rise to the bonding and antibonding energy bands of the solid, and the two remaining p-electrons from each atom form a non-bonding 'lone pair'. The lone pair energy band lies between the bonding and antibonding bands, and is the highest filled band in the solid, thus being the valence band.

Just as in amorphous Si or Ge, dangling bond defects will have energies between the valence and conduction bands, but in lone-pair semiconductors such disturbed energy states will now produce energy levels that fall within the lone pair band instead of within the bandgap. This makes it difficult to dope chalcogenide semiconductors, especially by donors, since the additional energy levels often lie within an existing energy band. Similar considerations apply to ternary chalcogenide alloys: these will have lone-pair bands arising from the group VI element, with additional energy levels from the other components (e.g., from group IV or V). However, amorphous chalcogenides will also contain three-fold coordinated chalcogenide atoms and these defects can associate in pairs or may interact with an adjacent two-fold coordinated atom. Defect centers with two oppositely charged chalcogens may form a 'valence alternation pair' (VAP).

So it may be seen that the semiconducting behavior of solids arises from the nearest-neighbor interactions of the atoms giving rise to allowed electron energy bands, and that long-range order is not necessary to produce the semiconductor's forbidden energy gap. Defects then produce localized energy levels that usually lie within the forbidden gap.

Preparation of Amorphous Silicon

Despite the early use of selenium as a photoconductive semiconductor in xerography, most amorphous semiconductors might have stayed as curious but technologically impractical materials, if an alternative preparation method to physical vapor deposition had not been discovered.

In the 1950s the preparation of high-purity single-crystal silicon was not yet established, and work at Standard Telecommunication Laboratories, UK, was attempting to use radio-frequency heating to produce silicon from low-pressure silane gas (SiH_4). It was noticed that a side effect was a deposit of resistive, noncrystalline material on cool parts of the container, close to unwanted plasma 'glow discharges'. In following up this observation, the technique was developed to produce a variety of other useful compounds from electrical discharges in appropriate gas mixtures, particularly silicon nitrides and oxides. With new scientific interest in amorphous semiconductors, the silicon coatings from silane were examined in more detail, and were shown to be significantly different from amorphous silicon produced by evaporating and condensing silicon. This new form was very photoconductive and could be doped n-type by adding phosphine to the silane. A more detailed study of doping led to the seminal papers by W.E. Spear and P.G. LeComber of Dundee University, UK, and to an explosive growth in the whole subject, once the potential commercial applications became apparent.

Despite many alternative synthesis methods, plasma chemical vapor deposition (CVD) remains the most widely used, and produces material that has not been surpassed in quality. The key features are a vacuum chamber into which a controlled flow of each of the gaseous sources is fed, and a heated plate on which the substrate is placed for coating. The decomposition of the gases is driven by an electrical discharge between two electrodes, which enables the gas temperature to be much lower than in conventional pyrolysis (thermal CVD). An important aspect of the process is the incorporation of hydrogen into the deposited material, which is thus more accurately known as hydrogenated amorphous silicon, a-Si:H. In the best material, hydrogen is mainly bonded singly as Si–H, but if the preparative conditions are not optimum then there may also be large amounts of $Si–H_2$ (as may be seen from the infrared absorption spectra). Alternative source gases usually contain hydrogen (e.g., Si_2H_6), and hydrogen may be added as a diluting agent for the doping gases (e.g., B_2H_6 or PH_3). Other precursors include fluorinated gases, but these have failed to give the predicted improvements in performance of the semiconductor in such applications as solar cells, and have greater chemical reactivity.

Since silane is spontaneously flammable in air, and the dopant gases are not only flammable but also very toxic, efforts have been made to produce a-Si:H by other means. Sputtering from undoped or doped silicon targets has been a popular choice, but it is still necessary to incorporate hydrogen into the deposited material, for instance by adding it to the usual argon sputter gas. This is because hydrogen serves to passivate unsatisfied 'dangling' Si bonds and to reduce the stresses that would occur in a pure silicon network. Photo-CVD, using ultraviolet light to drive the reaction without raising the temperature, can produce high-quality material but has the problem of unwanted silicon deposits on the illuminated reaction chamber window. Although conventional silane pyrolysis is not

suitable for producing a-Si:H, because the temperatures needed for useful deposition rates also remove much of the hydrogen if the substrate is too hot, it is possible to heat the gas separately from the substrate by means of a hot filament and so produce good-quality films.

Other hydrogenated silicon alloy semiconductors are synthesized by additions to silane of ammonia (giving amorphous silicon nitride, a-Si:N:H), or a hydrocarbon gas such as methane (giving a-Si:C:H), or germane (giving a-Si:Ge:H).

Preparation of Amorphous Chalcogenides

Unlike the tetrahedral group IV elements, useful members of this class of amorphous semiconductors may be produced by vacuum evaporation and condensation on to a relatively cool substrate. Compounds, such as As_2Se_3, do not evaporate congruently and so the deposit will change composition as the evaporation proceeds. More uniform compositions may be obtained by sputtering with argon, and glassy oxides may also be obtained in this way with oxygen added to the inert gas. Chalcogenide glasses may also be prepared by rapid cooling from a melt, or by chemical vapor deposition from heated vapor precursors. Since glasses are thermodynamically unstable, they will tend to transform to a more stable, sometimes crystalline structure. The rate for this process depends on the temperature and may be effectively too slow to observe.

Optical Properties of Amorphous Semiconductors

Optical absorption spectra of amorphous semiconductors clearly show a forbidden energy gap. By plotting a suitable function of the absorption coefficient, α, versus photon energy, $h\nu$, usually that given by J. Tauc, $(\alpha h\nu)^{1/2}$, the 'optical gap' may be determined from the intercept of the linear portion of the plot with the x-axis. Unlike crystalline semiconductors, amorphous ones have no k-selection rules to give indirect and direct transitions across the bandgap, thus a-Si:H behaves as a direct gap material. This 'optical gap' is greater than the bandgap determined from electrical conductivity measurements, since the latter will have contributions from both localized bandtail states (low electrical mobility) and extended states (high electrical mobility) (Figure 2).

The absorption spectra also have a portion at lower photon energies having an exponential dependence on photon energy, known as the Urbach tail. This is related to the structural disorder in the material and the tail states. Transitions involving defect bandgap states lie at even lower photon energy, and the density of defects can be determined from the weak absorption in this region of the spectrum, perhaps by using photothermal deflection spectroscopy to detect the weak optical absorption.

Photoluminescence also provides a probe of the defect content of many materials. Interpretation and identification of the actual defect (e.g., chemical impurity or structural imperfection) is aided by

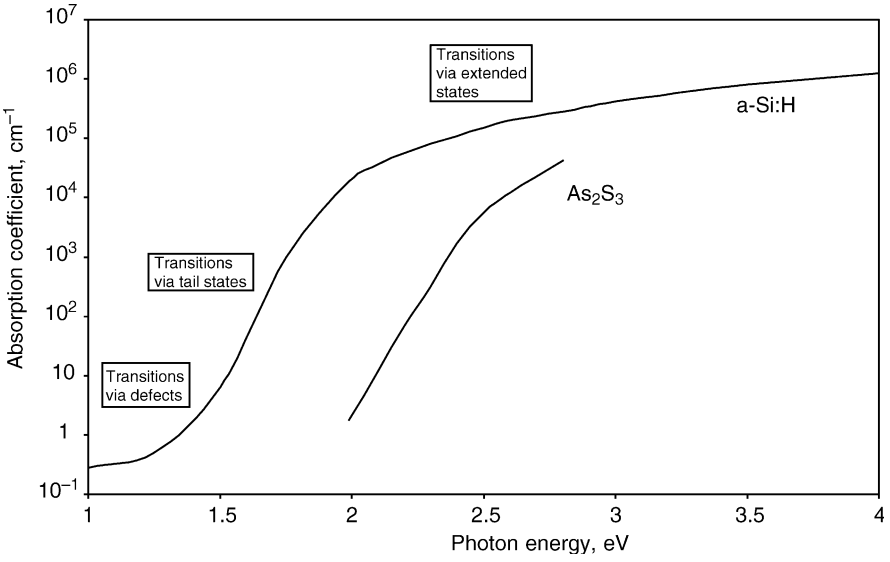


Figure 2 Typical optical absorption spectra of hydrogenated amorphous silicon and arsenic sulfide.

reference to the crystalline form. Both crystalline and amorphous forms of chalcogenide have a wealth of special effects, including reversible photoinduced bleaching and darkening, and laser-induced crystallization and amorphization. In addition the creation and dispersal of defects by appropriate illumination of some amorphous chalcogenides is accompanied by changes in the electron spin resonance (ESR) signal. Some of these effects have been the basis for applications in optical memories. More prosaically, chalcogenide glasses are important infrared transparent materials and amorphous chalcogenide thin films have had historical importance as the photoconductive layer in xerography, although here they are now being supplanted by less toxic amorphous silicon alloys.

Electrical Properties of Amorphous Semiconductors

The electrical conductivity of amorphous semiconductors is often controlled by the defects that are introduced by impurities or by structural flaws. For a-Si:H this provides alternative conduction mechanisms, via extended state electron and hole conduction (as in crystalline silicon), and via hopping of electrons or holes through localized energy levels within the bandtails or within the bandgap. The relative importance of each route depends on the temperature and on the density of states (DOS) in each energy range, as well as on the mobility of the charge carriers.

These different conduction mechanisms are clearly seen in the temperature dependence of the conductivity of a-Si:H and of many amorphous chalcogenides, since a semilog plot of conductivity versus $1000/T$ K will have portions with different slopes as each mechanism dominates a particular temperature range. Near room temperature, good-quality a-Si:H films in the dark exhibit a well-defined Arrhenius relationship with an activation energy that is the energy separation of the Fermi level from the conduction band extended states, electrons dominating the conductivity. The conduction band extended states do not have the high electron mobility found in crystalline silicon, due to the disruption of long-range order and the consequent increased electron scattering. Typical electron mobilities in a-Si:H are $\sim 1\,\mathrm{cm^2\,v^{-1}\,s^{-1}}$, and hole mobilities are even lower.

The band tail states have even worse carrier mobilities, since movement is by hopping to vacant sites that might lie at some distance, and hopping is unlikely between sites that have a large energy difference. (The process has an inverse exponential dependence on the energy difference between the initial and final states.) Hopping of carriers between states that lie within the bandgap is a significant contribution to the conductivity only at low temperatures, unless the density of such defects is high. In materials having weakly localized states, nearest-neighbor hopping is supplanted by variable range hopping, as carriers prefer to hop to a more distant site if it has a smaller energy difference from the starting site than do the neighboring sites. As the temperature is lowered, this hopping distance will increase and the conductivity has N.F. Mott's characteristic $\exp(-B/T^{1/4})$ dependence on temperature, T. Lower-quality films deposited by plasma CVD on to unheated substrates tend to be p-type due to hopping conductivity through the high density of defect states. The link between substrate deposition temperature and material quality is caused by the way in which hydrogen is incorporated into the growing material: high temperatures allow only singly bonded hydrogen, which passivates dangling bonds thus removing bandgap defect energy levels. Low temperatures also allow hydrogen to be incorporated as $Si-H_2$ and this disrupts the Si–Si lattice bonding, producing more defect levels. Confirmation of this effect is obtained from sputtered amorphous silicon, with and without hydrogen added to the argon.

When a-Si:H is illuminated it may show strong photoconductivity, depending on the preparative conditions. Free carriers are generated by the illumination but may fall into defect states (energies within the band gap) where they can be released or can recombine with the oppositely charged carriers. The detailed response to light (wavelength, intensity, duration) and temperature thus depends on the distribution and concentration of defects. Luminescence effects are similarly influenced. Most amorphous chalcogenides are only weakly photoconductive due to their high density of gap states, those near the middle of the gap (e.g., VAPs) being highly effective recombination centers.

The electrical properties of other hydrogenated amorphous tetrahedral semiconductors based on silicon are determined by their defect populations, and in general they are less well behaved. a-Si:Ge:H and a-Si:Sn:H have reduced bandgaps and a-Si:C:H and a-Si:N:H have increased bandgaps, which offer many possibilities for multilayer device design, provided that their normally high defect contents can be reduced or controlled.

The study of glassy semiconductors was stimulated in the 1960s by the discovery of electrical switching in sputtered multicomponent thin films.

These simple devices changed from a nonconducting to a conducting state when the applied electric field exceeded a certain threshold, and returned to their initial state when the current fell below a certain value. A variety of other switching types was discovered, especially under pulsed operation, including those having a memory effect based on the reversible crystallization and vitrification of a narrow filament between the contacts. A theoretical model of these devices comprises both electronic and thermal effects, which are responsible for initiating the switching event and for memory effects. A practical problem has been the variable jitter in the switching time.

Applications

Amorphous silicon alloy films are valuable as the active layers in thin-film photovoltaic cells, two-dimensional optical position detectors, linear image sensors (optical scanners), and thin-film transistors used in liquid crystal display panels. They also have uses as antireflection coatings and planar optical waveguides. Amorphous chalcogenide films have been used as electrical switches and memory devices, for optical data storage, and as dry photoresists and electron beam resists. Both types of amorphous material have been used in xerography and medical X-ray imaging detectors.

Amorphous silicon solar cells have power conversion efficiencies of ~12% for the most complicated structures. These are tandem cells that use different alloys (including a-Si:C:H) for the various layers, in order to enhance effective absorption of the solar spectrum. A serious drawback of using amorphous semiconductors in electronic devices is their short carrier diffusion length. In solar cells this is circumvented by ensuring that there is a built-in field to separate photogenerated electron–hole pairs before they can recombine. Because doped amorphous silicon alloys have high defect densities, it is difficult to make the effective pn junctions that provide such built-in fields in conventional crystalline silicon cells. Instead, amorphous silicon cells use pin structures, where the i-layer is effectively undoped and provides an extended electric field between the p-i and i-n junctions. Long periods of illumination increase the dark current in these devices, as additional defects are generated. This 'Staebler–Wronski effect' is believed to be associated with hydrogen movement at a passivated dangling bond (Si–H) and a nearby weak Si–Si bond. It is a cause of slow degradation in unstabilized solar cells although the change may be reversed by thermal annealing.

In contrast to 'minority carrier' photovoltaic devices, thin-film transistors necessitate the control of majority carriers alone. To address each pixel in a liquid crystal display, a switching element is needed and simple thin-film transistors fulfill this role. Although low carrier mobilities in amorphous semiconductors restrict the current passed and switching speeds, the requirement for coating large area glass panels at low temperatures make a-Si:H with a-Si:N:H gate dielectric almost the only options. Improved performance is possible by using pulsed laser irradiation to change the amorphous semiconductor to polycrystalline material.

Further Reading

Brodsky MH (ed.) (1985) *Amorphous Semiconductors*, 2nd edn. Berlin: Springer-Verlag.

Elliot SR (1984) *Physics of Amorphous Materials*. London: Longman.

EMIS Datareviews Series No.1 (1989) *Properties of Amorphous Silicon*, 2nd edn. London: INSPEC.

Joannopoulos JD and Lucovsky G (eds) (1984) *The Physics of Hydrogenated Amorphous Silicon*. Topics in Applied Physics, vols 55 and 56. Berlin: Springer-Verlag.

Journal of Non-Crystalline Solids (2000) *Proceedings of the 18th International Conference on Amorphous and Microcrystalline Semiconductors (1999)*, vol. 266–269. This journal has published the proceedings of most of these well-established annual meetings dealing with all aspects of amorphous semiconductors.

Kramer B and MacKinnon A (1993) Localization – theory and experiment. *Reports on Progress in Physics* 56: 1469–1564.

Luft W and Tsuo YS (1993) *Hydrogenated Amorphous Silicon Alloy Deposition Processes*. New York: Marcel Dekker.

Madan A and Shaw MP (1988) *The Physics and Applications of Amorphous Semiconductors*. San Diego, CA: Academic Press.

Mott NF (1985) Electrons in non-crystalline materials – the last twenty five years. *Contemporary Physics* 26: 203–215.

Pankove JI (volume ed.) (1984) Hydrogenated amorphous silicon. In: Willardson RK and Beer AC (series eds) *Semiconductors and Semimetals*, vol. 21, Parts A–D. Florida: Academic Press.

Zallen R (1983) *The Physics of Amorphous Solids*. New York: Wiley.

Band Structure Engineering

C R Pidgeon, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Direct gap semiconductors have many unique properties which are advantageous for optical and electronic applications, and in particular for detectors and emitters of infrared radiation associated with interband optical transitions. Gallium arsenide, GaAs, based structures have principally been used in the near-infrared, and so-called narrow gap semiconductors in the mid-infrared (MIR) spectral regions. More recently systems based on intersub-band transitions have shown great promise for MIR devices, with the obvious advantage for some applications that they can be constructed from large gap, better controlled materials (mainly InP- and GaAs-based to date) and have a peak operating frequency which is determined simply by the design of the particular quantum structure. Both approaches have been substantially advanced by band structure engineering techniques, reducing Auger losses and phonon losses, respectively, which is the subject of the present work.

The classic examples of narrow gap materials are the III–V semiconductors (InSb, InAs, and their related alloys), II–VI semiconductors (mainly $Hg_{1-x}Cd_xTe$), and the lead salt alloys. As a direct result of the small energy gap, these materials have a very low electron effective mass, very high electron mobility, even at room temperature, and a high saturation velocity. In addition they have a high ratio of stimulated to spontaneous transition probabilities and a large nonlinear susceptibility. However, they have the disadvantage of a high intrinsic carrier density at room temperature and thermal generation rate, so that a fundamental limitation to the performance of devices at elevated temperatures (preferably room temperature) is nonradiative losses resulting from Auger recombination processes. One solution is to cool these systems to 80 K or below which involves expensive and bulky cooling methods such as Stirling-cycle coolers, negating their advantages over other, otherwise inferior, devices. In the past 10 years there has been an intensive effort to circumvent this limitation by quantum structure and so-called 'band structure engineering' techniques made possible by epitaxial growth techniques, whereby the Auger processes are suppressed. Since the emphasis of the band structure engineering work has been mainly towards the optimization of III–V semiconductor lasers rather than detectors, we reflect that emphasis here. Although developed more recently than the bipolar diode lasers, the unipolar quantum cascade (QC) lasers provide perhaps the best example of systems that can be conveniently optimized by band structure engineering techniques. In this case, phonon processes dominate so that the design is to do with phonon suppression. Thus, we are concerned principally with suppression of Auger processes for interband devices, and phonon processes for intraband devices.

Nonradiative Processes and Band Structure Engineering

Suppression of Auger Recombination

Electron recombination processes are dealt with in detail elsewhere in the Encyclopedia (*see* **Semiconductor Physics:** Recombination Processes) but we briefly summarize the main points here. It is well known that there are three principal interband recombination mechanisms by which thermodynamic equilibrium is established in semiconductors: Shockley–Read recombination, radiative recombination, and Auger recombination. These are shown schematically in Figure 1. The Shockley–Read recombination occurs via extrinsic energy levels produced by lattice defects and impurities, Figure 1a, and can in principle be controlled by the procedure used to grow the material; hence it does not provide a fundamental limit to the

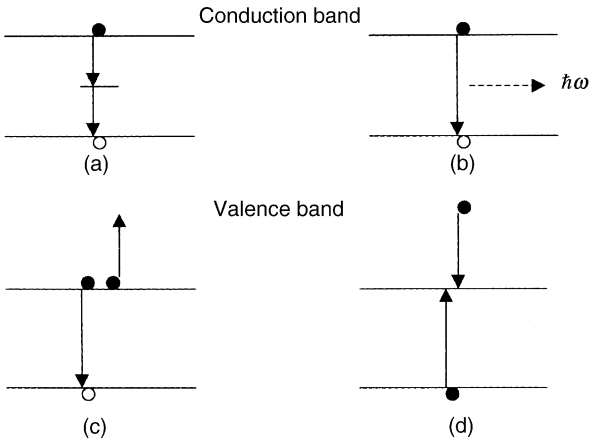


Figure 1 Electronic processes in semiconductors: (a) Shockley–Read recombination; (b) radiative recombination; (c) Auger recombination; (d) impact ionization.

carrier lifetime. The radiative and Auger mechanisms (Figure 1b,c, respectively) are, however, fundamental and are determined by the electronic structure of the material. The former is central to the operation of light-emitting diodes and is approximately inversely proportional to the electron and hole concentration. Optimum detector operation is achieved in materials that are limited by the radiative lifetime mechanism. Thus, in both cases it is desirable to minimize Auger processes.

By contrast with the two-particle radiative recombination process, Auger recombination is a three-particle interaction. For example in one so-called CHCC (or Auger-1) process an electron collides with a second electron; the first electron recombines with a hole in the valence band and the energy produced by the electron–electron collision excites the second electron higher up into the conduction band. This second electron eventually de-excites by phonon emission, and can cause observable heating of carriers by an Auger type of up-conversion. However, for the present article we are only interested in the first part of the process, so we do not consider the heating further here. The Auger process is the inverse of impact ionization where an energetic conduction electron collides with a valence electron, thereby creating an electron–hole pair as shown in Figure 1d. The final state consists of two low-energy electrons and one hole. Analogously, in Auger recombination the initial state has three particles (e.g., two electrons and one hole) and the final state one (e.g., one electron). Due to this three-body nature, the collision probability (or transition rate) per unit volume, R, varies with the cube of the carrier concentration and (if $n = p$) may be written $R = Cn^3$, where n is the electron density and C is referred to as the Auger coefficient. Note that the Auger rate is sometimes defined per carrier as $R' = Cn^2 = 1/\tau_A$, where τ_A is the Auger lifetime.

Narrow gap III–V semiconductors have the familiar three-band (so-called Kane energy band structure – (*see* **Semiconductor Physics:** Band Structure and Optical Properties) consisting of an s-like conduction band separated by a small energy gap from a triply degenerate p-like valence band, part of whose degeneracy is raised by spin–orbit splitting as shown in Figure 2. Thus, a large number of permutations of Auger processes become possible for n- and p-type semiconductors. Those which are initiated by a collision between two electrons are referred to as conduction, or n-n, Auger processes (e.g., the CHCC process), and those by a collision between two holes as valence, or p-p, Auger processes. It can be shown that the n-n process (so-called CHCC or Auger-1) illustrated in Figure 3a is dominant in bulk n-type and the p-p process

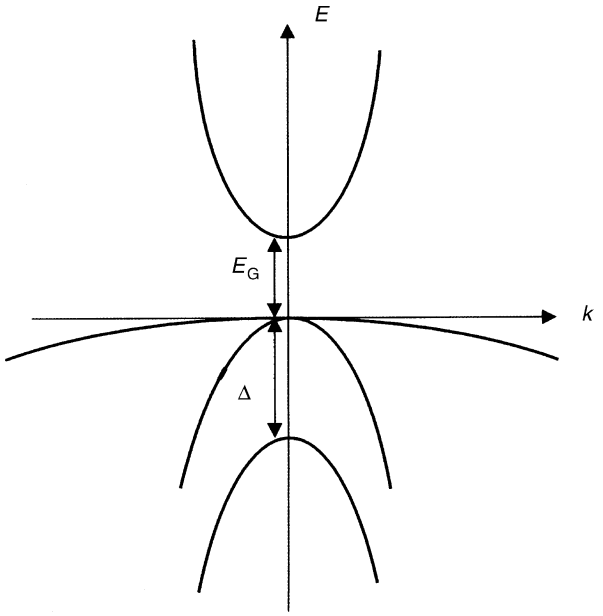


Figure 2 Schematic diagram of the conduction and valence band structure of InSb showing the energy bands around the Γ-point ($k = 0$) that are involved in Auger transitions.

(so-called CHLH or Auger-7) shown in Figure 3c is dominant in bulk p-type semiconductors, and so these have received the most attention. For larger gap (shorter wavelength) semiconductors there is sometimes a significant resonance between the bandgap and the spin–orbit splitting and the p-p process (CHSH) shown in Figure 3b becomes important. Clearly these simple descriptions require extension when the more complex band structure of strained layer superlattice quantum systems is considered.

With certain approximations, some analytic solutions for the Auger recombination rates are possible. This model is, while not appropriate for higher levels of band structure engineering, very important for giving an insight into how Auger effects should be studied, and how they themselves might be engineered. The approximations generally used are parabolic bands and Boltzmann statistics, but some further analytic results may also be obtained when these restrictions are relaxed. Bloch functions and Fermi's golden rule are utilized (*see* **Semiconductor Physics:** Recombination Processes) to describe the particle states, so obtaining the general expression for the Auger rate per unit volume. Energy and momentum conservation provide strong constraints on the resulting integral, and give rise to a considerable reduction in the Auger transition rate. However if, for example, the momentum conservation constraint is relaxed due to occupied levels existing with a broad range of momenta, such as at some band extremum with a

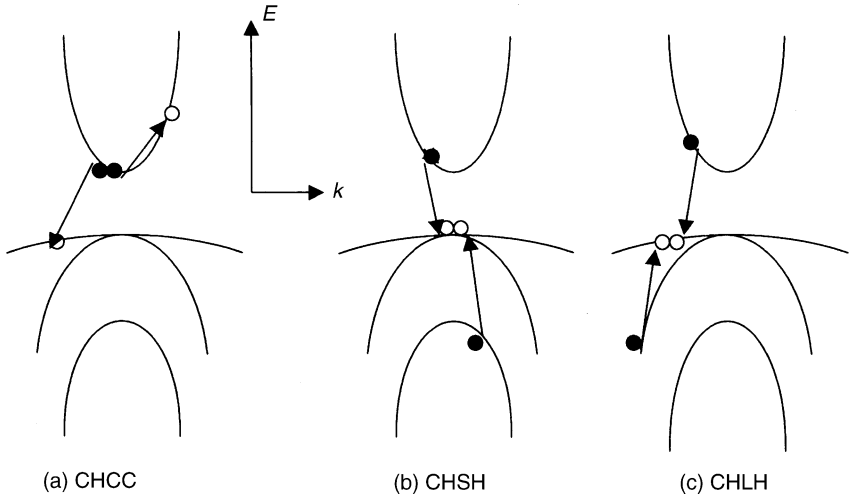


Figure 3 Important Auger recombination mechanisms for mid-infrared (MIR) semiconductor detectors and lasers: (a) CHCC or Auger-1; (b) CHSH; (c) CHLH or Auger-7.

large effective mass, then the Auger rate will be strongly enhanced. Or to put it the other way round, the Auger rate can be reduced by engineering the band structures such that the heavy hole effective mass is reduced (e.g., by stress – see below) to approach that of the conduction band. Furthermore, since we are concerned with nonradiative recombination across the bandgap with energy E_g, any energy levels (again band extrema in particular) separated from the conduction or valence band edges by E_g are particularly susceptible to the excitation part of the Auger process. The Auger rate may therefore be reduced by designing the structure so that there are no band extrema at energies E_g from the conduction and valence band edges.

Interband (Bipolar) Lasers

The technology of band structure engineering, which includes superlattices, quantum confinement, and intentional incorporation of strain, has now become sufficiently advanced that it can be employed in a routine manner. As originally proposed, the modified band structure of strained III–V quantum-well structures can lead to significant benefits for diode laser performance, including reduced threshold current density, improved efficiency, and enhanced dynamic response and speed. Predicted advantages have been demonstrated, although for near-infrared devices these have not been quite as substantial as originally hoped because of the presence of phonon-assisted Auger processes which are not so susceptible to band structure engineering techniques. Strained layer lasers are now commercially available at a variety of MIR wavelengths, where the dominant nonradiative loss is via the direct interband Auger scattering routes.

From the point of view of the present article the main contribution of 'designed' strain splittings and quantum confinement is in the reduction of the Auger coefficient itself, by removing states that easily satisfy the momentum and energy conservation conditions required for the process to take place. However, in addition to this there can be a substantial reduction in the threshold carrier injection level required for population inversion in a laser. In real III–V semiconductors the heavy hole valence band mass, m_h, is so much greater than the electron mass, m_e, that the usual laser picture of degenerate electron and hole distributions does not apply. At normal carrier injection levels the quasi-Fermi level of the holes is above the top of the valence band resulting in a classical heavy hole distribution. Thus, the lower laser levels are almost completely filled with electrons, by contrast with the ideal situation of two mirror bands. This leads to an injection level for the laser threshold of greater than 2:1 required for the actual case compared to the ideal case. Since the Auger recombination rate per unit volume is approximately proportional to the cube of n, this factor of ~ 2 reduction in injection level can produce a factor of almost an order of magnitude reduction in this loss mechanism. Large biaxial strains, either compressive or tensile, can lead to a considerable reduction in the hole mass and hence in the density of states at the top of the valence band, so that the laser characteristics may then approach those of an ideal semiconductor, with $m_e = m_h$.

A further important loss mechanism is free hole absorption of the heavy holes resulting from intervalence band transitions. This can be effectively

eliminated, and other improvements made to laser performance, by the introduction of a strained layer structure which raises the degeneracy of the light and heavy hole at $k = 0$ and reduces the mass of the heavy hole. One example of this comes just from the symmetry of the problem, where clearly the cubic symmetry of the p-like valence band is not matched to the uniaxial symmetry of the laser beam. Application of biaxial or uniaxial strain removes the cubic symmetry, changing the relative energies of the p_x, p_y, and p_z valence states, and making it possible for almost all of the injected holes to be in the states of the correct symmetry to take part in the lasing action. The one parasitic phenomenon in semiconductor lasers which is hardly affected by band structure engineering is spontaneous emission, which will always make some contribution to the current density required at threshold.

The band structure of a bulk unstrained direct gap tetrahedral semiconductor is shown again for comparison in Figure 2. The effect of biaxial tension and compression is shown respectively in Figure 4a,b. In each case the lowest conduction band is approximately parabolic near the zone center, but becomes anisotropic with the electron dispersion at small k. The axial strain breaks the cubic symmetry of the semiconductor, introducing a tetragonal distortion

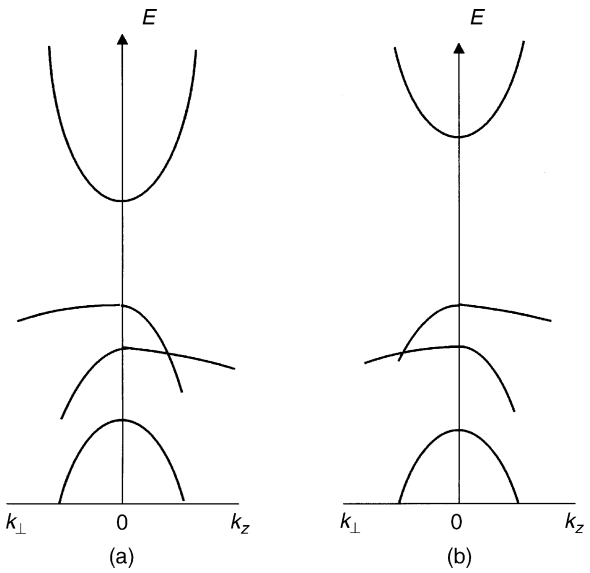


Figure 4 Schematic representation of the band structure of a direct gap tetrahedral semiconductor: (a) under biaxial tension the hydrostatic component of the tension reduces the mean bandgap, while the axial component splits the degeneracy of the valence band maximum and introduces an anisotropic valence band structure (the highest band being light along the strain axis, k_z and comparatively heavy perpendicular to that axis, $k_\perp$); (b) under biaxial compression the mean bandgap increases and the valence splitting is reversed.

which splits the degeneracy of the light and heavy hole states at the valence maximum, Γ, typically by about 60–80 meV for a 1% lattice mismatch. The resulting valence band structure is highly anisotropic (Figure 4a,b), with the band which is heavy along the strain axis, wavevector $k_\perp$, being comparatively light in the plane, along $k_\parallel$, and vice versa. The in-plane mass, $m_\parallel$, determines the density of states, and thus, under compression (Figure 4b), can be reduced according to the principles described above to improve conditions towards those for ideal laser operation: viz, (1) that there is only one band at the valence band maximum; (2) that it has an effective mass as close to that of the electron as possible; and (3) that above the Bernard–Duraffourg transparency condition, the polarization of the gain is anisotropic, with spontaneous emission and gain being suppressed along all directions except along the axis contributing to the laser beam.

Since the Auger recombination involves three carriers, the Auger current J_{NR} in a device with an undoped active region varies approximately as $J_{NR}(T) \propto C(T)n_{th}^3$, where $C(T)$ is the temperature-dependent Auger coefficient per unit volume and nth is the threshold carrier density. The influence of strain on Auger recombination has two aspects, as discussed above. First, J_{NR} is very sensitive to any reduction in n_{th} brought about by strain either through a decrease in h (heavy hole) mass or an increase in the optical transition strength. Secondly, strain may change the magnitude of $C(T)$ for the interband Auger processes described earlier (Figure 3). A simple calculation of Auger recombination in a quantum well heterostructure, assuming parabolic bands and Boltzmann statistics, shows that the coefficient increases exponentially with temperature as $C(T) = C_o\exp(-E_a/kT)$, where the activation energy, E_a, is approximately dependent on $(m_h)^{-1}$. Thus, both of these should be reduced by several orders of magnitude due to a reduction in the heavy hole mass, m_h, resulting from the strain effects discussed above.

Phonon Scattering: QWIPs and QC Lasers

Sub-band transitions are useful for MIR detectors and emitters because they make it possible to cover a broad range of MIR wavelengths using the same combination of large gap, InP-based and GaAs-based state of the art semiconductor alloys that are used for high-speed electronics and optical communications. They are the most advanced of the low-dimensional solids from the point of view of demonstrated detector performance. This superior control of materials results in high uniformity and has thus

allowed the demonstration of large 128×128 arrays with excellent imaging performance. Detectivities $D^* > 10^{10}\,\mathrm{cm\,Hz^{1/2}\,W^{-1}}$, low noise equivalent temperature differences, and excellent infrared video imaging can be achieved at 10 μm for temperatures $T = 10–80$ K.

Although the QWIP (quantum well infrared photodetectors) detectors showed promise for many years as alternatives to HgCdTe systems, the first intersub-band laser was not realized until the demonstration of the quantum cascade (QC) laser. Intersub-band (QC) lasers differ fundamentally from bipolar diode lasers in that they rely on only one type of carrier and on electronic transitions between conduction sub-band states arising from size quantization in semiconductor quantum heterostructures. These in-plane sub-bands have the same curvature (by contrast with the interband devices which have conduction and valence bands of opposite curvature), so that the joint density of states and corresponding gain spectrum (collision limited) is much narrower and approximately symmetric. Thus, the arguments which show that the Kane-type band structure is particularly disadvantageous for the overall Auger loss mechanism in interband diode lasers do not apply; the problems resulting from having $m_e \ll m_h$ are removed and the inversion is much less sensitive to thermal broadening of the electron distribution. In addition, there is no free hole absorption, and, since the component semiconductor materials are of comparatively wide bandgap, there is negligible intrinsic carrier concentration. However, the threshold current density tends to be rather large owing to the rapid nonradiative phonon relaxation of the population inversion.

In the original QC laser structure the electrons streamed down a potential staircase, sequentially emitting photons at the steps, consisting of coupled quantum wells in which the population inversion between discrete conduction band excited states was achieved by control of tunneling. The laser structure was grown with the $Al_{0.48}In_{0.52}As/Ga_{0.47}In_{0.53}As$ heterojunction material system lattice matched to InP. Electrical pumping was achieved by alternating 25 undoped coupled-well active regions with compositionally graded layers. The coupled-well region was essentially a four-level laser system, where a population inversion was achieved between the two excited states, $n = 3$ and $n = 2$. While the nonradiative phonon emission lifetime is short ($\tau_{32} \sim 4.3$ ps at the electric fields used in the tunneling process), strong coupling to an adjacent GaInAs well gives an even shorter lower state lifetime ($\tau_{21} \sim 0.6$ ps) from optical phonons with nearly zero momentum transfer between the strongly overlapped and closely spaced $n = 2$ and $n = 1$ states. This enables the achievement of a population inversion between these states. (Note that the separation between the $n = 2$ and $n = 1$ states, E_{21}, is designed to be approximately equal to the optical phonon energy.)

The radiative efficiency of these QC devices is unavoidably low ($<10^{-3}$) because most of the electrons in the upper sub-band relax within ~1 ps through nonradiative optical phonon emission before radiative transitions with a lifetime in the nanosecond range have a chance to occur. An alternative configuration has been proposed, the type II interband cascade laser (T2ICL), which retains the advantages of cascaded tunneling injection and wavelength tuneability but eliminates the nonradiative phonon path between valence and conduction bands of opposite curvature. In addition, band structure engineering ideas discussed above are incorporated to suppress the unwanted Auger processes; the design results in a light in-plane hole mass and the removal of resonances between the energy gap and intervalence transitions.

See also

Semiconductor Materials: Group IV Semiconductors, Si/SiGe; Type-II Quantum Wells and Superlattices. **Semiconductor Physics:** Band Structure and Optical Properties; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures; Recombination Processes.

Further Reading

Faist J, Capasso F, Sirtori C, Sivco DL, Hutchinson AL and Cho AL (1994) Quantum cascade laser. *Science* 264: 553–556.

O'Reilly EP and Adams A (1994) Band structure engineering in strained semiconductor lasers. *IEEE Journal of Quantum Electronics* 30: 366–379.

Pidgeon CR, Ciesla CM and Murdin B (1997) Suppression of non-radiative processes in semiconductor mid-infrared emitters and detectors. *Progress in Quantum Electronics* 21: 361–419.

Yablonovitch E and Kane EO (1988) Band structure engineering of semiconductor lasers for optical communications. *Journal of Lightwave Technology* 6: 1292–1299.

Youngdale ER, Meyer JR, Hoffman CA, Bartoli FJ, Grein CH, Young PM, Ehrenreich H, Miles RH and Chow DH (1994) Auger lifetime enhancement in InAs-GaInSb superlattices. *Applied Physics Letter* 64: 3160–3162.

Dilute Magnetic Semiconductors

R R Gałązka, Polish Academy of Sciences, Warsaw, Poland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction, General Remarks

Semimagnetic semiconductors (SMSs) are a group of solids at the interface between semiconductors and magnetic materials. Semimagnetic semiconductors, also referred to as dilute magnetic semiconductors, are semiconductor-based solid solutions where some of the cations are replaced by transition metals or rare earth elements. The crystallographic structure of the semiconductor is conserved; the lattice constant is a function of composition. From a magnetic point of view SMSs are disordered magnetic materials, since magnetic atoms are randomly distributed in the cation sublattice of the semiconductor compound. Table 1 presents the most complete list of bulk ternary SMSs. Quaternary SMSs are also investigated (e.g., PbSnMnTe or CdMnSeTe).

Generally speaking, in SMSs there coexist two interrelated and interacting subsystems: mobile delocalized charge carriers and localized magnetic moments connected with paramagnetic ions. Electronic (mainly optical) properties of SMSs have been the subject of intensive studies since early in 1970. Due to the strong spin exchange interaction between mobile carriers and localized magnetic moments (the exchange constant of the sp-d interaction is of the order of 1 eV for II–VI SMSs) significant changes of the band structure and behavior of the carriers were observed. A number of new physical phenomena were discovered, such as giant Faraday rotation (now important in magnetic field sensors and other applications in, for example, the automobile industry based on this strong magneto-optical effect), the magnetic field induced metal–insulator transition, the bound magnetic polaron and quantum thermomagnetic oscillations. Investigations of the influence of the electron subsystem on the magnetic properties of SMSs are connected with the observation of a ferromagnetic phase in PbSnMnTe induced by carrier concentration.

Molecular beam epitaxy (MBE) grown GaAs with Mn opened a new chapter of III–V SMSs showing ferromagnetic behavior with the highest transition temperature $T_c = 110\,\mathrm{K}$. In 1987 the first paper devoted to layered structure and magnetic properties of low-dimensional (LD) SMSs was published starting intensive studies of superlattices and other LD structures made of SMSs.

Energy Band Structure, Optical and Magneto-optical Properties of SMSs

Replacing cations such as Cd or Hg with a paramagnetic ion such as Mn in the same crystallographic structure does not markedly disturb the semiconductor properties of the material. The energy gap changes (increases) but the conduction and valence bands conserve their symmetry and character as in nonmagnetic semiconductor mixed crystals. The spin momentum of paramagnetic ions is connected with the 3d or 4f shell, for transition metals or rare earth elements, respectively. The energy level of 3d or 4f electrons lies below the top of the valence band and thus has negligible influence on its shape or that of the bottom of the conduction band. Thus all basic semiconductor optical properties connected with the band symmetry and topology are the same as for a typical semiconductor. These are described, for example, in articles on the lead salts (MS641) involving direct optical transitions at the L-point of the Brillouin zone, and on mercury cadmium telluride (MS639) with direct optical transitions at the Γ-point. The only influence of the magnetic subsystems on electrons comes from the spin exchange interaction between mobile carriers and localized magnetic ions. This interaction can be represented by a Heisenberg term

$$H_{\mathrm{ex}} = \bar{\sigma} \sum_i \bar{S}_i J(\mathbf{r} - \mathbf{R}_i) \qquad [1]$$

where $\bar{\sigma}$ and $\bar{S}$ are spin operators of the band electron and magnetic ion, respectively, the summation is over all lattice sites occupied by the magnetic ions, and J is the exchange constant.

Assuming mean field and virtual crystal approximations this term can be rewritten in the form which is periodic with a lattice constant:

$$H_{\mathrm{ex}} = x\langle \bar{S} \rangle_{\mathrm{av}} \bar{\sigma} \sum_R J(\mathbf{r} - \mathbf{R}) \qquad [2]$$

where x is the molar fraction of magnetic ions, $\langle \bar{S} \rangle_{\mathrm{av}}$ is the average over all magnetic ions and is directly

Table 1 Composition and crystal structure of ternary bulk SMSs

Material	*Crystal structure*	*Composition range*
II–VI (Mn)		
$Zn_{1-x}Mn_xS$	ZB	$0 < x \leq 0.10$
	W	$0.10 < x \leq 0.45$
$Zn_{1-x}Mn_xSe$	ZB	$0 < x \leq 0.30$
	W	$0.30 < x \leq 0.57$
$Zn_{1-x}Mn_xTe$	ZB	$0 < x \leq 0.86$
$Cd_{1-x}Mn_xS$	W	$0 < x \leq 0.45$
$Cd_{1-x}Mn_xSe$	W	$0 < x \leq 0.50$
$Cd_{1-x}Mn_xTe$	ZB	$0 < x \leq 0.77$
$Hg_{1-x}Mn_xS$	ZB	$0 < x \leq 0.37$
$Hg_{1-x}Mn_xSe$	ZB	$0 < x \leq 0.38$
$Hg_{1-x}Mn_xTe$	ZB	$0 < x \leq 0.35$
II–V (Mn)		
$(Cd_{1-x}Mn_x)_3As_2$	tetr.	$0 < x \leq 0.18$
$(Zn_{1-x}Mn_x)_3As_2$	tetr.	$0 < x \leq 0.15$
IV–VI (Mn, Eu, Gd)		
$Pb_{1-x}Mn_xS$	RS	$0 < x \leq 0.05$
$Pb_{1-x}Mn_xSe$	RS	$0 < x \leq 0.17$
$Pb_{1-x}Mn_xTe$	RS	$0 < x \leq 0.12$
$Sn_{1-x}Mn_xTe$	RS	$0 < x \leq 0.40$
$Ge_{1-x}Mn_xTe$	RH	$0 < x \leq 0.18$
	RS	$0.18 < x \leq 0.50$
$Pb_{1-x}Eu_xSe$	RS	$0 < x \leq 0.04$*
$Pb_{1-x}Eu_xTe$	RS	$0 < x \leq 0.32$*
$Sn_{1-x}Eu_xTe$	RS	$0 < x \leq 0.013$
$Pb_{1-x}Gd_xTe$	RS	$0 < x \leq 0.11$*
$Sn_{1-x}Gd_xTe$	RS	$0 < x \leq 0.09$*
II–VI (Fe, Co, Cr)		
$Zn_{1-x}Fe_xS$	ZB	$0 < x \leq 0.26$*
$Zn_{1-x}Fe_xSe$	ZB	$0 < x \leq 0.22$*
$Zn_{1-x}Fe_xTe$	ZB	$0 < x \leq 0.01$*
$Cd_{1-x}Fe_xS$	W	$0 < x < 0.09$*
$Cd_{1-x}Fe_xSe$	ZB	$0 < x < 0.20$
$Cd_{1-x}Fe_xTe$	ZB	$0 < x \leq 0.06$*
$Hg_{1-x}Fe_xS$	ZB	$0 < x \leq 0.037$*
$Hg_{1-x}Fe_xSe$	ZB	$0 < x \leq 0.20$*
$Hg_{1-x}Fe_xTe$	ZB	$0 < x \leq 0.02$*
$Zn_{1-x}Co_xS$	ZB	$0 < x \leq 0.15$*
$Zn_{1-x}Co_xSe$	ZB	$0 < x \leq 0.10$*
$Zn_{1-x}Co_xTe$	ZB	$0 < x < 0.06$*
$Cd_{1-x}Co_xS$	W	$0 < x \leq 0.064$*
$Cd_{1-x}Co_xSe$	W	$0 < x \leq 0.082$*
$Cd_{1-x}Co_xTe$	ZB	$0 < x \leq 0.04$
$Hg_{1-x}Co_xS$	ZB	$0 < x \leq 0.02$
$Hg_{1-x}Co_xSe$	ZB	$0 < x \leq 0.047$*
$Zn_{1-x}Cr_xSe$	ZB	$0 < x \leq 0.005$*
$Zn_{1-x}Cr_xTe$	ZB	$0 < x \leq 0.003$*
$Cd_{1-x}Cr_xS$	W	$0 < x \leq 0.002$*

ZB – zinc blende, W – wurtzite, RS – rock salt, RH – rombohedral. Asterisks denote SMSs with the highest reported x value; tetr, tetragonal structure.

Reprinted in part from Gałązka RR (1995) Influence of electron subsystem on magnetic properties of semimagnetic semiconductors. In: Tróc R, Morkowski J and Szymczak M (eds) Proceedings of the International Conference of Magnetism, Warsaw 1994. *Journal of Magnetism and Magnetic Materials*, 113–116. Copyright Elsevier Science.

related to the magnetization of the sample:

$$M = N_0 g \mu \langle \bar{S} \rangle_{\mathrm{av}} \qquad [3]$$

N_0 is the number of unit cells per unit volume, g is the Lande factor for magnetic ions, and μ is the Bohr magneton.

The Heisenberg term must be added to the effective-mass Hamiltonian to solve the energy eigenvalue problem for SMSs. Realizing this was the turning point in understanding their properties and provided a basis for this group of alloys to be distinguished from other semiconductor mixed crystals. Since that time SMSs have become the subject of intensive studies in Europe and elsewhere.

The exchange interaction depends on the ion spin and the exchange constant. The spin of Mn is the highest for transition metals and equals five Bohr magnetons, and whilst a typical value for the exchange coupling between magnetic ions $J_{\mathrm{d-d}}$ is 10^{-3} eV, the exchange constant for electrons and holes with magnetic ions $J_{\mathrm{sp-d}}$ is about 1 eV in SMSs.

The exchange interaction is also strongly temperature and magnetic field dependent. As the macroscopic magnetization of a sample is proportional to the thermodynamic average value of the magnetic ion spin, to a good approximation we can replace the ion spin operators in the Hamiltonian by their average value, calculated or taken from measurements of the magnetization. From typical magnetic functions we can thus obtain information on the electron behavior.

In semiconductors, an external magnetic field acts on both the orbital motion and the spin of electrons producing Landau quantization and spin splitting, to an extent dependent on the effective mass of the carriers. The exchange interaction acts on their spin only and is, in effect, mass independent. Because of this, and also the large value of $J_{\mathrm{sp-d}}$, the band structure in turn changes drastically under the influence of an external magnetic field and depends strongly on temperature. From **Figure 1** we can see that the very typical structure of the degeneracy of the light and heavy hole valence bands disappears in SMS when a magnetic field is applied. The change is accompanied by a drastic increase of hole mobility like that observed in low-dimensional structures. In SMS such as $Hg_{1-x}Mn_xTe$ or $Hg_{1-x}Mn_xSe$ (for $x \leq 0.07$) (**Figure 2**), the magnetic field produces an overlap of the valence and conduction bands, an effect never observed in nonmagnetic semiconductors.

Impurity levels are also influenced by exchange interactions: the ionization energy of an acceptor

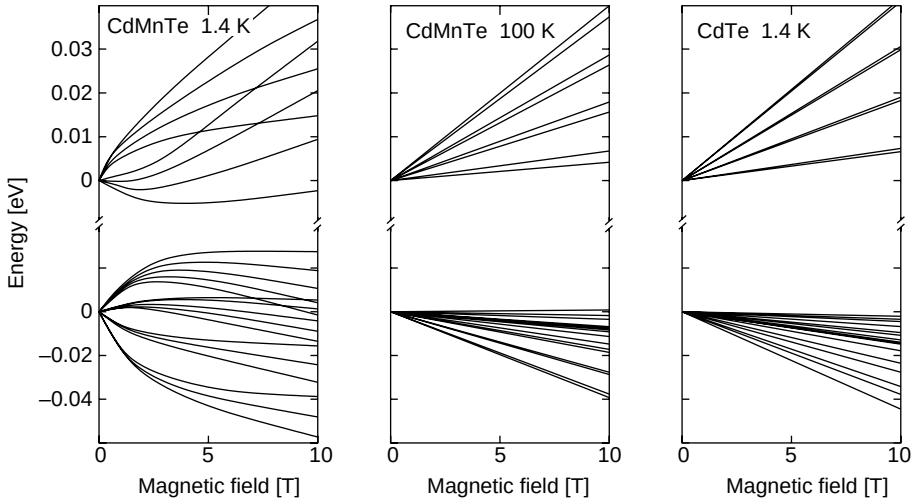


Figure 1 Conduction and valence band quantization in an external magnetic field for $Cd_{0.95}Mn_{0.05}Te$ and CdTe. Spin splitting, which is very weak in CdTe, is the main effect in CdMnTe. Notice the almost equal splitting of the heavy- and light-hole bands in CdMnTe. The influence of exchange interaction is visible even at 100 K. The picture of CdTe over this range of temperature is practically unchanged. Reproduced from Gałązka RR (1987) Semimagnetic semiconductors. *Europhysics News* 18(6): 90–92.

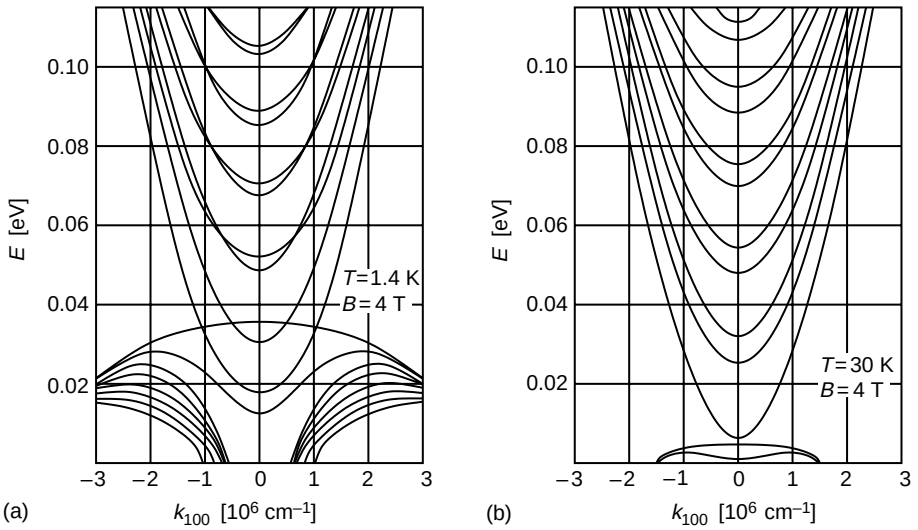


Figure 2 Energy band structure of $Hg_{1-x}Mn_xTe$, $x = 0.02$, calculated at 14 K (a) and 30 K (b) in a magnetic field of 4 T. After Gałązka RR (1978) Semimagnetic semiconductors based on HgMnTe and CdMnTe. Reproduced from Wilson BLM (ed) ICPS 14th, Edinburgh Institute of Physics, Conference series no. 43, 1978, p.133–140.

decreases and its wavefunction becomes anisotropic under the influence of a magnetic field. Both effects produce a giant negative magnetoresistance and a field-induced insulator–metal transition. Indeed all (particularly optical) properties susceptible to being changed by a magnetic field are very different in SMSs from those of standard semiconductors: the Faraday effect., Shubnikov–de Haas oscillations, interband magnetoabsorption, luminescence and spin flip Raman scattering (see below) become strongly dependent on temperature and magnetic field.

In the absence of a magnetic field, the average value of the ion spins is zero, the magnetization is zero too, and SMSs should behave as typical non-magnetic semiconductors. Whereas this is true for delocalized band electrons, when an electron is localized around an impurity, a bound magnetic

polaron can be observed. This concept was first introduced to explain the behavior of magnetic semiconductors (Eu chalcogenides) and then later was applied also to SMSs. A localized electron can produce a spontaneous magnetization within the range of its wavefunction via the exchange interaction. Local ferromagnetic ordering in this range produces a Stokes shift in spin-flip Raman scattering even in the absence of an external magnetic field and other related phenomena. Local fluctuations of magnetization and a detailed knowledge of electronic states are essential for a theoretical description of this effect. The bound magnetic polaron is a subtle example of the feedback between electronic and magnetic subsystems present in SMS.

Magnetic Properties

Magnetic properties of SMS connected with the exchange interaction between paramagnetic ions are similar to those of other dilute magnetic materials. The broad range of possible magnetic ion concentrations (up to $x = 0.80$ in $Zn_{1-x}Mn_xTe$ and $x = 0.70$ for $Cd_{1-x}Mn_xTe$) allows us to observe the evolution of magnetic interactions in the same crystallographic structure as a function of magnetic ion concentration. A continuous transition from diamagnetic behavior (for $x \leq 0.005$) of the host semiconductor, to paramagnetism, spin-glass (for $x \geq 0.02$) and finally antiferromagnetic order has been observed. Type III antiferromagnetic ordering is observed for $x > 0.6$ indicating that the magnetic elementary cell contains two elementary cells of the cation sublattice. From a crystallographic point of view the magnetic subsystem is disordered; the distribution of magnetic atoms on the cation sublattice is random.

A spin-glass-like state is observed in SMSs in the broadest range of paramagnetic ion concentrations ($0.02 \leq x \leq 0.60$) below and above the percolation threshold (Figure 3). Indeed the temperature dependence of the magnetic susceptibility below and above the transition temperature and their dynamic properties at low and high magnetic fields suggest rather the existence of two spin-glass phases arising from the competition between spin-glass and clustering behavior. Below the percolation threshold ($x \leq 0.17$ for fcc lattice) the mechanism responsible for a spin-glass phase is even less clear, although, for such a dilute system it must be long-range to produce freezing of the spins.

The antiferromagnetic order has been inferred from specific heat and magnetic susceptibility measurements, and additional information concerning this phase has come from neutron diffraction studies: only a certain fraction of the total number of magnetic ions (e.g., about 50% for CdMnTe, $x = 0.65$) is well ordered; the rest remain in a disordered spin-glass phase. Thus a mixed phase (antiferromagnetic and spin-glass coexisting together) rather than a truly antiferromagnetic phase is observed.

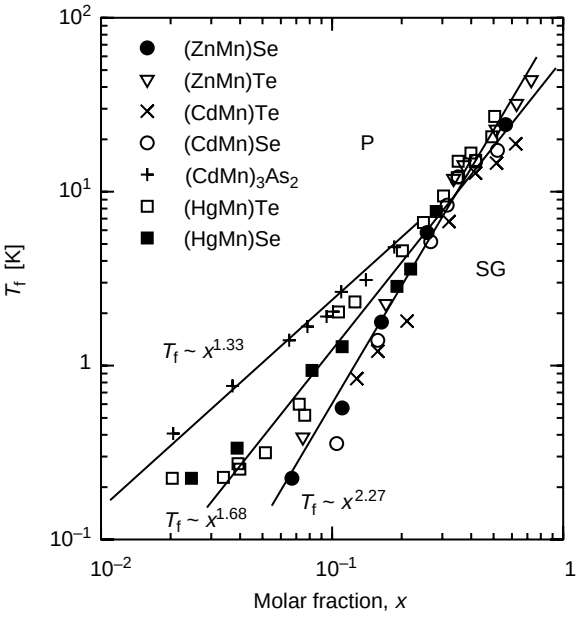


Figure 3 Magnetic phase diagram for the listed SMSs. P indicates the paramagnetic region, SG indicates the spin-glass region. Reprinted in part from Gałązka RR (1995) Influence of electron subsystem on magnetic properties of semimagnetic semiconductors. Reproduced with permission from Tróc R, Morkowski J and Szymczak M (eds) Proceedings of the International Conference of Magnetism, Warsaw 1994. *Journal of Magnetism and Magnetic Materials*, 113–116. Copyright Elsevier Science.

The influence of carrier concentration on magnetic properties has been extensively investigated in metallic alloys. Several IV–VI semiconductors like SnTe can exhibit quite high metallic-like carrier concentrations. In the quarternary alloy PbSnMnTe, the concentration of holes can be varied by annealing in the range 10^{20} to 5×10^{21} cm^{-3}. In $Pb_{0.25}Sn_{0.72}Mn_{0.03}Te$ an abrupt transition from the paramagnetic to ferromagnetic phase has been found for $p > 3 \times 10^{20}\,cm^{-3}$. The Curie temperature is a function of hole concentration, but the Mn ions conserve their magnetic moment of five Bohr magnetons. The magnetic phase diagram for this first ferromagnetic SMS (Figure 4) is thus three dimensional (T,x,n) in contrast to the

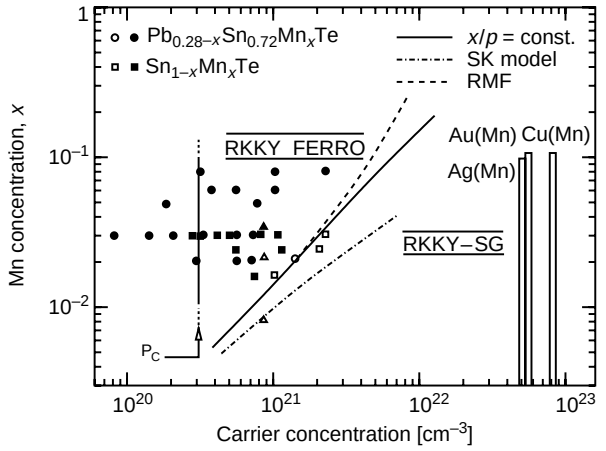


Figure 4 Magnetic phase diagram for PbSnMnTe (•) and SnMnTe (▪). The ferromagnetic phase is observed in the region above the critical carrier concentration p_c and the straight line defined by the equation $x/p = \text{const}$. Away from this region a spin-glass state is observed. Calculated phase boundaries between the ferromagnetic and spin-glass state are also shown for two theoretical models: SK: Sherrington–Kirkpatrick model and RMF: random mean field model. Canonical metallic spin-glasses are also shown. Reproduced from de Jonge WJM, Story T, Swagten HJM and Eggerkamp PJT (1992) Carrier-induced breakdown of ferromagnetism. *Europhysics Letters* 17: 631–636.

two-dimensional diagram (T,x) usually presented for magnetic alloys. Even a low concentration of spin polarized carriers when optically pumped can change the magnetization of HgMnTe owing to the strong exchange interaction of carriers with magnetic ions.

Low-Dimensional (LD) SMSs: Magneto-optical and Spintronic Effects

Modern technology such as MBE makes it possible to obtain quasi-3D SMSs in the whole range of x from 0 to 1. It also gives broad possibilities to create different combinations of semiconductor, SMS, and magnetic materials. In addition the dimensionality of LD structures influences the properties of both electronic and magnetic subsystems. Further there is almost always some strain in the system which also influences the properties of superlattices and other LD structures. Studies of LD SMSs not only extend the material base but open a new area of dimension-dependent phenomena and interactions in SMS systems.

Among many effects observed in LD SMS structures we will mention only a few phenomena, where exchange interaction is essential and which cannot be observed in nonmagnetic LD semiconductors. In all cases the giant (magneto-optical) Zeeman splitting caused by the exchange interaction plays a key role.

In superlattices made of ZnSe/ZnMnSe with a small amount of Mn, the band offset of conduction and valence bands is negligible at zero external magnetic field. In the presence of an external field the conduction and valence bands of ZnMnSe split creating a spin-superlattice because the spin-split states allow only one spin configuration. The possibility of the existence of such a spin superlattice was earlier suggested for SMSs and later experimentally observed.

If the superlattice, quantum well or heterostructure is made of SMS and non-SMS material (e.g. ZnTe/CdMnSe or CdTe/CdMnTe) one can exploit the drastic differences in the Zeeman splitting occurring in different layers to pinpoint the localization in space of specific electronic states. If one (hole or electron) or both states involved in optical transitions originate from the SMS layer, this state will show a very strong Zeeman splitting in the external magnetic field. Thus the SMS component can be used as a 'marker' in order to bring out the general properties of the wavefunction distribution.

Excitonic Zeeman splitting in non-magnetic quantum wells with SMS barriers can be applied to studies of the shape of interface profile and the influence of the growth conditions on the shape of the interface. If the interface is not completely flat, even a small fraction of magnetic atoms in the well close to the interface can significantly enhance the Zeeman exciton splitting in the quantum well. The sensitivity of this method allows one to detect the sharpness of an interface with one monolayer accuracy.

The MBE technique opens up possibilities of creating hybrid structures with SMSs, useful for spintronic devices such as spin filters, spin transistors and tunneling magnetoresistance sensors. Very importantly, it has been realized that SMSs are highly suitable to form the ferromagnetic spin injecting electrodes in such systems where simple metallic electrodes fail because of the large electrical resistance mismatch between the electrode and the semiconductor in which the spin is being processed. A further possibility is to fabricate hybrid metallic ferromagnetic–SMS structures. In such a case one can make use of the magnetic field exerted by the metallic ferromagnet positioned very close (on an atomic scale) to the SMS layer or quantum structure, thus affecting the spin splitting in the latter. One could use such a fringe field to shape the band edges in such a fashion that an additional localization of the carriers in a two-dimensional quantum well is induced

leading to formation of quantum wires or dots. In other applications a spin-coherence storage effect has been obtained for core electron polarization in Mn-doped ZnSe quantum wells. The optically pumped conduction electron spin polarization generates a nonequilibrium spin polarization of the paramagnetic core electron spins(3d in the case of Mn) through the core–conduction-electron coupling. This core electron spin polarization persists long after the conduction electron spin polarization has decayed away and presents an interesting avenue towards information storage.

The physics of SMSs is currently a well-established part of solid state physics. In the investigations of bulk crystals and LD structures many problems are solved and new problems are waiting further research. Particularly important seems to be investigation of layers and LD structures because of the broad technological possibilities and because of the new important factor (i.e., dimensionality) whereby spin-dependent phenomena can be intentionally regulated.

List of Units and Nomenclature

g	Lande factor for paramagnetic ion
H_{ex}	Heisenberg spin exchange Hamiltonian
J_{d-d}	exchange constant for coupling between paramagnetic ions (transition metals)
$J(\bar{r}-\bar{R})$	exchange constant
J_{sp-d}	exchange constant for electron or hole coupling with paramagnetic ions
LD	low dimensional
M	magnetization
MBE	molecular beam epitaxy
N_0	number of unit cells per unit volume
SMS	semimagnetic (diluted magnetic) semiconductor
$\langle\bar{S}\rangle_{av}$	average value of spin operator
T_c	Curie temperature
V_h	hybridization energy
x	molar fraction of paramagnetic ions
$\boldsymbol{\mu}$	Bohr magneton
$\overline{\boldsymbol{\sigma}}, \bar{S}$	spin operators

See also

Magneto-Optics: Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering. **Semiconductor Materials:** Lead Salts; Mercury Cadmium Telluride. **Semiconductor Physics:** Spin Transport and Relaxation in Semiconductors; Spintronics.

Further Reading

Bauer G, Pascher H and Zawadzki W (1992) Magneto-optical properties of semimagnetic lead chalcogenides. *Semiconductor Science and Technology* 7: 703–723.

Dietl T (1994) Magnetic semiconductors. Moss TS (ed.) *Handbook of Semiconductors*, vol. 3b, pp. 1251–1342. Amsterdam, The Netherland: North-Holland.

Dietl T (2002) Ferromagnetic semiconductors. *Semiconductor Science and Technology* 17: 377–392. (Special issue: Semiconductor Spintronics, guest ed Hideo Ohno).

Dobrowolski W, Kossut J and Story T (2003) Magnetic semiconductors–new bulk materials and low-dimensional quantum structures. In: Buschow KHJ (ed.) *Handbook of Magnetic Materials, II–VI and IV–VI*. vol. 15, pp. 289–378. Amsterdam: Elsevier North-Holland.

Furdyna JK (1991) In: Averous M and Balkanski M (eds) *Semimagnetic Semiconductors and Diluted Magnetic Semiconductors*. New York: Plenum.

Furdyna JK and Kossut J (eds) (1988) Semiconductors and Semimetals. *Diluted Magnetic Semiconductors*, vol. 25, 470 pp. San Diego, CA: Academic Press.

Gałązka RR and Kossut J (1980) Narrow-gap semimagnetic semiconductors. In: Zawadzki W (ed.) *Lecture Notes in Physics*, vol. 133, pp. 245–265. Berlin: Springer Verlag.

Gałązka RR and Kossut J (1982) Semimagnetic semiconductors. In: Madelung O, Schulz M and Weiss H (eds) *Landolt-Börnstein New Series*, vol. III/17b, pp. 302–318. Berlin: Springer Verlag.

Gałązka RR, Kossut J and Story T (1999) Semimagnetic semiconductors. In: Rössler U (ed.) *Landolt-Börnstein New Series*, vol. III/41B, pp. 650–721. Berlin: Springer Verlag (Update to chapter 5, Vol. III/17b).

Kossut J and Dobrowolski W (1993) Diluted magnetic semiconductors. In: Buschow KHJ (ed.) *Handbook of Magnetic Materials*, vol. 7, pp. 231–305. Amsterdam, The Netherlands: North Holland.

Kossut J and Dobrowolski W (1997) Properties of diluted magnetic semiconductors. In: Capper P (ed.) *Narrow-Gap II–VI Compounds for Optoelectronics and Electromagnetic Applications*, pp. 401–429. London: Chapman and Hall.

Kossut J (2001) Low dimensional structure of diluted magnetic (semimagnetic) semiconductors – a subjective review. *Acta Physica Polonica, A* 100 (supplement): 111–138.

Ohno H, Matsukura F and Ohno Y (2002) Semiconductor spin electronics. *Japan Society of Applied Physics (JSAP International)* 5: 4–13.

Story T (2003) Optoelectronic properties. In: Monasreh MO (series ed.) *Optoelectronic Properties of Semiconductors and Superlattices* vol. 18, Khokholov D. (volume ed.) *Lead Chalcogenides: Physics and Applications, Semimagnetic Semiconductors Based on Lead Chalcogenides*, pp. 385–425. New York, London: Taylor and Francis Books Inc.

GaAs Based Compounds

J J Finley and J P R David, University of Sheffield, Sheffield, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Background

Gallium Arsenide (GaAs) is an important semiconductor that has come to dominate the field of optoelectronics by virtue of its favorable electro-optical properties and the ease by which it can be controllably modified by extrinsic means; combining it with its large family of related alloys ($Al_xGa_{1-x}As$, $In_xGa_{1-x}As$, $GaAs_xP_{(1-x)}$, $(In_xGa_{(1-x)})_yAsN_{1-y}$) and via the growth of hetero-structures with reduced dimensionality (Quantum Wells-2D, Wires-1D, and Dots-0D). In such low dimensional semiconductor nanostructures, quantum mechanical effects are directly exploited to customize and dramatically enhance electro-optical properties. Such investigations of the fundamental electro-optical properties of GaAs-based compounds has resulted in the emergence of several new device concepts such as the Quantum Cascade laser, led to the discovery of new physical phenomena like the fractional Quantum Hall effect, and may pave the way for an entirely new class of opto-electronic devices with true quantum functionality.

In the following, we explore the fundamental physical and electro-optical properties of GaAs and its related alloys and illustrate the considerable impact this remarkable family of materials has had on modern semiconductor devices.

Fundamental Physical Properties

The crystal structure of GaAs (Figure 1 – inset) consists of two interpenetrating face centred cubic lattices that accommodate the group III and V atoms respectively. The sub-lattices of Ga and As are shifted from each other by $a/4$ along the [111] body diagonal, where a is the lattice constant, forming the well-known zincblende crystal structure. For stochiometric GaAs, the lattice constant is $a = 0.56536(1)$ nm at room temperature which translates to a Ga–As nearest neighbor separation (bond length) of 0.2308 nm.

Figure 1 summarizes the variation of the lattice constant for a number of GaAs related III–V semiconductor alloys, plotted as a function of their fundamental electronic bandgap. Particularly relevant is AlAs, which also crystallizes in the zincblende form and has a lattice constant (a(AlAs) = 0.5660(2) nm) almost identical to that of GaAs ($\Delta a < 1\%$). The sole fact that the $Al_xGa_{(1-x)}As$ alloy can be almost perfectly lattice matched to GaAs over its entire compositional range, whilst exhibiting near continuously variable electronic and optical properties, is the principal reason for the key role GaAs-related materials play in semiconductor opto-electronic devices. We continue by investigating the electronic and optical properties of the $Al_xGa_{(1-x)}As$ family of materials and discover why they are particularly suited to opto-electronic device applications.

Summary of Electronic Properties of GaAs and $Al_xGa_{1-x}As$

The optical properties of semiconductors are determined principally by inter-band radiative transitions involving electronic states in the vicinity of the conduction (CB) and valence band (VB) extrema. Consequently, in the following we focus on electronic properties of GaAs close to these points and substantiate our discussion using experimental data obtained from optical spectroscopy.

GaAs Bandstructure

GaAs is a direct-gap semiconductor ($E_g = 1.42$ eV at 300 K), both the conduction band minimum and

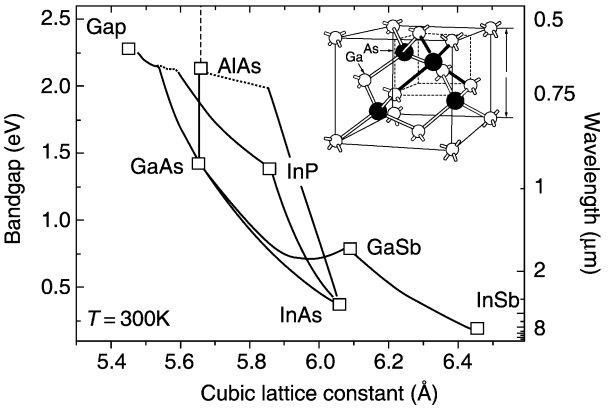


Figure 1 Fundamental bandgap of selected compounds within the III–V semiconductor family plotted as a function of the cubic lattice constant – a. The inset depicts the zincblende crystal structure within the GaAs unit cell. Data obtained from various sources, see for example Brozel MR and Stillman GE (eds) (1996) *Properties of Gallium Arsenide*, INSPEC, emis Datareviews Series No.16 and Loffe Institute, on NSM semiconductor data repository at http://www.ioffe.rssi.ru

valence band maximum occurring at the center of the Brillouin zone (Γ-point). Figure 2 summarizes our discussion of the electronic structure of GaAs, depicting the bandstructure schematically within the first Brillouin zone. The valence electron configurations of Ga and As atoms result in the band edge CB states being two-fold degenerate due to the electron spin and characterized by an almost isotropic effective mass of $m^*_\Gamma = 0.068\, m_o$. For GaAs heterostructures with reduced dimensionality, the very weak anisotropy of the conduction band states can generally be ignored when calculating the lowest energy confined electron states but has a significant impact on higher quantized levels and must, therefore, be included. In addition to the fundamental CB minimum at the zone center or Γ-point of the Brillouin zone, satellite minima occur at the six equivalent X-points ($k = (1/a, 0, 0)$) and eight equivalent L-points ($k = (1/a, 1/a, 1/a)$). The X-point and L-point minima lie $\Delta E_{\Gamma X} = 0.49$ eV and $\Delta E_{\Gamma L} = 0.32$ eV above the Γ-minimum at low temperatures ($T = 4.2$ K) and are characterized by strongly direction dependent effective masses parallel (∥) or perpendicular (⊥) to the principal axes. Electronic states in the vicinity of these minima are termed the X and L-valleys respectively and have a significant impact on high-energy electron transport in devices such as HEMTs and in avalanche carrier multiplication processes, but do not generally play a significant role in the optical properties of the binary – GaAs. However, satellite CB minima can strongly influence the optical properties of GaAs–AlAs superlattices and QWs that can exhibit a type-II band alignment for GaAs layer thicknesses < 6 nm. In this case, the electrons (holes) are localized in the AlAs (GaAs) layer and the nature of the AlAs X-states dominates the optical properties.

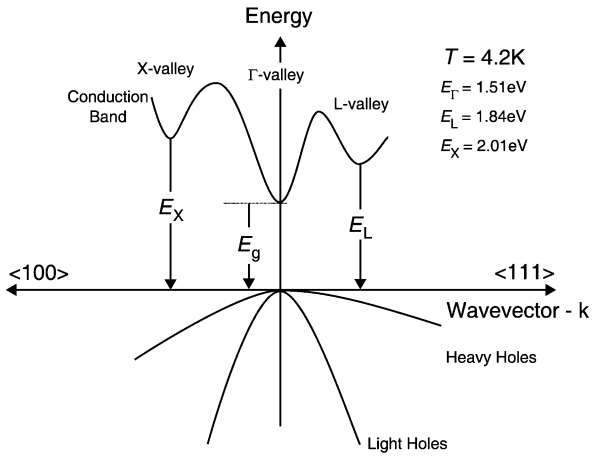


Figure 2 Schematic representation of the bandstructure of GaAs showing just a single conduction band over whole Brillouin zone in Γ-L and Γ-X directions only and neglecting the split off band that does not significantly influence electrical or optical properties.

Table 1 Low temperature GaAs effective masses for electrons (upper panel) and holes (lower panel). Values marked with * are inferred from measurements on GaAs based heterostructures and values marked ** are calculated

	Crystallographic axis	*Effective Mass (m_o)*
CB-Γ minimum	*Isotropic*	0.068(2)
CB-X minimum*	∥ [100]	1.9
	⊥ [100]	0.19
CB-L minimum**	∥ [111]	1.9
	⊥ [111]	0.075
VB-HH	[111]	0.72(3)
	[100]	0.33(6)
VB-LH	[111]	0.08(3)
	[100]	0.09(4)

The valence band (VB) states of GaAs are four-fold degenerate at the bandedge and, characteristic of most zincblende semiconductors the valence bandstructure, is significantly more strongly anisotropic than the CB states discussed above. Away from the Γ-point, the four-fold degeneracy of the valence band states is lifted into 2 doubly degenerate sets of bands termed the heavy (HH) and light hole (LH) bands by virtue of their zone center effective masses (Table 1). The HH band is considerably more strongly anisotropic than the LH band and is characterized by a much larger effective mass along the [100] direction when compared with [001], $m^*_h[111] > m^*_h[001]$. The anisotropy of the LH states is much weaker than for the HH band, with approximately equivalent effective masses along [100] and [111] crystallographic directions. Typical values for the CB and VB effective masses are summarized in Table 1, obtained using a variety of experimental techniques at low temperature.

Temperature Dependence

Quantitative values given above relate specifically to the bandstructure at liquid helium temperatures ($T = 4.2$ K). Increasing temperature results in a narrowing of many of the conduction–valence band gaps discussed above and lowering of the effective masses. This temperature dependence of the fundamental bandstructure arises from the anharmonic dilation of the lattice and from interactions with phonons, the population of which becomes significant at elevated temperature. In the simplest picture, one can visualize the effect of increasing temperature

Table 2 Varshnii coefficients for the Γ, L, and X CB minima in GaAs relative to the top of the VB. For the Γ minimum a number of experimental works are available and the value cited is the average

	$\alpha(10^{-4}$ eV/K)	β (K)
Γ point	5.45	210
X point	7.20	205
L point	6.05	204

as a shrinking of the first Brillouin zone as the lattice constant increases, and with a subsequent increase of the band curvature leading to ~5–10% lighter effective masses at room temperature than at 4.2 K. In addition, the dramatic increase of the phonon population at elevated temperatures results in a narrowing of the fundamental bandgaps. Phenomenologically, the temperature dependence of the fundamental bandgaps can be well described by the well-known Varshnii relation

$$E_g(T) = E_g^{T=0\ \mathrm{K}} - \frac{\alpha T^2}{(T+\beta)} \qquad [1]$$

where α and β are experimentally determined constants which are sensitive to the bandgap of interest (e.g. Γ, X, or L, etc.) as summarized in Table 2.

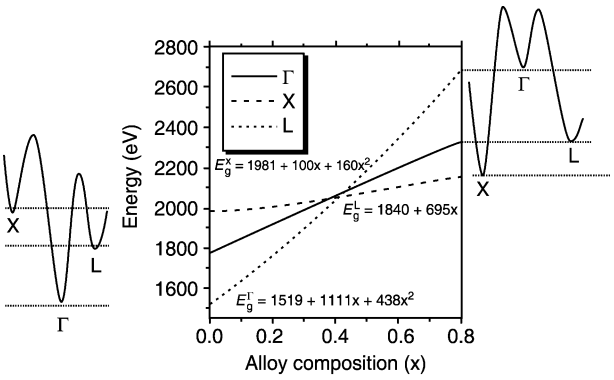


Figure 3 The absolute energy (at 4.2 K) of the Γ, X and L CB minima for $Al_xGa_{(1-x)}As$ as a function of aluminum composition up to $x \sim 0.8$. Higher compositions are not given due to the well-known unreliability of experimental data due to instability of high Al content $Al_xGa_{(1-x)}As$ due to oxidation. Polynominals describing each curve are also plotted. Data taken from the following sources: Γ-minimum, Bosio C, Staehli JL, Guzzi M, Burri G and Logan RA (1988) Direct-energy-gap dependence on Al concentration in $Al_xGa_{1-x}As$. *Physical Review* 38: 3263–3268, X-minimum, Guzzi M, Grilli E, Oggioni S, Staehli JL, Bosio C and Pavesi L (1992) Indirect-energy-gap dependence on Al concentration in $Al_xGa_{1-x}As$ alloys. *Physical Review* B45: 10951, and L-minimum, Henning JCM, Ansens JPM and Roksnoer PJ (1986) Spectroscopic determination of L6 conduction-band minima in $Al_xGa_{1-x}As$. *Journal of Physical Chemistry*, L335.

GaAs/$Al_xGa_{(1-x)}As$ – The Most Significant Alloy

Incorporation of aluminum into the GaAs crystal to form $Al_xGa_{(1-x)}As$ results in a number of profound and important modifications of the electronic properties compared with pure GaAs. Aluminum is more strongly electronegative than Gallium and consequently the electron charge distribution along the Al–As bonds are more strongly localized towards the group-III site than for Ga–As bonds which qualitatively results in a widening of the fundamental bandgap with increasing aluminum composition (x).

The variation of the $Al_xGa_{(1-x)}As$ energy gaps for the Γ, X, and L conduction band minima at low temperature are summarized in Figure 3 as a function of the alloy composition up to $x \sim 0.8$. The data of Figure 3 reflect the results of optical emission and absorption measurements and are most reliable for more dilute compositions since crystalline AlAs and high Al content $Al_xGa_{(1-x)}As$ oxidize rapidly under atmospheric conditions.[1] With increasing Al content, the Γ-minimum shifts rapidly to higher energy whilst the shift of the L and X-minima is significantly less pronounced. At $x \sim 0.4$ the Γ and X minima intersect and for $x > 0.4$, the lowest energy conduction band minimum shifts to the X-point and the fundamental bandgap becomes indirect. This effect is responsible for the short wavelength cutoff of $Al_xGa_{(1-x)}As$/GaAs based multi-quantum well LEDs at $\lambda \sim 630$ nm since the transition to an indirect bandgap is accompanied by a drastic reduction of the radiative efficiency. Shorter wavelength LED operation region of the spectrum is possible by moving to other related alloys in the III–V family such as $(Al_xGa_{1-x})_{0.5}In_{0.5}P$ (520–630 nm) and $(GaAs)_{0.6}P_{0.4}$ (~600 nm). We return to discuss the optical properties of $Al_xGa_{(1-x)}As$ below in Section 4.

As expected, the pronounced modifications of the bandgap induced by Al incorporation are accompanied by modifications of the carrier effective masses and temperature dependence of the bandgap. These compositional dependence of the Γ, X, and L conduction band and HH and LH valence band effective masses are summarized in Table 3.

It should be stressed that the relationships given in Table 3 should be considered to be accurate at the

[1] The ease by which high Al content $Al_xGa_{(1-x)}As$ can be oxidized is of great technological benefit for device fabrication. Examples include the formation of current apertures in injection lasers, $Al_xGa_{(1-x)}As$ etch stops etc. For a summary Brozel MR and Stillman GE (eds) (1996) *Properties of Gallium Arsenide*, INSPEC, emis Datareviews Series No.16.

Table 3 Compositional variation of the electron and hole effective masses for $Al_xGa_{(1-x)}As$ taken from various sources: Data drawn from the following sources: Lautenschlager Γ-P, Garriga M, Logothetidis S and Cardona M (1997) *Physical Review* B35: 9174 and Grilli E , Guzzi M and Zamboni R (1992) *Physical Review* B45: 1638. Thurmond CD (1975) L-minimum. *Journal Electrochemical Society,* 122: 1133 and Aspines DE (1976) L-Minimum. *Physical Review* B14: 5331

	Crystallographic Axis	*$Al_xGa_{(1-x)}As$ Effective Mass (m^0)*
CB-Γ minimum	*Isotropic*	$0.0682 + 0.0835x$
CB-X minimum	∥ [100]	$1.9 - 1.02x$
	⊥ [100]	$0.19 + 0.06x$
CB-L minimum	∥ [111]	$1.9 - 0.58x$
	⊥ [111]	$0.075 + 0.075x$
VB-HH	DOS	$0.62 + 0.14x$
VB-LH	DOS	$0.087 + 0.09x$

$\sim 5-10\%$ level, since reliable experimental data only exists for the GaAs binary endpoint and more dilute alloy compositions ($x < 0.3$).

The Varshnii coefficients of the temperature dependence of the bandgap vary with $x(0 < x < 0.5)$ according to $\alpha(x) = 5.5 + 3.35x(10^{-4}\,\text{eV/K})$ and $\beta(x) = 225 + 88x(\text{K})$ respectively. Recent photoreflectance measurements have indicated that these simple linear relationships do not hold well over the entire compositional range with evidence for some bowing of $\alpha(x)$ and $\beta(x)$ closer to the AlAs binary endpoint.

GaAs/$Al_xGa_{(1-x)}As$ Hetero-interfaces

Modern epitaxial growth techniques now enable the routine fabrication of semiconductor heterostructures consisting of a well-defined sequence of atomically precise layers of different materials, which together manipulate the opto-electronic properties to suit a particular device or application. Of particular technological relevance is the GaAs/$Al_xGa_{(1-x)}As$ hetero-interface since both materials are lattice matched for $0 < x < 1$ (Figure 1), and multiple layers of GaAs and $Al_xGa_{(1-x)}As$ can be readily deposited without introducing defects into the crystal which seriously degrade optical performance. In this section, we summarize the most salient features of the GaAs/$Al_xGa_{(1-x)}As$ hetero-interface, with particular emphasis on the electronic properties before continuing to discuss the optical properties of $Al_xGa_{(1-x)}As$ in Section 4.

The most important parameters that characterize a semiconductor hetero-interface are the magnitude and sign of the conduction (ΔE_c) and valence band (ΔE_v) offsets. These parameters determine how effectively electrons and hole are localized within specific epilayers and are related to the total difference in the bandgap ($\Delta E_g = E_g^{AlGaAs} - E_g^{GaAs}$) via $\Delta E_g = |\Delta E_c + \Delta E_v|$. Taking the mean of 21 separate optical and electrical investigations between 1985 and 1995, the currently accepted value for the band-offset ratio is $\Delta E_c/\Delta E_v \approx 64/36$, with an associated variance over the body of the data of approximately $\pm 8\%$. It should be stressed that the majority of the experimental investigations of the band offsets at the GaAs–$Al_xGa_{(1-x)}As$ hetero-interface have been performed for alloy compositions $x < 0.4$ and much less experimental data exists regarding the variation of $\Delta E_c : \Delta E_v$ with higher Al composition. A rather old review (*c.* 1992), collating data available, at the time has indicated that the absolute energies of the conduction and valence band offsets vary with composition as $\Delta E_c = (800 \pm 30)x$ meV for $(x < 0.4)$ and $\Delta E_c = (510 \pm 40)x\,(0 < x < 1)$. It is widely assumed that the band offsets do not change appreciably with temperature.

Optical Properties of GaAs and Significant Alloys

Optical Absorption

Monochromatic light of energy $E_p = \hbar\omega$ incident on the surface of GaAs may stimulate interband optical transitions in which a photon is absorbed and an electron from the valence band is promoted into the conduction band. In processes involving a single photon, the photon momentum is entirely negligible compared with the electron and optical transitions are vertical in k-space.

Following excitation, the electron–hole (e–h) pairs relax extremely rapidly ($\tau_{LO} \sim 150$ fs) by sequential emission of multiple longitudinal optical (LO) phonons until their energy becomes close to the bandedge and further LO-phonon emission is forbidden by energy conservation requirements. This initial phase of the carrier relaxation can be extremely rapid, the LO-phonon scattering time in GaAs being due to the semi-polar nature of the crystal. After this initial

rapid relaxation phase, the electron and hole populations relax principally via acoustic phonon emission ($\Delta E < 2$ meV) over much longer timescales ($\sim 10-300$ ps), which depend upon the lattice temperature, carrier type and density. Finally, the electron and hole populations form thermal distributions via carrier–carrier and Auger scattering over longer timescales up ~ 500 ps. During these final relaxation and thermalization phases, e–h pairs are depleted from the conduction and valence bands by interband radiative recombination.

For energies less than the bandgap, the semiconductor normally appears transparent, although the presence of impurities and defect levels can result in a weak absorption 'tail' below the gap. Optical absorption by impurities results in only electrons in the conduction band or holes in the valence band. The wavelength of light that just causes the creation of an electron–hole pair is given by $\lambda = hc/E_g$, where h is Planck's constant, c the speed of light and E_g the fundamental bandgap. As described in Section 3.1, this corresponds to approximately $\lambda = 870$ nm at room temperature, depending on the type and level of the doping. Light with wavelength much shorter than this, corresponding to energies much larger than E_g, will create electron–hole pairs higher up the band structure with higher energy and be strongly absorbed. These carriers, however, will rapidly lose this excess energy by interacting with the lattice and relax to the minima of the conduction and valence bands.

The optical absorption is usually quantified by the absorption coefficient, α, which is a function of the photon energy or wavelength and is shown for GaAs in Figure 4. The fraction of the light intensity that is absorbed in the semiconductor after passing through a thickness x is given by $f = (1 - \exp(-\alpha x))$ and the fraction of light intensity that is transmitted is then simply $\exp(-\alpha x)$. α is usually expressed in units of cm^{-1} and at energies above the bandgap has values typically between $10^4-10^5\,\mathrm{cm}^{-1}$. Light with wavelength corresponding to energies above the bandgap is usually absorbed very efficiently (>90%) within a few microns. Increasing the bandgap, say by adding aluminum, would have the effect of making the material more transparent at a given wavelength.

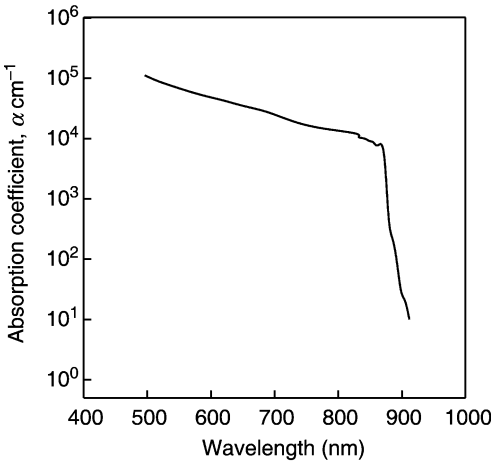


Figure 4 Variation of the absorption coefficient, α, as a function of wavelength in GaAs at room temperature.

Optical absorption is an important process as it allows light to be converted to free carriers, and hence a current in the semiconductor. This forms the basis of optical detection by most types of semiconductor photodetectors.

Refractive Index

The speed (more precisely the phase velocity) of light, of all wavelengths, in a vacuum is constant at $\sim 3.10^8\ \mathrm{m\,s}^{-1}$. The speed of light in a medium such as a semiconductor is slower due to the microscopic interaction of the electromagnetic wave with the crystal. The refractive index, R_I, is defined as the ratio of the speed in a vacuum to the speed in the semiconductor. Increasing the density of the medium decreases the speed. Changing the semiconductor, for example by adding aluminum, not only increases the bandgap as discussed above, but also decreases R_I for a given wavelength.

At wavelengths corresponding to an energy just below the bandgap of the semiconductor, the material is essentially transparent, i.e., the absorption coefficient is low.

Light can be made to travel in a waveguide comprising this semiconductor, provided that it was surrounded by semiconductor material of a lower R_I. Light is totally internally refracted at the interface between the GaAs and $Al_xGa_{1-x}As$ and so cannot escape. The semiconductors with a smaller R_I usually have a larger bandgap, so are even more transparent at the operating wavelength and help to physically confine the electrons and holes within the GaAs.

The R_I can be changed by injecting charge (electrons or holes) into the semiconductor, by changing the temperature (which changes the bandgap), and by applying an electric field. This forms the basis of operation of devices such as optical modulators. Unfortunately it also can cause 'chirping', i.e., the unwanted change in light

intensity or wavelength as the charge or temperature changes.

The R_I which is a function of the wavelength, is sometimes expressed as a complex number:

$$R_I = n + ik \quad [2]$$

where n is the real part and k is the imaginary part called the extinction coefficient, which is related to the absorption coefficient. Typical room temperature values of n for GaAs vary from about 3.37 at $\sim$1.55 μm to 3.87 at 620 nm.

Common GaAs Based Alloys for Novel Technological Applications

GaInP/AlGaInP – 'Red, Orange and Green Emission'
The alloy $(Al_xGa_{1-x})_{0.52}In_{0.48}P$ is becoming technologically important as it has the same lattice parameter as GaAs for all values of x, so can grow lattice matched on GaAs. This alloy system offers the largest direct bandgap of virtually any III–V (excluding the Nitride family), so has attracted considerable interest for visible lasers and LEDs in the wavelength range 530 nm–670 nm. At room temperature the direct Γ bandgap increases almost linearly from $\sim$1.85 eV for $Ga_{0.52}In_{0.48}P$ to $\sim$2.3 for $(Al_{0.5}Ga_{0.5})_{0.52}In_{0.48}P$. For values of $x > 0.5$, the bandgap becomes indirect as in $Al_xGa_{1-x}As$ ($x > 0.4$) when the X-level indirect bandgap becomes lower than the Γ-level direct bandgap. Earlier work in this alloy suggested that the point at which it became indirect was at $x = 0.66$. The X-level bandgap appears to be relatively insensitive to the aluminum composition and has a value of $\sim$2.3 eV at room temperature.

GaAs/GaInAsN – '1.3 μm and 1.55 μm for telecommunication applications'
One limitation of GaAs is that most of the compounds that are lattice matched to it, such as $Al_xGa_{1-x}As$ or $(Al_xGa_{1-x})_{0.52}In_{0.48}P$, have bandgap energies that are large, making them unsuitable for telecommunication applications which require narrow bandgap materials capable of operating in the range 1.3 μm to 1.55 μm. Adding indium (In) to GaAs can decrease the bandgap but also causes strain as the lattice parameter increases, which limits the amount of indium and total thickness of the structure.

Recently, it has been shown that while adding In to GaAs increases the lattice parameter, adding N to GaAs decreases it. Consequently GaInAsN can be made to be lattice matched to GaAs by controlling the In and N content. Unlike conventional alloy semiconductors, adding In or N both result in a reduction of the bandgap energy. Furthermore, GaInAsN is found to be a direct bandgap material with a type I band lineup when combined with $Al_xGa_{1-x}As$, which is a wide bandgap material also lattice matched to GaAs. These properties make GaInAsN ideally suited for lasers and there have been several reports of devices operating at 1.3 μm in the literature.

See also

Semiconductor Physics: Quantum Wells and GaAs-based Structures.

References

Bosio C, Staehli JL, Guzzi M, Burri G and Logan RA (1988) Direct-energy-gap dependence on Al concentration in $Al_xGa_{1-x}As$. *Physical Review* 38: 3263–3268.

Brozel MR and Stillman GE (eds) (1996) *Properties of Gallium Arsenide*, INSPEC, emis Datareviews Series No.16.

Data drawn from the following sources: Lautenschlager Γ-P, Garriga M, Logothetidis S and Cardona M (1997) *Physical Review* B35: 9174 and Grilli E, Guzzi M and Zamboni R (1992) *Physical Review* B45: 1638. Thurmond CD (1975) L-minimum. *Journal Electrochemical Society*, 122: 1133 and Aspines DE (1976) L-Minimum. *Physical Review* B14: 5331.

Guzzi M, Grilli E, Oggioni S, Staehli JL, Bosio C and Pavesi L (1992) Indirect-energy-gap dependence on Al concentration in $Al_xGa_{1-x}As$ alloys. *Physical Review* B45: 10951.

Henning JCM, Ansens JPM and Roksnoer PJ (1986) Spectroscopic determination of L6 conduction-band minima in $Al_xGa_{1-x}As$. *Journal of Physical Chemistry*, L335.

Ioffe Institute, on NSM semiconductor data repository at http://www.ioffe.rssi.ru/SVA/NSM/Semicond/index.html

Landolt-Börnstein (1986) *Numerical data and functional relationships in science and technology*. In: Madelung O (ed.) *Semiconductors*, vol 22. Berlin: Springer Verlag.

Nakwaski W (1995) InAs, AlAs and their ternary compounds. *Physica* B210: 1–25.

Varshni YP (1967) Temperature dependence of the energy gap in semiconductors. *Physica* B34: 149.

Group IV Semiconductors, Si/SiGe

D J Paul, University of Cambridge, Cambridge, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

SiGe is now a mature materials system with electronic circuits on the market place. While no optical devices are available yet, there is a substantial amount of research in the field. In this article, the fundamental materials properties will be reviewed. As SiGe is a strained system, we shall start with the basic material properties as this determines the Ge content and layer thicknesses which are available for use. The band structure and electronic behavior will then be reviewed before the optical properties of the material.

The main drive for SiGe research in optical systems is the ability to use the extensive Si processing knowledge and to integrate a number of components onto a Si chip. Conventional Si transistors may also be used as driver circuits for the optical components, reducing chip counts in systems and therefore reducing costs. As Si processing is substantially cheaper than other semiconductor materials, if SiGe devices can be made on Si processing lines, they have the potential to be substantially cheaper than comparable III–V or II–VI devices. The SiGe material properties that will be reviewed in this article should be considered as an expansion of bulk silicon.

Material Properties and Strain

Ge has a 4.2% larger lattice constant that Si (Table 1). When growing epitaxial films of germanium or the alloy $Si_{1-x}Ge_x$ $(0 < x < 1)$ on a silicon substrate, there exists a maximum thickness called the critical thickness above which it costs too much energy to strain additional layers of material into coherence with the substrate. Instead defects appear, in this case misfit dislocations, which act to relieve the strain in the epitaxial film. Relaxed Si, Ge or SiGe films form diamond, diamond and zincblende lattices respectively while strained heterolayers below the critical thickness form a tetragonal symmetry arrangement. The strain therefore will substantially modify the electronic and optical properties of the material as will be discussed later.

Most SiGe films are produced by either chemical vapor deposition or molecular beam epitaxy. Both these techniques allow the deposition of films at low temperatures (typically 670–970 K) which is essential to prevent strained layers relaxing and forming defects. Chemical vapor deposition uses a number of different gases such as silane, disilane, dichlorosilane, and germane (occasionally with a hydrogen carrier gas) as the main species, while phosphine or arsine are used as n-type dopants and diborane is the typical p-type dopant. With the high melting point of both silicon and germanium, molecular beam epitaxy requires electron beam evaporators for the host materials silicon and germanium while effusion cells of antimony (occasionally phosphorus) and boron are used for n- and p-type dopants, respectively.

There are a number of models used to calculate the equilibrium critical thickness. The most often quoted is the Matthews and Blakeslee model (Figure 1). Experimentally it has been shown that defect-free films with thicknesses above the critical thickness may be grown (Figure 1) although these layers are metastable and will relax if temperatures above the

Table 1 Materials properties of Si, Ge and relaxed $Si_{1-x}Ge_x$

Element		*Si*	*Ge*	*Bulk relaxed $Si_{1-x}Ge_x$*
Lattice		Diamond	Diamond	Zincblende
Lattice constant	a_o (nm)	0.5431	0.5657	$0.5431 + 0.0226x$
Lattice mismatch	η_o (%)	0	4.17	$0.04115x - 0.02500x^2 + 0.05826x^3 - 0.03379x^4$
Density	(kg m^{-3})	2329	5323	$2329 + 2994x$
Sound velocity	m s^{-1} (300 K)	8440	4900	$8440 - 3540x$
Poisson s ratio	ν	0.280	0.273	$0.280 - 0.007x$
Melting point	(K)	1685	1210	
Elastic moduli	c_{11} (GPa)	165.8	128.5	
Elastic moduli	c_{12} (GPa)	63.9	48.3	
Elastic moduli	c_{44} (GPa)	79.6	66.8	
Thermal conductivity	κ (W m^{-1} K^{-1}) (300 K)	140	60	
Dielectric constant	ε	11.9	16.2	$11.9 + 4.1x$

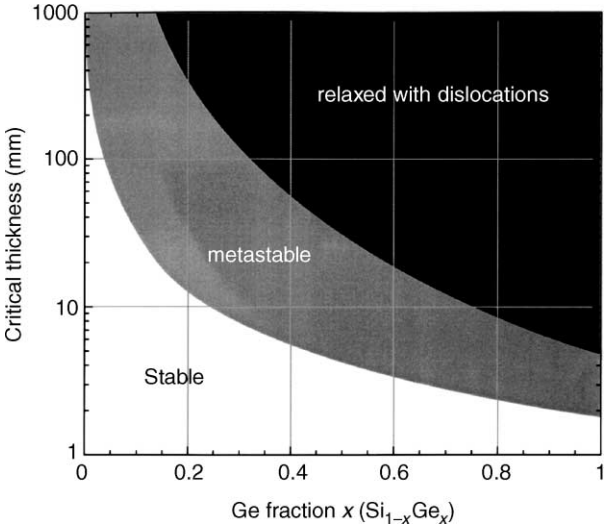


Figure 1 The critical thicknesses for growing a strained $Si_{1-x}Ge_x$ epitaxial layer on a bulk silicon substrate. The white area corresponds to the region below the equilibrium critical thickness as calculated with the Matthews and Blakeslee model while the gray region corresponds to a metastable layer which may relax after thermal treatments. The black area will have dislocations and strain relaxation using any growth technique or temperature.

growth temperature are applied to the heterolayers. For any manufacture, the layers cannot be metastable as high-temperature processes are normally employed in processing. The Matthews and Blakeslee equilibrium critical thickness, h_c for strained-$Si_{1-x}Ge_x$ on bulk Si is given by

$$h_c = \frac{0.55}{x}\ln[10h_c] \qquad [1]$$

while an approximate value can be obtained from

$$h_c \approx 1.7793x^{-1.2371}(\text{nm}) \qquad [2]$$

Figure 1 clearly demonstrates that for thick SiGe films, low Ge contents must be used.

There is also a substantial interest in producing low-defect relaxed $Si_{1-y}Ge_y$ films or strained-Si on a relaxed $Si_{1-y}Ge_y$ substrate which can provide bandgaps and band discontinuities not available to pseudomorphically grown alloy layers. Ideally the relaxed $Si_{1-y}Ge_y$ substrate is grown on a Si substrate as this allows integration with conventional complementary metal oxide semiconductor (CMOS) electronics. To date no high-quality bulk single-crystal $Si_{1-y}Ge_y$ substrates are available due to the strong segregation of Si and Ge phases as liquid SiGe crystallizes. Growing a thick $Si_{1-x}Ge_x$ layer on Si well above the critical thickness so that it relaxes produces a threading dislocation density of about $10^{12}\,\text{cm}^{-2}$ which is far too large to produce device-quality material for electronics or optics. By growing a linearly graded $Si_{1-y}Ge_y$ with the grading rate sufficiently low (typically 10% per μm), the strain relaxation mechanism changes from a nucleation relaxation mechanism to a glide relaxation mechanism resulting in dislocations with comparatively long misfit segments. The defect densities are typically 10^5–$10^6\,\text{cm}^{-2}$ (Figure 2) although recent thermal processes and other techniques have now reduced this to $100\,\text{cm}^{-2}$.

Substantial work has appeared in the literature on self-organized growth of islands. Such islands are of interest in many materials systems as the quantum confinement they may offer can potentially enhance luminescence properties in numerous opto-electronic devices. The dots are formed by the Stranski–Kranstanow growth mode in a strained system. Planar epitaxial layers grown under stress conditions can lower their elastic energy by morphological changes. For the growth of strained-Ge or $Si_{1-x}Ge_x$ on bulk Si, this corresponds to the formation of islands when the growth parameters in the system are correctly set. While the islands may have well-defined shapes, the formation of the dots relies on random nucleation followed by ripening and coalescence. These mechanisms have produced large distribution of sizes (up to 30% of the mean size), densities and positions. It is quite common to have distributions of two or more sizes of islands and the densities are typically of order of 10^8–10^9 islands per cm^2. While the heights of these islands may be as low as 3 nm, the

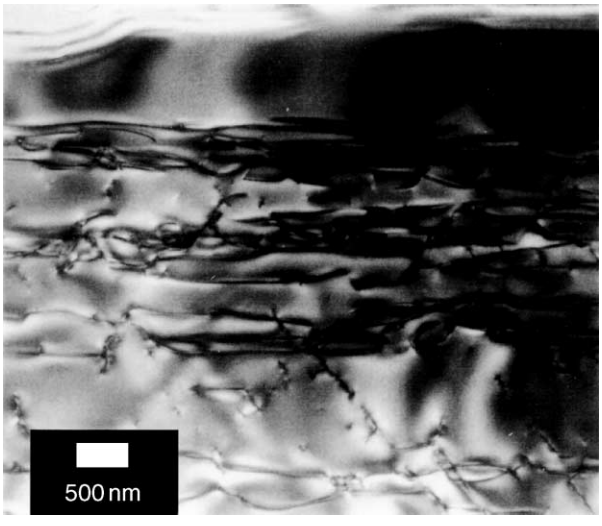


Figure 2 A transmission electron micrograph of a strain relaxation buffer grown by DERA, Malvern. The dense region of dislocations corresponds to an approximately 4 μm thick $Si_{1-y}Ge_y$ layer which is graded from $y = 0$ to $y = 0.23$. Above this is a 1 μm thick $Si_{0.77}Ge_{0.23}$ buffer which has no dislocation segments threading to the surface.

widths of the islands are typically 20 nm or greater, which is not yet small enough to produce the quantum confinement properties useful for significant improvements in opto-electronic devices.

Electronic Properties and Band Structure

As many of the optical properties and responses are derived or result from the band structure of materials, this section will concentrate on the band structure and electronic properties of SiGe layers. The bandgap of SiGe whether strained or unstrained is always below that of bulk Si. This potentially allows applications at longer wavelengths than comparable Si devices. Figure 3 shows the appropriate bandgaps for both the unstrained and strained $Si_{1-x}Ge_x$ grown on relaxed Si. For the unstrained case with a Ge fraction $x < 0.85$, the band structure is Si-like with six Δ-valleys forming the conduction band (Figure 4a). For $x > 0.85$ the band structure is Ge-like with four L-valleys. When strain is applied, both the conduction band and the valence band are split. Figure 4b shows the splitting of the conduction band with the Δ_4 valleys lower in energy and the Δ_2 valleys higher. In addition the valence band is split with the heavy-hole band being higher in energy than the light-hole band (Figure 3). The split-off band is also moved further away from the light and heavy hole bands as the strain in the system is increased. The indirect energy gap for the strained $Si_{1-x}Ge_x$ at 4.2 K as measured by photoluminescence is given by

$$E_g = 1.155 - 0.43x + 0.0206x^2 \text{ [eV] for } x < 0.85 \quad [3]$$

$$E_g = 2.010 - 1.27x \text{ [eV] for } x > 0.85 \quad [4]$$

while the direct bandgap is given by

$$E_g = 3.395 - 1.287x + 0.153x(1-x) \text{ [eV] (300 K)} \quad [5]$$

Due to the critical thickness, measurements of the bandgap of strained $Si_{1-x}Ge_x$ grown pseudomorphically on bulk Si substrates has been limited to low Ge concentrations. Using photoluminescence measurements at 4.2 K, Robbins has obtained a value for the exciton bandgap for strained $Si_{1-x}Ge_x$ layers for $x < 0.25$ to be

$$E_g = 1.155 - 0.874x + 0.376x^2 \text{ [eV]} \quad [6]$$

The bandgap for any strained $Si_{1-x}Ge_x$ layer grown on a relaxed $Si_{1-y}Ge_y$ substrate is shown in Figure 5. If the active strained $Si_{1-x}Ge_x$ layer is compressively strained then the conduction band is split into Δ_4 valleys being lower in energy and the Δ_2 valleys being higher (Figure 4b). For tensile strain (such as strained Si on a relaxed $Si_{1-y}Ge_y$ substrate), the Δ_2 valleys are lower in energy and the Δ_4 valleys are higher (Figure 4b).

The band alignments for typical heterolayer structures are shown in Figure 6. Pseudomorphically grown strained $Si_{1-x}Ge_x$ layers are only useful for forming quantum wells in the valence band (Figure 6a). Photoluminescence measurements do suggest a type I band alignment for these structures (quantum wells in both the conduction and valence bands) although it was later shown that the material is type II but due to charges producing band bending, the photoluminescence wrongly suggests a type I transition. For a quantum well for electrons, strained Si layers are grown on a relaxed $Si_{1-y}Ge_y$ substrate. In addition, quantum wells in the valence band can be produced for strained $Si_{1-x}Ge_x$ layers with high Ge content grown on relaxed $Si_{1-y}Ge_y$ substrates for $x > y$. In almost all cases the band alignments are type II although a small region exists where theory suggests type I is possible but this has not been demonstrated experimentally at present.

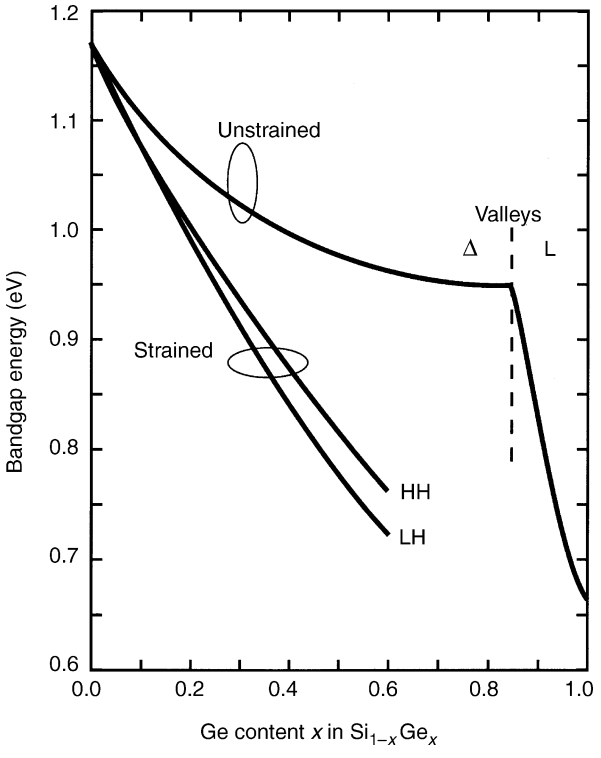


Figure 3 The bandgap energy for strained $Si_{1-x}Ge_x$ on a relaxed Si substrate and unstrained $Si_{1-x}Ge_x$. For the unstrained there is a discontinuity at $x \sim 0.85$ where the band structure changes from Si-like with Δ-valleys to Ge-like with L valleys (see Figure 4).

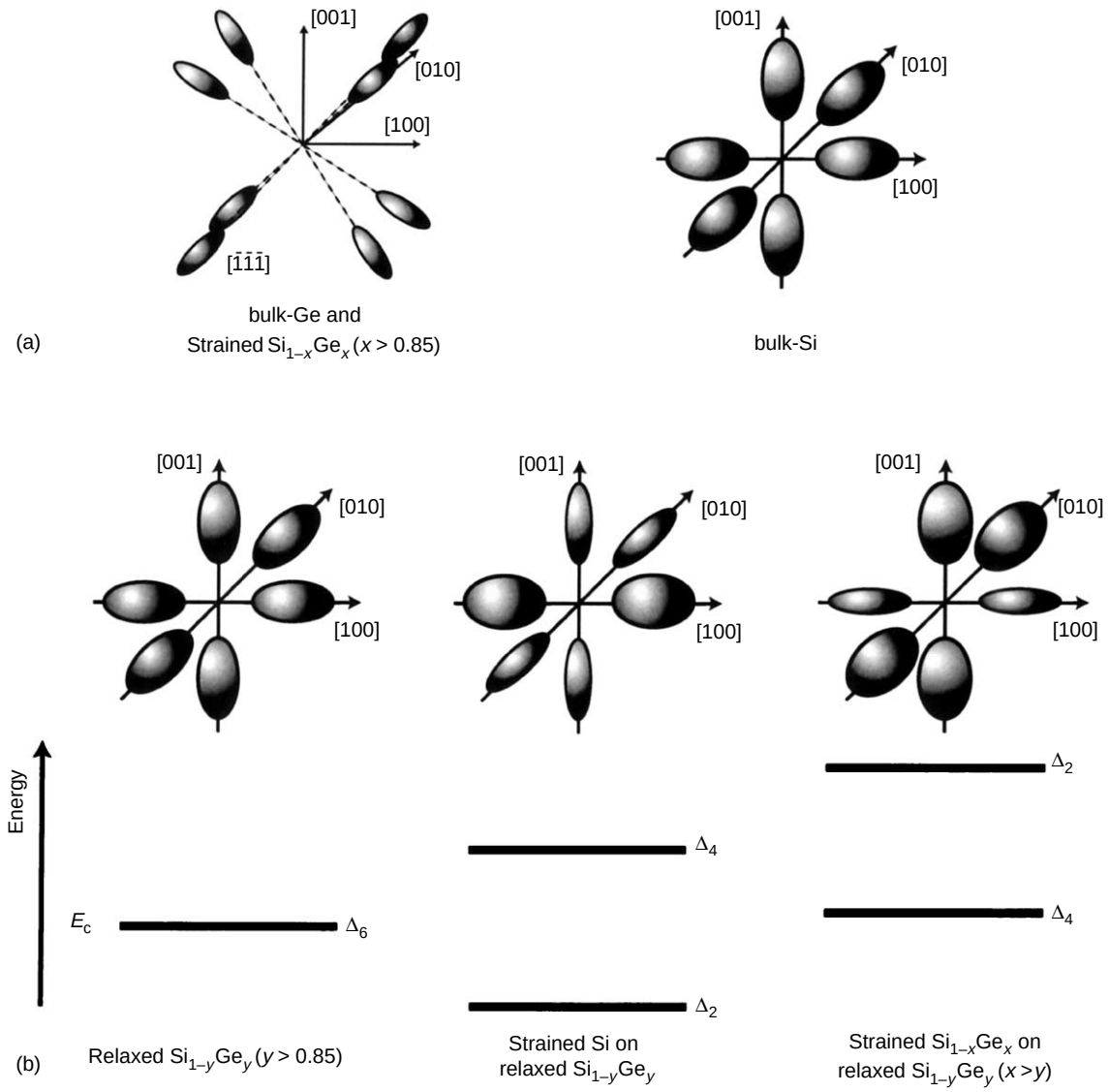


Figure 4 (a) A schematic diagram of the valleys for Ge-like and Si-like material. (b) A schematic diagram of the valley structures and energy levels of the conduction bands for relaxed, tensile strained (strained Si on relaxed $Si_{1-y}Ge_y$) and compressively strained (strained $Si_{1-x}Ge_x$ on relaxed $Si_{1-y}Ge_y$ ($x > y$)) material.

The valence band discontinuity is given by

$$\Delta E_v = (0.74 - 0.06y)(x - y)\,[\text{eV}] \qquad [7]$$

while the conduction band discontinuity for strained Si on a relaxed $Si_{1-y}Ge_y$ substrate is

$$\Delta E_c = 0.6y\,[\text{eV}] \qquad \text{for } x < 0.6 \qquad [8]$$

The complete set of ΔE_c and ΔE_v values for all x and y are shown in Figures 7 and 8 respectively. Table 2 gives some of the electronic data on Si, Ge, and relaxed $Si_{1-x}Ge_x$ including mobilities, bandgaps, and effective masses.

The mobilities for electrons and holes in relaxed $Si_{1-x}Ge_x$ layers are below the bulk Si and Ge layers due to alloy scattering (Table 2). For strained layers, mobility enhancements have been demonstrated in a number of different systems. Electron mobility enhancements result from the splitting of the valleys reducing intervalley scattering. Hole mobility enhancements result from the splitting of the valence band and reduction of effective masses. The hole mobilities for pseudomorphically grown strained $Si_{1-x}Ge_x$ quantum wells have shown values as high as $17\,000\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$ at a carrier density of $2 \times 10^{11}\,\text{cm}^{-2}$ for $Si_{0.87}Ge_{0.13}$ layers. Higher Ge contents reduce the mobility due to alloy scattering. At room temperature, only modest mobility improvements of about 20% have been observed giving $220\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$ at the high carrier

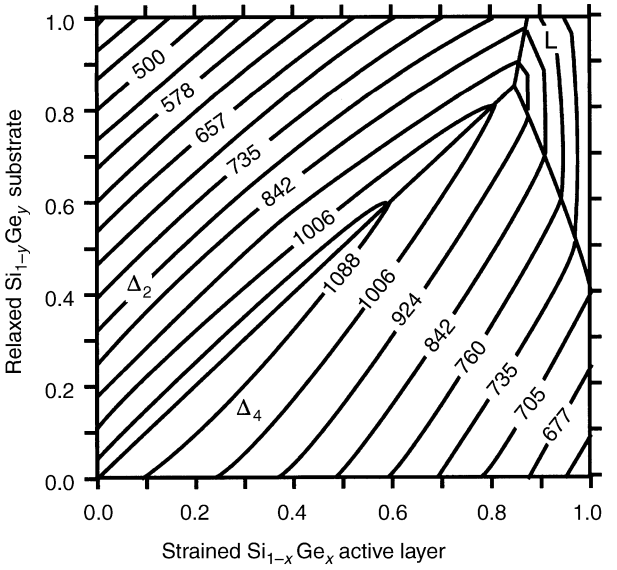


Figure 5 The bandgap energy (in meV) for strained $Si_{1-x}Ge_x$ grown on a relaxed $Si_{1-y}Ge_y$ substrate. Theoretical curves from Riger and Vogl have been calibrated to the experimental data from pseudomorphic layers.

densities required for electronic devices of about 2×10^{12} cm^{-2}. To remove alloy scattering, pure strained Ge channels grown on relaxed $Si_{1-y}Ge_y$ substrates have shown low-temperature mobilities of 55 000 cm^2 V^{-1} s^{-1} for densities of 5×10^{11} cm^{-2} with the room-temperature values reaching 1665 cm^2 V^{-1} s^{-1}. For electrons, a strained Si layer on relaxed $Si_{1-y}Ge_y$ substrates ($0.2 < y < 0.3$) also reduces the alloy scattering in the channel and produces much higher low-temperature mobilities of 390 000 cm^2 V^{-1} s^{-1} for densities of 4×10^{11} cm^{-2} while room-temperature values have peaked at 2830 cm^2 V^{-1} s^{-1}. All these values correspond to electrons or holes at heterointerfaces. Enhanced mobilities have also been observed for strained Si metal oxide semiconductor field effect transistors (MOSFET) where electron mobility enhancements of 75% compared to the bulk Si devices has been reported on relaxed $Si_{0.8}Ge_{0.2}$ although intrinsic transconductances are only increased by 60%. For holes, a strained Si layer on a $Si_{0.8}Ge_{0.2}$ relaxed substrate has demonstrated a 50% mobility enhancement over a bulk Si MOSFET.

The effective mass for electrons in strained Si layers on relaxed $Si_{1-y}Ge_y$ substrates for $0.1 < y < 0.4$ has produced a very small spread of effective masses with values of 0.19 m_0 (where m_0 is the free electron mass) for the longitudinal mass as measured by cyclotron resonance and the temperature dependence of Shubnikov–de Haas oscillations. The electron effective masses are very similar to the bulk values for

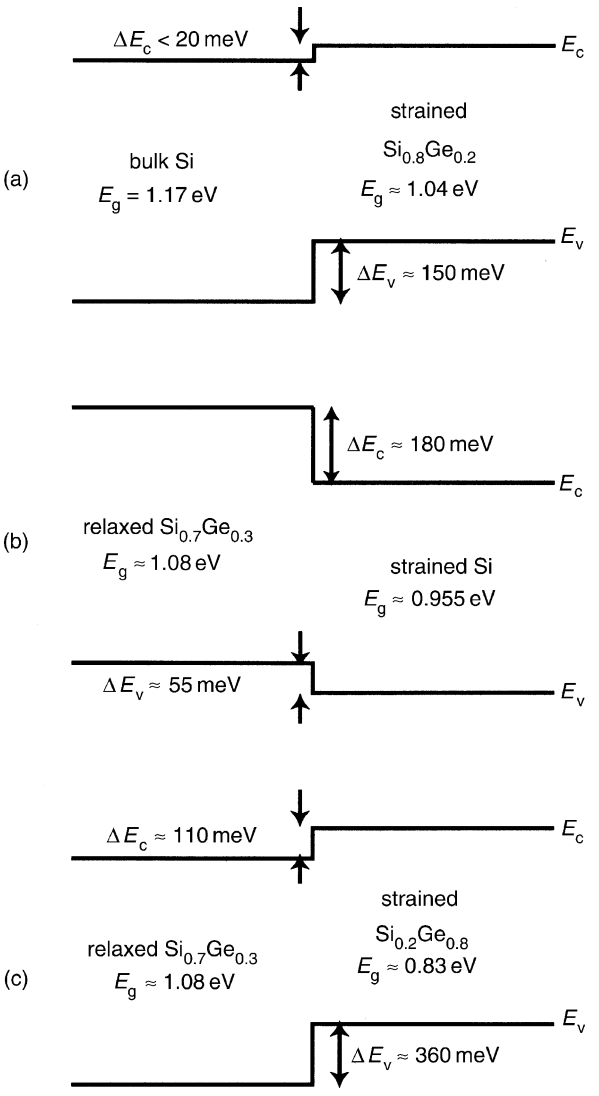


Figure 6 A schematic diagram of some typical SiGe heterolayer structures showing the bandgap, conduction band discontinuity, ΔE_c, and the valence band discontinuity, ΔE_v. (a) shows a quantum well in the valence band using a pseudomorphic strained $Si_{0.8}Ge_{0.2}$ layer; (b) shows a quantum well in the conduction band using a strained Si layer on a relaxed $Si_{0.7}Ge_{0.3}$ substrate; and (c) shows a quantum well in the valence band using a high Ge content layer on a relaxed substrate with lower Ge content.

strained Si quantum wells. Hole effective masses in pseudomorphically strained $Si_{1-x}Ge_x$ layers have produced a large spread of data even for similar alloy and strain values mainly due to the large nonparabolicity of the valence band (Table 3) and many of the measured samples having metastable $Si_{1-x}Ge_x$ quantum wells. The in-plane hole mass is theoretically expected to decrease as the strain in the system is increased but to date the experimental values have been much lower than theoretically calculated values. Experimental data (Table 3)

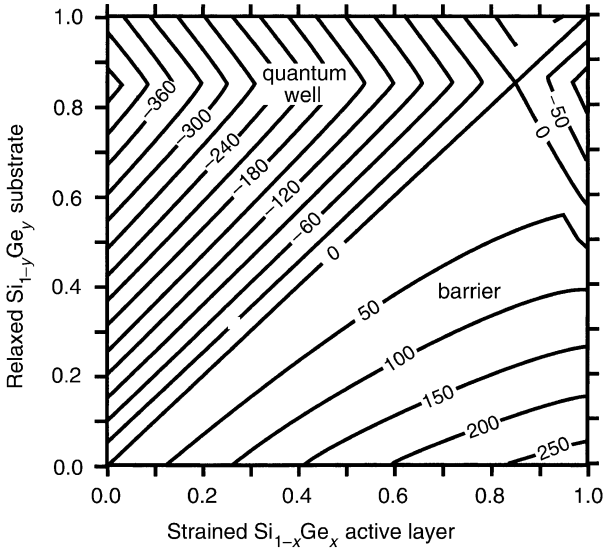


Figure 7 The conduction band discontinuity (in meV) for strained $Si_{1-x}Ge_x$ grown on a relaxed $Si_{1-y}Ge_y$ substrate. Positive numbers correspond to potential barriers to electrons while negative values represent quantum wells. The general trends and unknown experimental regimes are from the theoretical work of Rieger and Vogl while the quantum well data for $0.1 < y < 0.4$ and $x = 0$ has been recalibrated from experimental work. No experimental data are known to allow calibration of the barrier regime.

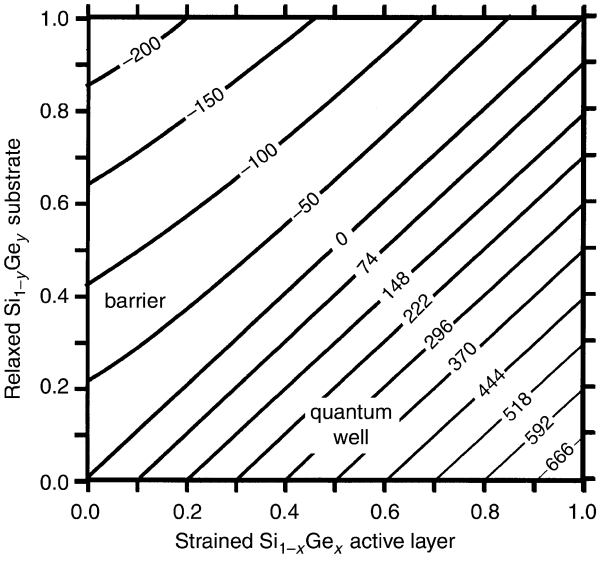


Figure 8 The valence band discontinuity (in meV) for strained $Si_{1-x}Ge_x$ grown on a relaxed $Si_{1-y}Ge_y$ substrate. Positive numbers correspond to quantum wells for holes while negative values represent potential barriers. The values from theory in the quantum well region agree well with experimental values but little experimental data exist to compare the barrier data from the theorertical values (plot from Rieger and Vogl) with experiment.

suggests that strained $Si_{0.88}Ge_{0.12}$ layers have a heavy effective mass of about $0.3\,m_0$ compared to the bulk value of $0.55\,m_0$.

Optical Properties

Due to the problems of producing high-quality single crystals of $Si_{1-x}Ge_x$, there are relatively little data on the optical properties of relaxed $Si_{1-x}Ge_x$ in the literature. The situation is even worse for strained $Si_{1-x}Ge_x$ as the layer thicknesses are limited to below 30 nm for Ge contents above 0.1, a value much smaller than the penetration depths in the system at most optical frequencies. There is, however, substantial interest in the optical properties, particularly the refractive index, as they suggest that SiGe can be used as a waveguide on Si substrates.

For the relaxed $Si_{1-x}Ge_x$ layers, most of the published work in the field has been from crystals grown by liquid phase epitaxy using melts of indium or tin. The samples had free hole concentrations of $10^{17}\,cm^{-3}$ along with indium or tin impurities and details can be found in the reviews by Humlícek. Some of the data are also from thick epitaxial layers grown well above the critical thickness where the threading dislocation density is likely to be about $10^{12}\,cm^{-2}$. The data below 2 eV were obtained by absorption measurements, those between 2 and 5.6 eV have been obtained using ellipsometry and the values between 6 and 12 eV were obtained by reflectance measurements.

Figures 9 and 10 show the real and imaginary parts of the complex dielectric function defined as $\varepsilon = \varepsilon_1 + i\varepsilon_2$ where $i = \sqrt{-1}$ for a number of different Ge contents. Figures 11 and 12 show the real and imaginary parts of the complex refractive index defined by $N = n + ik = \sqrt{\varepsilon}$. The peak starting at 3.395 eV and decreasing as the Ge content is increased corresponds to the E_1 transition across the direct gap and may be used to find the Ge content using the expression

$$x = 4.707 - \sqrt{6.538E_1 - 0.0397} \qquad [9]$$

The refractive index is high owing to the strong electronic interband absorption in the visible and ultraviolet regions of the spectrum.

The other important optical property of the material is the absorption. The absorption coefficient is defined as $K = 4\pi k/\lambda$ where k is the extinction coefficient for a wavelength λ. The intensity of a wave traversing a distance d in a homogeneous medium of extinction coefficient, k, is attenuated by a factor $\exp(-Kd)$. Therefore $1/K$ is the penetration depth. The absorption coefficient is shown in Figure 13.

Table 2 Electronic properties of Si, Ge and relaxed $Si_{1-x}Ge_x$

Element		*Si*	*Ge*	*Bulk unstrained $Si_{1-x}Ge_x$*
Direct bandgap	(eV) at 300 K	3.4	0.80	$3.395 - 1.287x + 0.153x\,(1 - x)$
Indirect bandgap	(eV) at 300 K	1.11 (D)	0.664 (L)	$1.155 - 0.43x + 0.0206x^2$ $(x < 0.85)$ $2.010 - 1.27x$ $(x > 0.85)$
Spin-orbit splitting	Δ_{so} (meV)	44	290	Discontinuity at $x \sim 0.85$
Electron mobility	μ_e ($cm^2\,V^{-1}\,s^{-1}$) at 300 K	1450	3900	$\mu_e = 1466.2 - 6425.9x + 6601.8x^2$ $+7716.5x^3 - 10944x^4 (x < 0.6)$ $\mu_e = 2.9235 \times 10^{-28} e^{71.105x} + 4790.1$ $-14246x + 11117x^2 (x > 0.6)$
Hole mobility	μ_h ($cm^2\,V^{-1}\,s^1$) at 300 K	505	1800	$\mu_h = 188.42 - 935.7x + 5455.5x^2 - 8553.7x^3$ $+ 5751.7x^4$ (all x) ($N_A = 1 - 50 \times 10^{16}$ cm^{-3})
Longitudinal electron mass	m_l^*	0.1905	0.082	Discontinuity at $x \sim 0.85$
Transverse electron mass	m_t^*	0.9163	1.58	Discontinuity at $x \sim 0.85$
Heavy-hole mass	m_{hh}^*	0.537	0.284	Discontinuity at $x \sim 0.85$
Light-hole mass	m_{lh}^*	0.153	0.044	Discontinuity at $x \sim 0.85$
Spin–orbit mass	m_{so}^*	0.234	0.095	Discontinuity at $x \sim 0.85$

Table 3 A comparison of effective masses in modulation-doped samples as measured by the temperature dependence of the Shubnikov–de Haas oscillations (SdH) or by cyclotron resonance (CR) at liquid helium temperatures. All samples have strained $Si_{0.88}Ge_{0.12}$ or $Si_{0.87}Ge_{0.13}$ quantum wells grown on bulk Si substrates. Therefore only the heavy-hole band should be populated at these low temperatures and strain values. ℓ is the mean free path which is the mean scattering distance of a hole in the system and is a quantitative measure of disorder

	Mobility ($cm^2\,V^{-1}\,s^{-1}$)	*p ($\times 10^{11}$ cm^{-2})*	*ℓ (nm)*	*m^* (m_0) SdH*	*m^* (m_0) CR*
DERA	6560	4.4	71	0.304	0.26
NRC	6800	5.9	86	0.31	–
IBM	6000	3.7	60	0.34	0.18
Warwick	9600	2.5	78	0.27	0.23, 0.29
Warwick	10500	2.1	78	0.27	0.24, 0.30

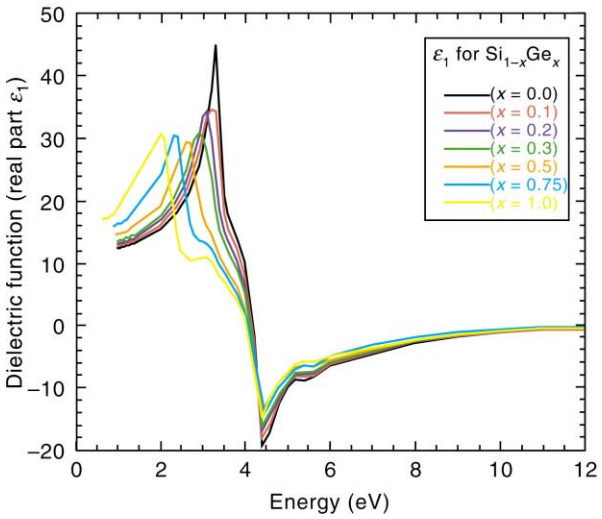


Figure 9 The real part (ε_1) of the complex dielectric function $\varepsilon = \varepsilon_1 + i\varepsilon_2$.

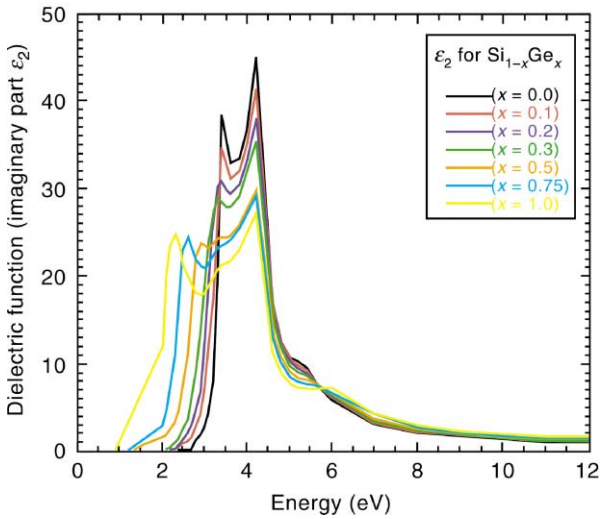


Figure 10 The imaginary part (ε_2) of the complex dielectric function $\varepsilon = \varepsilon_1 + i\varepsilon_2$.

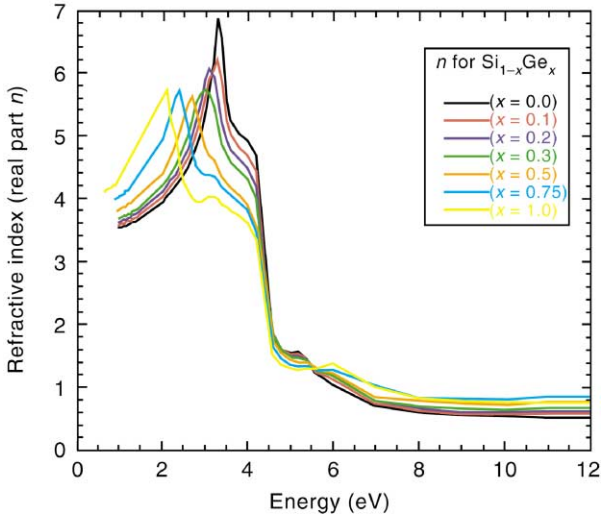


Figure 11 The real part (n) of the complex refractive index $N = n + ik$.

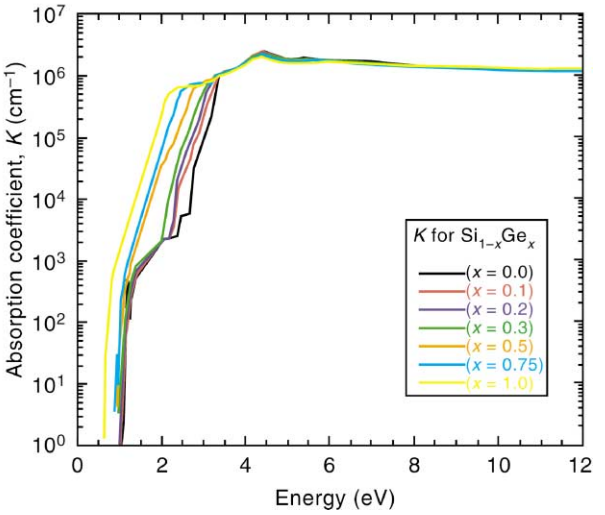


Figure 13 The absorption coefficient K.

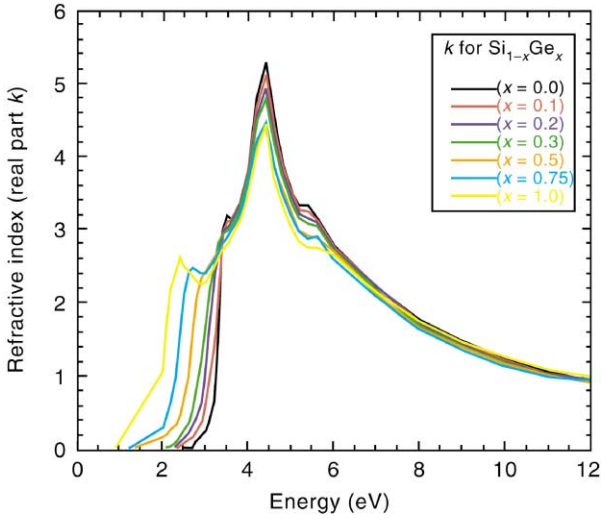


Figure 12 The imaginary part (k) of the complex refractive index $N = n + ik$.

Determining the optical properties of strained $Si_{1-x}Ge_x$ layers has been substantially more difficult. The major problem is that the absorption coefficient is about $10-20\,cm^{-1}$ around the fundamental bandgap (Figure 6) and therefore any shift due to strain effects becomes very difficult to observe using absorption techniques in thin SiGe layers. Additional complexity results as the substrate material has a fundamental bandgap close to the strained layer. Therefore photoluminescence is one of the few methods for extracting experimentally the optical changes due to strain.

The splitting of the conduction (Figure 8) and valence bands (Figure 9) has already been discussed. The photoluminescence signals will also shift with strain for the lowest lying bandgap energy. Almost all the experimental data for strained layers have been obtained for pseudomorphically grown strained $Si_{1-x}Ge_x$ alloys on bulk Si substrates and the data agree well with the bandgaps shown in Figures 3 and 5. The quantum well photoluminescence signal is found to shift by -1.4 meV per kbar of pressure. After removal of nonhydrostatic components, a value of -1.52 meV per kbar for the pure hydrostatic shift has been reported.

Conclusions

A review has been presented of the fundamental properties of relaxed and strained $Si_{1-x}Ge_x$ alloys. The materials properties were first reviewed and demonstrate the limitations imposed by the critical thickness. The electronic properties including the bandstructure were then discussed and demonstrate the lowered bandgap for $Si_{1-x}Ge_x$ alloy compositions compared to bulk Si. This property has been used in $Si_{1-x}Ge_x$ p-i-n photodetectors to move the absorption to longer wavelengths in an attempt to compete with III–V photodetectors for the important 1.3 μm and 1.55 μm fiber optic frequencies. In the final section the optical properties of relaxed and strained $Si_{1-x}Ge_x$ material were reviewed. The difference in refractive indices has opened research into SiGe on Si waveguides.

List of Units and Nomenclature

Absorption coefficient (K)	[cm^{-1}]
Band gap energy (E_g)	[eV]
Conduction band discontinuity (ΔE_c)	[meV]
Critical thickness	[nm]
Density	[$kg\ m^{-3}$]
Dielectric constant (ε)	
Effective mass [m_0 where $m_0 = 9.11 \times 10^{-31}\ kg$]	
Elastic moduli (c_{ij})	[GPa]
Extinction coefficient (k)	[cm^{-1}]
Lattice constant (a_0)	[nm]
Mean free path (l)	[nm]
Melting point	[K]
Mobility (μ)	[$cm^2\ V^{-1}\ s^{-1}$]
Pressure	[kbar]
Refractive index (N)	
Sheet carrier densities	[cm^{-2}]
Sound velocity	[$m\ s^{-1}$]
Temperature	[K]
Thermal conductivity (κ)	[$W\ m^{-1}\ K^{-1}$]
Valence band discontinuity (ΔE_v)	[meV]
Wavelength (λ)	[nm]

See also

Semiconductor Physics: Band Structure and Optical Properties; Outline of Basic Electronic Properties; Recombination Processes.

Further Reading

Humlícek J, Garriga M, Alonso MI and Cardona M (1989) Optical spectra of $Si_{1-x}Ge_x$ alloys. *Journal of Applied Physics* 65(7): 2827–2832.

Humlícek J, Lukes F and Schmidt E (1991) Silicon-germanium alloys (Si_xGe_{1-x}). In: Palik ED (ed.) *Handbook of Optical Constants of Solids*, vol. II, pp. 607–636. San Diego, CA: Academic Press.

Kasper E and Lyutovich K (eds) *Properties of Silicon Germanium and SiGe:Carbon*. London: INSPEC IEE.

Paul DJ (1999) Silicon-germanium strained layer materials in microelectronics. *Advanced Materials* 11(3): 191–204.

Rieger MM and Vogl P (1993) Electronic-band parameters in strained $Si_{1-x}Ge_x$ alloys on $Si_{1-y}Ge_y$ substrates. *Physical Review B* 48(19): 14276–14287.

Robbins DJ, Canham LT, Barnett SJ, Pitt AD and Calcott P (1992) Near band gap photoluminescence from pseudomorphic $Si_{1-x}Ge_x$ single layers on silicon. *Journal of Applied Physics* 71(3): 1407–1414.

III-Nitrides

K P O'Donnell, University of Strathclyde, Glasgow, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction to InGaN

The optical properties of wurtzite GaN are of interest mainly to band-structure theorists. Luminescence from pure InN has probably never been detected. It is therefore surprising that the combination of these binary III–N compounds produces a solid solution, InGaN, which has revolutionized blue–green optoelectronics in the last 10 years, and may form the basis of a new lighting technology to replace the tungsten lamp in the first half of the twenty-first century.

InGaN is also the material basis of the blue laser diode, which will be used for future generations of optical data storage and for many other applications that require a compact source of short-wavelength coherent light. The market for nitride light emitters has been growing at a rate of ~20% a year since the introduction of the first blue nitride light-emitting diode (LED) by Nichia Chemical in 1993. The first commercial lasers, with lifetimes in excess of 10 000 hours, were released in 1998.

So what is InGaN? The naïve response–that it is a semiconductor alloy–ignores the theoretical demonstration, based on simple equilibrium thermodynamics, that InN is insoluble in GaN at the temperatures usual for growth (using metalorganic vapor phase epitaxy, MOVPE). If we imagine a cation lattice, on which we distribute Ga and In atoms, in a fixed numerical proportion, in order to form a solid solution of given composition, it is thermodynamically probable that the In atoms will associate to form agglomerations in certain regions of the lattice. The size, shape, and properties of the resulting concentration fluctuations have been a topic of intense speculation.

Practically speaking, InGaN layers are grown either as rather thin (<3 nm) quantum wells (QWs) immersed in GaN or AlGaN or as thicker epilayers (10–500 nm) on top of thick GaN templates. Only the QWs are of use in devices, whereas the composition and structure of the thick epilayers are easier to measure in research applications. It is dangerous to suppose that the material retains

the same properties when grown in thicker layers. Recently, a third class of sample has emerged, in the form of self-organized 'quantum boxes' (originally French *bôites*), which we shall call QBs. These samples, grown at low temperatures by molecular beam epitaxy (MBE) comprise pancake-shaped disks of uncertain composition, about 50 nm across and 3–5 nm high.

Optical Spectra of Devices

There have been several 'generations' of nitride LED. We refer here to the Nichia Blue, Extrablue (actually turquoise) and Green diodes which set the scene in 1996, and the Amber diode which first appeared in 1998, although not yet available commercially. The spectral properties of these devices have been stable while their efficiencies have improved in the intervening years. The basic design appears to incorporate a single quantum well. Up to a point, the wavelength tuning depends upon the incorporation of different amounts of indium in the well. (Green and amber devices may have the same amount.)

In what follows, we prefer to specify composition by quoting the indium nitride content, x, according to the formula unit

$$\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{N} = x\mathrm{InN} + (1-x)\mathrm{GaN} \qquad [1]$$

Electroluminescence (EL) spectra of blue and green diodes are shown in Figure 1. Spectra are broad (compared to $k_\mathrm{B}T$ at room temperature) with exponential tails to high and low photon energies. The linewidth tends to increase as the device peak wavelength shifts further to the red.

Whereas it is difficult to measure directly the optical absorption spectrum of an ultrathin layer of material, an equivalent measurement can be made on an LED by monitoring the photocurrent (PC) generated by incident monochromatic light. PC spectra of the blue and green diode are shown in Figure 1, together with the absorption and photoluminescence (PL) spectra of a thick violet-light-emitting epilayer.

The spectra of Figure 1 transform into each other by 'stretching' the x-axis, using the bandgap energy of GaN as an anchor point. Comparison of many pairs of spectra produces evidence of a *universal scaling behavior*, which relates the details of each spectrum to a single controlling parameter. In the case of epilayers, this is undoubtedly the InN fraction, as we shall see later.

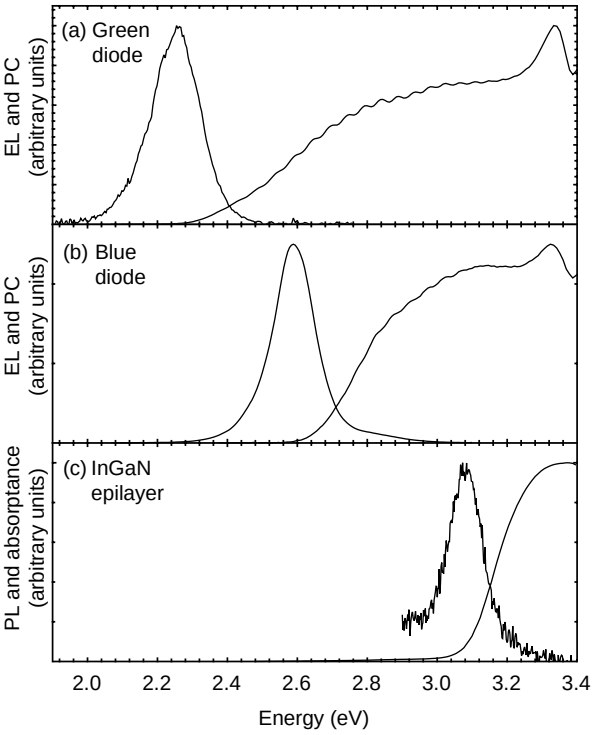


Figure 1 EL and PC (quasi-absorption) spectra of nitride LEDs, compared to PL and absorption spectra of a thick InGaN epilayer.

In order to describe the form of the absorption/excitation curve, we introduce the formal expression

$$\alpha = \frac{\alpha_0}{1 + \exp\left(\dfrac{E_\mathrm{g} - E}{\Delta E}\right)} \qquad [2]$$

where α_0 is the plateau absorption, observed experimentally, but not yet theoretically explained, E_g is the (effective) bandgap energy of the sample, E is the photon energy, and ΔE is an energy broadening term, similar to the well-known Urbach parameter.

Optical Energy Relationships

The spectral universality pointed out in the previous section requires the existence of certain relationships between the optical energies. We list the defining energies as the bandgap, E_g, the peak energy of EL or PL, E_p, and the difference between these energies, measured for a particular sample, which we call the Stokes shift:

$$\mathrm{SS} = E_\mathrm{g} - E_\mathrm{p} \qquad [3]$$

(While the energy broadening tends to increase in a regular fashion with the Stokes shift, there are irregularities in both ΔE and the PL linewidths of samples from different sources. Hence ΔE and PL

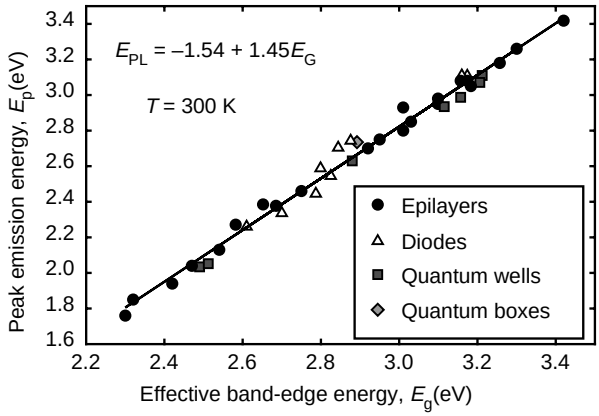


Figure 2 Linear dependence of luminescence peak energy on effective bandgap.

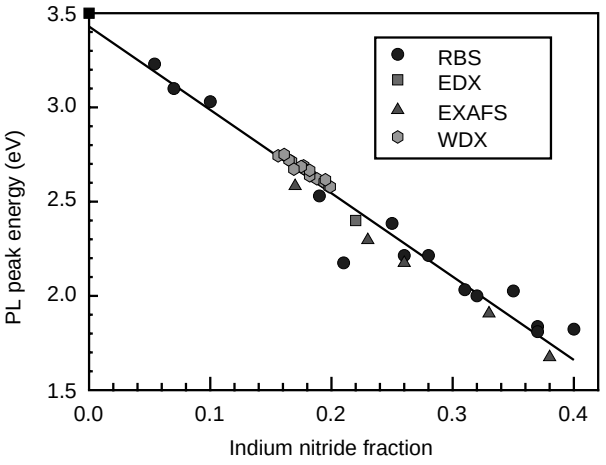


Figure 3 Linear dependence of the PL peak energy on the InN content of InGaN epilayers.

linewidth appear to be local rather than universal parameters.)

The relationship between E_g and E_p for a wide range of emitters is found to be linear (Figure 2):

$$E_p = 1.45E_g - 1.54 \qquad [4]$$

The bandgap energy and luminescence peak energy coincide at $1.54/(1.45 - 1) = 3.42$ eV, which corresponds to the energy bandgap of pure GaN. The simple linear dependence of E_p upon E_g implies a linear increase of the Stokes shift energy with increasing detuning from the GaN band edge.

Dependence of Optical Energies on *x*, the InN Fraction, for Thick Epilayers

As stated above, it is usual to measure the composition of thick layers, rather than quantum wells, for reasons of experimental convenience. Several techniques can be used to this end, including measurement of the lattice constants by X-ray diffraction (XRD) (one asymmetric and one symmetric reflections are required to take into account the strain state of the layer), Rutherford backscattering spectrometry (RBS), wavelength-dispersive (preferable) or energy-dispersive X-ray fluorescence (XRF), sputtered ion mass spectroscopy (SIMS), and extended X-ray absorption fine structure (EXAFS).

All usable techniques provide comparable results if carried out with sufficient care. We plot the peak energy of PL emission against the measured indium nitride fraction for a range of epilayer samples in Figure 3. Once more, the dependence appears to be linear in a restricted composition range. The best-fit line to a linear function $E_p(x)$ has an x-intercept which is close to the bandgap of GaN. The dependence of the optical energies on the indium nitride fraction can now be summarized by three interdependent linear relations, which are valid in the composition range $0 < x < 0.4$:

$$E_p = 3.4 - 4.4x$$

$$E_g = 3.4 - 3.0x \qquad [5]$$

$$SS = 1.4x$$

The measurements described above rely on separate determinations of composition and an optical energy for each sample. Since all samples are to some extent inhomogeneous, it would be preferable to make such measurements at one and the same time. This can be done by using an electron probe microanalyzer, modified to allow simultaneous cathodoluminescence (CL) spectroscopy. CL is analogous to PL. A recent study of samples with $0.05 < x < 0.25$ has yielded the result:

$$E_c = 3.40(1) - 3.91(5)x \qquad [6]$$

with error bars (shown) that are about ten times smaller than those of the previous work.

Bandgap of InN

It is dangerous to extrapolate experimentally determined relationships beyond the range of measurements used to derive them.

The bandgap of an alloy is best represented by a quadratic function of x, with the deviation from linearity described by a bowing parameter. The above measurements suggest a small bowing parameter for the InGaN alloy system. If we ignore bowing and equate E_p with E_c, the InN bandgap obtained by setting $x = 1$ is calculated to be 0.71 eV. This is certainly an underestimate, but is similar to recent values obtained experimentally for layers of

pure InN grown by Molecular Beam Epitaxy. Before the year 2000, the generally accepted value of the InN bandgap was close to 2 eV.

Origin of the Stokes Shift

Stokes shifts originate in systematic energy loss during the excitation–luminescence cycle. While the bandgap, as defined above, is the mean energy of electron–hole pairs across the bandgap, the PL peak is the modal energy of electron–hole recombination. The large Stokes shift in the InGaN system is probably due to the influence of giant internal electric fields that separate the carriers prior to recombination.

In the absence of a microscopic model, it is difficult to provide a quantitative description of energy loss in the excitation–emission cycle. Experiment shows that the shift can be as large as ~ 0.6 eV in an epilayer with $x = 0.4$. If such a layer were 100 nm thick and uniform in composition, it would be strained by about 4% when pseudomorphic to a GaN template, with an estimated vertical electric field of 6×10^6 V m^{-1}. Unit strain would correspond to a 15-eV shift (equivalent to a piezoelectric deformation potential). While these numbers are large, they seem reasonable in the light of sophisticated theoretical calculations of the magnitude of the piezoelectric effect in InGaN. But it is an unfortunate experimental fact that layers grown with different thicknesses, and presumably under different states of strain, may show the same value of the Stokes shift.

If the material has composition fluctuations, accompanied by strain variations, the recombination energy will be much more sensitive to the local environment. One can imagine a situation in which recombination always takes place at the lowest available energy, as carriers diffuse under the action of local strain/electric fields. This situation would be rather similar to that which pertains in quantum wells with spatial energy disorder, caused by well width fluctuations. A theoretical treatment of localization in three dimensions would be very difficult.

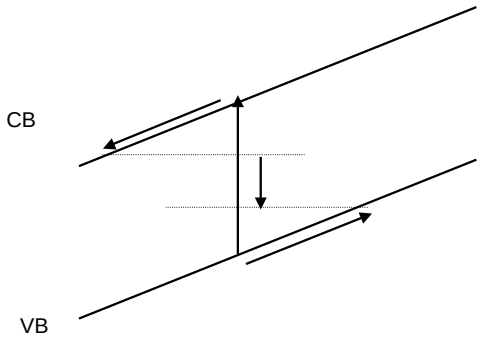


Figure 4 Excitation–emission cycle in the presence of an electric field. CB and VB are conduction band and valence band, respectively.

In any case, the physical situation of excitation and recombination in the presence of an electric field is that sketched in Figure 4.

Quantum Wells and Quantum Boxes

The incorporation of quantum wells has several advantages for light-emitting devices. In nitrides, the expected increase of oscillator strength for radiative transitions, due to the reduced dimensionality, is mitigated to some extent by the opposing effect of the internal electric fields, but the quantum well still acts as an energy sink for mobile excitations and as a site for carrier localization.

Quantum confinement can also act as a wavelength tuning mechanism in QW devices. The effect in blue-emitting nitride devices is rather small until the QW width falls below 3 nm. It should be clear that a reduction in the photon energy of recombination with increasing well width will accompany vertical separation of carriers in a piezoelectric model. Usually, as mentioned above, wavelength tuning in nitride devices is accomplished by changing the composition. What should we make of the fact that amber and green diodes apparently have the same composition? It would appear that the well widths are different. On the other hand, the amber emitters have anomalously narrow linewidths, which may imply that some other mechanism is in force. Here, we must let the matter rest, in the absence of detailed structural information.

Quantum boxes should be distinguished from quantum dots (QDs). The former are mesoscopic objects that form spontaneously under certain growth conditions, which usually involve the influence of strain. The best-characterized example in semiconductor physics is the Stranski–Krastanow growth of InGaAs structures on GaAs. While GaN in AlGaN forms recognizable nanocrystals, InGaN forms less well-defined structures of indeterminate shape, some 30–50 nm in lateral extent and 3–10 nm high. QDs, on the other hand, are truly zero-dimensional objects, in the sense that carrier motion is quantized simultaneously in three spatial dimensions. The utilization of reduced dimensionality to enhance recombination probabilities in active devices is only valid when QB structures meet the criteria of low dimensionality: the carrier motion must be restricted to a distance smaller than the appropriate Bohr radius.

Structure of InGaN Epilayers

The state of strain of InGaN epilayers is usually determined by XRD. Both symmetric and asymmetric

reflections are required to calculate the actual values of the out-of-plane and in-plane lattice constants, c and a, respectively. These can then be compared with those predicted for a relaxed alloy by Vegard's law, for example:

$$c(\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{N}) = xc_{\mathrm{InGaN}} + (1-x)c_{\mathrm{GaN}} \qquad [7]$$

The composition itself, as well as the state of strain, can be determined from measurements of c and a with one additional assumption: that the Poisson ratio of the alloy also obeys Vegard's law. If the composition is known from an independent chemical measurement, this assumption can be tested.

InGaN epilayers are usually grown on thick GaN formers. In pseudomorphic growth, the constraint, $a(\text{layer}) = a_{\mathrm{GaN}}$ applies. This is sometimes assumed to be the case when layers are thinner than about 50 nm, but nothing should be taken for granted when dealing with this material. Each layer is different. A similar warning applies to the assumption that thick InGaN layers will be relaxed. It has recently been shown that partially relaxed and pseudomorphic regions may actually coexist within the same InGaN film, which nevertheless has a composition that is uniform throughout. Other films may show a concentration gradient due to compositional pulling during growth.

The local structure of InGaN films has been measured using In-edge EXAFS. Using a tetrahedral two-shell model that does not distinguish the wurtzite and cubic crystal structures, we find that the In–N1 bond length is more or less independent of the alloy composition in the range $0.17 < x < 0.38$ of samples measured, having a value of 2.11(1) Å (compared to 2.16 Å in pure InN). The M1 shell of mixed In and Ga atoms is located on average at about 3.28 Å, as shown in Figure 5.

The straight line on the figure connects the values of the a lattice parameters (of GaN at $x = 0$ and InN at $x = 1$) and hence represents the predictions of Vegard's law for a relaxed or bulk alloy. The measured In–M1 separations are less dependent on x than Vegard predicts.

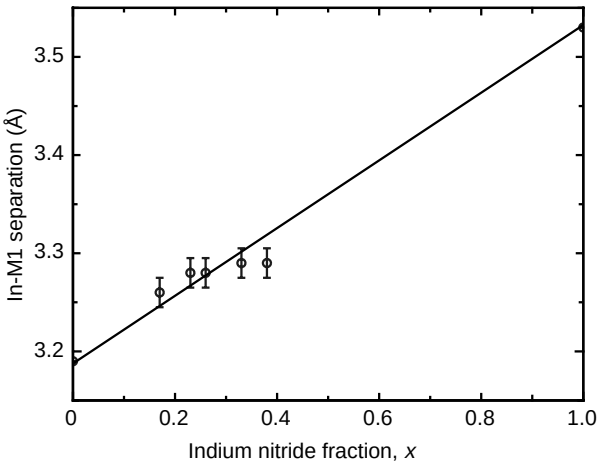


Figure 5 EXAFS-determined local structure parameters of thick InGan films.

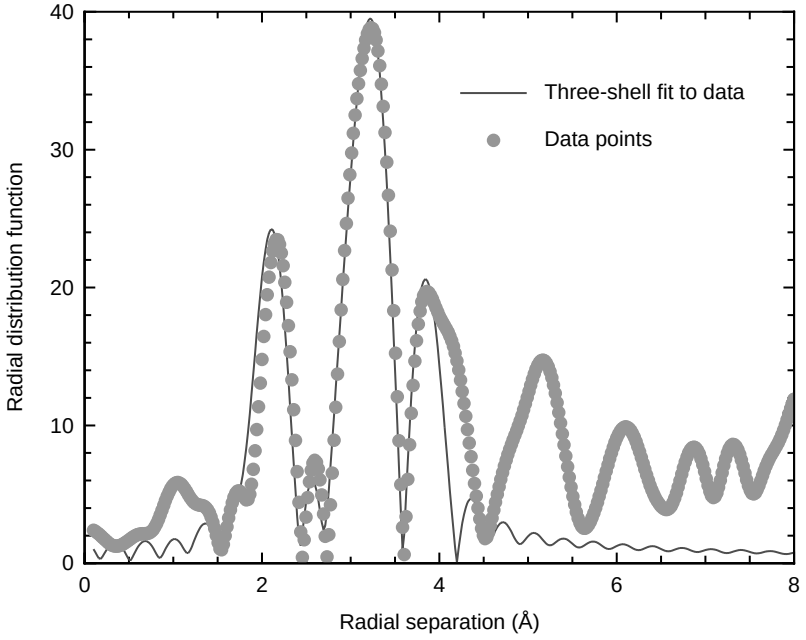


Figure 6 Fourier transform of the EXAFS spectrum yields the radial distribution function of an InGaN QB sample.

Local Structure of InGaN QW and QB

There has been a great deal of speculation in the literature concerning the local structure of InGaN alloys in the active layers of light-emitting devices. Much of this discussion paddles in the shallows of invalid comparison, as researchers extrapolate their findings from the results of measurements and calculations that can be made to those that cannot. One example is the large theoretical effort devoted to the defect physics of GaN, which is supposed somehow to 'explain' the optical properties of InGaN. ('Let us first suppose that all the horses in the race are spherical.')

We will try to avoid invalid comparisons in the brief discussion that follows. Recently, a Strathclyde University/Daresbury Laboratories/CHEA-CNRS Valbonne collaboration has succeeded in measuring the local structure of an InGaN QB sample, grown on a thick GaN template by MBE. The radial distribution function (RDF) around the target In probe atom is shown in Figure 6, together with a three-shell fit to the data.

The first shell of neighbors comprises four N atoms at 2.10(2) Å, the second shell is pure Ga at 3.22(3) Å, and the third shell pure In at 3.56(3) Å. These results suggest that there are two distinct In environments in such samples. One resembles the InGaN alloy, encountered in the epilayer samples, while the other is an environment of pure InN.

On the basis of these results, it is extremely likely that InGaN phase-separates in QB growth to form a two-component mixture comprising an In-poor and an In-rich phase. The manner in which these phases interpenetrate is a matter for speculation.

See also

Semiconductor Materials: Quantum Dots. **Semiconductor Physics:** Infrared Lattice Properties; Light Scattering; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Gil B (ed.) (1998) *Group III Nitride Semiconductor Compounds, Physics and Applications*. Series on Semiconductor Science and Technology 6. Oxford, UK: Oxford Science Publications.

Gil B (ed.) (2002) *Low Dimensional Nitride Semiconductors*. Oxford, UK: Oxford Science Publications.

Nakamura S and Fasol G (2000) *The Blue Laser Diode*, 2nd edn. Berlin: Springer-Verlag.

Large Gap II–VI Semiconductors

J F Donegan, Trinity College Dublin, Dublin, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Background

The chemical reaction of group IIA (Be, Mg), IIB(Zn, Cd) and group VI (O, S, Se, Te) atoms produces solid crystalline materials which are classified as semiconductors. The archetypal compound is ZnSe and this material will be used as the main example in this article. These are termed II–VI binary (involving two elements) compounds. It is also possible to grow alloy materials where more than one atom from each group is used; an example is $Zn_xCd_{1-x}Se$ where x is the proportion of Zn in the material. In general, there is a random arrangement of the Zn and Cd atoms within the crystal structure. Both binary and alloy semiconductor materials are single crystals so that the atoms are arranged in a periodic three-dimensional system called a lattice. The lattice may be represented by a unit cell; simply by repeating the pattern of the unit cell, we can generate the entire crystal lattice.

At the atomic level, the zinc (selenium) atom is bonded to 4 selenium (zinc) atoms which are arranged at the corners of a tetrahedron. The unit cell formed by this tetrahedral arrangement is termed the zincblende structure and is illustrated in Figure 1a. The dimension shown is called the lattice constant a and is characteristic of a particular II–VI semiconductor material. Some II–VI materials crystallize with a different unit cell which is termed the wurtzite structure, though again the atoms are bonded to four others in a tetrahedral arrangement as shown in Figure 1b. In this case, two lattice constants denoted a and c are required to characterize the unit cell. Some of the II–VI materials can occur in either of these two structures depending on how they were grown. The lattice constants of several II–VI compounds are given in Table 1.

II–VI Materials Growth

All of the II–VI family of semiconductors have high melting points due to the strong bonding that exists between the atomic species; values are given in

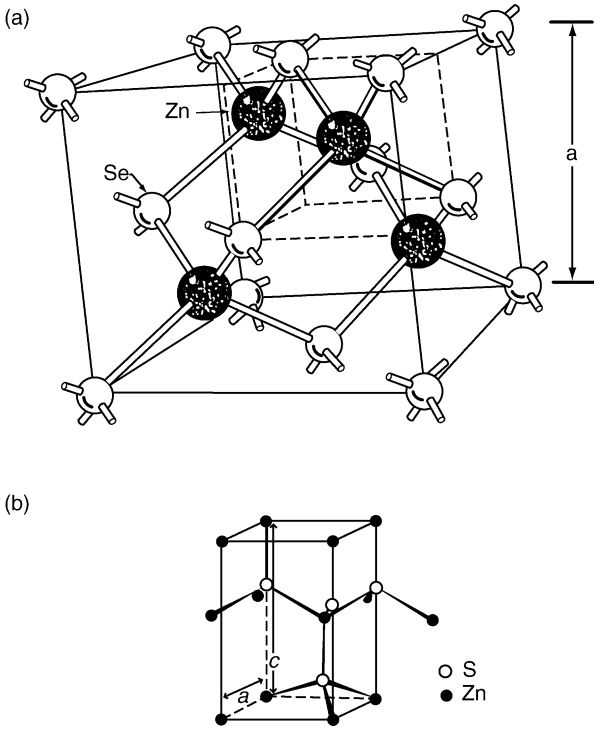


Figure 1 (a) Zincblende lattice; (b) wurtzite lattice. *a* and *c* are the lattice constants as given in Table 1.

Table 1. The crystals can be grown either by cooling a liquid (often called a melt) or by condensation from a gas (vapor). Crystal growth is a very precise and difficult process requiring the control of a large number of variables most notably temperature and pressure. As many of the II–VI materials can exist in either the zincblende or wurtzite structures, producing single crystals with only one structure is often quite difficult. Many of the crystals are 'twinned'; in a zincblende structure, a twin produces a wurtzite structure in one plane. The photograph in Figure 2 shows a ZnSe single crystal. This crystal took several days to grow indicating that very stable growth conditions must be maintained over long periods of time. The vapor phase technique takes advantage of the high vapor pressure of the II–VI compounds that allows for the growth of large single crystals at temperatures much below the melting point of these materials. Vapor growth is helped also by the congruent evaporation of the two constituents into the vapor state in a dissociated form. These two vapor constituents also recombine stoichiometrically into the solid crystal. The melting points and vapor pressures at the melting points of some II–VI semiconductors are given in Table 1.

It is also possible to grow materials in a layer-by-layer fashion onto an existing substrate; this is termed epitaxial growth. Recent developments have allowed the growth to proceed one atomic layer after the other. As a result, materials can be grown that are less than 100 Å in size. At these small sizes, the fundamental electronic and optical behavior of these materials is altered due to quantum mechanical effects. These materials are termed quantum wells. Growth temperatures in these cases are much below the melting point of these materials and allow for uniform single crystals without any 'twin' formation. In molecular beam epitaxy (MBE), atomic beams of the constituents impinge onto a heated substrate where they combine to form the crystalline material. Growth of ZnSe onto III–V substrates such as GaAs is well developed. The quality or purity of the single crystals can often be deduced from their optical properties as we will see later.

The addition of small amounts of other atoms into the crystal during the growth procedure is called doping. Semiconductors are materials whose electrical conductivity can be drastically altered by the addition of small amounts of dopants. This ability to control the electrical conductivity is the key to their use in the high-speed devices used in modern optical communications systems (*see* **Semiconductor Physics:** Impurities and Defects). In ZnSe, doping with Cl produces n-type material while doping with N produces p-type material. The ability to form a pn junction (a thin region between n and p type material) is the precursor for the production of optical devices such as light-emitting diodes (LEDs) or laser diodes (LDs) based on II–VI semiconductors. Doping also changes the optical properties of II–VI materials.

All materials are defective to some degree, and these II–VI materials are no exception. The number and type of defects are determined by thermodynamic and purity considerations during the growth process. There are native defects such as vacancies (missing atoms) and dislocations (extra or missing layers of atoms). There are also impurities present; foreign atoms due to the impurities in the starting growth materials. These impurities have an important effect on the ability of the material to emit light. Dopants and impurities may help to increase the amount of light emitted while dislocations usually reduce the emission efficiency in II–VI materials.

Optical Properties of II–VI Semiconductors: Basic Concepts

Introduction

When light is incident on a solid material, its electric field interacts with the electric field within the atoms comprising the solid. This interaction leads to a

Table 1 Structural, thermodynamic and optical properties of II–VI compound semiconductors

Material	*Lattice constant (Å)*	*Crystal structure*	*Melting point (mp) (°C)*	*Vapor pressure at mp (atm)*	*Refractive index*	*Bandgap (0 K)*	*Bandgap (300 K)*
ZnO	*a* 3.250 *c* 5.207	W	1975	22	2.0	3.44	3.37
ZnS	*a* 3.820 *c* 6.260	W	1718–1850	3.7	2.8	3.91	3.80
	a 5.410	Z			2.4	3.83	3.66
ZnSe	*a* 5.668	Z	1520	1.0	2.8	2.82	2.67
ZnTe	*a* 6.104	Z	1300	0.6	3.1	2.39	2.25
CdS	*a* 4.133 *c* 6.710	W	1405–1475	3.8	2.5	2.58	2.42
	a 5.832	Z			–	–	–
CdSe	*a* 4.299 *c* 7.015	W	1239–1264	0.4	2.6	1.84	1.74
	a 6.057	Z			–	1.77	–
CdTe	*a* 6.477	Z	–	–	2.75	1.60	1.49
MgS	*a* 5.89	Z	–	–	2.6	–	4.5
MgSe	*a* 5.62	Z	–	–	2.6	–	3.6
BeTe	*a* 5.627	Z	–	–	–	–	2.8

Data taken from many sources mentioned in the bibliography but primarily from Bhargava, R. (1997) (ed.) *Wide Bandgap II–VI Semiconductors*, The Institute of Electrical Engineers, London, Table 1 of Chapter 5.4 with permission.



Figure 2 ZnSe single crystal grown by vapor transport. Photograph courtesy of NASA.

number of effects that we can observe. The light will be reflected or scattered from the surface of the solid, depending on whether the surface is smooth or rough on a length-scale similar to the wavelength of light. The materials of interest here are high-quality crystals with a smooth reflective surface. When light enters a material, it is refracted, i.e., it changes direction within the material. The simplest example of this is when a stick is placed in a jug of water, it appears to be bent. It is of course the light that is bent due to its different interaction with the air and with water. Snell's law states that:

$$\frac{\sin \alpha_i}{\sin \alpha_{tr}} = n \qquad [1]$$

where α_i is the angle of incidence of the light and α_{tr} is the angle of refraction. n is known as the refractive index of the material and is characteristic of it. As well as being refracted, any light that enters the crystal is either absorbed within the material or transmitted through the material to emerge on the other side. If the light is absorbed the material will look colored if the absorption takes place in the visible region where our eyes are sensitive. The ZnSe crystal in Figure 2 is such a material; the orange color is due to the absorption of blue light from the white light shining on the material. Other materials look transparent; ZnO is such an example. Following absorption, some materials produce light of their own; the color of the light is characteristic of the material and this process is termed luminescence. Light may also be scattered within a crystal by imperfections and by vibrations of the atoms in the crystal.

When light enters a solid material, the electric field polarizes the atoms within the structure. The positive and negative charges within the atoms move slightly from the equilibrium positions. All of the optical effects mentioned above are related to the induced polarization. As the constituent atoms in the structure are unique, the induced polarization and the resulting optical properties are specific to particular materials, and this makes optical

characterization a very important tool for the investigation of semiconductor materials.

Reflection at the Air-to-Material Interface

Our experience of looking at reflection from glass tells us that these are weakly reflecting materials; the same is true of single crystal II–VI materials such as that in Figure 2. The interface between air and an optical material is a fundamental aspect of optical physics. The amount of light reflected can be determined from Maxwell's equations. The simplest equation is found where light is incident perpendicular to the surface of the material:

$$R = \frac{(n-1)^2}{(n+1)^2} \qquad [2]$$

where R is the reflection coefficient and n is the refractive index of the material. The value of n is characteristic of the particular material. Values of n for several II–VI materials are given in Table 1. The refractive index n for all these II–VI materials is not constant but varies slowly with the wavelength of the light impinging on the material. In Figure 3, we show the wavelength dependence of the refractive index of ZnSe.

Transmission and Absorption of Light within II–VI Materials

Once the light has passed through the interface with the air, it can either be transmitted through the material to emerge on the other side at a second material–air interface or be gradually absorbed within the body of the material. The ZnSe sample in Figure 2 is illuminated with white light, but it appears orange in color to our eyes, i.e., the orange component within the white light is transmitted through the material to our eye. The other components in the light, mostly green and blue, are absorbed within the material. To obtain more quantitative information, we take a spectrum of the absorption within the material. We illuminate the sample with light of different wavelengths and determine how much is transmitted or absorbed at each wavelength. The transmission spectrum of ZnSe is shown in Figure 4.

The visible spectral region extends from 0.4 to 0.7 microns and we see that there is a large change in transmission around 0.5 microns that is responsible for the color of the ZnSe crystal as seen in Figure 2. This spectral region around the onset of a large reduction in transmission at 0.5 microns is termed the bandgap of the material. As in the case of the refractive index, the value is unique to the material. II–VI materials are termed wide bandgap materials as they possess larger values of the bandgap energy than for Si, Ge and the III–V semiconductor materials. If we shine light of wavelength 0.5 microns onto ZnSe material, the transmission is very low. Under these conditions, the material would look black to our eyes. Once the light has been

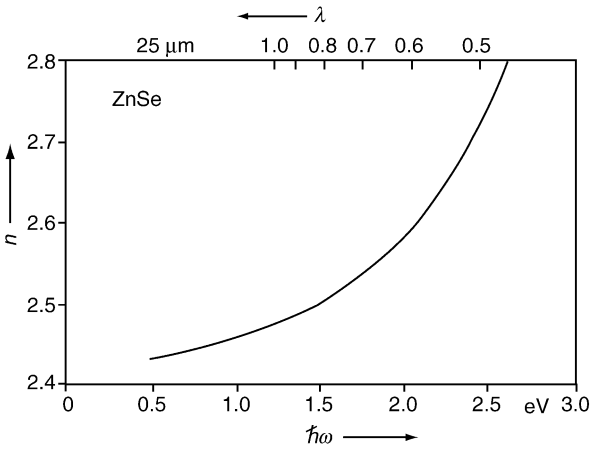


Figure 3 Refractive index n of ZnSe versus wavelength and photon energy at $T = 300\,K$. Adapted from Madelung E (ed.) (1982) *Numerical Data and Functional Relationships in Science and Technology, New Series, Group III: Crystal and Solid State Physics, vol. 17b: Physics of II–VI and II–VII Compounds, Semimagnetic Semiconductors*, Berlin: Springer-Verlag, Figure 52, with permission.

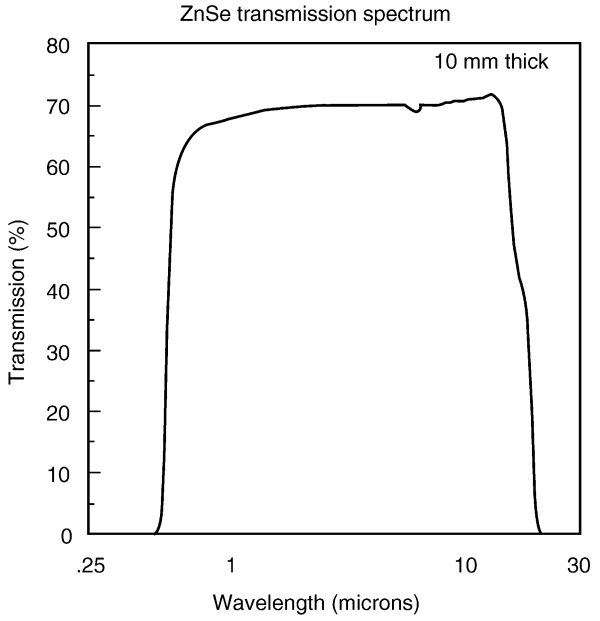


Figure 4 ZnSe transmission spectrum versus wavelength. Courtesy of II–VI Incorporated.

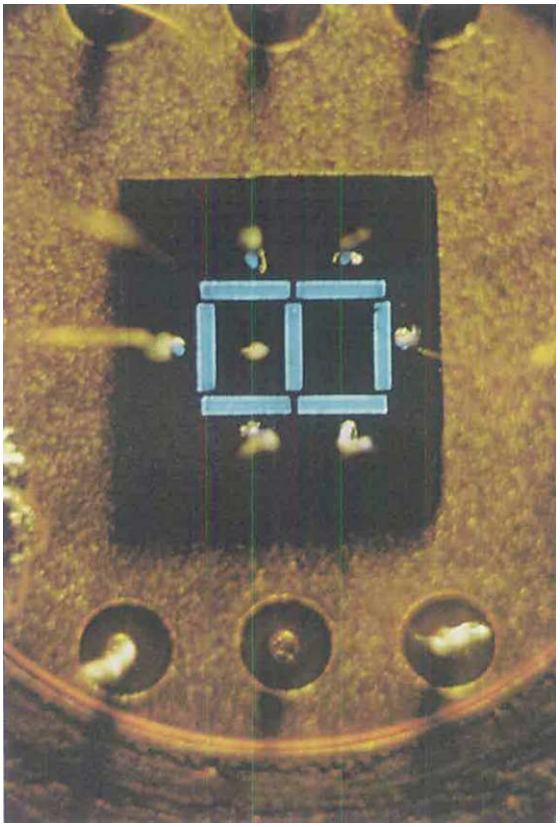


Figure 5 A seven-segment II–VI LED with blue emission taken at room temperature. Reproduced with permission from Hagerott M, Jeon H, Nurmikko AV, *et al.* (1992) Indium tin oxide as transparent electrode material for ZnSe-based blue quantum well light emitters. *Applied Physics Letters* 60: 2825–2827.

absorbed by the II–VI semiconductor, it can in turn produce light of its own. This light, known as luminescence, is again characteristic of the particular material and is the basis for the use of such materials in light-emitting diodes (LEDs) and laser diodes (LDs). In Figure 5 we show the luminescence from a II–VI LED. We see also from Figure 4 that at very long wavelengths, the ZnSe material also exhibits a large reduction in light transmission. This spectral region is termed the far infrared and the light absorbed at this spectral region produces vibrations of the atoms within the material, referred to as lattice absorption.

Light Scattering

Within the bulk of the material, small areas may have a different arrangement of atoms or may contain a substantial number of foreign atoms known as impurities. A large grouping of these impurities is known as an inclusion. In these regions, the optical properties will be different from the bulk of the material. Differences in both absorption coefficients and refractive indices can be found. These different refractive indices cause the light passing through the material to be scattered. Poor-quality material will often contain many inclusions. Even to the eye, these crystals will not look highly transparent. Changes in the refractive indices can also occur in highly pure materials due to vibrations of the atoms in the crystals. The lattice vibrations change the distance between the atoms in the material which alters the electric fields within the material and thereby changes the optical properties. Light can also be scattered due to these vibrations; experimental measurements of light scattering can provide useful information on the vibrations in these materials.

Optical Properties of II–VI Semiconductors: More Advanced Concepts

Light–Semiconductor Interactions

Light is an oscillating electromagnetic field whose interaction with semiconductor materials is in terms of the complex dielectric constant $\varepsilon_c = n_c^2$ where n_c is the complex refractive index defined as $n_c = n - ik$. k is known as the extinction coefficient and the value of k determines the amount of absorption in a material. The optical processes of reflection and absorption in semiconductors are governed by this complex dielectric constant. It turns out that there is a strong relationship between these processes. If the absorption spectrum is determined at all frequencies, then we can determine the refractive index spectrum at all frequencies by what is known as the Kramers–Kronig relations. From this description, we can now refine eqn [2] presented above:

$$R = \frac{(n-1)^2 + k^2}{(n+1)^2 + k^2} \qquad [3]$$

If the material is transparent at a particular wavelength, then $k = 0$ and we can determine the refractive index n. This is the case for values given in the table. The technique of spectroscopic ellipsometry can be used to determine the optical response of materials over a large wavelength range. This technique relies upon the change from linear polarization to elliptical polarization for light reflected from a material. From these measurements, the dielectric response can be determined.

If we look again at Figure 3, we see that there is a strong change in sample transmission around 0.5 μm. This corresponds to an energy of 2.18 eV

and is the bandgap energy of ZnSe. According to the details presented above, transitions are now being made from the full valence to the empty conduction bands of the semiconductor material. Once we have exceeded the bandgap energy, we can make further transitions between the bands. The spectrum shown in the figure is a broad band, reflecting the change from sharp atomic transitions to broad transitions in the solid state. Light with a wavelength greater than 0.5 μm does not have sufficient energy to effect the transitions between the bands. These wavelengths are not absorbed in the material. The material is transparent and the value of the extinction coefficient k is zero. The orange color of the crystal in Figure 2 corresponds to this transparent spectral region.

Cooling samples to low temperatures close to absolute zero has a profound effect on the optical properties of materials, particularly the absorption spectrum. In Figure 6, we show the absorption spectrum of a ZnSe sample at a temperature of 2 K. The absorption spectrum is determined from the measured transmission by the following equation

$$\alpha = \frac{1}{L} \ln \frac{I_0}{I_t} \qquad [4]$$

where α is the absorption coefficient, L the length of the sample, I_0 the light intensity incident on the sample and I_t the light intensity transmitted through the sample. In addition to the broad band spectrum there are now a number of sharp spectral lines near the bandgap energy.

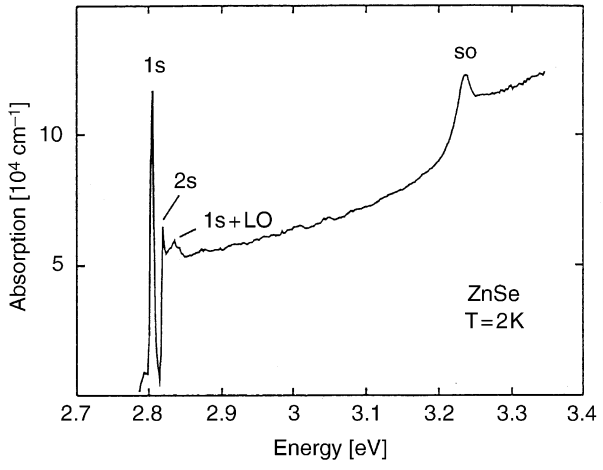


Figure 6 Low-temperature optical absorption spectrum of a ZnSe layer showing the 1S and 2S exciton features, the 1S + 1LO-phonon transition and the split-off (SO) exciton. Reproduced with permission from Bhargava R (ed.) (1997) *Wide Bandgap II–VI Semiconductors*, The Institute of Electrical Engineers: London, Figure 2 of Chapter 4.1.

These lines are due to exciton formation. An exciton is an electron–hole pair bound together by their mutual Coulomb interaction. It is the solid state equivalent of the Hydrogen atom (*see* **Semiconductor Physics:** Excitons). The bandgap energy at 2 K has increased from its value at 300 K. This is due to the thermal changes in the interatomic distance and the vibrational properties of the materials. Values of the bandgap energies determined from the position of the exciton lines in absorption spectra at low and at room temperatures are given in Table 1. At low temperatures, the luminescence is dominated by excitons, and we observe sharp line transitions. The width of the exciton transition is an excellent indicator of the quality of the material, and is often used by crystal growers to determine the quantity of defects in their materials.

It is possible for excitons and for free carriers (those that are not bound together) to recombine without giving out light. This nonradiative process usually involves the creation of local vibrations of the crystal lattice in the vicinity of defects in the material. We can define the radiative efficiency η_r as follows

$$\eta_r = \frac{R_r}{R_r + R_{nr}} \qquad [5]$$

where R_r is the radiative transition rate and R_{nr} is the nonradiative transition rate. In order to develop new optical devices based on II–VI semiconductors, it is essential that the value of R_{nr} be as small as possible. As R_{nr} is directly related to the concentration of defects in these materials, one can see why it is essential to make semiconductors with as few defects as possible. We can study these effects using a pulsed light source. Carriers are created during the pulse of light and recombine after the pulse has stopped. If we measure the luminescence intensity versus time, we can deduce the lifetime, which is given as

$$\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}} \qquad [6]$$

where τ_r is the radiative and τ_{nr} the nonradiative lifetime, respectively. By varying the carrier density we can separately determine the contribution to the lifetime from radiative and nonradiative effects. In II–VI materials at room temperature, lifetimes are of the order of 1 ns.

Quantum Wells, High Carrier Density Effects, and Lasers

Recent developments in the growth of semiconductor materials has allowed the production of samples with monolayer accuracy (3–5 Å). If we grow a layered structure with materials of high bandgap on either side of a material of lower bandgap, we can observe quantum mechanical effects. These materials are called quantum wells (*see* **Semiconductor Physics:** Quantum Wells and GaAs-Based Structures). As an example, a 75 Å ZnSe quantum well can be grown between layers of ZnSSe. Due to the small size of the active material, quantum mechanical effects come into play, most notably an increase in the bandgap energy as the width of the quantum well decreases. The effect of placing this material between ZnSSe layers that have a larger bandgap is to confine the electrons and holes to the small region of the well. In this small region, they are less likely to encounter defects and so quantum well materials usually have a higher radiative efficiency than bulk II–VI materials which are grown by standard techniques. In Figure 7, we show the exciton absorption and luminescence from a ZnSe quantum well at low temperature.

If we increase the excitation level in these materials, we can eventually cause them to operate as lasers. At high carrier densities, nonradiative transitions become saturated and radiative efficiency increases. As the refractive indices in II–VI materials are quite large, the reflectivity of the two ends of the sample are usually sufficient to act as the mirrors for the laser cavity. All the II–VI semiconductor materials have exhibited laser action due to their direct bandgap. Most of the early studies of lasing were carried out at low temperature. In recent times, lasing action has been observed at room temperature in CdZnSe quantum well materials and there are attempts at present to develop commercial II–VI lasers for use in optical storage devices and flat panel displays. In Figure 8 we show the transition from luminescence through gain and into lasing action for a CdZnSe quantum well material.

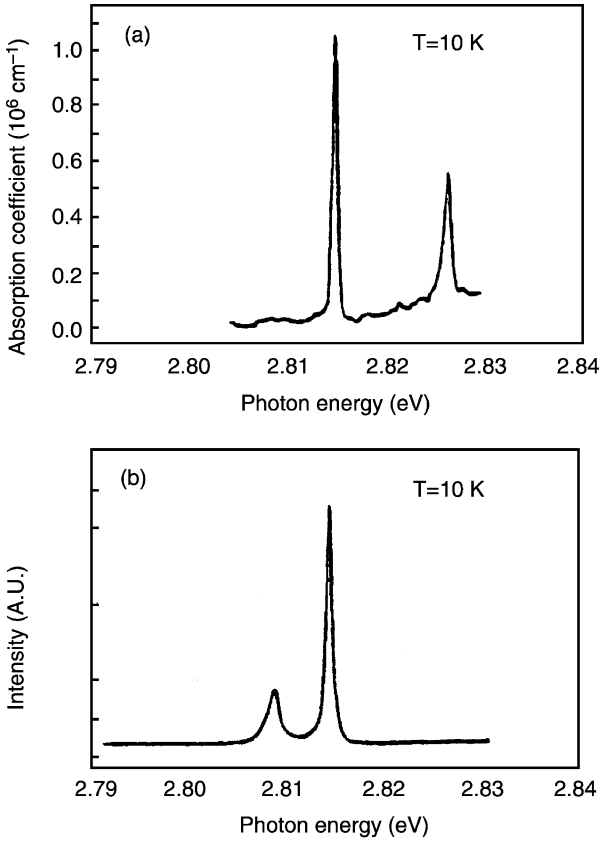


Figure 7 (a) Absorption and (b) luminescence from a single ZnSe quantum well at $T = 10$ K in the vicinity of the $n = 1$ quantum well exciton transition. Reprinted from Nurmikko AV, Jeon H, Gunshor RL and Han J (1996) II–VI lasers – directions. *Journal of Crystal Growth*, 159: 644–651 with permission from Excerpta Media Inc.

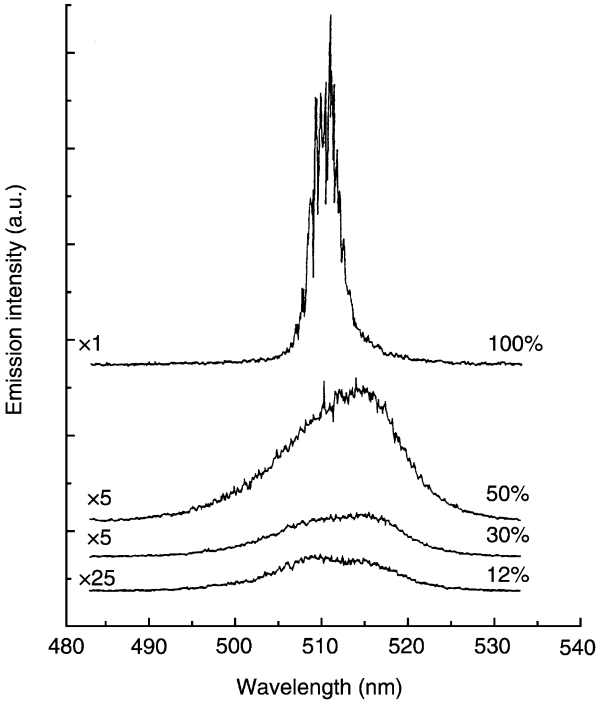


Figure 8 Emission spectra above and below lasing for a 40 Å $Cd_{0.28}/Zn_{0.72}$ Se/ZnSe system at room temperature. 100% corresponds to 1.1 times lasing threshold. Reprinted from Donegan JF, Jordan C, Rees P, Logue F, Heffernan JF and Hegarty J (1996) A study of internal loss in CdZnSe/ZnSe multiple quantum well materials. *Journal of Crystal Growth*, 159: 653–656 with permission from Excerpta Media Inc.

Dopants and Impurities

The key to the large-scale use of semiconductor materials is the large change in the electrical conductivity of the materials with the incorporation of small amounts of other elements. This process is known as doping and the foreign elements are known as dopants. The recent development of II–VI laser diodes in ZnSe materials is based on the p-n diode junction. Electrons injected into the n-type material recombine with holes injected into the p-type material at the junction between these two materials. A particular type of luminescence transition unique to semiconductor materials is the donor–acceptor pair (DAP) recombination. As donor and acceptor atoms are located in specific crystalline sites with respect to one another, we observe that the recombination energy is dependent on separation. Pairs that are close together often exhibit sharp lines while those far apart merge into a broad band. In Figure 9 we show the DAP recombination in ZnSe at 1.6 K.

It is also possible to incorporate foreign atoms in II–VI materials that have a small effect on the electrical properties but significantly change the optical properties. Such foreign atoms are termed impurities to distinguish them from dopants. An example is Mn^{2+} in ZnS. The Mn ions take the place of a small number of Zn atoms in the lattice. The luminescence from such a material is a broad emission in the visible region due to optical transitions on the Mn ions. ZnS:Mn materials are used as phosphors in flat-panel displays.

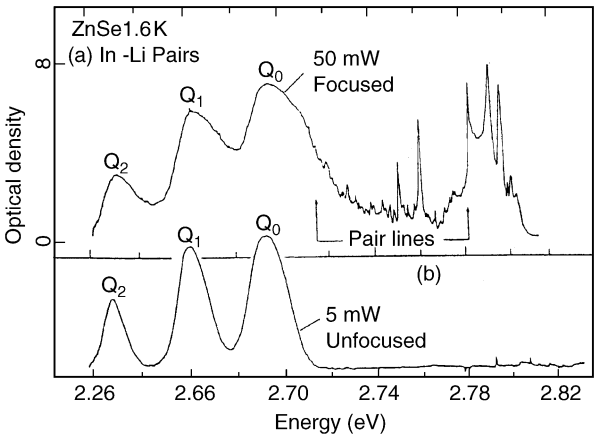


Figure 9 DAP recombination spectra of In-Li pairs in ZnSe showing pair lines, pair band (Q_0) and LO-phonon replicas (Q_1 and Q_2) at two different intensities (a) and (b). Adapted with permission from Merz JL, Nassan K and Shiever JW (1973) Pair spectra and the shallow acceptors in ZnSe. *Physical Review B* 8: 1444–1452.

Effects of Temperature, Pressure, Electric, and Magnetic Fields

The optical properties of II–VI materials are subject to variation due to temperature changes, external pressure and fields. Increasing temperature is found to lead to a decrease in the bandgap of II–VI materials. This decrease is due to the combined effects of an increase in the lattice spacing with temperature and the electron–phonon interaction. As the population of phonons is temperature dependent, this effect increases strongly as temperature increases and is the major contributor to the change in bandgap with temperature. A second effect of increasing the temperature is the change in the luminescence properties. In particular, it is found that the luminescence efficiency decreases and this is reflected in a reduced luminescence lifetime at room temperature compared with low-temperature values. Looking again at eqn [6], it is the nonradiative lifetime that changes rapidly with temperature and results in the decrease in the luminescence efficiency.

Pressure changes may be hydrostatic or uniaxial. In the hydrostatic case, the atoms are all brought closer together, and for the II–VI materials, this results in an increase in the bandgap energy. On the other hand, a uniaxial stress results in a decrease in the bandgap in the direction of the stress and an increase in the transverse directions. In II–VI materials with a wurtzite structure, the situation is more complex with two longitudinal and four transverse deformations possible.

External electric and magnetic fields can also be applied to II–VI semiconductors and thereby the optical properties can be changed. Electric fields polarize the materials and can lead to a reduction in the optical strength of transitions due to this polarization-induced carrier separation. Magnetic fields, on the other hand, lead to a splitting of the conduction and valence bands into Landau levels due to the circular motion of the carriers around the direction of the field.

See also

Semiconductor Physics: Band Structure and Optical Properties; Excitons; Impurities and Defects; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Adachi S and Taguchi T (1991) Optical Properties of ZnSe. *Physical Review B* 43: 9569–9577.

Bhargava R (ed.) (1997) *Wide Bandgap II-VI Semiconductors*. London: Institute of Electrical Engineers.

Madelung E (ed.) (1982) *Numerical Data and Functional Relationships in Science and Technology, New Series, Group III: Crystal and Solid State Physics, vol. 17b: Physics of II-VI and I-VII Compounds, Semimagnetic Semiconductors*. Berlin: Springer-Verlag.

Madelung E (ed.) (1987) *Numerical Data and Functional Relationships in Science and Technology, New Series, Group III: Crystal and Solid State Physics, vol. 22a: Intrinsic Properties of Group IV Elements and III-V, II-VI and I-VII Compounds*. Berlin: Springer-Verlag.

Mar H (1992) Optical properties of wide bandgap II-VI semiconductors. In: Ruda HE (ed.) *Widegap II-VI Compounds for Opto-electronic Applications*, pp. 167–188. London: Chapman and Hall.

Nurmikko AV and Gunshor RL (1993) Physics and application of II-VI quantum confined structures. In: Henneberger F, Schmitt-Rink S and Göbel EO (eds) *Optics of Semiconductor Nanostructures*, pp. 229–262. Berlin: Akademie Verlag.

Pankove J (1975) *Optical Processes in Semiconductors*, 1st edn. New York: Dover Publications.

Stanley RP and Hegarty J (1993) Physics of excitons in wide bandgap II-VI quantum wells. In: Henneberger F, Schmitt-Rink S and Göbel EO (eds) *Optics of Semiconductor Nanostructures*, pp. 263–290. Berlin: Akademie Verlag.

Yu PY and Cadona M (1996) *Fundamentals of Semiconductors*. Berlin: Springer Verlag.

II–VI Compounds 1999. Ninth International Conference. Elsevier, Netherlands. *Journal of Crystal Growth* vol. 214–215, 2000.

II–VI Compounds 1997. Eighth International Conference. Elsevier, Netherlands. *Journal of Crystal Growth* vol. 184–185, 1998.

II–VI Compounds 1995. Seventh International Conference. Elsevier, Netherlands. *Journal of Crystal Growth* vol. 159, 1996.

Lead Salts

G Bauer and G Springholz, Johannes Kepler Universität, Linz, Austria

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The binary IV–VI compounds formed from Pb and group-VI elements like S, Se, and Te are among the oldest known semiconducting materials, and have been used for many years for electronic and opto-electronic device applications. The first ever reported solid state diode was made from single crystalline PbS and its rectifying properties were exploited in early radio receivers. Later on, interest in IV–VI semiconductors shifted to mid-infrared opto-electronic device applications such as photon detectors operating in the 3–14 μm wavelength range, taking advantage of the narrow energy gap between the valence and conduction band of the IV–VI compounds. Soon after the demonstration of lasing in III–V semiconductor diode structures, the first mid-infrared p-n junction laser was reported using $Pb_{1-x}Sn_xTe$, and since then efficient mid- and far-infrared IV–VI compound diode lasers have been fabricated, finding their main applications for remote sensing of gaseous pollutants in trace gas sensing devices, for toxic gas analysis systems, for human breath analysis in medical diagnostics, and for fabrication process control. Recently, IV–VI multiquantum-well structures have also attracted a lot of attention for their potential as efficient thermoelectric devices.

Basic Properties

In lead salt compounds, both the valence band maximum and the conduction band minimum occur at the L-points of the Brillouin zone. Thus these many-valley semiconductors have a direct gap, but not in the center of the Brillouin zone. In addition the small values of the energy gaps ($E_g < 410$ meV at $T = 300$ K) lead to a strong $\mathbf{k}\cdot\mathbf{p}$ interaction which in turn results in a strong nonparabolicity of the $E(k)$ dispersion relations. The small energy gaps also lead to small effective masses of electrons and holes, and the strong spin–orbit interaction (Pb is a very heavy element) leads to large values of the g-factors.

The bandgaps of the lead compounds have a positive temperature coefficient ($dE_g/dT > 0$) and a negative pressure coefficient ($dE_g/dp < 0$), i.e., the gaps increase with temperature and decrease with hydrostatic pressure. These anomalous properties are related to peculiarities of the band structure, among them the fact that in the lead compounds the Pb 6s levels are below the top of the valence band, and thus are occupied, in contrast to the group IV, III–V and II–VI compounds.

Almost all optical device applications of the IV–VI semiconductors are based on the materials subgroup of cubic lead chalcogenides PbTe, PbSe, and PbS, and their quasibinary alloys with other chalcogenide such as compounds, like PbSnTe, PbSnSe, or ternary alloys like PbEuTe, PbSrTe, PbEuSe, PbSrSe, etc. These compounds all crystallize in the rocksalt structure

with sixfold coordination and lattice constants in the 5.9 to 6.5 Å range. Mechanically, the lead salt compounds are much softer than the tetrahedrally bonded semiconductors, which has significant implications for device fabrication processes.

A further unusual property of the lead compounds is their huge static dielectric constant ε_0. In particular PbTe is close to a structural phase transition from the NaCl structure to a rhombohedral modification with ferroelectric properties. GeTe and SnTe are indeed ferroelectric with Curie temperatures of about 700 K and 145 K respectively, and also PbGeTe and PbSnTe become ferroelectric when the Ge or Sn content exceeds 1% or 20%, respectively. PbTe exhibits a softening of the transverse optic phonon mode with decreasing temperature. Although PbTe remains cubic, the temperature dependence of the static dielectric constant $\varepsilon_0 \propto C/(T-\theta)$ can be associated with an extrapolated negative Curie temperature θ of -75 K. As a consequence, the static dielectric constant increases strongly with decreasing temperature, reaching a value of $\varepsilon_0 = 1350$ at 4 K. This is the reason for the strong increase of the static dielectric constant at low temperatures as shown in Table 1, in which important properties of PbTe, PbSe, and PbS are summarized.

Also, the high-frequency dielectric constants ε_∞ of the lead salts are extremely high, which leads to refractive indices of about 4.4 for PbS, to 5.0 for PbSe and to about 6.0 for PbTe at $T = 300$ K, for photon energies just below the respective energy gaps.

As a result of the huge static dielectric constants, the hydrogen-like impurity binding energies are in the range of μeV, and the carrier scattering at ionized impurities is effectively screened. Consequently, the low-temperature carrier mobilities are high, exceeding 10^6 cm^2/Vs in PbTe. For doping of the lead salts two alternative methods are used, namely (1) adjustment of the stoichiometry of the compounds, or (2) doping with various kinds of impurities. Because the electronic states of these vacancies are resonant within the conduction or valence bands, each metal (Pb or Sn) vacancy acts as a doubly charged acceptor, and each chalcogen (Te, Se, or S) vacancy acts as a doubly charged donor. As a result, background doping levels below 10^{17} cm^{-3} are difficult. For doping with impurities, atoms from group IIIa (In, Ga, Tl), group Va (As, Sb, Bi), as well as group Ib elements (Cu, Ag) are used. These dopants are amphoteric p- or n-type, depending on whether they are incorporated substitutionally on metal or chalcogen lattice sites.

Table 1 Physical properties of some of binary IV–VI semiconductors

Parameter	*PbTe*	*PbSe*	*PbS*
Lattice constant a_0 (Å)	6.462	6.124	5.936
Melting point (°C)	930	1080	1113
Energy gap E_g			
at 300 K (meV)	319	278	410
at 77 K (meV)	217	176	307
at 4 K (meV)	190	147	286
Static dielectric constant ε_0			
at 300 K	414	210	169
at 77 K	1000	227	181
at 4 K	1350	280	–
High-frequency dielectric constant ε_∞	34	21	17
TO phonon mode frequency	18 (4 K) 33 (300 K)	37	65

Table 2 Expansion coefficients (10^{-6} K^{-1})

Compound	*Expansion coefficient*
PbS	20.3
PbSe	19.4
PbTe	19.8
GaAs	6
Si	2.6
BaF_2	18.8

Other important aspects of the lead salts are their rather large linear thermal expansion coefficients of around 20×10^{-6} K^{-1} which are about a factor of three larger than that of GaAs, and six times larger than that of Si as shown in Table 2. The thermal conductivity is much lower than that of other semiconductors, which is favorable for their use in thermoelectric devices but causes problems in the removal of the dissipated heat in IV–VI compound diode lasers.

Band Structure

A main property of the electronic band structure of the lead salt compounds is the small and direct energy gap of less than 410 meV (see Table 1). Since the maxima of the valence band and the minima of the conduction band are located at the same wavevector in reciprocal space, optical photons can be directly absorbed or emitted at the band edges. This makes the lead salts very well suited for opto-electronic applications such as infrared photon detectors or emitters. The band extrema are located at the edges in the $<111>$ direction of the Brillouin zone (L-points). In contrast to most other semiconductors, the conduction and valence bands are nearly mirror symmetric with almost equal effective masses for the electrons and holes. Because of the narrow energy gaps, the energy bands are strongly nonparabolic and the effective masses of electrons as well as holes are rather small ($0.02-0.08\,m_0$ for the transverse masses,

Table 3 Band edge values of effective masses (in units of m_0) and g-factors (absolute values), $T = 2$ K

Parameter	*PbTe*	*PbSe*	*PbS*
m_t^{CB}	0.0207	0.037	0.081
m_l^{CB}	0.0213	0.0687	0.108
m_t^{VB}	0.0241	0.036	0.076
m_l^{VB}	0.263	0.066	0.108
g_t^{CB}	16.7	32.5	–
g_l^{CB}	66.1	44.1	12
g_t^{VB}	13.5	30.6	–
g_l^{VB}	65.6	–	13

Table 4 Pressure and temperature coefficients of the energy gaps

	PbTe	*PbSe*	*PbS*
dE_g/dp(eV Pa^{-1})	-7.5×10^{-11}	-9.1×10^{-11}	-9.1×10^{-11}
dE_g/dT (meV K^{-1})	0.457	0.457	0.462

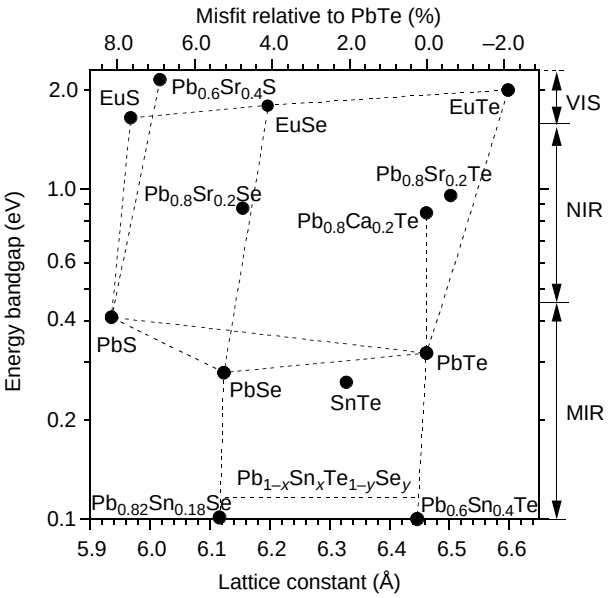


Figure 1 Energy bandgap versus lattice constant for several IV–VI compounds and alloys.

see Table 3). In addition, the bands are anisotropic, i.e., the Fermi surfaces are elongated ellipsoids of revolution around the $<111>$ axes, characterized by a longitudinal and transverse effective mass m_l and m_t parallel and perpendicular to the four-fold $<111>$ directions.

For PbTe the effective mass anisotropy is very large with a ratio of about 10 between longitudinal and transverse masses, whereas for PbSe and PbS this ratio is only 1.8 and 1.2, respectively (Table 2). Furthermore, in the alloy system $PbSe_{1-x}Te_x$ the energy gap shows a bowing, but not with a minimum value but rather a maximum gap for x at about 0.7. This is in contrast to the behavior found in direct-gap zincblende semiconductors: the bandgap $E_g(x)$ of an $A_{1-x}B_x$ alloy can generally be represented by $E_g(x) = (1-x)E_g(A) + xE_g(B) - bx(1-x)$ where b denotes the bowing coefficient. For III–V and II–VI compounds $b > 0$; whereas, for PbSeTe alloys $b < 0$.

Apart from the above-mentioned negative pressure coefficients of the direct gaps and their positive temperature coefficient which are tabulated in Table 4, there is a further intriguing anomaly in the chemical trends of the fundamental gap at the L-points: it decreases from PbS to PbSe, but increases from PbSe to PbTe, i.e., the direct gaps do not decrease monotonically with increasing atomic number of the anions. This decrease of the energy bandgap with decreasing temperature is substantial: i.e. it decreases from room temperature to 4 K by about 130 meV (see Table 1). This represents a large relative change with respect to the absolute value of the bandgap, and thus the emission of lead salt based diode lasers can be tuned over a broad wavelength range just by changing the operation temperature. This effect is actually utilized for spectroscopic applications of these lasers.

For adjustment of the fundamental absorption edge of infrared detectors as well as the adjustment of the wavelength emission of lead salt diode lasers, alloying of the lead salt compounds with other chalcogenide compounds is used. As illustrated in Figure 1, for this purpose many different ternary and quaternary alloys have been used. Important alloy systems for far-infrared applications are the lead–tin chalcogenide alloys. The tin chalcogenides (SnS, SnSe, SnTe) form single-phase pseudobinary alloy systems with the lead salts for all compositions. They have similar band structures as compared to the lead salt compounds; however, the arrangement of valence and conduction bands is exchanged. As a result, with increasing Sn content the energy bandgap of the ternary lead–tin chalcogenide alloys decreases and a zero bandgap semiconductor is reached for a certain ternary composition. For an operation temperature of 77 K, the bandgap is zero for a Sn content of $x_{Sn} \sim 40\%$ in the case of $Pb_{1-x}Sn_xTe$, and of $x_{Sn} \sim 20\%$ in the $Pb_{1-x}Sn_xSe$ case.

Optical Properties

The strong fundamental absorption in the lead salt compounds reflects the particularly high joint density of states at the direct gap at the L-points of the Brillouin zone. The main advantage of the lead

compounds with respect to other direct narrow-gap semiconductors like InSb, or HgCdTe alloys, with band extremes at the center, i.e., the Γ-point of the Brillouin zone, is related to the differences in the effective masses. Whereas the transverse effective masses of the IV–VI compounds have comparable values to the effective masses of the narrow-gap III–V and II–VI compounds, the longitudinal effective masses of the lead compounds are much larger. Thus the mass anisotropy together with the four-fold valley degeneracy at the L-point of the Brillouin zone leads to the substantially higher joint density of states and corresponding large absorption constants. Due to the rapid change of $\alpha(\omega)$ in the fundamental absorption range, a peak appears in the refractive index, $n(\omega)$, as shown in Figure 2 for PbTe, PbSe and for PbS. The shift of the peak with increasing temperature to higher photon energies reflects the increasing energy gaps.

In two-dimensional structures the changes of the absorption constant with frequency are particularly steep for energies which correspond to the onset of absorption between quantum confined electric subbands in the valence and conduction band. Such step-like features have indeed been observed in the absorption of PbTe/PbEuTe multiquantum-well (MQW) structures by Yuan *et al.* In Figure 3 the absorption constants and refractive indices are shown for a series of three MQW samples with different PbTe well widths. The steps in the absorption coefficients for the interband transitions associated to the longitudinal valleys with their main axis parallel to the [111] growth direction, denoted by 'l' as well as those associated by the oblique valleys, denoted by 'o' are accompanied by corresponding cusp-like enhancements of the refractive index. These cusps are of importance for the proper design of MQW laser structures.

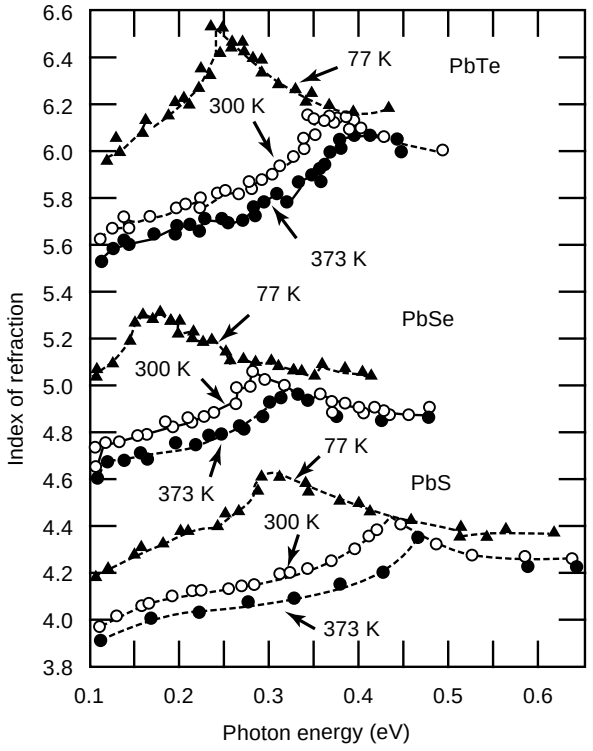


Figure 2 Refractive indices of PbTe, PbSe, and PbS as a function of photon energy at 77, 300, and 373 K. Reproduced with permission from Zemel JN, Jensen JD and Schoolar RB (1965) Electrical and optical properties of epitaxial films of PbS, PbSe, PbTe and SnTe. *Physical Review* 140: A330–A337. Copyright 1965, American Physical Society.

Because of the extremely high values of the refractive indices, the lead salts can easily be combined with materials of much lower refractive indices like EuTe, EuSe, or BaF_2 for efficient multilayer Bragg mirror structures. The high reflectivity of Bragg mirrors is caused by constructive interference of electromagnetic waves reflected at the subsequent interfaces of a multilayer structure. To obtain constructive interference, all interfaces have to be parallel, and, depending on the refractive indices of the media inside the Bragg mirror, the distance between subsequent interfaces should be an even or odd multiple of one quarter of the optical wavelength λ. In practice, stacks of two alternating layers with different refractive indices n_1 and n_2 and with a thickness equal to a quarter optical wavelength are used.

If for these λ/4 layers $Pb_{0.93}Eu_{0.07}Te$ and EuTe are used, a high index contrast of 68% is obtained and therefore very high reflectivities can be achieved already by a very small number of layer pairs. It was shown that a Bragg mirror with 3.5 periods of $EuTe/Pb_{0.93}Eu_{0.07}Te$ layer pairs for a target wavelength of 3.8 μm has only a total thickness as small as 2.55 μm. For comparison with results obtained for other material combinations it is useful to relate the stop bandwidth $\Delta\lambda$ to the center wavelength λ_T of the Bragg mirror. For $EuTe/Pb_{0.93}Eu_{0.07}Te$ this relative stop bandwidth $\Delta\lambda/\lambda_T$ is as large as 50%. An even wider stop bandwidth of 59% was demonstrated for a mirror with PbTe and EuTe λ/4 layers. This value represents, to the best of our knowledge, the highest relative stop bandwidth obtained for molecular beam epitaxy (MBE) grown Bragg reflectors.

Infrared Lasers

Among the infrared diode lasers, those based on IV–VI compounds play a particularly important role. They cover a wavelength range from 2.5 μm to about 30 μm. In fact, up to now most commercially

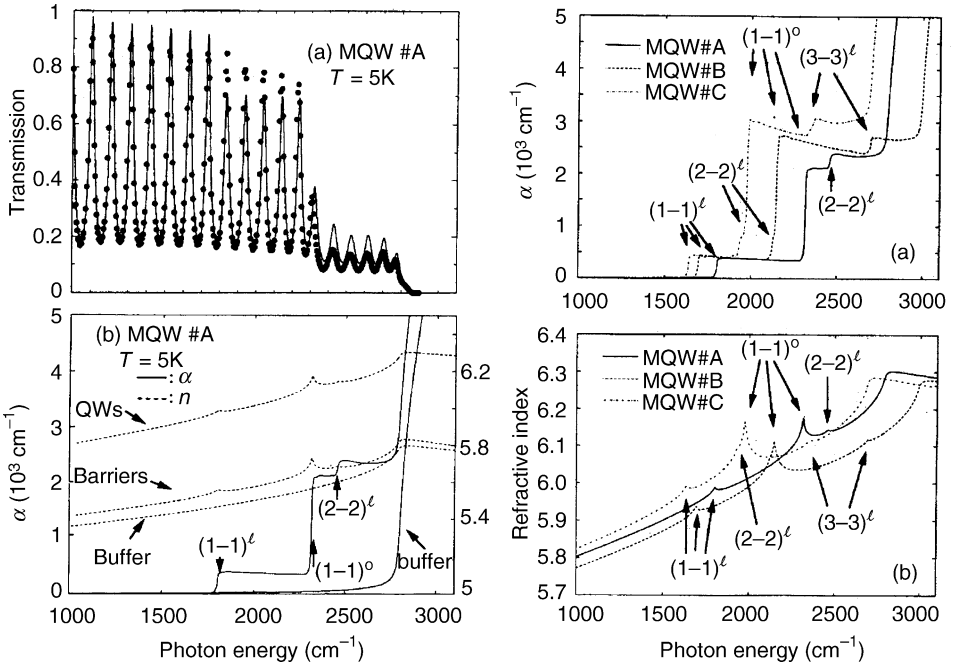


Figure 3 Left panel: (a) transmission versus energy for a PbTe/$Pb_{1-x}Eu_xTe$ multiquantum-well sample A (6.2 nm/62.1 nm, $x = 0.026$, 46 periods). Dots: experimental data; full line: calculation based on an envelope function model. (b) Frequency dependence of the absorption constant (full line) and of the refractive index (dotted lines). The arrows indicate interband transitions between electric sub-bands associated with valleys with main axis parallel to the [111] growth direction (l) and oblique to it (o). Right panel: (a) Comparison of the absorption constant of three PbTe/PbEuTe MQW samples (B: 9.3 nm/55.4 nm, $x = 0.03$, 30 periods; C: 11.8 nm/48.6 nm, $x = 0.024$, 30 periods). (b) Refractive index versus energy. Steps in the absorption constant due to onset of interband transitions are associated with cusps in the refractive index. Reproduced with permission from Yuan S, Springholz G, Bauer G and Kriechbaum M (1994) Electronic and optical properties of PbTe/$Pb_{1-x}Eu_xTe$ multiple-quantum-well structures. *Physical Review B* 49: 5476–5489, Copyright 1994, American Physical Society.

available mid- and far-infrared laser diodes are made from IV–VI compounds. The main material properties which are advantageous in this respect are (i) the electronic band structure with mirror-like bands at the L-point of the Brillouin zone and the resulting high joint density of states; (ii) the lower nonradiative Auger recombination rate in IV–VI compounds as compared to III–V and II–VI materials, and (iii) the high dielectric constant and the resulting weak detrimental influence of defects on nonradiative recombination losses. At wavelengths longer than 3 μm, IV–VI interband diode lasers are superior to their III–V counterparts for reaching high operating temperatures. So far IV–VI compound lasers hold extreme high cw operating temperature at wavelengths larger than 3 μm. With PbSe/PbSrSe diode lasers pulsed operation up to 60 °C was reported, whereas cw operation up to 223 K was achieved with separate confinement buried PbTe/PbEuSeTe heterostructure laser. Embedding PbTe quantum wells as the active medium in PbEuTe barriers between two dielectric Bragg mirrors optically pumped pulsed lasing was observed up to 65 °C. The recently developed III–V based quantum cascade lasers are rapidly improving and are becoming important competitors for IV–VI lasers. The commercial IV–VI lasers, however, are still based on conventional designs like homojunction, heterojunction, or single and multiple heterojunction lasers.

A unique property of lead salt laser diodes is the ease of tuning of the emission wavelength due to the strong change of bandgap of the IV–VI materials with increasing temperature. This, together with their narrow linewidths, makes these lasers ideally suited for spectroscopic applications, including research in molecular spectroscopy of ions or radicals in heterodyne and acoustic spectroscopy, or Doppler spectroscopy of molecular beams. So far the main applications are in time resolved combustion analysis in the automotive industry, in the monitoring of emissions of power plants, in exact monitoring of gas compositions in chemical plants, and spatially and time resolved *in situ* analysis.

Apart from the significant advances with IV–VI edge emitting diode lasers, vertical cavity surface

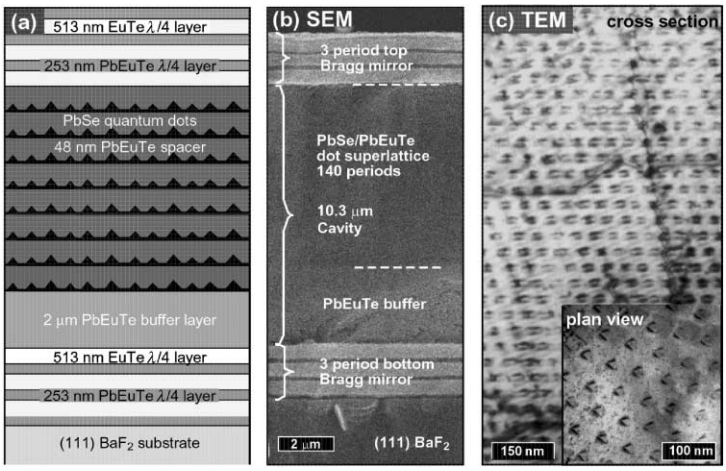


Figure 4 (a) Schematic cross-section of a PbSe quantum-dot VCSEL structure; (b) corresponding cross-sectional scanning electron micrograph; and (c) cross-sectional and plan-view transmission electron micrographs of a PbSe/PbEuTe dot superlattice sample reference sample. Reproduced with permission from Springholz G, Schwarzl T, Heiss W, *et al.* (2001) Midinfrared surface emitting PbSe/PbEuTe quantum dot lasers. *Applied Physics Letters* 79: 1225–1227. Copyright 2001, American Institute of Physics.

emitting lasers (VCSELs) have been successfully produced with optically pumped pulsed mode operation up to 50 °C. These lasers have attractive features like low-divergence circular output beam, low active volumes, and low thresholds.

High-finesse vertical cavity structures have also made it possible to realize mid-infrared surface-emitting PbSe/PbEuTe quantum dot lasers. As shown in Figure 4 in this case a self-organized PbSe/PbEuTe quantum dot superlattice was introduced into the active cavity region between Bragg mirrors consisting of three $\lambda/4$ layer pairs of 513 nm EuTe alternating with 253 nm $Pb_{0.94}Eu_{006}Te$. Optically pumped VCSEL emission at $\lambda = 4.2-3.9\ \mu m$ was achieved up to 150 K. The PbSe dots exhibit a face-centered cubic three-dimensional stacking in the superlattice. Lasing occurs simultaneously at the $m = 28^{th}$ and 29^{th} order cavity modes, which shifts with increasing temperature to the $m = 29^{th}$ and $m = 30^{th}$ mode as shown in Figure 5. This simultaneous stimulated two-mode laser operation is a result of the inhomogeneously broadened quantum-dot gain spectrum, due to PbSe dot size fluctuations.

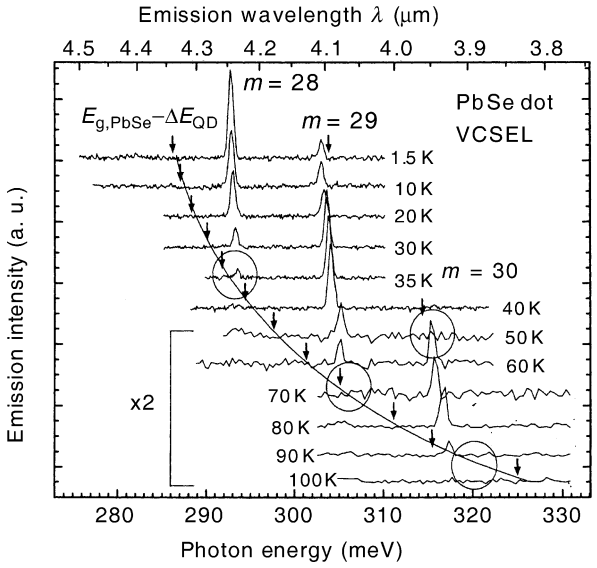


Figure 5 PbSe dot VCSEL emission spectra at various temperatures between 1.5 and 100 K for a pump power twice the threshold, showing the switching of the laser emission to higher cavity modes as the temperature increases. The arrows and the dashed line indicate the low-energy edge of the quantum-dot gain spectrum given by $E_{g,PbSe}(T) + \Delta E_{QD}$, where $\Delta E_{QD} = 141$ meV. Reproduced with permission from Springholz G, Schwarzl T, Heiss W, *et al.* (2001) Midinfrared surface emitting PbSe/PbEuTe quantum dot lasers. *Applied Physics Letters* 79: 1225–1227. Copyright 2001, American Institute of Physics.

Infrared Detectors

The photoconductive properties of PbS and PbSe have been used for the detection of infrared radiation already since about 1930. These detectors are still mass produced either by vacuum deposition or precipitation from aqueous solution in the presence of an oxidizing agent. The polycrystalline films deposited on glass or similar substrates are usually about 1 μm thick with grain sizes of about 1 μm.

Apart from polycrystalline thin-film detectors also single crystalline films have been used for infrared detection. In particular, the photoconductive response

of a sequence of n- and p-doped PbTe multilayers (so-called nipi structures) was shown to come close to that for the theoretical limit for the detectivity D^* of an ideal photodetector at temperatures between 77 and 90 K. Recently, lead selenide detectors have been used in advanced focal plane array (FPA) detectors and are commercially available as linear or bilinear FPAs with up to 526 detector elements. These elements are coupled to CMOS multiplexer readout circuits. The PbSe detector elements are thermoelectrically cooled, with entire power requirement for the FPAs and the cooler of less than 5 W. These structures are a low-cost alternative to the more elaborate FPAs based on HgCdTe technology.

The importance of thermal imaging by using focal plane arrays of many detector elements has recently stimulated further work with IV–VI epitaxial layers. In particular, the possibility of using Si substrates, as first shown by Zogg *et al.*, has brought a new impetus to the field of IV–VI infrared devices. With the use of Si wafers it is possible to integrate the detector of an infrared camera which consists of thousands of pixels to the read-out electronics which is usually based on standard silicon CMOS devices. In this case, the active lead salt infrared detector elements are grown by molecular beam epitaxy on CaF_2/BaF_2 buffer layers deposited on the Si (111) substrates after the fabrication of the electronic circuitry for the read-out. Conventional HgCdTe focal plane technology relies on hybrid structures with In bump bonds between the infrared devices and the Si substrate for each of the pixels. Due to the rather high dielectric constants of the IV–VI compounds as compared to HgCdTe, the high-frequency operation of IV–VI detector structures is limited to about 100 MHz. This upper frequency limit is, however, sufficient for infrared focal plane arrays. With IV–VI active epilayer thicknesses of few μm internal quantum efficiencies close to 100% and external efficiencies of about 60% have been achieved. Using materials like PbS, PbTe, PbEuSe, PbEuTe, or PbSnSe, the cutoff wavelength of the detector array can be varied from 4 to 14 μm at $T = 77$ K. The sensitivity of these detector structures is limited by generation-recombination noise for $T < 100$ K, whereas for higher temperatures these sensors are diffusion limited. The variation of quantum efficiencies between different pixels was shown to be less than about 3% for a 256 pixel device. Since all growth steps for these infrared sensors are performed at temperatures below 450 °C, a processing compatibility exists with standard Si-VLSI technology.

Summary

The peculiar structural, electronic, and optical properties of IV–VI semiconductors as compared to other semiconductor materials are a consequence of the ten valence electrons per atomic pair instead of the eight valence electrons typical for the tetrahedrally bonded group IV, III–V, and II–VI semiconductors. The most important group of IV–VI materials is the so-called lead chalcogenide (lead salt) compounds and their ternary and quaternary alloys. Their electronic band structure with the narrow direct minimum gap not at the center but at the edges of the Brillouin zone (L-points) exhibits a number of unique properties. The minimum energy gap increases with increasing temperature and decreases with hydrostatic pressure. The refractive indices are particularly large. As compared to narrow-gap III–V or II–VI semiconductors the Auger recombination rates are by two orders of magnitude smaller. All these properties provide the lead salts with great advantages for achieving infrared p-n junction lasing operation at comparatively high temperatures. Furthermore, efficient microcavity designs with extremely high reflectivities of 99% were realized with only three layer periods of combination like $PbSrTe/BaF_2$ or PbEuTe/EuTe due to their extremely high refractive index contrast. Further applications of the lead compounds are high sensitive infrared detectors, including focal plane array detectors which can be fabricated with comparatively low cost technologies, as well as thermoelectric devices.

List of Units and Nomenclature

Absorption *constant*	α
Absorption index	κ
Composition	x
Damping parameter	Γ_D
Detectivity	D^*
Dielectric function	ε
Energy	E
Energy gap	E_g
Free electron mass	m_0
High frequency dielectric constant	ε_∞
Kramers Kronig relations	KKR
Longitudinal effective mass	m_l
Longitudinal g-factor	g_l
Pressure	p
Refractive index	n
Static dielectric constant	ε_0
Temperature	T
Transverse effective mass	m_t
Transverse g-factor	g_t

See also

Nonlinear Optics, Basics: Kramers–Krönig Relations in Nonlinear Optics.

Further Reading

Bauer G (1980) Magnetooptical properties of IV–VI compounds. In: Zawadzki W (ed.) *Narrow Gap Semiconductors Physics and Applications*, pp. 470–494. Berlin: Springer Verlag.

Bauer G and Krenn H (1985) Lead selenide, lead telluride. In: Palik ED (ed.) *Handbook of Optical Constants*, pp. 517–523 and 535–546. New York: Academic Press.

Felix CL, Bewley WW, Vurgaftman I, *et al.* (2001) Low-threshold optically pumped $\lambda = 4.4$ μm vertical cavity surface emitting laser with a PbS quantum well active region. *Applied Physics Letters* 78: 3770–3772.

Holloway G (1980) Thin film IV–VI semiconductor photodiodes. In: Hass G and Francombe MF (eds) *Physics of Thin Films*, vol. 11, pp. 105–203. New York: Academic Press.

Jantsch W (1983) Dielectric properties and soft modes in semiconducting (PbSnGe)Te. *Springer Tracts in Modern Physics* 99: 1–50.

Katzir A, Rosman R, Shani Y, *et al.* (1989) Tuneable lead salt lasers. In: Cheo PK (ed.) *Handbook of Solid State Lasers*, pp. 227–347. New York: Marcel Dekker.

Khoklov D (2003) Lead chalcogenioles: Physics and applications. In: Manasreh MO (ed.) *Optoelectronic Properties of Semiconductors and Superlattices*, vol. 18. New York: Taylor and Francis.

Lucovsky G, Martin RM and Burstein E (1974) Structure, bonding and lattice dynamical properties of the IV–VI compounds. In: Rabii S (ed.) *Physics of IV–VI Compounds and Alloys*, pp. 93–102. New York: Gordon and Breach.

Nimtz G (1993) IV–VI compounds. In: Hellwege K and Madelung O (eds) *Numerical Data and Functional Relationships in Science and Technology*. Landolt-Börnstein, new series, group III, vol. 17, Part F, pp. 152–180. Berlin: Springer Verlag.

Partin DL and Heremans J (1994) Growth of narrow bandgap semiconductors. In: Moss TS and Mahajan S (eds) *Handbook on Semiconductors*, vol. 3A, pp. 369–450. Amsterdam: North-Holland.

Preier H (1990) Physics and applications of IV–VI compound semiconductor lasers. *Semiconductor Science and Technology* 5: S12–S20.

Ravich YI, Efimova BA and Smirnov IA (1970) Semiconducting Lead Chalcogenides. New York: Plenum.

Schwarzl T, Heiss W and Springholz G (1999) Ultra-high-finesse IV–VI microcavities for the midinfared. *Applied Physics Letters* 75: 1246–1248.

Springholz G, Shi Z and Zogg H (1998) Molecular beam epitaxy of IV–VI semiconductors. In: Liu AWK and Santos MB (eds) *Heteroepitaxial Thin Film Systems*, pp. 621–688. Singapore: World Scientific.

Springholz G, Schwarzl T, Heiss W, *et al.* (2001) Mid-infrared surface-emitting PbSe/PbEuTe quantum dot-lasers. *Applied Physics Letters* 79–81: 1225.

Tacke M (1995) New developments and applications of tunable IR lead salt lasers. *Infrared Physics and Technology* 36: 447–463.

Yuan S, Springholz G, Bauer G and Kriechbaum M (1994) Electronic and optical properties of PBTe/$PB_{1-x}Eu_x$Te multiple quantum well structures. *Physical Review* B49: 5476–5489.

Zemel JN, Jensen JD and Schoolar RB (1965) Electrical and optical properties of epitaxial films of PbS, PbSe, PbTe and SnTe. *Physical Review* 140: A330–A337.

Zogg H, Fach A, John J, *et al.* (1995) Photovoltaic IV–VI on Si infrared sensor arrrays for thermal imaging. *Optical Engineering* 34: 1964–1969.

Mercury Cadmium Telluride

M B Reine, BAE SYSTEMS, Lexington, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Mercury-cadmium telluride (HgCdTe or $Hg_{1-x}Cd_x$Te) is a pseudobinary alloy semiconductor that crystallizes in the zincblende structure. The energy bandgap $E_G(x,T)$ varies continuously, and nearly linearly, with alloy composition parameter x, ranging from 1.6 eV for the wide-gap semiconductor CdTe to −0.3 eV for the semimetal HgTe. The bandgap passes through zero at an alloy composition of $x = 0.16$ (for $T = 0$ K). This broad range of energy bandgap provides a continuum of materials whose absorption edge can be tailored to provide high-performance quantum detectors for wavelengths over the 1–30 μm spectral range. Direct allowed valence-to-conduction band transitions are responsible for large values of the optical absorption coefficient, which enable quantum efficiencies approaching 100% to be achieved in practical device designs. Long minority carrier lifetimes result in low thermal noise, and permit high performance to be achieved at operating temperatures that are the highest reported among infrared detectors of comparable cutoff wavelengths. HgCdTe is well-established today as

the most broadly applicable high-performance infrared detector material.

This article summarizes the fundamental properties of this versatile alloy semiconductor, and relates the material properties to its success as an infrared photoconductive and photovoltaic detector material.

Historical Background

It is noteworthy that HgCdTe is probably the first example of a semiconductor material that was intentionally created to satisfy well-defined theoretical device requirements. The importance of the infrared atmospheric transmission window at 8–12 μm was well known for thermal imaging systems, which enable night vision by imaging emitted infrared radiation from the scene rather than reflected visible radiation. The only available infrared detectors with the required sensitivity at 8–12 μm were extrinsic photoconductors such as copper-doped germanium, in which the photoinduced transitions were between a donor or acceptor level and the adjacent band extremum. However, such extrinsic detectors required cooling to prohibitively low temperatures, such as 10–15 K, which made thermal imaging systems bulky and impractical.

Theory showed that the reasons for such a low operating temperature were fundamental to the extrinsic detector material itself – essentially a short carrier lifetime that resulted in a large thermal generation rate and a correspondingly large thermal noise. Theory also showed that an intrinsic detector, in which the photoinduced transitions were direct (vertical) transitions between the valence band and the conduction band, could have significantly longer carrier lifetimes. This meant that an intrinsic detector, with a bandgap of 0.1 eV that matched the 8–12 μm band, could achieve the required sensitivity at much higher operating temperatures, as high as 77 K, that of liquid nitrogen. These theoretical and practical considerations motivated the search in the late 1950s for a semiconductor with a direct bandgap of 0.1 eV.

The first report of the synthesis of the semimetal HgTe and the wide-bandgap semiconductor CdTe to form the semiconductor alloy HgCdTe was published in 1959 by the Royal Radar Establishment in Malvern, England. This seminal paper reported both photoconductive and photovoltaic response at wavelengths extending to 12 μm, and made the understated observation that this material showed promise for intrinsic infrared detectors.

Soon thereafter, working under a US Air Force contract with the objective of devising an 8–12 μm background-limited semiconductor infrared detector that would operate at temperatures as high as 77 K, the group led by Paul Kruse at the Honeywell Corporate Research Center in Hopkins, Minnesota developed a modified Bridgman crystal growth technique for HgCdTe. In 1962 they reported both photoconductive and photovoltaic detection in rudimentary HgCdTe devices.

Fundamental Material Properties

Three fundamental properties determine the degree to which a semiconductor will make a successful infrared detector. First, the energy bandgap must be approximately equal to the smallest photon energy to be detected. There can be severe cooling penalties if the bandgap is significantly smaller than the smallest photon energy of interest. Second, the optical absorption coefficient must be large enough so that nearly all incident photons of interest are converted to electron–hole pairs for practical values of device thickness. Third, the inherent recombination mechanisms must allow long carrier lifetimes to be realized so that thermal generation rates, and hence thermal noise, can be minimized for practical devices and operating conditions.

The three key features that make HgCdTe the nearly ideal infrared detector material – tailorable energy band gap over the 1–30 μm range, large optical absorption coefficients that enable high quantum efficiency, and favorable inherent recombination mechanisms that lead to high operating temperature – are direct consequences of the energy band structure of this zincblende semiconductor. They apply no matter whether the device embodiment is a photoconductor, a photodiode or a metal–insulator–semiconductor detector. In addition, HgCdTe has many other favorable material properties, summarized in this article, that are technologically important for realizing practical infrared detectors. Tables 1 and 2 summarize various material properties for $Hg_{1-x}Cd_xTe$. **Table 1** summarizes material properties for $Hg_{1-x}Cd_xTe$ that vary with alloy composition x and temperature T: lattice constant a, energy band gap E_G, cutoff wavelength λ_{CO} ($= 1.24/E_G$), intrinsic carrier concentration n_i, conduction band effective mass ratio m_C/m_0, conduction band g-factor g_C, static and high frequency dielectric constants $\varepsilon_S/\varepsilon_0$ and $\varepsilon_\infty/\varepsilon_0$, index of refraction $n_R = (\varepsilon_\infty/\varepsilon_0)^{1/2}$, electron and heavy hole mobilities μ_e and μ_{hh}, mobility ratio b, radiative and Auger-1 lifetimes τ_{RAD} and $\tau_{Auger\text{-}1}$, and typical lifetimes $\tau_{typical}$ observed in n-type HgCdTe of low carrier concentration ($<1 \times 10^{15} cm^{-3}$).

Table 1 Material properties for selected compositions of $Hg_{1-x}Cd_xTe$, including the binary components HgTe and CdTe, and several technologically important alloy compositions. The radiative and Auger-1 lifetimes τ_{RAD} and $\tau_{Auger\text{-}1}$ are calculated for n-type HgCdTe with $N_D = 1 \times 10^{15}$ cm^{-3}

Property	*HgTe*	*$Hg_{1-x}Cd_xTe$*						*CdTe*
x	0	0.194	0.205	0.225	0.31	0.44	0.62	1.0
a (Å)	6.461	6.464	6.464	6.464	6.465	6.468	6.472	6.481
T (K)	77	77	77	77	140	200	250	300
E_G (eV)	−0.261	0.073	0.091	0.123	0.272	0.474	0.749	1.490
λ_{CO} (μm)		16.9	13.6	10.1	4.6	2.6	1.7	0.8
n_i (cm^{-3})		1.9×10^{14}	5.8×10^{13}	6.3×10^{12}	3.7×10^{12}	7.1×10^{11}	3.1×10^{10}	4.1×10^{5}
m_C/m_0		0.006	0.007	0.010	0.021	0.035	0.053	0.102
g_C		−150	−118	−84	−33	−15	−7	−1.2
$\varepsilon_S/\varepsilon_0$	20.0	18.2	18.1	17.9	17.1	15.9	14.2	10.6
$\varepsilon_\infty/\varepsilon_0$	14.4	12.8	12.7	12.5	11.9	10.8	9.3	6.2
n_R	3.79	3.58	3.57	3.54	3.44	3.29	3.06	2.50
μ_e (cm^2/V-s)		4.5×10^{5}	3.0×10^{5}	1.0×10^{5}				
μ_{hh} (cm^2/V-s)		450	450	450				
$b = \mu_e/\mu_{hh}$		1000	667	222				
τ_{RAD} (μs)		16.5	13.9	10.4	11.3	11.2	10.6	
$\tau_{Auger\text{-}1}$ (μs)		0.45	0.85	1.8	40	450	4.8×10^{3}	
$\tau_{typical}$ (μs)		0.4	0.8	1	7			

Table 2 Material properties for $Hg_{1-x}Cd_xTe$ that are independent of or relatively insensitive to alloy composition

Property	*Symbol*	*Value*
Kane matrix element	E_P	19 eV
Split-off band energy	Δ	0.93 eV
Heavy-hole effective mass ratio	m_{hh}/m_0	0.40–0.53
Valence band offset	E_v(HgTe) − E_v(CdTe)	0.35–0.55 eV

Energy Band Gap and Band Structure

The energy band gap $E_G(x, T)$ of $Hg_{1-x}Cd_xTe$ varies continuously, and nearly linearly, with alloy composition parameter x, ranging from 1.6 eV for the wide-gap semiconductor CdTe to −0.3 eV for the semimetal HgTe, as shown in Figure 1a. Also plotted in Figure 1a is the cutoff wavelength $\lambda_{CO}(x, T)$ ($\approx hc/E_G(x, T)$, where h is Planck's constant and c is the speed of light), defined as that wavelength at which the photoresponse falls to 50% of its peak value. The energy bandgap $E_G(x, T)$, defined at the difference between the Γ_6 and Γ_8 band extrema at $\Gamma = 0$, passes through zero at an alloy composition of $x = 0.16$ (for $T = 0$ K), as illustrated in Figure 1b. This broad range of energy bandgap provides a continuum of materials whose absorption edge can be tailored to provide high-performance quantum detectors for wavelengths over the 1–30 μm spectral range.

Optical Absorption Coefficient

Optical absorption coefficient data for several alloy compositions of $Hg_{1-x}Cd_xTe$ are plotted versus wavelength in Figure 2. Several trends are evident. The absorption strength generally decreases as the gap becomes smaller due both to the decrease in the conduction band effective mass and to the $\lambda^{-1/2}$ dependence of the absorption coefficient on wavelength λ. There is an exponential tail at energies just below the bandgap energy.

Carrier Lifetime Mechanisms

There are two fundamental mechanisms that determine the minority carrier lifetime in defect-free HgCdTe: radiative recombination and Auger recombination (Auger-1 in n-type and Auger-7 in p-type). Because they involve only interactions among electrons and holes in the valence and conduction bands, and do not require the intermediary of defect states, these mechanisms are regarded as inherent to the material itself, depending only on the band structure, the donor or acceptor concentration, and the temperature. The important advantage that HgCdTe enjoys as a high-performance detector material is that these fundamental mechanisms provide, for practically achievable

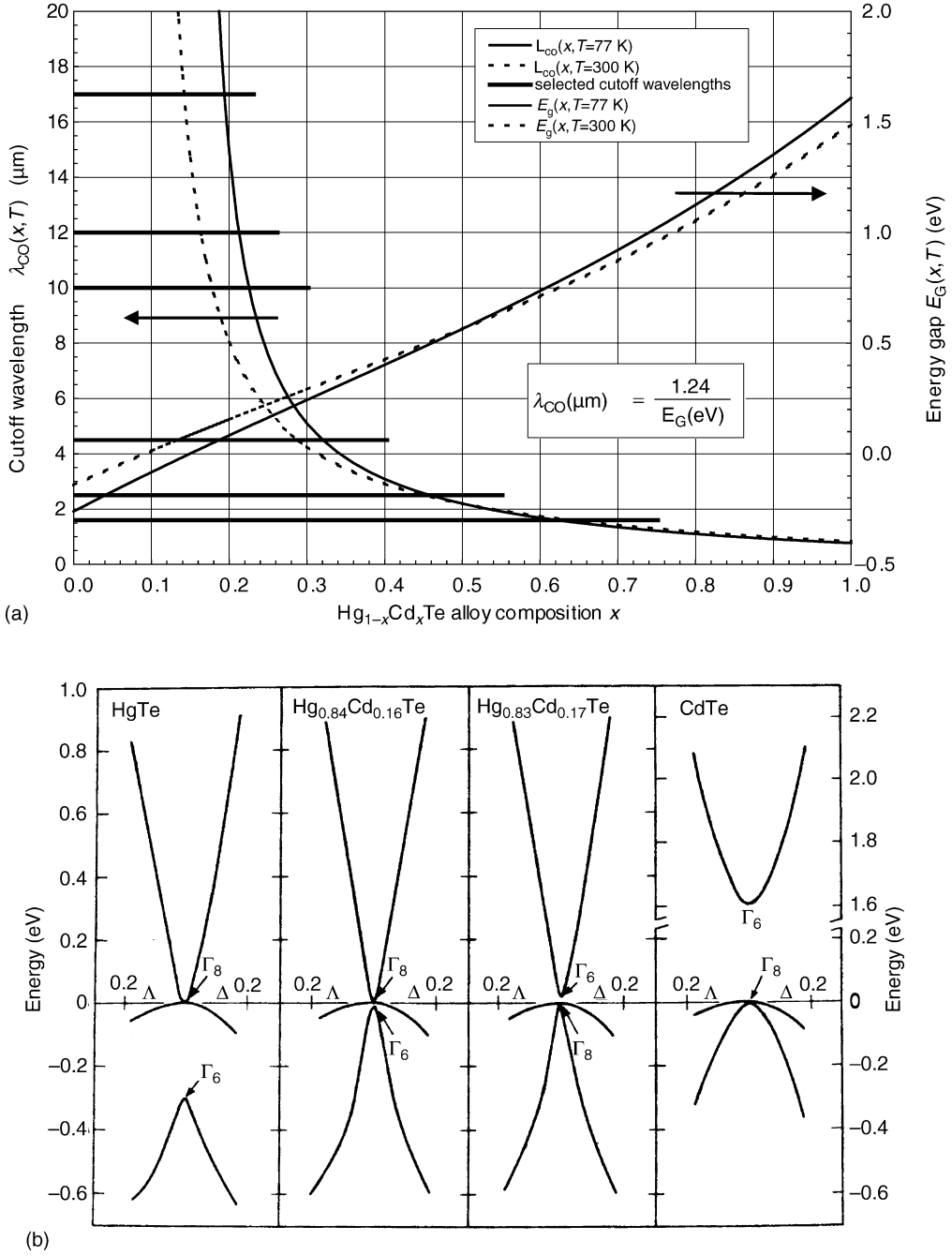


Figure 1 (a) Energy bandgap $E_G(x,T)$ and cutoff wavelength $\lambda_{CO}(x,T)$ for $Hg_{1-x}Cd_xTe$, plotted versus alloy composition parameter x, for temperatures of 77 K (solid curves) and 300 K (dashed curves). Horizontal bars indicate important cutoff wavelengths for HgCdTe detector technology. (b) Band diagrams near the Γ point calculated for HgTe, for two compositions of $Hg_{1-x}Cd_xTe$, and for CdTe, illustrating the transition from negative to positive energy bandgap. The energy bandgap is defined at the difference between the Γ_6 and Γ_8 band extrema at $\Gamma = 0$. Part (b) reproduced with permission from Chadi DJ and Cohen ML (1973) Electronic structure of $Hg_{1-x}Cd_xTe$ alloys and charge-density calculations using representative K points. *Physical Review B* 7: 692–699.

carrier concentrations, lifetimes that are long enough (and hence thermal generation rates that are low enough) that background-limited sensitivity is achieved at temperatures substantially higher than for other classes of infrared detectors, such as extrinsic detectors (As-doped Si, Cu-doped Ge) and quantum-well infrared photodetectors (QWIPs).

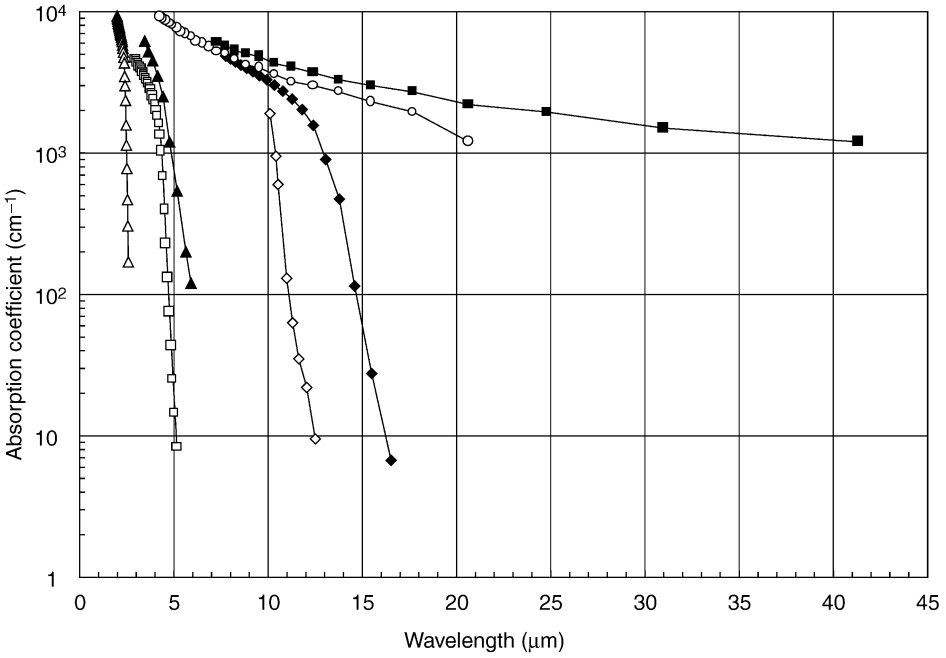


Figure 2 Optical absorption coefficient data for several $Hg_{1-x}Cd_xTe$ alloy compositions, for photon energies near the fundamental absorption edge, plotted versus wavelength.

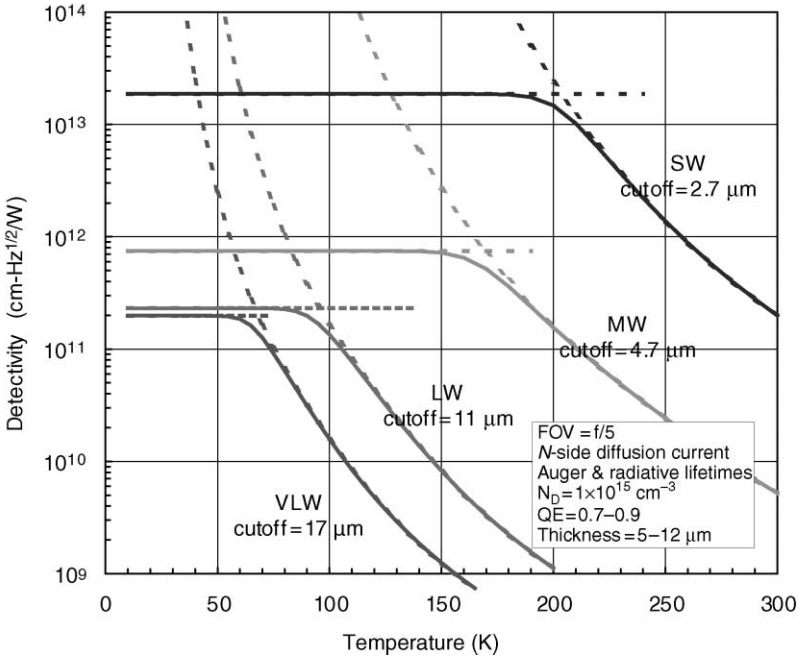


Figure 3 Calculated detectivity (D^*) for $Hg_{1-x}Cd_xTe$ infrared detectors for four technologically important wavelength regions, plotted versus operating temperature.

Figure 3 illustrates the sensitivities and operating temperatures that can be achieved for HgCdTe infrared detectors for four important wavelength regions. Detectivity, D^*, is the normalized signal-to-noise ratio that is the key figure of merit for the sensitivity of an infrared detector. At lower temperatures, detector thermal noise is negligible, and D^* is limited by detector noise due to

fluctuations in the arrival rate of photons from the thermal room-background radiation. This limit is referred to as the BLIP (background-limited infrared photodetector) limit, and is indicated in Figure 3 by the horizontal dashed lines. As detector temperature increases, the detector thermal noise increases exponentially, and eventually overcomes the background noise, causing D^* to decrease exponentially for further increases in temperature. Detector thermal noise is proportional to the thermal generation rate, which is inversely proportional to the carrier lifetime.

n-Type and p-Type Doping

HgCdTe can be made n-type or p-type by a number of relatively convenient methods, at carrier concentrations required for high-performance n-type photoconductor and p-n junction photodiode architectures. Common donors are indium and iodine. Arsenic has become the most commonly used acceptor dopant, although copper and gold are used in some cases. Some HgCdTe photodiode designs still employ native metal-vacancy point defects as acceptors, although the clear trend is toward the use of extrinsic acceptor doping to avoid the strong Shockley–Read recombination associated with the Hg vacancy. Residual electrically active impurity concentrations are generally less than $1 \times 10^{14}\,\mathrm{cm}^{-3}$, which allows controllable doping at low-$10^{14}\,\mathrm{cm}^{-3}$ concentrations.

Crystal Growth Methods

The first growth methods for HgCdTe were bulk growth. Liquid phase epitaxy (LPE) began to be developed around 1975, followed by metalorganic chemical vapor deposition (MOCVD) and molecular beam epitaxy (MBE) in the early 1980s. Bulk growth, LPE, MOCVD, and MBE techniques have been refined through the years, and each is still in use today. Each technique has found application for certain types of HgCdTe devices. Bulk growth methods are used for production quantities of n-type wafers, with diameters of 12–20 mm, for LW and VLW photoconductive arrays, which are more tolerant of dislocations. LPE is in use today for production quantities of photovoltaic (PV) HgCdTe arrays, and has also been used for certain photoconductive HgCdTe arrays. Vapor-phase epitaxy (VPE) methods such as MBE and MOVPE are in use for engineering and prototype quantities. Both run-to-run reproducibility and control as well as intra-wafer uniformity of HgCdTe alloy composition are well in hand, meeting or exceeding requirements.

LPE-based one-layer and two-layer photodiode technology is well established today, offering the lowest defect densities and highest operabilities for production quantities of high performance FPAs. The VPE methods, MBE and MOVPE, will probably not entirely replace LPE in the near future, but will continue to improve, particularly with respect to lower defect densities, offering *in situ* growth of advanced bandgap-engineered detector designs that are unwieldy or impossible with LPE, such as dual-band detectors, avalanche photodiodes, and planar buried junctions with *in situ* CdTe passivation. MBE has shown the potential for rapidly switching from growth of one HgCdTe alloy composition to another in successive growth runs, without the necessity of preparatory calibration growth runs. This compositional agility, important for the rapid and cost-effective adjustment from product to product, will be further enhanced by the ongoing efforts to implement and improve *in situ* real-time monitors and feedback controls.

Substrates for Epitaxial Growth

The lattice mismatch between HgTe and CdTe is small, approximately 0.3%. This allows epitaxial growth of high-quality HgCdTe films on IR-transparent CdTe or nearly-lattice-matched IR-transparent $Cd_{1-z}Zn_zTe$ ($z \approx 0.04$) substrates, with dislocation densities in the mid-$10^4\,\mathrm{cm}^{-2}$ range. It also allows the *in situ* growth of various isotype and anisotype heterojunctions with tolerably low or negligible densities of misfit dislocations. This has led to an increasing number of bandgap-engineered HgCdTe photodiode structures, such as dual-band or two-color detector arrays and Auger-suppressed photodiodes with the potential for greatly increased operating temperature.

Sapphire and silicon are IR-transparent substrates that are less costly, that are available in much larger areas, and that are more rugged than CdTe and CdZnTe. Both sapphire and silicon can be used for epitaxial growth of HgCdTe films with dislocation densities that are acceptably low (mid-$10^6\,\mathrm{cm}^{-2}$) for many important photodiode applications, such as for the MW and SW spectral ranges, and for the LW spectral range for high background photon fluxes. LPE growth of HgCdTe($x = 0.4$) is done on 3-inch diameter sapphire substrates with MOCVD-grown CdTe buffer layers. MBE growth of HgCdTe($x = 0.3$) on 4-inch diameter (211) silicon substrates, with ZnTe/CdTe buffer layers, has recently been reported.

Dielectric Constant

The relatively low dielectric constant ($\varepsilon_S = 18\varepsilon_0$) of HgCdTe, in contrast to those of PbTe and PbSnTe, permits low junction capacitance. This is important for fast response in laser pulse detectors where small $R-C$ time constants are needed. It is also important for suppressing preamplifier noise below the detector noise, and thereby achieving high detectivity (D^*), in photodiodes operating at low temperature and low background fluxes, where the junction resistance is sufficiently high that the junction capacitance becomes an important contribution to the junction impedance at frequencies of interest.

Surface Passivation

Passivation of the exposed HgCdTe surfaces is critical for both device performance and long-term stability. The favorable surface properties of HgCdTe have enabled practical surface passivation schemes to be developed for both n-type photoconductors and p-n junction photodiodes of both planar and mesa configurations.

n-type HgCdTe photoconductors are commonly passivated with a native oxide that accumulates the surface, resulting in surface recombination velocities as low as 50 cm s^{-1}, along with low $1/f$ noise (knee frequencies less than 50 Hz). The surface accumulation layer also provides a layer of surface electrons with high mobility, though lower than the bulk electron mobility, such that the shunt conductance of this layer is tolerably low compared to the bulk conductance of the detector itself. Several methods have been used for growing this native oxide, including plasma deposition and electrochemical (anodic) oxidation.

HgCdTe photovoltaic detectors require quite different passivation because usually both n-type and p-type surfaces are exposed as well as the depletion region. The CdTe/HgCdTe heterostructure is an important passivation for HgCdTe photovoltaic detectors and arrays. CdTe passivation has enabled arrays of HgCdTe photodiodes with negligible $1/f$ noise (knee frequencies less than 1 Hz), radiation hardness well in excess of the 20–40 kRad(Si) usually encountered in Earth-resource space missions, and excellent stability to thermal bake and to exposure to moisture.

CdTe has a number of important advantages as a passivant for HgCdTe junctions. It is a 'native' passivant that is chemically compatible with HgCdTe. Adhesion is excellent. The valence band edge of CdTe is approximately 0.4 eV below that of HgTe, thereby allowing repulsive barriers to form in the underlying HgCdTe for both electrons in the conduction band and holes in the valence band. Both mesa and planar photodiodes have been successfully passivated with CdTe. A wide variety of methods has been reported for deposition of CdTe passivating layers, including MBE, MOVPE, e-beam evaporation, sputtering, and hot wall epitaxy.

Electron-to-Hole Mobility Ratio

The small values for the conduction band effective mass ratio lead to large values for the electron mobility, as large as 4.5×10^5 cm^2/V-s for $x = 0.195$ at 77 K. The heavy hole effective mass ratio is quite large, approximately 0.5 for all alloy compositions, leading to low values for the heavy hole mobility, on the order of 500 cm^2/V-s. This gives large values for the electron-to-heavy hole mobility ratio, b, as high as 1000 for $x = 0.194$ at 77 K.

These large values of b were of great benefit to n-type HgCdTe photoconductors. Most such devices have small active areas, on the order of 50×50 μm^2, and the drift length for minority carriers (holes) becomes comparable to the interelectrode spacing at modest bias voltages. In this case, when the contacts have a high recombination velocity, the photoconductive gain saturates to the value $b/2$. High values of b in HgCdTe allowed sufficiently high photoconductive gains to be achieved to raise the detector noise above the preamp noise, an essential condition for achieving BLIP sensitivity.

HgCdTe Infrared Detector Configurations

Three different types of infrared detectors can be realized in the $Hg_{1-x}Cd_xTe$ alloy semiconductor: the photoconductive (PC) detector, the photodiode, also referred to as the photovoltaic (PV) detector, and the metal–insulator–semiconductor (MIS) detector. These are illustrated in Figure 4.

In all three types of HgCdTe detector – PC, PV, and MIS – excess electron–hole pairs are photogenerated by the same mechanism: valence-to-conduction band absorption of infrared radiation with photon energy greater than the bandgap energy E_G. All three types can achieve high quantum efficiencies, approaching 100% for well-designed devices. All three types have the same fundamental recombination mechanisms (Auger-1 and radiative) that determine the highest possible operating temperature to achieve a given sensitivity (signal-to-noise ratio).

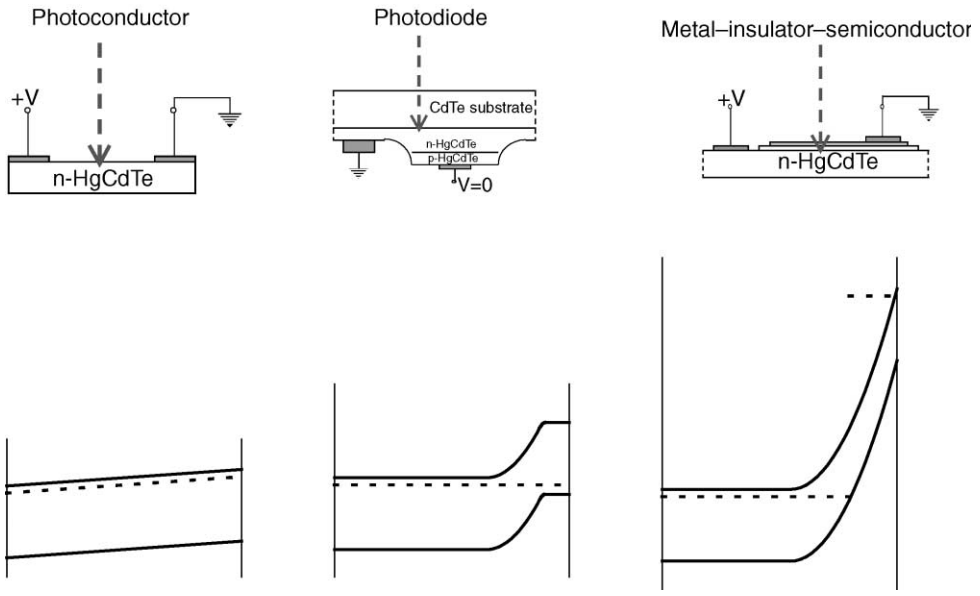


Figure 4 Three different types of infrared detectors have been realized in the $Hg_{1-x}Cd_xTe$ alloy semiconductor: the photoconductive (PC) detector, the photodiode or the photovoltaic (PV) detector, and the metal–insulator–semiconductor (MIS) detector. HgCdTe PC and PV detector arrays are being manufactured for a wide variety of applications. The MIS detector was abandoned in the mid-1980s because it placed unachievable demands on material quality.

The differences among PC, PV, and MIS HgCdTe detectors are due to the way in which the photogenerated electron–hole pairs are manifested electrically in the terminal characteristics of the device. In the PC detector, the excess electron–hole pairs are sensed as a small increase in the conductivity of what is basically a two-terminal resistor. In the PV detector, the excess electron–hole pairs are sensed as a photocurrent in a short-circuited p-n junction photodiode. In the MIS detector, the excess electron–hole pairs are sensed as a small change in the voltage due to minority photocarriers filling a transient charge storage well that has been biased to deep depletion.

PC HgCdTe Detectors

The HgCdTe PC detector is a low-resistance two-terminal device. Resistances are typically 25–100 ohms per square. A dc bias current is applied to convert the conductivity change to an observable voltage change. The device operates under near-equilibrium conditions, with small dc bias voltages typically on the order of 0.1 V. The dc bias electric field within the photoconductor is quite small, generally in the 20–50 $V\,cm^{-1}$ range. The low-resistance device requires a low-noise bipolar preamplifier, usually external to the dewar. Each PC detector element requires one lead through the dewar wall, thus limiting the practical number of PC elements in an array to about 200–300.

The basic HgCdTe photoconductor is a rectangular area, photolithographically defined in an n-type HgCdTe layer approximately 8–15 μm thick, with two ohmic contacts on opposite edges. Surface passivation is straightforward: a thin native oxide strongly accumulates the surface of n-type HgCdTe, thereby reducing the surface recombination velocity to negligibly small values.

Many useful variations on the simple rectangular geometry are possible because of the favorable material properties of HgCdTe. Several contact geometries have been devised to minimize the recombination of photocarriers at the contacts. Serpentine designs have been used to increase detector resistance and reduce bias power dissipation, especially important issues for large-area and very long-wavelength detectors. The SPRITE (signal processing in the element) design reduces the number of electrical connections and dewar leads for scanned arrays by performing both detection and integration within an elongated HgCdTe bar.

PV HgCdTe Detectors

The HgCdTe photodiode is a more complicated device than the photoconductor, requiring both n-type and p-type layers, and having a surface

depletion region that makes stringent demands on surface passivation technology.

The PV detector is a high-resistance p-n junction device. The photodiode is operated very close to equilibrium conditions, either at zero bias voltage or at a small reverse bias voltage whose magnitude is several kT, typically 20–30 mV. The absorber layer is usually n-type HgCdTe about 5–15 μm thick, with the thicker values required for longer-wavelength radiation. A p-n junction is formed by a thin p-type layer. A wide variety of junction formation methods are used. Wide-gap-p on narrow-gap-n heterojunctions are formed by two-layer LPE. Planar p-on-n or n-on-p junctions are formed by arsenic or boron implantation into n-type or p-type layers grown by LPE or MBE. Vertical-geometry p-on-n junctions are formed by a damage mechanism during ion beam milling. Both single-junction and multijunction devices are grown *in situ* by vapor phase epitaxial methods such as MBE and MOVPE. Surface passivation is generally accomplished by CdTe. The high-resistance PV device can match well with low-noise silicon CMOS preamplifiers that can be integrated with two-dimensional arrays to form large back-illuminated hybrid focal plane arrays (FPAs). Typical array formats for HgCdTe FPAs are 256×256 and 480×640, and can be as large as 1024×1024 and 2048×2048.

MIS HgCdTe Detectors

The metal–insulator–semiconductor (MIS) detector, also referred to as a photocapacitor, is usually formed on an n-type HgCdTe absorber layer. The insulator of choice is a thin native oxide. The gate electrode is a thin semitransparent metal film.

In contrast to the PC and PV detectors, the MIS detector operates in a strongly nonequilibrium mode.

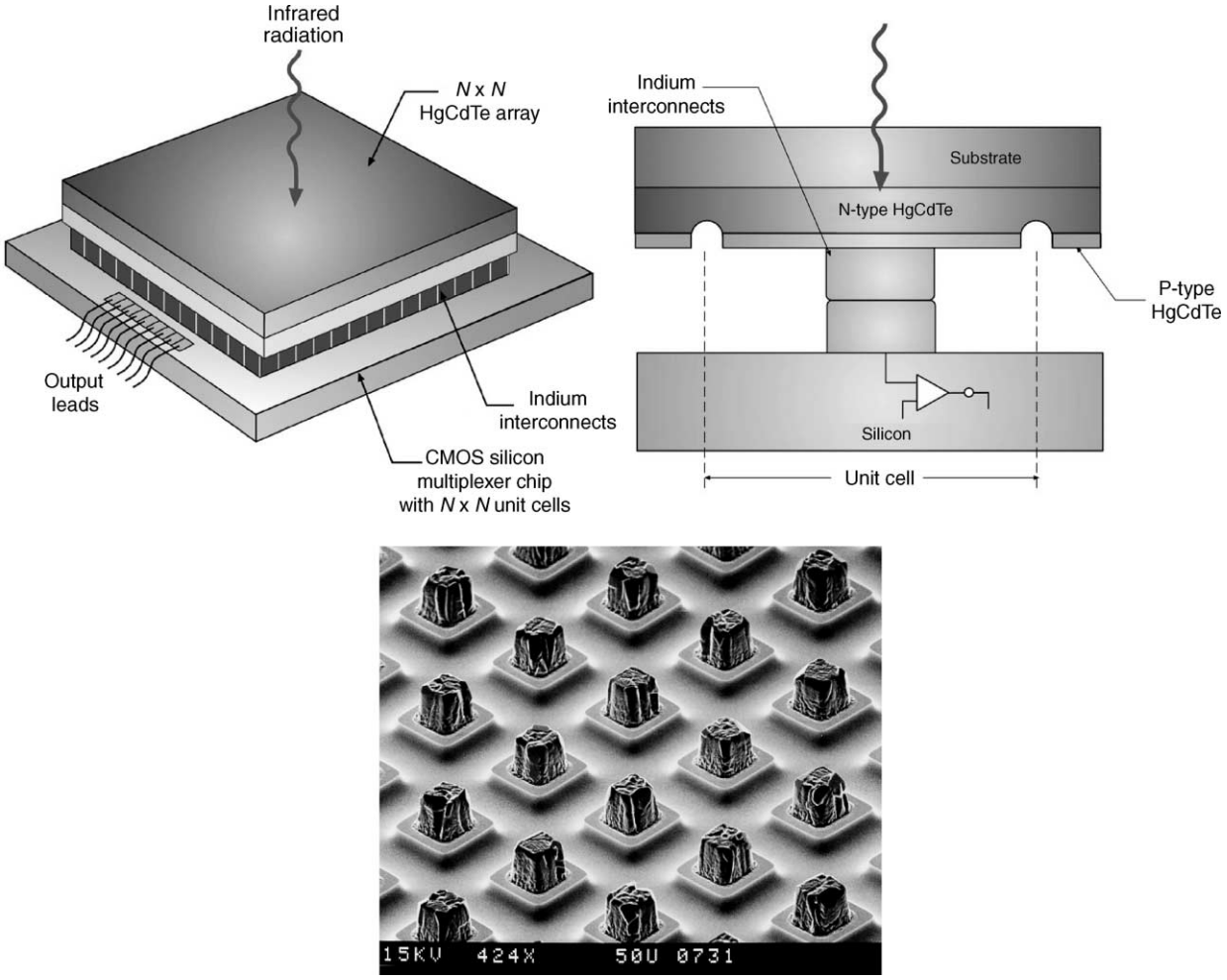


Figure 5a Back-illuminated bump-interconnected HgCdTe FPA architecture. The SEM photo shows an LPE P-on-n HgCdTe mesa photodiode array with 60×60 μm^2 unit cells and 33×33 μm^2 mesa (junction) areas.

A bias voltage pulse of several volts is applied across the capacitor to drive the HgCdTe surface into deep depletion. Excess electron–hole pairs are photogenerated in the absorber layer and diffuse to the edge of the depletion region, where the holes are collected in the charge well. After a period of time, called the integration time, the voltage across the capacitor is sensed, giving a measure of how much charge was collected, and the cycle is repeated.

Because of the nonequilibrium operation of the MIS detector, much larger electric fields are set up in the depletion region than in the p-n junction, resulting in defect-related tunneling dark current that is orders of magnitude larger than the fundamental dark current. The MIS detector requires much higher material quality than PV or PC detectors, which still has not been achieved. For this reason, all development of the HgCdTe MIS detector was abandoned in the mid-1980s.

Two-Dimensional HgCdTe Infrared Focal Plane Arrays

Among the most important applications of HgCdTe are large two-dimensional electronically scanned hybrid arrays, referred to as focal plane arrays (FPAs). A hybrid HgCdTe FPA consists of a two-dimensional HgCdTe photovoltaic detector array that is interfaced electrically, thermally and mechanically with a matching two-dimensional array of input circuits in a silicon CMOS ROIC chip. Each HgCdTe detector element in the array has one electrical connection to its matching input circuit node in the silicon ROIC.

The thermal expansion coefficient of HgCdTe is sufficiently close to that of silicon to allow several technologically viable hybrid arrangements of HgCdTe detector arrays and silicon multiplexer chips.

There are two main types of hybrid HgCdTe FPAs being developed and manufactured today: the back-illuminated bump-interconnected configuration, shown in Figure 5a, and the front-illuminated or 'loophole' configuration shown in Figure 5b. Each configuration successfully solves the problem of the thermal expansion mismatch between silicon and HgCdTe in different ways, and each requires fundamentally different HgCdTe photodiode designs and processing.

In the back-illuminated configuration, illustrated in Figure 5a, the HgCdTe detector array is bump-mounted onto the silicon ROIC by cold-welded or thermally reflowed indium interconnects. Incident radiation reaches the detectors through an infrared-transparent substrate such as CdTe, CdZnTe, sapphire, or silicon.

In the front-illuminated (or 'loophole') FPA architecture, shown in Figure 5b, the thin HgCdTe layer is epoxied to the silicon ROIC chip, and undergoes elastic deformation when the FPA is cooled. This allows large arrays to be made without need for engineering the thermal expansion of the silicon ROIC chip. Low-temperature processing techniques

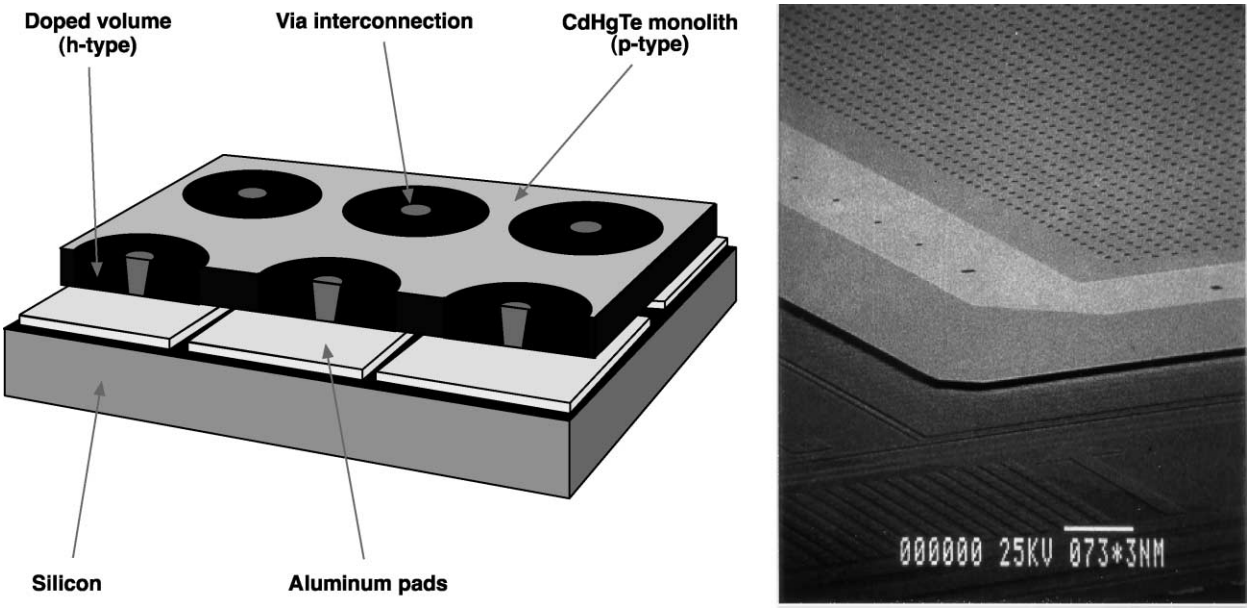


Figure 5b Front-illuminated loophole HgCdTe FPA architecture. The SEM photo shows one corner of a FPA. The metalized via holes are on 20 μm centers.

for junction formation and passivation are required because of the presence of the epoxy.

Conclusions and Trends

There has been enormous progress in HgCdTe material science and device technology over the past 45 years. This progress has been fueled by the many military and space applications for which HgCdTe PC and PV infrared detectors provide nearly ideal solutions. It is important to recognize that the progress so far in HgCdTe detector technology has been possible only through advances in basic materials growth and processing technology.

Further progress can be expected. The fundamental performance limits of HgCdTe photodiodes have not yet been reached. Continued reduction of material defects, both grown-in and process-induced, will reduce thermal generation rates, increase junction impedance, and allow higher operating temperatures for a number of important applications. Continued development of the *in situ* VPE growth methods – MBE and MOVPE – will allow bandgap-engineered heterojunction devices of increasing quality and complexity. Continued development of VPE growth on alternative substrates such as silicon will reduce the cost of two-dimensional arrays, and will make larger arrays practical.

List of Units and Nomenclature

BLIP	Background limited infrared photodetector
CdTe	Cadmium telluride
CdZnTe	Cadmium zinc telluride
CMOS	Complementary metal-oxide-semiconductor
FPA	Focal plane array
HgCdTe	Mercury cadmium telluride
$Hg_{1-x}Cd_xTe$	Mercury cadmium telluride
IR	Infrared
LPE	Liquid phase epitaxy
LW	Long wavelength
LWIR	Long wavelength infrared
MBE	Molecular beam epitaxy
MCT	Mercury cadmium telluride
MIS	Metal-insulator-semiconductor
MOCVD	Metal-organic chemical vapor epitaxy
MOVPE	Metal-organic vapor phase epitaxy
MW	Medium wavelength
MWIR	Medium wavelength infrared
PC	Photoconductive
PV	Photovoltaic
QWIP	Quantum well infrared photodetector
ROIC	Readout integrated circuit
SPRITE	Signal processing in the element
SW	Short wavelength
SWIR	Short wavelength infrared
VLW	Very long wavelength
VLWIR	Very long wavelength infrared
VPE	Vapor phase epitaxy

See also

Semiconductor Physics: Band Structure and Optical Properties; Impurities and Defects; Outline of Basic Electronic Properties.

Further Reading

Capper P (ed.) (1994) *Properties of Narrow Gap Cadmium-based Compounds*. EMIS Datareviews Series No. 10. London: INSPEC, Institution of Electrical Engineers. Reviews and compilations of fundamental and technological properties of HgCdTe and CdTe/CdZnTe/CdTeSe.

Capper P (ed.) (1997) *Narrow-Gap II–VI Compounds for Optoelectronic and Electromagnetic Applications*. London: Chapman and Hall. Twenty review chapters on growth techniques, materials characterization, and device applications of HgCdTe.

Capper P and Elliott CT (eds) (2000) *Infrared Detectors and Emitters: Materials and Devices*. Boston, MA: Kluwer Academic Publishers. Review chapters on HgCdTe materials, photoconductive and nonequilibrium HgCdTe detectors, photovoltaic HgCdTe detectors, and HgTe-CdTe superlattice detectors.

Rogalski A, Adamiec K and Rutkowski J (2000) *Narrow Gap Semiconductor Photodiodes*. Washington, DC: SPIE Press.

Willardson RK and Beer AC (eds) (1981) *Defects, (Hg,Cd)Se, (Hg,Cd)Te*. Semiconductors and Semimetals, vol. 16. New York: Academic Press. Review chapters on magneto-optical properties of HgCdTe and nonlinear optical effects in HgCdTe.

Willardson RK and Beer AC (eds) (1981) *Mercury Cadmium Telluride*. Semiconductors and Semimetals, vol. 18. New York: Academic Press. Seven review chapters: history of HgCdTe, preparation of high-purity Hg, Cd and Te, crystal growth of HgCdTe, Auger recombination in HgCdTe, and photoconductive, photovoltaic and metal-insulator-semiconductor HgCdTe detectors.

Modulation Spectroscopy of Semiconductors and Semiconductor Microstructures

Y-S Huang, National Taiwan University of Science and Engineering, Taipei, Taiwan

F H Pollak, Brooklyn College of the City University of New York, Brooklyn, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Modulation spectroscopy is a versatile and powerful optical technique for obtaining valuable information relevant to a large variety of semiconductor systems including bulk/thin films, surfaces/interfaces [semiconductor/air (vacuum), semiconductor/electrolyte, semiconductor/metal, semiconductor/semiconductor (homo- and heterojunctions)], micro- and nanostructures (quantum wells, multiple quantum wells, superlattices, quantum wires, quantum dots), the effects of growth/processing, as well as the characterization of actual device structures [heterojunction bipolar transistors (HBTs), pseudomorphic high electron mobility transistors (PHEMTs), edge-emitting and vertical-cavity surface-emitting (VCSEL) quantum well lasers, etc.].

Modulation spectroscopy deals with the measurement and interpretation of changes in the optical response of a sample which are effected by the modification of the measurement conditions. This can easily be accomplished by periodically modulating either (1) the measurement conditions themselves ('internal' modulation) or (2) some parameters applied to the sample ('external' modulation). The observed normalized changes are usually small so that the difference signals are closely related to a derivative of the absolute spectrum with respect to the modifying parameters. The derivative nature of modulation spectroscopy emphasizes structure localized in the photon energy region of the interband (intersub-band) transitions of semiconductors (semiconductor microstructures) and suppresses uninteresting background effects. As a result, weak features that may not be detected in the absolute spectra are often enhanced. Because of this derivative-like nature a large number of sharp spectral features may be observed, even at room temperature.

The ability to perform a lineshape fit is one of the great advantages of modulation spectroscopy. Since for the modulated signal the features are localized in photon energy it is possible to account for the lineshapes to yield accurate values of energies and broadening parameters of the associated interband transitions. For example, the energies of various interband (intersub-band) transitions may be evaluated to within a few meV, even at room temperature. Thus, the effects of static external perturbations such as electric and magnetic fields, temperature, hydrostatic pressure, uniaxial stress, etc., can be conveniently studied.

A particularly useful form of modulation spectroscopy is electromodulation (EM) since it is sensitive to surface/interface electric fields and can be performed in contactless modes that require no special mounting of the sample. Under appropriate conditions the EM spectrum may exhibit above bandgap oscillatory features, called Franz–Keldysh oscillations (FKOs), which are a direct measure of the relevant electric field (built-in and/or applied).

Two of the most widely used forms of EM are photoreflectance (PR) and contactless electroreflectance (CER). In PR, modulation of the built-in electric field is caused by photo-excited electron–hole pairs created by a pump source of ~3–5 mW power of either (1) an internally modulated laser diode or (2) a mechanically chopped dc laser. The modulating frequency is typically ~100–200 Hz. CER utilizes a condenser-like system consisting of a front wire grid electrode with a second metal electrode separated from the first electrode by insulating spacers, which are ~0.1 mm larger than the sample thickness. The sample is placed between these two capacitor plates. Thus, there is nothing in direct contact with the front surface of the sample. The probe beam is incident through the front wire grid. EM is achieved by apply an ac voltage (~1 kV peak to peak at ~200 Hz) across the electrodes.

Instrumentation

Shown in Figure 1 is a schematic drawing of the experimental arrangement for a general 'external' modulated reflectance experiment (except for PR). Light from an appropriate light source (xenon arc or tungsten halogen lamp) passes through a monochromator. The exit intensity at wavelength λ, $I_o(\lambda)$, is focused onto the sample by means of a lens (or mirror). The modulation (electric field, stress, temperature) is applied to the sample at frequency Ω_m. The reflected light is collected by a second lens (mirror) and is focused onto an appropriate detector

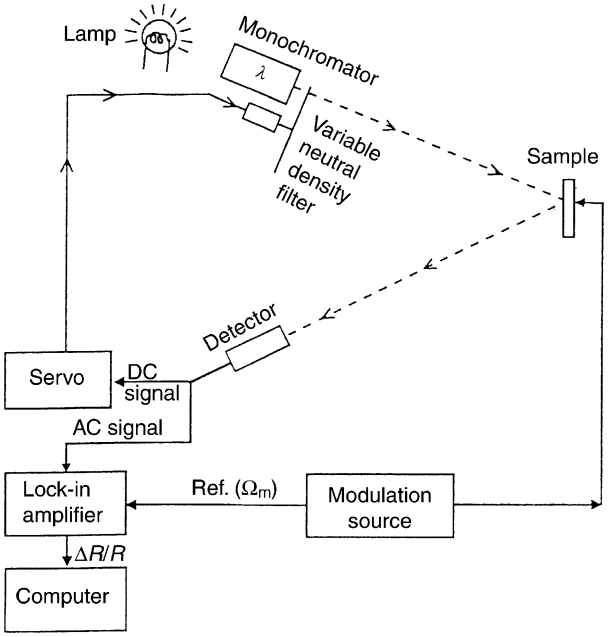


Figure 1 Schematic representation of external modulation experimental apparatus.

(photomultiplier or photodiode). For the sake of simplicity the two lenses (mirrors) are not shown.

The light striking the detector contains two signals: the dc is given by $I_o(\lambda)R(\lambda)$, where $R(\lambda)$ is the dc reflectance of the material while the modulated value (at frequency Ω_m) is $I_o(\lambda)\Delta R(\lambda)$, where $\Delta R(\lambda)$ is the change in reflectance produced by the modulation source. The ac signal from the detector, proportional to $I_o(\lambda)\Delta R(\lambda)$, is measured by a lock-in amplifier. Typically $I_o\Delta R$ is 10^{-4}–10^{-6} of I_oR. In order to evaluate the quantity of interest, i.e., the relative change in reflectance $\Delta R/R$, a normalization procedure must be used to eliminate the uninteresting common feature $I_o(\lambda)$. This can be done in several ways, one of which is shown in Figure 1, where the normalization is performed by a variable neutral density filter (VNDF) connected to a servo mechanism. The dc signal from the detector, which is proportional to $I_o(\lambda)R(\lambda)$, is fed into the servo which moves the VNDF in such a manner as to maintain a constant $I_o(\lambda)R(\lambda)$, i.e., $I_o(\lambda)R(\lambda) = C$. Under these conditions the ac signal $I_o(\lambda)\Delta R(\lambda) = C\Delta R(\lambda)/R(\lambda)$. Subsequently, the signal to the lock-in amplifier is proportional to the quantity of interest, i.e., $\Delta R(\lambda)/R(\lambda)$.

Bulk/Thin-Film Material

Modulation spectroscopy has been used to study the properties of bulk/epitaxial layered semiconductors such as ordering in $GaInP_2$, piezoelectric fields in strained GaN, the influence of isotopes, the properties of low-temperature grown GaAs, the space charge region, carrier concentration, the nature of the band bending (carrier type), impurity effects, deep levels, alloy composition, the influence of static external or internal perturbations such as temperature, external uniaxial stress, lattice-mismatch strain, hydrostatic pressure, the effects of crystal growth/processing/annealing procedures, and amorphous and microcrystalline materials.

The dashed lines in Figure 2 are the piezoreflectance (PZR) spectra in the region of the direct gap of ReS_2 in the temperature range $25\,K < T < 450\,K$. The data at low temperatures exhibit two well-resolved features labeled E_1^{ex} and E_2^{ex}. The solid lines are lineshape fits to the first derivative of a Lorentzian profile which yields the energies indicated by arrows and also the related broadening parameters. The solid/open squares and circles in Figure 3 are the temperature dependence of E_1^{ex}/E_2^{ex} of ReS_2 and $ReSe_2$, respectively. The solid, dot-dashed, and dashed lines in the figure are fits to the Varshni, O'Donnel–Chen, and Bose–Einstein expressions, respectively. The Varshni equation is:

$$E_i^{ex}(T) = \frac{E_i^{ex}(0) - \alpha_i T^2}{\beta_i + T} \qquad [1]$$

where $i = 1$ or 2, $E_i^{ex}(0)$ is the excitonic transition energy at 0 K and α, β are the Varshni coefficients. The O'Donnel–Chen expression is:

$$E_i^{ex}(T) = E_{i0} - S_i\langle\hbar\Omega_i\rangle\left(\frac{\coth\langle\hbar\Omega_i\rangle}{2kT} - 1\right) \qquad [2]$$

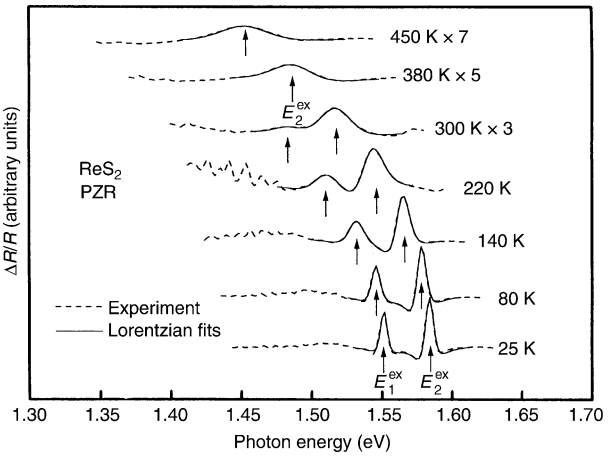


Figure 2 The PZR spectra (dashed lines) in the region of the direct gap of ReS2 in the temperature range $25\,K < T < 450\,K$. The solid lines are lineshape fits yielding the energies indicated by the arrows.

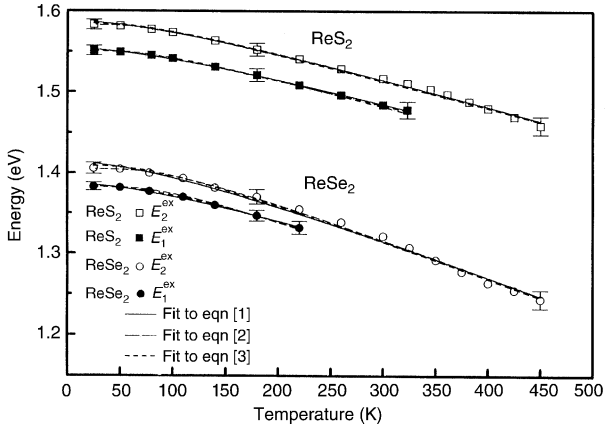


Figure 3 The temperature dependence of the excitonic transition energies of ReS_2 and $ReSe_2$. The solid, dot-dashed, and dashed lines in the figure are fits to the Varshni, O'Donnel–Chen, and Bose–Einstein expressions, respectively.

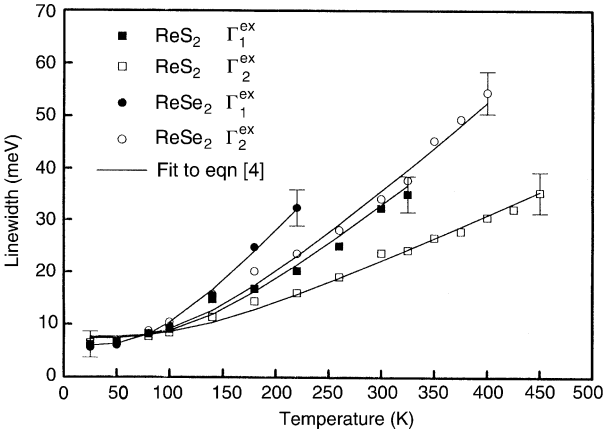


Figure 4 The solid/open squares and circles are the temperature dependence of the broadening parameters of E_1^{ex}/E_2^{ex} of ReS_2 and $ReSe_2$, respectively. The solid lines are fits to the Bose–Einstein-type expression.

where E_{i0} is the excitonic transition energy at 0 K, S_i is a dimensionless coupling constant related to the strength of the electron–phonon interaction, and $\langle \hbar\Omega_i \rangle$ is the average phonon energy. The Bose–Einstein-type expression can be written as

$$E_i^{ex}(T) = E_{iB} - a_{iB}\left[1 + \frac{2}{\exp(E_{ap}/kT) - 1}\right] \quad [3]$$

where a_{iB} represents the strength of the electron (exciton)–phonon interaction, and E_{ap} corresponds to the average phonon energy.

The temperature dependence of the broadening parameters, $\Gamma(T)$, is due only to the electron–phonon interaction and can also be described by a Bose–Einstein-type equation. For the lowest lying direct gap only the longitudinal optic (LO) phonon is involved. In Figure 4 the solid/open squares and circles are the temperature dependence of the broadening parameters of E_1^{ex}/E_2^{ex} of ReS_2 and $ReSe_2$, respectively. The solid lines are fits to the Bose–Einstein-type expression:

$$\Gamma_i(T) = \frac{\Gamma_i(0) + \Gamma_{i,ep}}{\exp(E_{i,LO}/kT) - 1} \quad [4]$$

where $i = 1$ or 2. The first term of eqn [4] corresponds to broadening mechanisms due to intrinsic lifetime, electron–electron interaction, impurity, dislocation, and alloy scattering effects. The parameter Γ_{ep} is an electron (exciton)–phonon coupling coefficient and E_{LO} is the zone-center LO phonon energy. The obtained values of Γ_{ep} for these materials have been compared with those for the direct gap in GaAs and ZnSe.

Surfaces/Interfaces

Almost since the inception of modulation spectroscopy it has been recognized that EM could be used as an effective probe of surfaces and interfaces. For EM this is true not only because of the sharp, derivative-like spectral features but also due to its sensitivity to electric fields, i.e., FKOs. In addition, PR and CER are of considerable interest and usefulness since they are contactless, employ low light levels, require no special mounting of the sample, and can be performed in any transparent ambient including ultrahigh vacuum (UHV).

Of particular usefulness in the study of Fermi level pinning at surfaces are UN(SIN)/UP(SIP) and δ-doping configurations. The UN(SIN)/UP(SIP) structures are made by fabricating an undoped layer of thickness L (~1000 Å) on a buried doped $n^+(p^+)$ buffer on a doped $n^+(p^+)$ substrate. In the $n^+(p^+)$ buffer/substrate the Fermi level occurs near the conduction (valence) band edge. At the surface the Fermi level is pinned at some value, V_F. Therefore, there exists in the undoped region, which has a small broadening parameter, a large, almost constant electric field, F. A similar constant field can be created by placing a δ-doping layer a distance L from the surface.

The relation between F, as observed from the FKOs, and V_F is given by:

$$FL = V_F - (kT/q) - SCC - V_p(T) \quad [5]$$

where the second, third, and fourth terms on the right-hand side are the Debye length, space charge layer, and temperature dependent photovoltaic effect corrections, respectively.

The properties of the metal–semiconductor interface have been reported by a number of investigators using PR and/or ER. Jonker *et al*. have demonstrated that an epitaxial Fe film on GaAs (UN structure) provides a ferromagnetic contact, suppresses midgap state formation, and does not pin the Fermi level. Four samples were investigated: a native oxide reference sample, the same sample after sulfur passivation, Al/GaAs-(2 × 4), and Fe/GaAs-(2 × 4). The width of the undoped GaAs layer was 1070 Å for the first three samples and 1500 Å for the last sample. These results are shown in Figure 5a–d, respectively. The fields obtained from the FKOs are 58 kV/cm, 61 kV/cm, 55 kV/cm and 60 kV/cm. For the first three samples the measured barrier height is about 0.64 eV (not corrected for the photovoltaic effect) while for the Fe/GaAs material it is about 40% higher (~0.9 eV). Another significant feature of the Fe/GaAs spectrum is the large number of FKOs (in relation to the other samples) extending to about 2 eV. The significant reduction in damping indicates a substantial increase in carrier lifetimes.

Micro- and Nanostructures

Modulation spectroscopy is a very powerful tool for investigating many of the fundamental aspects of compositional single quantum wells (SQWs)/multiple quantum wells (MQWs), superlattices (SLs), quantum wires, and quantum dots (QDs). Various phenomena can be studied, including band offset, well and barrier widths, excitons, strain, coupling and decoupling between wells, miniband formation, two-dimensional electron gas (2DEG) effects, zone-folding in short-period SLs, built-in electric fields, etc. Work on a large variety of GeSi, III–V and II–VI systems has been reported, including both lattice-matched and strain-layer configurations. In addition the effects of various external perturbations can be evaluated.

The properties of a vertically coupled InAs/GaAs QD-based laser structure fabricated by a self-assembled technique have been investigated by CER (300 K and 20 K). Signals have been observed from all the relevant regions of the sample. The dotted lines are the CER spectra at 300 K and 20 K in Figure 6a,b, respectively. The solid lines are lineshape fits which yield the energies indicated by arrows. The low-energy features, labeled QD_0 (lateral coupling), QD_1 (vertical coupling) and QD_2 (vertical coupling), originate from the QDs while the resonances designated W_1 and W_2, correspond to the first conduction to first heavy- (1C-1H) and light-hole (1C-1L) transitions, respectively, in the QW (approximately one monolayer) formed by the

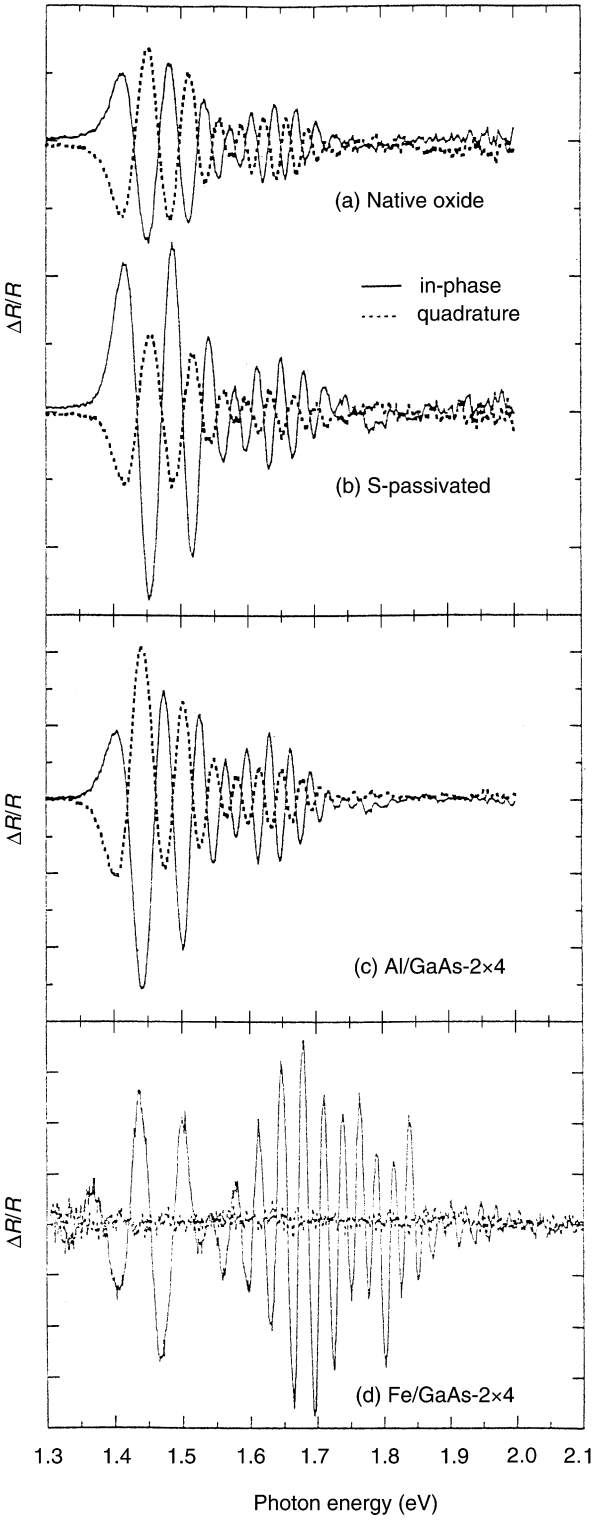


Figure 5 The PR spectra at 300 K for (a) a native oxide reference sample, $F = 58$ kV/cm; (b) the same sample following sulfur passivation, $F = 61$ kV/cm; (c) Al/GaAs-(4 × 2) sample, $F = 55$ kV/cm, and (d) Fe/GaAs-(2 × 4) sample, $F = 60$ kV/cm. The thickness of the undoped spacer region is 1070 Å for (a)–(c) and 1500 Å for (d).

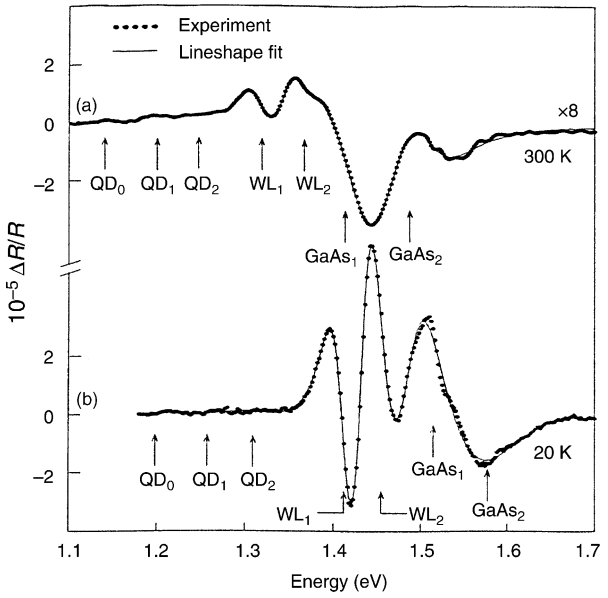


Figure 6 CER spectra (dotted lines) of an InAs/GaAs quantum dot-based laser structure at (a) 300 K and (b) 20 K. The solid lines are a lineshape fit to a first derivative Gaussian.

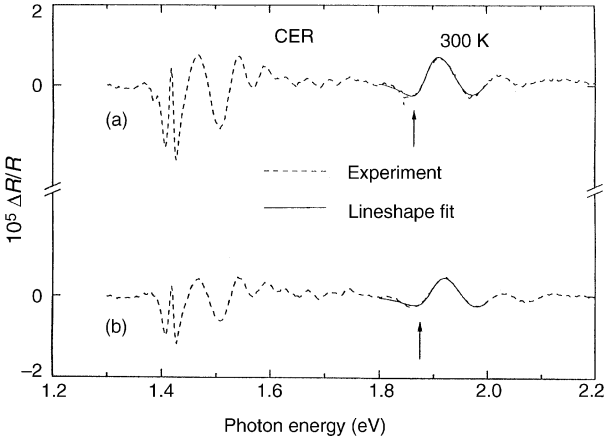


Figure 7 Room-temperature CER spectra (dashed lines) for a GaInP/GaAs HBT sample fabricated by metalorganic chemical vapor deposition for incident light polarized (a) $\boldsymbol{E}\|[110]$ and (b) $\boldsymbol{E}\|[1\bar{1}0]$. The solid lines are a lineshape fit which yields the bandgap energies denoted by arrows.

wetting layer (WL). Although the oscillator strength of the QD features is small compared to the WL resonances, they are clearly visible even at 300 K. The $GaAs_1$ and $GaAs_2$ features are attributed to GaAs-related transitions. The former at both 300 K and 20 K corresponds to the bandgap energy of GaAs and probably comes from the undoped 400 Å-thick GaAs layers which envelop the QD region. The $GaAs_2$ feature is most likely related to the p-doped 6000 Å-thick GaAs cap layer. The heavy doping of this layer induces a Burstein–Moss blue shift of the bandgap.

Device Structures

The methods of PR and CER are valuable tools in the evaluation of important device parameters for structures such as heterojunction bipolar transistors (HBTs), pseudomorphic high electron mobility transistors (PHEMTs), edge-emitting and vertical cavity surface-emitting (VCSEL) quantum well lasers, multiple quantum well infrared detectors, solar cells, superlattice optical mirrors, resonant tunneling structures, metal-oxide-semiconductor (MOS) configurations, among others.

The dashed lines in Figure 7a,b are the 300 K CER spectra of a $GaInP_2$(emitter)/GaAs(collector) HBT (fabricated by metalorganic chemical vapor deposition) for incident light with electric field vector $\boldsymbol{E}\|[110]$ (polarization A) and $\boldsymbol{E}\|[1\bar{1}0]$ (polarization B), respectively. Both collector and emitter signals exhibit well-pronounced FKOs. The solid lines are least-squares fits to a function which contains Lorentzian broadened electro-optic functions, yielding 1.873 ± 0.001 eV and 1.885 ± 0.001 eV for the $GaInP_2$ bandgaps for polarizations A and B, respectively, as indicated by the arrows. The ordering parameter of 0.3 deduced from the differences in the measured bandgaps for the two polarizations was consistent with transmission electron microscope measurements. The electric fields evaluated from the emitter and collector region FKOs were compared with a computer simulation of the field profiles using a comprehensive, self-consistent model, including the photovoltaic effect. The deduced emitter and collector doping densities were in good agreement with the intended growth conditions.

Summary

Modulation spectroscopy is a very powerful tool for investigating many of the fundamental and applied aspects of a wide variety of semiconductors and semiconductor micro- and nanostructures, including bulk/thin films, surfaces/interfaces, single quantum wells, multiple quantum wells, superlattices, quantum wires, quantum dots, etc. In addition the effects of various perturbations such as alloy composition, temperature, electric and magnetic fields, hydrostatic pressure, internal and/or external stain, processing, etc., can be evaluated. Recently these methods have also been applied for the characterization of actual device structures such as HBTs, PHEMTs, quantum well lasers, etc.

See also

Semiconductor Materials: Quantum Dots. **Semiconductor Physics:** Band Structure and Optical Properties; Impurities and Defects; Outline of Basic Electronic Properties; Quantum Wells and GaAs-based Structures; Surface Photovoltage Spectroscopy of Semiconductors.

Further Reading

Aspnes DE (1980) Modulation spectroscopy/electric field effects on the dielectric properties of semiconductors. In: Balkanski M (ed.) *Handbook on Semiconductors*, vol. 2, pp. 109–154. Amsterdam, The Netherlands: North-Holland.

Glembocki OJ and Shanabrook BV (1992) Photoreflectance spectroscopy of microstructures. In: Seiler DG and Littler CL (eds) *Semiconductors and Semimetals*, vol. 36, pp. 222–289. New York: Academic Press.

Pollak FH (1994) Modulation spectroscopy of semiconductors and semiconductor microstructures. In: Balkanski M (ed.) *Handbook on Semiconductors*, vol. 2, pp. 527–635. Amsterdam, The Netherlands: North Holland.

Pollak FH (1995) Electromodulation spectroscopy of semiconductor surfaces and interfaces. In: Halevi P (ed.) *Photonic Probes of Surfaces*, pp. 175–209. Amsterdam, The Netherlands: North-Holland.

Pollak FH (1996) Modulation spectroscopy of quantum wells and superlattices. In: Bhattacharya P (ed.) *Properties of III–V Quantum Wells and Superlattices*: *EMIS Data Reviews Series #15*, pp. 232–240. London: INSPEC.

Pollak FH (1998) Modulation spectroscopy of the group III-nitrides. In: Gil B (ed.) *Group III Nitride Semiconductor Compounds*, pp. 158–181. Oxford, UK: Clarendon.

Pollak FH (2001) Modulation spectroscopy characterization of semiconductors and semiconductor microstructures. In: Buschow KH, Cahn W, Flemmings MC, *et al.* (eds) *Encyclopedia of Materials: Science and Technology*, vol. 6, pp. 5743–5748. Amsterdam, The Netherlands: Elsevier.

Ramdas AK and Rodriguez S (1992) Piezospectroscopy of semiconductors. In: Seiler DG and Littler CL (eds) *Semiconductors and Semimetals*, vol. 36, pp. 137–220. New York: Academic Press.

Quantum Dots

R J Warburton, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Quantum dots are nanometer-sized clusters of semiconductor material. An island of low-bandgap material is embedded in a sea of high-bandgap material so that both electrons and holes are spatially confined in all three directions. In an unstructured semiconductor, the electron energy levels of the individual atoms hybridize to form energy bands. In a quantum dot, however, there are discrete energy levels. This can be understood from the band structure of the unstructured material where there is a so-called dispersion relation connecting the energy, E, and the momentum, k. The confinement potential of a quantum dot quantizes the k-values, giving rise to discrete energy levels. If the energy levels of a quantum dot are separated by more than the typical energy level broadenings and by more than the thermal energy then the quantum dot will behave quite differently to the bulk material. The quantum dot will exhibit atom-like properties – discrete energy levels with small degeneracies – despite the fact that it is embedded in a semiconductor matrix. It is this ability to prepare a variety of atom-like systems in a semiconductor environment which has driven the research into semiconductor quantum dots.

In some sense, nanosized quantum dots have been the subject of research for quite some time. Color filters, for instance, consist of nanosized particles dispersed in a glass matrix, and this is a very mature technology. Also, lithography of layered semiconductor systems already possessing confinement in one dimension has been pursued for at least 15 years. However, developments in the past few years have been rapid, all based on the discovery of self-assembly of quantum dots directly in the growth of semiconductor heterostructures. These quantum dots can be produced with standard materials in standard growth systems, and this allows an easy integration of quantum dots into devices such as diode lasers. Furthermore, the self-assembled quantum dots are of extremely high quality and this is of great value not just for devices but also for fundamental physics experiments, aimed at studying or exploiting the coherence of excitons confined in the quantum dots. For these reasons, this short review concentrates on the properties of self-assembled quantum dots, giving just brief descriptions of other systems. It should be noted that the bibliography is extremely limited and

in no way does justice to the large number of research groups which have made valuable contributions to this field. Instead, the bibliography is simply a list of suggestions for finding out more about this fascinating research area.

Quantization in Quantum Dots

Quantum dots are on the one hand large enough that the bulk band structure of the host material can be used to describe the material's electronic properties. In the effective mass approximation, this amounts to giving both electrons and holes a quadratic dispersion on momentum, just as for a free electron, but with an effective mass, typically about 0.1 for the electrons, and 0.3 for the holes. On the other hand, quantum dots are small enough that confinement effects are important. For instance, the energy of an exciton in a nanometer-sized quantum dot is a strong function of the quantum dot size. As the dots become smaller, confinement effects become larger, and the energy separating electron and hole levels increases. This does not increase indefinitely, however. There is a finite-size energy barrier to the matrix surrounding the dot such that in the limit of small dot sizes, the wavefunction spreads out into the barrier region and confinement effects become very weak. Ideally, an electron in a quantum dot has a confinement energy of at least several meV, and an energy barrier to the surrounding matrix of several 100 meV so that confinement effects are visible even at room temperature.

The motivation for the study of quantum dots is to combine the atom-like properties with the flexibility of a solid-state environment. This is perfectly illustrated by considering a quantum dot as the gain medium for a laser. In a conventional semiconductor diode, a band of levels has to be inverted in order to achieve population inversion. This is costly in terms of the threshold current and the temperature stability. In a quantum dot, inversion can be achieved simply by inverting the population of the two-fold degenerate ground state, which is clearly trivial. Furthermore, a strong confinement energy and a large barrier to the surrounding matrix imply that this population inversion is stable against rises in temperature, implying better room-temperature performance. This example also exemplifies the motivation for investigating quantum dots using well-developed inorganic semiconductor technology as in this case the dots can be integrated into established device geometries. Generally speaking, the important concept in quantum dot physics is spatial confinement of electrons in all three directions, giving rise to discrete energy levels.

Lithography for Quantum Dots

Developments in semiconductor growth technology, largely during the 1970s and 1980s, enabled thin layers of different semiconductors to be grown on top of each other. There are two principal growth techniques, molecular beam epitaxy (MBE), in which a heated substrate is bombarded with pure species in a high vacuum, and metal organic chemical vapor deposition (MOCVD), in which chemical precursors flow over a heated substrate and react. In the simplest structure, termed a quantum well, a thin layer of low-bandgap material is sandwiched between high-bandgap material. The discontinuities in the potentials at the interface give rise to confined electronic states for both carrier types, electrons and holes. The quantum well is an example of a two-dimensional system: there is confinement in one direction, the growth direction, but free dispersion in the other two directions. To make a quantum dot, it is necessary to confine the electron and hole in all three directions. An obvious approach is to take a quantum well, with perfect confinement in one direction, and process it laterally with lithographic techniques to complete the quantization. Note that an analogy can be made to the physics of the quantum Hall effect, where a two-dimensional electron gas is subjected to a very high magnetic field. The magnetic field quantizes the lateral motion in a similar way to the electrostatic potential in a quantum dot.

This approach to the formation of optically active quantum dots obviously has some advantages, notably that the shape and position of the quantum dots can be chosen in the lithography. However, there are some significant problems associated with these techniques. As most of these problems do not plague the properties of self-assembled quantum dots, self-assembled quantum dots are more suitable for advanced photonic devices.

A number of different lithographic techniques has been used to process quantum-well heterostructures into quantum dots. For research applications, electron-beam lithography is the method of choice. The sample is coated with a resist, for instance PMMA, and exposed with electrons in an area with nanometer size. A protective layer, either a metal or insulator, is transferred to the exposed area by developing the resist, by depositing the protective layer material, and by removing the unwanted material with a lift-off step. The wafer is then etched, typically in a dry environment, to produce a free-standing column.

The photoluminescence of these quantum dots has been measured by various research groups, and although the results vary quite widely there are some systematic trends. The photoluminescence

efficiency decreases rapidly for pillars smaller than a few microns. This is somewhat improved at low temperature, where reasonably strong emission can be observed down to 100-nm diameter pillars. For pillars with micron diameters, the lateral confinement effects are weak, and it is debatable if true quantum dots have been achieved. There are a number of effects which degrade the emission. First, pinning of the Fermi energy by surface states generates large electric fields near the surface, which can ionize electron–hole pairs. This is well known for quantum wells: if a GaAs quantum well for example is placed less than about 25 nm from the surface, the photoluminescence is very weak because of the surface electric field. In an etched pillar, these effects are clearly magnified by the new surface area generated. Secondly, the etching process damages the material to a greater or lesser degree, and the damaged region contains centers for nonradiative recombination.

Other lithography techniques have also been used to generate semiconductor quantum dots for optical experiments. An elegant idea is to deposit small regions of a highly strained film onto the sample surface. In this way, a laterally varying strain film in an underlying quantum well can be set up, generating quantum dots through the effect of strain on the band gap. (This idea has also been married with self-assembly, by using Stranski–Krastanow-grown InP stressors on a GaAs surface, generating lateral confinement in an underlying InGaAs quantum well.) A further idea is to generate a quantum dot by deliberately growing a quantum well with rough interfaces. In such a system, there will inevitably be regions which confine an electron–hole pair through localization. Both of these approaches clearly circumvent the problems generated by the exposed surface in an etched quantum dot. However, they generate small lateral confinement potentials, typically a few meV. This is of value in a research environment where fundamental experiments can be carried out at low temperature, but in a possible device at room temperature, where the thermal energy is 25 meV, quantization effects will not be discernible.

Self-Assembled Quantum Dots

A solution to the difficulties inherent in quantum dots made by lithographic techniques began to emerge in the early 1990s. It was discovered that quantum dots can be produced directly in the semiconductor growth by depositing a material onto a substrate with a substantially smaller lattice constant. The most well-known example is the growth of InAs (lattice constant 6.05 Å) on GaAs (lattice constant 5.65 Å) where In-rich quantum dots form. Because billions of dots form spontaneously in this process, the dots are referred to as self-assembled quantum dots. Since the early experiments, considerable progress has been made, such that the state-of-the-art quantum dots possess many of the hoped-for properties: they are coherent (meaning that they contain no dislocations), highly homogeneous (meaning that fluctuations in shape and composition from dot to dot are small), and have large quantum efficiencies for the emission of light.

Growth of Self-Assembled Quantum Dots

When growing one material on top of another, three principal growth mechanisms have been identified: Frank–van der Merwe, Volmer–Weber, and Stranski–Krastanov. In the Frank–van der Merwe mode, material is deposited layer by layer, as in the growth of the lattice-matched pair GaAs and AlAs. In the Volmer–Weber mode, island formation occurs: the material does not wet the surface because it is energetically unfavorable. Another possibility arises if there is a lattice mismatch. In this case, the epilayer is strained, and the strain energy obviously increases with increasing epilayer thickness. In the Stranski–Krastanow mode, the initial growth is layer by layer, as in Frank–van der Merwe, but beyond a certain thickness, islands form, as shown schematically in Figure 1. On the one hand, the islands limit the strain energy, because some strain relaxation is allowed, which is not possible in a thin film. On the other hand, island formation results in an increase of the

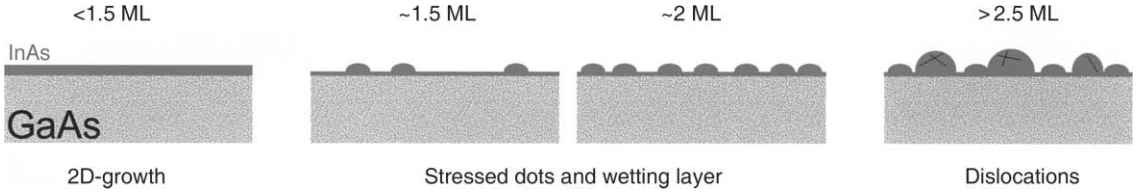


Figure 1 A schematic of the Stranski–Krastanow mode. The material deposited has a substantially larger lattice constant than the substrate, for example InAs on GaAs. Initially, a thin and uniform layer is deposited. At the critical thickness (1.5 monolayers for InAs on GaAs), islands spontaneously form. Further growth results in an increase in the number of islands. At large coverages, large and dislocated islands appear.

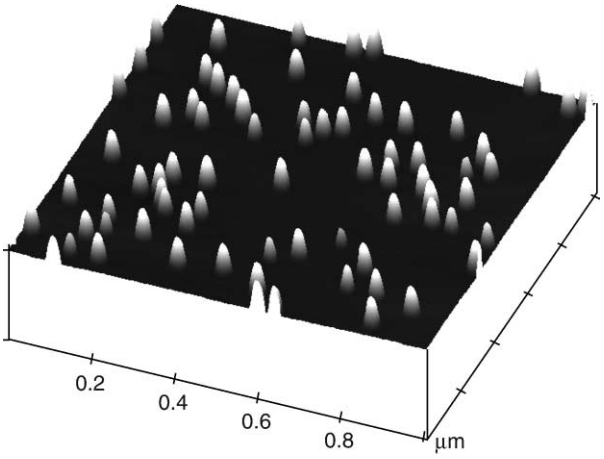


Figure 2 An atomic force micrograph of InAs quantum dots on a GaAs substrate. The InAs dots were grown by MBE at a growth temperature of 530 °C. The image was taken by Axel Lorke.

surface energy. This means that there is a size at which the total energy is minimized.

It is now well established that many combinations of semiconductors grow in the Stranski–Krastanow mode, and this can be utilized for self-assembly of quantum dots. The most studied example is InAs on GaAs where high-quality quantum dots can be grown with both MBE and MOCVD. Dots form at the so-called critical thickness of 1.5 monolayers, corresponding to just 4 Å of material. For thicknesses less than the critical thickness, the covering is not completely uniform; there are island-like structures, elongated along the $[0\bar{1}1]$ direction. At the critical thickness dots form, leaving a thin InAs layer (the wetting layer) between the dots. As more InAs is added, the density of dots increases with only small changes in the dot size. As an example, Figure 2 shows InAs dots on a GaAs substrate.

The dots grown in this Stranski–Krastanow mode are free of dislocations, which is an essential prerequisite in III–V compounds for a high quantum efficiency for optical emission. Furthermore, the dots are remarkably homogeneous. Statistical fluctuations during growth will always give rise to a distribution in dot size and height. Nevertheless, the fluctuations in dot size can be as small as ~10%. The dots are randomly arranged in the lateral plane if the growth proceeds on a flat substrate, and densities are between 10^9 and 10^{11} cm^{-2}. The shape and composition of the dots depend on the growth parameters. For instance, InAs/GaAs dots can have facets along particular crystal directions, or they can be lens-shaped without facets, depending on the growth technique, and in particular the growth temperature. In fact, measuring the shape and composition is a very challenging task, particularly for buried dots.

Stranski–Krastanow growth has turned out to be a robust phenomenon for III–V materials; quantum dots have been produced in this way for a number of material combinations. The most important condition is that the deposited material has a substantially larger lattice constant than the substrate. InAs forms dots not just on GaAs but also on InP. The InAs/InP dots are particularly interesting because they emit at the technologically important 1.5 μm wavelength. InP can also form the dot material by using GaInP substrates where the bandgap is pushed up into the red region of the spectrum. Dots can also be grown with II–VI materials; the growth of CdSe on ZnSe is analogous to the growth of InAs on GaAs. These II–VI dots emit in the green. Finally, dots can also be formed in the nitrides, for instance by depositing InGaN on an AlGaN surface. In fact, the high emission efficiency of GaN, despite the large defect density, has been attributed to the formation of quantum dot-like structures in the active layers.

It is a challenge to arrange the dots in the plane of the substrate. An easy route with self-assembly has not been found to accomplish this. Nevertheless, dots can be encouraged to grow at particular positions by pre-patterning the substrate. This obviously limits the separation of the dots in any such array to the limits set by lithography. However, self-assembly does provide a means to order the dots in the growth direction: if the separation between successive layers is small enough (less than 15 nm for InAs/GaAs) then the dots grow directly above one another, as shown in Figure 3. In a new layer, the residual strain field from the dots in the layer below is sufficient to seed the dots. In this way, vertical stacks of dots can be built up. In fact, it has been argued that the dots tend to become more laterally ordered with each layer. The point is that if two dots are close together in one layer, their strain fields overlap, and provide only one seeding center in the subsequent layer.

Optical Properties

The main interest in quantum dot physics is the quantization in all three spatial directions, giving rise to atomic-like energy levels. For a system of self-assembled quantum dots, it is important to measure the fundamental bandgap, the energetic separation between the confined states both for electrons and for holes, the barriers to the electron continuum and hole continuum, and the oscillator strengths for the various interband transitions. All this information can be gleaned with optical spectroscopy.

Perhaps the simplest experiment to perform, if not to understand, is photoluminescence in which electron–hole pairs are excited with a laser beam.

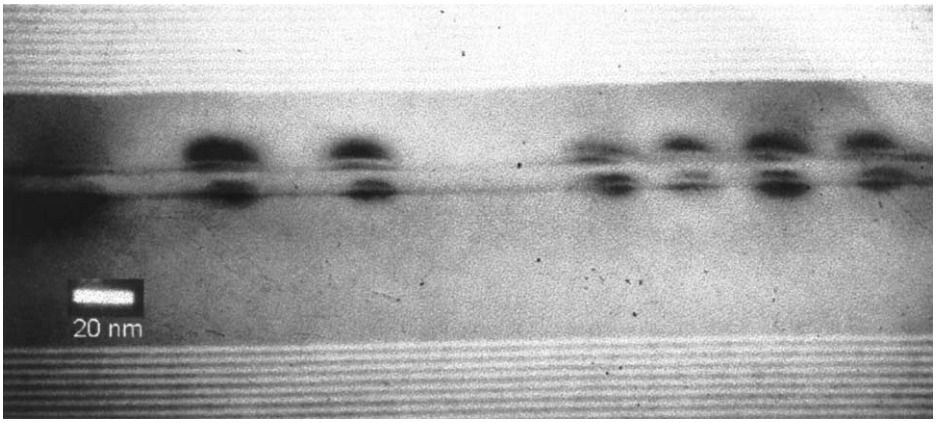


Figure 3 A transmission electron micrograph of two layers of InAs quantum dots in GaAs showing how the dots in the upper layer lie directly above the dots in the first. The image was taken by Gio Medeiros-Ribeiro.

The laser energy can be tuned to lie just above the bandgap of the wetting layer in which case the electron–hole pairs are excited in the wetting layer itself. Alternatively, if the laser energy is larger than the bandgap of the barrier material, electron–hole pairs are excited in the barrier. In both cases, the carriers relax into the dots, and then down the ladder of states in the dot, so that the dot electron and hole ground states are occupied. Emission is detected when an electron and a hole recombine. As such, the experiment measures only the energy of the ground-state exciton. However, if the pump intensity is increased such that electron–hole pairs are generated faster than they can recombine, the excited states of the dots are also occupied and recombination can also occur from excited states. An example of such an experiment is shown in Figure 4. At low pump powers, only the ground state emission is detected; at higher pump powers, the first excited state emerges, and at higher powers still, the second excited state also emerges, and so on. The spectra are broadened by the inhomogeneous broadening, i.e., from fluctuations from dot to dot, but the broadening is small enough that the excited state emission can be easily distinguished from the ground state emission. For these particular quantum dots, the energetic separation between the ground and excited state emission is 15 meV, and this corresponds to the sum of the electron and hole confinement energies. The separation between the other emissions is also about 15 meV, implying that the states are approximately equally spaced, implying in turn that the confining potential is approximately parabolic. Furthermore, in this experiment, emission from the wetting layer can also be made out, and this enables the electron and hole confining potentials to be estimated.

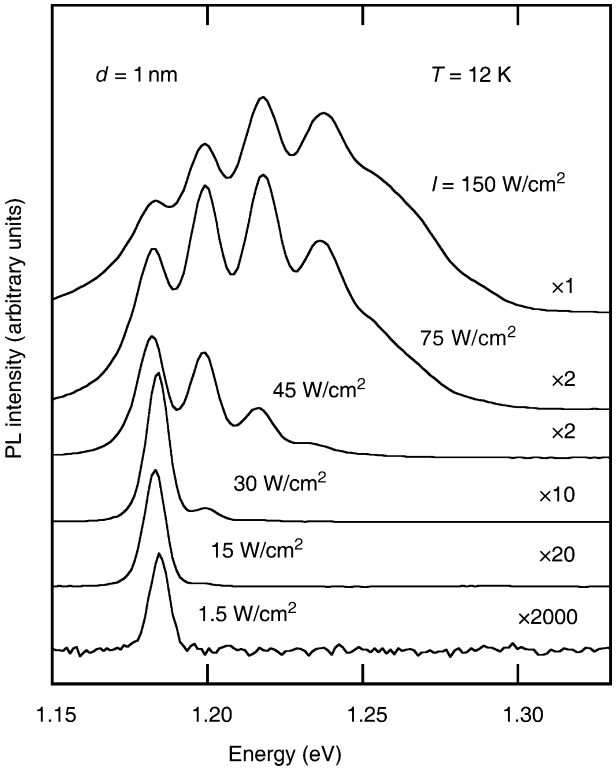


Figure 4 Photoluminescence from an ensemble of quantum dots as a function of laser excitation power. The measurement temperature was 12 K. The quantum dots are induced in an InGaAs/GaAs quantum well by a strain field generated by InP stressors on the sample surface. The stressors are grown in the Stranski–Krastanow mode, and are 1 nm away from the quantum well. At low excitation power, only the ground state emission is observed. As the power increases, emission from excited states is also observed through state filling. Reprinted with permission from Lipsanen H, Sopanen M and Ahopelto J (1995) Luminescence from excited states in strain-induced $In_xGa_{1-x}As$ quantum dots. *Physical Review* 51: 13868. Copyright 1995 American Physical Society.

Photoluminescence invariably yields the sum of electron and hole energies. In order to measure just the electronic properties, it is possible to occupy the dots with, say, two electrons (so that in each dot the two-fold degenerate ground state is fully occupied) and to excite electric dipole transitions between the electron levels. This has been accomplished with so-called charge-tunable heterostructures, with which electrons can be controllably loaded into the dots from an electron reservoir. The interlevel absorption lies in the far-infrared region of the spectrum; for InAs/GaAs quantum dots for instance, the transition is at a wavelength of 25 μm (an energy of 50 meV), as shown in Figure 5. The results of this experiment also point to an approximately parabolic confining potential from the fact that as electrons are added to the dots, the energy of the far-infrared absorption does not change much. (In the limit of a perfect parabolic potential, a famous result known as Kohn's theorem states that the long-wavelength absorption should show no energy dependence on the electron occupation.)

In self-assembled quantum dots, the confining potential can be thought of as very steep in the growth direction, and relatively shallow in the lateral plane. In other words, the dots are essentially two-dimensional disks. The electron ground state is, in analogy with atomic physics, s-like, and the excited state p-like. However, because of the disk-like nature of the dots, the p-state is four-fold degenerate, and not six-fold degenerate as in a conventional atom. One way to demonstrate this is to perform interband absorption experiments on charge-tunable structures, as shown in Figure 6. The first transition corresponds to the transition from the hole s-state to the electron s-state, and disappears when the dots are occupied with two electrons. This is simply because the Pauli principle forbids the transition once the final level is fully occupied. The second transition corresponds to the transition from the hole p-state to the electron p-state, and disappears when the dots are occupied with four electrons, confirming the degeneracies of

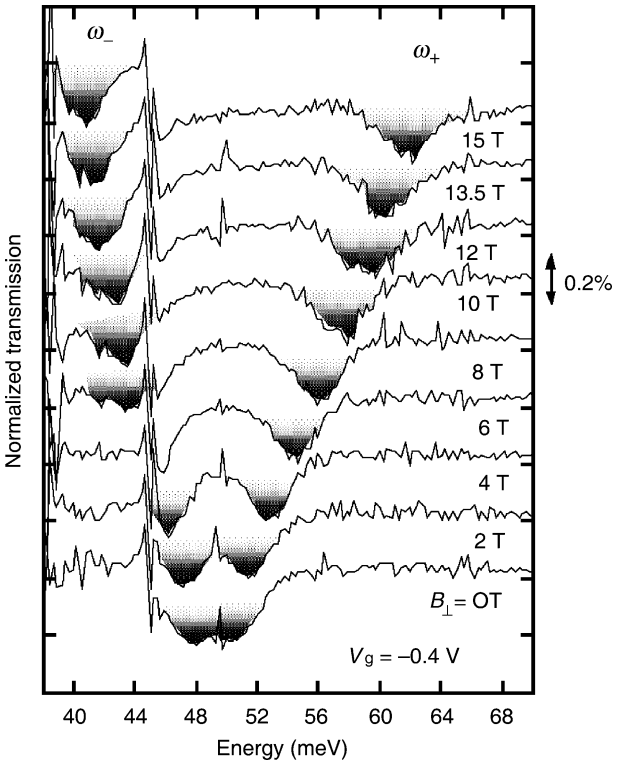


Figure 5 Transmission in the far infrared of an ensemble of InAs/GaAs quantum dots where each quantum dot is occupied with two electrons. The data were taken at a temperature of 4.2 K. The shaded transmission minima correspond to absorption from the quantum dots. The transitions are between the electron ground state and the first excited state. Two peaks are observed even at zero magnetic field, suggesting that the dots are slightly oval in shape. In a magnetic field, the two peaks move apart, which is a consequence of the Zeeman effect. The strong feature at 45 meV arises from an electronic interaction with an interface phonon. Reprinted with permission from Fricke M, Lorke A, Kotthaus JP, Medeiros-Ribeiro G and Petroff PM (1996) Shell structure and electron-electron interaction in self-assembled InAs quantum dots. *Europhysics Letters* 36: 197. Copyright 1996 European Physical Society.

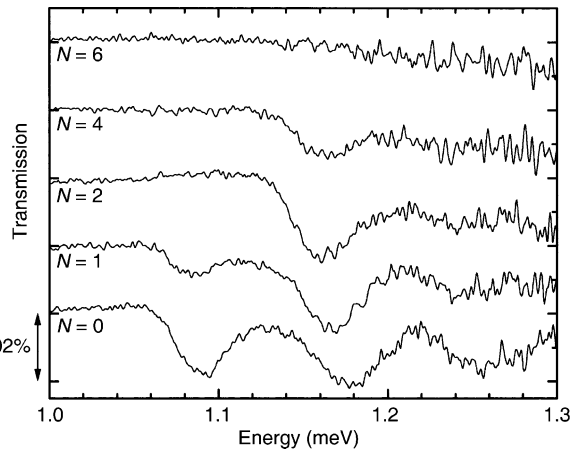


Figure 6 Transmission in the near infrared of an ensemble of InAs/GaAs quantum dots at 4.2 K. Spectra are shown for different values of N, the electron occupation. At $N = 0$, there are three transitions corresponding to transitions from a valence state to an electron state. These are the s–s, p–p and d–d transitions, where the states are labeled in analogy to atomic physics. At $N = 2$, the electron ground state is fully occupied, and the s–s transition disappears because it is blocked, a consequence of the Pauli principle. Similarly, the p–p transition is blocked at $N = 6$ when both the electronic s and p states are fully occupied. Reprinted with permission from Warburton RJ, Dürr CS, Karrai K *et al.* (1997) Charged excitons in self-assembled semiconductor quantum dots. *Physical Review Letters* 79: 5282. Copyright 1997 American Physical Society.

the confined states. This experiment, unlike a photoluminescence experiment, also yields a value for the oscillator strength. The result can be understood simply in terms of the overlap integral between the hole and electron states. In other words, these dots are in the strong confinement regime where the confinement energies are larger than the exciton binding energy. In the other limit, the weak confinement regime, the exciton binding energy dominates over the confinement energies, and it makes sense only in this limit to think of an exciton moving as a coherent entity throughout the dot. In the weak confinement regime, the oscillator strength is actually larger than in the strong confinement regime.

An important issue in interpreting photoluminescence experiments is the rate at which a highly excited electron–hole pair can relax to the ground state of a quantum dot. Relaxation in the continuum is known to occur on a picosecond time-scale: electrons and holes relax by emitting LO-phonons if possible; and if not, acoustic phonons. Capture by a dot also occurs quickly (apart from in low-density samples where capture may take a few hundred picoseconds). It was initially thought that relaxation of carriers once captured by a quantum dot would be hindered by the discrete nature of the density of states. The argument is that the energy separation between the levels is too large for acoustic phonon emission; and there is no reason why the separation between the levels should match the LO-phonon energy (which has a very weak dispersion). Nonetheless, experiments have shown that relaxation from excited dot states to the ground state is rapid, occurring on a picosecond time-scale, pretty much as for quantum wells. Auger processes can be important, whereby an electron in a dot loses energy by promoting an electron in the continuum to a higher energy. But relaxation can also be fast without a large population of mobile carriers in the wetting layer, and this is thought to arise through multiphonon processes. The radiative lifetime is typically 1 ns for self-assembled quantum dots, depending to some extent on the size and material system. As for quantum wells, relaxation occurs much faster than recombination.

Lasers and Self-Assembled Quantum Dots

The motivation for using quantum dots as the gain medium for semiconductor lasers is that they should have advantageous properties over quantum wells. These include a lower threshold current, better temperature stability, and a higher modulation frequency. All of these possible advantages stem from the quantization. In a quantum dot, there are discrete levels each with low degeneracy so that population inversion is easy to achieve. Also, there are large potential barriers to the continuum of states in the surrounding matrix, and an energy separation between levels in the dot at least comparable to the thermal energy at room temperature. This should make a quantum dot device relatively insensitive to temperature. The threshold current of a semiconductor laser usually behaves as $I = I_0 e^{T/T_0}$ where T is the temperature, T_0 is the characteristic temperature, and I_0 the low-temperature threshold current. Generally speaking, lasers are desirable with a small I_0 and a large T_0.

The main obstacle to realizing the benefits of full spatial confinement in self-assembled quantum dots is the inhomogeneity. For a particular lasing wavelength only a fraction of the dots can contribute to the lasing signal. Additionally, the dot densities are typically about 10^{11} cm^{-2} for these applications, implying that in the layer plane there is more unoccupied space than dot material. All in all, the gain at a particular wavelength is limited. Nevertheless, it would appear that even with inhomogeneous dot ensembles the advantages inherent in zero-dimensional states can be exploited.

In addition to the density of states, there is an important difference in the nature of the gain spectrum between quantum well and quantum dot lasers. In a quantum well, gain exists over a range of energies through band filling. If an electron–hole pair is stimulated to emit by the lasing field, then the vacancies will be quickly filled by carrier–carrier scattering in the bands, maintaining a quasi-thermal equilibrium in the conduction and valence bands. In quantum dots at low temperature, capture into a particular dot is a random process, and given the rapid relaxation and the high barriers to the continuum, an electron–hole pair is very likely to emit from the dot where it was created. In other words, the system is not in thermal equilibrium. At higher temperatures, thermal escape out of the dots becomes likely, enabling the system to come into thermal equilibrium. The transition into thermal equilibrium will depend critically on the barrier height, and with appropriate design may take place close to, or even above, room temperature.

Quantum dot lasers have now been realized, and emit continuous wave at room temperature. Progress has been rapid in the past few years. In fact the best quantum dot lasers now have the lowest I_0 and the highest T_0 of any semiconductor laser diode.

Even if the intrinsic advantages of quantum dots are largely offset by the inhomogeneous broadening for laser applications, there are some other advantages. The InAs/GaAs quantum dot emission wavelength is

significantly higher than that of InGaAs/GaAs quantum wells, and with subtle changes in the growth can be increased to 1.3 μm, corresponding to one of the low-loss windows in optical fiber. Also, much higher powers have been realized with quantum dot lasers than with quantum well lasers. In quantum well lasers, device failure occurs at high current, often due to degradation of the end mirrors. It would appear that the carrier localization in quantum dots is sufficiently strong to prevent diffusion to the end facets, allowing the devices to withstand higher currents.

Colloidal Quantum Dots

Progress has been made recently in making inorganic quantum dots with techniques from organic chemistry. Organometallic precursors are employed. The growth temperature is typically below 200 °C which has been found to give material with quite high crystalline quality. A crucial development has been the use of various organic groups to passivate the surface so that the quantum dots have bright photoluminescence. The II–VIs, being more polar than the III–Vs, are easier to produce with these techniques, and most work has concentrated on CdS and CdSe quantum dots which emit in the visible. Nevertheless, it is also possible to produce III–V quantum dots with this technology, and although the bulk GaAs bandgap lies in the near-infrared, the quantum confinement effects can be so high as to push the photoluminescence into the visible. Furthermore, HgCdTe has also been developed as a material system for quantum dots, and the emission can be tuned to the all-important 1.5 μm.

The spectral width of emission from these colloidal quantum dots tends to be very large because of a large spread in particle size. However, techniques are being developed to isolate a specific size. With a selective precipitation technique for instance, the spectral width can be reduced considerably to the extent that a fine structure can be observed in the optical response.

The advantage of these quantum dots lies in the possibility of mass-production for applications such as solar cells. The ease with which one material system can be used to cover a large spectral bandwidth is already attracting a lot of interest, for instance in using passivated and water-soluble quantum dots as biological labels. By using different materials, a huge spectral range can be covered with this technology, as illustrated in Figure 7. Of course, there are still challenges to be met in this field. The quantum dots tend to degrade in the course of time, losing their high quantum efficiency for emission,

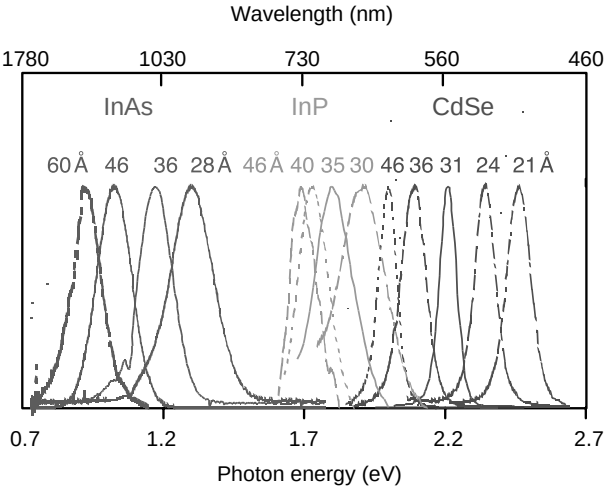


Figure 7 Room temperature emission spectra of surfactant-coated colloidal quantum dots. The quantum dots consist of InAs (left), InP (middle), and CdSe (right). Selection of the size has been used to give a particular wavelength in each band. Reprinted (abstracted/excerpted) with permission from Bruchez M Jr, Moronne M, Gin P, Weiss S and Alivisatos AP (1998) Semiconductor nanocrystals as fluorescent biological labels. *Science* 281: 2013. Copyright 1998 American Association for the Advancement of Science.

presumably as a result of changes to the surface structure. Improvements have been made recently by capping the dots with a high-bandgap semiconductor, such as ZnS, which hinders interactions of the exciton with the surface states. The emission from organic-based dots also tends to show temporal fluctuations in wavelength, revealed experimentally from the emission of a single dot. This is thought to come about through the charging and discharging of traps in the vicinity of the quantum dot. Fundamentally, there are difficulties injecting carriers across an interface into the colloidal dots. However, progress can be expected in all these areas, and these quantum dots may well find applications in novel photonic devices.

Single-Dot Spectroscopy

No matter how quantum dots are prepared, there is always a distribution in emission energy. This constitutes an inhomogeneous broadening mechanism. For self-assembled InAs dots on GaAs, the lowest emission width from an ensemble of dots is about 20 meV. This energy is large compared to several other energies in the system. For example, the binding energy of a charged exciton, or of an exciton–exciton complex (a biexciton), is about 2 meV in this system. Also, the width of single dot emission is two to three orders of magnitude smaller

than the ensemble width at low temperature. The width of single-dot emission is inversely related to the time over which the exciton maintains its coherence, and this should be as long as possible for future applications of quantum dots which manipulate the phase of the excitonic wavefunction. In order to investigate the properties of quantum dots in detail it is clearly necessary to measure the optical response of a single quantum dot rather than the response of a large ensemble of dots. The development of this field has been aided by the techniques developed for single-molecule spectroscopy which was first reported in the late 1980s and has subsequently become almost a standard experimental technique.

Typical densities of self-assembled quantum dots correspond to 100 per μm^2. Given that the emission wavelength is around 1 μm, it is clear that an optical technique with subwavelength spatial resolution is required to excite and detect a signal from just one quantum dot. The techniques employed include cathodoluminescence, where a focused electron beam provides the spatial resolution, and near-field optics where a subwavelength-sized aperture in a metal mask determines the spatial resolution. In the latter case, the aperture is either permanently connected to the sample surface, or it is formed on a tapered optical fiber positioned just a few nanometers away from the sample surface. An example is shown in Figure 8: with a confocal microscope the spatial resolution is not high enough to isolate spectroscopically a single dot. With increased spectral resolution, however, just a small number of peaks remain, each corresponding to the emission from a single dot. A further technique is to etch nanometer-sized mesas in the sample such that each mesa contains on average just one quantum dot. In some cases, single quantum dot emission has been detected with a conventional optical microscope (usually a confocal microscope) by adjusting the growth conditions to give particularly low dot densities. This can be achieved with colloidal nanocrystals by spinning a dilute suspension of the dots onto a substrate, a technique which has been frequently used in single-molecule spectroscopy. In each case, the photoluminescence signal from a single quantum dot is not large, and a detector capable of counting individual photons with a high quantum efficiency is required. A cooled silicon CCD camera is very convenient, giving not only a very low dark count and high quantum efficiency, but also a huge multiplexing advantage. However, the response of a Si detector tails off in the visible, and decreases rapidly at about 1050 nm in the near infrared, so it is by no means a universal solution.

The most striking feature of the emission from a single inorganic quantum dot is its spectral purity. Linewidths as small as a few μeV (a few hundred MHz) have been reported at low temperature for

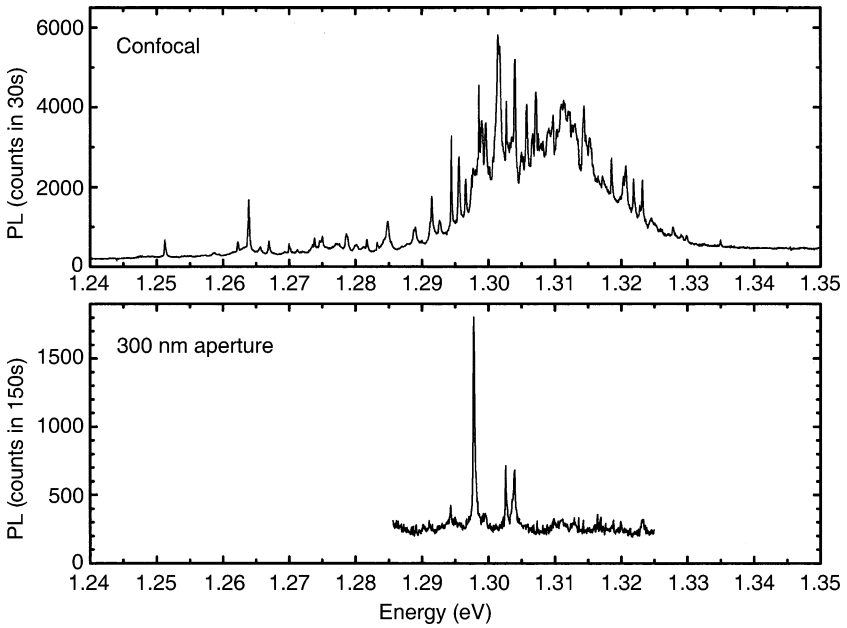


Figure 8 Emission from quantum rings from a 1 μm^2 area of the sample, and from a 0.07 μm^2 area of the sample, both taken at 4.2 K. The quantum rings are generated with a Stranski–Krastanow growth procedure using InAs and GaAs. The first spectrum was taken with a low-temperature confocal microscope; the second with a near-field technique. It can be seen how large statistical fluctuations arise in the first spectrum, because the number of rings probed is not sufficient to give a smooth distribution function. In the second spectrum, just a few peaks can be discerned above the noise level, each arising from the emission from an individual quantum ring.

quantum dots formed in a localizing potential in a quantum well. These linewidths can be interpreted as de-phasing times of the excitons. The maximum possible de-phasing time is obviously the radiative lifetime, in which case the line is homogeneously broadened. This limit has been observed at low temperature on natural quantum dots, but not at elevated temperature: as the temperature increases, scattering of the excitons by phonons decreases the coherence time, observed through a broadening of the emission.

The sharp lines observed in single-dot spectroscopy give access to fine structure on a meV energy scale. Most of this fine structure arises through Coulomb interactions between the electrons and holes. For example, two recent experiments have explored the emission as a function of electron occupation, and exciton population. In both cases, new lines emerge when a single dot is populated by more than one electron–hole pair.

Perspectives

The field of semiconductor quantum dots has made rapid advances in the past decade largely through the discovery of self-assembly in the growth of lattice-mismatched inorganic semiconductors. The most immediate application of these defect-free nanostructures is as the gain medium for semiconductor laser diodes, the aim being to reduce the threshold currents beyond those achievable with quantum wells. This approach is already proving highly successful. However, there are also other possibilities for novel photonic applications of quantum dots. These include memory elements for image processing, detectors and sources for single photons, and perhaps even elements for the coherent manipulation of wavefunctions.

See also

Semiconductor Materials: Band Structure Engineering; GaAs Based Compounds. **Semiconductor Physics:** Band Structure and Optical Properties; Excitons; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures; Recombination Processes.

Further Reading

Bayer M, Stern O, Hawrylak P, Fafard S and Forchel A (2000) Hidden symmetries in the energy levels of excitonic artificial atoms. *Nature* 405: 923.

Bimberg D, Grundmann M and Ledentsov NN (1998) *Quantum Dot Heterostructures*. Chichester, UK: Wiley.

Bruchez M, Jr, Moronne M, Gin P, Weiss S and Alivisatos AP (1998) Semiconductor nanocrystals as fluorescent biological labels. *Science* 281: 2013.

Empedocles SA and Bawendi MG (1997) Quantum-confined Stark effect in single CdSe nanocrystalline quantum dots. *Science* 278: 2114.

Fricke M, Lorke A, Kotthaus JP, Medeiros-Ribeiro G and Petroff PM (1996) Shell structure and electron–electron interaction in self-assembled InAs quantum dots. *Europhysics Letters* 36: 197.

Gaponenko SV (1998) *Optical Properties of Semiconductor Nanocrystals*. Cambridge, UK: Cambridge University Press.

Leonard D, Pond K and Petroff PM (1994) Critical layer thickness for self-assembled InAs islands on GaAs. *Physical Review B* 50: 11687.

Lipsanen H, Sopanen M and Ahopelto J (1995) Luminescence from excited states in strain-induced $In_xGa_{1-x}As$ quantum dots. *Physical Review B* 51: 13868.

Tersoff J, Teichert C and Lagally MG (1996) Self-organization in growth of quantum dot superlattices. *Physical Review Letters* 76: 1675.

Warburton RJ, Dürr CS, Karrai K, *et al.* (1997) Charged excitons in self-assembled semiconductor quantum dots. *Physical Review Letters* 79: 5282.

Warburton RJ, Schäflein C, Haft D, *et al.* (2000) Optical emission from a charge-tunable quantum ring. *Nature* 405: 926.

Weisbuch C and Vinter B (1991) *Quantum Semiconductor Structures*. San Diego, CA: Academic Press.

Xie Q, Madhukar A, Chen P and Kobayashi NP (1995) Vertically self-organized InAs quantum box islands on GaAs(100). *Physical Review Letters* 75: 2542.

Type-II Quantum Wells and Superlattices

I Vurgaftman and J R Meyer, Naval Research Laboratory, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

When thin layers of two semiconductor materials are combined to form a heterojunction, quantum well (QW), or superlattice, the properties of the resulting 'quantum heterostructure' are to a large extent governed by the alignment of the conduction and valence bands at the interfaces. Usually the overall conduction-band (CB) minimum lies in the same material as the overall valence-band (VB) maximum, as for example in the $GaAs/Al_{0.3}Ga_{0.7}As$ QW illustrated in Figure 1a. In such a type-I structure,

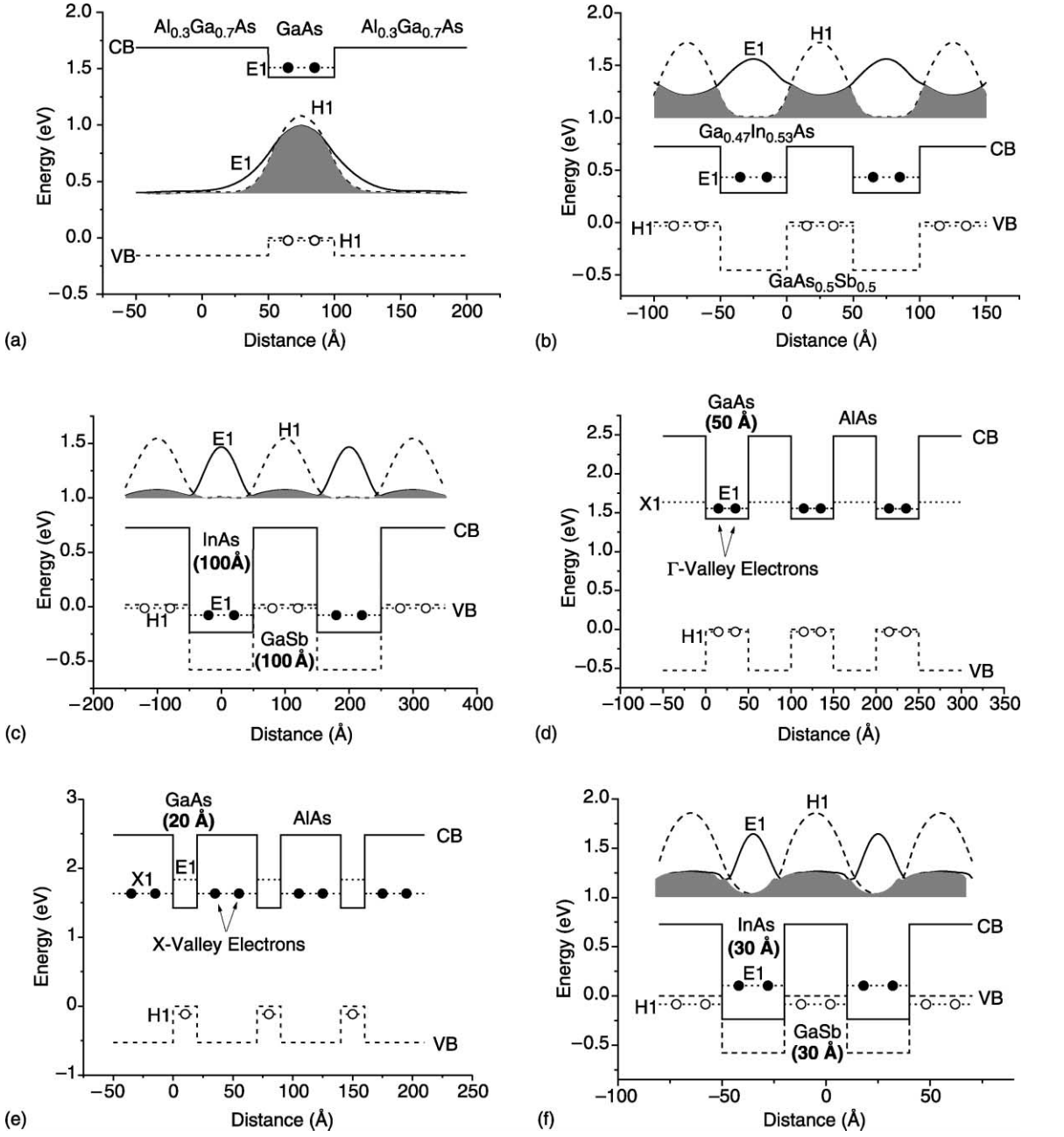


Figure 1 Conduction band (CB) and valence band (VB) profiles along with selected sub-band energies and wavefunctions for: (a) 50 Å GaAs/$Al_{0.3}Ga_{0.7}As$ type-I quantum well (GaAs substrate); (b) 50 Å/50 Å $Ga_{0.47}In_{0.53}As/Ga_{0.5}As_{0.5}Sb$ type-II staggered-gap superlattice (InP substrate); (c) 100 Å/100 Å InAs/GaSb type-II broken-gap semimetallic superlattice (GaSb substrate); (d) 50 Å/50 Å GaAs/AlAs type-I superlattice (GaAs substrate); (e) 20 Å/50 Å GaAs/AlAs type-II superlattice with indirect band alignment both in real and momentum spaces; and (f) 30 Å/30 Å InAs/GaSb type-II semiconducting superlattice. E1, H1, and X1 represent the first sub-band in the Γ-valley conduction band, Γ-valley valence band, and X-valley conduction band, respectively. The overlap between the electron and hole wavefunctions is shown in gray.

the electrons and holes both reside mostly in the GaAs and their wavefunctions overlap almost perfectly as shown in the figure.

There are cases, however, when the CB minimum is in one layer while the VB maximum is in the other. For this so-called type-II alignment, which is exemplified by the $Ga_{0.47}In_{0.53}As/GaAs_{0.5}Sb_{0.5}$ superlattice of Figure 1b, the electrons and holes tend to separate spatially. It should not be surprising that this significantly influences many of the basic optical and electronic properties. Just one example is the band-edge absorption, whose strength is proportional to the square of the wavefunction overlap integral.

In contrast to the staggered type-II heterostructure of Figure 1b, another possibility is the broken-gap type-II alignment of the InAs/GaSb superlattice shown in Figure 1c. Since the CB minimum in InAs actually lies below the VB maximum in GaSb, this combination becomes a semimetal even though the bulk constituents are both semiconductors.

The nature of the band alignment can also be strongly influenced by shifts of the electron and hole energy levels induced by quantum confinement. Figure 1d illustrates a type-I GaAs/AlAs superlattice with relatively thick GaAs layers. Notice that the same superlattice becomes type-II if, as in Figure 1e, the GaAs layers are thin enough that the GaAs Γ-valley electron level (center of the Brillouin zone) is pushed to a higher energy than the AlAs X-valley level (boundary of the Brillouin zone). Similarly, quantum confinement can turn the broken-gap (semimetallic) type-II InAs/GaSb superlattice of Figure 1c into a staggered (semiconducting) type-II superlattice as in Figure 1f.

Although many different type-II heterostructures have been identified over the last three decades, only a few are likely to have long-term technological importance for optics or electronics. The system that has received the most attention in terms of both fundamental scientific studies and opto-electronic device development is InAs/GaSb and related alloy combinations such as InAs/GaInSb. Following a brief overview of wider-gap type-II heterostructures, this article will concentrate primarily on that interesting narrow-gap system.

Wide-Gap Type-II Heterostructures

The best-known wide-gap type-II heterostructure is the GaAs/AlAs superlattice of Figure 1e. The band lineup is type-II provided the GaAs thickness is no greater than ≈10 monolayers (28 Å) and the AlAs layer is no thinner than the GaAs layer. Since the CB minimum is in the AlAs X-valley and the VB maximum is in the GaAs Γ-valley, this structure has an indirect energy gap in both real space and momentum space. In type-II GaAs/AlAs, the X-valley states at the CB minimum form one or more superlattice minibands, which display relatively low resistance at small biases, negative differential velocity, etc. A related structure is the GaAs/AlAs/GaAs/AlAs/GaAs double-barrier intervalley resonant tunneling diode, in which resonant or sequential phonon-assisted tunneling proceeds via intervalley transitions whenever the Fermi level in the GaAs emitter lines up with the X-valley minimum in AlAs. Structures containing AlGaAs do not necessarily need quantum confinement to establish a type-II alignment, since it occurs naturally in $Al_xGa_{1-x}As$/AlAs heterojunctions with $x > 0.3$.

The device potential of wide-gap type-II heterostructures is rather limited, owing to their low luminescence efficiencies (due in large part to the doubly indirect band alignment) and low mobilities (due to the heavy X-valley effective mass). A few proposals have focused on modulation of the band alignment between the type-I and type-II regimes. The $Ga_{0.47}In_{0.53}As/GaAs_{0.5}Sb_{0.5}$ superlattices of Figure 1b and related structures have been investigated as alternative active regions for semiconductor lasers emitting in the important 1.3–2 μm wavelength range.

InAs/GaSb/AlSb-Based Type-II Superlattices and Quantum Wells

Molecular beam epitaxy (MBE) was used to fabricate the first InAs/GaSb-based superlattices in 1977, only a few years after the initial GaAs/AlGaAs superlattice growths. It was quickly demonstrated that for relatively thick layers the band alignment is broken-gap type-II, with 100–150 meV overlap between the conduction band minimum in InAs and the valence band maximum in GaSb. In this semimetallic superlattice, GaSb valence-band electrons spill over into the InAs and leave behind holes (see Figure 1c). It was also observed that at equal layer thicknesses of ≈90 Å quantum confinement induces a semimetal-to-semiconductor transition (Figure 1f), and that the gap becomes increasingly positive when the thickness is reduced further. This is no mere scientific curiosity, since being able to tune in a wide range of energy gaps spanning 0 to >0.5 eV has enormous practical consequences for infrared (IR) emitters and detectors. Such a capability cannot be replicated by any other III–V heterostructure system.

Another happy accident is that the lattice constants of InAs (6.058 Å) and GaSb (6.096 Å) are matched closely enough to allow them to be grown together by MBE or metalorganic chemical vapor deposition (MOCVD). A third binary member of the 'antimonide' or '6.1-Å' family of growth-compatible heterostructures is AlSb (6.136 Å), which is valuable in that it serves as a quantum barrier to both InAs conduction states and GaSb valence states. The extended family includes a variety of ternary alloys (GaInSb, GaAlSb, InAsSb, AlAsSb, etc.), whose energy gaps as a function of lattice constant are given by the curves in Figure 2, and also at least three quaternary alloys: GaInAsSb, AlGaAsSb, and InAlAsSb. Taken together, this menagerie of materials

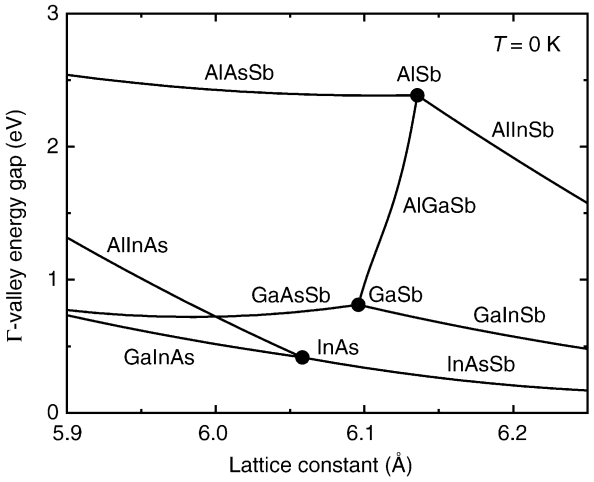


Figure 2 Direct energy gaps at zero temperature for III–V compound semiconductors and their ternary alloys with lattice constants in the vicinity of 6.1 Å.

provides the type-II IR device designer with an exceptional degree of configurational flexibility.

Type-II IR structures are typically grown on either a GaSb or InAs substrate. Since even a small amount of strain build-up can produce dislocation defects that seriously degrade the performance, one usually balances the net tensile strain (associated with layers having a lattice constant slightly larger than that of the substrate) with a compensating amount of net compressive strain (from layers whose lattice constant is slightly smaller than that of the substrate).

Band structures for complicated type-II layering configurations have been calculated by a variety of methods, including $k \cdot p$, tight-binding, and empirical pseudopotentials. While all reproduce the main features of the electron and hole sub-bands and dispersion relations near the center of the Brillouin zone, the relative merits of the different approaches remain controversial. Due to the small energy gaps, band-mixing effects are especially important in this system.

One consequence of the small type-II antimonide energy gap is that the electron and hole bands repel each other and both effective masses can be lighter than in any other III–V semiconductor (e.g., smaller than in the bulk constituents). This leads to high mobilities at room temperature, and promising possibilities for high-electron-mobility transistors (HEMTs). Also of interest are resonant tunneling diodes (RTDs) based on InAs/AlSb/GaSb/AlSb/InAs and related structures, in which interband tunneling via the GaSb valence states results in negative differential resistance with a very high peak-to-valley ratio. Since these devices are primarily electronic rather than optical, we will not discuss them further here.

While interfaces play an important role in any semiconductor heterostructure, they are especially crucial in InAs/GaSb and related type-II systems because the two materials on opposite sides of the interface have neither a common cation nor a common anion (unlike GaAs/AlGaAs, for example, in which both wells and barriers have As anions). By timing the shutter sequences in the growth, one can in principle end the InAs on an In layer and begin the GaSb on Sb, to produce InSb-like interface bonds. Alternatively, one can end the InAs on As and begin the GaSb on Ga to form GaAs-like bonds. Detailed characterizations by a variety of methods have shown that the device properties and even the band structure can be relatively sensitive to whether the interface bonds are InSb-like or GaAs-like.

With regard to opto-electronic devices, a potential limitation of InAs/GaSb superlattices is that in order to reach a long wavelength (small energy gap), the layers must be rather thick (e.g., ≈90 Å for zero gap). In such a structure the overlap between the electron and hole wavefunctions becomes quite small (see Figures 1c and 1f), and consequently the interband optical interactions are weak. However, Mailhiot and Smith pointed out that this issue can be circumvented to a large extent by employing $Ga_{1-x}In_xSb$ (with $x \leq 0.4$) rather than GaSb for the hole QWs. Compressive strain in the GaInSb produces a large splitting of the heavy and light hole sub-bands, along with a substantial increase of the type-II offset energy. The result is that much thinner layers (with large wavefunction overlaps) may be employed in conjunction with energy gaps corresponding to long IR wavelengths. In fact, the overlap for a short-period InAs/GaInSb superlattice such as that illustrated schematically in Figure 3 can be large enough to yield interband optical interaction strengths comparable to those in bulk materials and type-I superlattices.

While practical constraints on the minimum reproducible layer thickness set a lower wavelength limit on the order of 2 μm, the upper limit governed by uniformity considerations is quite high, perhaps in the $\lambda = 25-100$ μm range. This makes type-II antimonide quantum heterostructures a natural choice for mid-IR to far-IR photodetectors and lasers.

Type-II IR Photodetectors

Because IR detectors sense the radiant black-body emission from warm objects rather than relying on reflected visible light, they can be used for 'night vision.' Both photoconductive (PC) and

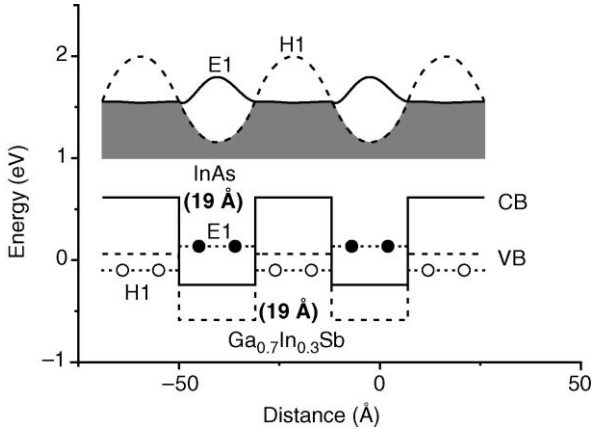


Figure 3 Conduction and valence band profiles along with electron (E1) and hole (H1) sub-band energies and wavefunctions for a 19 Å/19 Å InAs/$Ga_{0.7}In_{0.3}Sb$ type-II superlattice (GaSb substrate) with an energy gap equal to that of the InAs/GaSb superlattice in **Figure 1d**.

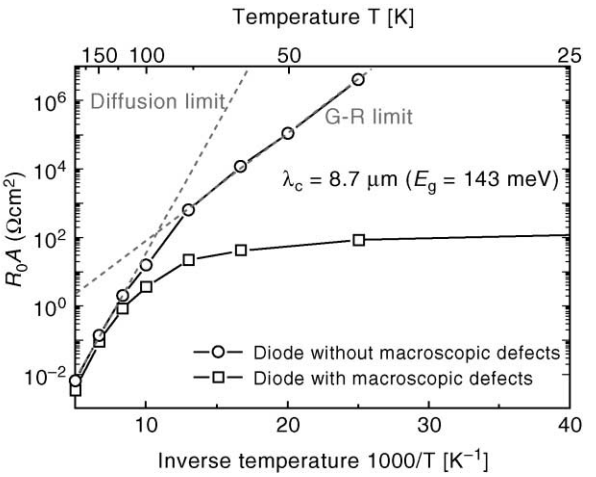


Figure 4 Temperature dependence of the dynamic impedance at zero bias R_0A for two different InAs/GaInSb superlattice diodes with cutoff wavelengths of 8.7 μm. The generation-recombination (G-R) and diffusion-limited R_0A values are indicated by the dashed lines. Reproduced with permission from Bürkle *et al.* (2000) *MRS Symp. Proc.*, 607, pp. 77–82.

photovoltaic (PV) IR detectors are now being developed with type-II active regions, usually consisting of simple InAs/GaInSb superlattices. The PV geometry, in which a reverse bias is applied to a p–n junction, is ultimately preferred since it is easily adaptable to pixelation in a 2D array with many discrete detector elements. That allows an IR camera to construct a detailed thermal picture of any given scene.

An optimized PV photodetector must have high absorption in the active region, unimpeded vertical transport to assure collection of the photogenerated carriers, and a low dark current. The thin layers of a type-II InAs/GaInSb active region help to satisfy the first two criteria, since they assure both a large wavefunction overlap (**Figure 3**) and rapid electron and hole tunneling through the superlattice minibands.

The dark current in a high-quality diode tends to be dominated by generation-recombination currents at low temperatures and diffusion currents at higher *T*. However, trap-assisted tunneling via dislocations and other macroscopic defects can significantly increase the dark current. **Figure 4** plots the resistance–area product (R_0A) for diodes with and without macroscopic defects. Surface leakage resulting from inadequate sidewall passivation can also lead to a substantial contribution that is proportional to the perimeter of the mesa. The diffusion current scales as $1/\tau^{1/2}$, where τ is the recombination lifetime. To date, the defect-mediated Shockley–Read lifetimes of <100 ns have been significantly shorter than in competing HgCdTe IR materials (>1 μs). However, Auger lifetimes are substantially enhanced, as will be discussed further in the context of type-II IR lasers. In an Auger event, the energy and momentum of the recombining electron–hole pair is transferred to a third carrier, which scatters to a higher-lying state in either the same or a different band. Because three different carriers must interact simultaneously, the recombination rate scales with the cube of the carrier density. Usually Shockley–Read recombination dominates at lower temperatures and Auger processes at higher *T*.

Typically the n-side of a type-II diode is heavily doped ($10^{18}\,cm^{-3}$ range) and relatively thin, while most of the absorption occurs on the lightly doped ($10^{16}\,cm^{-3}$ range) p-side. A standard figure of merit is the detectivity (D^*), which is essentially the signal-to-noise ratio in a device of standard size. For 77 K operation, type-II PV detectors with cutoff wavelengths of 7.5 μm and 12 μm achieved D^* of 1×10^{12} cm $Hz^{1/2}$/W and 5×10^{10} cm $Hz^{1/2}$/W, respectively. Those values are nearly as high as typical results for the much more mature HgCdTe FPA technology.

It is anticipated that type-II detectors will be particularly attractive at very long wave IR (VLWIR), at which it becomes increasingly difficult to maintain adequate control over the HgCdTe compositional uniformity. Recent PC detectors operated out to 22 μm, and PV devices to 16 μm. Due in part to the strong suppression of Auger recombination, type-II detectors are also expected to be advantageous at noncryogenic temperatures where future systems will increasingly operate.

Type-II IR Lasers

At visible and near-IR wavelengths out to $\approx 2\ \mu m$, the QW diode laser is by now a mature technology that dominates the commercial marketplace due to its low cost (<$1 per laser in some cases), compactness, and reliability. However, the technical challenges become much more severe in the mid-IR spectral region ($3-5\ \mu m$). Before the advent of the type-II laser, no interband III–V semiconductor device emitting at $\lambda \geq 3\ \mu m$ had ever operated at room temperature. Alternatives are the lead salt laser, which is limited to very low-output powers, and the quantum cascade laser (QCL) that is based on optical transitions between two electron QW sub-bands rather than interband electron–hole processes. Mid-IR lasers are needed for high-power military applications such as IR countermeasures against heat-seeking missiles and for the sensitive detection of trace chemicals using the spectroscopic 'fingerprint' that distinguishes each species from any other.

From the 1960s to the late 1990s, conventional III–V double heterostructure (DH) and type-I QW lasers for the mid-IR never achieved the desired levels of performance because they failed to overcome certain fundamental limitations. Those included:

1. inadequate band offsets, which at higher temperatures made it difficult for the QW or DH active regions to retain the injected electrons and holes;
2. rapid Auger nonradiative decay, which led to excessive threshold current densities at higher temperatures; and
3. restricted spectral range, limited to wavelengths shorter than the cutoff for bulk $InAs_{1-x}Sb_x$.

However, it has more recently become apparent that all of these limitations can be either minimized or removed by employing a type-II InAs/GaSb-based active region.

For example, electrical confinement is far more effective because the GaSb conduction band confines InAs electrons and the InAs valence band confines GaSb holes (see **Figure 1f**). Auger recombination is also significantly suppressed, in part because the holes are nearly as light in the plane as the electrons, making it much more difficult to conserve both momentum and energy in the three-carrier process. Nearly an order-of-magnitude suppression of the room-temperature Auger rate has been reported, which translates directly into a similar reduction of the threshold current density. And as was discussed above in connection with IR photodetectors, very long wavelengths are readily attainable because the energy gap in a type-II active region can be arbitrarily small. Type-II lasers have already operated cw at $\lambda > 7\ \mu m$, and cryogenic operation out to wavelengths as long as $100\ \mu m$ has been projected theoretically.

In 1986 the first type-II semiconductor laser, which emitted at $\lambda = 1.86\ \mu m$ from single liquid-phase epitaxy (LPE)-grown heterojunctions, was demonstrated. Since the mid-1990s, $3-4\ \mu m$ GaInAsSb/InAs-based single-interface type-II diodes have been developed that operate up to 205 K in pulsed mode.

In 1995, lasing at $\lambda = 3.2-3.5\ \mu m$ was reported from type-II $InAs/Ga_{0.75}In_{0.25}Sb$ superlattice diodes grown by MBE. Subsequent work on these devices extended the temperature range to 255 K in pulsed mode and 180 K for cw operation, and the emission wavelength to $4.3\ \mu m$.

While the wavefunction overlap and hence the gain for a two-constituent type-II superlattice active region can be quite favorable, penetration of the electron wavefunctions into the thin GaInSb layers is in fact so great that a wide miniband forms along the growth axis. The electrons then have strong energy dispersion in all three dimensions, whereas present-generation near-IR and visible semiconductor lasers nearly always have QW active regions with 2D carriers. The stepped quasi-2D density of states yields a much higher differential gain per injected carrier at threshold, and also narrower spectral emission lines than are attainable with a 3D electron population. The most straightforward approach to blocking the formation of a miniband in the superlattice is to add a high electron barrier between the periods, e.g., InAs/GaInSb/AlSb. However, the drawback of that approach is that the wavefunction overlap becomes much smaller than in the simple superlattice, resulting in significantly lower gain.

On the other hand, if a second electron well is added on the other side of the hole well, the resulting structure with four constituents per period (e.g., InAs/GaInSb/InAs/AlSb) combines the large wavefunction overlap of the two-constituent superlattice with 2D electron and hole confinement. **Figure 5** illustrates band profiles, energy levels, and wavefunctions for the 'W' laser, which takes its name from the shape of the conduction-band profile.

Optically pumped type-II W lasers display high-power pulsed operation at temperatures well above ambient. Wavelengths have spanned the range $2.6-7.3\ \mu m$, and mid-IR vertical-cavity surface-emitting lasers (VCSELs) were also fabricated. Optical pumping is viewed primarily as a means of achieving high cw or quasi-cw output powers in the mid-IR, and in that regard W lasers have exceeded all other semiconductor approaches with 1.5 W peak output power being obtained at 71 K for $100\ \mu s$

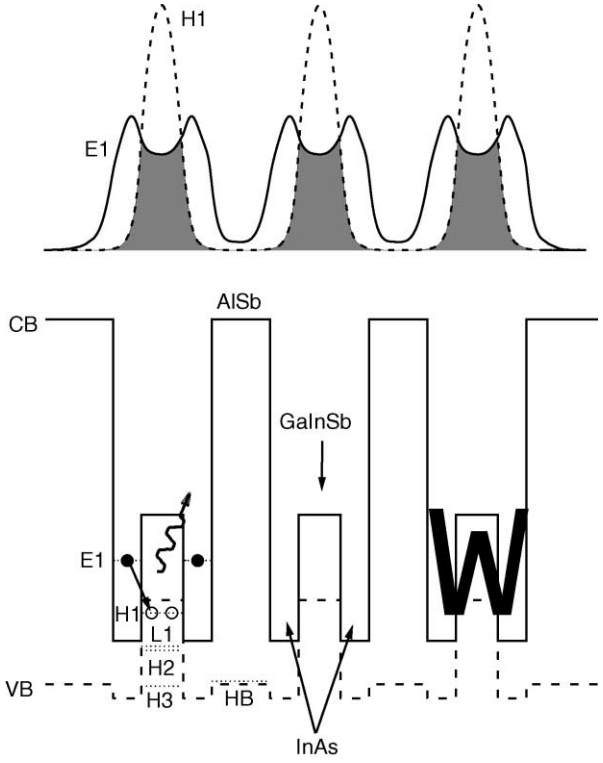


Figure 5 Conduction and valence band profiles along with selected sub-band energies and wavefunctions for InAs/$Ga_{0.7}In_{0.3}Sb$/InAs/AlSb type-II 'W' structure (GaSb substrate). Electron (E1), heavy hole (H1, H2, H3), light hole (L1), and barrier heavy hole (HB) sub-bands are indicated.

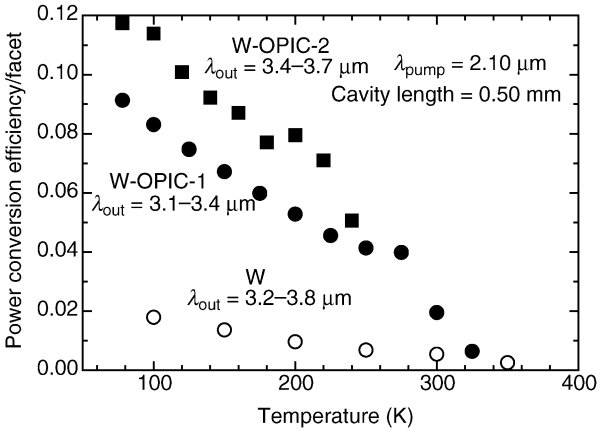


Figure 6 Power conversion efficiency per uncoated facet as a function of temperature for two different W-OPIC lasers emitting in the wavelength ranges of 3.1–3.4 μm (filled circles) and 3.4–3.7 μm (filled squares), respectively. For comparison, conversion efficiencies for a typical 'W' laser without an OPIC emitting in the same spectral range (open circles) are also shown. All of the lasers were optically pumped in pulsed mode by a 2.1 μm Ho:YAG laser, and the cavity length for all three devices was 0.5 mm.

pulses with a 10% duty cycle. The average output at 82 K for that device was up to 360 mW for 20 μs pulses and a 50% duty cycle.

Cw operation at $\lambda = 3.0$ μm was observed nearly to room temperature (290 K), and even a 6.1 μm emitter operated cw to 210 K. The cw output power of 540 mW/facet from a device with 80 W QW periods was the highest ever reported for 78 K operation in the mid-IR. Despite the rather high output power, however, the differential power conversion efficiency was only 2.5% at 78 K and fell to 1% at 140 K. Much higher efficiencies were obtained using the optical pumping injection cavity (OPIC) approach, in which the active region was surrounded by semiconductor quarter-wave mirror stacks that formed an etalon cavity resonant at the 2.1 μm pump wavelength. For pulsed operation, OPIC devices with uncoated facets and only 10 active QWs yielded lower thresholds, lower internal losses, and higher power conversion efficiencies than any previous results at $T \geq 200$ K. Figure 6 illustrates that the efficiencies were 7.1% at 220 K for one device and 4% at 275 K for another.

An alternative approach to maximizing the pump absorbance despite a small number of QWs is the 'integrated absorber' (IA) concept. In this approach, the W QWs are separated by intermediate-gap GaInAsSb absorbing layers which donate optically generated electrons and holes to the active wells. At 80 K with 35 μs pulses at 2.5% duty cycle, an IA laser with only five QWs separated by ≈1600-Å-thick absorber layers had a power conversion efficiency of 9.8% and peak output power of 2.1 W. Thus far the IA has not done as well as the OPIC at high temperatures, due mostly to an inadequate valence band offset between the absorber layers and the active QWs.

The beam quality for optically pumped type-II lasers was substantially improved by using the angled-grating distributed-feedback (α-DFB) approach pioneered at shorter wavelengths. Since both the grating and the facets are needed to produce optical feedback, only a narrow angular cone is fed back into the tilted cavity. For a pump-stripe width of 50 μm, the near-diffraction-limited output beam had 15 times narrower angular divergence (1.4° full width at half-maximum (FWHM)) than the double-lobed beam emitted from a conventional Fabry–Perot laser occupying an unpatterned portion of the same bar. Even more promising results have been reported using the photonic-crystal distributed-feedback (PCDFB) laser, in which the grating is defined on a two-dimensional rectangular lattice that is tilted with respect to the facets as in the α-DFB. For a pump-stripe

width of 200 μm, the beam quality for the direction along the laser stripe was four times the diffraction limit.

Although optical pumping is the simplest way to inject electrons and holes into the active region, type-II diode lasers will ultimately be much more inexpensive, convenient, compact, and energy efficient. However, generalization of the W configuration to electrical pumping is not entirely straightforward. This is because holes injected from the p-type cladding layer may have poor growth-axis transport, and hence hole populations in the various GaInSb QWs could be highly nonuniform. Whereas the holes in a superlattice are required to tunnel through only one InAs electron well at a time (see Figure 3), the barrier for the structure in Figure 5 consists of two InAs layers in addition to the AlSb barrier. To overcome this problem the AlSb was replaced with a specially designed quaternary $Al_{0.15}Ga_{0.85}As_{0.05}Sb_{0.95}$ barrier, which raised the L1 light-hole sub-band to within ≈100 meV of the H1 heavy-hole valence band maximum. That permitted thermally excited light holes to provide the needed tunneling transport to all of the QWs. Another key design element was the insertion of relatively thick (0.6 μm) undoped $Al_{0.35}Ga_{0.65}As_{0.03}Sb_{0.97}$ 'broadened-waveguide' layers on each side of the active region, in order to minimize free-carrier absorption in the doped $Al_{0.9}Ga_{0.1}As_{0.07}Sb_{0.93}$ cladding layers. Some of the devices also contained InAs/AlSb 'hole-blocking' layers, which prevented hole leakage into the n-type broadened waveguide layer.

A pulsed $\lambda = 3.27$ μm W-diode of this type with 10 active QW periods achieved the first room temperature operation (to 310 K) of an electrically pumped III–V interband laser in the mid-IR. Another device with five QWs operated in cw mode to 200 K, the highest T_{max} to date for any interband III–V laser beyond 3 μm. The threshold at 78 K was 63 A/cm^2, and Figure 7 shows that the cw output was 140 mW at 78 K and 20 mW at 180 K. Following optimization of the turn-on voltages, series resistances, and internal losses, cw diodes with one–three QWs are projected to operate in the thermo-electric cooler range and ultimately at room temperature.

A more exotic variation on the type-II diode emitter is the interband cascade laser (ICL) that was first proposed by RQ Yang. Like the intersub-band QCL, the ICL generates multiple photons per injected electron by making an optical transition at each step of a staircase-like QW structure. While the threshold current density for the QCL is high because of the inherently short lifetimes associated with intersub-band phonon scattering, the interband configuration of the ICL eliminates that nonradiative relaxation

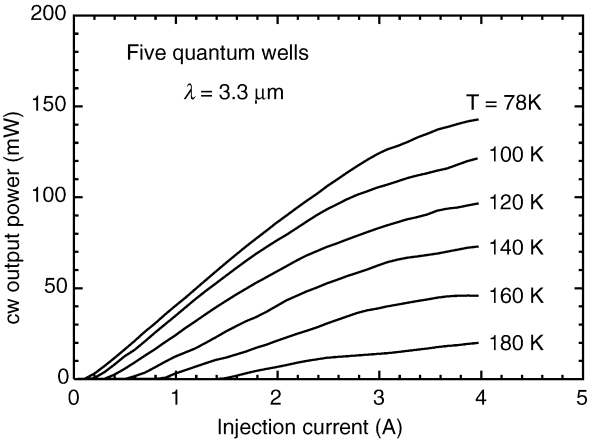


Figure 7 Cw output power as a function of injection current at temperatures between 78 K and 180 K for a 'W' diode laser with a 2-mm-long cavity and 100-μm-wide stripe emitting at a wavelength of 3.3 μm.

path. Early projections that the ICL should display low thresholds, high cw operating temperatures, high output powers, and high differential quantum efficiencies (well above the conventional limit of one photon/electron) have all been confirmed experimentally.

Figure 8 illustrates conduction and valence band profiles for a typical single stage of the ICL staircase. While this example employs a W configuration, in which two InAs electron QWs (19 Å and 13 Å) surround the active GaInSb hole QW (32 Å), some designs employ only a single InAs active well. Electrons tunneling into the InAs QWs from the preceding injection region at left emit a photon by making spatially indirect radiative transitions to the GaInSb valence band. They next tunnel into the valence states of the adjacent GaSb and GaInSb hole wells, whose function is to provide a thick barrier to prevent electron leakage from the active InAs QWs directly to the 70 Å InAs well that begins the next injection region. From the final GaInSb hole QW (40 Å), the carriers undergo near-elastic interband scattering into conduction states of the 70 Å InAs QW. Electrons traversing the superlattice miniband of the injection region, which in this case consists of eight digitally graded InAs wells, finally tunnel into the active electron QWs of the next period. This series of events is repeated at each step of the ICL staircase.

Following the first ICL demonstration in 1997, there have been reports of threshold current densities 13 A/cm^2 at 80 K, maximum operating temperatures of up to 300 K for pulsed operation and 150 K for cw (with epitaxial-side-up mounting), pulsed output powers of 6 W/facet and slope quantum efficiencies exceeding 600%. A cw wall plug efficiency of 22%

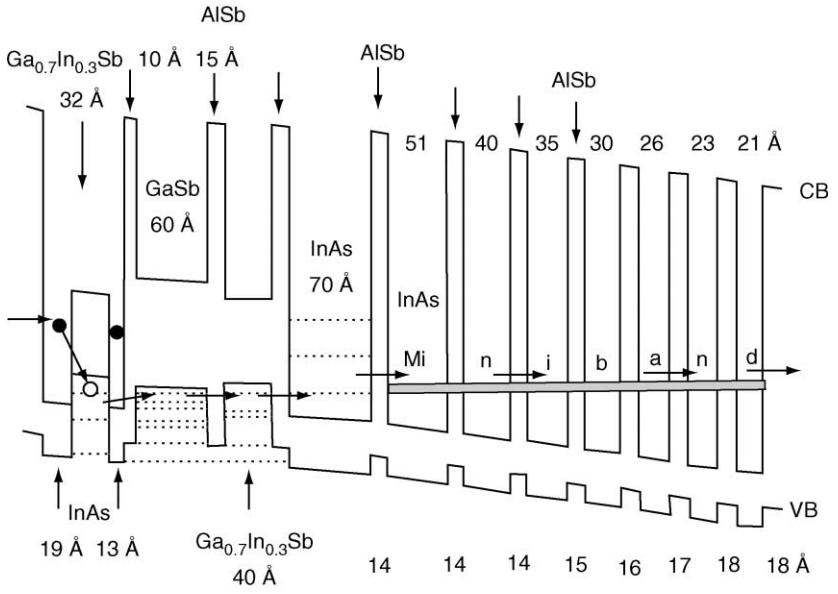


Figure 8 Conduction and valence band profiles for one period of an interband cascade laser with a W active region and three hole wells. Sub-band energy levels are also indicated.

at 80 K was recently reported. While substantial further improvements are expected, the primary disadvantage of the ICL is its complexity.

Despite numerous dramatic advances by type-II mid-IR lasers since their first demonstration in 1995, the performance has not yet progressed to the level that will be required to support a robust commercial technology. One key issue is the cw operating temperature, which must increase to at least the thermo-electric cooler range if the potential market is to extend beyond niche applications. Higher-temperature operation will result automatically once the number of QWs, and hence the threshold current density, can be reduced. Further suppression of the Auger decay rate would also be highly beneficial if it is proven possible.

The laser output power and efficiency are governed by the internal loss, which in type-II QWs tends to be dominated by resonant free-hole intersub-band transitions. The details of the valence sub-band structure are not well known at present, although recent experiments have apparently identified a number of the transition energies. As in the case of the threshold, reducing the number of QWs is highly beneficial since it reduces the number of free carriers that can absorb the lasing photons. The optically pumped OPIC and IA structures discussed above achieved dramatic reductions in the internal losses by following that strategy.

The beam quality attainable from a semiconductor laser is governed by the linewidth enhancement factor (LEF), which is proportional to the variation in the refractive index induced by a small change in the carrier density. A large positive value of the LEF leads to filamentation in broad-area lasers as tiny peaks in the optical intensity cause carrier depletion and a local increase in the refractive index, which in turn stimulates further focusing of the output beam. The output then consists of an incoherent superposition of many rapidly diverging filaments, resulting in a beam quality that is several orders of magnitude worse than the diffraction limit. While theoretical LEFs for optimized type-II W structures can be in the 1.5–2 range even at room temperature, recent Hakki–Paoli measurements have yielded values in the 2.5–5 range. This is possibly due to additional inhomogeneous broadening of the gain spectrum induced by interface roughness.

See also

Semiconductor Materials: Band Structure Engineering; Quantum Dots. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures; Recombination Processes.

Further Reading

Mailhiot C and Smith DL (1989) Long-wavelength infrared detectors based on strained InAs-GaInSb type-II superlattices. *Journal of Vacuum Science and Technology* A7: 445–449.

Meyer JR, Malin JI, Vurgaftman I, Hoffman CA and Ram-Mohan LR (1997) Antimonide-based quantum heterostructure devices. In: Manasreh MO (ed.) *Strained*

Quantum Wells and Their Applications, pp. 235–272. Amsterdam, The Netherlands: Gordon and Breach.

Olafsen LJ, Vurgaftman I and Meyer JR (in press) Antimonide mid-IR lasers. In: Choi HK (ed.) *Long-Wavelength Semiconductor Lasers*. New York: Wiley.

Ram-Mohan LR, Vurgaftman I, Meyer JR and Dossa D (2000) Wavefunction engineering: a new paradigm in quantum nanostructure modeling. In: Nalwa HS (ed.) *Handbook of Nanostructured Materials and Nanotechnology*, vol. 2, pp. 707–739. New York: Academic Press.

Razeghi M (1998) Current status and future trends of infrared detectors. *Optoelectronics Review* 6: 155–194.

Vurgaftman I, Felix CL, Aifer EH and Meyer JR (2000) Interband III–V mid-IR semiconductor lasers. In: Perera AGU and Liu HU (eds) *Handbook of Thin Film Devices*, vol. 2, pp. 171–193. San Diego, CA: Academic Press.

Vurgaftman I, Felix CL, Bewley WW, *et al.* (2001) Mid-IR "W" lasers. *Philosophical Transactions of the Royal Society A* 359: 489–504.

Vurgaftman I, Meyer JR and Ram-Mohan LR (2001) Band structure parameters for III–V compound semiconductors and their alloys. *Applied Physics Reviews, Journal of Applied Physics* 89: 5815–5875.

Wilson BA (1988) Carrier dynamics and recombination mechanisms in staggered-alignment heterostructures. *IEEE Journal of Quantum Electronics* 24: 1763–1777.

Yang RQ, Bradshaw JL, Bruno JD, Pham JT and Wortman DE (2002) Mid-IR type-II interband cascade lasers. *IEEE Journal of Quantum Electronics* (Feature Section on Mid-IR Quantum Cascade Lasers) 38: 559–568.

SEMICONDUCTOR PHYSICS

Contents

Outline of Basic Electronic Properties
Band Structure and Optical Properties
Excitons
Impurities and Defects
Infrared Lattice Properties
Light Scattering
Polarons
Quantum Wells and GaAs-Based Structures
Recombination Processes
Spin Transport and Relaxation in Semiconductors; Spintronics
Surface Photovoltage Spectroscopy of Semiconductors

Outline of Basic Electronic Properties

C R Pidgeon, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

We briefly introduce the main concepts underlying the detailed treatment of specialist aspects of semiconductor physics in ensuing sections, viz: the free electron model, density of states, Fermi-Dirac statistics, energy bands and energy gaps, effective mass, electrons and holes, intrinsic and extrinsic conductivity, direct and indirect interband optical transitions, excitons and low-dimensional systems.

Introduction

A proper treatment of optoelectronic devices needs to include the physical properties of the materials from which the devices are made, and this subject is covered under the section of the Encyclopedia entitled 'Semiconductor Fundamentals'. This includes sections on the basic physics common to all semiconductors useful in optoelectronic applications, the experimental techniques used to investigate the optical properties of

these semiconductors and the more detailed properties of the most widely used semiconductors, and semiconductor systems. The properties of greatest importance here are those of electrons in crystalline solids, but the interaction between the electrons and the atoms of the crystal lattice is also significant. The combination of these microscopic effects is responsible for determining the macroscopic optical and electrical properties of semiconductors, and in this section we give the necessary background to these topics.

The classical treatment is extremely useful in describing the electrical conductivity of solids; the electrons are regarded as charged particles obeying Newton's laws of motion, and electrical resistance is assumed to arise from collisions with the vibrating lattice and/or impurities. However, the classical theory does not give a method of predicting the natural frequencies, or energies, of these electronic oscillators. To do this we need to use quantum mechanics, and the elements of this are reviewed here. We start with the free electron theory of metals, where the electrons are treated as particles constrained to move in a potential box. For convenience we begin by considering the situation in one dimension (1D).

Free Electron Theory of Metals

Free electrons in a 1D potential box of length L in the x-direction are described by a plane wave function that is normalized within the box, $\psi = (1/L)e^{ikx}$, and satisfies the Schrödinger equation in the x-direction:

$$\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + E\psi = 0 \qquad [1]$$

where E is the electron energy. This gives the familiar parabolic E–k dispersion relation (see Figure 1), $E = \hbar^2k^2/2m$ (with electron momentum $p = \hbar k$). In order to evaluate physical problems it is necessary to define the boundary conditions for the wave function. We assume fixed boundary conditions (i.e., $\psi = 0$ at $x = 0, L$) leading to the quantization condition $k = n\pi/L$ where n is a positive integer.

Density of States

Physical properties of solids depend on the number of allowed k-values per unit range of k (or E-values per unit range of E), a quantity referred to as the density of states, D(k) or D(E) – for example, the number of allowed states between k and $k + dk = D(k)dk$. From the fixed boundary condition ($k = n\pi/L$, the number of allowed values of k per unit range of k in 1D is just

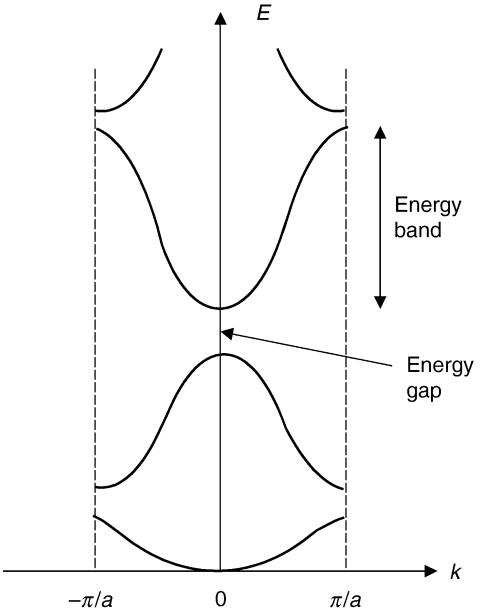


Figure 1 Schematic representation of energy bands in one dimension, characterized by gaps at the high symmetry points of the Brillouin zone.

$D(k) = L/\pi$. In terms of energy, using the E–k dispersion relation, we can write (including the factor 2 for two possible spin states):

$$D(E)dE = D(k)(dk/dE)dk$$
$$= (L/2\pi)(2m/\hbar^2)^{1/2}E^{-1/2}dE \qquad [2]$$

i.e., the density of states depends on $E^{-1/2}$ for the 1D situation.

For a cube of side L, in three dimensions we generalize the identity for fixed boundary conditions:

$$\frac{L^2k^2}{\pi^2} = n_x^2 + n_y^2 + n_z^2 = R^2 \qquad [3]$$

where $k_x = n_x\pi/L$ etc., and the states fill up the positive quadrant of a sphere of radius ($R = Lk/\pi$) in k-space. Thus, the number of states in a given volume of k-space is numerically equal to that volume, so that the number of states between R and $R + dR$ in the positive quadrant (hence $1/8^{th}$) of a spherical shell of radius R in k-space is

$$D(R)dR = 4\pi R^2 dR(1/8) \qquad [4]$$

i.e., $(1/8)^{th}$ of the surface area of the shell multiplied by the thickness of the element. Changing the variable to k, the number of states between k and $k + dk$ is

$$D(k)dk = D(R)\frac{dR}{dk}dk = \frac{V}{2\pi^2}2k^2dk \qquad [5]$$

where the volume $V = L^3$, and a factor of 2 has again been included to account for the two possible spin states of the electron. This in turn, through the relation between E and k, can be expressed in terms of the density of states per unit energy so that the number of states between E and $E + \mathrm{d}E$ is given by:

$$D(E)\mathrm{d}E = D(k)\frac{\mathrm{d}E}{\mathrm{d}k}\mathrm{d}E = \frac{2V}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{\frac{3}{2}}E^{\frac{1}{2}}\mathrm{d}E \quad [6]$$

In two dimensions, by the same argument:

$$\frac{L^2k^2}{(\pi)^2} = n_x^2 + n_y^2 = R^2 \quad [7]$$

Therefore $D(R)\mathrm{d}R = (2\pi R/4)\,\mathrm{d}R$, where the factor 1/4 implies the positive quarter of the 2D ring between R and $R + \mathrm{d}R$ in k-space, $A = L^2$ and

$$D(k)\mathrm{d}k = D(R)\frac{\mathrm{d}R}{\mathrm{d}k}\mathrm{d}k = \frac{A}{\pi}k\mathrm{d}k \quad [8]$$

From the E–k relation we have the number of states between E and $E + \mathrm{d}E$ as:

$$D(E)\mathrm{d}E = D(k)\frac{\mathrm{d}E}{\mathrm{d}k}\mathrm{d}E = \frac{A}{\pi}\left(\frac{m}{\hbar^2}\right)\mathrm{d}E \quad [9]$$

Thus, in 2D systems we have the important result that $D(E)$ is a constant.

Fermi-Dirac Distribution of Electrons

At zero temperature an atom exists in what is known as the ground state. All the electrons rest in the lowest energy states allowed to them by the Pauli principle, and fill the allowed states up to an energy known as the Fermi energy, E_F. Above absolute zero of temperature this is no longer the case, and thermal energy allows some electrons to exist in higher energy states. The probability $f(E)$ that a state of energy E will be occupied for an ideal electron gas in thermal equilibrium (thermal energy kT) is governed by Fermi-Dirac statistics:

$$f(E) = \frac{1}{1 + \exp\left(\dfrac{E - E_\mathrm{F}}{kT}\right)} \quad [10]$$

For a metal at room temperature the difference between the Fermi energy and the energy of the conduction band minimum is much greater than the thermal energy, $(E_\mathrm{F} - E_\mathrm{co}) \gg kT$; this is also true for a strongly n-type semiconductor (see below) and the distribution is said to be degenerate.

Energy Bands and Energy Gaps: Semiconductors

In the presence of the (perturbing) periodic crystal potential, $V(r)$, the above simplicity is lost: this is the so-called 'nearly free electron model', and the perturbed Schrödinger equation is now:

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} + V(x)\psi = E\psi \quad [11]$$

for the x-direction. This topic is discussed in all undergraduate texts on solid state physics, but we list the main results of the model as follows:

1. The electrons that we are concerned with are confined within a periodic lattice. Consequently the wave function ψ is modulated with the periodicity of the lattice. By symmetry arguments the correct wave function can be shown to be:

$$\psi_k(x) = \frac{1}{\sqrt{L}}\mathrm{e}^{ikx}u_k(x) \quad [12]$$

   where $u_k(x)$ is a function that has the symmetry of the lattice. $\psi_k(x)$ is known as the Bloch function and is a satisfactory solution to the Schrödinger equation. Because of the periodic nature of the Bloch equation, it is always possible to select the wavevector index k of any Bloch function to lie within the range $\pi/\mathrm{a} < k < \pi/a$. This region is referred to as the first Brillouin zone, and the representation where the E–k dispersion is mapped back entirely into the first Brillouin zone is called the 'reduced zone scheme' (Figure 1).
2. Regions of E appear where solutions $\psi_k(x)$ are not allowed, corresponding to Bragg reflections of the electron waves by the lattice; these are the so-called energy gaps (forbidden energy regions) which separate energy bands (allowed energy regions). In 3D these discrete bands of energy form the conduction and valence bands within materials. The forbidden energy gap separates these bands. An additional quantum number n is required to label the individual energy bands corresponding to successive Brillouin zones in the wavevector, $2n\pi/a$ (where $n = 0, 1\ldots$). The position of the Fermi level determines the electron population of the energy bands, and the character of the solid (see below).
3. The parabolic E–k dispersion relation is retained near $k = 0$, and at other symmetry points of the Brillouin zone, but with the substitution of an effective mass, m^* (see below) for the free electron mass, i.e., $E = \hbar^2k^2/2m^*$ for $k \ll \pi/a$. This is referred to as the effective mass approximation.

The nearly free electron model is one limit for the description of the formation of energy bands and energy gaps in a solid. In real crystals the energy bands derive from, and retain the symmetry of, the parent atomic states of the host atoms. In this limit, the model that starts out from the wave functions of the free atoms is known as the tight binding model or the LCAO (linear combination of atomic orbitals) approximation. As neutral separated atoms are brought together to form a crystal, their wave functions overlap and the Coulomb interaction between the atom cores and the electrons splits the atomic energy levels, spreading them into bands. Each state of given quantum number of the free atom is spread in the crystal into a band of energies; a given s, p, d-state in N separate atoms will produce the corresponding N-orbital s, p, d-energy band in the solid. Including the factor of 2 for spin, a full s-band would contain 2N electrons, a full p-band 6N electrons and so on. The width of the band is proportional to the strength of the overlap interaction between neighboring atoms. Each energy band will not have the same energy as any other band over any substantial range of the wavevector.

Metals, Insulators and Semiconductors

The position of the Fermi level distinguishes whether the solid is a metal, insulator or semiconductor. Generally those bands that derive from full atomic orbital states will be fully occupied with electrons, those that derive from empty atomic states lead to empty energy bands, and those from partially full atomic states lead to partially full energy bands. The highest band of those that are filled with electrons is called the valence band; the next highest band is called the conduction band (the fundamental energy gap, E_G, is the valence/conduction band separation, referred to from now on as simply the energy gap). For a metal, E_F is in the conduction band at all temperatures; for an insulator, E_F is close to the middle of the energy gap. Semiconductors are a class of crystals with small energy gaps; they are insulators at $T = 0\,K$, but are conductors at finite temperatures due to either thermal excitation of electrons across the gap or doping, n- or p-type (see below).

Effective Mass

Consider what happens to an electron when an external electric field F is applied. Suppose the electron is initially in a state k. When the field has acted on the electron for a small time dt, it has gained an energy:

$$dE = eFv\,dt = (eF/\hbar)(dE/dk)dt \qquad [13]$$

since the group velocity of the de Broglie wave (energy, $E = \hbar\omega$) is $v = \hbar^{-1}(dE/dk)$. Or simplifying we have: $dk/dt = eF/\hbar$. This gives for the resulting acceleration:

$$a = dv/dt = (1/\hbar)(d^2k/dt^2)(dk/dt)$$
$$= (eF/\hbar^2)(d^2E/dk^2) \qquad [14]$$

Comparing this result for the acceleration of a free electron of mass m under the influence of F, $a = eF/m$, it follows that the electron behaves as if it had an effective mass, m^*, given by

$$\frac{1}{m^*} = \frac{1}{\hbar^2}\frac{d^2E}{dk^2} \qquad [15]$$

Thus, the electron effective mass m^* gives a measure of the conduction band curvature near the minimum at $k = 0$. We note that the same result is achieved by differentiating twice the 'nearly free electron' E–k relation, $E = \hbar^2k^2/2m^*$.

Electrons and Holes

Consider the situation where a single electron is excited either optically or thermally from the top of the filled valence band to the bottom of the conduction band of a 1D lattice, and consider the collective behavior of the valence band when an external electric field is applied. Denoting the charge of an electron by $-e$ and the velocities of the electrons by v_i, we may write for the current associated with all electrons in the completely filled valence band in the absence of an external field:

$$I = -e\sum_i v_i = -e[v_j + \sum_{i\neq j} v_i] = 0 \qquad [16]$$

Thus, if the electron j were missing, we should have current

$$I' = -e\sum_{i\neq j} v_i = ev_j \qquad [17]$$

Applying an external field F, the rate of change of the current due to the field is:

$$dI'/dt = e(dv_j/dt) = -e^2F/m_j^* \qquad [18]$$

Since the holes are at the top of the valence band, m_j^* is negative and the right hand side of the equation is positive – i.e., a positive external field F gives a positive change of current at the top of the valence band. Thus, a band in which an electron is missing behaves as a 'positive hole' with an effective mass $m_h = -m_e$, and positive charge $+e$. Since m_e is negative near the top of the valence band, m_h is positive.

Impurities, Intrinsic and Extrinsic Conductivity

The key point about semiconductors from a device point of view is that one can control their conductivity. The basic method is by doping the host crystal with impurities either of valence greater than the host (donors giving excess electrons, called n-type) or of valence less than the host (acceptors giving excess holes, called p-type). The conductivity is referred to as intrinsic conductivity (from intrinsic carrier concentration, N_i) and extrinsic conductivity (from extrinsic carrier concentration, N_e) for undoped and doped material, respectively.

The impurities are chosen to have very low binding energies, so that the electrons (holes) are substantially ionized at room temperature to provide n-type (or p-type) conductivity. These shallow impurities can be treated using the simple Bohr theory of the hydrogen atom, substituting the electron orbiting around the proton in the hydrogen atom under the Coulomb potential e/r, by the electron in the crystal (electron effective mass m^*) orbiting round the ionized (donor/acceptor) impurity atom under the Coulomb potential $e/\varepsilon r$ in the crystal – where ε is the static dielectric constant of the crystal. As with the hydrogen atom, this gives a set of discrete energy levels:

$$E = \frac{m^*/m_o}{\varepsilon^2}\frac{13.6}{n^2}\text{eV} \qquad [19]$$

where 13.6 eV is the binding energy of the H-atom and n is the principal (Bohr) quantum number. The binding energy is then

$$E_B = \frac{m^*/m_o}{\varepsilon^2} \times 13.6\ \text{eV} \qquad [20]$$

For typical semiconductors (Ge, Si…) we have $m^*/m_o \sim 0.05$, $\varepsilon \sim 16$. Thus, E_B is in the range of a few meV (i.e., much less than E_G, and kT at room temperature). The energy level scheme for a doped semiconductor is shown schematically in real space (at room temperature) in Figure 2.

The total conductivity will be $\sigma = \sigma_i + \sigma_e$, where the intrinsic conductivity $\sigma_i = n_i e^2\tau/m^*$ and the extrinsic conductivity $\sigma_e = n_e e^2\tau/m^*$. τ is the so-called carrier scattering time. The intrinsic carrier concentration is related to temperature by $n_i = n_0 \times \exp(-E_G/2kT)$, where n_0 is a constant. In the extrinsic regime ($n_e \gg n_i$), at temperatures such that $kT \gg E_d$ (so that the impurities are all ionized), the conduction (valence) carrier concentration $n_e \rightarrow n_d$ (or n_a), the donor (or acceptor) concentration.

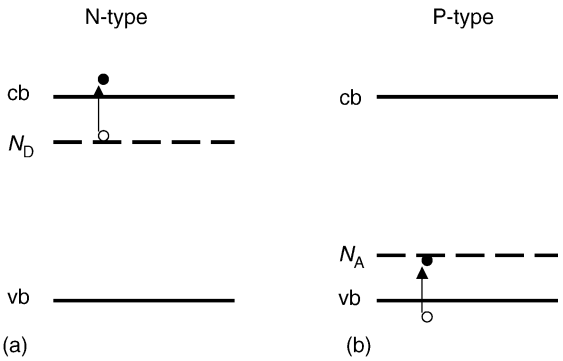


Figure 2 Doping in semiconductors. In n-type material (a), carriers are excited thermally from the donor states to the conduction band where they can contribute to the conductivity. In p-type material (b), electrons are effectively 'captured' by the acceptor states, leaving behind holes in the valence band which can also contribute to conduction. N_D and N_A are the concentration of donor and acceptor impurities, respectively.

Infrared Absorption – Interband Optical Transitions, Excitons

Apart from thermal promotion of electrons and holes via interband (valence to conduction band) transitions responsible for intrinsic conductivity, it is possible to directly excite such transitions optically by subjecting the sample to radiation with $\hbar\omega > E_G$. If the lowest conduction band minimum is situated vertically above the valence band maximum (usually at $k = 0$), then the optical transition process is referred to as a direct transition. If the lowest conduction band minimum is situated at some other point in k-space than that of the valence band maximum, the process is referred to as an indirect transition; this requires the participation of optical phonons (i.e., quantized lattice vibrations) to satisfy momentum conservation, and is therefore a weaker (higher order) process. The semiconductors are respectively referred to as Direct Gap and Indirect Gap Semiconductors. Simplified energy band diagrams for direct and indirect transitions are shown in Figure 3.

The optical spectrum corresponding to the onset of interband electric dipole transitions is referred to as the absorption edge. The most widely used indirect gap materials are the group IV semiconductors Si and Ge. The best known direct gap materials are the III–V's (InSb, GaAs, GaP, etc.), II–VI's (HgTe, CdTe, ZnSe, etc.) and IV–VI's (lead salts PbTe, PnSe, SnTe) and alloy combinations of all these. An important additional effect, which becomes progressively stronger as the energy gap becomes larger in a semiconductor is the Coulomb interaction between the electrons and holes created by interband optical transitions. The electron is raised into an excited state but remains bound in a hydrogen-like orbit

around the positive hole creating a two particle excitation called an exciton. The composite particle can then move throughout the crystal. In an ideal crystal, therefore, the absorption spectrum should consist of a series of discrete lines corresponding to the ground and excited exciton states, followed at somewhat high energies by the continuum associated with interband electronic transitions. The energy levels of the exciton spectrum measured from the beginning of the ionization continuum will be given by the Bohr formula:

$$E_{ex} = \frac{\mu}{\varepsilon^2}\frac{13.6}{n^2}\text{eV} \qquad [21]$$

where μ is the reduced mass given by: $1/\mu = 1/m_e + 1/m_h$ (m_e, m_h are electron and hole effective masses). A schematic diagram of the absorption edge is shown in Figure 4 (dashed line – in the absence of excitons; solid line – in the presence of excitons).

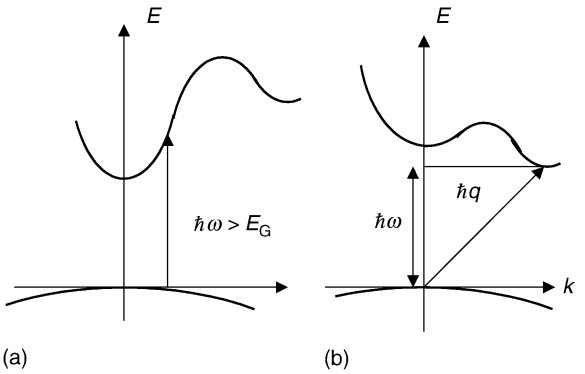


Figure 3 (a) Direct and (b) indirect gap semiconductors. Direct gap transitions involve only the absorption or emission of a single photon and are therefore very nearly 'vertical' transitions in k-space. On the other hand, indirect transitions have a 'two-stage' nature, involving the participation of both a photon and an optical phonon that carries sufficient momentum $\hbar q$ to allow the conduction band electron created in the absorption process to have a large non-zero wavevector.

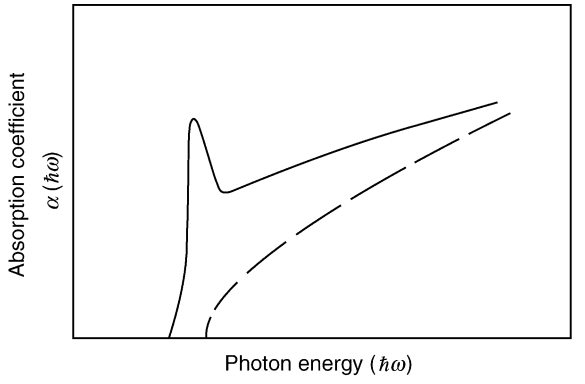


Figure 4 Absorption edge spectra without (dashed curve) and with (solid curve) the excitonic effect.

Conduction and Valence Band Parameters in Typical Semiconductors

As has been described earlier in real semiconductors the conduction and valence bands retain the symmetry of the parent atomic states from which they derive. For example, the conduction and valence bands of silicon can be regarded as deriving from 3s and 3p electrons, and it is generally true that for all group IV, III–V and II–VI semiconductors commonly used in electronic applications the conduction band has s-like symmetry and the valence band has p-like symmetry. This means that the valence band has the additional complication that it is triply degenerate (c.f. the p_x, p_y, p_z components of the hydrogen atom p-orbital); and further, in the presence of spin-orbit coupling (familiar from atomic physics) part of this degeneracy is raised, resulting in doubly degenerate (at $k = 0$) so-called heavy hole (hh) and light hole (lh) bands, and a band split-off by the spin-orbit splitting, Δ:

$$E_{hh} = -\hbar^2k^2/2m_{hh}\ ;\ E_{lh} = -\hbar^2k^2/2m_{lh}\ ;$$
$$E_{so} = -\Delta - \hbar^2k^2/2m_{so} \qquad [22]$$

The resulting conduction and valence bands near $k = 0$ are shown in Figure 5.

Low Dimensional Systems: Quantum Wells

So far we have been concerned almost entirely with the bulk properties of crystalline solids. Such properties can normally be specified in terms of coefficients that are independent of the shape and size of a particular specimen. Thus, for example the electrical

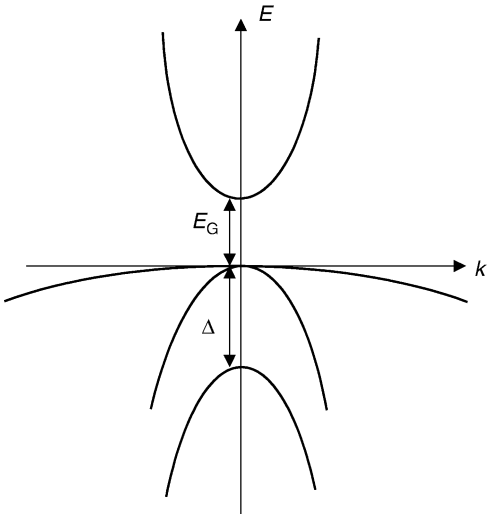


Figure 5 Energy bands (s-like conduction and p-like valence bands) of a direct gap semiconductor near the Brillouin zone center.

conductivity and resistivity, or the optical absorption coefficient and refractive index are properties only of the material and not of the sample dimensions. When one or more dimensions of a solid are reduced sufficiently, the properties are no longer given by these bulk coefficients. The sample is then described as a low-dimensional system (LDS). Such systems form the basis of many optoelectronic devices. LD systems can be classified according to the number of dimensions that are small: thin films (and heterojunctions) are two-dimensional since only the thickness is reduced, and these form so-called quantum wells; fine wires are one-dimensional since two dimensions are reduced, forming quantum wires; dots or specks are zero-dimensional, forming quantum dots.

From an optoelectronic point of view, the most important example is that of electrons confined to a semiconductor film or junction; this system is called the two-dimensional electron gas (2-DEG) and the structure a quantum well. Thus, it is possible to create a narrow electric potential junction region so that the electrons are restricted in 1D (say z) according to the quantum mechanics of a particle in an infinite potential well (described earlier), but 'free' within the xy junction plane.

See also

Semiconductor Materials: Band Structure Engineering; Quantum Dots. **Semiconductor Physics:** Band Structure and Optical Properties; Impurities and Defects; Infrared Lattice Properties; Quantum Wells and GaAs-Based Structures; Recombination Processes.

Further Reading

Hook JR and Hall HE (1991) *Solid State Physics*, 2nd edn. Chichester, UK: John Wiley and Sons, Inc.

Kittel C (1996) *Introduction to Solid State Physics*, 7th edn. Chichester, UK: John Wiley and Sons, Inc.

Seeger K (1989) *Semiconductor Physics: an Introduction*, 4th edn. Berlin: Springer.

Band Structure and Optical Properties

W Zawadzki, Polish Academy of Sciences, Warsaw, Poland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The purpose of this article is to describe the influence of the electronic energy band structure of semiconductors on their optical properties. By the band structure one understands the energy–wavevector dispersion relation $\varepsilon(\mathbf{k})$ for the electron energy bands and the relative positions of band maxima and minima in the Brillouin zone.

The main problem in calculating the band structure of a semiconductor is that of the periodic potential of the lattice in which the electrons move. It is, in fact, a self-consistent problem since the moving electrons partly screen the potential. Different approximations have been developed to deal with the question, and in all of them the symmetry of the lattice is of fundamental importance. Thus, in the so-called empty-lattice approximation one deals exclusively with the symmetry and periodicity of the lattice without specifying the potential. This can give qualitative ordering and symmetry of the bands but no quantitative results. In the opposite limit one uses the tight binding approximation, in which the bands are constructed from the atomic states of separate atoms. This method gives quite a good description of lower (valence) bands but poor results for higher (conduction) bands. A powerful way to describe real energy bands is obtained by various forms of pseudopotential calculation, where one approximates the potential near actual atoms by simple parametrized forms and then adjusts the parameters to fit experimental (mostly optical) data. The pseudopotential methods give a good overall description of various bands in the entire Brillouin zone, but they do not provide sufficiently precise results near the band extrema. A semi-empirical way to serve the latter purpose is called $\mathbf{k}\cdot\mathbf{p}$ theory which we describe below. Various methods of band structure calculation are reviewed in books on solid state physics as outlined in the suggestions for Further Reading at the end of this article.

In the majority of important semiconductor materials (Si, Ge, many III–V and II–VI compounds) the maxima of the valence bands are at the center of the Brillouin zone (Γ point). The minima of the conduction bands in Si and Ge are not at the Γ point. This means that the fundamental optical interband transitions in these materials are indirect (in the ε–$\mathbf{k}$ diagram), i.e., they require phonon assistance. On the other hand, the minima of conduction bands in important III–V and II–VI compounds are at the Γ point, so that the interband optical transitions are direct and they do not require phonon assistance.

Both systems are utilized in opto-electronic devices, particularly for detectors. However, for emitters (light-emitting diodes and lasers) we are mostly concerned with the second case, on which we concentrate here.

Bloch States

The Bloch theorem states that if the potential $V(\mathbf{r})$ in which the electron moves is periodic with the periodicity of the lattice, then the solutions $\Psi(\mathbf{r})$ of the Schrödinger wave equation

$$\left[\frac{\mathbf{p}^2}{2m_0} + V(\mathbf{r})\right]\Psi(\mathbf{r}) = \varepsilon\Psi(\mathbf{r}) \qquad [1]$$

are of the form $\Psi(\mathbf{r}) = \exp(i\mathbf{k}\cdot\mathbf{r})u_k(\mathbf{r})$, where $u_k(\mathbf{r})$ is periodic with the periodicity of the direct lattice, and $\mathbf{k}$ is the wavevector ($\hbar\mathbf{k}$ is the pseudomomentum). The proof of this theorem can be found, for example, in undergraduate textbooks on solid state physics.

k·p Theory

The underlying idea of semi-empirical $\mathbf{k}\cdot\mathbf{p}$ theory is to describe the energy bands in the vicinity of a given point of the Brillouin zone (usually near a band extremum). Symmetry properties are used to minimize the number of unknown band parameters (effective mass, energy gap, etc.) which are then determined by experiment. The initial Schrödinger equation for an electron in the periodic potential $V(\mathbf{r})$ of the crystal lattice reads

$$[\mathbf{p}^2/2m_0 + V(r) + H_{so}]\Psi = \varepsilon\Psi \qquad [2]$$

where m_0 is the free electron mass and H_{so} is the spin–orbit interaction. Since $H_{so}(\mathbf{r})$ is also periodic with the lattice periodicity, the solutions of eqn [2] are given by the Bloch states above.

We use $\mathbf{k}\cdot\mathbf{p}$ perturbation theory, as follows, to describe the energy bands near $\mathbf{k} = 0$ (i.e., the so-called Γ point of the Brillouin zone). Since we are interested in small values of $\mathbf{k}$, we expand our unknown cell periodic functions, $u_{l\mathbf{k}}(\mathbf{r})$, in terms of the set of Γ-point functions, $u_{lo}(\mathbf{r})$, whose symmetry we know. One defines a representation $\Phi_{l\mathbf{k}} = \exp(i\mathbf{k}\cdot\mathbf{r})u_{lo}(\boldsymbol{r})$, where $u_{lo}(\mathbf{r})$ is the periodic $\mathbf{k}$-independent function satisfying the equation

$$[\mathbf{p}^2/2m_0 + V + \mathrm{H_{so}}]u_{lo} = \varepsilon_{lo}u_{lo} \qquad [3]$$

whose solution we know in terms of the band-edge energies, ε_{lo}. This representation is sometimes referred to as the Luttinger and Kohn (LK) representation. The u_{lo} functions are in general mixtures of spin up and spin down states because of the spin–orbit interaction. One looks for the solutions of eqn [2] in the form

$$\Phi_{n\mathbf{k}} = \exp(i\mathbf{k}\cdot\mathbf{r})\sum_l c_l^{(n)}(\mathbf{k})u_{lo}(\mathbf{r}) \qquad [4]$$

where the sum is over all bands and $c_l^{(n)}(\mathbf{k})$ are the coefficients to be determined. Inserting the above form into eqn [2], performing the operation $\mathbf{p}^2$ (i.e., operating twice on the Bloch product function with the operator $\mathrm{p} = -i\hbar\nabla$) and using the property [3] one eliminates the unknown potential $V(\mathbf{r})$. Multiplying the resulting equation from the left by $u_{l'o}^*$, integrating over the unit cell Ω and using the orthonormality of u_{lo}, one finally obtains

$$\sum_l\left[(\varepsilon_{lo} - \varepsilon')\delta_{l'l} + \frac{\hbar}{m_0}\mathbf{k}\cdot\mathbf{p}_{l'l}\right]c_l^{(n)}(\mathbf{k}) = 0 \qquad [5]$$

for $l' = 1, 2, 3, \ldots$ Here $\delta_{l'l}$ is the Kronecker delta function, $\varepsilon' = \varepsilon - \hbar^2k^2/2m_0$, and $\mathbf{p}_{l'l}$ is the so-called matrix element of the momentum given by $\mathbf{p}_{l'l} = (1/\Omega)\int_\Omega u_{l'o}^*\mathbf{p}u_{lo}\,\mathrm{d}\mathbf{r}$. The index l' runs over all energy bands.

Equation [5] formulates the famous $\mathbf{k}\cdot\mathbf{p}$ theory. The nondiagonal part $\mathbf{k}\cdot\mathbf{p}_{l'l}$ can be eliminated by the perturbation procedure, and the method is referred to as $\mathbf{k}\cdot\mathbf{p}$ perturbation theory. In the second order of perturbation the bands are separated and in each band the carriers are described by a dispersion relation $\varepsilon_l = (\hbar^2/2)\mathbf{k}(\widehat{1/m_{lo}^*})\mathbf{k}$ where $(\widehat{1/m_{lo}^*})$ is an inverse effective mass tensor at the band edge in question. The inverse mass is generally a 3×3 tensor, but for cubic materials in the center of the Brillouin zone it is a scalar, so that $\varepsilon_l(\mathrm{k}) = \hbar^2\mathrm{k}^2/2\mathrm{m}_{lo}^*$ (where $1/\mathrm{m}_{lo}^* \equiv 1/\mathrm{m}_0 + (2/\mathrm{m}_0^2)\Sigma_{l\neq l'}(\mathrm{p}_{ll'}^2)/[\varepsilon_{l0} - \varepsilon_{l'0}]$). We then say that the band is spherical and parabolic. The second order of perturbation is a good approximation if the energy ε (as counted from a non-degenerate band edge) is small compared to forbidden energy gaps. For the triply degenerate p-like valence band one has to do degenerate perturbation theory and treat the bands together as a 3×3 matrix equation.

In the same approximation the wavefunction for a given band is: $\Psi_{l\mathbf{k}}(\mathbf{r}) = \exp(i\mathbf{k}\cdot\mathbf{r})\,u_{lo}(\mathbf{r})$. In the absence of spin–orbit coupling, H_{so}, the u_{lo} function would just have the symmetry of the parent band (i.e., labeled S for the s-like conduction band, and X, Y or Z for the triply degenerate p-like valence band – each with a single spin up or spin down component). In the presence of H_{so}, the triply degenerate valence band states become mixtures of X, Y and Z with mixed spin character and part of the degeneracy is raised.

In atomic notation the presence of H_{so} has transformed the basis from the (l,s) to the (J,m_J) representation. This results in the well-known light and heavy hole bands (degenerate at $k=0$) and the spin–orbit split-off band.

Under the influence of a radiation field, of vector potential $\mathbf{A}'$ and frequency ω, the optical transition probability for an electron to be raised from state Ψ_i (initial) to Ψ_f (final) is proportional to the square of $M=(e/mc)\langle\Psi_f|\mathbf{A}'\cdot\mathbf{p}|\Psi_i\rangle$ i.e., it is determined by the same momentum matrix element, p_{fi}, which governs the effective mass. For interband transitions one immediately gets the selection rule $\Delta k=0$ (i.e., direct transitions). The polarization selection rules are determined by the $u_{lo}(r)$ components of the initial and final states, through the momentum matrix elements.

Narrow-Gap Semiconductors

Semiconductors such as InSb and $Hg_{1-x}Cd_xTe$ (with $x<0.3$) have small energy gaps and are referred to as narrow gap semiconductors (NGSs). They are commonly used in opto-electronic applications which may be accessing energies in the conduction or valence band which are a significant fraction of the energy gap. Under these circumstances it is not valid to cut off the expansion to order k^2, and one has to deal with so-called nonparabolic bands. Thus, in NGSs (or indeed in any situation where the energy becomes an appreciable fraction of ε_g) the above procedure is not valid and an alternative approximation is in order. Following semidegenerate perturbation theory one includes exactly in eqn [5] a finite number of bands (near each other in energy) and neglects distant bands. This is referred to as the Kane model, and the energy bands near $k=0$ are shown for InSb-type III–V compounds and for HgTe-type II–VI compounds in Figure 1. In this case we include only the conduction and valence band in our set of states (eqn [5]). Including spin and degeneracy of the Γ_8 symmetry the set [5] then has eight equations. They can be solved analytically and the resulting energies are given by

$$\varepsilon'(\varepsilon'+\varepsilon'_g)(\varepsilon'+\varepsilon'_g+\Delta)-\left(\varepsilon'+\varepsilon'_g+\frac{2\Delta}{3}\right)\kappa^2k^2=0 \quad [6]$$

where $\varepsilon'=\varepsilon-\hbar^2k^2/2m_0$. We see that for this restricted set of states the only unknown parameters are the energy gap, ε_g, the spin–orbit splitting energy, Δ, and $\kappa=-i(\hbar/m_0)\langle S|p_z|Z\rangle$ (note that this is the only non-vanishing matrix element of the momentum). The fourth energy is $\varepsilon'=-\varepsilon_g$. All four energies are spin degenerate. In NGSs one may neglect the free electron term, i.e., set $\varepsilon'\approx\varepsilon$. For $\varepsilon\ll\varepsilon_g+(2/3)\Delta$ the resulting quadratic equation for the conduction and the light-hole bands is

$$\varepsilon(1+\varepsilon/\varepsilon_g)=\hbar^2k^2/2m_0^* \quad [7]$$

where $1/m_0^*=(4\kappa^2/3\hbar^2\varepsilon_g)(\varepsilon_g+2\Delta/3)/(\varepsilon_g+\Delta)$ defines the effective mass at the band edge. The root for the heavy holes is $\varepsilon=-\varepsilon_g$ i.e., this band is not correctly described within the three-level model. The bands given by eqn [6] are spherical but nonparabolic. For $\varepsilon\ll\varepsilon_g$ one recovers the standard dispersion, $\varepsilon=\hbar^2k^2/2m_0^*$.

For HgTe-type materials, setting the zero of energy at the Γ_8 edge and replacing ε_g by $-\varepsilon_0$ (cf. Figure 1), one obtains

$$\varepsilon'(\varepsilon'+\varepsilon_0)(\varepsilon'+\Delta)-\left(\varepsilon'+\frac{2\Delta}{3}\right)\kappa^2k^2=0 \quad [8]$$

For $\varepsilon'\ll(2/3)\Delta$ the dispersion [7] is valid with $1/m_0^*=4\kappa^2/3\hbar^2\varepsilon_0$. In this case the conduction and the heavy hole bands ($\varepsilon\equiv0$) are degenerate at $k=0$, i.e., the thermal gap is zero.

The important property of the above model is that it holds also in the limit of $\varepsilon_g\to0$ (i.e., for mixed $Hg_{1-x}Cd_xTe$ crystals). The effective mass in eqn [7] tends to zero, but the dispersion described by eqn [6] is valid and it gives $\varepsilon=(2/3)\kappa k$, i.e., a linear band.

The electron and hole wavefunctions in NGSs resulting from semidegenerate perturbation theory are given by eqn [4], in which the sum runs over

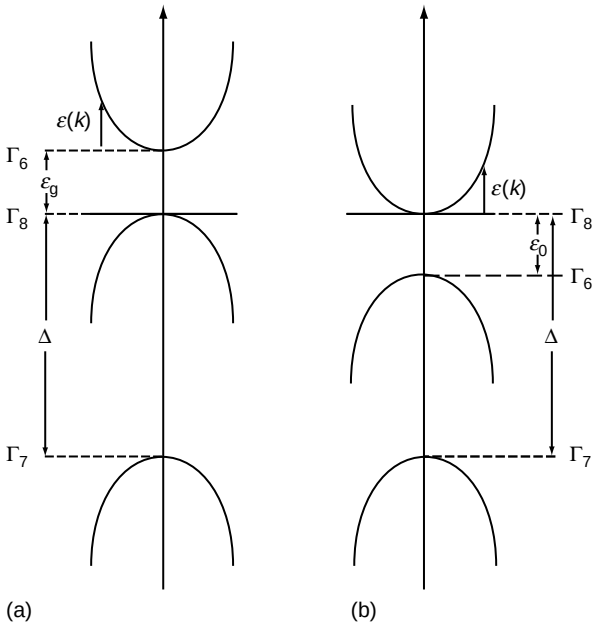


Figure 1 Three-level model of band structure near the Γ point of the Brillouin zone: (a) InSb-type semiconductors; (b) HgTe-type semiconductors.

all the included states. Thus they involve truly **k**-dependent amplitudes of the Bloch states, cf. eqn [1]. In addition, these functions are mixtures of spin-up and spin-down states. These features have important consequences for optical and dc transport phenomena.

If the conduction band minimum is not at the Γ point (Ge and Si), there are at least two different matrix elements of momentum and the resulting band is ellipsoidal.

Effective Mass

The nonparabolic dispersions $\varepsilon(k)$ in NGSs require generalizations of important relations. The momentum mass $\widehat{m}^*$ is introduced as a connection between the psuedomomentum and the velocity: $\hbar\mathbf{k} = \widehat{m}^*\mathbf{v}$. Since $v_i = \delta\varepsilon/\delta(\hbar k_i)$, one obtains for a spherical band

$$\frac{1}{m^*} = \frac{1}{\hbar^2 k}\frac{\mathrm{d}\varepsilon}{\mathrm{d}k} \qquad [9]$$

The above mass is a scalar, but it depends on the energy (or momentum). This mass is more useful than the one usually defined in textbooks, relating force to acceleration. The latter is not a scalar even for a spherical band. However, for the standard parabolic band, $\varepsilon = \hbar^2 k^2/2m_0^*$, both masses are the same and are equal to m_0^*. The mass [9] is related via velocity to the electric current, so that it enters naturally in transport phenomena (also free carrier optics). In particular, this mass enters the definition of mobility $\mu = q\tau/m^*$, where q is the charge and τ is the relaxation time. It enters the density of states (see below). Finally, it is the momentum mass [9] which is measured in cyclotron resonance.

For the dispersion [7] (the so-called two-level model), the mass [9] in the conduction band is

$$m^*(\varepsilon) = m_0^*\left(1 + \frac{2\varepsilon}{\varepsilon_\mathrm{g}}\right) \qquad [10]$$

Again, for $\varepsilon/\varepsilon_\mathrm{g} \ll 1$ the energy dependence of the mass is negligible, which represents the standard regime. Figures 2 and 3 illustrate energy dependences of the electron masses in InSb and HgSe, as measured by various optical and transport methods. The increase of the mass with energy is very strong in both cases.

Density of States

The density of states in energy space is calculated beginning with pseudomomentum space. For the spherical band one obtains

$$\rho(\varepsilon) = \frac{k^2}{\pi^2}\frac{\mathrm{d}k}{\mathrm{d}\varepsilon} = \frac{m^* k}{\pi^2\hbar^2} \qquad [11]$$

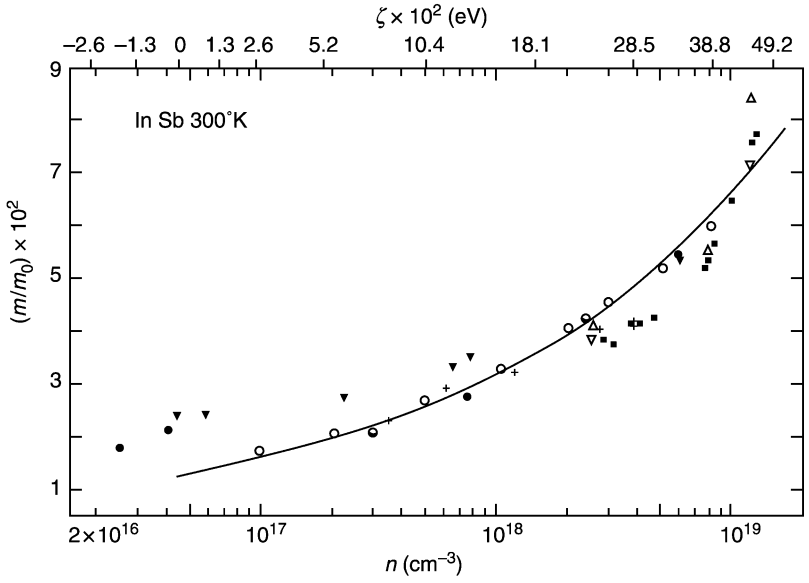


Figure 2 Electron effective mass in InSb at room temperature versus free electron concentration. The solid line, calculated for the two-level model, represents the mass value at the Fermi energy, as indicated on the upper abscissa. The experimental results of various authors are also indicated. Reproduced from Zawadzki W (1974) Electron transport phenomena in Sm all-gap semiconductors. *Advances in Physics* 23: 435–512.

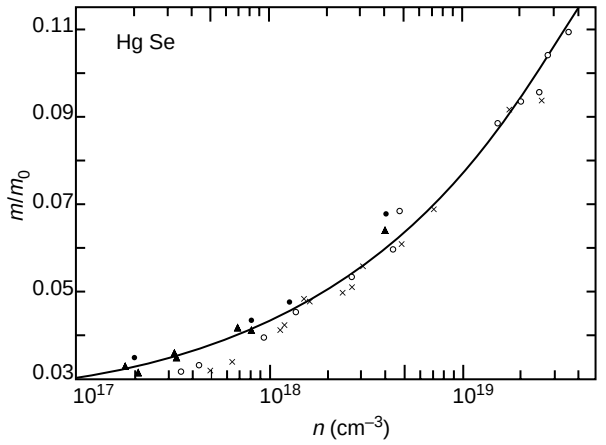


Figure 3 Electron effective mass in zero-gap material HgSe versus free electron concentration. The solid line is calculated for the three-level model. Experimental results of various authors are also indicated. Reproduced from Zawadzki W (1974) Electron transport phenomena in Sm all-gap semiconductors. *Advances in Physics* 23: 435–512.

The momentum mass [9] enters naturally into this important quantity. For the parabolic case $m^* = \text{const.}$ and $k \sim \varepsilon^{1/2}$, so that the standard result is recovered. In the general case one should use eqn [10] for $m^*(\varepsilon)$ and determine k from eqns [6] or [7].

Electron–photon Interaction

The electron–photon interaction can be introduced into the theory on two levels. If one replaces in the initial equation [2] the momentum $\mathbf{p}$ by $\mathbf{p} + (\text{e/c})\mathbf{A}'$, the interaction term is obtained in the scalar form $H_{\text{int}} = (e/m_0 c)\mathbf{A}'\cdot\mathbf{p}$, where $\mathbf{A}'$ is the vector potential of radiation. The wavefunctions to be used for the calculation of matrix elements are of the form [4], i.e., they include the periodic LK amplitudes. In fact, the matrix elements of momentum for optical transitions are identical to those of the $\mathbf{k}\cdot\mathbf{p}$ theory, as pointed out earlier. Since (as noted above) the wavefunctions for all bands have the same form [4], differing only by the coefficients $c_l^{(n)}(\mathbf{k})$, the matrix elements for interband and intraband optical transitions have the same form.

The electron–photon interaction can also be introduced directly to the $\mathbf{k}\cdot\mathbf{p}$ theory of eqn [5] replacing $\hbar\mathbf{k}$ by $\hbar\mathbf{k} + (e/c)\mathbf{A}'$. If the free electron term $\hbar^2k^2/2m_0$ is neglected, the interaction matrix involving $\mathbf{A}'$ terms is a number matrix. In this procedure the wavefunctions for initial and final states are simply columns and rows of $c_l^{(n)}$ coefficients and LK amplitudes no longer come into play. Here the $\mathbf{p}_{l'l}$ elements of $\mathbf{k}\cdot\mathbf{p}$ theory determine directly the electron–photon interaction. Both procedures described above give the same results.

Quantum Wells

If charge carriers are placed in a quantum well described by a potential $U(z)$, the motion in the z-direction is quantized while the motion in the x–y plane remains free. One takes the same LK basis (cf. eqn [3]), but the general form of solution is given by

$$\Psi = \sum_l f_l(\boldsymbol{r})u_{l0}(\boldsymbol{r}) \qquad [12]$$

since, in the presence of an external potential, $\mathbf{k}$ is not a good quantum number and the envelope functions $f_l(\mathbf{r})$ are unknown. If $U(z)$ is a slowly varying function within the unit cell the potential appears only on the diagonal of the set [5] and $\hbar\mathbf{k}$ is replaced by $\mathbf{p}$. One obtains

$$\sum_l \left[\left(\varepsilon_{l0} + \frac{\mathbf{p}^2}{2m_0} + U - \varepsilon\right)\delta_{l'l} + \frac{\mathbf{p}_{l'l}\cdot\mathbf{p}}{m_0}\right] f_l(\mathbf{r}) = 0 \qquad [13]$$

One can now eliminate the nondiagonal terms applying the standard Luttinger and Kohn scheme of second-order perturbation theory and arrive at the decoupled band equations with effective masses. This results in the standard quantization due to electric (and magnetic, if present) fields, and wavefunctions are of the form $\Psi_{lm} = f_m u_{l0}$ for the m-th sub-band of the l-th band.

For NGSs this procedure is a poor approximation and one should proceed as before including exactly a finite number of bands (cf. eqn [12]). This results in a set of coupled differential equations for the envelope functions $f_l(\mathbf{r})$. In some important cases one can find the eigenenergies without finding the functions by using the semiclassical approximation (the so-called WKB method).

We shall omit the calculations with coupled differential equations by using a semiclassical procedure also in another sense. Namely, we shall generalize the nonparabolic dispersion relation [7] to include the presence of external fields. To this end we observe that, including the potential $U(\mathbf{r})$ on the diagonal of eqn [13], one replaces $-\varepsilon$ by $-\varepsilon + U$. Further, if a magnetic field is introduced to the $\mathbf{k}\cdot\mathbf{p}$ theory one replaces $\hbar\mathbf{k}$ by $\mathbf{P} = \mathbf{p} + (e/c)\mathbf{A}$, where $\mathbf{A}$ is the vector potential of the magnetic field. Thus the semiclassical equation resulting from eqn [7] is

$$\left[\frac{\mathbf{p}^2}{2m_0^*} - (\varepsilon - U)\left(1 + \frac{\varepsilon - U}{\varepsilon_{\text{g}}}\right)\right]\Psi = 0 \qquad [14]$$

It can be seen that for $\varepsilon - U \ll \varepsilon_{\text{g}}$ one recovers the standard one-band approximation mentioned above. However, below we consider the general situation described by eqn [14].

Let us first consider the case of no magnetic field, i.e., $\mathbf{P} = \mathbf{p}$. For $U(\mathbf{r}) = U(z)$ one can separate the variables by looking for solutions in the form $\Psi = \exp(ik_x x + ik_y y)\Phi(z)$. One can now transform eqn [14] into

$$\left[\frac{1}{2m_0^*}p_z^2 - \frac{1}{\varepsilon_g}(\varepsilon - \varepsilon_\perp - U)(\varepsilon_g + \varepsilon + \varepsilon_\perp - U)\right]\Phi(z) = 0 \qquad [15]$$

where

$$\varepsilon_\perp = \frac{\varepsilon_g}{2} + \left[\left(\frac{\varepsilon_g}{2}\right)^2 + \varepsilon_g\frac{\hbar^2 k_\perp^2}{2m_0^*}\right]^{1/2} \qquad [16]$$

in which $k_\perp^2 = k_x^2 + k_y^2$. Equation (15) is suitable for the WKB semiclassical quantization with which one determines the eigenenergies ε by one integration.

For a magnetic field applied along the z-direction (parallel electric and magnetic fields) the effects of both fields are separated and the form of eqn [15] is valid, with $\hbar^2 k_\perp^2/2m_0^*$ in eqn [18] replaced by $\hbar\omega_c(n + 1/2) \pm (1/2)|g_0^*|\mu_B B$. Here $\omega_c = eB/m_0^* c$ is the cyclotron frequency, $n = 0, 1, 2, \ldots$ is the Landau quantum number, g_0^* is the spin splitting factor, and μ_B is the Bohr magneton. This results in nonequal spacing of Landau levels for a fixed value of B and in nonlinear B dependence of the Landau energies as functions of B (similarly but not identically to the bulk case). These features are illustrated by the cyclotron resonance experiment and theory shown in Figure 4.

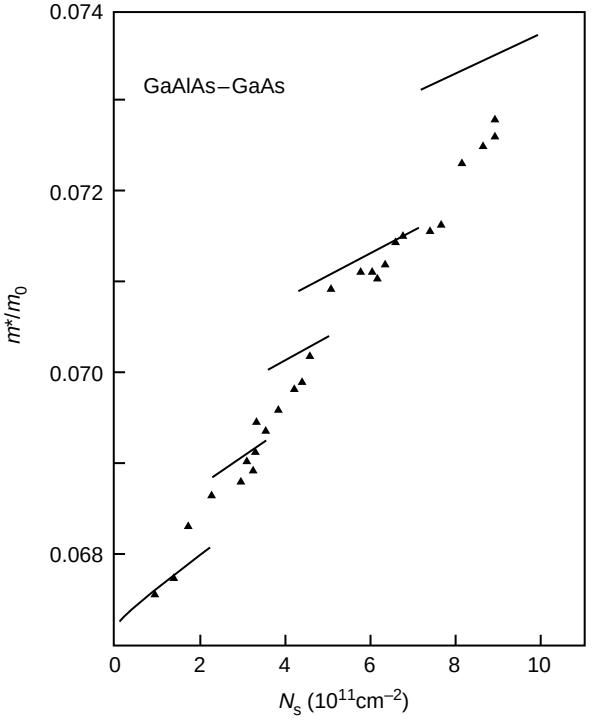


Figure 4 Electron effective mass versus two-dimensional electron density in GaAs/GaAlAs heterostructures, as measured by far infrared cyclotron resonance. The solid lines are calculated using the effective two-level model for GaAs. Reproduced from Zawadzki W, *et al.* (1994) Cyclotron emission study of electron masses in GaAs-GaAlAs heterostructures. *Semiconductor Science and Technology* 9: 320–328.

Selection Rules for Intersub-band Optical Transitions

For the one-band effective mass approximation the electron–photon interaction is $H_{int} \sim \mathbf{A}'\cdot\mathbf{p}$ (cf. above), and the wavefunctions for the initial and final sub-bands are $\Psi_m = \exp(ik_x x + ik_y y)\Phi_m(z)$, in which the $\Phi_m(z)$ must be orthogonal to each other since they describe different energies. It is then clear that the matrix element $\langle fn|\mathbf{A}'\cdot\mathbf{p}|in\rangle$ is finite only if $\mathbf{A}'$ is polarized along the z-direction (also called the growth direction) since for $\mathbf{A}'$ parallel to x or y the integral over z is $\langle\Phi_{m'}|\Phi_m\rangle = 0$ giving a vanishing matrix element. Experimentally this is a serious problem for spectroscopic applications involving intraband (i.e., intersub-band) transitions in n-type semiconductors, since it is not possible to access these using normal incidence radiation.

This selection rule is somewhat relaxed for p-type semiconductors as a result of the mixed spin states of the complex valence band. In addition the narrow-gap band structure introduces interesting new possibilities into the selection rules. The wavefunctions have the form [12] and for the light polarizations A'_x or A'_y the momentum operators p_x or p_y act also on the periodic amplitudes $u_{l0}(\mathbf{r})$, which leads to a nonvanishing transition probability for intersub-band transitions. This has indeed been observed in narrow-gap materials.

See also

Magneto-Optics: Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering. **Semiconductor Materials:** Mercury Cadmium Telluride; Quantum Dots. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Ashcroft NW and Mermin ND (1976) *Solid State Physics*. Holt, Rinehart and Winston.

Kane EO (1980) Band structure of narrow gap semiconductors. In: Zawadzki W (ed.) *Narrow Gap Semiconductors. Physics and Applications*, pp. 13–31. Berlin: Springer.

Luttinger JM and Kohn W (1955) Motion of electrons and holes in perturbed periodic fields. *Physical Review* 97: 869–883.

Zawadzki W (1974) Electron transport phenomena in Sm all-gap semiconductors. *Advances in Physics* 23: 435–512.

Zawadzki W (1984) Electric and magnetic quantisation of two-dimensional systems – elementary theory. In: Bauer G, *et al.* (eds) *Two-Dimensional Systems, Heterostructures, and Superlattices*, pp. 2–11. Berlin: Springer.

Zawadzki W (1991) Intraband and interband magneto-optical transitions in semiconductors. In: Landwehr G and Rashba EI (eds) *Landau Level Spectroscopy*, pp. 483–512. Amsterdam: North-Holland.

Excitons

I Galbraith, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

In semiconductors electron–hole pairs can lower their energy by correlating their motion in an exciton state. Such states dominate the bandedge optical properties of direct-gap semiconductors. The resultant absorption peaks are affected by quantum confinement, electric, and magnetic fields and carrier scattering.

Introduction

In the simplest, single-particle picture, the ground state of a semiconductor consists of a full valence band and an empty conduction band. The lowest lying excitation above the ground state in such a scheme is produced by excitation of one electron from the valence band to the conduction band. To achieve this optically requires a photon energy greater than the bandgap, E_g. However in many semiconductor materials this simple picture is unable to explain the observed absorption spectrum in the neighborhood of the fundamental bandgap. The origin of the discrepancy lies in the neglect of the Coulomb interaction between the negatively charged excited electron and the hole which is left in the valence band.

The term exciton refers to a two-particle excitation, consisting of a bound electron and hole. Such excitations dominate the bandedge optical spectra of most direct gap semiconductors, especially at low temperatures. In particular there exists a series of hydrogen-like bound states lying below the bandedge. These states are bound by energies

$$E_b^{3D} = \frac{\mu e^4}{8\mathrm{h}^2\varepsilon_r^2\varepsilon_0^2}\frac{1}{n^2} \qquad [1]$$

where $\mu = m_e m_b/(m_e + m_b)$, is the reduced mass, $m_{e(b)}$ the electron (hole) effective mass, e is the electronic charge, h is Planck's constant, ε_0 is the permittivity of free space, ε_r the relative permittivity of the material and $n = 1, 2, 3, \ldots\infty$ is the principal quantum number.

The binding energy of the lowest lying exciton state varies considerably from one semiconductor to another being, e.g., 0.5 meV in InSb and over 60 meV in Zn0. Similarly the spatial extent of the 1 s electron-hole relative wave function or effective Bohr radius is given by

$$a_0^{3D} = \frac{\hbar}{\sqrt{2\mu E_b^{3D}}} \qquad [2]$$

and ranges from 750 μm in InSb to ≈2 nm in Zn0. Typically, in semiconductors the exciton spatial extent is much larger than the lattice constant of around 0.6 nm, and such excitons are classed Wannier excitons. At the opposite limit, which occurs in many molecular materials, the exciton is localized around a particular atomic site and is termed a Frenkel exciton. Wannier excitons are characterized by the hydrogen-like quantum numbers of their relative motion and an overall center of mass momentum which describes the wave-like motion of the bound entity through the crystal.

Another type of exciton is often referred to as a bound exciton. This consists of an electron–hole pair bound to a neutral impurity center. Such excitons are localized around the impurity and cannot move through the crystal in the same way as a free exciton.

Optical Properties

Excitons manifest themselves primarily as strong modifications to the bandedge optical properties of semiconductors. In particular the bound states give rise to discrete absorption lines at lower energy than the bandgap energy. Several peaks corresponding to the different hydrogen-like bound states

(1s, 2s, etc.) can be seen provided the lines are sufficiently narrow. A theoretical calculation of this is shown in Figure 1.

Associated with these absorption changes are concomitant changes in the refractive index. Contributions to the spectral width of the absorption peaks are characterized as either homogeneous or inhomogeneous. Homogeneous broadening refers to broadening associated with the finite lifetime of a particular state, e.g., due to phonon scattering. Inhomogeneous broadening is due to non-uniformity in the system, e.g., in a sample having a spatially non-uniform strain field.

There is also enhancement of the absorption at photon energies above the bandgap due to the residual influence of the Coulomb attraction on the electron-hole scattering states. Although not bound, electrons and holes in these states still have an enhanced probability of being found close together. This is called the Sommerfeld Enhancement Factor (see Figure 1).

It is important to note that although there may be an excitonic peak in the absorption spectrum this does not imply that excitons form a stable population in such a sample. Often, especially at room temperature, the lifetime of an optically generated exciton is determined by the scattering time with optical phonons which may be very short (<1 ps). In this case the excitons are quickly ionized and the quasi-thermal equilibrium which is formed is dominated by essentially plasma-like behavior. In photoluminescence experiments, where carriers are generated high in the band and the spectrum of the resulting emitted luminescence is measured, peaks associated with the free exciton are often observed and taken as a signature of the presence of excitons. This view has been challenged recently by theoretical calculations which show that a photoluminescence peak at the exciton energy can be also be produced from an uncorrelated plasma if proper account of the Coulomb interaction is taken. This issue remains controversial.

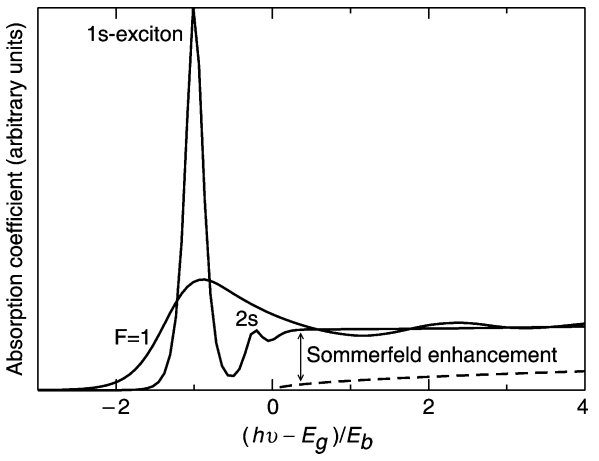


Figure 1 Calculated bulk absorption spectra including (solid-line) and neglecting (dashed line) the Coulomb interaction. The Sommerfeld enhancement is seen as photon energies above the bandgap. Also shown is the spectrum with an applied electric field of one exciton Rydberg per Bohr radius.

In bulk samples, experiments such as absorption and reflectivity are complicated by the existence of polariton effects. An exciton-polariton is a quantum mixture of the propagating photon inside the semiconductor sample and the material excitation. Where the photon dispersion and the exciton energy dispersion meet there is an anti-crossing and this is manifested in, for example, the appearance of two lines in the reflectivity spectrum.

It was hoped during the 1970s that laser action in wide bandgap material such as ZnSe would be possible based on excitonic lasing. In this process a quasi-equilibrium exciton population would form the injected excitation and stimulated recombination would occur with the associated emission of a scattering partner which would take up the necessary momentum to ensure overall momentum conservation. This leads to a light emission wavelength below the absorption edge which is advantageous for minimizing losses. In fact, this process has only been seen at low temperature for scattering with LO-phonons, electrons and other excitons. Each mechanism has its own characteristic lasing threshold and temperature dependence.

Influence of Quantum Confinement

The advent of Molecular Beam Epitaxy enabled the growth of semiconductor layers of thickness similar to the electronic de Broglie wavelength. By sandwiching a low bandgap semiconductor between two high bandgap layers, one can make quantum well structures in which the electronic motion is effectively confined to the plane of the layer. In such structures the excitonic properties are also radically altered by this confinement. An example of the absorption spectrum of a ZnCdSe quantum well embedded in ZnSe is shown in Figure 2. The heavy and light hole excitons correspond to different spin quantum numbers for the electron and hole making up the exciton.

The exciton binding energy is enhanced by a factor of 4 compared to the bulk case and the spatial extent of the 1s wavefunction is also reduced by half:

$$E_b^{2D} = \frac{\mu e^4}{8\mathrm{h}^2\varepsilon_r^2\varepsilon_0^2}\frac{1}{(n-\frac{1}{2})^2} \quad n = 1, 2, 3\ldots \qquad [3]$$

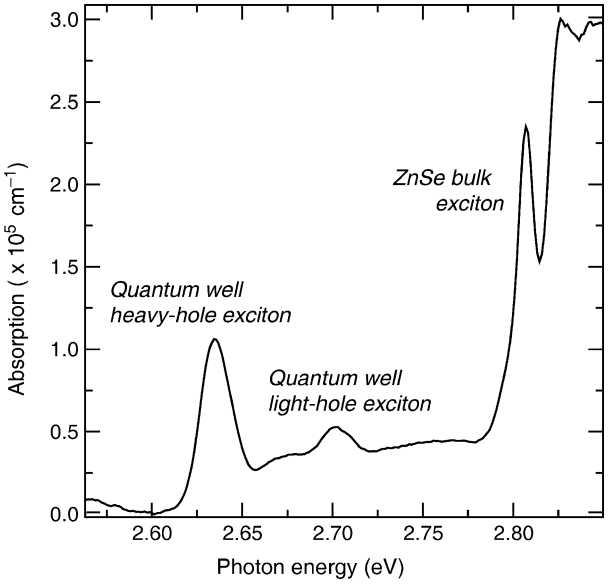


Figure 2 Heavy and light hole exciton absorption spectrum for ZnCdSe quantum wells embedded between ZnSe barriers. Also shown is the bulk exciton of the ZnSe barriers.

This enhancement can be understood since the particles are confined to lie in the plane and are hence closer together on average than in the three-dimensional case. In practice, a four-fold enhancement in the binding energy is never observed since the quantum wells are neither infinitely deep nor infinitely thin. Both these limitations lead to a reduction in the actual enhancement from 4 to about 2, depending on the structure. For example, the binding energy in a realistic quantum well corresponds to the bulk value of the well material for very wide wells and to the bulk value of the barrier material for very narrow wells. In between the value is enhanced by the quantum confinement. The maximum enhancement typically occurs when the well width is around half the bulk Bohr radius for the well material.

Accurate calculation of the exciton binding energy is complicated by the fact that the quantum confinement influences the band structure, making it strongly anisotropic and splitting heavy and light holes. A complete inclusion of these effects is at the limit of current computational power due to the need to calculate the electron-hole interaction matrix elements. Approximate schemes, including variational approaches, have proved successful in fitting a variety of samples, including GaAs and ZnSe based quantum wells. The limitation to such calculations is often the requirement for accurate values of the material parameters such as effective masses and energy band offsets.

The Sommerfeld enhancement of the absorption into the continuum states of quantum wells leads to a doubling of the absorption at the bandedge. This enhancement reduces only slowly as the photon energy is increased.

If multiple quantum wells with very thin barrier layers are used, the electronic states in the neighboring wells couple together to form a superlattice. The superlattice dispersion in the growth direction behaves like a heavy mass and the resulting exciton is three-dimensional but strongly anisotropic.

More extreme quantum confinement is possible making quantum wires, by etching or selective growth of quantum well samples. In these quasi-one dimensional structures, as in quantum wells, the exciton binding energy is again increased further. The Coulomb problem in one-dimension is pathological in that the ground-state binding energy diverges logarithmically, and higher states are pairwise degenerate. However, accounting for the finite cross-section of realistic wires eliminates this problem and recovers a finite binding energy. One consequence of this is that there is no simple enhancement factor which corresponds to the four-fold enhancement in going from bulk to two-dimensional excitons. The density of states in one-dimensional systems diverges at the bandedge, implying the existence of an infinite absorption in a perfect quantum wire. However, the influence of excitonic correlations cancels this divergence, resulting in an enhanced, but not divergent, bandedge absorption feature.

The ultimate limit in quantum confinement can be achieved in quantum dots, where the carriers are confined in all three directions. In this case the term exciton should be used with some care as there are no bandedge states to compare with. The discrete energy levels dictated by the confinement are affected by the Coulomb interaction and the energy is renormalized – in some cases substantially. The absorption spectrum should consist of a series of sharp atomic-like transitions corresponding to s-s, p-p, d-d,... transitions and indeed luminescence spectra obtained from single quantum dots do display such fine structure. Even in the best controlled material systems, such as InAs quantum dots, the spectra are strongly inhomogeneously broadened due to variations in the dot size, shape, and alloy composition.

In large quantum wells, where the width is larger than the exciton Bohr radius the center-of-mass part of the wavefunction can be confined by the barriers. This leads to a quantized motion for the exciton as a whole, rather than for the individual electron and hole separately. One implication of this is that for

such excitons the center-of-mass momentum is no longer a good quantum number.

The above effects are a controlled case of disorder induced effects in semiconductors. This is an area which is only now beginning to be understood. The presence of local potential fluctuations will influence the energy spectrum of excitons in that vicinity. Clearly both the depth and spatial extent of these fluctuations will be important – very short range (less than Bohr radius), shallow potential fluctuations will not have any influence on the exciton. While deep potentials can lead to the center of mass confinement above.

Influence of Electric and Magnetic Fields

Just as an externally applied electric field affects the absorption edge via the Franz–Keldysh Effect, broadening the edge itself and inducing oscillations in the continuum part of the spectrum, so, in the case of an excitonic bandedge, there is substantial modification due to the presence of the field. At low fields, when the Coulomb interaction is strong compared to the potential drop across the exciton diameter, the exciton will be little affected by the applied field. At higher fields the exciton will become gradually more polarized by the field until the exciton is field ionized. This ionization process is manifested in the field broadening of the exciton line (see **Figure 1**).

Of more technological importance is the influence of an electric field applied perpendicular to the plane of a quantum well. The field forces the electron and hole in opposite directions leading to a reduction in the exciton binding energy. This would lead to a blue shift in the exciton absorption peak were it not for an even larger reduction in the single particle electron and hole energy levels within the quantum wells. Overall a strong red-shift of the excitonic absorption peak is observed with increasing field. This is accompanied by a reduction in the oscillator strength due to the reduced electron-hole overlap. This is termed the Quantum Confined Stark Effect and is the basis for the operation of some semiconductor electro-optic modulators. An electric field applied in the plane of the quantum well gives rise to Franz–Keldysh oscillations analogous to the bulk case.

The influence of magnetic fields on excitons is more subtle and richer than the electric field. This is because the magnetic field seeks to induce a circular cyclotron orbit motion on the carriers, but this is influenced by their mutual Coulomb interaction. Two regimes exist where either the cyclotron energy:

$$\hbar\omega_c = \hbar(eB/\mu) \qquad [4]$$

or the exciton binding energy dominates. In the former strong field case, one can treat the Coulomb interaction as a small perturbation on the electron and hole states in the presence of the magnetic field. In exciton states, which have a finite magnetic moment at zero field, we find a linear Zeeman shift in the magneto-exciton energy. In the other case we must use the magnetic field as the perturbation which produces a mixing of the zero-field exciton states. For example, for the ground state the magnetic field induces an admixture of the p-like excited state with the s-like ground state. The total angular momentum of the mixed state is proportional to B and as the energy of a magnetic dipole in a magnetic field is also proportional to B the shift of the magneto-exciton is proportional to B^2. This distinction between linear and quadratic shifts has been used to identify the nature of carrier populations (excitons or unbound free carriers) in a variety of samples. Clearly there exists an intermediate regime where both the magnetic and Coulomb energies are comparable and in this case the precise energy shifts need to be evaluated numerically.

In quantum well samples the orientation of the magnetic field with respect to the confinement direction is crucial in determining the physics. When the field is perpendicular to the confinement plane the cyclotron orbits exist as before and essentially the same phenomenon as in the bulk emerge. For other orientations, the behavior is much more complex and beyond the scope of this article.

Exciton Scattering

An electron and hole in an excitonic bound state execute a correlated motion which can be disturbed by scattering of either partner. This may lead to the ionization of the exciton and the destruction of the correlation or, alternatively, it may change the center of mass momentum of the exciton as a whole. Almost all of the important scattering mechanisms for excitons arise from the charged nature of the electron and hole. Via the Coulomb interaction, excitons can scatter with lattice phonons, other excitons, free carriers, and impurities. Each of these scattering processes has its own regime of dominance, dependent on, for example, temperature or material quality.

When the scattering rate with other particles is much larger than the recombination rate for electrons

and holes, a quasi-equilibrium state is reached. The detailed nature of such an interacting electron/hole plasma remains a a long-standing open question. The reasons for this can be traced to interplay between the intrinsic Coulomb interactions within the plasma, which give rise to both bound and scattering states, and Pauli exclusion arising from the fermionic nature of the electrons and holes. This leads to a complex phase diagram which encompasses, e.g., electron-hole droplets, the ionized electron-hole plasma, biexciton phase, and the exciton gas which, some suggest, may undergo Bose–Einstein condensation. The parameters of this phase diagram are carrier density, temperature, and the semiconductor material parameters. Interest in this essentially fundamental question has remained high, stimulated by the stream of technological benefits that even partial answers have brought.

In this context it is worth mentioning that there have been two theoretical approaches to exciton physics which each have their advantages and problems. One approach is to treat excitons as bosonic quasi-particles and derive results from Hamiltonians formulated using bosonic operators. This approach has the appeal of simplicity and produces acceptable results in the low-density regime. It does, however, omit a key feature of the constituent particles making up an exciton namely the fermionic nature of electrons and holes. A more rigorous approach based around electron and hole operators has been followed but this is considerably more complex and numerically involved.

Excitonic Molecules

Just as two hydrogen atoms can lower their total energy by forming a bound hydrogen molecule, so two excitons can make a bound complex which has lower energy than the two isolated excitons. Such complexes are called biexcitons and are mostly important in high quality material and at low temperatures. Their binding energy is less than that of the exciton by a factor of about 0.1–0.3, depending on the ratio of electron to hole effective masses.

Three particle complexes have also been observed consisting of two electrons bound to a hole. These are termed trions or charged excitons. Their binding energy lies intermediate between excitons and biexcitons.

See also

Semiconductor Materials: GaAs Based Compounds; Large Gap II–VI Semiconductors; Type-II Quantum Wells and Superlattices. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Bastard G (1998) Wave Mechanics Applied to Semiconductor Heterostructures. New York: Halstead John Wiley.

Haug H and Koch SW (1993) *Quantum Theory of the Optical and Electronic Properties of Semiconductors*. Singapore: World Scientific.

Klingshirn CF (1997) *Semiconductor Optics*. Berlin: Springer.

Rashba EI and Sturge MD (eds) (1982) *Excitons*. Amsterdam: North Holland.

Schmitt-Rink S, Chemla DS and Miller DAB (1989) Linear and nonlinear optical properties of semiconductor quantum wells. *Advances in Physics* 38: 89–188.

Ueta M, Kanzaki H, Kobayashi K, Toyozawa Y and Hanamura E (1986) *Excitonic Processes in Solids*. Berlin: Springer.

Yu PY and Cardona M (1999) *Fundamentals of Semiconductors*. Berlin: Springer.

Impurities and Defects

K A Prior, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A simple definition of a defect is any feature that disturbs the periodicity of the perfect lattice. This definition is sufficiently broad to encompass a whole range of different entities, which are normally divided into three categories based on their dimensionality, or extent in real space.

0D defects are those which have an extent of the order of only a few atomic spacings at most in any direction. They include all cases where one, or at most a few, atoms are removed or displaced, or are replaced by atoms of a different type.

1D defects are linear features within the crystal that extend over many atomic dimensions. The most

important type of 1D defect is the dislocation and in a single crystal it is possible for a dislocation to extend from one side of the crystal to the other.

2D defects include a variety of different features that occur primarily in one plane. Any junction between two different types of crystal in an epitaxially grown layer (a heterointerface) fits this description, as of course does the sample surface. There are also crystal defects which exist in two dimensions of which the most important is the stacking fault.

Finally, there are also 3D defects. This category includes voids and inclusions, which are both major perturbations in a crystal structure. In most pure crystals grown by modern techniques these are eliminated and they will not be discussed further.

In the sections below the main types of defect will be discussed according to the scheme described above. Despite the substantial research effort to understand the basic properties of many semiconductors, and of course the huge effort devoted to producing semiconductor devices, it remains true that the defect structure of most semiconductors is still poorly understood, with the exceptions of silicon and gallium arsenide. This is especially true in the case of the nitride semiconductors with the wurtzite structure which have recently been the subject of much commercial interest. In the following sections examples will mainly be taken from gallium arsenide.

Lastly, it is important to note that defects of different dimensionality interact. Several possible interactions are discussed in the last section.

0D Defects

Dopants

There are two basic types of 0D defects to consider: chemical defects, or impurities, and native or point defects caused by missing or misplaced atoms. Table 1 gives the major types of 0D defect found in compound semiconductors together with the symbol used in the common notation for point defects and impurities.

Impurity atoms are those which are introduced either accidentally during the production of the bulk crystal or epitaxial layer, or deliberately by the crystal grower in precisely controlled amounts to modify some property of the material. Usually, this means that the material added is a dopant, and has been added to create a surplus of either free electrons or holes in the semiconductor. Typically, the amounts of dopants added are between 10^{-6} and 10^{-3} mole fraction, giving doping levels between $10^{22}\,\mathrm{m}^{-3}$ and $10^{25}\,\mathrm{m}^{-3}$.

Table 1 Impurities and point defects in a compound *AB* with the zincblende lattice structure

Symbol	*Description*
X_A^n, X_B^n	Impurity atom, *X*, on *A* or *B* site with charge *n*
X_i^n	Interstitial impurity atom with charge *n*
V_A^n, V_B^n	Vacancy on *A* or *B* site with charge *n*
A_i^n, B_i^n	Interstitial *A* or *B* atom with charge *n*
A_B, B_A	*A* atom on *B* site or *B* atom on *A* site (antisite)

In general, any atom added to a material will introduce a localized state into the band structure of the semiconductor. This state may rarely occur in either the conduction or valence band, but is more likely to be located in the bandgap. If the localized energy level is within a few kT of the valance or conduction band edge, then the state is an electron acceptor or donor. If it is much further from the band edge then it is usually known as a deep level. In this case its main effect on the electrical properties will be to act as a trap for electrons or holes, and introduce noise into the current flow. Its effect on the optical properties of the material will be to reduce the radiative recombination rate for band-to-band transitions by acting as an alternate recombination pathway for either radiative or nonradiative transitions.

It is generally assumed that dopant atoms occur only on the site of the appropriate sublattice and are well behaved. While it is true that some elements are well behaved substitutional dopants, such as sulfur in gallium arsenide, which can be denoted S_{As}^{+} when acting as a singly charged donor, the behavior of other species is more complex. In many III–V compounds, silicon is an amphoteric dopant, occupying sites on both the metal and nonmetal sublattices which ionize as shown:

$$\mathrm{Si}_{\mathrm{Ga}}^{0} = \mathrm{Si}_{\mathrm{Ga}}^{+} + \mathrm{e}^{-}$$

$$\mathrm{Si}_{\mathrm{As}}^{0} = \mathrm{Si}_{\mathrm{As}}^{-} + \mathrm{h}^{+}$$

In molecular beam epitaxy (MBE) growth on GaAs(100) the ratio between the donor $\mathrm{Si}_{\mathrm{Ga}}^{+}$ and the acceptor $\mathrm{Si}_{\mathrm{As}}^{-}$ varies as the concentration of doping is increased. At low levels silicon is a donor, while at high concentrations the fraction which is an acceptor increases rapidly. The semiconductor is therefore compensated, effectively limiting the maximum concentration of free carriers.

Another type of behavior occurs with beryllium in III–V semiconductors and lithium in II–VI semiconductors. Here, the dopant atom is so small that it can easily be displaced from the lattice site. Thus, for example, in InGaAs, it is energetically favorable for the acceptor $\mathrm{Be}_{\mathrm{In}}^{-}$ atom to move from the lattice site to

an interstitial position:

$$Be_{In}^{-} = V_{In}^{-} + Be_{i}^{+} + e^{-}$$

The interstitial Be ion is a donor and causes some compensation, but even worse, is highly mobile. It can move a substantial fraction of a micron in an epitaxial device and in so doing completely destroys a carefully engineered abrupt p-n junction. It can be seen therefore that dopant interactions with the host lattice are important in determining device performance.

Native Defects

A large variety of native defect species are possible in semiconductors. Interstitial atoms and vacancies are found in both the diamond structure (silicon) and the zincblende structure (most III–V and II–VI semiconductors). In zincblende structures there is also the possibility of placing an atom on the wrong atomic site, or sublattice; for example in GaAs a Ga atom may be found on an As site. This type of defect is known as an antisite.

Bulk semiconductor crystals are grown under conditions close to equilibrium. In the long periods of time needed to grow the crystals, equilibrium concentrations of point defects are established. Epitaxial growth techniques (in particular molecular beam epitaxy, MBE and metalorganic chemical vapor deposition, MOCVD) operate under conditions far from equilibrium and the point defect concentrations may differ markedly from the equilibrium values. Some semiconductor processing operations can also increase point defect concentrations. Etching semiconductors with ion beams or energetic plasmas is a very efficient way of creating in the subsurface region high concentrations of point defects well above the equilibrium values.

It is possible to anneal crystals in the vapor of one of the constituent elements to alter the point defect concentrations. Interstitial and vacancy concentrations can be changed relatively rapidly over the course of several hours (interstitials) or days (vacancies) as these species are much more mobile than dopant atoms. Usually, annealing is done to reduce point defect concentrations, but on occasions, for example with the alloy HgCdTe, it is done to increase the concentration of a type of defect, in this case mercury vacancies which are used to dope the material.

Like dopant atoms, point defects can also introduce defect levels into the band structure of the semiconductor, and some defects, for example the gallium vacancy V_{Ga}^{n-}, exist in multiple charge states. In some materials, the native defects act either as dopants or as compensating traps, and are present in concentrations that can severely limit the maximum obtainable doping levels. The gallium vacancy, for example, is believed to limit the maximum doping level obtainable in n-type GaAs, and its concentration increases dramatically as the concentration of n-type dopant is increased.

It should be noted that the concentrations of the various types of defect are not independent of one another, and at any given growth (or annealing) temperature and applied partial pressures of the constituent elements, the equilibrium concentrations of the defects are fixed. Examples of the interactions between the types of defect may be seen by considering the displacement of a gallium atom from its lattice site to create a vacancy and an interstitial (the Frenkel reaction):

$$Ga_{Ga}^{0} = Ga_{i}^{0} + V_{Ga}^{0}$$

and the interaction of an arsenic interstitial with a gallium vacancy to create an arsenic antisite:

$$As_{i}^{0} + V_{Ga}^{0} = As_{Ga}^{0}$$

The relative importance of the types of defect varies between semiconductors. In compounds where one of the elements is very volatile, then high concentrations of vacancies on that particular sublattice are to be expected. Antisites occur only at very low concentrations in the ionic II–VI compounds, as the energy involved in placing an atom on the wrong sublattice is very high. However, in the more covalent III–V compounds, there is not such a large energy penalty, and it has been suggested that in GaAs the As_{Ga} antisite may be the most common point defect.

In addition to the simple defect types described above, there are other possibilities. One of these is that complexes of defects may form. The best known are the complexes that occur between dopant atoms and vacancies, for example the $[Te_{As}V_{Ga}]^{-}$ complex found in Te doped GaAs (**Figure 1**). In this material, partial compensation of the n-type dopant occurs with the formation of gallium vacancies, as described above:

$$\frac{1}{2}As_2(v) = V_{Ga}^{0} + As_{As}$$

$$V_{Ga}^{0} = V_{Ga}^{2-} + 2h^{+}$$

These vacancies are mobile and diffuse in the material. In the vicinity of Te_{As}^{+} they experience a strong Coulombic attraction and the complex forms when the vacancy and dopant atom are on nearest neighbor sites:

$$V_{Ga}^{2-} + Te_{As}^{+} = [V_{Ga}Te_{As}]^{-}$$

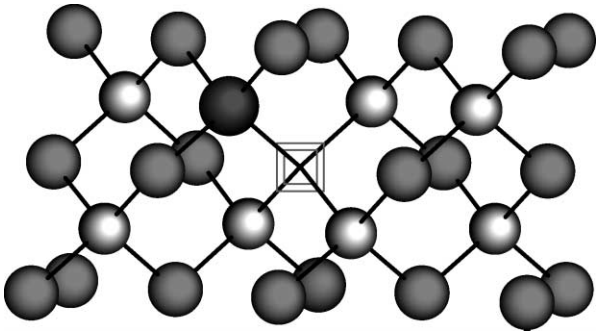


Figure 1 Diagram of the tellurium atom–vacancy complex in GaAs. Gallium and arsenic atoms are shown in mid-gray and white, respectively, while the tellurium atom is shown in dark gray.

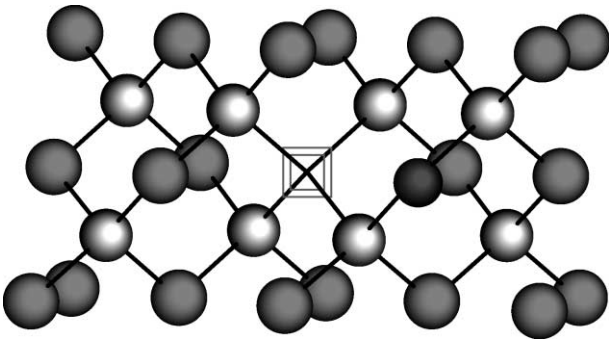


Figure 2 Diagram of the silicon atom–vacancy complex in GaAs. Gallium and arsenic atoms are shown in mid-gray and white, respectively, while the silicon atom is shown in dark gray.

A similar complex also forms in Si doped GaAs: $[Si_{Ga}V_{Ga}]^{-}$ (Figure 2), although here both species exist on the same sublattice, and consequently occur as next nearest neighbors in the complex.

In the description above it is assumed that all atoms which are located on a particular site are centered on that site. However, it is possible for a dopant atom to replace an atom of the host semiconductor, but not occupy the lattice site. *Ab initio* calculations of the bonding around the dopant atom suggest that it can be displaced away from one of its nearest neighbors towards the other three atoms (Figure 3). The dopant can be considered as having one bond weaker than the other three.

In the case of n- and p-type doped semiconductors the corresponding defects are known as DX and AX centers respectively. DX centers occur in many semiconductors but are particularly important in AlGaAs, where Si (or other n-type dopant) atoms form DX centers that partially compensate the material and reduce the electron mobility. However, irradiation with white light excites the DX center and converts it back to a normal dopant centered on the lattice site. In AlGaAs the DX center is the stable

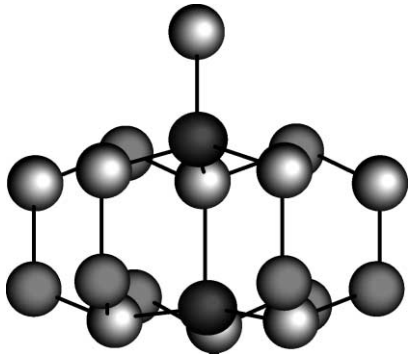


Figure 3 A GaAs crystal shown with [111] oriented vertically. Two Ga atoms have been substituted by silicon atoms which are shown in dark gray. One Si atom is on a substitutional site (top atom) and the other is displaced along [111] away from one of its nearest neighbors (bottom atom). This is one of the structures suggested for the DX center.

species and the dopant atom is a higher energy state. Conversion back to the DX center requires surmounting an energy barrier and thus the normal dopant is metastable. The irradiated semiconductor has a lower concentration of DX and a higher doping concentration than before, and consequently has a higher conductivity than the unirradiated material. This results in persistent photoconductivity in the irradiated material, which is a characteristic feature of semiconductors containing DX centers. At low temperatures, this effect can persist for many hours.

1D Defects

Within any crystal, there must exist a set of planes where the interplanar bonding is weakest. In both the diamond and zincblende crystal structures these are the {111} planes. When a crystal is subjected to a shear stress above its elastic limit, then the material accommodates that stress by changing shape, alternately breaking and reforming bonds across these planes where the bonding is weakest. The resulting movement is easily observable under a microscope in many materials and is known as slip or glide; the planes on which this happens are known as slip or glide planes.

Movement on a slip plane does not happen over the whole crystal at once, and there is always a boundary between the material in which slip has occurred and material where it has not. This boundary is the dislocation. A section through a piece of material where slip has occurred is shown in Figure 4. The curved line DD′ marks the dislocation and slip has occurred in the region marked DQD′. Because it marks a boundary, the dislocation may not end within the crystal and the only permitted ways that a dislocation may terminate are to finish at the surface of

the crystal (as shown) or to meet another dislocation, forming either a closed loop or a junction.

Across the slip plane, the material has moved by a certain amount, which is, typically, a lattice vector within the crystal. The direction of slip is known as the Burgers vector of the dislocation, and it is a constant for any one dislocation. The Burgers vector at any point may be determined by describing a simple loop in the crystal passing through adjacent lattice points lying in one plane. If a dislocation lies within the loop then the ends will not close, and the displacement needed to close the loop is the Burgers vector. This can be seen in Figure 4 on the left-hand face of the crystal around point D, where the Burgers vector is parallel to the line of the dislocation. It can also be seen on the right hand face of the crystal at point D′, where the Burgers vector, which has not changed direction, now lies perpendicular to the dislocation line. Dislocations where the Burgers vector is parallel or perpendicular to the dislocation line are known as screw and edge dislocations, respectively.

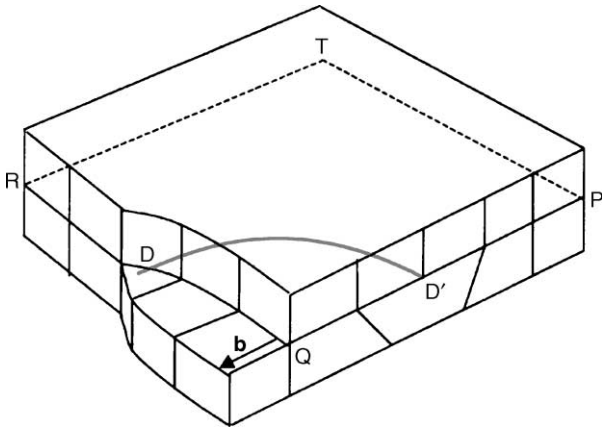


Figure 4 Diagram of a crystal showing a region where slip has occurred. For clarity only the lattice points are shown. The slip plane is PQRT and the slipped region is DQD′. The curve DD′ is the line of the dislocation and has a Burgers vector **b**.

Examples of edge and screw dislocations are shown in Figures 5 and 6, respectively. Edge and screw dislocations are considered to be 'pure' types of dislocations. If the Burgers vector makes any other angle to the dislocation line, the dislocation is referred to as a mixed dislocation. In semiconductors mixed dislocations are the most common type. There are many different ways of describing dislocations; however, the most common one is to give the angle between the dislocation line and the Burgers vector. In this system screw and edge dislocations are referred to as 0° and 90° dislocations.

At the core of the dislocation there is a small region, typically only one or two atoms wide, where the atoms have highly distorted bonding. Around this

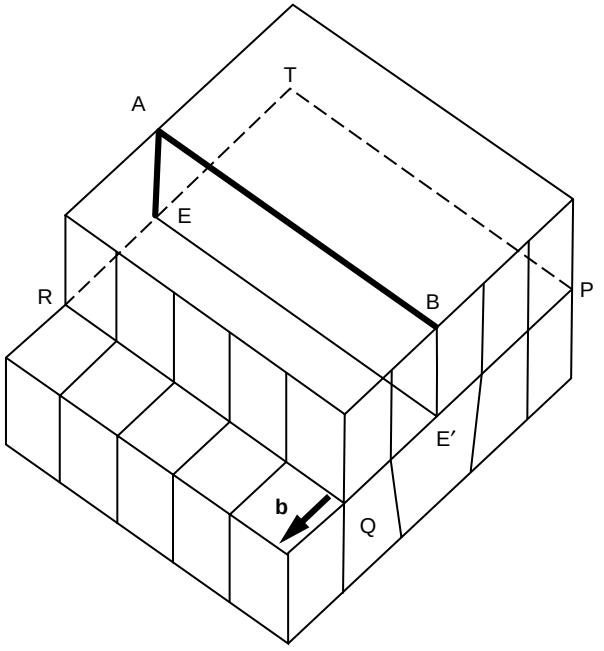


Figure 5 A schematic diagram of an edge dislocation. Only the lattice points are shown for clarity. The slip plane is PQRT and contains the Burgers vector denoted **b**. The line of the edge dislocation is EE′, which can be thought of as terminating a half-plane of atoms denoted ABEE′.

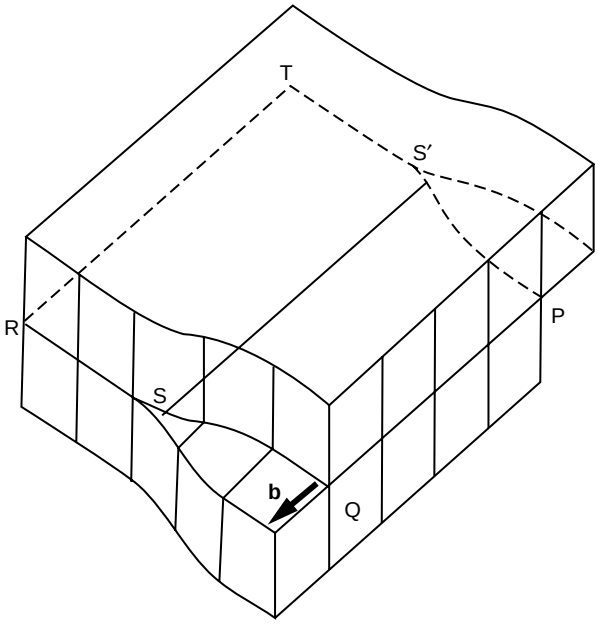


Figure 6 A schematic diagram of a screw dislocation. Only the lattice points are shown for clarity. The slip plane is PQRT and contains the Burgers vector denoted **b**. The line of the screw dislocation is SS′.

core the atoms adopt the normal (tetrahedral) bonding, although they are strained for many nanometers around the dislocation. The strain field around the dislocation means that low dislocation densities are very easy to detect using a technique such as X-ray topography, which images the strain introduced locally in a set of crystal planes by a dislocation. Using this technique, individual dislocations may be determined within crystals. Currently, it is possible to obtain silicon crystals which are essentially dislocation free, while the best GaAs crystals have dislocation densities of $<10^5$–$10^6\,m^{-2}$. In other materials the densities may be 100–1000 times higher.

The highly distorted bonding at the core of the dislocation inhibits dislocation motion. Under the action of an external stress, dislocations normally move on their glide plane to relieve the stress and deform the material plastically. This is dislocation glide. In covalent semiconductors, such movement means breaking and reforming strong bonds, and does not occur in silicon below the crystal growth temperature. Dislocations are also immobile in GaAs at room temperature, although they can be induced to move in softer, more ionic II–VI semiconductors. Dislocation densities are therefore usually considered 'frozen in' at room temperature.

In addition to macroscopic shear forces, each dislocation experiences the stresses introduced by other dislocations and point defects present in the crystal, and is either attracted or repelled by them. At high densities, therefore, the motion of dislocations is inhibited by the presence of the surrounding defects – the dislocations are said to be pinned.

Generation and movement of dislocations occurs in most semiconductors easily at their growth temperature, and is important in the relief of strain introduced in growing a material on a substrate with a different lattice constant. Edge dislocations always introduce, or remove, extra planes of atoms (Figure 5), and so an array of dislocations can add or remove atoms to accommodate the strain between two semiconductor layers of different lattice constant. These are known as mismatch dislocations and they occur frequently in epitaxial growth when strain relief has occurred.

Typically, growth in MBE or MOCVD occurs on the (001) surface. Mismatch dislocations lie on the {111} planes, which intersect this plane, and at low densities occur in crystals regularly spaced along [110] and $[1\bar{1}0]$ directions. The Burgers vectors are also of $\langle 110\rangle$ type and are typically at 60° to the dislocation line. The edge component of this vector, projected on to the (001) plane, gives the amount of strain relief to be expected.

Despite the distorted bonding that exists around the dislocation core, there are thought to be remarkably few broken bonds there. An exception to this is thought to be when the dislocation line is not straight, but kinked, the atoms at the kinks having broken bonds. As a result, nonradiative transitions involving dislocations are not strong. They are important, however, during operation of lasers and LEDs. Dislocations can be seen in electroluminescence of light-emitting devices as dark lines. These so-called dark-line defects (DLDs) are the result of nonradiative recombination occurring at dislocations. Not only do they reduce the efficiency of the device, but the energy released by the recombination of the electron and hole is deposited at the dislocation core as thermal energy. Under these conditions, dislocation movement becomes easy and DLDs propagate, eventually destroying the device.

2D Defects

In a simple description of metallic structures, the hexagonal close packed structure is described as a sequence of hexagonal sheets with the stacking pattern *ABABABAB*..., while that of the face centered cubic structure is *ABCABCA*... In a similar manner, we can describe the wurtzite crystal structure by the stacking sequence *ABABABAB*..., although the layers are no longer close packed and the stacking units comprise one metal and one nonmetal atom. The corresponding stacking sequence for the zincblende structure then becomes *ABCABCA*... These structures are shown in Figures 7a and b, respectively. It can be seen that the two atom units are oriented perpendicular to the stacking sequence, and it is possible to rotate the structure about the vertically oriented bonds between any two given planes. A rotation of 180° would place all the atoms back in registry but on the wrong site. For example, in a zincblende crystal, a rotation after the third layer in the sequence *ABCABCA*... would result in *ABCBCAB*... This means that the top part of the crystal is related to the bottom half by a rotation of 180°, and the two parts are said to be twinned. The twin plane extends throughout the crystal, and is a two-dimensional defect. Both above and below the twin plane, the stacking sequence is the usual *ABC*... found in the zincblende structure; however, at the junction there is a region denoted *BCB*. This region is three planes of the wurtzite stacking sequence, and is an example of what is called a stacking fault. In zincblende crystals, stacking faults introduce small regions of the wurtzite structure, while the reverse happens in wurtzite crystals.

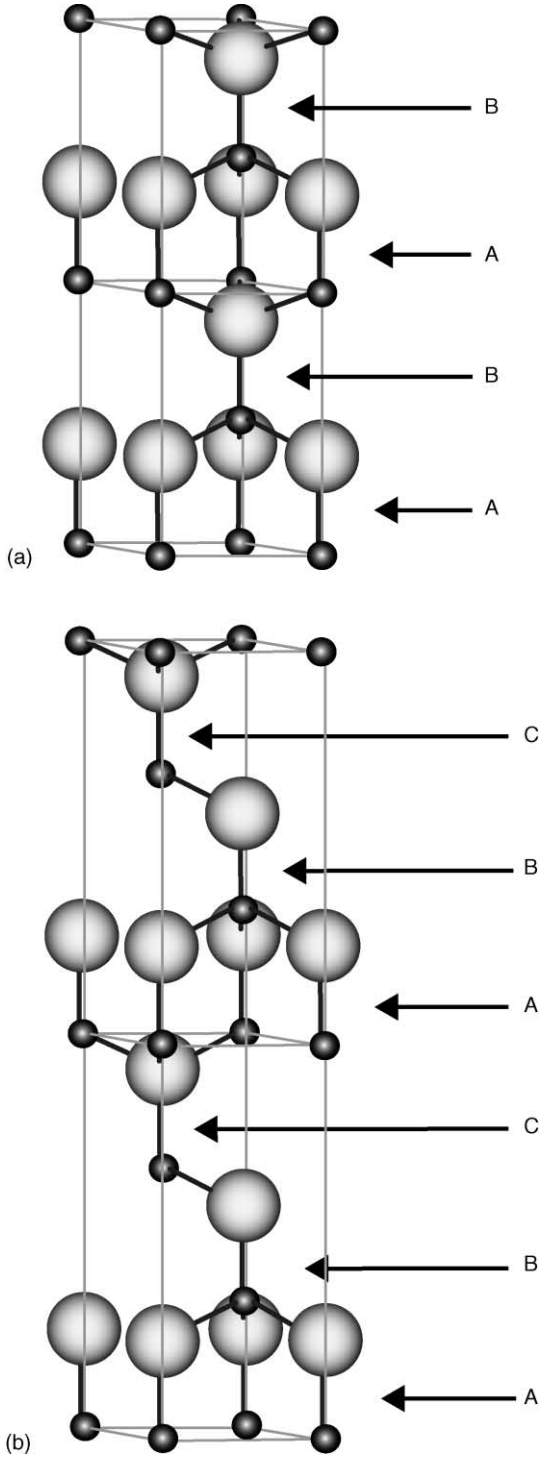


Figure 7 Diagrams of the stacking sequence of layers for two different semiconductor structures. In both cases the metal atoms are small and dark while the larger pale ones are the nonmetal atoms. The structures are: (a). The wurtzite crystal structure. The atoms which occupy the A, B and C sites are marked. Here the vertical direction in which the planes stack is the [0001] direction. (b). The zincblende crystal structure. The atoms which occupy the A and B sites are marked and the vertical direction corresponds to one of the ⟨111⟩ directions in the crystal.

The density of stacking faults in a material is primarily a function of the ease of placing the crystal planes in the wrong order. Under nonequilibrium growth conditions, this can occur easily for many different materials, but for slow growth near equilibrium, the prevalence of stacking faults is governed by the stacking fault energy. This is a measure of the wrong stacking sequence occurring and is also related to the energy of transformation between the zincblende and wurtzite crystal structures. For GaAs, the energy is high and stacking faults are rare, while for II–VI semiconductors the stacking fault energies are low and stacking faults are relatively common.

Typically, stacking faults are observed in TEM planar view micrographs, and on (001) oriented epitaxial samples are clearly visible as triangular regions lying on {111} planes, as described above, nucleated at the interface between the substrate and epitaxial layer.

While stacking faults are accidental introductions in a crystal structure, other 2D defects are introduced deliberately. Heterojunctions between different semiconductor alloys are an obvious example and are responsible for most of the useful properties of advanced semiconductor devices. However, by far the most important 2D crystal defect is the surface of the crystal. This creates a high density of broken bonds, each of which can introduce a defect level in the band structure. Surfaces usually lower their free energy by moving atoms from the bulk equilibrium positions to new sites and forming extra bonds in the plane of the surface. This removes some of the defect levels, but it is still possible to have surface defect densities as large as $10^{18}\,\mathrm{m}^{-2}$ which are associated with enhanced rates of nonradiative recombination of electrons and holes.

Interactions of Defects

Earlier it was assumed that the Burgers vector associated with a dislocation is a lattice vector. The energy of a dislocation per unit length is proportional to the magnitude of the Burgers vector squared:

$$E \propto b^2$$

In some crystals, however, it is possible for the displacement to be composed of two smaller steps, neither of which is a complete lattice vector:

$$\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2$$

where $\mathbf{b}$ is a lattice vector and $\mathbf{b}_1$ and $\mathbf{b}_2$ are the smaller displacements. In this case, the total energy

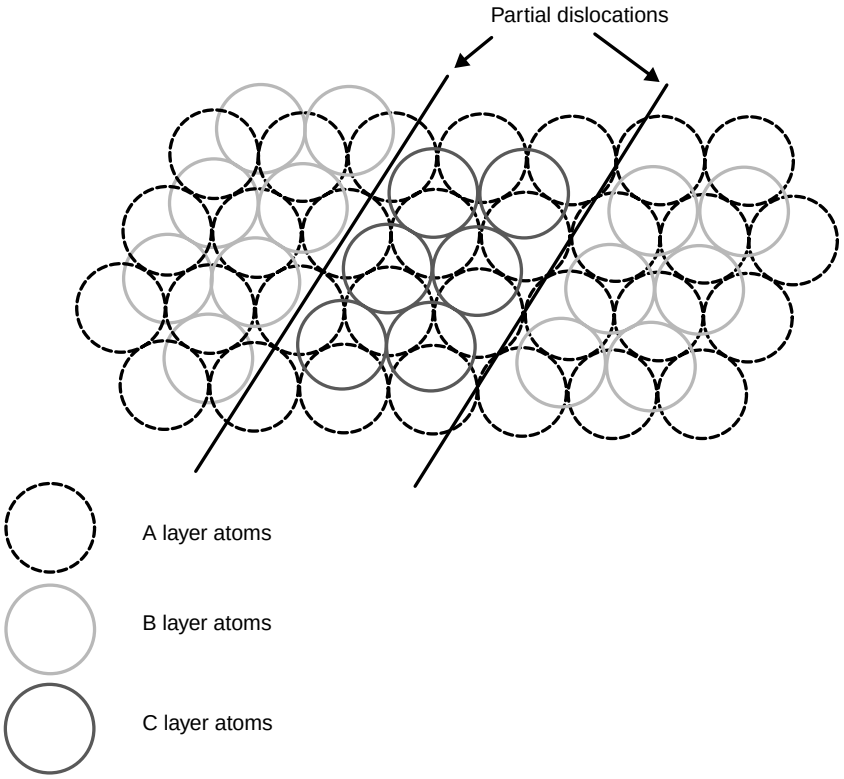


Figure 8 An example of a close packed structure where a region with a stacking fault lies between two partial dislocations.

can be reduced if the dislocation splits into two:

$$E_1 + E_2 \propto b_1^2 + b_2^2 < E$$

These dislocations are known as *partial* dislocations. A normal dislocation splits into two partial dislocations if it is energetically favorable to do so. Figure 8 shows an example in a close packed structure, which illustrates the point without the problem of introducing two different types of atoms. Two partial dislocations run through the top layer. The material lying between them has a different stacking sequence (AC) from the material around it (AB) and must contain a stacking fault. This shows that a 2D defect, such as a stacking fault, is bounded by two 1D defects, in this case partial dislocations.

There also exists an intimate relationship between point defects and dislocations, which may be seen by examining the edge dislocation shown in Figure 5. The edge dislocation may be considered as an extra half-plane of atoms, ABEE′ which terminates along line EE′. The last row of atoms in the plane can be removed from the sheet by one of two processes. Firstly, the atoms at the dislocation core can become interstitials and diffuse away from the dislocation. Alternatively, vacancies can diffuse towards the dislocation. In either case the result is the same; the dislocation line EE′ moves upwards towards AB. The opposite sequence of events (diffusion of interstitials towards the dislocation or vacancies away from it) is also possible and results in a movement of line EE′ downwards. Such processes are known as dislocation climb. Climb is not as common as glide, but it is important as the dislocation acts as a source or sink for point defects. This process contributes to the reduction in radiative efficiency in LEDs and lasers, particularly as the devices age.

See also

Semiconductor Physics: Outline of Basic Electronic Properties; Quantum Wells and GaAs-based Structures.

Further Reading

Beanland R, Kiely CJ and Pond RC (1992) Dislocations in heteroepitaxial films. In: Mahajan S (ed.) *Handbook on Semiconductors. Materials, Properties and Preparation*, vol. 3a, pp. 1149–1225. Amsterdam: North-Holland.

Fukuda M (1991) Reliability and Degradation of Semiconductor Lasers and LEDs. Boston, MA: Artech House.

Hirth JP and Lothe J (1982) Theory of Dislocations. New York: Wiley.

Hull D (1965) Introduction to Dislocations. Oxford, UK: Pergamon Press.

Hurle DTJ (1999) A comprehensive thermodynamic analysis of native point defect and dopant solubilities

in gallium arsenide. *Journal of Applied Physics* 85: 6957–7022.
Kelly A, Groves GW and Kidd P (2000) Crystallography and Crystal Defects, revised edn. Chichester, UK: Wiley.
Kossevich AM (1999) The Crystal Lattice. Berlin: Wiley-VCH.
Malloy KJ and Khachaturyan K (1993) DX and related defects in semiconductors. In: Willardson RK and Beer AC (eds) *Semiconductors and Semimetals*, pp. 235–291. New York: Academic Press.
Schubert EF (1993) Doping in III–V semiconductors. In: Ahmed H (ed.) *Cambridge Studies in Semiconductor Physics and Microelectronic Engineering*, vol. 1. Cambridge, UK: Cambridge University Press.
Skowronski M (1992) Impurity- and defect-centres in compound semiconductors. In: Mahajan S (ed.) *Handbook on Semiconductors. Materials, Properties and Preparation*, vol. 3a, pp. 1343–1398. Amsterdam: North-Holland.

Infrared Lattice Properties

T J Parker and S R P Smith, University of Essex, Colchester, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The vibrations of atoms in solids were first studied theoretically at the beginning of the twentieth century to explain the observed temperature dependence of the heat capacity. Following earlier work by Einstein and Debye, Born and von Karman showed in 1912 that the vibrational properties of a linear chain of alternate light and heavy atoms connected by springs account very well for the observed thermal behavior of solids. It was shown many years later that, if the atoms are alternately positively and negatively charged, the model also gives a good description of the optical properties.

In this article the vibrational properties of a linear diatomic chain of atoms are first described. The model gives a good description of the basic vibrational properties of solids with two atoms per unit cell, such as Si, Ge, and the binary III–V (e.g., GaAs) and II–VI (e.g. CdTe) semiconductors, and explains the origin of optic and acoustic branches of the phonon dispersion curves. This leads naturally to an understanding of the basic far-infrared reflectivity of binary semiconductors. The model is then further developed to account for the effects of impurities and the optical properties of alloys.

The far-infrared properties of semiconductors are then reviewed. We begin with the group IV semiconductors Si and Ge in which there is no first-order interaction with infrared radiation, and explain the origins of second-order interactions, and the effects of disorder and impurities. Examples of the first-order reflection spectra of bulk single crystal binary semiconductors and their alloys are then presented. Localized vibrational modes and phonon combination bands are discussed next. We finish with a brief review of more exotic excitations that occur in layered semiconductors: confined optic modes, surface phonon polaritons, guided waves, and interface modes.

Background Theory of Phonons in Semiconductor Materials

Phonons in a Diatomic Linear Chain Lattice

Consider a line of atoms of alternating type A and B, separated by a distance a, as shown in Figure 1. Suppose that the nth A atom is displaced by a small amount $u_{A,n}$ from its equilibrium position. The forces attempting to return it to its equilibrium position will be proportional to the distance it has moved relative to the neighboring atoms. If we assume that the only significant force involves the nearest neighbor B atoms n and $n+1$, the equation of motion for the nth A atom is

$$M_A \frac{d^2 u_{A,n}}{dt^2} = C(u_{B,n} + u_{B,n+1} - 2u_{A,n}) \qquad [1]$$

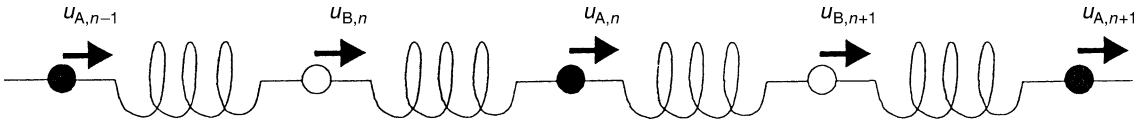


Figure 1 Atomic displacements in a diatomic linear chain.

where M_A is the mass of the atom, and C is the Hooke's law nearest-neighbor force constant. There is a similar equation

$$M_B \frac{d^2 u_{B,n}}{dt^2} = C(u_{A,n-1} + u_{A,n} - 2u_{B,n}) \qquad [2]$$

for the motion of the nth B atom. We assume that the solution of the equations of motion has the wavelike form

$$u_{A,n} = U_A \exp\{i(qan - \omega t\} \qquad u_{B,n} = U_B \exp\{i(qan - \omega t\} \qquad [3]$$

in which U_A and U_B are amplitude factors, ω is the angular frequency, and q is the wavevector, so that qa is the phase factor by which the phase of the wave changes from one site n to the next site $n+1$. The equations of motion are satisfied so long as the angular frequency ω is given by

$$\omega^2 = \frac{C}{M_A M_B}\left[M_A + M_B \pm \sqrt{(M_A + M_B)^2 - 4M_A M_B \sin^2(qa/2)}\right] \qquad [4]$$

There are two possible solutions here because there are two ions in the unit cell. A plot of ω as a function of q is shown in Figure 2a, with q in the range $-\pi/a$ to $+\pi/a$. In the lower branch, the *acoustic* phonon branch, the vibrations at long wavelength (equivalent to small q) are essentially the same as ordinary sound waves. The A and B atoms vibrate in phase, and at long wavelength there is a linear relation between ω and q:

$$\omega \approx v_s q \qquad [5]$$

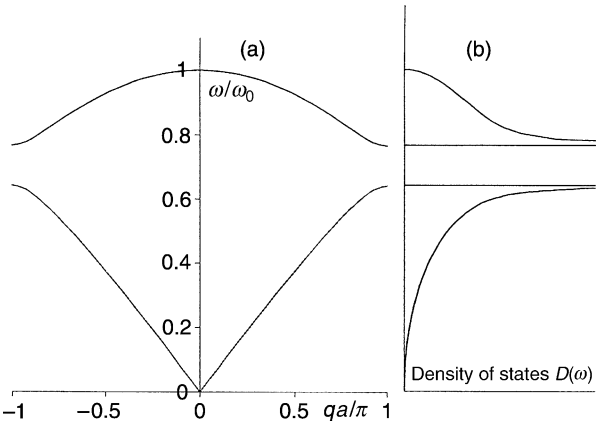


Figure 2 (a) Phonon frequency ω versus wavevector, q, in a one-dimensional diatomic lattice; (b) three-dimensional density of states $D(\omega)$.

where $v_s = \omega_0 a/2$ is the velocity of sound of the acoustic phonons.

The upper branch in Figure 2a, which is of more interest in the present context, is called the *optic* phonon branch. Here the A and B atoms vibrate in antiphase, and because the A and B atoms have equal but opposite charges there is an electric field associated with the vibration. The optic phonon vibrations will consequently affect the propagation of electromagnetic waves through the medium – hence the name – although the vibration frequencies (around 10 THz) are in the far infrared rather than the visible part of the electromagnetic spectrum. At this frequency, the infrared wavelength λ is about 30 μm. The propagation of the infrared will be affected only by phonons whose wavevector q matches the infrared wavevector $k = 2\pi/\lambda$. Since λ is around 10^4 times larger than the typical lattice spacing a in semiconductor materials, it follows that the optical properties of the material may be largely understood merely by considering the long-wavelength phonons with wavevector $q \approx 0$. Thus, in the diatomic case, only the optic phonon of wavevector $q \approx 0$ and frequency

$$\omega_0 = \sqrt{\frac{2C(M_A + M_B)}{M_A M_B}} \qquad [6]$$

is of importance to the propagation of electromagnetic radiation.

Many semiconductor materials are of the diatomic AB type – for example group III–V materials such as GaAs, and II–VI semiconductors such as CdTe. Their phonons are described in detail by three-dimensional versions of the simple linear chain model discussed above. The most basic semiconductor materials, Si and Ge, are clearly monatomic, but their crystal structure is identical to the zincblende structure of GaAs, and they show two phonon bands as in Figure 2a. However, because there is only one type of atom, these materials differ from the diatomic materials in that the optic phonon branch has no electric dipole moment. In Si and Ge, there is no dielectric anomaly around the frequency ω_0 of the type discussed in the following section. In these materials, the only significant optic effects occur when phonons combine in pairs with equal but opposite wavevector q, so that the resultant has very small wavevector that can match the infrared wavevector $k = 2\pi/\lambda$. The significant factor here is the density of states $D(\omega)$, representing the number of phonon states per unit frequency range. In Figure 2b, we show $D(\omega)$ for the diatomic lattice. This is

obtained as

$$D(\omega) = D(q)\frac{\mathrm{d}q}{\mathrm{d}\omega} \qquad [7]$$

where $D(q)$ is the density of states with respect to the wavevector, which is proportional to q^2 for a three-dimensional lattice. Note that $D(\omega)$ has peaks at frequencies corresponding to the zone boundary phonon frequencies because the dispersion curves are flat in these regions. These features are known as critical points, and they have well-defined shapes. Their spectroscopic importance is discussed later.

In bulk three-dimensional crystals, both transverse and longitudinal optic and acoustic phonons occur, each represented by different branches of the phonon dispersion curves, and the transverse branches in cubic materials are doubly degenerate. In three dimensions, more elaborate theoretical models are required, such as the rigid ion model and the shell model, and improved results can sometimes be achieved by including next-nearest-neighbor atoms in the model. As an example, the phonon dispersion curves of GaAs are shown in Figure 3.

It should be noted that the dispersion curves, and hence the associated optical properties, differ for propagation in different crystallographic directions, e.g., [100], [110], and [111].

Born–Huang Optic Phonon Model

The infrared lattice properties of materials at long wavelength are usually described by using the relative permittivity ε (often called the 'dielectric constant'), which is defined through the relation

$$D = \varepsilon\varepsilon_0 E \qquad [8]$$

between the electric field E in a medium and the displacement field D (here, ε_0 is the permittivity of free space). D is the sum of the field E and the polarization P of the medium:

$$D = P + \varepsilon_0 E \qquad [9]$$

P has contributions from lattice displacements (phonons), free electrons (plasmons), and the basic atomic polarizabilities. Since these quantities depend on frequency, ε is also a function $\varepsilon(\omega)$ of angular frequency ω. The charges of the A and B sites are taken as $+Q$ and $-Q$ respectively, and their equations of motion [1] and [2] are extended to include the electric forces $\pm EQ$. As a result of these considerations, the dielectric function can be written in the form

$$\varepsilon(\omega) = \varepsilon_\infty\left[1 + \frac{\omega_\mathrm{L}^2 - \omega_\mathrm{T}^2}{\omega_\mathrm{T}^2 - \omega^2}\right] = \varepsilon_\infty\frac{\omega_\mathrm{L}^2 - \omega^2}{\omega_\mathrm{T}^2 - \omega^2} \qquad [10]$$

Here, ε_∞ is the high-frequency dielectric constant due to the atomic polarizabilities, and ω_L and ω_T are frequencies that are close to the natural vibrational frequency ω_0 that occurs in the absence of electric field effects; the difference $(\omega_\mathrm{L}^2 - \omega_\mathrm{T}^2)$ is of order $\omega_0^2 Q^2/(Ca^3\varepsilon_0)$.

Figure 4 shows a plot of $\varepsilon(\omega)$ as a function of ω. Note that $\varepsilon(\omega)$ diverges at $\omega = \omega_\mathrm{T}$, and that $\omega = 0$ at $\omega = \omega_\mathrm{L}$ ($\omega_\mathrm{L}/\omega_\mathrm{T} = 1.25$ in Figure 4). ω_T is called the transverse optic (TO) phonon frequency: it is the resonance frequency of propagation of transverse

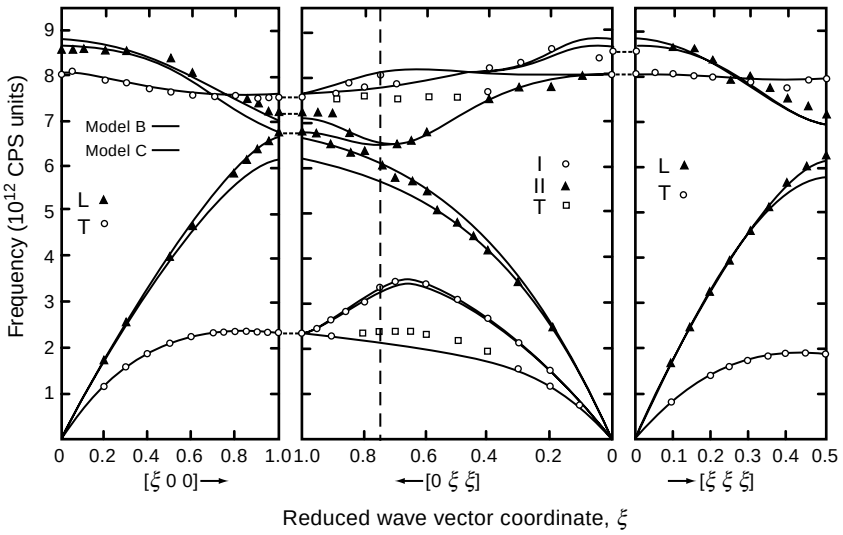


Figure 3 Phonon dispersion curves of GaAs determined by neutron spectrometry (points) compared with calculation (dotted and solid lines). Reproduced with permission from Waugh JLT and Dolling G (1963) Crystal dynamics of gallium arsenide. *Physical Review* 132: 2410.

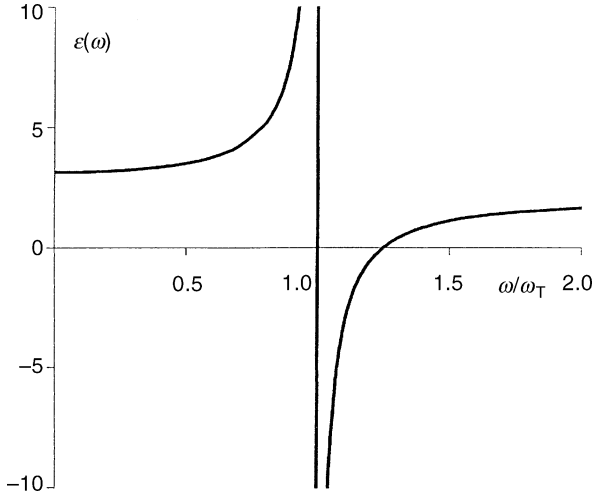


Figure 4 Dielectric function $\varepsilon(\omega)$ as a function of angular frequency, ω.

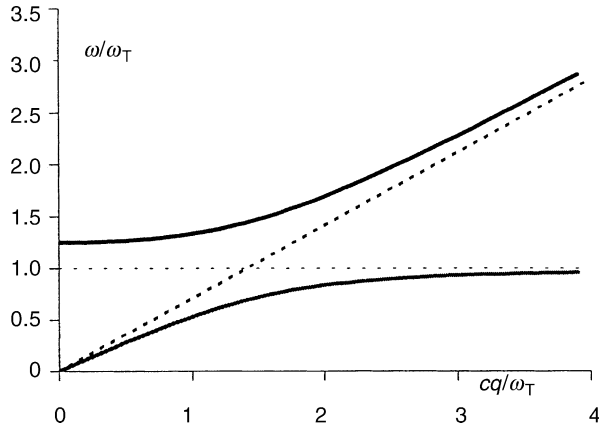


Figure 5 The dispersion curves for phonon polariton modes (solid lines) for a diatomic linear chain of atoms showing the gap in bulk phonon modes between the frequencies ω_T and ω_L.

electromagnetic waves in the material. ω_L is called the longitudinal optic (LO) phonon frequency: it is the frequency of the longitudinal electrostatic wave. Physically, the dielectric constant determines the velocity v of propagation of electromagnetic waves in the material

$$v^2 = \frac{\omega^2}{q^2} = \frac{c^2}{\varepsilon(\omega)} \qquad [11]$$

where c is the velocity of light. Equation [11] has two solutions for ω as a function of wavevector: the two solutions are shown in Figure 5. Note that there is no solution with ω in the range between ω_T and ω_L. The excitations in the region where the graphs are curved are called 'polaritons' – a combination of polar phonon and photon.

The refractive index n of the material is the ratio of v, the velocity of light in the medium, to c, the

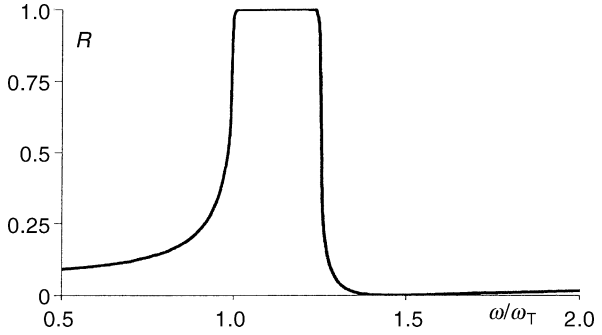


Figure 6 Calculated reflectivity plotted versus angular frequency for a diatomic linear chain of atoms for the case of zero damping.

velocity in free space; thus $n = \sqrt{\varepsilon}$. n is often found from a measurement of the optical reflectivity R. At normal incidence

$$R = \left|\frac{1-n}{1+n}\right|^2 \qquad [12]$$

A graph of R is shown in Figure 6; note that the reflectivity is 100% for ω in the range $\omega_T < \omega < \omega_L$.

Impurity Phonon Modes

Impurity atoms in crystal lattices can significantly affect the infrared properties. The simplest case is when an impurity atom has the same binding but a different mass. Consider the monatomic, one-dimensional case, in which the impurity atom has a mass $M + \delta M$, where M is the normal mass. Let the impurity be at the site $n = 0$. The equations of motion are

$$-M\omega^2 u_n = C(u_{n-1} + u_{n+1} - 2u_n) \quad \text{for } n \neq 0$$

$$-(M+\delta M)\omega^2 u_0 = C(u_{-1} + u_1 - 2u_0) \quad \text{for } n = 0 \qquad [13]$$

These equations have a solution provided that ω satisfies the equation

$$N^{-1}\frac{\delta M}{M}\sum_q \frac{\omega^2}{\omega_q^2 - \omega^2} = 1 \qquad [14]$$

where N is the number of atoms in the linear chain, and ω_q the unperturbed normal mode frequencies. This equation has solutions ω that lie close to the bulk phonon frequencies ω_q. However, if $\delta M < 0$ (i.e. the impurity atom is lighter than the bulk atoms), one additional mode is possible at a frequency ω that is higher than the maximum frequency $\omega_0 = \sqrt{C/M}$ of the bulk phonons. This is the so-called impurity or localized vibrational mode (LVM – see Figure 7).

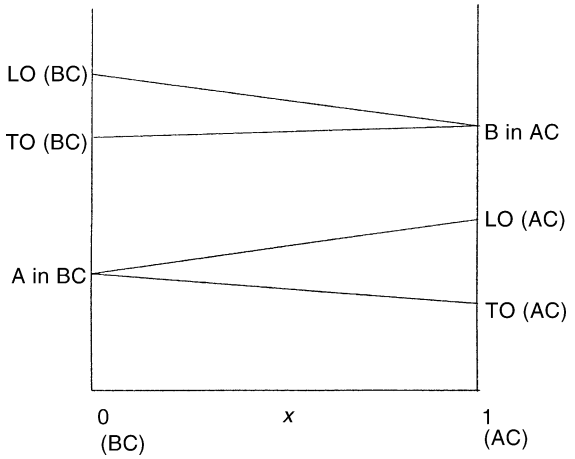


Figure 7 Representation of TO and LO phonon frequencies in a mixed crystal $A_xB_{1-x}C$ plotted versus composition x.

The impurity atom, being lighter than the material atoms, vibrates at a higher frequency than the bulk frequencies; the amplitude of the vibration is greatest at the impurity site, and decreases the further away one goes from the impurity site. If the impurity atom is heavier than the bulk atoms ($\delta M > 0$), there is no distinguishable local mode. In the case of diatomic lattices, an analogous situation pertains: if the impurity is lighter than the bulk atoms, an impurity mode is distinguishable above the top of the optic phonon band; if the impurity is heavier, there is the possibility that a distinct impurity mode may appear in the gap between the acoustic and optic phonon branches.

Mixed Crystals

Mixed compound semiconductor crystals are in widespread use because the bandgap can be adjusted by varying the composition. For example, the ternary material $A_xB_{1-x}C$ is an alloy of a fraction x of the material AC with a fraction $1 - x$ of the material BC. A typical example is $Al_xGa_{1-x}As$. The infrared lattice properties of these materials are classified as 'one-mode' or 'two-mode' depending on how many distinct optic phonon frequencies are observed. The distinction can easily be understood in the context of the discussion above of impurity modes. If an A atom in the bulk material BC has a distinct local impurity optic phonon mode frequency, and similarly for a B impurity in AC, then there will usually be two distinct optic phonon modes throughout the whole composition range $0 < x < 1$: one mode is AC-like, and the other BC-like, as shown in Figure 7. On the other hand, if neither A in BC, nor B in AC, provide distinct optic phonons, there is only one distinguishable optic phonon in the alloy $A_xB_{1-x}C$.

Review of Far-Infrared Properties of Semiconductors

Group IV Semiconductors

We begin with Si and Ge, both of which crystallize in the diamond structure. As mentioned earlier, there is no first-order interaction between far infrared radiation and the TO mode. However, there are several other mechanisms that result in absorption. The dipole moment of the zone center TO phonon in an ionic solid is an example of a symmetry property that, in group theory, is described by a set of selection rules for each different crystal lattice. Other mechanisms that can lead to the creation of a dipole moment are:

Disorder

Disorder destroys the translational symmetry, and hence the selection rules. Consequently, phonon modes that, in a perfect lattice, are not infrared active will become active. The most general case occurs if the crystal becomes completely disordered, in which case all phonon modes become dipole-active and the far-infrared absorption spectrum resembles the phonon density of states. This is illustrated in Figure 8, in

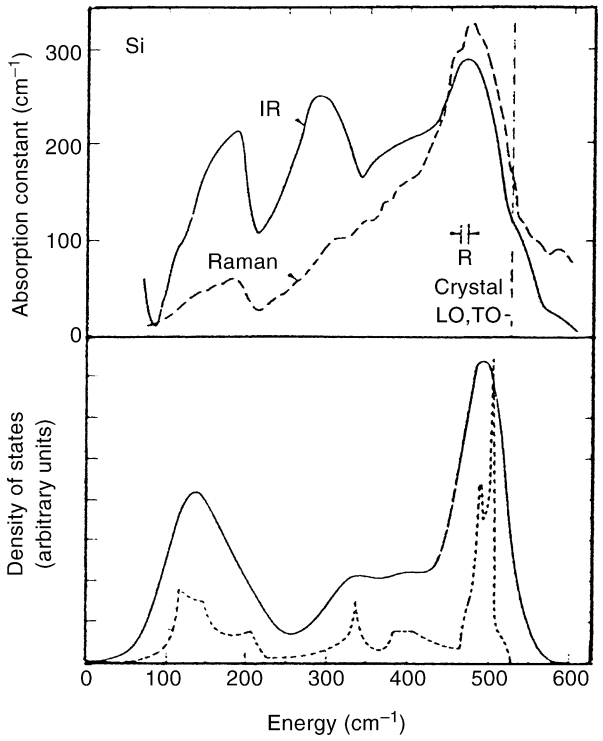


Figure 8 Top: infrared absorption coefficient (solid line) and Raman spectrum (dashed line) of amorphous Si. Bottom: calculated density of states (solid line), and broadened density of states (dashed line). Reproduced with permission from Bradskyn H and Lunio (1974) Infrared vibrational spectra of amorphous Si and Ge. *Physical Review B* 9: 1646, ADS.

which measured infrared and Raman spectra of amorphous Si are compared with the calculated one-phonon density of states.

Phonon combination bands

A combination of the lattice distortions due to two phonons excited simultaneously can produce a resultant displacement of the ions in the lattice with a dipole moment. Such phonon combinations can be excited simultaneously by a far-infrared photon, and they are observed as weak bands in the infrared absorption spectrum. Apart from the dipole moment selection rule, it is also necessary for wavevector conservation to occur. Since $q \sim 0$ for the photon, the two phonons must either be at the zone center, or have equal and opposite wavevectors. An additional consideration is that, if the process is to lead to observable absorption, it should have a high probability as determined by the density of states, so observable absorption features usually arise from critical-point phonon combinations. An example of phonon combination bands in Si is shown in Figure 9. Assignment of the features in the spectrum is usually possible by inspection of the phonon dispersion curves at the zone boundaries, a procedure known as critical-point analysis. Alternatively, the two-phonon density of states can be calculated, taking into account the selection rules, and then compared with the measured spectrum. Similar considerations apply to higher-order phonon combination bands, e.g., three or four phonons, etc., but the strength of the combination bands usually decreases rapidly as the order of interaction increases.

The presence of impurities

As mentioned above, dipole-active localized vibrational modes (LVMs) can occur if the sample is doped with lighter elements as this destroys the local translational symmetry. For instance, an LVM occurs if Si is doped with B, as shown in Figure 10. The LVM frequency is proportional to the ratio $(C/M)^{1/2}$ as seen earlier where C is the force constant and M is the isotopic mass of the substituted atom. As such modes have narrow linewidths, the features in the spectrum due to different isotopes of the same element are easily distinguished, as shown in Figure 10.

Group III–V and II–VI Semiconductors: Bulk Crystals

The group III–V and group II–VI semiconductors crystallize in either the cubic (zincblende) structure, or the hexagonal (wurtzite) structure. Many mixed crystal materials, such as $Al_xGa_{1-x}As$, are grown for

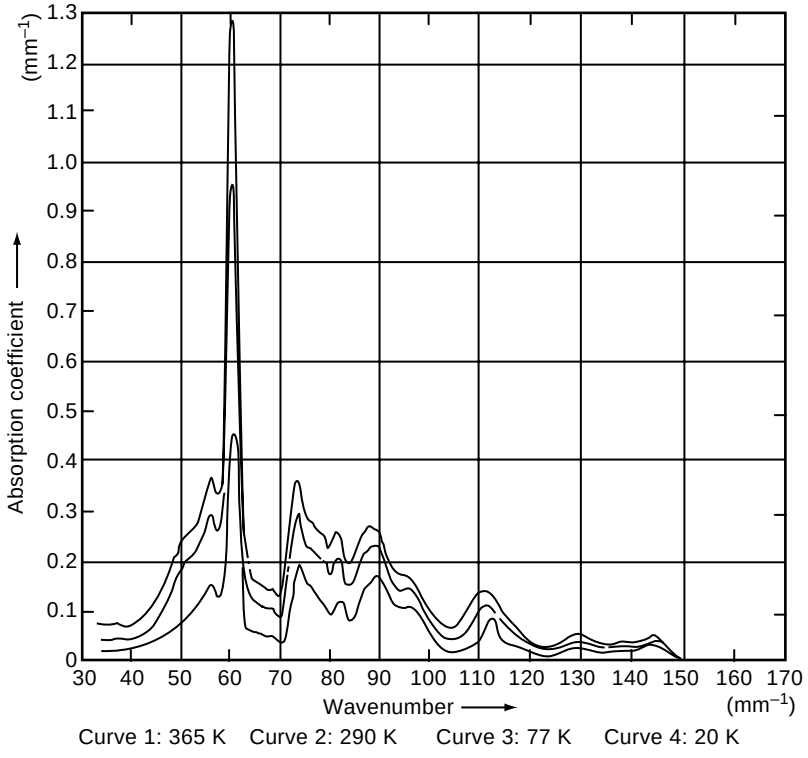


Figure 9 Measured lattice combination bands in a single crystal of silicon at 290 K (top), 77 K (middle), and 20 K (bottom). (Note that 100 mm^{-1} = 1000 cm^{-1}.) Reproduced with permission from Johnson FA (1959) Lattice absorption bands in silicon. *Proceedings of the Physical Society* 73: 265.

electronic applications that require materials with bandgaps not found naturally. Furthermore, a host of low-dimensional structures, such as epilayers, multiple quantum wells (MQWs) and superlattices, are fabricated by molecular beam epitaxy (MBE) or metalorganic vapor phase epitaxy (MOVPE) to create devices that exploit the novel properties which occur when one or more of the physical dimensions are small.

Reflectivity from a bulk single crystal

In the reststrahl frequency range bulk phonon modes cannot propagate and incident radiation is perfectly reflected, i.e., $R = 1$, as shown in the calculated spectrum in Figure 6. Figures 4 and 6 represent the ideal case of a harmonic crystal, in which there is no interaction between the phonon modes, so that all modes have infinite lifetime or, equivalently, spectral features have zero linewidth, corresponding to zero damping. This is close to the behavior of a crystal at absolute zero but, in practice, the phonon modes are coupled, so each mode will decay into combinations of other phonon modes. This leads to finite phonon lifetimes and finite linewidths of all features in the measured spectra. This damping has the effect of

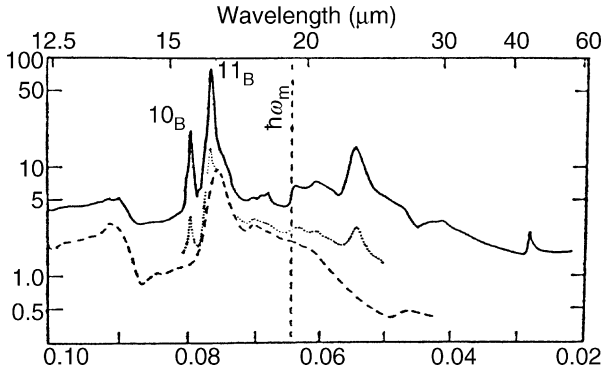


Figure 10 The far-infrared absorption spectrum of Si doped with B ($^{10}B + {}^{11}B$). Reproduced with permission from Angress JF, Goodwin AR and Smith SD (1965) A study of the vibrations of boron and phosphorus in silicon by infrared-red absorption. *Proceedings of the Royal Society A* 287: 64.

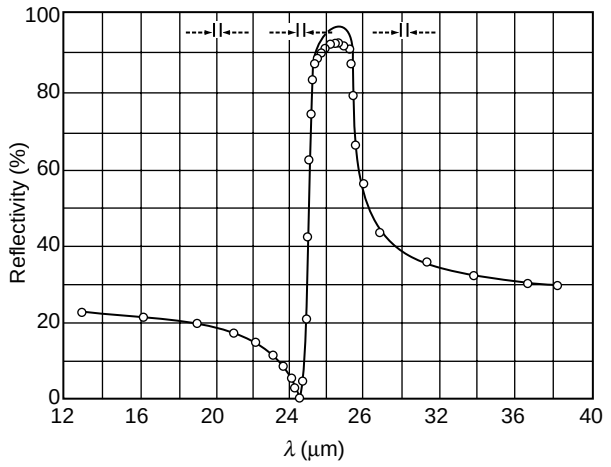


Figure 11 The reststrahlen band reflectivity of GaP, measured (points) and calculated from a damped simple harmonic oscillator (line). Reproduced with permission from Kleinman DA and Spitzer WG (1960) Infrared lattice absorption of GaP. *Physical Review* 118: 110.

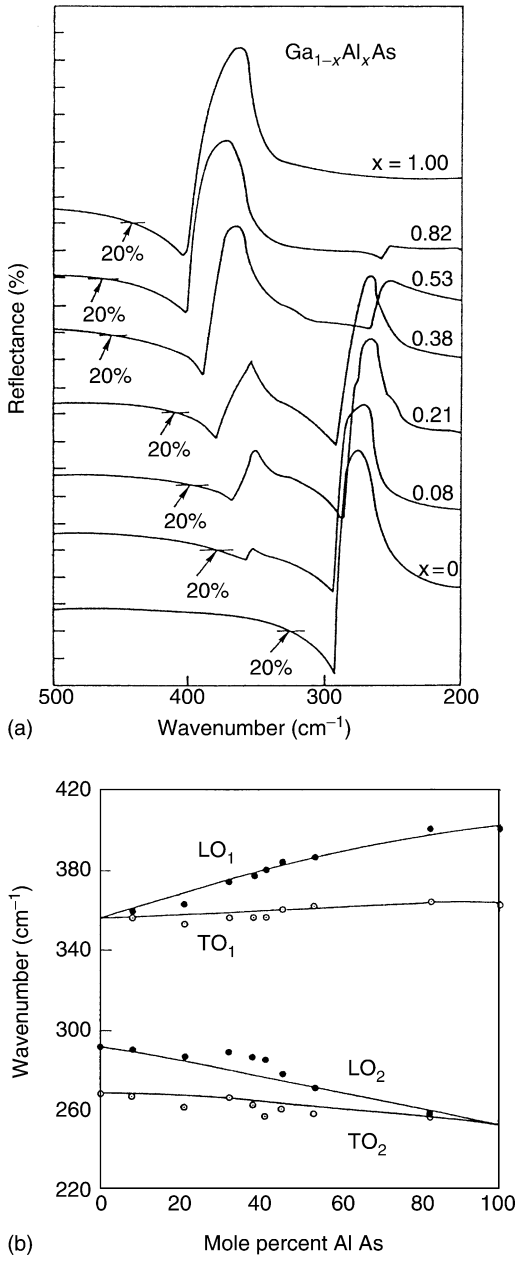


Figure 12 (a) Reflection spectra for GaAs, AlAs, and five $Ga_{1-x}Al_xAs$ mixed crystals. The curves have been displaced for clarity. (b) TO and LO mode frequencies obtained analytically from the spectra, plotted versus alloy composition. Reproduced with permission from Ilegens M and Pearson GL (1970) Infrared reflection spectra of $Ga_{1-x}AL_xAs$ mixed crystals. *Physical Review B* 1: 1576.

rounding off features in the reflectivity compared with Figure 6, and it can be represented analytically by the equation:

$$\varepsilon(\omega) = \varepsilon_\infty + \frac{(\varepsilon(0) - \varepsilon_\infty)\omega_T^2}{\omega_T^2 - \omega^2 - i\gamma\omega} \qquad [15]$$

This differs from eqn [10] only by the addition of the damping coefficient, γ, which takes account in a phenomenological way of the interactions between the phonon modes. Here $\varepsilon(0)$ is the dielectric function at zero frequency, and $\varepsilon(0) - \varepsilon_\infty$ is the oscillator strength (i.e., the dipole strength) of the mode. We shall consider γ again later.

The reflectivity of GaP is shown in Figure 11, and this should be compared with Figure 6. In pure diatomic crystals like the simple ionic solids, e.g., NaCl and KCl, or the binary semiconductors, e.g., GaAs and CdTe, this is the only region of high reflectivity at frequencies below the visible or ultraviolet, and it is often described as the reststrahlen band, from the German term for residual ray.

Reflectivity from mixed crystals of binary semiconductors

The spectra of mixed crystals were discussed above and an example is shown in Figure 12a, which shows the reflection spectra of GaAs, AlAs and five $Ga_{1-x}Al_xAs$ mixed crystals. The TO and LO phonon frequencies were obtained from the spectra using a Kramers–Kronig analysis and classical dispersion theory. Changes in the mode frequencies with composition are interpreted using the so-called random element isodisplacement (REI) model. In Figure 12b, the TO and LO frequencies are plotted versus composition, x, and it can be seen that this system displays typical two-mode behavior, following closely the behavior shown in Figure 7.

Localized vibrational modes due to impurities

LVMs have been widely investigated in binary semiconductors because they offer a useful diagnostic tool for the investigation of impurities. In addition to the identification of isotopic species, high-resolution transmission spectroscopy can be used to determine the sites of the impurity atoms in the lattice. In Figure 13 the LVM absorption band for ^{12}C in GaAs is shown at different resolutions. The ^{12}C atoms are substitutional impurities on the As sites and the different components of the band arise from the different nearest-neighbor isotopic arrangements

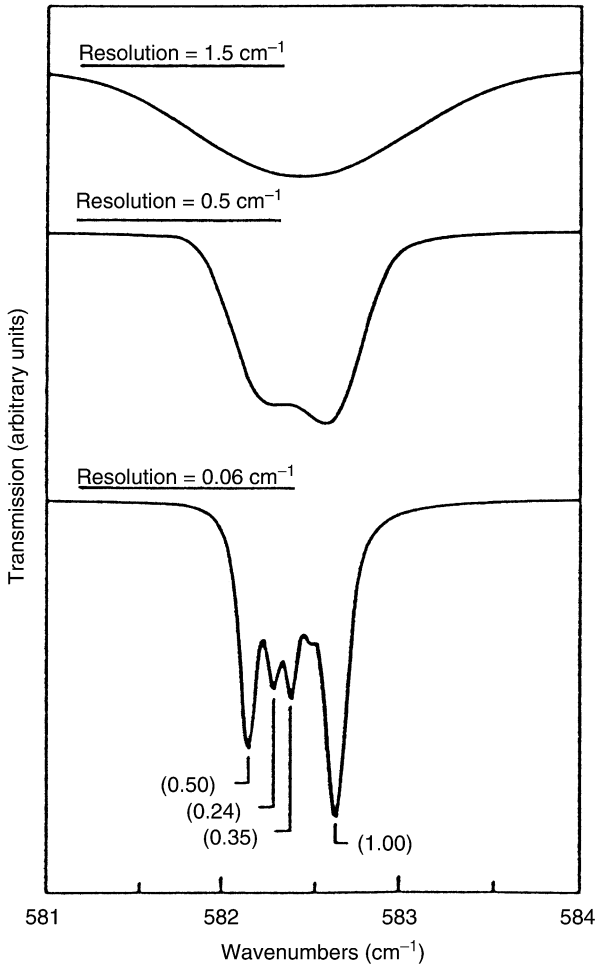


Figure 13 Localized vibrational mode (LVM) absorption band for ^{12}C in GaAs measured at 80 K. The numbers in brackets are estimated band strengths. Reproduced with permission from Theis WM, Bajaj KK, Litton CW and Spitzer WG (1982) Direct evidence for the site of substitutional carbon impurity in GaAs. *Applied Physics Letters* 41: 70.

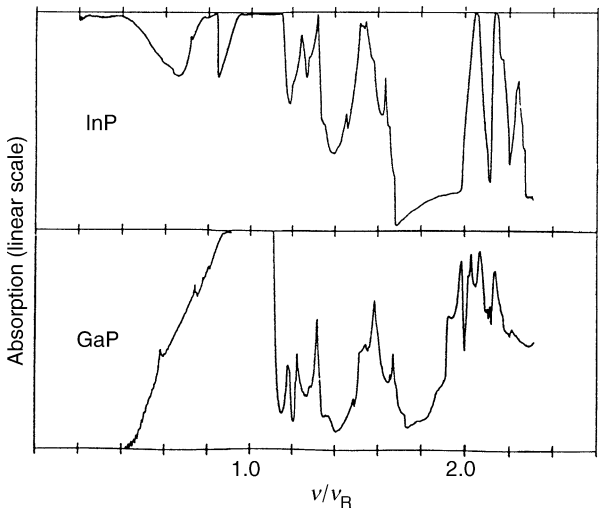


Figure 14 High-resolution far-infrared absorption spectra of GaP and InP with the frequency scale normalized in each case to the reststrahl frequency ($\nu/\nu_R = 1$ corresponds to $\omega = \omega_T$). Reproduced with permission from Koteles ES and Datars WR (1976) Two-phonon absorption in InP and GaP. *Solid State Communications* 19: 221.

surrounding the different ^{12}C atoms in the lattice. Note that ^{12}C is much lighter than either of the host lattice atoms, Ga or As, as required theoretically.

Phonon combination bands

Weak features due to the simultaneous excitation of two or more phonons are observed in the spectra of all semiconductors, and there are two excitation mechanisms. The first is direct excitation of two or more phonons by an incident photon via nonlinear terms in the dipole moment. The second process is indirect: the incident photon first excites a zone center TO phonon and the TO phonon then decays into a combination of phonons elsewhere in the Brillouin zone, a process described as anharmonic decay of the TO phonon. Mathematically, it is described by a complex frequency-dependent damping coefficient, $\gamma(\omega)$, in the oscillator equation [15]. Clearly, the latter mechanism is not available in Si and Ge as the TO phonons are not dipole-active. Anharmonicity is very important in strongly ionic crystals like the alkali halides (i.e., group I–VII compounds like NaCl), leading to prominent phonon combination bands, but it becomes progressively less important in semiconductors as the interatomic bonding becomes less ionic and more covalent, which is the case in the sequence through group II–VI (e.g., semiconductors like CdTe or ZnSe), group III–V (e.g., GaAs or InP), and group IV (Ge and Si), which are entirely covalent. Nearly all two-phonon combinations are dipole-active in zinc-blende structure crystals with the result that the transmission spectrum either side of the reststrahlen band is very rich in weak phonon combination bands, as shown in **Figure 14** for GaP and InP.

Layered Semiconductor Structures

Confined optical phonons

Semiconductor superlattices are fabricated for a variety of electronic and optical applications, and the

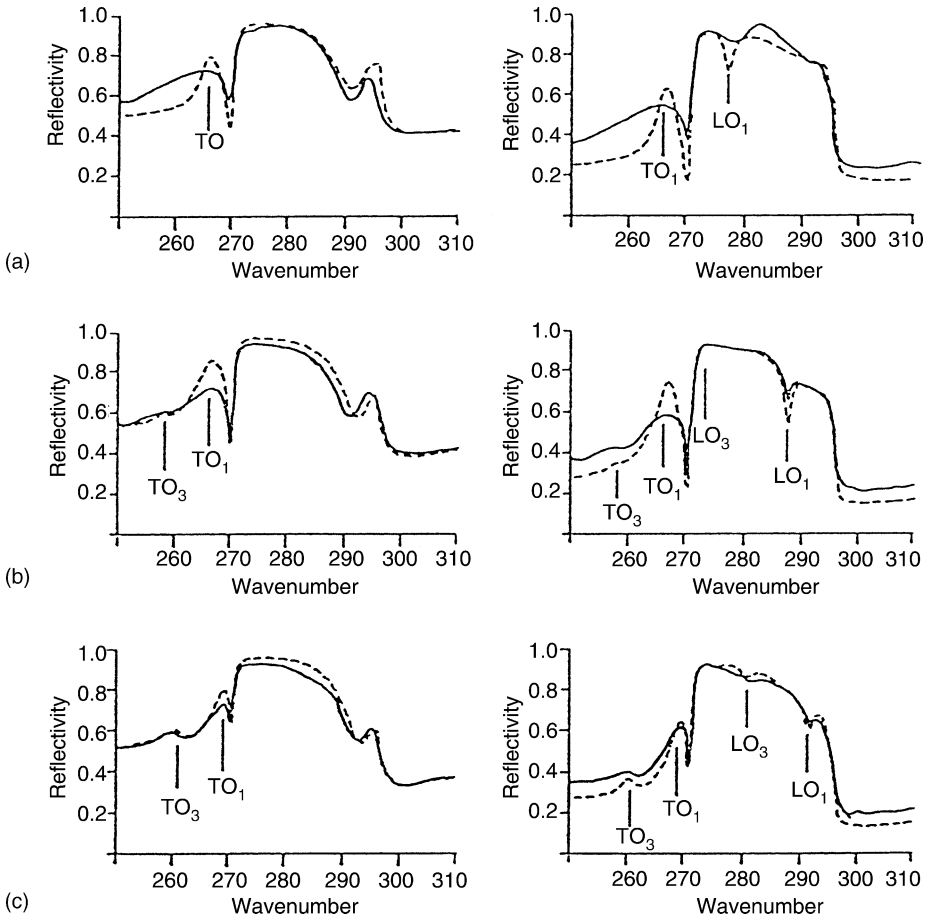


Figure 15 Measured (solid lines) and calculated (dashed lines) oblique incidence (45°) reflectivity spectra at 77 K of three short-period $(GaAs)_n/(AlAs)_n$ superlattices on GaAs substrates: n denotes the number of monolayers. Left curves: s-polarization, right curves: p-polarization. Note that the strong reststrahl reflectivity is from the GaAs substrates and the fine structure is due to confined modes. The subscripts 1, 2, and 3 denote the orders of the confined modes. Reproduced with permission from Dumelow T, Hamilton AA, Parker TJ, Tilley DR, Foxon CT, Hilton D and Moore KJ (1990) Far infrared measurements of bulk and surface phonons in GaAs/AlAs superlattices. *International Journal of Infrared and Millimeter Waves* 11: 901.

quality of the interfaces in such structures is important as it determines the amount of electronic scattering at the interfaces. Some structures are fabricated from alternate layers of materials such as GaAs and AlAs, in which there is no overlap between the frequencies of the optical branches of the phonon dispersion curves. This has the effect that optical phonons excited in one type of layer, e.g., in GaAs, cannot propagate into the other type of layer, i.e., AlAs, and vice versa, so that the optical phonon modes are confined in their respective layers. The confined modes behave like standing waves, with frequencies determined by the thicknesses of the layers and the bulk optical phonon dispersion curves. Thus, if the superlattice structure is of poor quality, and the interfaces are broad due to alloying, rather than abrupt, there will be a change in the effective widths of the layers. This leads to small shifts in the frequencies of the confined phonons which can be measured and used to obtain a quantitative estimate of the amount of interface broadening. An example is shown in Figure 15. The degree of interface broadening determined from these spectra was 1.4 atomic spacings, in agreement with estimates using other techniques.

Surface phonon polaritons and guided waves

The reststrahl reflection band (Figure 6) occurs because there is a gap between ω_T and ω_L in the polariton dispersion curves (Figures 4 and 5), with no solutions for bulk phonon modes. Surface phonon polaritons can, however, propagate in this frequency band, but they have wavevectors that are larger than those, ω/c, of far-infrared radiation, so special techniques are required to observe them. The necessary wavevector enhancement can be obtained in two ways. Prism coupling can be used, as shown in Figure 16. When the angle of incidence of the incident beam at the base of the prism exceeds the critical angle for total internal reflection an evanescent wave with advanced wave vector, $q > \omega/c$, is set up at the base of the prism. This wave can couple to a surface polariton, and the energy used to excite the polariton results in a dip in the reflectivity at the frequency of the surface polariton, hence the name of the technique: attenuated total reflection (ATR). An example is shown in Figure 16. Alternatively, wavevector enhancement can be achieved by placing a grating on the surface of the sample. The incident photon wavevector is then enhanced by $+2m\pi/d$, where d is the grating period and m is an integer.

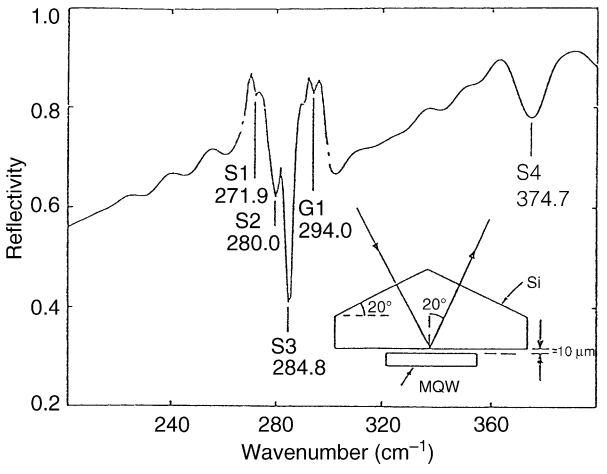


Figure 16 ATR reflection spectrum with dips due to surface phonon polaritons (S) and guided waves (G) in a GaAs/$Al_{0.38}Ga_{0.65}As$ MQW. Inset: schematic diagram of the ATR stage. Reproduced with permission from Dumelow T, Hamilton AA, Parker TJ, Tilley DR, Foxon CT, Hilton D and Moore KJ (1990) Far infrared measurements of bulk and surface phonons in GaAs/AlAs superlattices. *International Journal of Infrared and Millimeter Waves* 11: 901.

Interface phonons

Under certain conditions phonons can propagate along the interface between two different media that are in intimate contact, as for instance, in the case of thin epilayers on substrates, or in multilayered

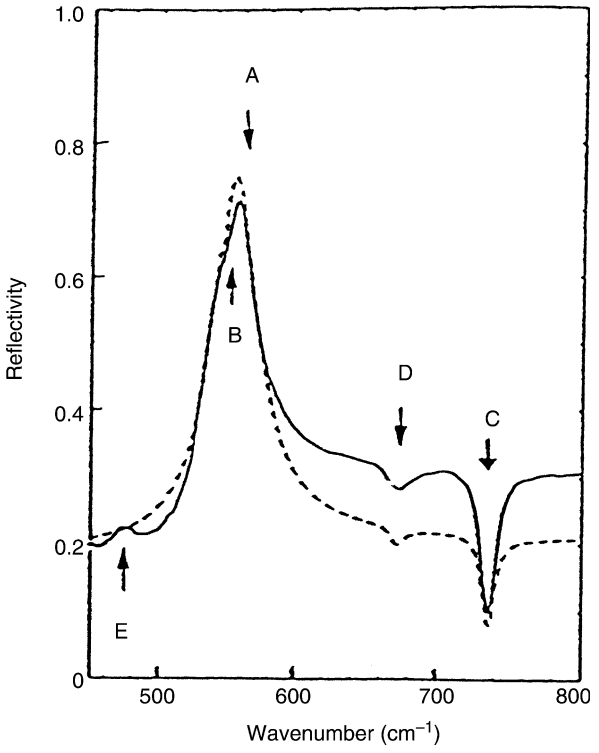


Figure 17 Far-infrared reflectivity of a GaN epilayer on a GaAs substrate at 290 K with dips due to a Brewster mode at C and a Berreman mode at D. Reproduced with permission from Mirjalili G, Parker TJ, Farjami Shayesteh S, Bulbul MM, Smith SRP, Cheng TS and Foxon CT (1998) Far infrared and Raman analysis of phonons and phonon interface modes in GaN epilayers on GaAs and GaP substrates. *Physical Review B* 57: 4656.

structures. There are two common sets of criteria for the existence of such interface modes. The first, leading to so-called Brewster modes, is when the numerator of the Fresnel reflection coefficient passes through zero due to impedance matching in the two media. The second, leading to a Berreman mode, is when the real part of the dielectric constant has a zero crossing in one of the media at a frequency where it is negative in the other medium. Both modes are observed as dips in the reflectivity. A reflection spectrum for a GaN epilayer on a GaAs substrate is shown in Figure 17. The two interface modes occur at frequencies close to the frequencies of GaN LO phonons; the Berreman mode is close to the GaN LO phonon frequency at the base of the GaN epilayer, which is shifted down by about 50 cm^{-1} from the bulk GaN value (735 cm^{-1}) due to disorder arising from the large lattice mismatch between the epilayer and the substrate. Both interface modes couple strongly to free carriers (i.e., plasmons) propagating normal to the interface, and thus act as useful probes of the structural and electronic properties of the epilayer. No other technique is available for probing the electronic properties of the epilayer for propagation normal to the layer.

See also

Semiconductor Materials: GaAs Based Compounds. **Semiconductor Physics:** Band Structure and Optical Properties; Light Scattering; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Angress JF, Goodwin AR and Smith SD (1965) A study of the vibrations of boron and phosphorus in silicon by infrared-red absorption. *Proceedings of the Royal Society A* 287: 64.

Barker AS Jr and Sievers AJ (1975) Optical studies of the vibrational properties of disordered solids. *Reviews of Modern Physics* 47 (Suppl. 2): S1.

Berreman DW (1963) Infrared absorption at longitudinal optic frequency in cubic crystal films. *Physical Review* 130: 2193.

Born M and Huang K (1954) *Dynamical Theory of Crystal Lattices*. Oxford: Clarendon Press.

Dumelow T and Tilley DR (1993) Optical properties of semiconductor superlattices in the far infrared. *Journal of the Optical Society of America* 10: 633.

Dumelow T, Hamilton AA, Parker TJ, Tilley DR, Foxon CT, Hilton D and Moore KJ (1990) Far infrared measurements of bulk and surface phonons in GaAs/AlAs superlattices. *International Journal of Infrared and Millimeter Waves* 11: 901.

Johnson FA (1959) Lattice absorption bands in silicon. *Proceedings of the Physical Society* 73: 265.

Kittel C (1996) *Introduction to Solid State Physics*, 7th edn. New York: John Wiley.

Kleinman DA and Spitzer WG (1960) Infrared lattice absorption of GaP. *Physical Review* 118: 110.

Koteles ES and Datars WR (1976) Two-phonon absorption in InP and GaP. *Solid State Communications* 19: 221.

Mirjalili G, Parker TJ, Farjami Shayesteh S, Bulbul MM, Smith SRP, Cheng TS and Foxon CT (1998) Far infrared and Raman analysis of phonons and phonon interface modes in GaN epilayers on GaAs and GaP substrates. *Physical Review B* 57: 4656.

Newman RC (1974) *Studies of Crystal Defects*. London: Taylor and Francis.

Theis WM, Bajaj KK, Litton CW and Spitzer WG (1982) Direct evidence for the site of substitutional carbon impurity in GaAs. *Applied Physics Letters* 41: 70.

Waugh JLT and Dolling G (1963) Crystal dynamics of gallium arsenide. *Physical Review* 132: 2410.

Light Scattering

M Balkanski, Université Pierre et Marie Curie, Paris, France

© 2005, Elsevier Ltd. All Rights Reserved.

When a monochromatic light beam passes through a transparent medium, a small part of the light is scattered out of the incident direction with a change in frequency. The scattering processes are usually classified into three groups. The light scattering is named:

Brillouin scattering when the light is scattered with a small frequency shift that varies continuously with scattering angle.

Raman scattering when the light is scattered with a relatively large frequency shift that is independent of scattering angle. The possibility of observing a given transition, however, depends on the orientation of the crystal relative to the polarization of the incident light.

Rayleigh scattering when the light is scattered without frequency shift. In this case the scattering is elastic with the incident and scattered frequencies equal: $\omega_i = \omega_s$. Rayleigh scattering is particularly useful in the study of critical phenomena or aspects related to the size and polarizability of particles.

Brillouin Scattering

Brillouin scattering is caused by the interaction of light with the acoustic modes of vibration of the crystal. Consider an acoustic wave of frequency ω_a propagating with velocity $\pm v$ as shown in Figure 1. Light with incident frequency ω_i interacts strongly with acoustic waves satisfying the Bragg condition $\omega_a = 2\omega_i(v/c)n\sin(\theta/2)$, where n is the refractive index of the material and θ is the angle of deviation of the scattered beam. Since the acoustic waves are moving with velocity $\pm v$ the scattered light suffers a Doppler shift in frequency and exhibits a frequency doublet at the angle θ given by the Brillouin equation

$$\omega_i \pm \omega_a = \omega_i \pm 2\omega_i(v/c)n\sin(\theta/2)$$

In terms of the quantum picture, Brillouin scattering corresponds to scattering of incident photons of frequency ω_i and wavevector $\mathbf{k}_i$ into scattered photons of frequency ω_s and wavevector $\mathbf{k}_s$ with the emission or absorption of an acoustic phonon of frequency ω_a and wavevector $\mathbf{q}$. The predicted Brillouin spectrum is shown in Figure 2. It consists of Stokes ($\omega_s = \omega_i - \omega_a$) and anti-Stokes ($\omega_s = \omega_i + \omega_a$) components corresponding to phonon emission and absorption, respectively.

The frequency shift $|\omega_i - \omega_s|$ for a well-defined scattering angle θ directly gives the acoustic phonon frequency and velocity. The width Γ of the Brillouin component of the spectrum is a measure of the damping or attenuation of the wave.

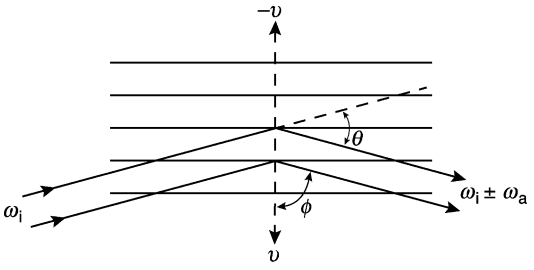


Figure 1 Light scattering at angle ϕ and Bragg reflection at angle θ by moving thermal waves.

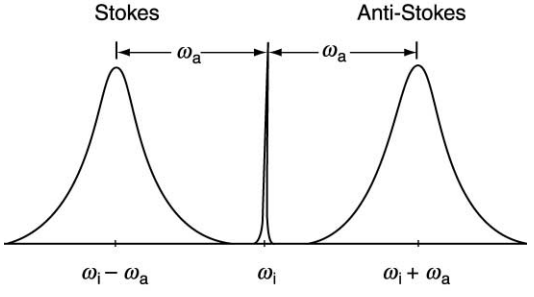


Figure 2 Schematic diagram of a Brillouin spectrum.

Raman Scattering

The essential difference between Brillouin and Raman scattering is that in Raman scattering the incident light beam is scattered with relatively large frequency shift independent of the scattering angle. The same basic considerations apply to Raman scattering that apply to Brillouin scattering. The Raman spectrum has Stokes and anti-Stokes branches corresponding to the emission and absorption, respectively, of an elementary excitation. A variety of elementary excitations are important. They include optical phonons and, in the case of magnetic materials, magnons. Also of interest are electronic excitations such as intraband single-particle excitations, interband excitations, and collective excitations (plasmons).

Raman Scattering by Phonons

An inelastic light scattering event involves the destruction of a photon of frequency ω_i incident from a light source, the creation of a scattered photon of frequency ω_s and the creation or destruction of an optical phonon of frequency ω_o.

Destruction of a photon involves interaction of the radiation field with matter by which a pair of free carriers is created in the semiconductor corresponding to a virtual transition between the valence and conduction bands. The excited state, free electron in the conduction band, interacts with the lattice via the optical deformation potential. This interaction results in the creation or annihilation of an optical phonon of frequency ω_s. When the resulting process corresponds to the creation emission of a phonon, the frequency of the scattered photon is $\omega_s = \omega_i - \omega_a$ and is referred to as the Stokes component of the spectrum. When the process is the annihilation or absorption of a phonon the scattered frequency is $\omega_s = \omega_i + \omega_a$, referred to as the anti-Stokes component.

These two first-order scattering events can be described in terms of Feynman diagrams as shown in Figure 3.

Considering these elementary processes one can calculate, using time-dependent perturbation theory, the Raman amplitude and hence the differential scattering cross-section. The remarkable feature here is that for incident frequencies approaching the frequencies of an electronic excited state of the material the Raman amplitude diverges. This leads to the phenomenon of resonant Raman scattering.

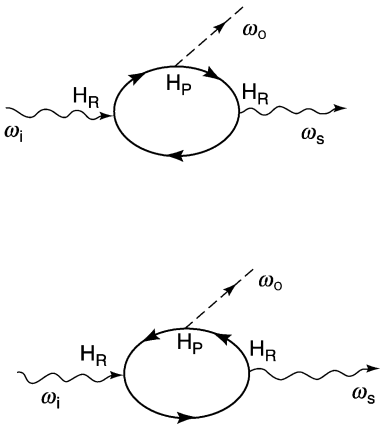


Figure 3 Feynman diagrams for two first-order scattering events.

The Raman scattering efficiency, which is the scattering cross-section per illuminated area can be calculated numerically. Numerical estimates give efficiencies typically of order 10^{-6} or 10^{-7}.

In polar semiconductors the macroscopic electric field leads to an additional contribution to the scattering cross-section for the longitudinal optical (LO) phonon modes. The macroscopic field is a consequence of the Coulomb field of the ionic charges and gives rise to an electron–phonon interaction.

There is no simple relationship between the scattering intensities for transverse optical (TO) and LO phonons in polar semiconductors.

When the probing frequency coincides with the transverse optical mode frequency, i.e., for the regime $cq \cong \omega_{TO}$ (where c is the light velocity, q the photon wavevector and ω_{TO} the transverse optical phonon frequency) in polar semiconductors it is necessary to take into account the polariton nature of the coupled photon–TO-phonon modes and their associated macroscopic electrical field. The dispersion relation of the coupled photon–TO mode, called a polariton, consists of two branches: an upper branch ω_+, in which the electro-optical contribution dominates, and the lower branch ω_-, which contains essentially the mechanical contribution. This leads to a contribution to the Raman amplitude from both mechanical and electro-optical origins. An exception occurs for the lower polariton branch with $\omega = \omega_{TO}$; then the electro-optic contribution is very small compared to the mechanical contribution. If $cq \gg \omega_{TO}$ as in the upper polariton branch, destructive interference between the mechanical and electro-optic contributions can occur.

Selection Rules in Raman Scattering

Not all elementary excitations in semiconductors scatter light. The Raman active modes are determined by selection rules established using group-theoretical methods. The various normal modes in a given crystal correspond to various symmetries of the vibrations of the atoms in the crystal and are characterized by the irreducible representations of the space group of the crystal lattice. One can show that a normal mode can participate in a first-order Raman transition if and only if its irreducible representation is the same as one of the irreducible representations that occur in the reduction of the representation of the Raman tensor.

An important result of group theory is the rule of mutual exclusion which states that, in crystals with a center of inversion, excitations that are active in the first-order infrared spectrum are inactive in the first-order Raman spectrum, and conversely, excitations that are active in the first-order Raman spectrum are inactive in the first-order infrared spectrum. In particular, the first-order spectra due to optical phonons in NaCI are infrared active, but Raman inactive, whereas in Si they are infrared inactive, but Raman active. This difference is related to the fact that each atomic site in NaCI is a center of inversion and the active optical phonons have odd parity whereas each midpoint between two nearest-neighbor atomic sites in Si is a center of inversion and the active optical phonons have even parity.

In a crystal with the zinc blend structure, there is no center of inversion, and the crystal is both infrared and Raman active. The only non-vanishing elements of the Raman tensor $R(\hat{e})$ with $\hat{e} \| z$ are the xy and yx elements. To observe the Raman effect of an LO phonon propagating in the z-direction, one can arrange the polarization of the incident light parallel to the x-axis and observe the scattered light with its polarization parallel to the y-axis or vice versa. An analogous set of constraints applies to TO phonons. A typical spectrum is illustrated in Figure 4.

For higher-order Raman processes in which several phonons participate, it is the product of the irreducible representations of the phonons involved that must be the same as the irreducible representation of the Raman tensor.

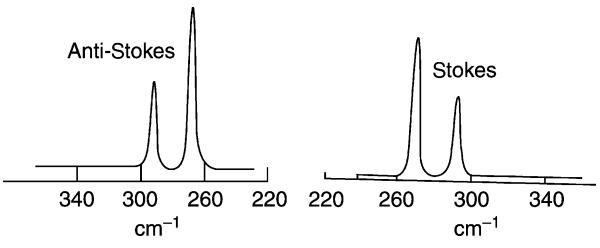


Figure 4 Raman spectrum for optical phonons.

Geometrical Aspects of First-Order Raman Scattering

The observation of light scattering by optical phonons in transparent crystals is usually done in a geometry in which the linearly polarized incident light beam is directed along, say, the x-axis and the scattered beam is observed along the y-axis (see Figure 5). When the crystal is not transparent, as is often the case in semiconductors, the observation is made in the backscattering geometry in which the scattered beam is observed in the opposite direction to the incident beam.

The scattering geometry affects the range of phonon wavevectors that is accessible in first-order Raman scattering. The condition of wavevector conservation is

$$\mathbf{K}_i - \mathbf{K}_s = \pm\mathbf{q}$$

where $\mathbf{q}$ is the phonon wavevector and the plus (minus) sign refers to the Stokes (anti-Stokes) process. The Stokes geometry is shown in Figure 6 and satisfies the following relation:

$$q^2 = k_i^2 + k_s^2 - 2k_i\, k_s \cos\theta$$

Forward scattering is characterized by $\theta = 0$ and a minimum value of q given, for isotropic media, by

$$q_{min} = \frac{n(\omega_i)\,\omega_i - n(\omega_s)\omega_s}{c}$$

where $n(\omega_i)$ and $n(\omega_s)$ are the refractive indices of the crystal for the incident and scattered light, respectively, and $k_{i(s)} = n(\omega_{i(s)})\omega_{i(s)}/c$.

Back scattering is characterized by the maximum value of q when $\theta = 180°$ and is given by

$$q_{max} = \frac{n(\omega_i)\omega_i - n(\omega_s)\omega_s}{c}$$

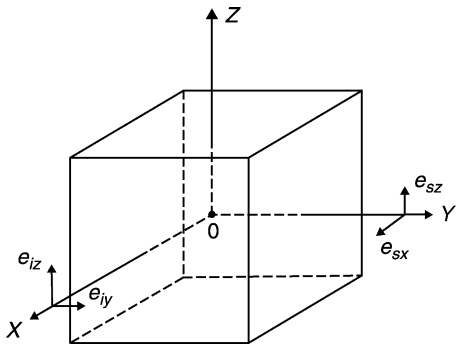


Figure 5 Geometry for light-scattering experiments in transparent crystals.

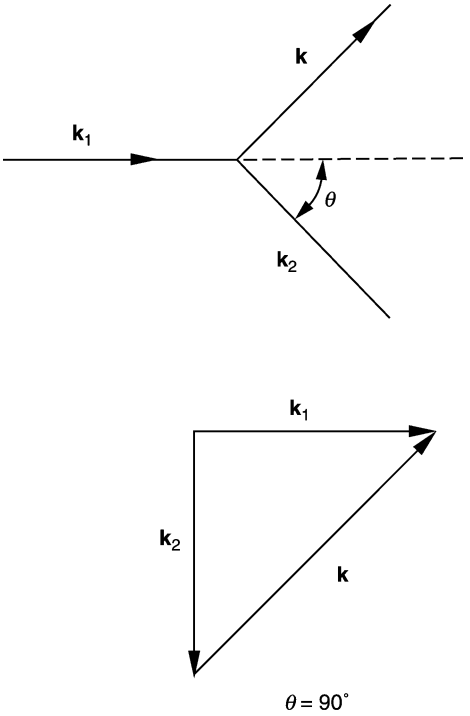


Figure 6 Diagram for Stokes processes.

For typical light-scattering experiments in the visible region the range of the incident wavevector is $0 < k_i < 10^6\ \mathrm{cm}^{-1}$. This implies that for first-order scattering processes the accessible range of $\mathbf{q}$ under conditions of wavevector conservation is small compared to a nonzero reciprocal lattice wavevector. Light-scattering experiments yield the frequencies of optical modes at essentially the center of the Brillouin zone.

The energy and momentum conservation rules have to be modified when the lifetime of the crystal excitations are strongly limited by their decay into other crystal excitations. Momentum conservation breaks down in imperfect crystals, in solids lacking translational symmetry like amorphous materials, and in crystals which are opaque to incident and scattered light.

In those cases where the incident and scattered waves are damped inside the scattering volume, such as occur in small-gap semiconductors that are opaque at the light frequencies involved, $\mathbf{k}_i$ and $\mathbf{k}_s$ are complex. The inelastic scattering is due to excitations having a range of wavevector

$$\Delta q = |\mathrm{Im}\,\mathbf{k}_i| + |\mathrm{Im}\,\mathbf{k}_s|$$

about $\mathbf{q} = \mathrm{Re}(\mathbf{k}_i - \mathbf{k}_s)$. Effects associated with such a wavevector uncertainty have been reported in

Raman scattering spectra of III–V semiconductor compounds.

Second-Order Raman Scattering

Second-order Raman scattering involves two phonons rather than a single phonon as in first-order scattering. The possible second-order processes are the following: two phonons may be created, giving a Stokes component in the scattered light, or one is created and the other destroyed, giving a Stokes or anti-Stokes component, or both may be destroyed giving an anti-Stokes component. For each of these cases there are the possibilities of a pair of first-order electron–phonon interactions or a single second-order electron–phonon interaction.

In the case of second-order Raman scattering the wavevector conservation condition is $\mathbf{q}_1 \pm \mathbf{q}_2 \cong 0$, where $\mathbf{q}_1$ and $\mathbf{q}_2$ are the wavevectors of the phonons involved. This condition places no restriction on the magnitudes of the individual wavevectors, other than $|q_1| \cong |q_2|$, which is in contrast to first-order scattering where $q \cong 0$.

The possible phonon frequency pairings associated with second-order processes are $\omega_1 + \omega_2$ (combination band), $\omega_1 - \omega_2$ (difference band), and $2\omega_1$, (overtone band). The second-order scattering spectrum covers a broad range of frequencies. The overtone spectrum provides a measure of the phonon density of states.

Resonant Light Scattering

General Formulation

Resonant Raman scattering occurs when the incident or scattered photon energy is close to the energy of an intermediate electronic state relative to the ground electronic state. Certain terms in the Raman amplitude then diverge leading to a very large scattering cross-section. Divergences occur in the Stokes spectrum when $h\omega_i = h\omega_a$, and $h\omega_s = h\omega_b$: the photon energies *resonate* with the excitation energies of the intermediate states *a* or *b*. If *a* and *b* are the same state, resonance occurs nearly simultaneously for both ω_i and ω_s, and gives rise to a particularly strong enhancement of the scattering efficiency. It is thus clear that the resonance phenomenon is specific to the nature of the intermediate state, and its investigation leads to basic information concerning the electronic states of the system. We shall now examine the resonance behavior in several cases where the intermediate states are carrier Bloch states, free excitons, or bound excitons.

Resonance at the Fundamental Bandgap

The divergent behavior of the Raman amplitude is associated with the factor $(h\omega_i - E_g)^{-1}$ where E_g is the lowest direct bandgap of the material.

Resonance at Free Exciton States

The Raman intensity as a function of the incident photon energy has a Lorentzian line shape in the range of a single intermediate state. It is centered at the resonance frequency and has a width determined by the lifetime of the intermediate state. The distinction between scattering at resonance and scattering off resonance in the wings of the Lorentzian is that at resonance the intensity is determined by the exciton lifetime, whereas off resonance it is determined by the frequency separation from resonance. Since in many instances excitons have very long lifetimes, they can lead to a very large Raman cross-section at resonance.

Resonance at Bound Exciton States

Scattering induced by the presence of impurities involving bound exciton states has the following essential features: (a) observation of sharp resonance at the energies of the bound exciton which form discrete levels below the free exciton resonance; (b) dependence upon impurity concentration; (c) involvement of LO phonons of wavevector of the order of the inverse of the impurity state radius. Resonance at bound exciton energies has been observed in impurity-induced resonant Raman scattering by LO phonons in CdS. The resonance is centered at the absorption peak for the impurity state and has approximately Lorentzian shape. If the impurity gives rise to a localized vibrational mode, bound exciton resonance in Raman scattering due to the local mode can occur.

Anharmonic Effects on Raman Spectra

In a perfect crystal whose vibrations are harmonic and which is transparent to the incident and scattered light, the Raman lineshape is a delta function centered on the optical phonon frequency. Experimentally one finds that even in crystals of very high quality, the Raman line is broadened into a roughly Lorentzian shape with a width that increases with increasing temperature. This width can be attributed to anharmonic terms in the vibrational Hamiltonian.

The temperature dependence of the line width is determined by the phonon occupation factors. As the temperature approaches zero, the phonon

occupation factor approaches zero also, and the width approaches the value due to the zero-point motion of the nuclei. At temperatures above the Debye temperature, the width due to cubic anharmonicity becomes proportional to T. Quartic anharmonicity imparts a T^2 dependence to the width which becomes significant at very high temperatures.

In addition to line broadening, anharmonicity causes a shift of the frequency of peak intensity to lower values as the temperature increases. Contributions to the shift proportional to T and T^2 arise in the high-temperature regime from cubic and quartic anharmonicity, respectively.

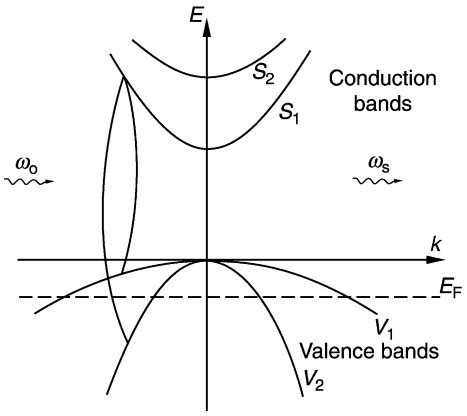


Figure 7 Electronic transitions occurring in intervalence band scattering.

Light Scattering due to Electronic Excitations

Light Scattering by Plasmons

Plasmons are similar to LO phonons in that they have a macroscopic electric field associated with them. The field modulates the electric susceptibility and gives rise to Raman scattering. The difference in frequency of the incident and scattered light is equal to the plasma frequency ω_p specified by $\omega_p = e^2 n/\varepsilon_0 \varepsilon_\infty m^*$, where n is the electron concentration, m^* is their effective mass, and ε^∞ is the high-frequency dielectric constant.

When the plasma frequency is close to the LO phonon frequency, there is a plasmon–phonon interaction via the macroscopic electric fields that leads to coupled modes and forces the frequencies apart.

Light Scattering Associated with Interband Transitions

The plasmons dealt with in the preceding section are an intraband type of excitation. Also of interest are interband transitions in which a carrier undergoes a transition from one band to a different band during the scattering process. An example is provided by p-type silicon in which an electron makes a transition from an occupied state in the light-hole band to an unoccupied state in the heavy-hole band via the conduction band as shown in Figure 7.

For the case of isotropic bands the scattering efficiency spectrum at 0 K has sharp edges. If the anisotropy of the valence bands is taken into account, the sharp edges of the spectrum are smoothed out.

The experimental Raman spectrum for p-type Si exhibits a broad continuum due to the electronic transitions but in the same spectral range there are also Raman active optical phonons which superpose on the electronic continuum and distort it. The resulting distortion of the spectrum is associated with resonant interference that occurs between the discrete phonon line and the electronic continuum.

See also

Scattering: Raman Scattering. **Semiconductor Physics:** Outline of Basic Electronic Properties.

Further Reading

Balkanski M (1972) In: Abeles F (ed.) *Optical Properties of Solids*. Amsterdam: North-Holland.

Balkanski M, Jain KP, Beserman R and Jouanne M (1975) *Phys. Rev.* B12: 4328.

Birman JL (1974) *Encyclopedia of Physics*, vol XXV 12b. Berlin: Springer-Verlag.

Born M and Huang K (1954) *Dynamical Theory of Crystal Lattice*. Oxford, UK: Oxford University Press.

Brillouin L (1922) *Ann. Physique* 17: 88.

Buchler S and Burstein E (1974) *Phys. Rev. Letter* 33: 908.

Daman TC and Shah J (1971) *Phys. Rev. Letter* 27: 1506.

Hayes W and Loudon R (1978) Scattering of Light by Crystals. New York: John Wiley.

Haro E, Balkanski M, Wallis RF and Wanser KH (1986) *Phys. Rev.* B34: 5358.

Kanehisa MA, Wallis RF and Balkanski M (1982) *Phys. Rev.* B25: 7619.

Wallis RF and Balkanski M (1986) Many-Body Aspects of Solid Spectroscopy. Amsterdam: North-Holland.

Yu PY, Pilkhun MH and Evangelisti F (1978) *Solid State Communications*, vol. 25, p. 371.

SEMICONDUCTOR PHYSICS

Polarons

J T Devreese, Universiteit Antwerpen, Antwerp, Belgium

© 2005, Elsevier Ltd. All Rights Reserved.

The Polaron Concept

If an electron (or hole) is placed in a polarizable medium (like an ionic crystal or a polar semiconductor), it induces a deformation of polarization field around itself. On the other hand, the electron is attracted to the potential well of this disturbed polarization field. Thus, a feedback loop between the electron and the polarization field arises, which leads to a highly correlated state of the charge carrier and the polarization field. A conduction electron (or hole) together with its self-induced polarization in an ionic crystal or in a polar semiconductor forms a quasiparticle, which is called a polaron. The polaron concept was introduced by Landau in 1933.

The physical properties of the polaron differ from those of the band carrier. In particular, the polaron is characterized by its binding (or self-) energy E_0, effective mass m^* and by its response to external electric and magnetic fields (e.g., dc mobility and optical absorption coefficient).

If the spatial extension of the polaron is large compared to the lattice parameters of the solid, the latter can be treated as a polarizable continuum. This is the case of a 'large polaron', to which most of this article is confined.

The polaron characteristics are determined by the coupling between an electron and the long-wavelength longitudinal optical (LO) phonons. The strength of this coupling is expressed in terms of the constant introduced by Fröhlich:

$$\alpha = \frac{e^2}{\hbar c}\sqrt{\frac{m_b c^2}{2\hbar\omega_{LO}}}\left(\frac{1}{\varepsilon_\infty} - \frac{1}{\varepsilon_0}\right) \qquad [1]$$

In this definition, ω_{LO} is the frequency of LO phonons, c is the velocity of light, and ε_∞ and ε_0 are, respectively, the electronic and the static dielectric constant of the polar crystal.

In order to find the coupling constant according to eqn [1], it is crucial to know the electron (hole) band mass m_b. This unknown parameter can, in principle, be determined from experiment. However, in ionic crystals and polar semiconductors, experiment measures the polaron effective mass m^* rather than the band mass, because of the polaron effect. The problem of determining the band mass is therefore far from trivial and should be tackled on the basis of the results of polaron theory. In Table 1 the coupling constants are indicated for a number of crystals.

Landau and Pekar were the first to investigate the self-energy and the effective mass of the strong-coupling polaron, when $\alpha \gg 1$. As seen from Table 1, in substances with large values of the electron–phonon coupling constant (like alkali halides, TlBr, CdF_2, $SrTiO_3$) α ranges between 2.5 and 4.5. Though, strictly speaking, polarons in such materials are of intermediate coupling, their properties can be approached within the strong-coupling approximation.

The cases when $\alpha \ll 1$ (like InSb, InAs, GaAs) are referred to as the weak-coupling regime. Polarons in many other substances with $\alpha < 1$ (e.g., GaP, CdTe, ZnSe, CdSe, CdS, CdO) are well described in the weak-coupling approximation.

In the weak-coupling regime, the states of the polaron can be imagined as those of a band carrier perturbed by the electron–phonon interaction. Fröhlich provided in 1954 the first weak-coupling perturbation theory results for the polaron characteristics:

$$E_0 = -\alpha\hbar\omega_{LO} \qquad [2]$$

$$m^* = \frac{m_b}{1 - \alpha/6} \qquad [3]$$

Feynman formulated the polaron problem in the Lagrangian (path integral) form of quantum mechanics and obtained an upper bound for the polaron self-energy at all values of the α, which at weak and strong coupling gave accurate limits. Over the years the Feynman all-coupling model

Table 1 Electron–phonon coupling constants

Material	α	*Material*	α
InSb	0.02	AgCl	1.8
InAs	0.052	KI	2.5
GaAs	0.068	TlBr	2.55
GaP	0.201	KBr	3.05
CdTe	0.286	RbI	3.16
ZnSe	0.43	CdF_2	3.2
CdSe	0.49	KCl	3.5
CdS	0.52	CsI	3.67
CdO	0.74	RbCl	3.81
AgBr	1.6	$SrTiO_3$	4.5

for the polaron has been proven to be in many respects the most successful approach to the polaron problem.

Optical Properties and Internal Structure of Polarons

Optical Absorption of Polarons at Weak Coupling

At zero temperature and in the weak-coupling limit, the optical absorption is due to the following elementary polaron scattering process, schematically shown in Figure 1. An incoming photon is absorbed by a polaron. The polaron emits a phonon during the absorption process and takes recoil energy from the incident light. At zero temperature, the absorption coefficient for absorption of light with frequency Ω can be expressed in terms of elementary functions in two limiting cases: in the region of comparatively low frequencies $(\hbar(\Omega - \omega_{LO})/\zeta \ll 1)$

$$\Gamma(\omega)=\frac{1}{\varepsilon_0 nc}\frac{2^{1/2}N^{2/3}\alpha}{(3\pi^2)^{1/3}}\frac{e^2}{(\hbar m_b\omega_{LO})^{1/2}}\frac{\omega-1}{\omega^3}\Theta(\omega-1) \quad [4]$$

and in the high-frequency region $(\hbar(\Omega-\omega_{LO})/\zeta \gg 1)$

$$\Gamma(\omega)=\frac{1}{\varepsilon_0 nc}\frac{2Ne^2\alpha}{3m_b\omega_{LO}}\frac{(\omega-1)^{1/2}}{\omega^3}\Theta(\omega-1) \quad [5]$$

Here $\omega=\Omega/\omega_{LO}$, ε_0 is the dielectric permittivity of the vacuum, n is the refractive index of the medium, N is the concentration of polarons, and ζ is the Fermi level for electrons. The step function

$$\Theta(\omega-1)=\begin{cases}1 & \text{if } \omega>1\\ 0 & \text{if } \omega<1\end{cases} \quad [6]$$

reflects the fact that at zero temperature absorption of light accompanied by emission of a phonon can occur only if the energy of a photon is larger than that of a phonon ($\omega>1$). In the weak-coupling limit, according to eqn [4], the absorption spectrum consists of a one-phonon line.

Experimentally, the one-phonon line has been clearly seen for free polarons in the infrared absorption spectra of CdO-films (see Figure 2). In CdO, which is a weakly polar material with $\alpha \approx 0.74$, the optical absorption band is observed in the spectral region between 6 and 20 μm (above the LO phonon frequency). As seen from Figure 2, this optical absorption band can be attributed to the weak-coupling polaron absorption as described by eqn [4].

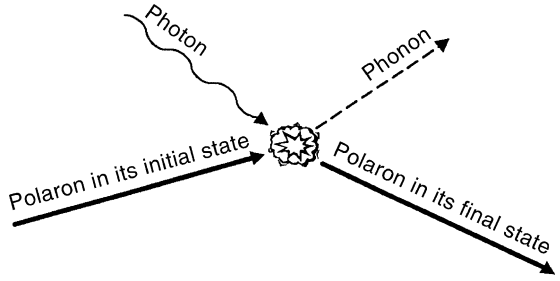


Figure 1 The elementary polaron scattering process leading to absorption of an incoming photon and to generation of an outgoing phonon.

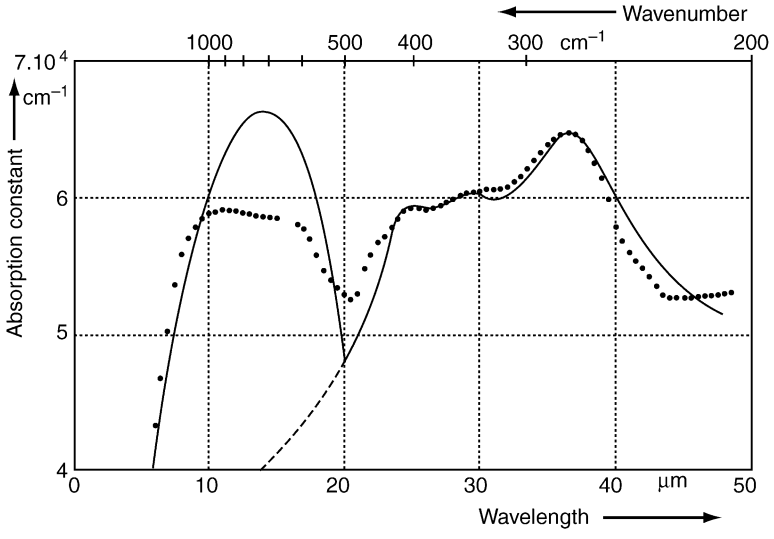


Figure 2 Optical absorption spectrum of a CdO film with the carrier concentration $N = 5.9 \times 10^{19}$ cm^{-3} at $T = 300$ K. The theoretical results are shown with (solid curves) and without (dashed curves) the polaron contribution of eqn [4], and compared to the experimental data (solid dots). (Reproduced with permission from Finkenrath H, Uhle N and Waidelich W (1969) The influence of phonons and polarons on the infrared absorption of cadmium oxide[*1]. *Solid State Communications* 7: 11–14.)

With strengthening coupling, it becomes possible that absorption of a photon is accompanied by emission of two, three, …, K phonons, thus giving rise to two-, three-, …, K-phonon sub-bands in the optical absorption spectra. At high temperatures, absorption of a photon can be accompanied not only by emission, but also by absorption of one or more phonons.

Optical Absorption of Polarons at Arbitrary Coupling

The theoretical absorption spectrum of a single large polaron (at all electron–phonon coupling strengths) is given by the expression

$$\Gamma(\omega) = \frac{1}{n\varepsilon_0 c} \frac{\mathrm{Im}\Sigma(\omega)}{[\omega - \mathrm{Re}\Sigma(\omega)]^2 + [\mathrm{Im}\Sigma(\omega)]^2} \qquad [7]$$

The so-called memory function $\Sigma(\omega)$ contains the dynamics of the polaron and depends on α and ω. This function is related to the impedance function $Z(\omega)$ of the polaron through $-\mathrm{i}Z(\omega) = \omega - \Sigma(\omega)$.

Optical Absorption of Polarons at Strong Coupling

The absorption of light by free large polarons was treated by Kartheuser, Evrard and Devreese using the adiabatic strong-coupling method by Landau and Pekar. The polaron ground state in this scheme has the energy

$$E_0 = -\frac{\alpha^2}{3\pi}\hbar\omega_{\mathrm{LO}} = -0.106\alpha^2\hbar\omega_{\mathrm{LO}} \qquad [8]$$

If the lattice polarization is allowed to relax or adapt to the electronic distribution of the excited electron (which itself then adapts its wavefunction to the new potential, etc., leading to a self-consistent final state), the so-called relaxed excited state (RES) results. Its energy is

$$E_{\mathrm{RES}} = -0.041\alpha^2\hbar\omega_{\mathrm{LO}} \qquad [9]$$

It was argued by Kartheuser, Evrard and Devreese, that for a sufficiently large α ($\alpha > 3$), the (first) RES of a polaron is a stable state, which can participate in optical absorption transitions. This idea was important in order to understand the optical absorption spectrum in the strong-coupling regime. In particular, the following scenario of a transition, which leads to a zero-phonon peak in the absorption by strong-coupling polaron, can be suggested. If the frequency of a photon is equal to

$$\Omega_{\mathrm{RES}} \equiv \frac{E_{\mathrm{RES}} - E_0}{\hbar} = 0.065\alpha^2\omega_{\mathrm{LO}} \qquad [10]$$

then the electron goes from the ground state (following the analogy with the hydrogen atom, let us call it '1s') to an excited state ('2p'), while the lattice polarization in the final state is adapted to the '2p' state of the polaron.

Each RES of the polaron can be accompanied by 1, 2, 3, … free optical phonons in the lattice. Such states are called scattering states (ScS). They are shown in Figure 3. Starting with the photon energy

$$\Omega_{\mathrm{RES}} + \omega_{\mathrm{LO}} \qquad [11]$$

a transition of the polaron towards the first ScS, belonging to the RES, becomes possible. In the course of this transition, while absorbing a photon, the polaron goes from the ground state to the RES with emission of one extra phonon. This process is called one-phonon sideband absorption.

The one-, two-, …, K-phonon sidebands give rise to the broad structure of the absorption coefficient. These transitions occur without lattice relaxation. The (unstable) polaron state, in which the lattice polarization corresponds to the electron ground state, while the electron is excited ('a frozen lattice'), is referred to as the *Franck–Condon* (FC) state of a polaron. Within the adiabatic approach, the energy of the lowest FC state is

$$E_{\mathrm{FC}} = \frac{\alpha^2}{9\pi}\hbar\omega_{\mathrm{LO}} = 0.035\alpha^2\hbar\omega_{\mathrm{LO}} \qquad [12]$$

The superposition of the one-, two-, …, K-phonon sidebands is centered at the FC transition frequency

$$\Omega_{\mathrm{FC}} \equiv \frac{E_{\mathrm{FC}} - E_0}{\hbar} = 0.141\alpha^2\omega_{\mathrm{LO}} \qquad [13]$$

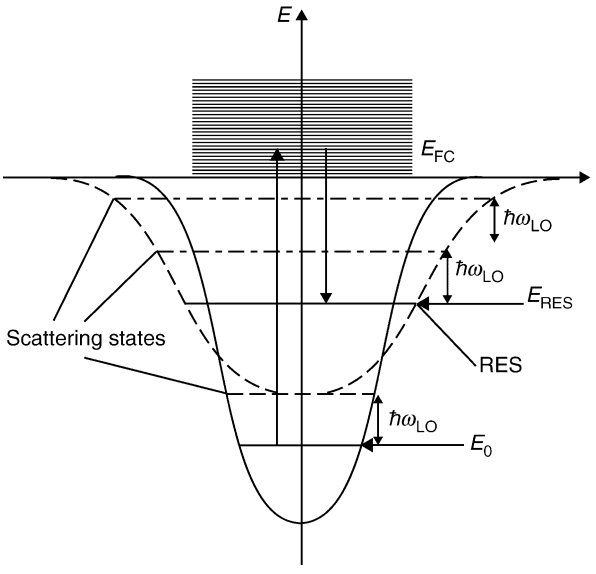


Figure 3 Internal excitations of a polaron at strong coupling: E_0 – the ground state, E_{RES} – the (first) relaxed excited state; the Franck–Condon states (E_{FC}) are shown with a set of horizontal lines above the zero energy. In fact, both the Franck–Condon states and the relaxed excited states lie in the continuum and, strictly speaking, are resonances.

The optical absorption spectrum at strong coupling is, thus, characterized by the following features.

(a) There is an intense absorption peak ('zero-phonon line'), which corresponds to a transition from the ground state to the first RES at Ω_{RES}.
(b) At $\Omega_{RES} + 1$, a phonon sideband structure appears (namely, a K-phonon sideband at $\Omega_{RES} + K$). The maximum of this structure occurs at the FC transition frequency Ω_{FC}.

The qualitative behavior predicted by Kartheuser, Evrard and Devreese, namely, an intense zero-phonon (RES) line with a broader sideband at the high-frequency side, was revealed using a general all-coupling expression for the optical absorption coefficient [7] at $\alpha = 5, 6, 7$.

Internal Structure of Polarons at Arbitrary Coupling

The optical absorption spectra calculated according to the theory of Devreese *et al.* are displayed in Figure 4 for values of α in the range from 1 to 6.

The absorption spectrum for $\alpha = 1$ consists of a one-phonon line, similar to the weak-coupling result. The absorption peak for $\alpha = 3$ might correspond to a one-phonon peak at this intermediate coupling. It was also argued that part of the oscillator strength in the absorption for $\alpha = 3$ is due to lattice relaxation in the final states, be it to a smaller extent than for $\alpha = 5$.

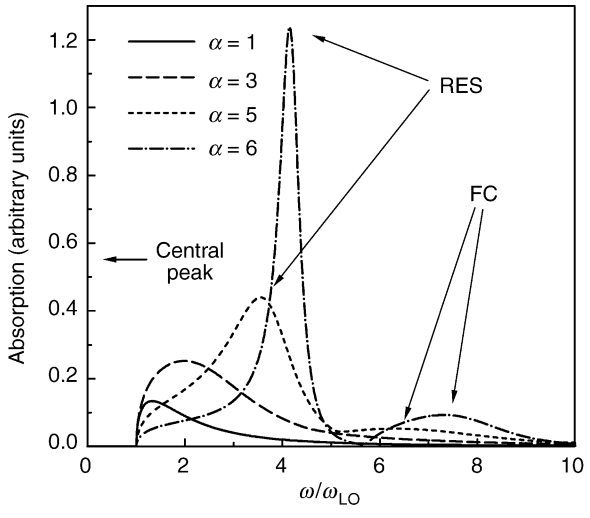


Figure 4 Polaron optical absorption spectra at zero temperature, calculated within the path integral method (according to eqns [11a], [12], [13] of the paper (Devreese JT, De Sitter J and Goovaerts M (1972) Optical absorption of polarons in the Feynman–Hellwarth–Iddings–Platzman approximation. *Physical Review B* 5: 2367–2381)) for different values of the electron–phonon coupling constant α. The peaks are labeled in the picture according to their physical origin as discussed in the text. A δ-like central peak is schematically shown by a vertical line.

It is remarkable that from [7] the three different kinds of polaron excitations appear in the calculated optical absorption spectra for polarons at $\alpha \geq 5$:

- scattering states where, e.g., one real phonon is excited (the structure starting at $\omega = 1$);
- relaxed excited state (RES);
- Franck–Condon (FC) states.

Besides that, at zero temperature, the optical absorption spectrum for one polaron contains a central peak [$\sim\delta(\omega)$] at the origin. For nonzero temperature, this central peak smears out and the optical absorption spectrum consists of both a broad envelope and an anomalous Drude-type low-frequency component.

For example, in Figure 4 the main peak of optical absorption for $\alpha = 5$ at $\omega = 3.51$ is interpreted as due to transitions to RES. The shoulder at the low-frequency side of the main peak is attributed to one-phonon transitions to scattering states. The structure centered at about $\omega = 6.3$ is attributed to a FC band. As seen from Figure 4, when increasing the electron–phonon coupling constant to $\alpha = 6$, the RES peak at $\omega = 4.3$ becomes very intense compared with the FC peak centered at $\omega = 7.5$.

The free-carrier polaron effects were revealed through measurements of optical constants, of cyclotron resonance and of mobility in semiconductors and insulating photoconductors. Experimentally, only the scattering states have been seen for free polarons. However, the full structure of eqn [7] has been revealed through cyclotron resonance measurements.

Magneto-absorption of Polarons

Polaron Cyclotron Resonance

A polaron in a magnetic field is referred to as a *magnetopolaron*. The most powerful technique to study the electron–phonon coupling in semiconductors is cyclotron resonance. Changing the magnetic field B allows one to vary the cyclotron frequency $\omega_c = eB/m_b c$. When it is brought closer to resonance with the phonon frequency ω_{LO}, a strongly renormalized magnetopolaron state is formed. This results in a splitting of the cyclotron resonance, which is a measure of the electron–phonon (polaron) coupling constant α.

A theory of the polaron effects in cyclotron resonance was proposed by Larsen. In particular, the variational approach of Larsen is based on an intermediate-coupling theory to calculate the energy levels (modified Landau levels) of a polaron in a

magnetic field. The polaron mass is then defined from the energy differences between the polaron (Landau-) energy levels.

A rigorous way to find the polaron cyclotron mass is to calculate the magneto-optical absorption spectrum of the polaron (the quantity which is actually measured) and to define the polaron mass, in the same way as the experimentalists do, from the peak positions in the spectrum. The magneto-optical absorption of polarons for all α and ω_c at $T = 0$ was calculated by Peeters and Devreese. They extended the memory-function formalism described above to study the response of a polaron in a magnetic field. The magneto-absorption was then obtained from

$$\lim_{\varepsilon \to 0} \mathrm{Re} \frac{1}{z - \omega_c - \Sigma(z)} \qquad z = \omega + \mathrm{i}\varepsilon \qquad [14]$$

where $\omega_c = eB/m_b$ is the cyclotron frequency for a rigid-lattice band mass m_b, and $\Sigma(z)$ is the memory function, which takes into account all the polaron internal states as well as all the Landau levels and depends on the coupling constant α and on the magnetic field. The explicit results for $\Sigma(z)$ depend on the optimal parameters of the anisotropic Feynman polaron model, which are found by a variational calculation of the polaron ground-state energy.

At $T = 0$ and when $\omega < 1$, it follows that $\mathrm{Im}\Sigma(\omega) = 0$. According to eqn [14], the position of the cyclotron resonance line is determined by the equation $\omega - \omega_c - \Sigma(\omega) = 0$. If solutions of this equation are denoted as $\omega = \omega_c^*$ then the polaron cyclotron mass m_p^* is calculated from the relation $\omega_c^* = eB/m_p^*$.

Cyclotron Resonance of Polarons in Silver Halides

In pioneering experimental studies, Brown and coworkers have combined mobility experiments and cyclotron resonance measurements to clearly demonstrate the polaron effect in AgBr. From a theoretical plot of polaron mobility versus band mass compared to experimental Hall mobility data at a given temperature, they estimate the band mass. This allows them to calculate α and the polaron mass m^*. This value of m^* can then be compared to the measured cyclotron mass in order to test the polaron theory. For AgBr, the experimental value $m^*/m_e = 0.27 \pm 0.01$ obtained from the cyclotron resonance at millimeter wave frequencies and at 18 K is shown to agree well with the value $m^*/m_e = 0.27 \pm 0.05$ determined by comparison of mobility experiment and theory.

Precise cyclotron mass measurements in AgBr and AgCl covered the range from zero magnetic field to 16 T (see Figure 5). Several polaron theories were compared in analyzing these experimental data. It should be pointed out that the weak-coupling theories (Rayleigh–Schrödinger perturbation theory, Wigner–Brillouin and its improvements) fail (and are all off by at least 20% at 16 T) to describe the

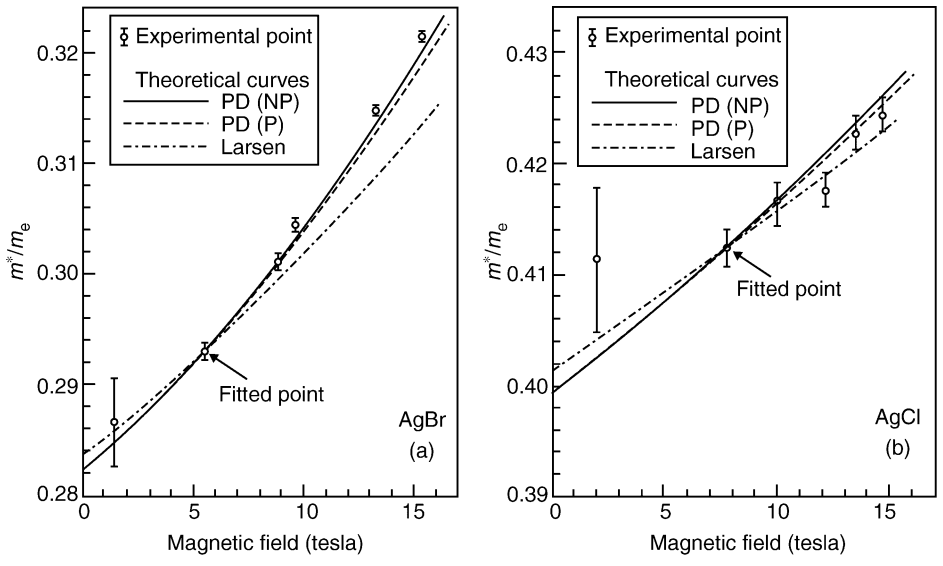


Figure 5 The polaron cyclotron mass in AgBr (a) and in AgCl (b): comparison of experiment and theory. Larsen: (Larsen D (1974) *Journal of Physics C* 7: 2877–2889); PD: (Peeters FM and Devreese JT (1986) *Physical Review B* 34: 7246–7259) (P) – with parabolic band, (NP) – with corrections of the two-band Kane model. In each case the band mass was adjusted to fit the experimental point at 525 GHz. (Reproduced with permission from Hodby JW, Russell G, Peeters F, Devreese JT and Larsen DM (1987) Cyclotron resonance of polarons in the silver halides. *Physical Review Letters* 58: 1471–1474). ©(1987) American Physical Society.

experimental data for the silver halides. The variational approach by Larsen underestimates the polaron cyclotron mass by 2% at 15.3 T. It turns out that the magneto-absorption calculated by Peeters and Devreese leads to the best quantitative agreement between theory and experiment as was analyzed for AgBr and AgCl. This analysis provides a clear confirmation of the Fröhlich description of the polaron in the case where weak-coupling approximations are adequate.

Cyclotron Resonance of Polarons in CdTe

The early infrared transmission study of hydrogen-like shallow donor impurity states in *n*-CdTe was reported by Cohn, Larsen and Lax (see Figure 6). By studying the Zeeman splitting of the $(1s \rightarrow 2p, m = \pm 1)$ transition in the Faraday configuration at magnetic fields up to ~160 kOe, they performed the first quantitative determination of polaron shifts of the energy levels of a bound electron. The experimental data were shown to be in good agreement with the weak-coupling theory of the polaron Zeeman effect. In this comparison, however, the value $\alpha = 0.4$ was used instead of $\alpha = 0.286$, which comes from the definition [1]. Similarly, the value $\alpha \sim 0.4$ was suggested to explain the measured variation of the cyclotron mass with magnetic field in CdTe. This discrepancy gave rise to a discussion where important contributions have been, in particular, due to Harper and Zawadzki.

Grynberg *et al.* applied the far-infrared photoconductivity technique to study the energy spectrum of shallow In donors in CdTe layers and obtained experimental data over the energy range relevant to study the magnetopolaron effect, as shown in Figure 7. An overall good agreement is found between experiment and the theoretical approach, where the electron–phonon interaction is treated within a second-order improved Wigner–Brillouin perturbation theory and a variational calculation is performed for the lowest-lying donor states ($1s$, $2p^{\pm}$, $2s$, $2p_z$, $3d^{-2}$, $4f^{-3}$). It is to be stressed that this agreement is achieved with the coupling constant $\alpha = 0.286$ as follows from eqn [1] with the known material parameters of CdTe.

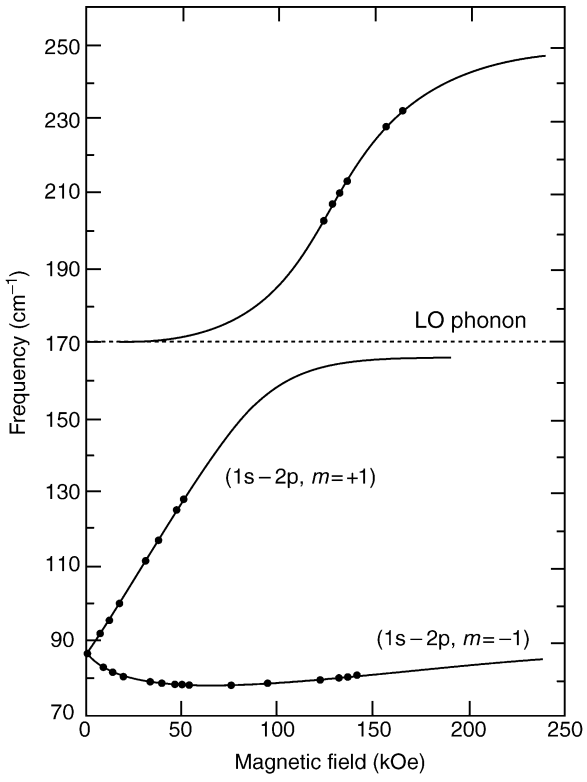


Figure 6 Plot of the experimentally determined magnetic field dependence of the $(1s \rightarrow 2p, m = \pm 1)$ transition frequencies in n-CdTe. The solid lines represent the theoretical results for the effects of the electron–LO-phonon interaction upon the shallow donor impurity levels. α treated as a fitting parameter is taken to be 0.4. (Reproduced with permission from Cohn DR, Larsen DM and Lax B (1972) Polaron Zeeman effect in CdTe. *Physical Review B* 6: 1367–1374). ©(1972) American Physical Society.

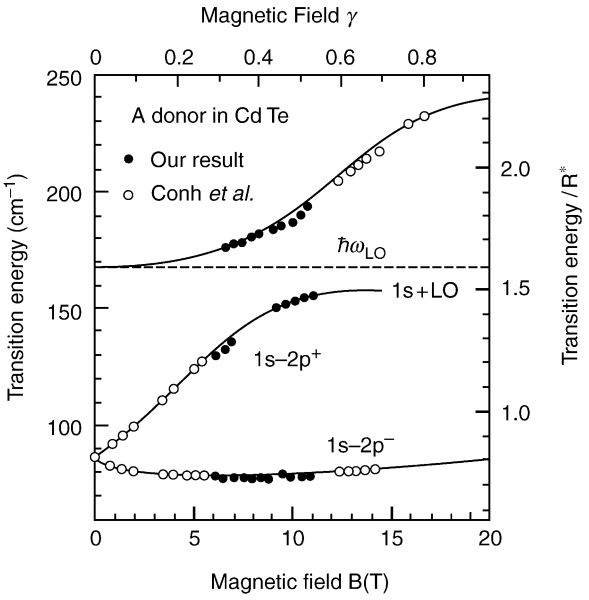


Figure 7 Plot of the experimentally determined magnetic field dependence of the $1s \rightarrow 2p^{\pm 1}$ transition energies in CdTe layers grown by molecular beam epitaxy. The solid lines represent the results of the calculation described in the text without any fitting parameters. The solid dots are the experimental data of the present work and the open circles represent the data of Cohn, Larsen and Lax. (Reproduced with permission from Grynberg M, Huant S, Martinez G, *et al.* (1996) Magnetopolaron effect on shadow indium donors in CdTe. *Physical Review B* 54: 1467–1470). ©(1996) American Physical Society.

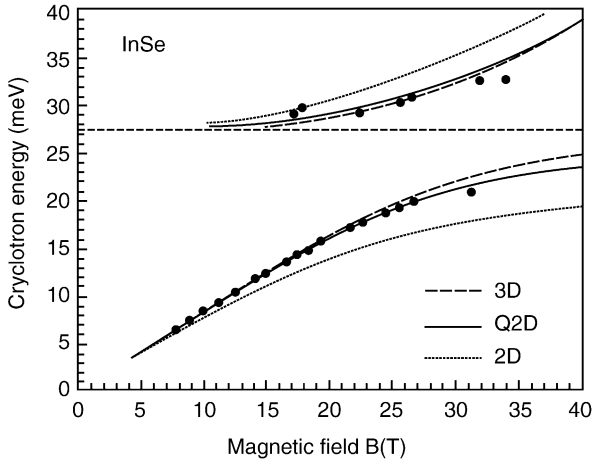


Figure 8 The cyclotron resonance position plotted as a function of magnetic field for InSe. (Reproduced with permission from Nicholas RJ, Watts M, Howell DF, *et al.* (1992) Cyclotron resonance of both magnetopolaron branches for polar and neutral. *Physical Review B* 45: 12144–12147). ©(1992) American Physical Society.

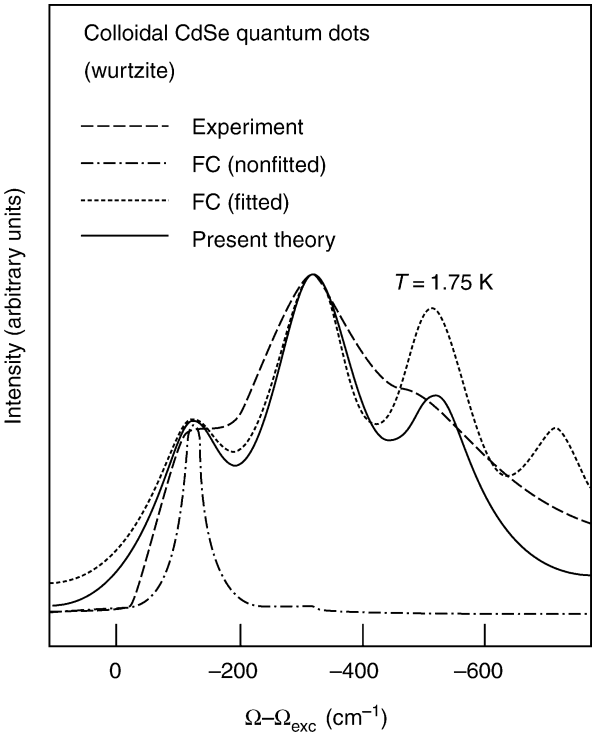


Figure 9 Fluorescence spectra of CdSe quantum dots with wurtzite structure at the average radius $\langle R \rangle = 1.25$ nm. The dashed line represents the experimental data of Nirmal M, Murray CB, Norris DJ and Bawendi MG (1993) *Z. Phys. D* 26: 361–363, the dot-dashed line displays a Franck–Condon progression with the Huang–Rhys parameter $S = 0.06$ calculated using the adiabatic approximation, the dotted line shows another Franck–Condon progression with the Huang–Rhys parameter $S = 1.7$, which is obtained by fitting the ratio of one-phonon and zero-phonon peak heights to the experimental value, and the solid line results from the nonadiabatic theory. (Reproduced with permission from Fomin VM, Gladilin VN, Devreese JT, *et al.* (1988) Photoluminescence of spherical quantum dots. *Physical Reviews B* 57: 2415–2425). ©(1998) American Physical Society.

Polarons in the Layered Compound InSe and in Superlattices

Nicholas *et al.* provided a clear demonstration of the polaron coupling by the cyclotron resonance in a two-dimensional electron gas (2DEG), which naturally occurs in the polar semiconductor InSe. One clearly sees, over a wide range of magnetic fields ($B = 18-34$ T), the two distinct magnetopolaron branches separated by as much as 11 meV ($\sim 0.4\omega_{LO}$) at resonance (Figure 8). The lines show the results of theoretical calculations for coupling to the LO phonon in bulk (3D), sheet (2D) and after correction for the quasi-2D systems at $\alpha = 0.29$ (for motion perpendicular to the c-axis). The agreement is reasonable for the 3D case, and is better for the quasi-2D system, where the finite spatial extent of the 2D electron gas in the symmetric planar layer is taken into account.

For GaAs/Al_xGa_{1-x}As quantum wells and superlattices, the polaron effect is found to decrease the energy of the shallow donor states at low magnetic fields and to lead to a resonant splitting of the energies at high magnetic fields. The results are in very good agreement with available experimental far-infrared optical data in the whole magnetic-field range.

Optical Properties of Quantum Dots: Effects of the Polaron Interaction

In order to interpret the phonon-assisted optical transitions in semiconductor quantum dots, a theory was developed by Fomin, Gladilin, Devreese *et al.* comprising the exciton interaction with both adiabatic and Jahn–Teller phonons, and also the external nonadiabaticity (pseudo-Jahn–Teller effect). The effects of nonadiabaticity of the exciton–phonon system are shown to lead to a significant enhancement of phonon-assisted transition probabilities and to multiphonon optical spectra which are considerably different from the Franck–Condon progression (see Figure 9). The calculated relative intensity of the phonon satellites and its temperature dependence compare well with the experimental data on the photoluminescence of CdSe quantum dots, both colloidal and embedded in glass. These results clearly demonstrate a breakdown of the adiabatic approximation when describing the polaron interaction in quantum dots.

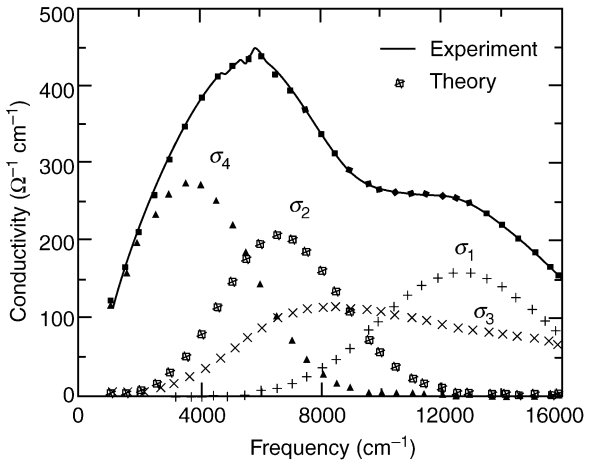


Figure 10 Experimental and theoretical conductivity data at 250 K for $Pr_2NiO_{4.22}$ and four contributions to the theoretical curves resulting from the transitions between the following polaron states: σ_1 – small → small; σ_2 – large → small; σ_3 – small → large; σ_4 – large → large. (Reproduced with permission from Eagles DM, Lobo RPSM and Gervais F (1995) Infrared absorption in oxides in the presence of both large and small polarons. *Physical Review B* 52: 6440–6450). ©(1995) American Physical Society.

Small Polarons

An electron or a hole trapped by its self-induced atomic (ionic) displacement field in a region of linear dimension, which is of the order of the lattice constant, is called a 'small polaron'. As distinct from large polarons, small polarons appear due to short-range forces.

Experimentally small-polaron effects have been analyzed, e.g., in KCl, LiF, NiO, MnO, TiO_2, $BaTiO_3$, $SrTiO_3$, $LaCoO_3$. More recently, Alexandrov and Mott (1996) surveyed both the principles and the main results of the small-polaron theory in the context of the (bi)polaronic approach to describe the physics of high-T_c superconductors.

Spectroscopic manifestations of both large and small polarons have been found by Eagles *et al.* in the infrared optical absorption spectra of $Pr_2NiO_{4.22}$. The infrared reflectivity spectra on a crystal of this material were measured between 30 and 20 000 cm^{-1}. Mid-infrared conductivity data for frequencies between 1000 and 16 000 cm^{-1} were obtained by Kramers–Kronig transformation of those reflectivity spectra. The resulting optical absorption spectrum was quantitatively interpreted (see Figure 10) using the assumption that in $Pr_2NiO_{4.22}$ two types of polarons, small and large, coexist but do not mix with each other. The contribution due to the transitions between the large-polaron states, which dominates the absorption in the low-frequency region, was calculated using the arbitrary coupling theory by Devreese *et al.* for the coupling constant $\alpha = 3$.

In summary, the polaron concept has been proven to be one of the corner-stones of modern solid-state optics.

Acknowledgments

This work has been supported by the BOF NOI (UA-UIA), GOA BOF UA 2000, IUAP, FWO-V projects G.0287.95, 9.0193.97, G.0274.01 and the W.O.G. WO.025.99N (Belgium).

See also

Semiconductor Materials: Quantum Dots.

Further Reading

Alexandrov AS and Mott N (1996) *Polarons and Bipolarons*. Singapore: World Scientific.

Devreese JT (ed.) (1972) *Polarons in Ionic Crystals and Polar Semiconductors*. Amsterdam: The Netherlands North-Holland.

Devreese JT (1996) Polarons. In: Trigg GL (ed.) *Encyclopedia of Applied Physics*, vol. 14, pp. 383–413. Weinheim: VCH.

Devreese JT, De Sitter J and Goovaerts M (1972) Optical absorption of polarons in the Feynman–Hellwarth–Iddings–Platzman approximation. *Physical Review B* 5: 2367–2381.

Feynman RP (1955) Slow electrons in a polar crystal. *Physical Review* 97: 660–665.

Kartheuser E, Evrard R and Devreese J (1969) Mechanism of absorption of light by free continuum polarons. *Physical Review Letters* 22: 94–97.

Kuper CG and Whitfield GD (eds) (1963) *Polarons and Excitons*. Edinburgh, UK: Oliver and Boyd.

Landwehr G and Rashba EI (eds) (1991) *Landau Level Spectroscopy*, 2 vols. Amsterdam, The Netherlands: North-Holland.

Peeters FM and Devreese JT (1986) Magneto-optical absorption of polarons. *Physical Review B* 34: 7246–7259.

Pekar SI (1951) *Issledovanija po Ekektronnoj Teorii Kristallov*. Moscow: Gostekhizdat (in Russian); German translation: (1954) *Untersuchungen über die Elektronentheorie der Kristalle*. Berlin: Akademie Verlag.

Quantum Wells and GaAs-Based Structures

P Blood, Cardiff University, Cardiff, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A quantum well is a potential minimum within a semiconductor structure which is sufficiently thin to localize charge carriers on a length-scale similar to their de Broglie wavelength which, for an electron in GaAs at room temperature, is about 30 nm. When electrons are localized in this way their electronic and optical properties are determined by quantum mechanical aspects of their behavior which are not apparent in larger-scale structures. The potential well is usually formed by a sandwich of a thin layer of narrow-gap semiconductor between two layers of a wider gap material as depicted in Figure 1. Typical quantum wells have widths of about 5 nm and the key to their routine production has been the development of advanced epitaxial semiconductor crystal growth techniques such as metalorganic vapor phase epitaxy (MOVPE) and molecular beam epitaxy (MBE). The most significant application of these structures has been in opto-electronic devices, especially laser diodes, and in this article I describe how electrons behave in quantum wells with particular reference to optical properties.

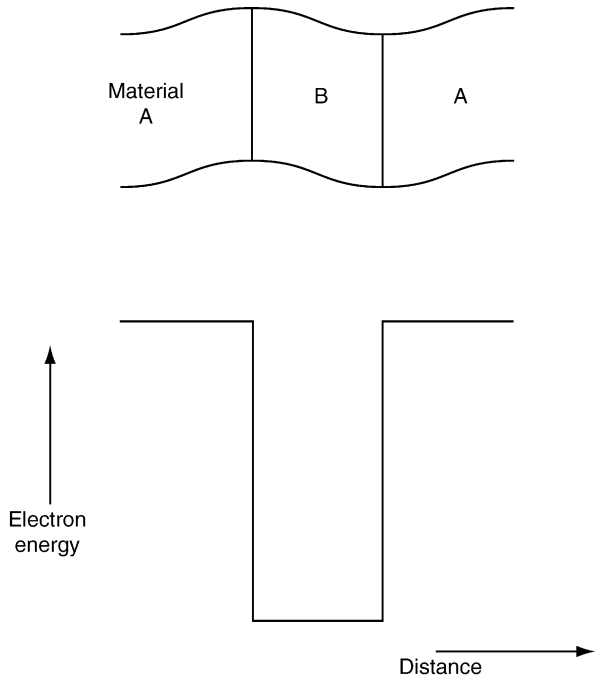


Figure 1 Electron energy diagram illustrating the formation of a potential well by sandwiching a layer of narrow gap semiconductor material (B) between two wider gap layers (A).

The optical properties of semiconductors are usually considered in terms of transitions of electrons between quantum mechanical energy levels, as in the Bohr atom. An alternative description is to consider the spatial displacement of electrons in the oscillating electric field of the electromagnetic radiation. These two descriptions are linked through Schrödinger's equation. For each energy state there is an associated wavefunction which determines the spatial probability distribution of electrons in the state. Consequently, a change in quantum state implies a change in spatial distribution and energy of the electron system. Both approaches are used: here I adopt the viewpoint of transitions between energy states.

We begin with a quantum mechanical description of electrons in a 'bulk' material, as appears in many textbooks. Here electrons are constrained by a three-dimensional potential well of extent given by the sample dimensions, typically millimeters or centimeters and therefore large compared with the de Broglie wavelength. Then we examine what happens when the size of the sample is reduced in one direction: in the limiting case of a quantum well electrons are able to move in only two directions. Quantum confinement is also possible in two or three dimensions to produce a quantum wire or quantum dot, respectively, and it is easy to use the concepts developed in this article to determine the properties of such systems.

Electron States in Bulk Material

We first consider the behavior of an electron in one direction (the x-direction) within a potential well with large dimensions (side length L), then generalize this to three directions (see Figure 2). The wavefunction ψ_n and energy E_n (measured with respect to the bottom of the well) are given by solutions of Schrödinger's equation, which within the well is:

$$\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi_n}{\mathrm{d}x^2} = E_n \qquad [1]$$

It is assumed that the potential is infinitely deep compared with the energy of the electrons.

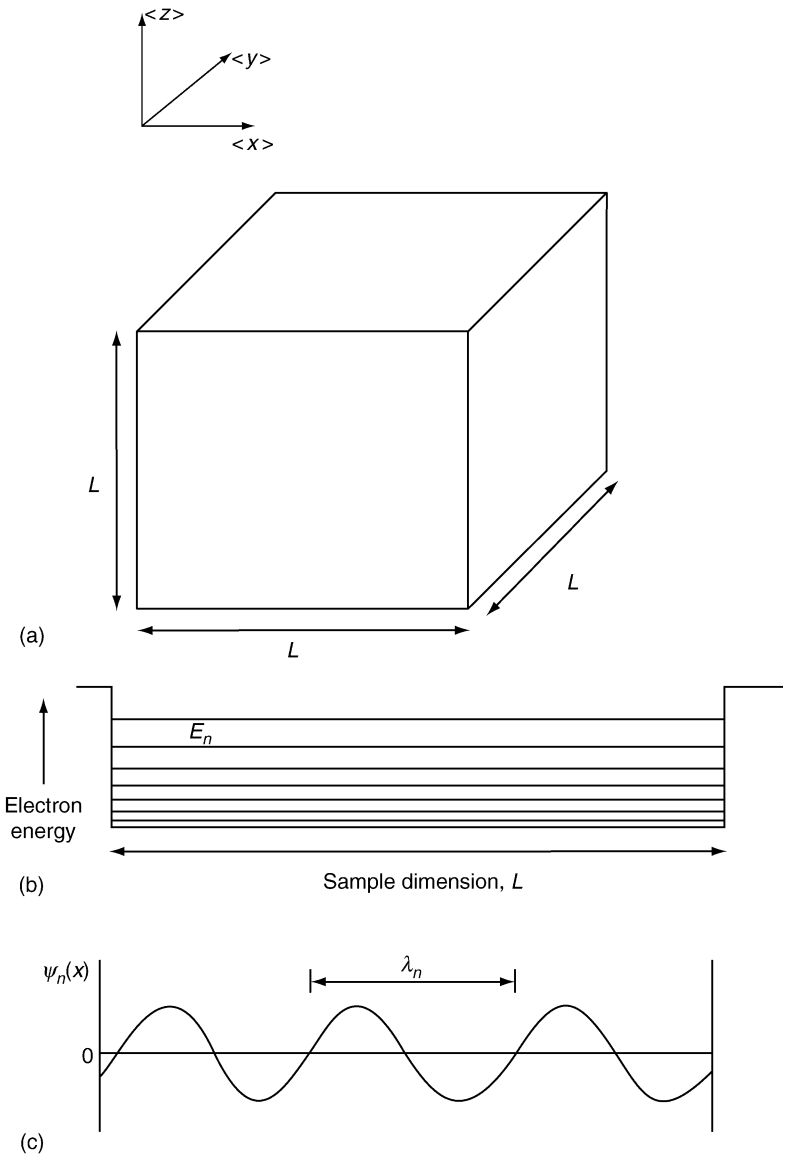


Figure 2 The properties of electrons in a large cubic sample (a) can be obtained by solving Schrödinger's equation for the potential well formed by the sample in each direction (b). This potential well keeps the electrons within the material. The wavefunctions $\psi(x)$ shown in part (c) satisfy cyclic boundary conditions as described in the text, leading to a series of electron energy levels, E_n, given by eqn [3], shown in (b) of the figure.

Within the sample the electrons are described by plane waves of the form $\psi_n = A_n \sin(k_n x)$, where k_n is a wavevector ($= 2\pi/\lambda_n$, where λ_n is the wavelength), and substitution into eqn [1] gives the corresponding energy eigenvalues as

$$E_n = \frac{\hbar^2 k_n^2}{2m} \qquad [2]$$

where m is the mass of the electron. [Substituting $E_n = k_B T$ at room temperature (0.025 eV) and using an effective mass of $0.067m_0$ (for GaAs) in eqn [2] gives $k_n = 2.1 \times 10^8$ m^{-1} and $\lambda = 30$ nm.] Applying cyclic boundary conditions which permit traveling-wave solutions of eqn [1] the wavelength must satisfy the condition $\lambda_n = nL$, i.e., $k_n = 2\pi n/L$, as illustrated in Figure 2, where n is an integer. The dimensions of a typical sample are much greater than the 'size' of an electron so n can be a very large number. The energy eigenvalues are therefore given by

$$E_n = \frac{\hbar^2}{2m}\left(\frac{2\pi n}{L}\right)^2 \qquad [3]$$

This treatment can be extended by solving eqn [1] for motion in three orthogonal directions, x, y, and z. In this case electron motion is described by a wavevector $\mathbf{k}$ comprising components k_x, k_y, k_z, each of which satisfies the cyclic boundary condition in the respective direction. We take the sample to be of dimension L in each direction with no loss of generality. Thus (representing unit vectors by $\langle\mathbf{x}\rangle$, etc.)

$$\begin{aligned}\mathbf{k} &= k_x\langle\mathbf{x}\rangle + k_y\langle\mathbf{y}\rangle + k_z\langle\mathbf{z}\rangle \\ &= \frac{2\pi n_x}{L}\langle\mathbf{x}\rangle + \frac{2\pi n_y}{L}\langle\mathbf{y}\rangle + \frac{2\pi n_z}{L}\langle\mathbf{z}\rangle \end{aligned} \quad [4]$$

the energy is

$$E = \frac{\hbar^2}{2m}\left\{k_x^2 + k_y^2 + k_z^2\right\} \quad [5]$$

and the wavefunction takes the vector form

$$\psi_k(\mathbf{r}) = A_k \sin(\mathbf{k}\cdot\mathbf{r}) \quad [6]$$

$\psi_n^2(\mathbf{r})$ represents the probability of an electron being at the location $\mathbf{r}$. Each allowed electron state obtained by solution of Schrödinger's equation is specified by a unique combination of the numbers $(n_x\, n_y\, n_z)$, which take both positive and negative values corresponding to plane waves traveling in positive and negative directions. The Pauli exclusion principle states that only one electron of each spin can occupy a given quantum state so the amplitude of each wavefunction is normalized such that

$$\int \psi_k^2(\mathbf{r})\, \mathrm{d}\mathbf{r} = \int [A_k \sin(\mathbf{k}\cdot\mathbf{r})]^2\, \mathrm{d}\mathbf{r} = 1 \quad [7]$$

where the integrals are evaluated over the volume of the sample. Since the amplitude of these solutions is constant throughout the sample electrons may be anywhere in the sample with equal probability. The motion of an *individual* electron of momentum $\hbar\mathbf{k}$ is represented by a wavepacket formed by combination of a number of wavefunctions having similar values of $\mathbf{k}$.

Because the sample dimensions are large the energies of the states allowed by Schrödinger's equation take a series of very closely spaced values for increasing integer values of n as shown in Figure 2b. Substituting a typical sample size of $L = 1$ mm into eqn [3], the $n = 1$ and 2 states are separated by only 7×10^{-11} eV in GaAs and these values are very small compared with thermal energies: at room temperature, $k_B T = 0.025$ eV. The allowed values of $\mathbf{k}$ and E are discrete and there is a finite but very large number of allowed states in the sample. Since these are very closely spaced we can regard $\mathbf{k}$ and E as continuous variables and the allowed energy states form a continuum. We can calculate the number of allowed energy states $\mathrm{d}N$ in a small energy interval, $\mathrm{d}E$, and hence determine the density of states in energy $g = (\mathrm{d}N/\mathrm{d}E)$ (the number per unit energy interval) at any value of energy.

Electron States in Quantum Wells

Now imagine a sample where the length of one side is reduced in the z-direction (L_z), as illustrated in Figure 3. Gradually the allowed values of k_z become more widely spaced as a consequence of the boundary conditions (eqn [4]). Eventually the width of the potential well becomes similar to the wavelength of the plane-wave state corresponding to the lowest energy level in the well. In these circumstances electron motion is not possible in the z-direction and the wavefunctions become standing waves with boundary conditions $n(\lambda_n/2) = L_z$ in an infinitely deep potential (where n_z takes only positive values) and the amplitude of the wavefunction outside the well must be zero. The lowest energy state, given by eqn [5] with $n_x = n_y = n_z = 1$, no longer lies near the bottom of the potential well. Since the energies associated with motion in the x and y directions remain very small (of order 10^{-10} eV derived earlier) because L_x and L_y are large, the energy of the lowest state is effectively determined by the z-dimension because it is very small, so eqn [5] gives

$$E_{n_z} = \frac{\hbar^2}{2m}\left(\frac{n_z\pi}{L_z}\right)^2 \quad [8]$$

for the lowest energy state in the well. For $L_z = 5$ nm (and $m = 0.067$) E_1 is 0.22 eV above the bottom of the well. Equation [8] shows that this energy can be changed by choice of the well thickness. A simple interpretation of this behavior is as follows. When L_z becomes very small it is no longer possible for the z-component of the wavefunction of the lowest energy state to satisfy the boundary condition of zero amplitude at opposite sides of the sample. The condition can only be satisfied by wavefunctions which have a smaller wavelength, and consequently a higher energy.

Electrons can occupy states defined by all values of $(n_x\, n_y\, n_z)$, thus for $n_z = 1$ there is a continuum of allowed states at increasing energies corresponding to increasing values of $(n_x\, n_y)$ and corresponding to motion in the (x, y) plane (Figure 4a). Since L_x and L_y are both large these states are closely spaced and form a continuum. There is a similar continuum of energy

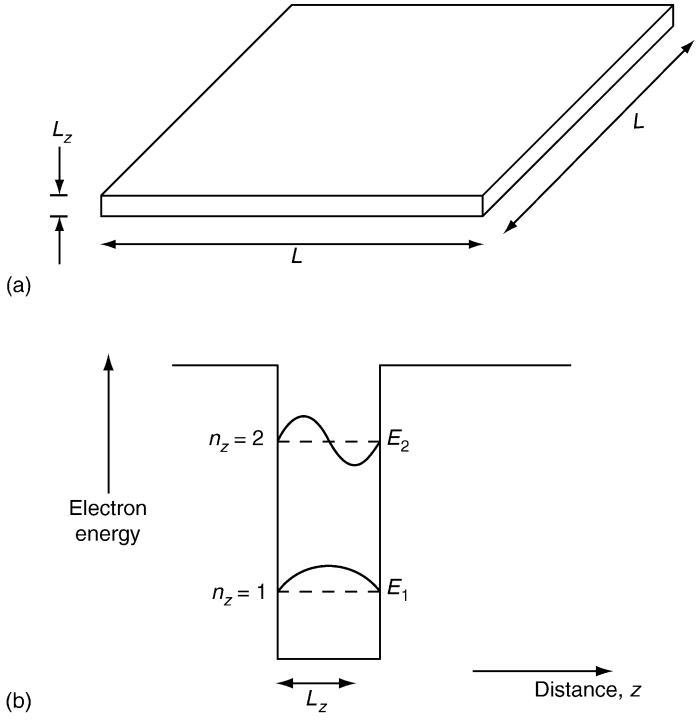


Figure 3 Electron energy diagram (b) for a sample (a) in which one dimension, L_z, is very small. The electron motion in the z-direction is constrained and the associated energy levels become widely spaced, defined by the integers $n_z = 1, 2, \ldots$ in eqn [8]. The wavefunctions in the z-direction, $\psi(z)$, are also illustrated for the first two electron levels.

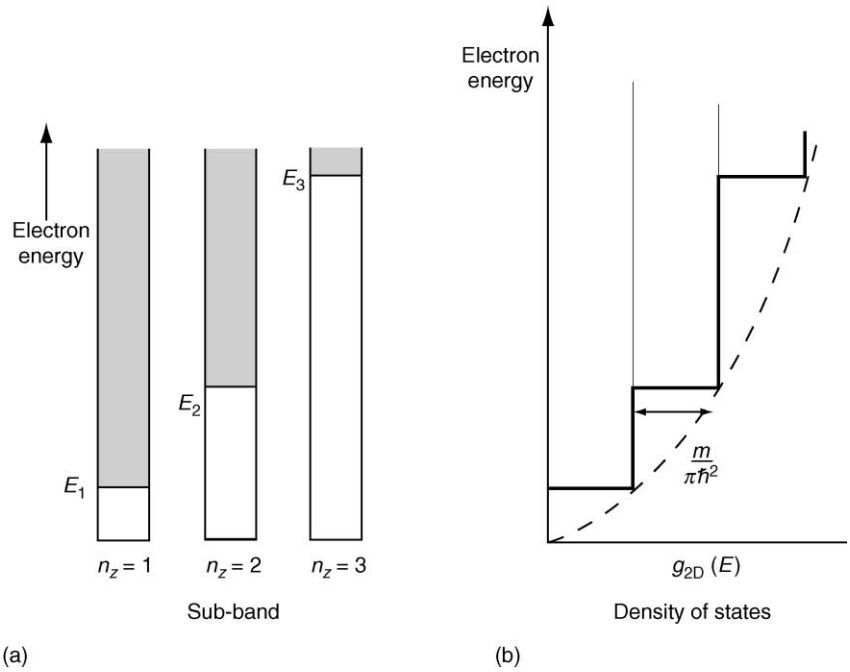


Figure 4 (a) Electron energy diagram showing the energy states associated with unconstrained motion in the (x, y) plane for each sub-band formed by localizing the electrons in the z-direction, defined by $n_z = 1, 2, 3$, etc. When all these allowed states are summed at any energy and expressed as a number of states per unit energy interval we obtain the density of states function $g_{2D}(E)$ shown in (b). This has a series of steps corresponding to each sub-band edge.

states above the energy levels for $n_z = 2, 3$, etc. Thus the energy level diagram is a series of sub-bands, each defined by n_z and with a continuum of states associated with motion in two dimensions in the (x, y) plane, as shown in Figure 4. The number of continuum energy states in a given sub-band in a small energy interval gives the density of states for both spin directions

$$g_{2D}(E) = \frac{m}{\pi\hbar^2} \qquad [9]$$

per unit energy interval per unit sample area.

The density of states is independent of n_z and therefore the same for all sub-bands, and is independent of energy. The density of states function is therefore a series of steps of height given by eqn [9] at each sub-band energy as illustrated in Figure 4b. The density of states per unit area within a given sub-band is independent of L_z as a consequence of dealing with a two-dimensional system.

In real samples the quantum well is not infinitely deep: typically the well depth is in the range 100–400 meV, consequently electrons are not totally confined to the well but are able to penetrate into the barrier material by quantum mechanical tunneling. Solution of Schrödinger's equation for a finite well (treated in many quantum mechanics textbooks) again yields a series of sub-bands, finite in number and with an energy spacing smaller than that given by eqn [8]. In a typical GaAs well the lowest energy state may be about 100 meV from the bottom of the well. The density of states in each sub-band, which is due to the (x, y) motion, remains unchanged. The wavefunction in the z-direction $F(z)$ comprises a sine wave-like standing wave within the well and an exponentially decaying wave in the barrier representing tunneling (Figure 5). The wavefunction is therefore made up of (x, y) plane-wave components and the localized z component and is of the form

$$\Psi_k(\mathbf{r}) = A_k \sin(\mathbf{k}_{xy} \cdot \mathbf{r}_{xy}) F(z) \qquad [10]$$

where $\mathbf{k}_{xy}$ and $\mathbf{r}_{xy}$ are the wavevector and position vector in the (x, y) plane. This wavefunction is again normalized according to eqn [7] such that each state is occupied by one electron of a given spin direction. We see from Figure 5 that we cannot regard the electron as being localized in the well: the electron distribution in the z-direction is specified by the probability distribution $F^2(z)$.

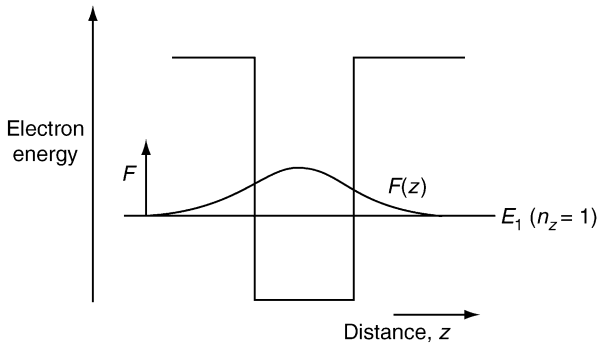


Figure 5 Illustration of the envelope wavefunction $F(z)$ in the z-direction for an electron localized in the $n_z = 1$ sub-band of a well of finite depth. The electron is able to penetrate the barrier by tunneling and this is represented by the decaying part of the wavefunction outside the well.

Occupancy of States in Quantum Wells

The density of states function describes the energy distribution of allowed states in the quantum well. Whether or not particular states are occupied by electrons is determined by the electron concentration and the temperature through the thermodynamic properties of the system. In most cases electron–electron scattering is sufficiently rapid to bring the electron population into an internal equilibrium. Furthermore the interaction between the electrons and the crystal lattice is able to redistribute energy between the two systems so that the lattice and the electron distribution have the same temperature. The electron distribution can therefore be described by Fermi–Dirac statistics for which the probability of a state at energy E being occupied with an electron is

$$f(E) = \frac{1}{1 + \exp\left\{\dfrac{E - E_f}{k_B T}\right\}} \qquad [11]$$

where E_f is the chemical potential (which can be equated with the Fermi level for electrons), T is the lattice temperature and all energies are positive quantities measured with respect to the same arbitrary zero. The energy distribution of electrons in the well, $n(E)$, is then the product of the density of states function, $g_{2D}(E)$, and the occupation probability, $f(E)$, as depicted in Figure 6. The total number of electrons, n, is the integral over energy. Combining eqns [9] and [11], for a single sub-band gives:

$$n_a = \frac{mk_B T}{\pi\hbar^2} \ell n\left\{1 + \exp\left(-\frac{E_n - E_f}{k_B T}\right)\right\} \qquad [12]$$

per unit area, single sub-band.

We cannot specify the carrier density in the well 'per unit volume' since the extent of the distribution in the z-direction is not defined because electrons

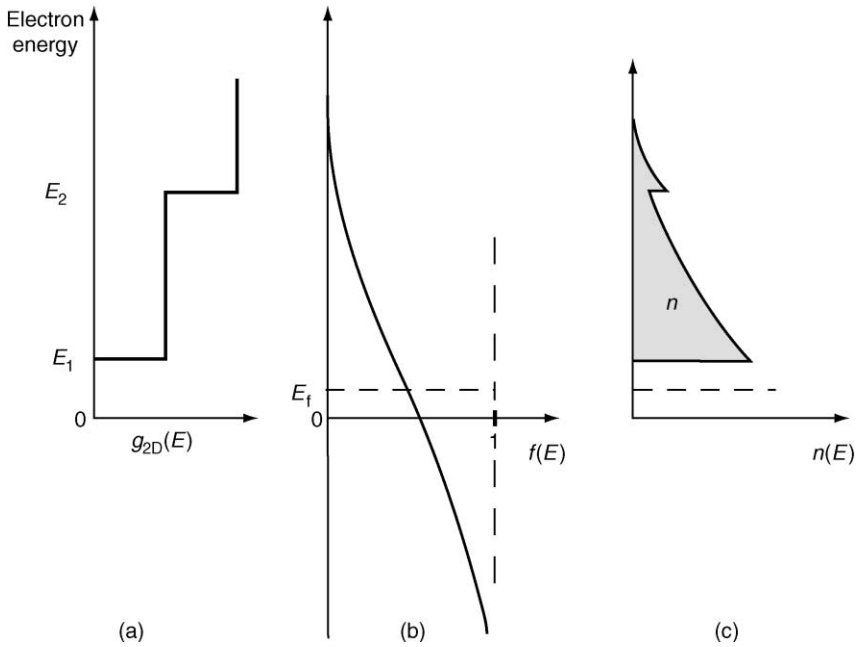


Figure 6 (a) The density of states function for a quantum well, with the value of $g_{2D}(E)$ on the horizontal axis indicating how the density of states varies with increasing electron energy. (b) is an illustration of the probability that a state is occupied by an electron, showing how $f(E)$ varies with increasing electron energy, given by eqn [11]. The actual density of electrons in a given energy interval at any energy $n(E)$ is given by the product of the density of available states and the probability that the state is occupied as shown in (c). The integral of $n(E)$ over energy gives the total density of electrons given for each sub-band by eqn [12].

may tunnel into the barrier material. The electron distribution is properly specified as a number per unit area having a probability distribution in the z-direction given by the function $F(z)$. Where more than one sub-band is populated the total electron density is obtained by adding the contributions from each sub-band each given by eqn [12] with the same Fermi energy and the appropriate sub-band energy. We always use Fermi factors to specify the probability of occupation of a state by an *electron*. The probability that a state is empty is $(1 - f)$.

Formation of Quantum Wells

A quantum well is formed by sandwiching a narrow gap semiconductor layer (E_{g1}) between two layers of wider gap material (E_{g2}). To avoid the formation of defects and dislocations, which have a deleterious effect on many properties of the structure, the constituent materials must have the same lattice parameter. Figure 7a is an energy band diagram for an n-type double heterostructure with layers of micron dimensions. At each interface there are discontinuities in the conduction and valence bands, ΔE_c and ΔE_v, respectively, which account for the difference in the band gaps of the two materials, ΔE_g. While the value of ΔE_g is known for any system, the manner in which this difference is apportioned between the two band edges is not known a priori and the values of ΔE_c and ΔE_v must be obtained from experiment. The discontinuities are often expressed as fractions Q_c, Q_v of the band gap difference:

$$\Delta E_c = Q_c \Delta E_g, \qquad \Delta E_v = Q_v \Delta E_g \qquad [13]$$

In equilibrium the Fermi energy is constant across the diagram and there is a band-bending region on each side of the discontinuity such that the conduction band edges return to their equilibrium energy values at large distances from the interface. The extent of these regions is determined by the discontinuity and the doping level and is typically 200–1000 nm. At each single interface a triangular well is formed which localizes electrons and whose precise shape is determined by the doping density, the carrier distribution and discontinuity through simultaneous solution of Schrödinger's equation and Poisson's equation. Single heterointerfaces only localize the majority carrier. Quantum confinement of both types of carrier can be achieved by a very thin narrow gap semiconductor layer.

When the layer of narrow gap material is made thinner than the band-bending regions, the

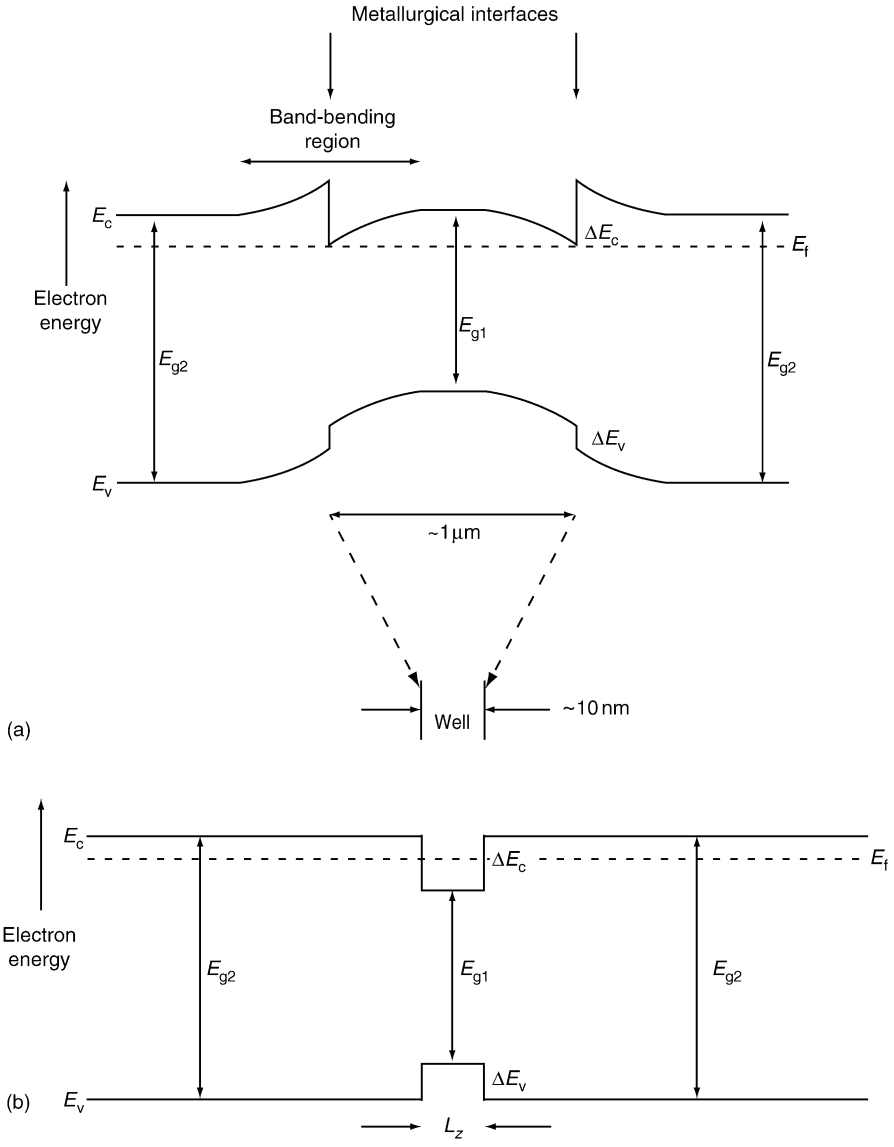


Figure 7 Electron energy diagrams showing the formation of a quantum well as a very thin double heterostructure made up of materials of band gaps E_{g1} and E_{g2}. When the layers are thick the energy diagram is influenced by band bending at the interfaces as shown in (a). When the narrow gap layer is made thinner than the band-bending distances a rectangular well is formed in the conduction and valence bands with depths determined by the respective band offset as shown in (b). On the small distance scale of the well the bands are flat in each region of the structure.

conduction band edge in this layer is not able to regain its bulk equilibrium position and the two triangular wells coalesce as illustrated in Figure 7b. When the well is sufficiently thin, say L_z less than 20 nm, the band-bending across this layer is negligibly small and the well becomes effectively rectangular. This well accumulates electrons from the surrounding barrier material and its potential relative to the Fermi level rises such that there is band bending in the barrier material on each side of the well inhibiting further accumulation. Since the band bending in the barrier occurs on a distance of hundreds of nanometers (being determined by the doping density) it is not apparent over the distance scale of the diagram of the quantum well, which is on the order of 10 nm. Consequently, as shown in Figure 7b, the well can be drawn as a rectangle to a good degree of approximation. This thin, narrow gap layer localizes both electrons and holes. The form of the well is determined by the respective band discontinuities, ΔE_c and ΔE_v, and the thickness of the layer.

Potential wells formed in this way are the basis for most quantum-confined device structures.

Other forms are possible. In certain material combinations ΔE_c or ΔE_v may be negative, meaning that, at the interface, both discontinuities are in the same sense and only one carrier type is confined in the thin narrow gap material. These are known as 'type II' structures. If there is an electric field across the well due to doping or strain (piezoelectric) effects the well becomes triangular. Wells can also be engineered with steps in the potential profile and multiple well systems can be grown in which the states are quantum mechanically coupled. All these variants provide opportunities to engineer the properties of the structure.

One important development of the rectangular quantum well deserves mention. It is possible for the well material to have a different lattice parameter to that of the surrounding barrier material provided that the strain energy can be accommodated elastically within the structure. This means that the strain energy in the layer must be below the energy necessary for formation of dislocations and this translates into an upper limit on the layer thickness (the critical thickness) for a given mismatch. Incorporation of elastic strain is significant because relaxing the lattice match requirement widens the choice of well and barrier materials and because strain modifies the properties of the electronic states in the quantum well and these effects can be used to advantage in the design of devices.

Here we concentrate on the rectangular potential well. Such wells are widely used, particularly in opto-electronic devices, and this structure provides the basic concepts which lie at the heart of all quantum-confined systems. We consider the 'model' lattice-matched quantum well system: a GaAs well bounded by AlGaAs barriers.

GaAs/AlGaAs Quantum Well Structures

As Al is substituted for Ga in the $Al_xGa_{1-x}As$ alloy system the direct band gap increases from 1.424 eV in GaAs ($x = 0$) to 3.018 eV in AlAs ($x = 1$) (Figure 8) while the lattice parameter remains unchanged. A quantum well is formed using a narrow GaAs layer sandwiched between barrier layers of $Al_xGa_{1-x}As$ having a composition chosen to give the desired well depth. The whole structure is grown on a GaAs substrate. The band discontinuity ratio between GaAs and $Al_xGa_{1-x}As$ is independent of alloy composition (ΔE_c and ΔE_v are constant fractions of ΔE_g as x is varied) with $Q_c = 0.66$ and $Q_v = 0.33$. For a typical barrier composition of $x = 0.3$,$\Delta E_g = 0.374$ eV and ΔE_c and ΔE_v are 0.247 eV and 0.127 eV, respectively.

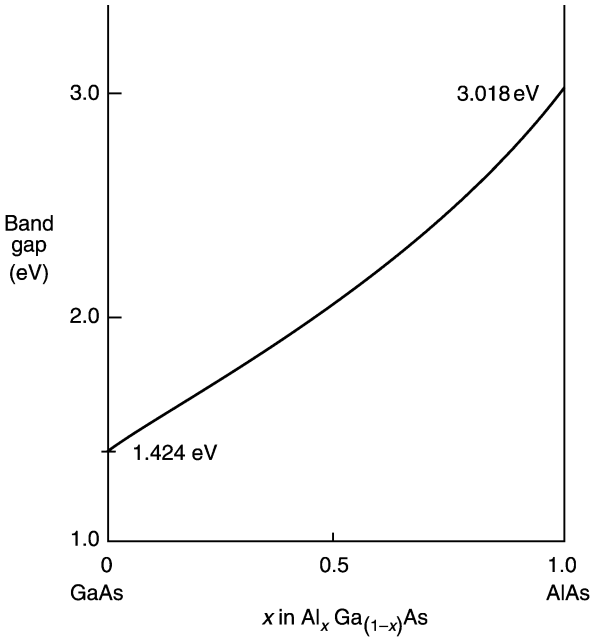


Figure 8 Variation of the direct band gap of $Al_xGa_{(1-x)}As$ with Al content (x).

Optical Properties of AlGaAs/GaAs Quantum Wells

General Principles

Figure 9 illustrates a transition of an electron between an occupied state at E_1 in the $n_z = 1$ sub-band in the conduction band and an empty state (i.e., a hole) at E_2 in the $n_z = 1$ sub-band in the valence band, resulting in emission of a photon of energy $h\nu = E_1 - E_2$ as required by energy conservation. This is the process of luminescence (or spontaneous emission), which requires external excitation such as illumination (photoluminescence) or biasing a p-n junction (electroluminescence, as in a light-emitting diode). Alternatively, incident light may promote an electron from an occupied state in the valence band to an empty state in the conduction band with the consumption of the energy of a photon. This is the process of optical absorption.

At typical temperatures the electrons and holes are at energies near to their respective band edges so the photon energy $h\nu$ is close to the separation of the $n_z = 1$ sub-bands. Since these energies are determined by the well width it is possible to change the photon energy for emission or absorption by changing the well width. As the well width is reduced the energy separation increases and the wavelength of light emitted by the structure is reduced. This ability to engineer the emission wavelength by simply adjusting

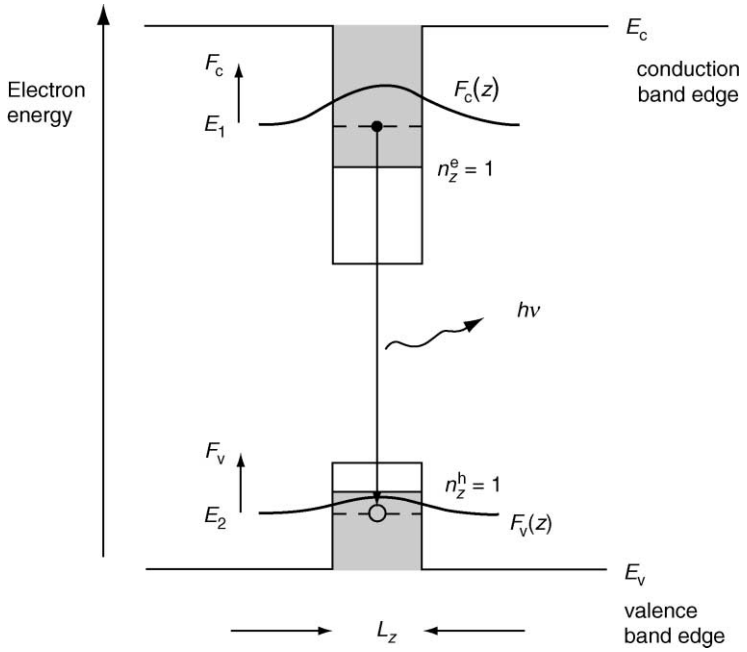


Figure 9 Illustration of a downward optical transition of an electron from a state in the $n_z = 1$ conduction sub-band to an empty state (hole) in the $n_z = 1$ valence sub-band resulting in emission of a photon of energy $h\nu = E_1 - E_2$. The spatial probability distribution of the electron and hole is specified by their respective envelope functions $F_c(z)$ and $F_v(z)$ and the probability of the transition occurring is proportional to their overlap integral, eqn [14].

the well width, without changing the chemical composition of the constituent materials, is one of the major attractions of quantum-confined structures. The shortest wavelength possible for a given material combination is determined by the band gap of the wide gap barrier material alongside the well.

The strength of the luminescence signal or the absorption is determined by the rates at which the electronic transitions occur. In general these rates are determined by the following intrinsic factors.

1. *The quantum mechanical probability of an electron making a transition between a full and an empty state*. This depends upon the 'matrix element' for the transition, which is a fundamental property of the well material and the overlap of the envelope functions. The electron and hole in the initial and final states must be in the same region of space and if the spatial distributions of electrons and holes, determined by F_c and F_v, do not overlap the probability for an electron to make a transition to the hole state is very low. The overlap is almost complete in a rectangular well (see Figure 9), but is only partial in a triangular well. The transition probability is proportional to

$$I_{\text{overlap}} = \left\{ \int F_v(z) F_c(z)\, dz \right\}^2 \qquad [14]$$

2. *The probability of occupation of initial and final states by electrons, specified by the Fermi factors* (eqn [11]). The initial state must be occupied and the Pauli exclusion principle prohibits an electron entering a final state which is already occupied. Thus for a downward transition resulting in emission of light the rate is proportional to $f_c(1 - f_v)$.
3. *The density of initial and final states* (eqn [9]). The more states there are within a given small energy interval, the greater the emission rate per unit energy.
4. *Photon distribution*. For processes such as optical absorption, which are 'induced' by the presence of a photon, the transition rate is also proportional to the photon density in the region of the well. Because the electrons are not wholly localized in the well, and because the spatial extent of the optical field is larger than that of the carriers, the coupling between the photon field (intensity $I_{\text{ph}}(z)$) and the carrier probability distributions is expressed as

$$I_{\text{coupling}} = \frac{\left\{ \int F_v(z) [I_{\text{ph}}(z)]^{1/2} F_c(z)\, dz \right\}^2}{\int I(z)\, dz} \qquad [15]$$

5. *Optical mode density*. The emission rate also depends upon the density of optical modes

available to receive the recombination radiation. In most situations the emission occurs into the cavity of free space which is very large compared with the wavelength of the radiation so the modes are very closely spaced and the cavity can accept radiation at any transition energy (this is the well-known Raleigh–Jeans theory of cavity radiation). However, in some special situations the cavity has dimensions comparable with the wavelength of the radiation (microcavities) and then the modes are widely spaced. In such a microcavity, the optical cavity effectively controls the electronic transitions.

Optical Absorption

Figure 10 is an optical absorption spectrum for a GaAs quantum well measured with light incident normal to the plane of the quantum wells. A series of steps can be seen which correspond to absorption at increasing photon energy between increasing orders of sub-bands, $E_{1e} \rightarrow E_{1h}$, $E_{2e} \rightarrow E_{2h}$, etc. Normally $n_z = 1$ to $n_z = 2$ transitions are forbidden in rectangular wells. At each step there is a pair of peaks (not always resolved) and above each step the absorption is roughly constant, following the step-like density of states function sketched in Figure 4. The peaks are due to the formation of excitons which are weakly bound electron–hole pairs. In bulk materials excitons are only observed at low temperatures whereas in quantum wells the localization increases the binding energy (E_{bX}, typically 6–8 meV) such that they are observed at room temperature. The excitonic peak is at an energy E_{bX} below the sub-band separation. There are in fact two different kinds of electrons in the valence band with different masses so there are pairs of $n_z = 1, 2, 3, \ldots$ sub-bands due to the different confinement energies for the different valence electrons. Spectra of this kind provided the first evidence for the distinctive sub-band structure of quantum well systems produced by carrier confinement. Absorption measurements provide data on the energy spacing of the sub-bands in the well, and it is possible to determine the well width and depths, from which the band offsets can also be determined.

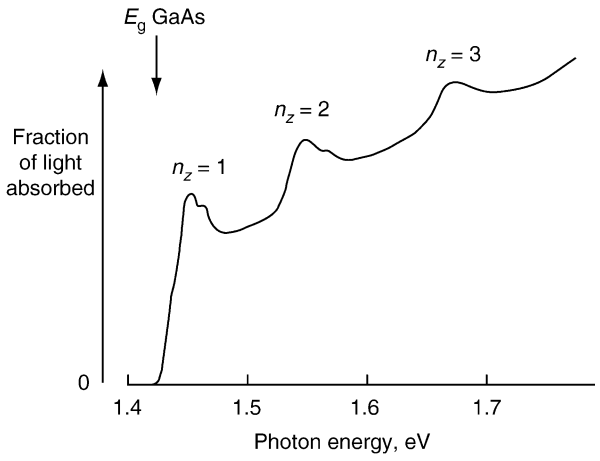


Figure 10 Absorption spectrum of a multiple quantum well structure having 10 nm GaAs wells measured at room temperature. The absorption edges corresponding to transitions between $n_z = 1, 2$, and 3 pairs of sub-bands can be identified. The peaks are due to absorption by formation of excitons.

The strength of the absorption is also of interest because it provides a measure of the optical matrix element if the Fermi factors f_c and f_v are known. If the incident optical beam is very weak such that the absorption process excites very few electrons then the system remains close to thermal equilibrium so the upper states are empty ($f_c = 0$) and the lower states are full ($f_v = 1$).

In *bulk material* the change in photon flux over a small distance Δx is proportional to the flux Φ and the distance traveled

$$\Delta\Phi = -\Phi\alpha\Delta x \qquad [16]$$

where α is the absorption coefficient which has units L^{-1}. This results in an exponential decrease of photon flux with distance traveled through the material

$$\Phi(x) = \Phi_0 \exp(-\alpha x) \qquad [17]$$

For a quantum well structure with light incident in the z-direction perpendicular to the plane of the quantum well, the fraction of light absorbed by a single well by transitions between a single sub-band pair (e.g., transitions between the $n_z = 1$ conduction and valence sub-bands) is independent of the thickness of the well because the density of states for a single sub-band is independent of well thickness. The strength of the absorption cannot be expressed by an absorption coefficient but by the fraction of light absorbed per well:

$$\Delta\Phi = -\gamma_w\Phi \qquad [18]$$

This arises because the quantum well is effectively a sheet in the z-direction: changing its thickness does not change the density of states per unit area (eqn [9]) and we cannot specify the electron concentration along the z-direction. This behavior also occurs because the well is much thinner than the wavelength of the light and it is not possible to

specify the variation of photon flux on a distance scale smaller than the wavelength. The barrier material surrounding the well has a wider band gap and does not absorb at the photon energies absorbed by the well. The well thickness does affect the sub-band spacing and at a fixed photon energy the absorption increases as the well width is decreased as light is absorbed by transitions between more sub-bands. Equation [18] applies individually to each sub-band because the density of states is the same for each sub-band.

In a multiple well system, the fraction of light transmitted through N independent wells is

$$T_N = \{1 - \gamma_w\}^N \qquad [19]$$

so the fraction of light absorbed by such a system is

$$\frac{\Delta\Phi}{\Phi} = -\left[1 - \{1 - \gamma_w\}^N\right] = -N\gamma_w \qquad [20]$$

which is proportional to the number of wells when $\gamma_w \ll 1$.

The fraction of light absorbed by transitions between a single pair of sub-bands for a rectangular GaAs quantum well is predicted to be about 0.01, and a similar value has been obtained from absorption measurements. In a rectangular well the envelope function overlap is complete; however, in a triangular well of the same material the absorption is reduced by the smaller overlap of the envelope functions.

Photoluminescence and Spontaneous Emission

We turn now to discuss the optical emission from quantum wells as a result of external excitation. An external light source with photon energy above the effective band gap produces excess electron–hole pairs by excitation of electrons from the valence band. These rapidly lose their excess energy and thermalize to the sub-band edges to take up Fermi energy distributions. The electrons subsequently recombine to vacant valence band states by spontaneous emission. The low energy edge of this spectrum corresponds to the sub-band separation and the shape of the spectrum at higher energies corresponds to the thermal distribution of carriers in the bands. At very low excitation intensities where the excess electron population is small the luminescence may be due to recombination of electrons within an exciton and occur at an energy E_{bX} below the sub-band separation. Generally the sub-band separation is several k_BT so only the lowest sub-band is populated and photoluminescence is not seen from higher sub-bands. In this respect absorption measurements provide a more complete characterization of the structure.

While the energy position of features in the photoluminescence spectrum provides useful information, it is difficult to use the strength of the photoluminescence signal to provide quantitative information about the rate of optical transitions within the material because the excitation power and volume within the sample must be known and a known fraction of emitted light must be collected and measured with a calibrated detector. The relative intensity of photoluminescence under standard conditions does provide comparative information and the linewidth of the spectrum at low intensity provides a qualitative indication of 'quality' (e.g., well width variations in the structure). A quantitative determination of the internal recombination rate can be obtained by measuring the decay of the luminescence signal following short-pulse excitation. Fuller discussion can be found in books dealing with the characterization of semiconductor structures.

Concluding Remarks

We have shown that an understanding of the electronic structure of quantum wells follows naturally from quantum mechanical treatments of electrons in bulk materials. When one of the dimensions of the structure is reduced to be similar to the de Broglie wavelength, quantization of the electron motion in this direction becomes apparent. The sub-bands which are formed have a profound effect on the electronic and optical properties when their spacing exceeds thermal energies such that only one band is significantly populated with carriers. The density of states function is a series of steps, each corresponding to a sub-band, and this produces characteristic step features in the optical absorption spectrum. The effective band gap of the structure is controlled by the width of the well. Quantum confinement also increases the binding energy of excitons compared with bulk materials and under low excitation conditions absorption and luminescence spectra are dominated by excitonic features even at room temperature.

One of the most successful application areas of quantum well structures has been in laser diodes. Figure 11 shows experimental data for laser emission wavelengths as a function of well width, showing the dramatic reduction which can be achieved simply by changing the width of the quantum well. Furthermore the step density of

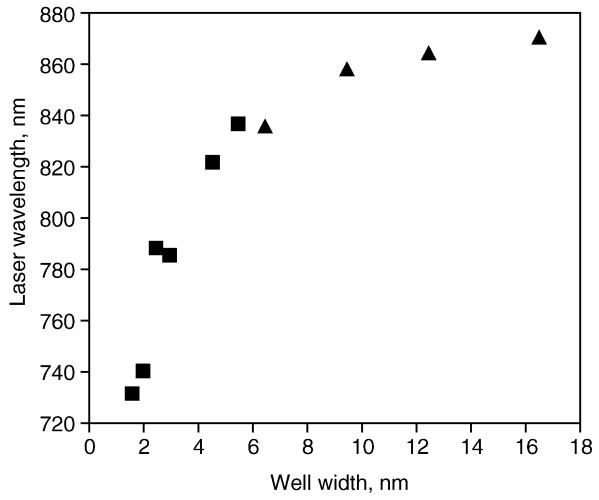


Figure 11 Measured values of the emission wavelength of GaAs/AlGaAs quantum well lasers as a function of well width, showing how the wavelength can be engineered by choice of the well width. Data from: squares: Woodbridge K, Blood P, Fletcher ED and Hulyer PJ (1984) Short wavelength (visible) GaAs quantum well lasers grown by molecular beam epitaxy. *Applied Physics Letters* 45: 16–18; and triangles: Chen HZ, Ghaffari A, Morkoç H and Yariv A (1987) Effect of substrate tilting on molecular beam epitaxial grown AlGaAs/GaAs lasers having very low threshold current densities. *Applied Physics Letters* 51: 2094–2096.

states function, characteristic of quantum wells, is one of the fundamental reasons for the reduction in threshold current of quantum well lasers relative to bulk GaAs devices.

We have considered GaAs because it is a model material system for production of quantum well structures. The concepts developed in this chapter are, however, quite general and can be applied to other material systems, other forms of quantum well and quantum wires and dots.

List of Units and Nomenclature

E	energy
E_f	Fermi energy
E_g	bandgap
F	envelope wavefunction
$f(E)$	Fermi–Dirac distribution function
g_{2D}	two-dimensional density of states $[\text{energy}]^{-1}\,[\text{area}]^{-1}$
$\hbar$	Planck's constant divided by 2π
I_{ph}	optical field intensity
$\mathbf{k}$	wavevector
k_B	Boltzmann's constant
m	electron mass
n	carrier density, expressed per unit area in quantum well structures
Q_c, Q_v	heterostructure band offset ratios
$\mathbf{r}$	position vector
t	time
T_N	fraction of light transmitted at normal incidence through N quantum wells
$\langle \mathbf{x} \rangle$	unit vector along x-direction, similarly for y, z directions
α	optical absorption coefficient in a bulk material, per unit length
γ	fraction of light absorbed by a quantum well at normal incidence to the plane of the well
λ	wavelength
Φ	photon flux (photons per unit time crossing unit area)
Ψ	electronic wavefunction
ν	light frequency

See also

Semiconductor Materials: Band Structure Engineering; Quantum dots; Type-II Quantum Wells and Superlattices. **Semiconductor Physics:** Excitons; Outline of Basic Electronic Properties.

Further Reading

Bastard G (1988) *Wave Mechanics Applied to Semiconductor Heterostructures*. New York: Les Editions de Physique, Halstead Press.

Bimberg D, Grundmann M and Ledentsov NN (1999) *Quantum Dot Heterostructures*. Chichester, UK: Academic Press.

Blood P (1991) Heterostructures in lasers. In: Morgan DV and Williams RH (eds) *Physics and Technology of Heterojunction Devices*. Stevenage, UK: Peter Peregrinus.

Blood P (2000) On the dimensionality of optical absorption, gain and recombination in quantum-confined structures. *Journal of Quantum Electronics* 36: 354–362.

Blood P and Orton JW (1992) *The Electrical Characterisation of Semiconductors: Majority Carriers and Electron States*. London: Academic Press.

Casey HC Jr and Panish MB (1978) *Heterostructure Lasers*. San Diego, CA: Academic Press.

Dingle R, Gossard AC and Weigmann W (1975) Direct observation of superlattice formation in semiconductor heterostructures. *Physical Review Letters* 34: 1327–1330.

Eisberg R and Resnick R (1985) *Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles*, 2nd edn. New York: John Wiley.

Hook JR and Hall HE (1991) *Solid State Physics*. Chichester, UK: John Wiley.

Kelly MJ (1995) *Low-Dimensional Semiconductors*. Oxford, UK: Clarendon Press.
Orton JW and Blood P (1990) *The Electrical Characterisation of Semiconductors: Measurement of Minority Carrier Properties*. London: Academic Press.
Perkowitz S (1993) *Optical Characterisation of Semiconductors*. London: Academic Press.
Wolfe CM, Holonyak N Jr and Stillman GE (1989) *Physical Properties of Semiconductors*. Englewood Cliffs, NJ: Prentice Hall.

Recombination Processes

P T Landsberg, The University of Southampton, Southampton, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In studies of recombination there occur considerable complications; many are associated with interactions between electrons, electrons and phonons, excitons, bi-excitons, impurity centers, etc. Most of these are not required in the present exposition. Modern work normally assumes that the basics of recombination physics are understood and the present exposition offers an appropriate outline.

Basic Assumptions

Electron–Electron Interactions for Electrons in Bands

The recombination problem in semiconductors is greatly complicated by the interaction of the electrons with each other. This allows one to speak only of the quantum states of the semiconductor crystal as a whole. However, as in metals, so also in semiconductors, a simplified picture is successful. In this, the electron interactions, and other interactions, are first neglected, but are later taken into account as a perturbation. Thus, the electrons are first treated as if they can move through each other; the fact that they collide and deflect each other by virtue of their Coulomb interaction is treated as a perturbation. The transitions are still described within the framework of the single-particle states of the unperturbed problem.

The Effect of Electron–Boson Interactions

Our first approximation is to neglect most (but not all) electron interactions. Later we take into account the two-electron transitions that arise. This two-particle recombination process is referred to as Auger recombination and its inverse as the generation of current carriers by impact ionization. In addition, electrons interact with the radiation and lattice fields and emit or absorb photons or phonons. These electron–boson interactions result in transitions of single electrons in an energy band scheme. We shall attach a superfix 'S' (for single-electron) to the recombination coefficients for such processes. These are then denoted by B^{S} or T^{S} depending whether only bands are involved or whether traps are also involved. They are illustrated in Figure 1, where the solid horizontal lines represent the conduction and valence band edges and the dashed line refers to traps. Note that the two-electron transitions do not have the superfix 'S', while n and p refer to the electron and hole concentrations; N_0 and N_1 are the concentration of unoccupied and occupied trap states, respectively.

Electron–Electron Interaction in Traps

Electron–electron interactions can at least formally be taken into account in connection with electrons trapped in a center: the spectrum is a function of the number, $r\,(=1,2,3,\ldots,M)$, of electrons captured. The 'irremovable' electrons can be included with the ion core. Given that the center captured r electrons (say), it can still be in a variety of quantum states, and they will be denoted by the symbol ℓ, yielding a set of quantum states (ℓ, r) for a center. The energy of such a state, divided by kT to make it dimensionless, is denoted by $\eta(\ell, r)$, yielding the canonical partition function

$$Z_r = \sum_{\ell} \exp[-\eta(\ell, r)] \qquad [1]$$

The location of a center in space will here be considered to be of no significance.

Centers can capture several electrons. This brings in a need for the chemical potentials (or Fermi levels) for r-electron centers. They are here denoted by γ (when divided by kT) and the suffix 'eq' denotes an equilibrium value. Thus, the equilibrium probability of finding an r-electron center in state ℓ is given by:

$$P(\ell, r)_{\mathrm{eq}} = \frac{\exp[r\gamma_{\mathrm{eq}} - \eta(\ell, r)]}{\sum_{s=0}^{M} (\exp s\gamma_{\mathrm{eq}}) Z_S} \qquad [2]$$

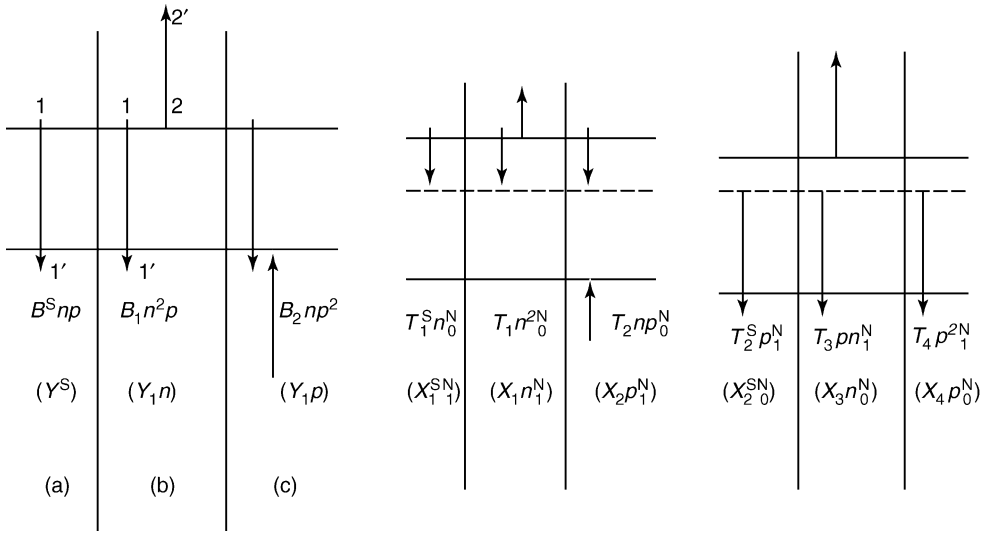


Figure 1 Definition of recombination coefficients. Transition rates per unit volume are stated with each process, and, in brackets, for the reverse process. Thus B_1, B_2, T_1-T_4 refer to Auger processes; B^s, T_1^s, T_2^s refer to single-electron recombination; Y^s, X_1^s, X_2^s refer to carrier generation processes; Y_1, Y_2, X_1-X_4 refer to impact ionization processes. Arrows indicate transitions made by electrons.

One can identify the dominant energies from optical or electrical experiments.

To understand the properties of these centers, suppose we can arrange for the equilibrium Fermi level to rise from the valence band to the conduction band. At first practically no electrons are captured ($r = 0$). As the Fermi level rises the states corresponding to $r = 1$ begin to appear. The states ℓ of the center which correspond to the ground states in equilibrium are always more highly populated than the states ℓ corresponding to excited states, so that we can often confine attention to them. As the Fermi level rises, the ground states for $r = 2$ become more important, and there may now be hardly any centers which have captured fewer electrons. If a larger number of electrons cannot be captured, by the time the Fermi level reaches the conduction band, then states of the center with $r > 2$ can (normally) be neglected as unstable. That this is a satisfactory picture is our second approximation.

Assumptions for Nonequilibrium Statistics

The now much simplified recombination problem is still complex because of the many states available to electrons in bands and centers. A key simplification arises from the fact that it is often possible to talk about a small number of groups of quantum states: let $I\,(= 1, 2, \ldots)$ label quantum states in a group i, let $J\,(= 1, 2, \ldots)$ label quantum states in a group j, etc. Within each group it is supposed that the transitions are much more rapid than they are between groups. In that sense electrons in each group are in equilibrium among themselves. Their quasi-equilibrium is then characterized by what is called a quasi-Fermi level. The dimensionless version, obtained by dividing it by kT, is denoted by γ_i, γ_j, etc., for groups i, j, etc. That this is reasonable is our third assumption. Thus γ_c, γ_v are quasi-Fermi levels which refer to the quasi-equilibrium of the conduction and valence band, respectively, and γ_I to an I-electron center. Groups of states with different quasi-Fermi levels are not in equilibrium with each other.

Recombination problems can now be discussed by neglecting transitions within one (quasi-equilibrium) group, because they proceed at exactly the same rate as their reverse transitions. The number of transition types to be considered is thereby greatly reduced. Away from equilibrium we shall adopt eqn [3] instead of [2] in order to allow for distinct quasi-Fermi levels:

$$P(\ell, r) = \frac{\exp[r\gamma_r - \eta(\ell, r)]}{\sum_{s=0}^{M} (\exp s\gamma_S) Z_S} \qquad [3]$$

Equation [3] is not always correct, but is an approximation arising from the quasi-Fermi level assumption. Among the improved theories are, for example, the 'cascade' theories.

The assumption of a quasi-Fermi level for each state of charge of a center implies that all the excited states of an r-electron center are populated according to eqn [2].

The Main Recombination Rates

General Results and The Two-Band Case

Recombination and its converse, generation, consist of a transition of an electron from one state to another. The observed rate is the net rate of recombination and is the algebraic sum of the recombination and generation processes. During these processes both total energy and total momentum must be conserved. This is achieved by creation or absorption of photons, or phonons, excitation of secondary electrons, etc.

The transition probability per unit time, S_{IJ}, for a single-electron transition from state I in a band to state J requires that state I should be occupied, with probability p_I say, and state J should be vacant with probability q_J say. The general expression for the average rate of the transition from I to J then takes the form $p_I S_{IJ} q_J$. For the reverse process electron state J has to be occupied and state I vacant. Thus the rate of this reverse process is $p_J S_{JI} q_I$. The net recombination rate per unit volume of the process I to J can then be written as:

$$u_{IJ} = (p_I S_{IJ} q_J - p_J S_{JI} q_I) V^{-1} \qquad [4]$$

By the principle of detailed balance, this expression vanishes at equilibrium. If one puts $X_{JI} \equiv S_{IJ} p_I q_J / S_{JI} p_J q_I$, one has $u_{IJ} = p_J S_{JI} q_I (X_{JI} - 1) V^{-1}$. In equilibrium $X_{JI} \rightarrow (X_{JI})_{\text{eq}} = 1$, and the recombination rate is zero.

One can say a little more if one assumes that states I and J are in conduction or valence bands, each with a quasi-Fermi level (divided by kT to make it dimensionless). If these are denoted by γ_e and γ_h, then $p_I = [1 + \exp(\eta_I - \gamma_e)]^{-1}$ for a conduction band. For the total recombination rate per unit volume between the bands i and j one finds:

$$u_{ij} \equiv \left[\sum_{I \in i} \sum_{J \in j} p_J S_{JI} q_I \right] \left[\exp\left(\frac{|e|\phi}{kT} \right) - 1 \right] V^{-1} \qquad [5]$$

where ϕ is essentially the difference between the quasi-Fermi levels: $\gamma_e - \gamma_h$. The first factor depends on the bands involved and it has been assumed that the transition probability ratio S_{IJ}/S_{JI} is independent of the excitation.

A transition rate, when multiplied by the charge of current carriers, is a current, and when divided by the area of the surface involved, is a current density. If i and j denote the states of the conduction and valence bands of a semiconductor, excitation independence may often be assumed, and one then expects a current density proportional to $\exp(|e|\phi/kT) - 1$. This is characteristic of the current through pn junctions, metal semiconductor junctions, etc. In these configurations the Fermi level difference between the ends of the device determines the voltage across it. When radiation is involved, however, excitation dependence of some of the parameters introduced above (e.g. S_{JI}) tends to spoil this simple story.

The Case of Defects

So far we have considered only two bands. When a trap is involved matters are rather different. Because of the interactions among the electrons on a center it is not possible to talk of the same level being occupied or vacant. Consequently, identification of forward and reverse processes in terms of levels becomes impossible. Instead one deals with a center, say an r-electron center, as a whole; we then need the probability that a given center is an r-electron center. For example, the capture of an electron converts an $(r-1)$-electron center into an r-electron center. Thus the p's and q's must be replaced by more complicated expressions. It is convenient to denote u_{ij} by u_{cv} in the simple two-band case and its structure is given (with a sign change) by a recombination coefficient:

$$u_{cv} = B^S np[1 - \exp(\gamma_h - \gamma_e)] \equiv B^S np - Y^S \qquad [6]$$

using Figure 1a. Here, n and p are the electron and hole concentrations. Auger effects in Figures 1b and 1c can also be included, at least formally. Then B^S has to be replaced in eqn [6] by $B^S + B_1 n + B_2 p$. Analogous replacements are found for recombination processes involving defects.

In the simplest case of one type of localized defect one finds a steady-state recombination rate per unit volume of the form:

$$u = \frac{np - (np)_0}{\tau_{n0}(p + p_1) + \tau_{p0}(n + n_1)} \qquad [7]$$

Here τ_{n0}, τ_{p0} are parameters with the dimension of time and p_1, n_1 are parameters with the dimension of concentration. The recombination increases with the defect concentration, which is actually in the denominators of τ_{n0} and τ_{p0}. This is an old and much used result associated with the names of W Shockley and W T Read.

Radiative Transitions

Spontaneous and Stimulated Emission

Quantum theory was initiated by Planck's law for black-body radiation at temperature T. This gives the spatial energy density as a function of frequency ν in this system as

$$f(\nu, T) = \frac{8\pi\nu^2}{c^3} \frac{h\nu}{\exp(h\nu/kT) - 1} \qquad [8]$$

For low frequencies one finds the Rayleigh–Jeans law which makes [8] proportional to kT in agreement with the classical equipartition theorem. For high temperatures [8] diverges, as required.

The result [8] may also be obtained by writing

$$N = [A + Bf(\nu, T)]N_u \qquad [9]$$

for the emission rate of photons by atoms, where N_u is the number of atoms in the upper of two states which are separated by the energy $h\nu$. The first term on the right-hand side is due to the normal decay of an excited state ('spontaneous emission'). The second term refers to additional emission ('stimulated emission') induced by the radiation of frequency $h\nu$ itself.

The first factor in [8] is due to the density of states and the second factor is due to the fact that the energy is considered. If one considers the number of photons of energy $h\nu$ at temperature T one comes up with

$$N_\nu = \left[\exp\frac{h\nu - \mu}{kT} - 1\right]^{-1} \qquad [10]$$

(Bose–Einstein distribution)

Here μ is a possible chemical potential of the radiation which is non-zero only in nonequilibrium situations such as in a semiconductor laser.

In a pn junction the two bulk materials several diffusion lengths away from the junction are approximately in equilibrium even if a modest current is flowing. On one side one has then just one quasi-Fermi level, say μ_I, and on the other side one has just one quasi-Fermi level, say μ_J. Then the difference

$$\mu \equiv \mu_I - \mu_J = q\varphi \qquad [11]$$

corresponds to the applied voltage φ. This is in agreement with eqn [10], in that $\varphi = 0$ implies no current and hence thermal equilibrium is possible.

Statistical mechanics teaches one the rule that, in equilibrium, occupation probabilities of individual quantum states are always less for states of higher energies. This ensures that the total energy of a quantum system converges, the occupation probability p being a kind of convergence factor. However, if one is away from equilibrium, the above rule can be suspended and one can have population inversion.

For $h\nu = \mu$ there is trouble with [10] because the steady-state photon occupation diverges. This does not correspond to a 'death ray', but is the result of imperfect modeling; for example, the leakage of photons from the cavity may have been neglected. A formula of type [10] is also needed in connection with solar cells.

Donor–Acceptor Pair Recombination

A striking demonstration of radiative donor–acceptor transitions in GaP at low temperature (1.6 K) was revealed by sharp lines (Figure 2) at photon energies $h\nu_i$ given by

$$h\nu_i = E_G - E_A - E_D + e^2/\varepsilon R_i\,[-E] \qquad [12]$$

where R_i is the distance between the ith-order nearest neighbors.

The discrete nature of the peaks is due to the fact that the impurities involved settle in general on lattice sites so that only definite separations R_i are possible. The energy gap is E_G; the value of $E_A + E_D$ can be inferred from the experimental lines by extrapolation to $R_i \to \infty$. The energy term E is sometimes neglected.

The ZnS phosphors were the first materials in which donor–acceptor radiation was hypothesized. However, it is hard to control its stoichiometry and its impurity content, and it has relatively large carrier mass and hence relatively large impurity activation energies. In such cases spectra such as those shown in Figure 2 are hard or impossible to obtain. This applies in general to many II–VI compounds.

Quantum Efficiency

The quantum efficiency is the radiative recombination as a fraction of the total recombination. It can also be regarded as the average number of electrons produced per incident photon. Consider an intrinsic semiconductor, i.e., one in which the electron and hole concentrations are equal. Then with the notation of Figure 1 the radiative band–band recombination rate is $B^s n^2$. The nonradiative rate is $(B_1 + B_2)n^3$. We shall add another nonradiative rate, An say, proportional to the injected carrier density $n \sim p$. The quantum efficiency is then

$$\eta = \frac{B^s n^2}{An + (B_1 + B_2)n^3 + B^s n^2} \qquad [13]$$

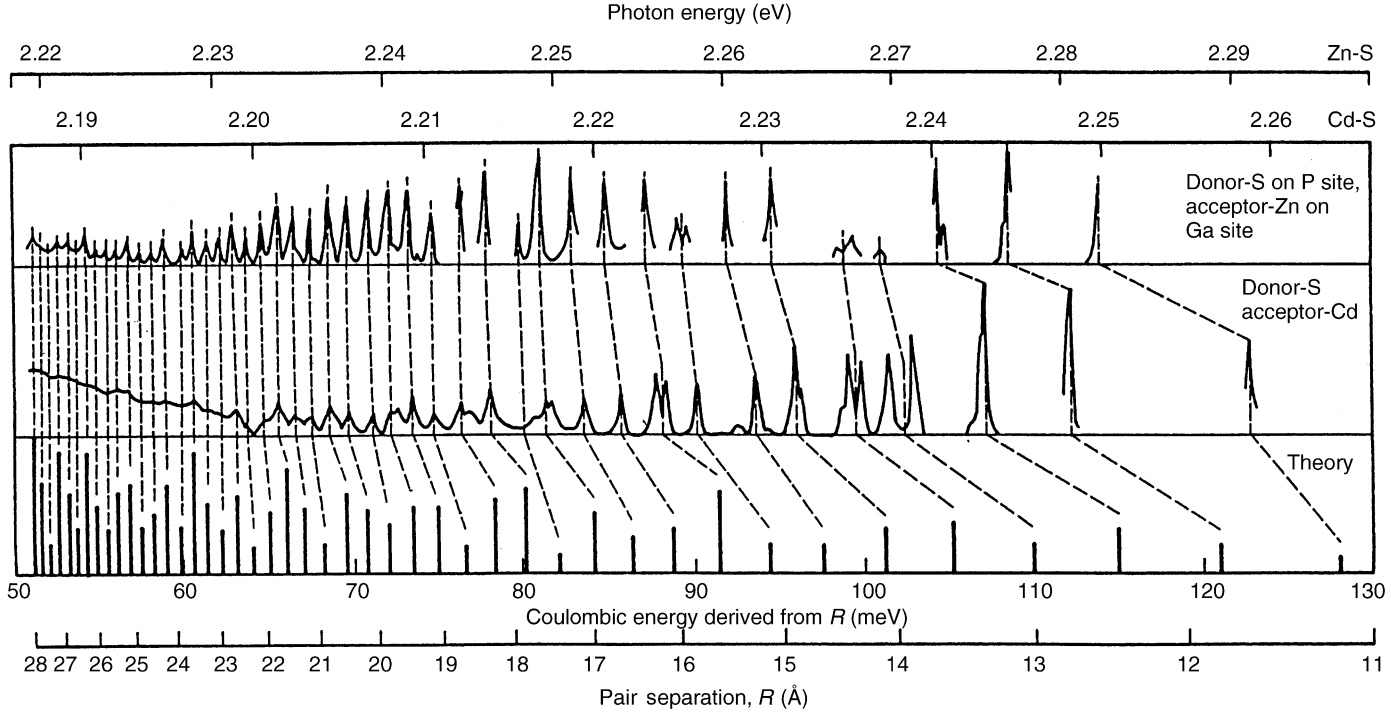


Figure 2 Comparison of the positions and intensities of the sharp line spectra at 1.6 K corresponding to both ZnS and CdS acceptor–donor pairs with the predicted pair distribution. The lower scales show the pair separation (R) and the Coulombic energy derived from R. The emission energy scales for the two measured spectra are shown at the top. Reproduced with Gershenzon M, Logan RA, Nelson DF, *et al.* (1968) *Proceedings of the International Conference on Luminescence*. Budapest, Hungary: Akademia Kiada.

Traps are here neglected, and one finds a maximum

$$\eta_{\max} = \left\{1 + \frac{2}{B^s} A^{1/2} (B_1 + B_2)^{1/2}\right\}^{-1} \qquad [14]$$

For a high-quality epitaxial AlGaAs/GaAs double heterostructure of great purity we may take

$$A \sim 0.5 \times 10^6 \text{ s}^{-1}, \quad B \sim 10^{-10} \text{ cm}^3 \text{ s}^{-1},$$
$$B_1 + B_2 \sim 10^{-29} \text{ cm}^6 \text{ s}^{-1} \qquad [15]$$

whence $n_1 \sim 2 \times 10^{17}$ cm^{-3}, $\eta_{\max} \sim 0.96$.

Detailed Balance

It is clear that the radiative recombination rate from a material should be obtainable in terms of its optical 'constants', the absorption coefficient $\alpha(\lambda)$ and the refractive index $\mu(\lambda)$. Both are functions of the wavelength. This connection can be formalized by comparing the optical absorption process with the emission process and this can be done by using the principle of detailed balance. This equates the rate of disappearance by absorption of photons with the rate of production of photons by radiative recombination. The radiative recombination rate can thus be obtained as an integral over the optical functions. This relationship, pioneered by van Roosbroeck and Shockley in 1954, has been used a great deal because the optical quantities are often known with some accuracy.

This result corresponds to balancing the rate $B^s np$ and the rate Y^s in **Figure 1a**. Analogous detailed balance results are obtainable for the other pairs of processes in **Figure 1**. Thus, the Auger recombination rate can be related to an integral over the impact ionization rate. However, this connection is less useful since impact ionization data are generally less well known than the optical absorption information.

Nonradiative Processes

Auger Effects

The Auger effect was discovered in 1925 in gases by Pierre Auger. An atom is ionized in an inner shell. An electron drops into the vacancy from a higher orbit and a second electron takes up the energy which is used to eject it from the atom. In solids the effect is roughly analogous. One of its characteristics is that it is not radiative. Since radiation is what is seen in many experiments, the Auger effect is suspected if and when there is less radiation than expected in the first place. It is rather harder to investigate than radiative transitions.

The effect has proved to be important as it limits the performance of semiconductor lasers, light-emitting diodes and solar cells, and it can be crucial in transistors and similar devices whose performance is governed by lifetimes. When heavy doping is required, as it is in the drive towards microminiaturization, its importance tends to increase, since the Auger recombination rate behaves roughly as n^2p or p^2n (see **Figure 1**) compared with the radiative rate which behaves more like np. The inverse process is impact ionization, and is important in the photodiode, the impact avalanche transit time (IMPATT) diode, and hot electron devices.

Impact ionization can be regarded as an autocatalytic reaction of order one:

$$\text{e} \rightarrow 2\text{e} + \text{h} \quad \text{or} \quad \text{h} \rightarrow 2\text{h} + \text{e} \qquad [16]$$

i.e., one extra particle is produced of the type present in the first place. This is a key feature for impact-induced nonequilibrium phase transitions in semiconductors.

In the theory one needs wavefunctions for four states of two electrons which are relevant after the many-electron problem has been reduced to a two-electron problem. In a matrix element calculation the four wavevectors imply a 12-fold integration in $\boldsymbol{k}$-space to cover all possible states of the two electrons. However, momentum and energy conservation usually reduce this to an eight-fold integration. Such a calculation is hard, for it requires (1) good wavefunctions and (2) accurate integrations.

Here we shall merely concentrate on the broad principles. In addition, the many-electron nature of the problem implies that another approximation is often inherent in the treatment apart from the use of perturbation theory. This is clear if one considers that the electron interactions are screened twice: once by an exponential screening factor and a second time by the dielectric constant. So there is some double counting, and the treatment of the effect as uncorrelated electronic transitions mediated by screened Coulomb interactions is another approximation. The collective effects that enter require more sophisticated field theoretic methods which take care of electron–hole correlations, plasmon effects and the effect of free excitons, the so-called excitonic Auger effect. These calculations broadly confirm the results obtained by the simpler methods used for energy gaps large compared with the plasmon energies,

provided the high-frequency dielectric constant is used and doping is not too heavy. Significant corrections are required for the narrow gap lead compounds, for example, and a considerable computational effort is required.

Let us now consider lifetimes for high carrier concentrations, as limited by band–band processes. They are best studied by looking at the emitted radiation. In this connection we recall the striking rise of the threshold current density J in a semiconductor laser as the material is changed from a direct material like GaAs to an indirect material, by mixing it with a compound such as GaP. Figure 3 shows this spectacular rise for $GaAs_{1-x}P_x$ at 77 K near $x = 0.38$. It is due to the drop in the radiative transition probabilities, as the substance becomes indirect, thus requiring a higher current density for threshold. So we should start with band–band processes in direct materials as most favorable for the experimentally accessible radiative transitions.

The essential point here is that the injected carrier density behaves as:

$$n_{\text{inj}} = \frac{J\tau}{qd} \sim \frac{(5\times10^{3}\ \text{A cm}^{-2})(10^{-9}\ \text{s})}{(1.6\times10^{-19}\ \text{C})(10^{-4}\ \text{cm})} \sim 10^{17}\ \text{cm}^{-3} \quad [17]$$

where J/q is the particle current density at threshold, τ is the lifetime of the carriers, and d is the thickness of the active layer. As active layers are made thinner n_{inj} increases to $10^{19}\ \text{cm}^{-3}$ or so, far in excess of the defect concentration in the (undoped) material.

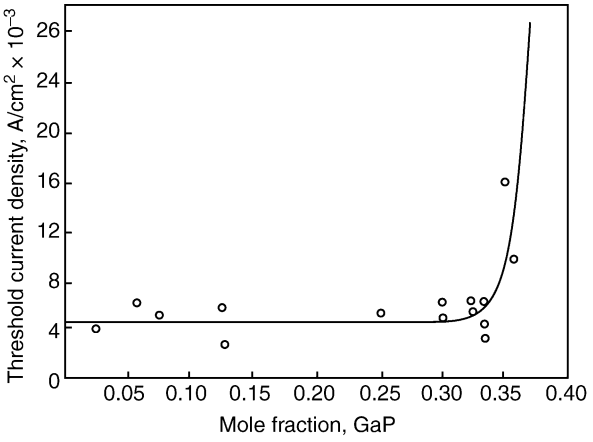


Figure 3 Lowest values of laser threshold current density at 77 K as a function of mole fraction of GaP for $Ga(As_{1-x}P_x)$ junction lasers. Reproduced from Neill CJ, Stillman GE, Sirkis MD, *et al.* (1966) Gallium arsenide-phosphide: Crystal, diffusion and laser properties. *Solid-State Electronics* 9: 735–742, with permission from Elsevier.

This brings in band–band Auger effects. These have also been invoked to explain the undesirable increase in J with temperature. Thus the interest in direct band–band Auger effects in III–V compounds is fuelled by the need for better and smaller opto-electronic devices which work at long wavelengths (1.3–1.5 μm). It matches the interest in indirect band–band Auger and impurity Auger effects in silicon due to the importance of heavy doping in VLSI (very large scale integration) and transistor technology. Figure 4 shows a typical Auger process, called CHHS, as the conduction band (C) and the split-off band (S) are involved. Two relevant states are in the heavy hole band (H). State $2'$ is referred to as the Auger carrier, as it has more kinetic energy than the others.

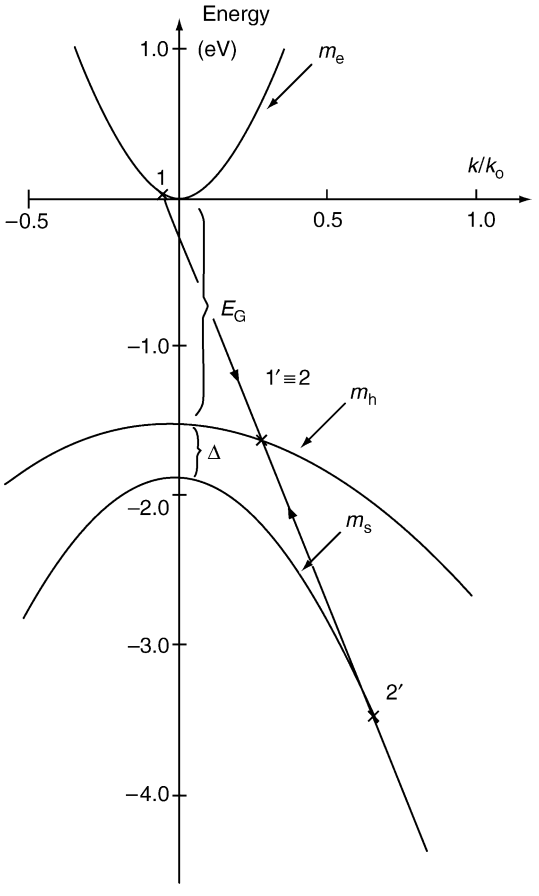


Figure 4 Conduction band, heavy hole and split-off valence bands of GaAs, all treated as parabolic with $\hbar^2k_0^2/2m_h \equiv E_G - \Delta$. x denotes a quadruplet of states for a most probable transition. The two states in the heavy hole band are not shown separately. The arrows indicate electron transitions. Reproduced from Neill CJ, Stillman GE, Sirkis MD, *et al.* (1966) Gallium arsenide-phosphide: Crystal, diffusion and laser properties. Solid-State Electronics 9: 735–742, with permission from Elsevier.

Impact Ionization

The CHHS and CHCC processes and their inverses can be written as

$$e_C + 2h_H \leftrightarrow h_s \quad 2e_C + h_H \leftrightarrow e_C \qquad [18]$$

where the suffix refers to the band. Viewed from left to right these are Auger processes. Viewed from right to left they are autocatalytic and impact ionizations.

Such processes are important for the impact-induced nonequilibrium phase transition in semiconductors. Also reaction rates with autocatalytic elements imply nonlinear equations in the concentrations and this gives rise to much interesting behavior as regards stability and bifurcation phenomena.

Momentum and energy conservation are very restrictive conditions on the four states involved in [18], and it is easily seen that they cannot normally be satisfied if in a direct band semiconductor the recombining electron drops from the band minimum to the valence band maximum. This effect raises all the kinetic energies of possible processes. The Auger electron (on the right-hand sides of [18]) must also have a minimum energy in order that the impact-ionization process can proceed. This leads to an activation energy for the process. For Figure 4, for example, and in the case of nondegeneracy, the energetic hole has a kinetic energy at threshold of

$$E_{th} = \frac{m_e + 2m_h}{m_e + 2m_h - m_s}(E_G - \Delta) \quad \text{(CHSS)} \qquad [19]$$

As the band gap increases we see that E_{th} goes up and so the Auger rate for simple parabolic bands decreases. Values of E_{th} for other transitions can be obtained from equation [19].

The total kinetic or threshold energy E_{th} can be converted to an activation energy by subtracting the basic energy which must under all conditions be part of the Auger particle energy. In the case of eqn [19] one finds for $E_G > \Delta$:

$$\begin{aligned} E_a &= E_{th} - (E_G - \Delta) \\ &= \frac{m_s}{m_e + 2m_h - m_s}(E_G - \Delta) \end{aligned} \qquad [20]$$

This can result in a strong temperature dependence in accordance with an Arrhenius factor $\exp(-E_{th}/kT)$. However, this is again lost, and the direct band–band Auger effect will certainly be important and only weakly temperature dependent, if $E_G \sim \Delta$. This can occur, for example, for InAs, GaSh and their solid solutions.

The connection between Auger processes and impact ionization has also been formalized. Of all Auger quadruplets of states in a set of nondegenerate bands, the most probable Auger transition involves the quadruplet, which yields the threshold for impact ionization. The three crosses in Figure 4 indicate such a quadruplet (the central cross represents two states).

Identification of Auger Effects

How does one know that an Auger effect has occurred? In pure but highly excited materials there is the original solid-state Auger effect which leads to a lifetime broadening of the electronic states at the band edge. For a semiconductor this effect causes fuzziness in the band edge which is a contributory factor to the overall bandgap shrinkage. The mechanism is illustrated in Figure 5, which shows how the lifetime of a vacant electron state near the band edge is shortened by the Auger processes within this band. Under normal conditions this effect is small in a semiconductor. But under degenerate conditions the effect leads by lifetime broadening to a low-energy tail in emission.

The blanket term 'Auger effect' for all Coulombically excited two-particle transitions has been used here. In semiconductor device work the 'Auger effect' has become associated with transitions in which one electron bridges an energy gap. But there is no need to limit the concept in this way. Its competition with radiative effects means that it is often detrimental and a full discussion of means of suppressing it has recently been given by Pidgeon *et al.*

More spectacular and more convincing is the *detection* of the energetic Auger electron or hole. For this purpose one may look for the weak luminescence emitted when this carrier recombines radiatively. This has been done for the band–band process in Si and for the band–impurity process

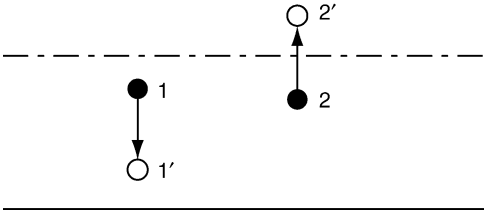


Figure 5 The filling of a state k by the Auger effect. The arrows indicate electron transitions.

in GaAs. It leads to so-called $2E_G$-emission. Its rate-limiting step is the rate of the Auger process per unit volume, $B_1 n^2 p$, which populates the high-energy level. The radiative process then proceeds at a rate $B'_1 n^2 p^2$ per unit volume with $B'_1 \sim 10^{-54}$–10^{-51} cm^9 s^{-1}.

More usual is the identification of the band–band Auger recombination mechanism from the minority carrier lifetime τ, which behaves as

$$\frac{1}{\tau_p} = B_1 n^2 \quad \text{or} \quad \frac{1}{\tau_n} = B_2 P^2 \qquad [21]$$

where B_1 and B_2 are Auger coefficients when the Auger particle is an electron or hole, respectively. A more complete expression allows also for trapping processes.

The departure from parabolic bands plays an important part for the energetic Auger particle (in state 2′). Theory shows that when this effect is taken into account, it tends to lower the band–band Auger coefficient. In fact in GaAs the CHCC process is ruled out altogether at 0 K for nonparabolic bands, suggesting that it should be an ideal material for radiative transitions. However, the possibility of phonon participation brings the effect back again, though it is still comparatively weak. The difficulty of conserving electron energy and momentum, which gives rise to the activation energies (eqn [20]) is greatly alleviated if phonons can take up some of the momentum in a direct gap semiconductor.

We have seen that the activation energy barrier against the band–band Auger effect can be overcome by a suitable disposition ($E_G = \Delta$) of the three direct bands and by phonon participation. It can also be overcome if there is an indirect minimum near a Brillouin zone edge at about half the direct energy gap.

Acknowledgments

The author wishes to acknowledge support from NATO under their grant PST.CLG 975758.

See also

Semiconductor Physics: Band Structure and Optical Properties; Excitons; Impurities and Defects; Infrared Lattice Properties; Outline of Basic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Fraser DA (1986) *The Physics of Semiconductor Devices*, 4th edn. Oxford, UK: Clarendon Press.

Hangleiter A (1985) Experimental proof of impurity Auger recombination in silicon. *Physical Review Letter* 55: 2976–2978.

Harrison D, Abram RA and Brand S (1999) Characteristics of impact ionization rates in direct and indirect gap semiconductors. *Journal of Applied Physics* 85: 8186.

Landsberg PT (1991) (paperback edn., 2003) *Recombination in Semiconductors*. Cambridge, UK: Cambridge University Press.

Nimtz G (1980) Recombination in narrow gap semiconductors. *Physics Reports* 63: 265.

Pidgeon CR, Ciesla CM and Murdin BN (1997) Suppression of non-radiative processes in semiconductor mid-infrared emitters and detectors. *Progress in Quantum Electronics* 21: 361–419.

Spin Transport and Relaxation in Semiconductors; Spintronics

M E Flatté, University of Iowa, Iowa City, IA, USA

D D Awschalom, University of California, Santa Barbara, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Control of the generation and transport of coherent electron spin in semiconductors suggests new ways to probe the fundamentals of quantum decoherence in solids and to explore device applications that rely on coherence, including quantum computation. The focus here will be on the physical phenomena that govern spin transport and relaxation, with a particular emphasis on those phenomena that are amenable to direct manipulation. Manipulation of material properties in general has been taken to its most advanced level in the design of semiconductor electronic devices, in which operations are performed by perturbing the motion of electrons through their electric charge; hence, 'spintronics' is a term introduced to describe both the manipulation of electrons by their spin and control of their spin properties, such as g-factors.

The study of the evolution of spin coherence first requires the generation and detection of spin coherence. Here the ultrafast optical technique for generating and detecting spin-polarized populations will be emphasized. The next logical component to treat will be the persistence of spin coherence, followed by the transport of spin coherence from one region of a material to another. Finally the direct ultrafast manipulation of spin-polarized populations will be described. These components are likely to play a prominent role in any attempt to apply the fundamental physical phenomena of spin coherence to specific device technologies.

Optical Generation of Spin-Polarized Distributions

Optical pulses with definite circular polarization can be used to generate spin-polarized distributions of excited electrons in crystals. In order for this process to work one of the energy bands involved in the optical transition must be characterized by a substantial spin–orbit interaction. The selection rules for circularly polarized photons require a change in the orbital angular momentum of the electron projected along the propagation direction of the photon. For the optical transitions of an atom in free space this azimuthal orbital angular momentum l_z must increase or decrease by $\hbar$, permitting the coupling of the p-state with $l_z = 1$ to an s-state for left circular polarization, and of the p-state with $l_z = -1$ to that s-state for right circular polarization. Thus the optical selection rules pick out a particular orbital angular momentum projection.

These selection rules by themselves do not explain how spin-polarized populations are generated, for both spin directions are possible for each value of l_z. In the presence of spin–orbit interaction, however, the energy of a state with a particular l_z is correlated with the spin orientation. The six p-states, which are degenerate in the absence of spin–orbit interaction, split into two groups: one of four degenerate states and the other of two degenerate states. The higher-energy group of four degenerate states consists of two states with **L** and **S** parallel and two states with a one-third probability for **L** and **S** to be antiparallel and a two-thirds probability for **L** and **S** to be perpendicular. Thus for a transition involving the four degenerate states, a circularly polarized photon (that selects a particular initial l_z) is three times more likely to generate a carrier with spin parallel to the initial l_z than antiparallel to it.

In a solid, although full rotational symmetry is lost, the band symmetries preserve some of these selection rules. In crystals of tetrahedral symmetry, such as zincblende semiconductors, the selection rules for transitions from the valence to the conduction band are identical to the selection rules between p and s states for the atom in free space. For example, the valence heavy-hole states are those for which **L** and **S** are parallel. Thus by setting the optical photon energy to the fundamental gap, the resulting excited conduction electrons and valence holes will be spin polarized. In quantum wells, where the energy of the heavy hole is split from that of the light hole by quantum confinement (and, sometimes, strain), fully spin-polarized carrier distributions are possible. Commonly used group IV, III–V, and II–VI semiconductors have a sufficient spin–orbit interaction to permit the straightforward optical generation of spin-polarized distributions. Shown in Figure 1 are the selection rules

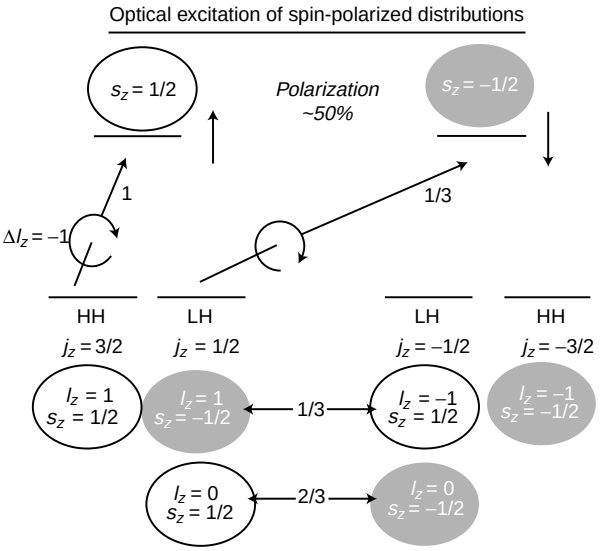


Figure 1 Selection rules for bulk zincblende crystals. Circularly polarized light changes the azimuthal orbital quantum number l_z by $+1$ or -1. Here $\Delta l_z = -1$ is shown. The orbital and spin quantum numbers for the two degenerate conduction band edge states and the four degenerate valence band edge states are shown – unshaded areas correspond to spin up and shaded to spin down. A transition with $\Delta l_z = -1$ couples one heavy-hole state to the conduction band up-spin state, and one light-hole state to the conduction band down-spin state. The orbital angular momentum of the heavy-hole state is entirely parallel to its spin angular momentum. In contrast only one-third of the light-hole state has antiparallel orbital and spin angular momentum. The rest of the light-hole state has perpendicular orbital and spin angular momentum. Thus the oscillator strength of the transition creating the conduction up-spin state is three times larger than that of the transition creating the conduction down-spin state, yielding a 50% spin polarized distribution. In quantum wells for which the heavy–light splitting is larger than the linewidth of the light source driving the transition, only the heavy-hole transition is excited and the spin polarization approaches 100%.

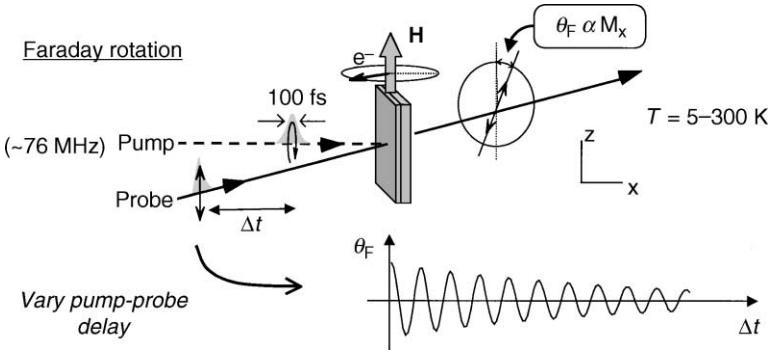


Figure 2 Diagram of the experimental geometry for time-resolved Faraday rotation. A pulse of circularly polarized light illuminates the sample, followed by a probe of linearly polarized light. The linearly polarized light can be considered as a sum of equal parts left and right circularly polarized light. The spin-polarization of the electron population (along the photon propagation direction) induces a difference in index of refraction for left and right circularly polarized light (circular dichroism). This produces a phase change between the left and right polarized components of the probe, which is measured as a rotation (θ_F) of the linear polarization direction. Monitoring the angle of rotation of the probe as a function of time directly indicates the spin polarization along the propagation axis of the photon as a function of time. Reproduced with permission from Awschalom DD, Samarth N and Loss D (eds) (2002) *Semiconductor Spintronics and Quantum Computation*. New York: Springer

which permit circularly polarized optical pulses to generate spin-polarized carriers in the conduction and valence band.

The evolution of a coherent spin-polarized population can be monitored through optical means as well. If the optical transitions used to generate the spin-polarized population are pumped hard enough, the transitions will be bleached. By measuring the polarization dependence of this bleaching the decay of the spin-polarized population can be determined. This can be done either by monitoring the transmission of time-delayed probe pulses of circularly polarized light, or by monitoring the rotation angle (Faraday rotation) of a pulse of linearly polarized light. The first of these depends directly on the imaginary component of the spin-polarized dielectric function, whereas the second depends directly on the real components. Shown in Figure 2 is the geometry for observing Faraday rotation from a spin-polarized distribution in the presence of an applied magnetic field perpendicular to the spin polarization.

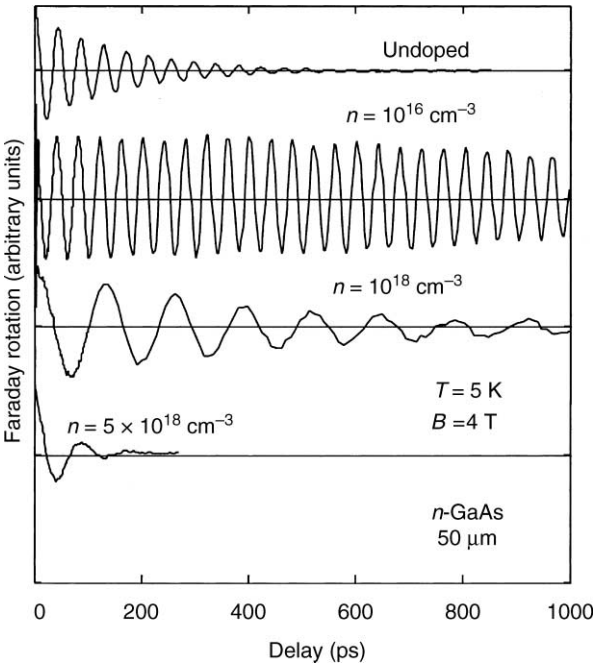


Figure 3 Time-resolved Faraday rotation measurements on bulk GaAs at several doping levels, indicating a sensitive doping dependence of the transverse spin coherence time T_2. T_2 for ideal doping densities can exceed 150 ns at 5 K. The magnetic field is 4 T, and the sample is 50 μm thick. Reproduced with permission from Kikkawa JM and Awschalom DD (1998) Resonant spin amplification in *n*-type GaAs. *Physical Review Letters* 80: 4313–4316.

Mobile Electron Spin Time Evolution: Decoherence and Precession

Two natural phenomena in the time evolution of a coherent spin-polarized population are decay of that population (decoherence) and rotation of the macroscopic magnetization in the presence of an applied magnetic field (precession). The time-scale of decay of the spin coherence can be very much longer than time-scales for the decay of other coherent phenomena in solids. Typical orbital coherence times, for example, are 100 fs, whereas spin coherence times can exceed microseconds. In Figure 3 are shown time-resolved Faraday rotation data for three bulk GaAs

crystals with different doping levels. The oscillations are characteristic of the spin precession, and the oscillation frequency is the Larmor frequency of the electron spin in this crystal. The crystal with the middle doping level, 10^{16} cm^{-3}, has oscillations that persist for the longest time, and therefore this material has the longest spin coherence time.

The dominant mechanism for the decay of these spin polarized distributions works via the same spin–orbit interaction that permits the generation and detection of these spin-polarized distributions. For mobile electrons in homogeneous environments, the mechanisms of spin coherence relate to the crystal momentum-dependence of the eigenstates of the electronic system. Because of time reversal invariance the electronic states with zero momentum (at the Γ point in the Brillouin zone) are doubly degenerate. The electronic states for momentum $\mathbf{k}$ near the Γ point occur in nearly degenerate pairs. Although it is common usage to refer to these eigenstates as degenerate spin eigenstates (correct in the absence of spin–orbit interaction), with spin–orbit interaction the states are not spin eigenstates and are not necessarily degenerate. As a consequence, when a spin-polarized carrier is scattered by ordinary, non-spin-dependent scattering from one momentum state $\mathbf{k}$ to another $\mathbf{k}'$, and then back again to $\mathbf{k}$, the effective orientation of the carrier spin will have rotated by an angle dependent on the intermediate momentum state $\mathbf{k}'$. For a population of carriers in thermal equilibrium this process will lead to misorientation of the individual spin moments from the mean, and thus a decoherence of the spin population. This process is characterized by an increasing decoherence rate with increasing orbital scattering. Thus the dirtier a material is made, the faster the spin orientations decohere.

In crystals with inversion symmetry the pairs of nearly degenerate states are precisely degenerate. Lack of inversion symmetry, and the consequential energy splitting, can be viewed as an effective (crystal) magnetic field, in which the approximate spin eigenstates precess. The magnitude and direction of this effective field depends on the orbital momentum $\mathbf{k}$, and because of time reversal invariance, the effective fields for $\mathbf{k}$ and $-\mathbf{k}$ are opposite. Thus an initially spin-polarized population will dephase due to experiencing different effective magnetic fields. Ordinary spin-independent scattering can now be seen as randomly changing the precession axis for the spin in this effective field. Once the information about the precession axis orientation has been lost the dephasing is not reversible and the spin orientations have decohered. As the precession angle increases linearly between scattering events, the longer the time between such scattering events the faster the spin population decoheres. Thus for this process the cleaner a material is made, the faster the spin orientations decohere. This precessional decoherence process dominates for mobile conduction electrons in bulk and quantum-well III–V semiconductors near room temperature.

The influence of the spin–orbit interaction is far stronger for the valence electrons than the conduction electrons for two reasons. First, as the effective magnetic fields come from relativistic transformations of internal electric fields, and the electric fields are largest near the nucleus, the valence electrons, which are closer to the nucleus than the conduction electrons, experience a stronger spin–orbit interaction. Second, the valence electrons have a dominant p-like character, whereas the conduction electrons have dominant s-like character (and thus lack orbital angular momentum to be coupled to the spin). Thus the spin coherence times of valence electrons (or holes) are commonly much shorter than the coherence times of the conduction electrons. Spin-flip exchange between conduction electrons and valence holes can therefore be the dominant source of decoherence for conduction electrons. This process can dominate at low temperatures in clean III–V semiconductors, and even up to room temperature in III–V quantum wells grown along the unusual (110) growth direction.

As a consequence one would expect the spin coherence times of electrons in n-doped semiconductors to be longer than those of electrons in p-doped or undoped semiconductors (in which excess holes must also be optically generated). This expected trend is present in Figure 3, for the n-doped materials have longer spin coherence times than the semi-insulating material. However, as the effect of the spin–orbit interaction increases with increasing momentum $\mathbf{k}$, the spin coherence times of the more heavily doped material are shorter than those of the lightly doped material. The doping corresponding to the longest spin coherence times may lie near the metal–insulator transition, although the theory of spin coherence times on the insulating side of that transition is not as well developed as on the metallic side.

Transport and Manipulation of Spin Coherence

An essential component of the motion of spin coherence in semiconductors is the response of carriers to electric fields. Outside of very unusual electronic states in semiconductors, such as the quantum Hall state, there is no spin–charge

separation of elementary carriers. The carriers in the semiconductors will move under the influence of an electric field, and if they happen to be spin-polarized, then a spin current will exist. Although there are numerous possible regimes of spin transport (at least as many as there are regimes of charge transport), only two will be presented here, both in the low-field diffusive limit.

Mobilities and diffusion constants are the material properties that describe the low-field charge transport in a homogeneous material. Spin transport can be described similarly, and the spin mobility and diffusion constant can be related to those of charge motion. The relationship between the two emerges from the nature of carrier packets that can be maintained under low-field conditions. Deviations from local charge neutrality cause very large space-charge fields in the semiconductor, and thus under low-field conditions approximate local charge neutrality must be maintained. Shown in Figure 4 are charge and spin packets in an n-doped semiconductor. The charge polarization packet, incorporating extra conduction electrons and valence holes, is shown in Figure 4a. The electron spin polarization packet shown in Figure 4b is qualitatively different from the charge packet as well as the hole spin polarization packet (shown in Figure 4c). Only the electron spin polarization packet does not incorporate both conduction electrons and valence holes.

The difference in packet structure profoundly changes the mobility and diffusion properties of the packet. For packets consisting of both electrons and holes the low-conductivity carrier species dominates the packet motion (in this case, the holes). For the electron spin packet, however, majority carriers dominate the packet motion. Thus motion of this spin packet is very rapid, and also sensitive to the higher Fermi temperature of electrons relative to holes. Experimentally this has been seen by optically creating a local spin-polarized packet of conduction electrons (an electron spin packet) in a thin sample of n-doped GaAs. The application of a lateral electric field drags that spin packet, and an optical probe measures the transport distance (thus the mobility) and the packet spread (yielding the diffusion constant). Such an approach is very similar in concept to the electrical probe of charge packet motion developed by Shockley and Haynes to measure minority hole mobility and diffusion in n-doped Ge.

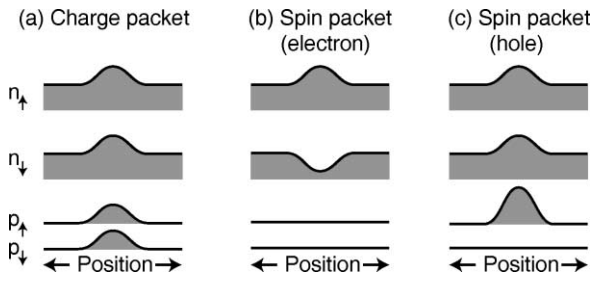


Figure 4 Inhomogeneous carrier packets in n-doped semiconductors in the low-field regime. The charge polarization packet (a) is constructed of approximately equal numbers of electrons and holes. The electron spin polarization packet (b) can be constructed without any excess holes, whereas the hole spin polarization packet (c) requires both excess electrons and excess holes. The presence of minority carriers in a packet significantly alters the mobility and diffusion properties of the packet.

The picture of approximately charge-neutral packets is very helpful for describing spin motion in homogeneous materials, but most semiconductor devices are based on inhomogeneous configurations of semiconducting material. For example, the most ubiquitous charge-based semiconductor device is the p–n diode, whose nonlinear current–voltage characteristics have been used as building blocks for numerous other devices. Such devices rely on the ability of a current to traverse an interface between two dissimilar materials – either different in doping density or different in host structure. Experiments in metallic ferromagnetic systems indicated that, whereas charge current was easily passed from one ferromagnetic metal to another through an interface, the same was not always true for the spin current. Differences in band structure or interface roughness are usually expected in such systems to strongly diminish the efficiency of spin current transfer across the interface.

Semiconductors, however, offer a possible new regime. The bulk band structures of most III–V semiconductors are very compatible – the variations in bandgap may be substantial, but the orbital wavefunctions are very similar. Current growth techniques permit the fabrication of atomically smooth interfaces between two semiconductors, so long as the lattice constants are almost identical. Thus one might expect the spin current to be transferred across such interfaces without tremendous degradation.

The transfer process for spin current across the GaAs/ZnSe interface has been investigated in detail. Here the ZnSe material was n-doped, and GaAs layers that were n-doped, p-doped, and semi-insulating were investigated. Electron spin polarization was generated optically in the GaAs layer. For n-doped GaAs in the absence of an electric field only a small signal for spin-polarized electrons in the ZnSe was detected. An applied electric field, however, which drove the electrons within the GaAs towards the interface, produced a substantial spin-polarized population in the ZnSe. This indicates that spin-polarized packets can be created in one

semiconductor, and then controllably moved into a second semiconductor with an applied electric field. This ability may become a key component of manipulating spin in semiconductors.

One reason the p–n diode is a paradigm of charge-based semiconducting electronics is that it exhibits the crucial phenomenon of minority spin injection. Manipulation of minority spin injection permits great control of charge current, for its amplitude is very sensitive to barriers (such as those introduced by gates). Spin-polarized minority carrier injection has not been shown yet experimentally, however the complementary process of minority carrier extraction has. This process corresponds to optically generating spin-polarized electrons in p-doped GaAs, where the electrons are minority carriers. The built-in electric field of the p–n junction drives these electrons out of the GaAs into the n-doped ZnSe with great efficiency. Thus one may imagine that spin-polarized minority carrier injection is plausible in these systems.

A hybrid of these two situations is the spin-polarized light-emitting diode. Here spin-polarized carriers (either electrons or holes) are electrically injected into the insulating region of p–i–n structure. In the insulating region they recombine with unpolarized carriers of the other species and emit light. Due to the selection rules described before, the emitted light is circularly polarized, and the degree of circular polarization is an indication of the spin-polarization of the recombining carriers. This device demonstrates that large numbers of spin-polarized carriers can be accumulated in the insulating region, although the spin-polarized carriers are not minority carriers.

Methods of manipulating electron spin coherence externally without electric fields are also under investigation. One demonstrated technique relies on the shift of energy levels in a semiconductor under illumination by laser light with frequencies below the fundamental bandgap. When the laser light is circularly polarized this 'ac Stark effect' shifts the energy levels of spin-up states and spin-down states differently, generating an effective magnetic field. Rotations of electron spin by angles greater than $\pi/2$ have been demonstrated in this ultrafast effective magnetic field in semiconductor quantum wells.

Transfer of Spin Coherence from Mobile Electrons to Other Spin Systems

Another fundamental phenomenon in spin coherence is the transfer of spin coherence from one system to another. If one spin system has an extremely long coherence time, this transfer can be viewed as establishing a reservoir of spin coherence that can be drawn on at later times. Explorations of transfer phenomena have focused on establishing spin coherence in a system that is coupled to the conduction electrons, such as core electron polarization or nuclear polarization. The polarization of these other spin systems can be detected because, when polarized, they exert an effective field on the conduction electrons. For example, when the nuclear system is polarized it exerts an effective field on the electrons through the hyperfine coupling.

It is also possible to polarize these other systems indirectly through their interaction with the conduction electron system. When the conduction electron system is driven far from equilibrium by optically pumping the spin polarization, the coupled spin systems also become spin polarized. In the case of the nuclear spin system this effect is the Overhauser effect. While continually generating conduction electron spin polarization with optical pumping the polarization of the nuclear system begins to rise with a time-scale characterized by the coupling between the nuclear spins and the conduction electron spins (hyperfine coupling). Because of the substantial difference between the nuclear spin coherence times parallel to and perpendicular to the applied magnetic field, the effective nuclear polarization is generated parallel to the applied magnetic field independent of the orientation of the conduction electron spin polarization. In bulk GaAs at 5 K the percentage polarization achieved was a few per cent, yielding an effective field through the hyperfine interaction of ~ 0.4 T. These nuclear polarizations and effective fields are increased over an order of magnitude in certain quantum wells, suggesting that nanostructure design can be used to control the effective fields. The induced nuclear polarization persists for time-scales characteristic of the nuclear longitudinal spin coherence time, on the order of minutes. If the optical pumping is turned off, the conduction electron spin polarization decays away rapidly, whereas the nuclear spin polarization persists for time-scales characteristic of the nuclear longitudinal spin coherence time, on the order of minutes. This nuclear hyper-polarization, and its manipulation through an optically pumped time-dependent conduction electron spin polarization, suggests a possible general technique of ultrafast all-optical nuclear magnetic resonance in solids.

A similar spin-coherence storage effect has been obtained for core electron polarization in Mn-doped ZnSe quantum wells. Again, the optically pumped

conduction electron spin polarization generates a nonequilibrium spin polarization of the paramagnetic core electron spins (3d in the case of Mn) through the core-conduction electron coupling. This core electron spin polarization persists long after the conduction electron spin polarization (and, in undoped systems, long after the conduction electron density) has decayed away. The magnetic polarization of the paramagnetic Mn spins dephases over the very long time-scales determined by spin–lattice relaxation, and thus is an interesting avenue towards information storage.

The above systems are not ferromagnetic – the spin systems involved in the process are paramagnetic spin systems coupled to the conduction electrons, which are 'super-polarized' through the conduction electron spin polarization. An alternative approach to manipulating the conduction electrons has recently been demonstrated using a thin layer of a ferromagnetic material placed on top of n-doped GaAs. Here the presence of this ferromagnetic material, even if only a few nanometers thick, leads to a remarkable enhancement of the nuclear spin polarization in the GaAs. For example in the presence of a mere 1000 G applied magnetic field, the ferromagnetic layer generates an effective field on the electrons from the nuclei of up to 9000 G. This effective field also depends on the orientation of the magnetism in the ferromagnetic layer. This intriguing effect has yet to be fully explored, but may yield a new type of magnetic 'gate' on the mobile electrons in a semiconductor without a direct form of electrical spin injection.

The Future of Semiconductor Spintronic Devices

Semiconductor charge-based electronics (e.g., transistors and integrated circuit chips) and magnetoelectronics (e.g., hard disk read heads) are two of the most successful electronic technologies of the twentieth century. A union of elements of these two technologies is anticipated in the area of semiconductor spintronics. Immediate approaches involve placing hybrid magnetic elements within known semiconductor devices. Semiconductor spintronic devices, however, are unlikely to compete with the most successful devices from each of these two areas – the semiconductor transistor is an extremely effective and efficient device, as is the magnetoelectronic hard disk read head. As has occurred with many other new technologies perhaps the most important applications of semiconductor spintronics have not even been envisaged yet.

Despite this, one significant application without a dominant technology suggests itself – that of quantum computation. Recent advances in the theoretical understanding of quantum information technology, including novel algorithms, error-correcting codes, and plausible scalable device architectures, suggest that calculations involving quantum mechanical coherent states are possible. Here the extremely long spin coherence times in semiconductor systems, and the demonstrated ability to manipulate that coherence on ultrafast optical times, indicate that spins in semiconductors (whether electron or nuclear) may constitute a good basis for a physical realization of a quantum computer.

See also

Quantum Optics: Quantum Computing with Atoms. **Semiconductor Physics:** Outline of Basic Electronic Properties.

Further Reading

Abragam A (1961) *Principles of Nuclear Magnetism*. Oxford, UK: Oxford University Press.

Awschalom DD and Kikkawa JM (1999) Electron spin and optical coherence in semiconductors. *Physics Today* 52: 33–38.

Awschalom DD, Samarth N and Loss D (eds) (2002) *Semiconductor Spintronics and Quantum Computation*. New York: Springer.

Ivchenko EL and Pikus GE (1997) *Superlattices and Other Heterostructures*, 2nd edn. New York: Springer.

Jeffries CD (1963) *Dynamic Nuclear Orientation*. New York: Wiley.

Meier F and Zachachrenya BP (eds) (1984) *Optical Orientation*. Amsterdam, The Netherlands: North-Holland.

Slichter CP (1992) *Principles of Magnetic Resonance*, 3rd edn. New York: Springer.

Wolf SA, Awschalom DD, Buhrman RA, *et al.* (2001) Spintronics: a spin-based electronics vision for the future. *Science* 294: 1488–1495.

Yafet Y (1963) g factors and spin-lattice relaxation of conduction electrons. In: *Solid State Physics*, vol. 14. New York: Academic Press.

Surface Photovoltage Spectroscopy of Semiconductors

L Kronik, Weizmann Institute of Science, Rehovot, Israel

Y Shapira, Tel-Aviv University, Tel-Aviv, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Surface photovoltage spectroscopy (SPS) is a well-established contactless technique for semiconductor characterization, which relies on analyzing illumination-induced changes in the surface voltage. SPS traces its origins to the pioneering work of Brattain and Bardeen in the early 1950s, and was extended into a powerful spectroscopic tool by Gatos, Lagowski and Balestra in the early 1970s. It has been used as an extensive source of surface and bulk information on various semiconductors and semiconductor interfaces. In the following, we present the basic theory behind SPS, its experimental setup, and a range of its applications.

Principles of Operation

In general, a surface is defined as a boundary of media with different physical properties. The periodic structure of an ideal crystalline semiconductor results in the appearance of allowed energy bands separated by forbidden energy gaps. In particular, semiconductors are characterized by a nearly full 'valence band', separated by a forbidden bandgap from a higher lying, nearly empty 'conduction band.' Electron transport takes place within the latter and hole transport takes place within the former. The mere termination of this periodic structure at the surface, as well as the chemical changes in the surface associated with this termination, allow for the formation of surface-localized electronic states within the semiconductor bandgap. Electrons or holes in these local states are referred to as 'trapped carriers', as opposed to the 'free carriers' in the conduction and valence bands.

The appearance of surface-localized states induces charge transfer between bulk and surface in order to establish thermal equilibrium between the two. The free carrier density in the vicinity of the surface therefore deviates from its equilibrium value in the bulk. This results in a surface space charge region (SCR), which is electrically non-neutral, implying a nonzero electric field in it and therefore a potential drop across it. Thus, even under equilibrium conditions the surface potential, V_s, is different from the electric potential far away in the bulk. As a specific example which illustrates these concepts, consider an n-type semiconductor with a depleted surface. The majority carriers are electrons and some of them have been trapped in surface states, such that their concentration in the vicinity of the surface is smaller than its equilibrium value but larger than that of the holes (see **Figure 1a**). The potential drop across the SCR is manifested by the bending of the semiconductor bands, which is such that electrons are repelled from the surface and holes are attracted to it, due to the trapped surface electrons. We note that, by definition, the energy band edge is lower the higher the electrical potential, so that a positive V_s corresponds to downward-bent bands.

Generally, the photovoltaic effect comprises an illumination-induced change in the equilibrium potential distribution and is typically the result of some charge transfer and/or redistribution within the device due to the incident illumination. This is the basic principle behind devices ranging from photodiode detectors to photovoltaic solar panels. A specific variant of the photovoltaic effect is the *surface* photovoltaic effect, which is at the center of this text. It is important to note that the formation of a surface photovoltage (SPV) occurs only if carrier photogeneration *per se* is followed by net charge redistribution. Usually, no significant driving force for such redistribution is found beyond the SCR and the underlying bulk remains quasineutral. Thus, we are interested primarily in the surface SCR.

The SPV mechanism depends strongly on whether the incident photon energy is larger or smaller than the semiconductor bandgap (denoted below as the super-bandgap and sub-bandgap, respectively). Dominant SPV mechanisms are demonstrated below for a depleted n-type surface. The dominant mechanism by which super-bandgap illumination results in surface photovoltage generation is shown in **Figure 1b**. The electric field in the SCR causes excess electrons to be swept away from the surface and excess holes to be swept towards it. This serves to reduce the density of surface-trapped electrons and decreases the band-bending. In a second mechanism, shown in **Figure 1c**, either electrons or holes are preferentially trapped at surface defects. This effectively charges the surface and thus modifies the

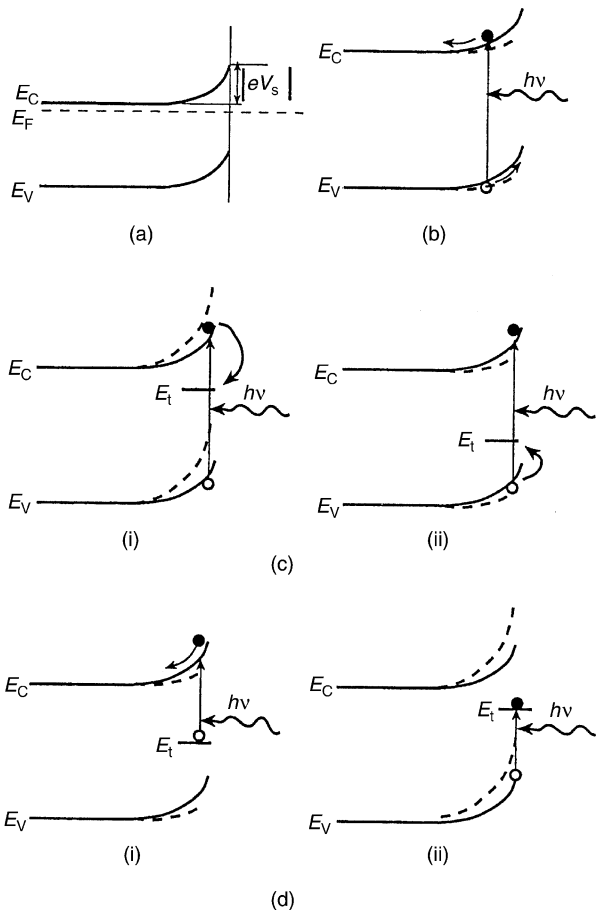


Figure 1 Schematic band diagrams of the surface space charge region at a depleted n-type semiconductor surface under different conditions: (a) thermal equilibrium; (b) super-bandgap illumination with carrier separation under an electric field; (c) super-bandgap illumination with preferential trapping of: (i) electrons, (ii) holes; (d) sub-bandgap illumination with excitation of trapped: (i) electrons, (ii) holes. In all diagrams, solid and dashed lines indicate band positions in the dark and under illumination, respectively. Straight arrows denote carrier generation and curved arrows denote carrier trapping. Adapted from Gal D, Mastai Y, Hodes G and Kronik L (1999) Bandgap determination of semiconductor powders via surface photovoltage spectroscopy. *Journal of Applied Physics* 86: 5573. Reproduced by permission of the American Institute of Physics.

surface potential. This second mechanism is usually apparent in bulk samples only if the equilibrium surface band-bending is fairly small, but increases in importance in polycrystalline materials with decreasing crystallite size, because of the increasing surface to volume ratio.

The most common mechanism for sub-bandgap SPV involves the direct modification of the surface charge, and hence potential, by excitation of trapped carriers, as shown in Figure 1d. Illumination by photons with energy $h\nu > E_c - E_t$ may produce electron transitions from a surface state at an energy E_t into the conduction band, where they are quickly swept to the semiconductor bulk by means of the surface electric field. Hence, the negative surface charge is reduced and the band-bending decreases. Conversely, illumination by photons with energy $h\nu > E_t - E_v$ may produce electron transitions from the valence band into a surface state situated at an energy E_t above the valence band maximum, E_v (which are equivalent to hole transitions from the surface state to the valence band). Such transitions increase the surface negative charge and therefore the surface band-bending. For a SPV to develop in this case, it is necessary to have a significant diffusion of the excess holes into the bulk and/or a significant recombination of electrons and holes inside the SCR, as there is no field-assisted driving force of holes into the bulk.

A typical SPV spectrum shows the SPV as a function of incident photon energy. Its most important features – the ones to watch for when attempting to interpret an SPV spectrum – are the sharp slope changes ('knees') associated with an abrupt onset of an additional carrier excitation mechanism. Consider, as an illustrative example, the top SPV curve in the right-hand part of Figure 2 (the rest of the figure is discussed below). In this curve, two 'knees' are shown. The first 'knee' (marked by an arrow) occurs at sub-bandgap photon energies and corresponds to the onset of carrier excitation from surface states (as in Figure 1d). The second 'knee' occurs in the vicinity of the bandgap energy and corresponds to the onset of band-to-band carrier excitation (as in Figure 1b,c). The proper interpretation of the energy position of these 'knees', their sign (negative or positive SPV change), and the relation of both to the SPV mechanism at work, is elaborated below in the 'Applications' sections.

So far, our discussion has been devoted solely to a study of semi-infinite, uniform bulk samples. However, actual semiconductor samples may be thin, i.e., not much larger than the diffusion length and/or the absorption length, so that excess carriers may be present near their back part as well. Many optically and electronically interesting structures (advanced detectors, transistors, lasers) consist of a multilayer structure, which has several buried interfaces.

Any change in the potential drop across a buried surface will also be measurable at the front surface, because potential drops are additive. For example, the photo-induced potential drop across a photodiode is measured at its contacts, which can be many microns away from its active region. Thus, the sensitivity of SPV measurements to a certain region in the semiconductor is limited only by the

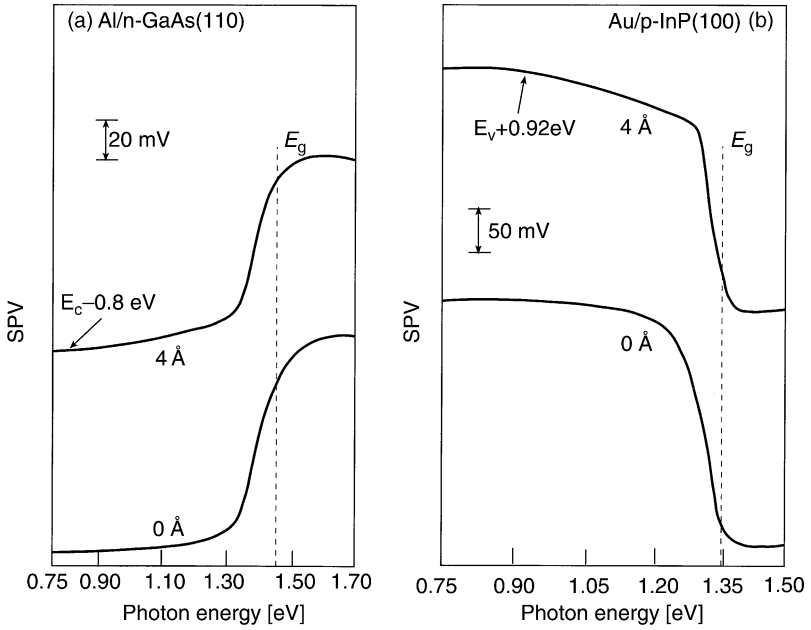


Figure 2 SPV spectra of: (a) an n-type GaAs(110) surface, UHV cleaved (top curve) and with an Al overlayer (bottom curve); (b) a p-type InP(110) surface, UHV cleaved (top curve) and with an Au overlayer (bottom curve). Arrows denote the onset of optically induced transitions and labels denote the corresponding defect level. Adapted from Burstein L, Bregman J and Shapira Y (1991) Characterization of interface states at III–V compound semiconductor–metal interfaces. *Journal of Applied Physics* 69: 2312. Reproduced by permission of the American Institute of Physics.

absorption length of the photons, i.e., by the possibility of introducing an excess of free carriers in the region of interest.

Experimental Details

SPV measurements are nontrivial because the surface potential is a *built-in* potential, rather than an external potential. In other words, the surface potential is not equal to the difference in Fermi levels between the front and back surfaces. Therefore, it cannot be simply measured using some form of voltmeter. Moreover, in the case of a free surface, the application of any contact to indirect electrical measurements of the built-in voltage (e.g., current–voltage or capacitance–voltage analyses) will invariably alter the surface properties and hence the quantity under measurement. Consequently, most of the prevalent techniques for measuring the surface potential (or at least changes of it upon illumination) are based on capacitive coupling schemes that do not require a direct electrical contact.

The two most commonly used methods rely on bringing a metallic probe close to the free surface, but never touching it, thus forming a parallel-plate capacitor. An alternating current, from which surface potential changes can be calculated, is then excited by (a) vibrating the metallic probe, which results in alternating capacitance, or (b) using 'chopped' illumination. The two approaches are known as the Kelvin probe and the ac-SPV schemes, respectively. In principle, they provide similar information, but their somewhat different physical principles result in several relative strengths and weaknesses, which make the method of choice application-dependent. In a nutshell, the Kelvin probe is more suitable for following phenomena with large time constants (in particular surface states with long relaxation times) and does not perturb the true band-bending of the free surface. The ac-SPV technique is more suitable for following fast relaxation phenomena, allows for a systematic modification of the surface band-bending in the dark as an extra degree of freedom, and does not require an ohmic back-contact.

Both Kelvin probe and ac-SPV measurements may also be employed in a scanning mode, which allows for recording lateral variations in the SPV, by scanning a small probe or a well-focused light beam, respectively, across the surface. Both approaches are typically limited to a lateral resolution roughly of the order of $\sim$1 μm. SPV measurements can also be performed by a limited modification of two high-resolution techniques for topographic

mapping of surfaces that have emerged in the 1980s: scanning tunneling microscopy (STM) and atomic force microscopy (AFM). This opens new horizons for SPV mapping with a resolution on the nanometer scale or even the atomic scale.

A schematic view of a complete, generic SPS setup is given in Figure 3. The sample under study is typically placed inside a metallic box, which serves as both a dark box and a Faraday cage. This box must be such that it effectively includes the probe.

Clearly, for spectroscopic purposes one needs to generate the SPV signal using a broadly tunable light source. The default choice would therefore be the use of a 'white-light' source in conjunction with a monochromator. Because the typical super-bandgap absorption coefficient is typically orders of magnitude larger than the sub-bandgap coefficient, SPS is vulnerable to the spurious contribution of high-order diffraction peaks and stray light. Use of a monochromator where these effects are as small as possible is therefore highly recommended. For some applications, use of an auxiliary light source (e.g., white light, laser light, ultraviolet light, etc.) allows for a 'photobias' of the system and/or for additional quantitative analysis.

The actual recording of an SPV spectrum is coordinated by a real-time computer program. This program controls the monochromator motor and thus wavelength changes, operates the control electronics of the probe, reads the measurement results by interfacing with the probe read-out electronics via an analog-to-digital converter, and finally stores the obtained spectrum for subsequent display and analysis.

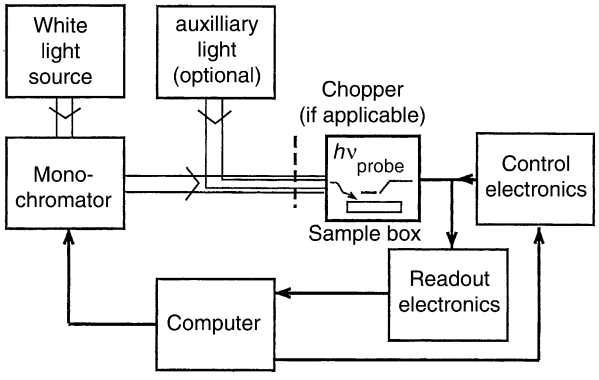


Figure 3 Schematic block diagram of a generic SPS setup. Adapted from Kronik L and Shapira Y (2001) Surface photovoltage spectroscopy of semiconductor structures: at the cross-roads of physics, chemistry and electrical engineering. *Surface and Interface Analysis* 21: 954. New York: John Wiley and Sons, Inc.

Applications I: Uniform Semiconductors

An elementary application of SPS is the approximate determination of the semiconductor bandgap. This determination is based on the large increase in absorption coefficient near the bandgap energy, E_g, found in most semiconductors. This increase brings about a significant change of the SPV signal, which is easily identified as a sharp change in slope of the SPV curve and is often the most significant one in a given spectrum. An example is shown in Figure 2, where the slope changes related to the bandgaps of GaAs and InP are clearly identified. Indeed, the bandgaps of most known semiconductors have been analyzed using this approach.

The use of SPS for extraction of E_g is at heart nothing more than an emulation of an absorption spectrum. However, as opposed to, e.g., transmission spectroscopy, SPS does not require light collection. It can therefore be performed on arbitrarily thick samples (or on layers sandwiched within a hetero-structure – see below) and does not require the sample to be removed from the substrate or grown on a transparent one. It is also inherently insensitive to reflection and scattering and is thus eminently useful for micro- and nanocrystallites.

One must bear in mind that the value of E_g obtained from an SPS plot by inspection is only approximate. Indeed, detailed experimental comparisons between SPV and absorption spectra reveal that the two are often similar, but never identical. Thus, the nominal bandgap is nearly always found within the onset of the largest SPV signal, but is usually relatively broad, the exact position of E_g within it is by no means obvious, and the error in E_g is often ~ 0.1 eV. However, it has been shown that this error can be reduced significantly – often by one to two orders of magnitude – by a careful quantitative analysis of the experimental data.

We now turn our attention to determining the type (p or n) of the semiconductor. This may be achieved with the help of the sign of the knee associated with the onset of the super-bandgap SPV. Most semiconductor surfaces are depleted, which means that the bands of p-type semiconductors are bent downwards towards the surface, whereas the bands of n-type semiconductors are bent upwards. Since super-bandgap illumination typically tends to decrease the surface band-bending, this would result in a positive SPV in n-type semiconductors and a negative SPV in p-type semiconductors. For example, the SPV spectra of n-type GaAs and p-type InP, shown in Figure 2, clearly feature opposite onset signs which obey the above rules.

While in many cases the type of the semiconductor is known a priori, this is not always the case and SPS can become very useful in determining the semiconductor type. For example, SPS was used for fast and nondestructive verification of the semiconductor type in GaN films, where the p-doping is known to be nontrivial and subject to compensation by its native n-type doping.

We note that under conditions where the SPV is dominated by trapping of carriers, as shown in Figure 1c, the sign convention described above clearly does not apply. For example, etched CdSe quantum dot films were found to exhibit a p-type response in a humid ambient and an n-type response in a dry ambient, on account of preferential trapping of holes or electrons in a dry or humid ambient, respectively.

A key strength of SPS is in the determination of the energy position and type of surface states. Because photons of sufficient energy may excite charge carriers from a surface state to a band or vice versa, we expect a knee in the SPV spectrum whenever the photon energy exceeds the threshold energy of a certain transition. If the latter involves excitation of electrons from the surface state to the conduction band (Figure 1d(i)),a positive change in the surface charge is induced, and hence a positive SPV is expected. Conversely, excitation of holes to the valence band (Figure 1d(ii)) makes the surface charge more negative and a negative SPV is expected. Thus, the combination of the SPV threshold energy and slope sign makes the determination of the approximate position of the surface state within the bandgap possible.

As an illustrative example, consider how SPS can be used to monitor metal-induced surface states, as shown in Figure 2. After cleaving in ultrahigh vacuum, neither GaAs nor InP surfaces display any knees associated with deep surface states, as expected for these surfaces. However, upon deposition of a very thin metallic overlayer, distinct sub-bandgap knees emerge. These knees, associated with positive and negative SPV changes at the Al/GaAs and Au/InP interfaces, respectively, indicate the formation of surface states situated ~0.8 eV below the conduction band edge and ~0.92 eV above the valence band edge, respectively.

Because SPS is contactless and nondestructive, it can be applied not only to clean surfaces, but also to real ones, in practically any ambient. This makes SPS highly suitable for a direct correlation of surface electronic structure with chemical treatments. One such area of active research is the study of surface passivation treatments. An illustrative example, in which the passivating effects of various liquids on a specific p-InP surface (denoted as (100)) was studied *in situ*, is shown in Figure 4. The negative

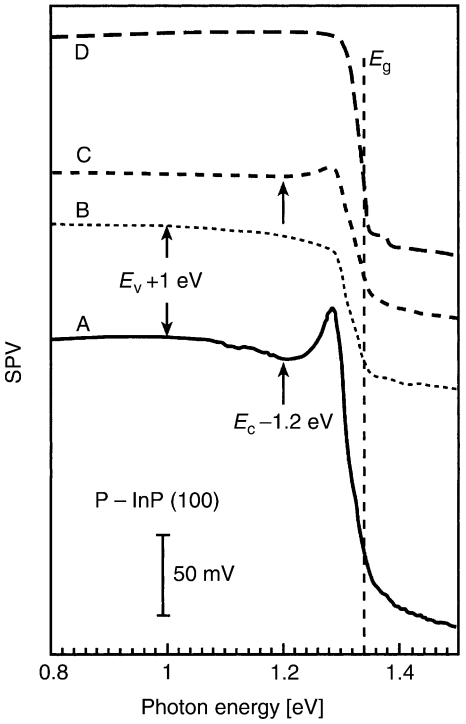


Figure 4 SPV spectra of the p-InP(100) surface in air and after successive immersion in (from top to bottom in the lower energy part): A – 18 MΩ H_2O; B – 0.4% HF; C – 30% H_2O_2; D – a 1.3 M Na_2S solution. Arrows denote the onset of optically induced transitions and labels denote the corresponding defect level. Adapted from Bastide S, Gal D, Cahen D and Kronik L (1999) Surface photovoltage measurements in liquids. *Review of Scientific Instruments* 70: 4032. Reproduced by permission of the American Institute of Physics.

SPV change at ~1.3 eV found in all spectra shown is clearly due to the onset of band-to-band transitions in the InP sample, with its sign indicating a p-type sample, as appropriate. The SPV spectrum of the water-exposed sample (which is similar to that obtained in air (not shown)) features two surface states: the negative SPV slope change, at ~1 eV, is due to a photo-induced population of a surface state situated ~1 eV above the valence band edge. The positive SPV slope change at ~1.2 eV is due to a photo-induced depopulation of a surface state situated ~1.2 eV below the conduction band edge. Interestingly, exposure to HF eliminates the latter state, but not the former. The opposite is true for exposure to H_2O_2. Finally, exposure to a Na_2S solution results in removal of both surface states, in agreement with the known passivating action of this solution. This is just one example of many illustrating the strength of SPS for revealing the link between chemical/physical treatments and device-affecting electronic properties, already at the material level.

While we have so far considered only surface states, SPS is sensitive to bulk states too, as long as charge excited from or trapped in them can be separated (typically under the electric field of a space charge region). Accordingly, SPS has been used for studying, e.g., Cr-related defects in GaAs or Fe-related defects in InP. The simplest distinction between surface and bulk states in SPS is obtained by examining whether the pertinent spectral feature disappears after surface (e.g., etching) or bulk (e.g., doping) treatment, respectively.

We note that in recent years, SPS has also been applied successfully along the lines outlined here for the study of bandgaps, doping type, and surface and bulk defects in the emerging field of organic semiconductors and organic treatments of semiconductors, as well as for nanostructures, such as fullerenes and nanotubes.

Additional, quantitative analyses of the SPV spectra allow for a determination of densities, thermal and optical transition rates, and the spatial site of gap states. Such analyses are usually based on the SPV response to a systematic modification of one or more external parameters, e.g., time (for transient responses), illumination intensity, temperature, etc.

Applications II: Multilayer Structures

Beyond the study of surface properties, the application of SPS to heterojunctions, multilayer structures, and actual device structures has been most rewarding in recent years. The SPS of multilayer structures is illustrated by Figure 5. This figure features the SPV spectra of a ZnO:Al/ZnO/CdS/Cu(In,Ga)Se_2 layered structure, used for the fabrication of thin-film solar cells. Distinct spectral knees at ~1 eV, ~2.4 eV, and ~3.1 eV, corresponding to the bandgaps of Cu(In,Ga)Se_2, CdS, and ZnO, respectively, are clearly observed. The unequivocal identification of all three bandgaps confirms the above-mentioned theoretical prediction of SPS sensitivity to buried interfaces. Interestingly, the magnitude of the SPV signal is seen to increase with increasing annealing time (in air, at 200 °C), in agreement with the increase in overall solar cell efficiency with such annealing.

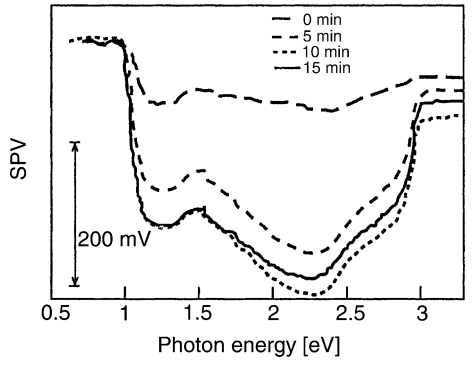


Figure 5 SPV spectra of ZnO:Al/ZnO/CdS/CIGS structures for various annealing times, in air, at 200 °C. Adapted from Gal D, Beier J, Moons E, *et al.* (1995) Band diagram and band line-up of the polycrystalline Cds/Cu(In,Ga)se_2 heterojunction and its response to air annealing. *AIP Conference Proceedings* 353: 453. Reproduced by permission of the American Institute of Physics.

Another feature of growing importance in SPS is that it allows for the characterization of quantum wells and other low-dimensional heterostructures. An illustrative example of the SPS of a multiple-quantum-well (MQW) GaAs/AlGaAs structure is shown in Figure 6a. The spectral knees at ~1.4 eV and 1.78 eV correspond to the onset of band-to-band absorption in the GaAs epilayer and the AlGaAs cap layer, respectively. In between these two knees, the shape of the SPV spectrum resembles typical absorption spectra of MQW structures and the observed MQW-related maxima agree well with calculated energy values of the hole level-electron level transitions denoted as 1HH-1E and 2HH-2E transitions. Note that before etching, a nonnegligible sub-bandgap SPV signal was apparent at photon energies below 1.38 eV, indicating the presence of optically active gap states. Since no such signal was observed after etching, these states were located at the external AlGaAs surface and were not associated with the MQW region. Moreover, the SPV spectrum in the MQW absorption energy range has not changed at all (other than undergoing a uniform shift) after etching, showing that the signal SPV in the MQW absorption range is not interfered by surface effects. Thus, all allowed hole–electron transitions have been resolved, at room temperature.

As SPV analyses of multilayer structures offer the possibility of performing contactless and nondestructive electrical and optical characterization of buried interfaces, they have found significant uses in the field of process monitoring and quality control of actual device structures. This is because the possibility of assessing the quality of a layered structure before any actual device processing takes place presents an opportunity for significant savings in time and money. Recent applications include the identification of defective batches and prediction of ultimate performance of solar cells and x-ray detectors, the prediction of energy levels and lasing wavelength in modern laser heterostructures, the assessment of bandgap narrowing and ultimate gain of heterojunction bipolar transistors, and the

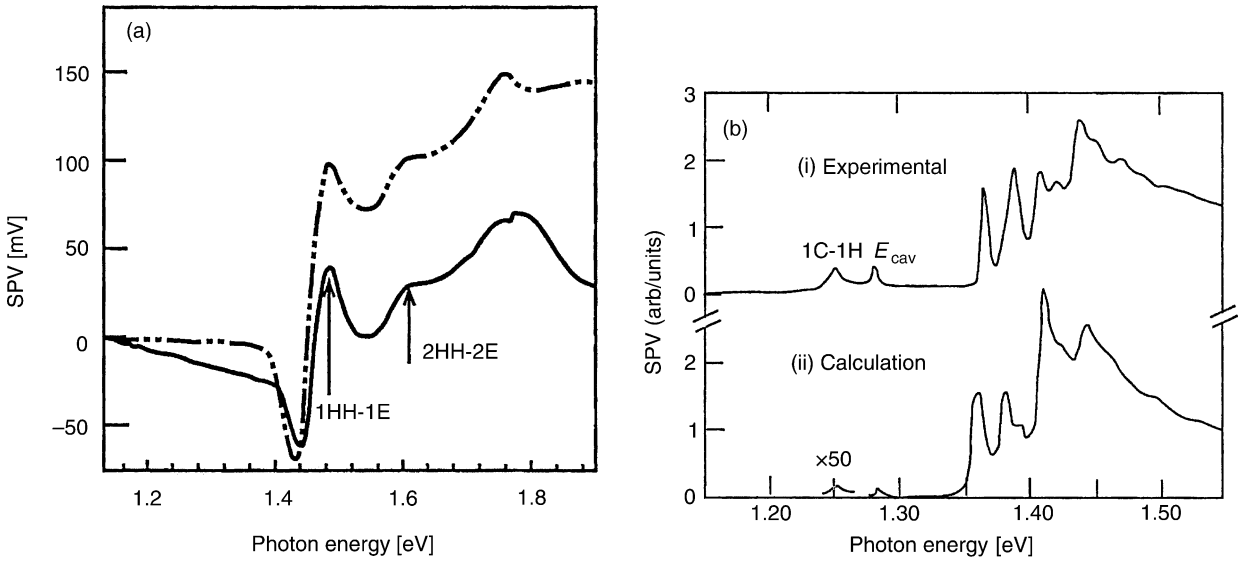


Figure 6 (a) SPV spectra of a GaAs/AlGaAs MQW sample before (solid curve) and after (dashed curve) etching. Adapted from Bachrach-Ashkenasy N, Kronik L, Shapira Y, *et al.* (1996). Surface photovoltage spectroscopy of quantum welts and superlattices. *Applied Physics Letters* 68: 879. Reproduced by permission of the American Institute of Physics. (b) Room temperature experimental SPV spectra (i) and calculated photo-excited carrier spectra (ii) of an InGaAs/GaAs/AlGaAs vertical-cavity surface-emitting laser structure. Adapted from Huang YS, Malikova L, Pollak FH, *et al.* (2000). Surface photovoltage spectroscopy, photoreflectance, and reflectivity characterization of an InGaAs/GaAs/GaAlAs vertical-cavity surface-emitting laser including temperature dependence. *Applied Physics Letters* 77: 37. Reproduced by permission of the American Institute of Physics.

characterization of the electronic structure of high electron mobility transistors and vertical cavity surface emitting lasers.

As an example of such studies, consider the recent characterization of an In-GaAs/GaAs/AlGaAs vertical-cavity surface-emitting laser, shown in Figure 6b. It is readily observed that SPS resolves both the gain providing 1C-1H transition and the lasing-wavelength-related cavity mode, as well as a rich oscillatory structure above ~1.37 eV due to interference effects from the distributed Bragg reflector stacks. The measured SPV spectrum is also in very good agreement with the calculated dependence of the photo-excited carrier density on photon energy, also shown in Figure 6b. In fact, SPS is the first contact-less technique to resolve all of these features in one measurement, and it also allows for a temperature-resolved determination of these quantities.

Summary

SPS features a theory that is relatively straightforward (at least at the qualitative level), and offers typical experimental setups that make the method contactless, nondestructive, ambient insensitive, and capable of any lateral resolution. It can provide a detailed picture of the electronic structure at surfaces, interfaces, heterojunctions, and even complicated multilayer structures. The wide range of SPS applicability also makes it easy to compare and link changes in electronic structure to changes in chemical/physical properties on the one hand, and electrical/optical device properties and performance on the other.

See also

Semiconductor Materials: Modulation Spectroscopy of Semiconductors and Semiconductor Microstructures. **Semiconductor Physics:** Band Structure and Optical Properties; Impurities and Defects; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Brattain WH and Bardeen J (1953) Surface properties of germanium. *Bell System Technical Journal* 32: 1–41.

Brillson LJ (1982) The structure and properties of metal–semiconductor interfaces. *Surface Science Reports* 2: 123–326.

Gatos HC and Lagowski J (1973) Surface photovoltage spectroscopy – a new approach to the study of high-gap

semiconductor surfaces. *Journal of Vacuum Science and Technology* 10: 130–135.
Kronik L and Shapira Y (1999) Surface photovoltage phenomena: theory, experiment, and applications. *Surface Science Reports* 37: 1–206.
Kronik L and Shapira Y (2001) Surface photovoltage spectroscopy of semiconductor structures: at the crossroads of physics, chemistry and electrical engineering. *Surface and Interface Analysis (special issue on Optical Techniques)* 21: 954–965 (main source for the present text).
Lüth H (1993) *Surfaces and Interfaces of Solids*, 2nd edn. Berlin: Springer-Verlag.
Many A, Goldstein Y and Grover NB (1971) *Semiconductor Surfaces*, 2nd edn. Amsterdam, The Netherlands: North-Holland.
Pollak FH (2000) Modulated electric fields produce semiconductor spectra. *Laser Focus World* 36: S9.

SOLITONS

Contents

Bright Spatial Solitons
Optical Fiber Solitons, Physical Origin and Properties
Soliton Communication Systems
Temporal Solitons

Bright Spatial Solitons

C Conti and G Assanto, University 'Roma Tre', Rome, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical spatial solitons, once considered an exotic and rather abstract concept belonging to the realm of mathematics or mathematical physics, nowadays are pursued both experimentally and theoretically in terms of their rich dynamics and variegated features, with important perspectives in applications to signal processing and switching, routing, filtering and – generally – light-by-light controlling. In an intuitive description of a bright spatial soliton, a self-focusing response balances natural diffraction of a light beam, allowing for its invariant propagation in transverse size and intensity. Spatial solitons can be regarded as the stationary or time-independent self-guided 'eigenwaves' of Maxwell equations in a material system with a nonlinear optical response. Basically, 'bright' spatial solitons (we will not address 'dark' cases here, i.e., a dip on an illuminated background) bear the same relationship to nonlinear optics in space as temporal, or fiber, solitons do in time. While both temporal and spatial solitons are special cases of light localization in the $(3 + 1)$ dimensions (light bullets) encompassing time and two transverse coordinates in space in addition to propagation; the spatial case requires particular attention compared to the fiber inasmuch as, in general, two transverse coordinates (2D) complement the direction of propagation. In other words, the scenario is potentially one-dimension richer than for 'standard' temporal solitons, where light confinement, i.e., the effect of the nonlinearity, takes place along a single coordinate. In this respect, the latter case finds a direct counterpart with spatial solitons in planar waveguides, i.e., $(1 + 1)$D solitons, where light confinement occurs linearly in one spatial dimension, providing group-velocity dispersion in time is replaced by beam diffraction in space, respectively.

When both transverse dimensions are involved, i.e. when a light beam propagates in bulk, in the simplest case of Kerr self-focusing, the nonlinearity can become overcritical and cause filamentation and catastrophic collapse. Additional mechanisms (such as saturation, nonlocality, or higher-order effects) or inherently different nonlinear responses (parametric or molecular), need therefore to be involved in the description of spatial solitons. In several instances, the mathematics differs from the nonlinear Schrödinger equation originally invoked to define a 1D bright optical soliton, leading to nonintegrable systems and

to a wider class of solutions, called solitary waves. Since solitary waves constitute the vast majority of the observed spatial light-localization phenomena, we will refer to spatial solitons even in those cases when their model is nonintegrable, defining them as non-diffracting beams in a self-focusing material. While retaining the distinction between 1D and 2D cases, in the following sections we will discuss bright spatial solitons and their features, grouping them on the basis of the underlying nonlinear mechanism and corresponding model, trying to provide a direct link with experimental observations. In doing so, we will leave out discussions on dark solitons and modulationally unstable phenomena, as well as more involved subjects ranging from higher-order to vector, vortex, photonic-crystal, and incoherent spatial solitons.

The General Model of Bright Spatial Solitons

Assuming the paraxial approximation to be valid, the general structure of the evolution equations which lead to spatial solitons propagating along z takes the so-called Foch–Leontovich form:

$$2ik\frac{\partial A}{\partial z} + \nabla_\perp^2 A = -2k^2\frac{\Delta n}{n_0}A \qquad [1]$$

where A is the complex amplitude of the linearly polarized electric field, $k = \omega/c = 2\pi n_0/\lambda$, with λ the wavelength, $\omega = 2\pi\nu$ the pulsation, n_0 the bulk refractive index, and $\Delta n(\mathbf{r})$ the index change induced by the nonlinear effect. In anisotropic materials, both n_0 and Δn may depend on the polarization of the envelope A. The latter is normalized such that the optical intensity is $I = |A|^2$. Hence the whole optical field is

$$E_{\text{opt}} = \text{Re}\left[\sqrt{\frac{2Z_0}{n_0}}A\text{e}^{-\text{i}\omega t + \text{i}kz}\right] \qquad [2]$$

with Z_0 being the vacuum impedance.

Hereafter we will neglect material losses and assume $\Delta n(\mathbf{r})$ to be real-valued. In the simplest case it is a function of the field intensity I calculated in $\mathbf{r}$. Typical models are the Kerr medium with $\Delta n = n_2 I$, or the saturable Kerr with $\Delta n = n_2 I/(1 + I/I_s)$ and I_s a saturation value; they are particular cases of local nonlinear media defined by a relation of the type $F[\Delta n(\mathbf{r}), I(\mathbf{r})] = 0$, with F a 'well-behaved' function.

There is also the possibility that the refractive index modification depends on the electromagnetic perturbation (e.g. the intensity) in a finite region, as happens with nonlocal solitons. In the latter case, F needs to be substituted by an integro-differential system of equations, i.e., it represents a functional between $\Delta n(\mathbf{r})$ and $I(\mathbf{r})$.

Relevant exceptions to eqn [1] are spatial solitons due to a quadratic nonlinearity, where a system of two coupled equations must be taken into consideration, and a description in terms of a refractive index change only holds in an approximate, and by no means physical, sense. Quadratic solitons will be discussed later. Here, instead of embarking into a general classification, we will describe the relationships to some representative cases. In this simple framework, spatial solitons are defined as solutions of the type $A = a\exp(\text{i}\beta z)$, with a real-valued and satisfying the equation:

$$\frac{1}{2k^2}\nabla_\perp^2 a + \frac{\Delta n(I)}{n_0}a = \frac{\beta}{k}a \qquad [3]$$

Equation [3] is the z-independent counterpart of eqn [1]. It has the form of the stationary Schrödinger equation (with a field-dependent Hamiltonian operator), where the kinetic energy and the potential are given by the diffraction term and by the nonlinear correction to the refractive index, respectively. Its solutions correspond to solutions of eqn [1] due to the balance between diffraction and nonlinearity. β is the nonlinear correction to the dispersion relation $k = n_0\omega/c$, i.e., it yields the resulting longitudinal wavevector $K = k + \beta$. Solitary solutions are often classified in terms of their dimensionality. Solutions of eqn [3], which depend on either 1 or 2 transverse coordinates, define 1D or 2D solitons, respectively. 1D solitons refer to situations in which a bulk nonlinear medium is illuminated with a highly elliptic beam (such that the diffraction along a transversal dimension, denoted by y, is negligible), or a planar waveguide is excited with a linear mode fixing the transverse profile in one of the dimensions (denoted by y). The nonlinear dynamics and the corresponding steady-state solutions are then described by eqns [1] and [3], with $\partial_y = 0$.

Spatial Solitons in Kerr Media

1D Kerr Spatial Solitons and the Nonlinear Schrödinger Equation

The simplest example of a spatial soliton is an optical beam diffracting in one transverse dimension in the presence of the so-called Kerr effect, i.e., a refractive index that increases linearly with the local intensity I:

$$n = n_0 + n_2 I \qquad [4]$$

where $n_2(>0)$ is measured in $m^2\ W^{-1}$ and has typical values of 10^{-20} for silica and 10^{-17} for semiconductors at photon energies below half-bandgap, i.e., leaving aside two photon absorption and higher-order effects. For 1D Kerr spatial solitons, eqn [1] becomes

$$2ik\frac{\partial A}{\partial z} + \frac{\partial^2 A}{\partial x^2} = -2k^2\frac{n_2}{n_0}|A|^2 A \qquad [5]$$

Equation [5] is known as the nonlinear Schrödinger equation (NLS), and is the basic model in describing temporal soliton, i.e., nondispersing pulses propagating in optical fibers, provided the variable x is intended as the temporal coordinate. The importance of eqn [5] is due to several remarkable, in some sense unique, mathematical properties, and to the fact that it can be considered a fundamental model for diverse nonlinear phenomena. This circumstance is often referred to as the 'universality of the NLS'.

In the framework of optical spatial solitons, the NLS relevance is moderated by the fact that most experimentally explored environments involve processes significantly more complicated than the Kerr effect. Nevertheless, it settles the basic principles and constitutes a reference point in optical soliton assessment.

Trying to gain some physical insight, the starting point is the beam spreading when $n_2 = 0$ in eqn [5], i.e., the linear regime corresponding to low intensities. The diffraction is inherent to the spectrum of plane waves in a finite beam (along x). To underline the role of the nonlinear response, one can consider the NLS without the diffraction term (the second derivative with respect to x), obtaining:

$$A(x,z) = \sqrt{I(x)}\,\exp\left[\mathrm{i}k\frac{n_2}{n_0}I(x)z\right] \qquad [6]$$

Equation [6] shows how the nonlinearity yields a phase-front curvature related to the sign of n_2: the medium is self-focusing (a converging beam) for $n_2 > 0$ or self-defocusing (a diverging beam) for $n_2 < 0$ (see **Figure 1**). In the absence of diffraction, i.e., for a plane wave, the Kerr effect modifies the field phase by increasing ($n_2 > 0$) or decreasing ($n_2 < 0$) the local refractive index (and hence the effective optical path) proportional to the intensity. For a beam of finite transverse extent, nonlinearity and diffraction act together. In a self-focusing medium, the concave phase-front due to diffraction can be balanced by the nonlinearity, and the natural tendency of a beam to spread can therefore be compensated. For this to happen, the length-scales associated to the diffraction, i.e., the Rayleigh range $L_R = kw_0^2$, and the Kerr effect must be comparable. A characteristic length $L_N = n_0/(kn_2I_0)$, over which the nonlinear phase shift equals 1, can be associated

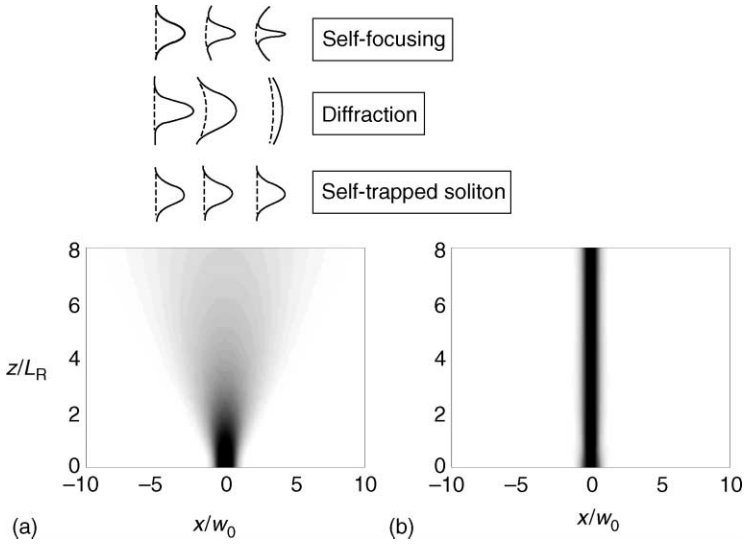


Figure 1 Kerr spatial solitons. Top: sketch of the mechanism leading to Kerr soliton formation. In the absence of diffraction, self-focusing of a Gaussian beam provides a converging phase-front (dashed lines). Conversely, diffraction by itself causes beam divergence. When these two effects act jointly, they can balance each other and soliton generation can be observed, with a flat phase-front. (Adapted with permission from Stegeman GI, Christodoulides DN and Segev M (2000) Optical spatial solitons: historical perspectives. *IEEE Journal of Selected Topics in Quantum Electronics* 6:1419–1426). Bottom: beam propagation method used to simulate the propagation of a one-dimensional Gaussian beam by integrating the nonlinear Schrödinger equation, for (a) low-power input, (b) input high enough to generate a self-trapped solitary wave. The transverse coordinate x is expressed in units of the beam waist, the propagation coordinate z in Rayleigh lengths.

to the nonlinearity, with I_0 the peak intensity across the beam profile. When the two lengths L_R and L_N are nearly equal, i.e., $w_0^2 I_0 \approx n_0/(k^2 n_2)$, the balance between the phase-front curvatures, due to diffraction and nonlinearity, yields a self-localized nondiffracting beam. Such a situation is described by the exact solutions of eqns [1] and [3], which takes the form:

$$A(x,z) = a(x)e^{i\beta z} \quad \text{with} \quad a(x) = \sqrt{I_0}\,\mathrm{sech}\left(x\sqrt{\frac{k^2 n_2}{n_0} I_0}\right) \qquad [7]$$

and $\beta = n_2 I_0/2n_0$ is determined by the peak intensity or, equivalently, by the power P_1 carried by the soliton per unit of transverse wavefront along y:

$$P_1 = \int_{-\infty}^{\infty} |A|^2\,dx = \sqrt{\frac{4n_0 I_0}{k^2 n_2}}$$

Note that the overall phase-front is flat, and the spatial extension in width yields $1/(n_2 k^2 I_0)^{1/2}$, as anticipated. The relationship between the soliton width – which determines the strength of diffraction – and the intensity – responsible for the nonlinear phase shift – is known as the existence curve, and is characteristic of the specific nonlinear mechanism.

Equation [7] is the fundamental soliton of the NLS and the paradigmatic 1D spatial soliton. Its sech shape specifically results from the type of nonlinearity; as discussed below, other effects (including parametric generation) can yield different profiles.

Figure 1 shows the simulation of a 1D Kerr soliton generated from a Gaussian input. Self-trapping is attained when the peak intensity or, equivalently, the power, reaches a threshold value depending on experimental details, the size of the nonlinear response, and the beam profile.

Figure 2 shows experimental results on the formation of a 1D Kerr soliton in a semiconductor (AlGaAs, $n_0 \approx 3.5$) waveguide. At low power (center panel) the output beam exhibits diffraction, whereas the soliton (bottom) reproduces the input beam profile (top panel) at the end facet of the sample. These results were obtained at $\lambda = 1.55$ μm, using pulsed laser beams, with temporal pulse durations of about 1 ps and peak powers of about 500 W; the generated 1D solitons had waists of tens of microns.

In a physically intuitive sense, self-trapping is due to the refractive index increase across the beam profile. Such an increase gives rise to a graded index optical waveguide able to confine all the plane wave components with angles of propagation within the numerical aperture of the input beam. Large spatial-frequency components k_x cannot be trapped and are emitted as radiation during soliton formation. It is possible to experimentally verify that this self-induced waveguide is also able to guide a weak probe uncorrelated to the soliton, provided it is above the cutoff for dielectric confinement. The use of spatial solitons for intensity-controlled or all-optical lightwave circuitry can therefore be envisaged. The first experiments on optical 1D Kerr spatial solitons date back to 1985, and were performed in carbon disulfide. Since then, 1D Kerr spatial solitons have been reported in glasses, semiconductors, and organic materials.

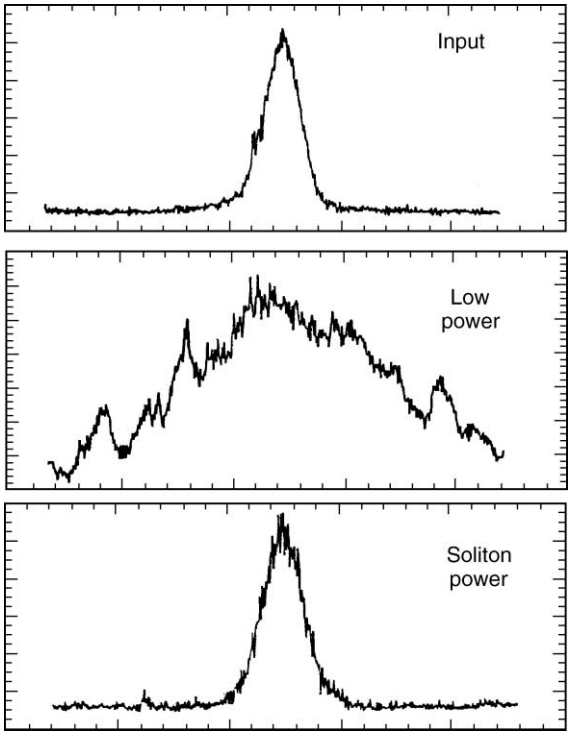


Figure 2 1D Kerr soliton in a planar AlGaAs waveguide. Transverse beam profiles at the input (top) and output facets of a planar waveguide in AlGaAs, at low (center) and high (bottom) power, showing diffraction and soliton formation, respectively. (Adapted with permission from Aitchison JS, Al-hemyari K, Ironside CN, Grant RS and Sibbett W (1992) Observation of spatial solitons in AlGaAs waveguides. *Electronic Letters* 28: 1879.)

Kerr Spatial Solitons in Two Transverse Dimensions: Saturable and Nonlocal Media

When the dynamics of the optical beam encompasses two transverse dimensions, the idealized discussion above no longer applies. As long as the medium is described by the simple Kerr law (eqn [4]), 2D propagation does not admit stable self-trapped

stationary beams. Because of this, two alternatives are available: (i) the beam diffracts; or (ii) the beam progressively self-focuses until it undergoes catastrophic collapse: in the absence of additional limiting phenomena, such as plasma formation or nonparaxial/vectorial coupling, its waist keeps shrinking with a corresponding unbound increase in peak intensity (beam power is conserved). Which regime actually applies, depends on whether the excitation is lower or higher than a self-focusing critical power $P_c = (0.61\lambda)^2/8n_0n_2$. The latter corresponds to a diffraction-driven divergence equal to the critical angle for total internal reflection in the induced waveguide. The self-focusing instability, first predicted by Kelly in 1965, is related to the so-called Derrick's theorem: particularly related to spatial solitons in a positive ($n_2 > 0$) Kerr medium, it states that the equilibrium between diffraction and nonlinearity in 2D is unstable.

The situation is markedly different when saturation is present: self-focusing arrests at the maximum value of the nonlinear refractive index increase, and stable self-trapped beams are allowed. The simplest model for a saturating nonlinearity yields the following stationary-state equation:

$$\frac{1}{2k^2}\nabla_\perp^2 a + \frac{n_2}{n_0}\frac{I}{1+I/I_s}a = \frac{\beta}{k}a \qquad [8]$$

Its solutions, while retaining a bell-shaped profile, depart from the purely Kerr model corresponding to $I_s \to \infty$, and can only be obtained numerically.

Gases, and specifically sodium vapors, were the first media in which spatial solitons of a saturable Kerr-like nonlinearity were observed in 1974. Historically, these were the first 2D optical spatial solitons.

Most recent developments on optical spatial solitons involve more complicated environments and nonlinear responses. In the next sections we will consider those most investigated experimentally: photorefractive solitons, nonlocal solitons in liquid crystals and parametric solitons.

Spatial Solitons in Photorefractive Media

In recent years, research on spatial optical solitons in photorefractives has been among the most active. Self-trapping in photorefractive materials was first predicted in 1992 and observed in 1993. However, a clear assessment of this result came with the demonstration of the so-called steady-state screening photorefractive solitons in 1994.

The principle underlying the photorefractive response can be described as a feedback mechanism relating to the optical induction of a static electric field **E**. In specific materials, such as semiconductors or ferroelectrics, a light beam generates free charges which diffuse and drift under an externally applied voltage. Owing to free carriers and donor/acceptor sites, the resulting net spatial charge produces a static electric field which, in turn, modifies the refractive index n_0 via the electro-optic effect, phenomenologically described by

$$\Delta n_{ij} = -\frac{1}{2}n_0^3 r_{ijk}E_k - \frac{1}{2}n_0^3\varepsilon_0^2(\varepsilon_r - 1)^2 g_{ijkl}E_kE_l \qquad [9]$$

Equation [9] is a series expansion of the index change in powers of the static electric field: all terms are present if the medium is noncentrosymmetric, otherwise only even-powered terms need be included. In essence, eqn [9] entails the feedback mediated by the static electric field on the optical wave.

The feedback leads to stable solutions under appropriate conditions. These are steady-state solitons if the inhomogeneous light-induced index change makes a suitable waveguide for the beam. The three main types of such self-localized beams are screening, photovoltaic, and centrosymmetric photorefractive solitons. Boundary conditions determine the induced static field to be inserted in eqn [9] and in the propagation eqn [1]. In the simplest cases we consider, the light-wave is linearly polarized and the tensorial nature of the index perturbation in eqn [9] can be neglected.

The equations encompassing most of the physics of the photorefractive effect constitute the Kukhtarev model and can be shown as:

$$\begin{aligned} &s(I + I_b + I_{dark})(N_d - N_d^i) - \gamma_r \rho N_d^i = 0 \\ &\nabla\cdot\mathbf{J} = \nabla\cdot[q\mu\rho\mathbf{E} + k_B T\mu\nabla\rho + \beta_{ph}(N_d - N_d^i)I\mathbf{c}_{opt}] = 0 \\ &\nabla\cdot\mathbf{E} + \frac{q}{\varepsilon_S}(\rho + N_A - N_d^i) = 0 \end{aligned} \qquad [10]$$

The relevant quantities in eqns [10] are: the current density **J**; the free charges (electrons) density ρ; the density of ionized (total) donors N_d^i (N_d), the density of negatively charged sites N_A, the photo-ionization cross-section s, the background density I_b (due to background laser light), the dark irradiance I_{dark} (sI_{dark} is the dark generation rate, typically $I_b \gg I_{dark}$), the recombination rate γ_r, the electron mobility μ, the low frequency dielectric constant $\varepsilon_S = \varepsilon_0\varepsilon_r$, the Boltzmann constant k_B, the versor $\mathbf{c}_{opt}$ of the optic axis, and the temperature T. In the second of eqn [10], the pertinent element β_{ph} of the photovoltaic tensor needs be taken into account only for the photovoltaic solitons described below. The above must be completed by boundary conditions for the bias V – the line integral of

E – and the geometry of the metal contacts. Typically, $N_d \approx 10^{18}$ cm$^{-3} \gg N_a$, $I_{dark} \approx 10^{-3}$–10^{-6} W/cm^2, $I_b \approx 10$ W/cm^2, $\varepsilon_r \approx 10^3$ and solitons are observed in the visible.

Equations [10] determine the field **E** in terms of the optical intensity I in steady-state, and **E** is inserted in eqn [9] to yield the optically induced index change. A major signature of photorefractive solitons is the cumulative process they rely on: the space-charge field **E** builds up in time along with the distribution of carriers, until equilibrium is reached. Hence the phenomenon does not require high optical intensities but, conversely, a sufficiently long time (of the order of minutes or hours). Photorefractive solitons are typically observed at modest excitations, as low as a few microwatts, with typical waists of tens of microns.

Screening and Photovoltaic Solitons

When an externally applied bias lowers the refractive index of a crystal through its electro-optic response, a light beam can screen this effect, defining a self-waveguide supporting a spatial soliton. In a noncentrosymmetric photorefractive, the beam raises the conductivity in the illuminated region, locally lowering the voltage drop compared to dark regions. The induced space-charge screens the applied static field, and it may bring the refractive index back to its unbiased value. This can be seen as a voltage divider formed by a small optically controlled resistance and two large ones corresponding to the dark regions. The resulting solitary waves are called screening photorefractive solitons. For them to be stable (in time), additional illumination of intensity I_b at the same wavelength is required on the crystal in order to increase the conductivity of the (dark) regions surrounding the soliton. The charge accumulation in the beam-tails would otherwise flatten the induced index distribution and destroy the soliton.

The simplest description of screening solitons is provided by the saturating relationship between the electric field E and the light intensity I:

$$E = -\frac{V}{L}\frac{1}{1 + I/I_b} \qquad [11]$$

to be used in eqn [9] to get the index perturbation via the electro-optic effect. L is the width across which the voltage is applied. Since the quadratic term in eqn [9] (weighed by the tensor g_{ijkl}) can be neglected in noncentrosymmetric crystals:

$$\Delta n \approx -\frac{1}{2} n_0^3 r_{eff} E \qquad [12]$$

where r_{eff} is the relevant component of the electro-optics tensor; typically $n_0 \approx 2$, $r_{eff} \approx$ tens or hundreds of pm/V, and the overall index perturbation is of the order of 10^{-4}–10^{-3}, with $V/L \approx 10^3$ V/cm. This description applies to 1D solitons; for 2D solitons, sometimes referred to as 'needle' solitons, the whole set of constitutive equations must be solved numerically. Nevertheless, in most cases the saturating model, even if not rigorous, describes well solitons and related phenomena both theoretically and experimentally, especially when nonlocal effects can be ignored.

Figure 3 shows the propagation of a 2D spatial screening soliton (top) compared to the linear case (bottom). The diffractive regime is realized by artificially 'turning off' the nonlinearity acting on the applied voltage (when $V = 0$ in eqn [11] then $\Delta n = 0$, corresponding to no induced perturbation).

Another class of photorefractive spatial solitary waves is that of photovoltaic solitons. In this case the electric field in the crystal has an additional component through the photovoltaic effect, the latter being relevant in materials such as $LiNbO_3$. To sustain bright solitons the index perturbation due to the photovoltaic current, which induces the space-charge field, must be positive. This is the case for Cu:KNSBN, where 2D photovoltaic solitons were first observed.

The space-charge field, due to the photovoltaic effect, can be derived from the Kukhtarev model:

$$E = -E_{ph}\frac{I/I_b}{1 + I/I_b} \qquad [13]$$

with E_{ph} is a constant depending on the material ($E_{ph} \approx 10^3$ V/cm) and I_b is the dark-irradiance controlled by the background laser light. Typical

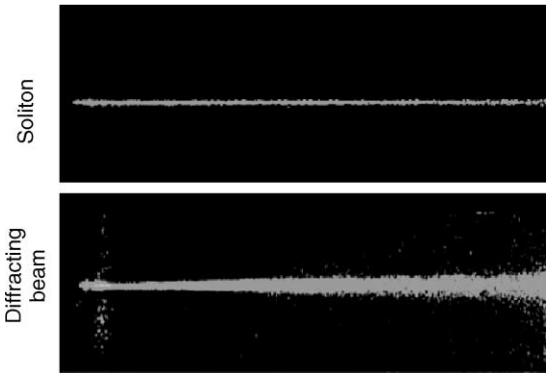


Figure 3 2D photorefractive soliton. Top view photograph of a 10 mm wide spatial soliton propagating in strontium barium niobate (top), and for comparison, the same beam diffracting when the nonlinearity if 'turned off' (bottom). (Adapted with permission from Stegeman GI, Christodoulides DN and Segev M (2000) Optical spatial solitons: historical perspectives. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1419–1426).

geometries account for the tensorial nature of the effect, and are arranged to make I_b of the same magnitude of the beam intensity I. If $E_{ph} > 0$, as in Cu:KNSBN, eqn [13] in eqn [12] shows that a positive index change of the saturating type can be induced and support bright solitons. This model is successfully used for both 1D and 2D bright photovoltaic spatial solitons.

Photorefractive Solitons in Centrosymmetric Media

Most experiments on screening solitons were performed in ferroelectric noncentrosymmetric crystals. When brought above the Curie temperature T_c, such materials exhibit a phase-transition to a centrosymmetric paraelectric state. Crystals such as Potassium Lithium Tantalate Niobate have room-temperature phase-transitions, conveniently enabling experimental investigations. In paraelectrics, the second-order electro-optic tensor r_{ijk} is zero because of symmetry, but the photorefractive effect is still available through the third-order coefficient g_{ijkl} in eqn [9]. By properly choosing the orientation, eqn [12] can then be replaced by

$$\Delta n = \frac{1}{2} n_0^3 \varepsilon_0^2 (\varepsilon_r - 1)^2 g_{eff} E^2 \qquad [14]$$

Operation near T_c greatly enhances the relevant electro-optic coefficient $g_{eff} > 0$ ($g_{eff} \approx 0.1\ m^4\ C^{-2}$), making the observation of photorefractive screening solitons possible. Using eqn [11] in eqn [14] the model is saturating and works for both 1D and 2D bright solitons.

Thanks to the low powers involved and a mature crystal technology, the photorefractive effect and related soliton phenomena have contributed several impressive results in optical morphology of nonlinear light propagation, including incoherent and white light excitations.

Spatial Solitons in Nonlocal Media, Liquid Crystalline Media

A link, such as the Kerr law (eqn [4]), between the index perturbation and the optical intensity is 'local', because Δn at a given point (x, y, z) only depends on the intensity at the same location. Otherwise the medium is said to be 'nonlocal', e.g., one of those described by the general Kukhtarev model (eqn [10]) although, sometimes well approximated by eqns [11] or [13].

There are various mechanisms yielding nonlocality, mainly through spatial diffusion of the index perturbation far from the excitation. Remarkable nonlocal effects are expected when the spatial features of the beam, e.g., its waist, are significantly smaller than those of the induced polarization.

One of the most important effects of nonlocality is the stabilization against the catastrophic self-focusing in 2D, as experimentally investigated by Suter and Blasberg in 1992. Evidence of a highly nonlocal behavior has been very recently reported in liquid crystals, where the nonlinearity can be thermal or re-orientational. In the former case the index perturbation is much wider than the beam waist due to thermal diffusion associated to (absorption mediated) heat transfer. In the latter case the elastic properties of the medium produce a widening of the self-waveguide far from the beam axis. In nematic liquid crystal (NLC), the re-orientational response stems from the optically induced tilt of the elongated molecules towards the linearly polarized field versor. The light forced re-orientation increases the refractive index, thus creating the self-guiding conditions for spatial solitons. However, since in the nematic phase the molecules are linked together in a wide-scale order, the region (kernel) affected by this process is determined by the elastic properties of the medium and can extend to tens of microns.

An effective approach to describe spatial solitons in a highly nonlocal medium (where the kernel is much larger than the beam) postulates that the punctual relationship between Δn and I is replaced by a link between Δn and the beam power P. In such a way the transverse integral of the intensity is fully embedded in the index perturbation.

Assuming $n = n(P, r)$ with the beam mainly subject to the refractive perturbation near its axis, one can write:

$$n^2(r, P) = n_0^2(P) - r^2 \gamma(P) \qquad [15]$$

with $n_0^2(P) \cong n_0^2$ the on-axis value. $\gamma(P)$ depends on the specific geometry and determines the wavefront convexity at the origin. The overall profile is basically determined by the constitutive relationships of the medium. The propagation equation can be shown as

$$2ik \frac{\partial A}{\partial z} + \nabla_\perp^2 A = \frac{k^2}{n_0^2} \gamma^2(P) r^2 A \qquad [16]$$

Equations like [16] also describe the transverse profile of a beam linearly propagating in optical fibers, as well as the quantum harmonic oscillator (interpreting z as time). The soliton is its nondiffracting solution, a particular stationary case of a more general Gaussian beam with waist variable in propagation.

For intensities:

$$I(z) = \frac{P}{\pi w^2(z)} e^{-r^2/w^2(z)} \qquad [17]$$

the spot-size obeys:

$$\frac{w^2(z)}{w_0^2} = 1 + \left[\frac{1}{k_0^2 w_0^4 \gamma^2(P)} - 1 \right] \sin^2 \left[\frac{n_0 z}{\gamma(P)} \right] \qquad [18]$$

Therefore, for a Gaussian excitation, the waist oscillates along z between the values w_0 – in general determined by the launch conditions – and $w_m = w_0 \gamma(P_S)/\gamma(P)$. P_s is the 'soliton power' for which $w_m = w_0$ and $k^{02} w_0^4 \gamma^2(P_S) = 1$, hence for $P < P_s$ or $P > P_s$ the beam pulsates cyclically with z (breathes). Such a dynamic balance between diffraction and nonlinearity stems from the spectral broadening associated to an increased beam shrinking: when the spectrum widens sufficiently, diffraction starts to prevail and the beam spreads again. This process repeats itself in propagation.

The stationary solution for $P = P_s$ corresponds to a point in the existence curve of waist versus power, i.e.:

$$k_0 w_0^2 \gamma(P_S) = 1 \qquad [19]$$

Such exact balance is critical, and experimentally observed solitary waves in highly nonlocal media appear as breathers.

Figure 4 displays the intensity profile of a soliton propagating in nematic liquid crystals, as acquired by a digital camera placed on the top of the cell and retrieving the scattered light (see Figure 5). The waist oscillation in the self-trapped beam and the dependence of the period on excitation are apparent.

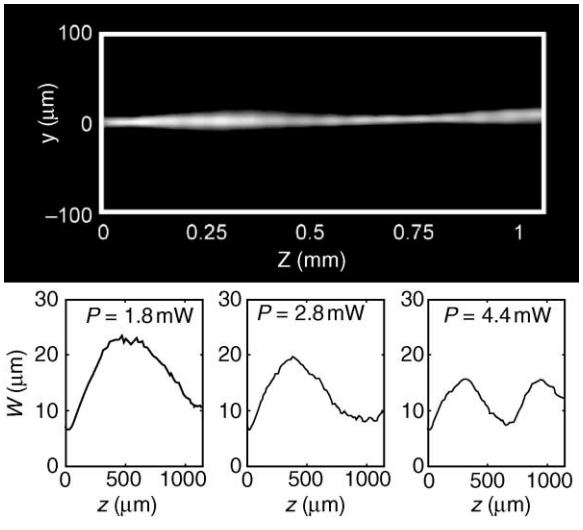


Figure 4 Nonlocal breather. (Top) top view photograph of a spatial soliton in NLC. Beam waist oscillations are apparent in propagation, as graphed (bottom panels) versus z for various excitations.

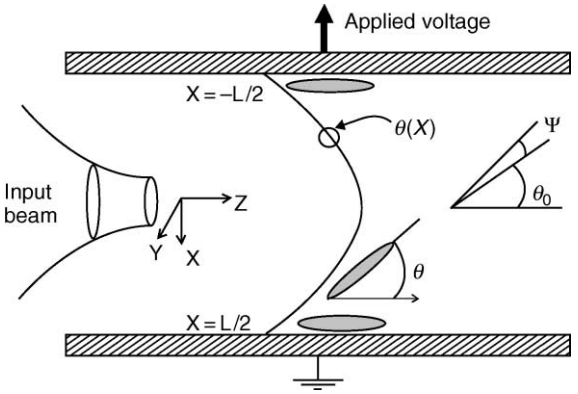


Figure 5 Sketch of a liquid crystal cell for the observation of nonlocal solitons. $\theta(x)$ is the orientation of the main axis of the ellipsoidal molecules, fixed by an external bias. The alignment is planar, with molecules parallel to the cell facets. The optical beam propagates in the center of the cell, where $\theta(x) = \theta_0$, and induces an additional re-orientation ψ. (Adapted from Conti C, Peccianti M and Assanto G (2003) Route to nonlocality and observation of accessible solitons. *Physics Review Letters* 91: 073901).

A specific highly nonlocal nonlinear system is formed by a linearly polarized beam in NLC with molecules pre-tilted (by an external electric or magnetic field) at $\theta_0 = \pi/4$ with respect to the input polarization. As sketched in Figure 5, the light wave induces an additional tilt Ψ. At the lowest order of approximation, the relevant set of equations can be shown as:

$$\begin{aligned} 2ik \frac{\partial A}{\partial z} + \nabla_\perp^2 A &= -k^2 \frac{n_a^2}{n_0^2} \Psi A \\ K \nabla_\perp^2 \Psi - G\Psi + \frac{n_a^2}{2cn_0} |A|^2 &= 0 \end{aligned} \qquad [20]$$

with K the NLC elastic constant ($K \approx 10^{-11}$ N, for NLC of the E7 type), $n_0 \approx 1.5$, n_a^2 the optical anisotropy ($n_a \approx 1$), i.e., the difference between relative permittivities along the two principal axes, G a constant determined by the specific geometry of the NLC cell and the pre-tilt angle θ_0. Since, for $K = 0$, eqn [20] represents a Kerr medium, it can be considered one of the simplest nonlocal generalizations of local models for spatial solitons. The length $R_c = \sqrt{K/G}$ quantifies the diffusion of the index perturbation away from the beam axis, such that Ψ decays according to the modified Bessel function $K_0(r/R_c)$. Typically, $R_c \approx 20\ \mu$m and solitons are observed in the visible and near-infrared.

Solitary wave profiles are found with the ansatz $A = a\exp(\mathrm{i}\beta z)$, obtaining:

$$\begin{aligned} \frac{\nabla_\perp^2 a}{2k^2} + \frac{n_a^2}{2n_0^2}\Psi a &= \frac{\beta}{k}a \\ R_c^2\nabla_\perp^2\Psi + \frac{R_c^2 n_a^2}{2Kcn_0}a^2 &= \Psi \end{aligned} \qquad [21]$$

No analytical solutions of eqns [21] are known in 2D, hence the profiles have to be derived numerically. Conversely, it can be shown that 1D sech^2 solutions exist (see next section on parametric solitons). From eqn [21], the whole family can be parameterized by a dimensionless $\alpha = 1/(2k\beta R_c^2)$. Since $(1/2k\beta)^{1/2}$ measures the soliton waist (see eqn [7]), α relates the beam size and the extension of the induced perturbation. Small α solutions access the highly nonlocal regime, as shown in Figure 6, where the ratio R between the index perturbation and the soliton waist is graphed. The insets show sample profiles of a and ψ in two limits, respectively. On the other hand, when $K = 0$, $R_c = 0$, and $\alpha \to \infty$, the local limit yields an index perturbation comparable in size to the beam waist.

It is worth pointing out that a model such as eqn [16], also derived in the framework of plasma physics, predicts unconditionally stable self-trapped beams. Nonlocal solitons realize an energy minimum, thus preventing beam collapse such as in saturable (Kerr) media. The oscillations observed in experiments can be regarded as fluctuations around the lowest energy state.

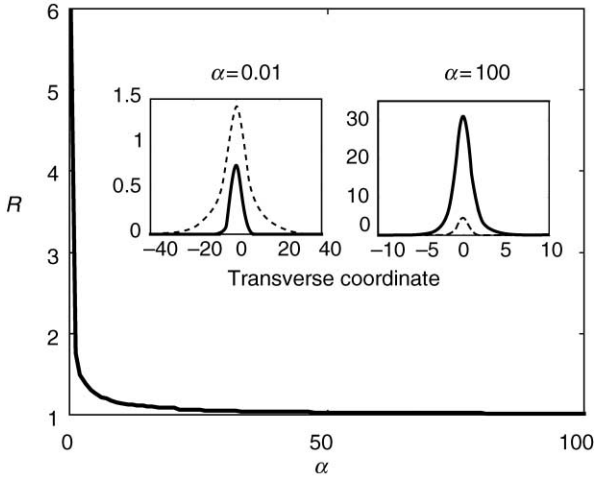


Figure 6 Spatial solitons versus nonlocality. Graph of the ratio R between the waist of the induced index perturbation in a nonlocal medium and the soliton waist versus parameter α. The insets show (in normalized units) soliton (solid line) and perturbation profiles (dashed line) for two different α, i.e., in the local ($\alpha = 100$) and highly nonlocal ($\alpha = 0.01$) regimes. (Adapted from Conti C, Peccianti M and Assanto G (2003) Route to nonlocality and observation of accessible solitons. *Physics Review Letters* 91: 073901).

Parametric Spatial Solitons in Quadratic Media

Solitons due to parametric processes, such as second-harmonic generation, belong to a different class with respect to previous sections. In fact, no refractive index modification can be physically associated to parametric effects. Parametric solitons are characterized by a power spectral density peaked around at least two carrier harmonic frequencies. It is the interaction between these components which provides the diffraction beating mechanism. Their existence was predicted in the early 1970s by Sukhorukov and Karamzin, while the first experimental evidence was reported in 1995.

The simplest process to address is second-harmonic generation (SHG): a fundamental frequency (FF) wave at ω interacts with a component at 2ω. The latter can be spontaneously generated in noncentrosymmetric crystals (e.g., $LiNbO_3$, $KNbO_3$, BBO, KTP, etc.) or externally seeded.

For plane waves, this process is ruled by:

$$\begin{aligned} \mathrm{i}\frac{\mathrm{d}A_1}{\mathrm{d}z} + \frac{\omega d_{\mathrm{eff}}}{c}A_1^* A_2\,\mathrm{e}^{-\mathrm{i}\Delta kz} &= 0 \\ \mathrm{i}\frac{\mathrm{d}A_2}{\mathrm{d}z} + \frac{\omega d_{\mathrm{eff}}}{c}A_1^2\,\mathrm{e}^{\mathrm{i}\Delta kz} &= 0 \end{aligned} \qquad [22]$$

with $\Delta k = k_1 - 2k_2 = 2\omega/c(n_1 - n_2)$ the phase mismatch, d_{eff} the nonlinear coefficient, n_2 and n_1 the indices at SH and FF, respectively, and

$$E_{\mathrm{opt}} = \mathrm{Re}\left[\sqrt{\frac{2Z_0}{n_1}}A_1\mathrm{e}^{-\mathrm{i}\omega t+\mathrm{i}k_1 z} + \sqrt{\frac{2Z_0}{n_2}}A_2\mathrm{e}^{-2\mathrm{i}\omega t+\mathrm{i}k_2 z}\right]$$

the overall electric field. d_{eff} relates to the quadratic susceptibility $d_{\mathrm{eff}}^{(2)}$ (measured in mV^{-1}) through to $d_{\mathrm{eff}} = d_{\mathrm{eff}}^{(2)}\sqrt{2Z_0}/(n_1\sqrt{n_2})$, and typically $d_{\mathrm{eff}} \approx 10^{-10}\,\mathrm{m}^2\,\mathrm{W}^{-1/2}$.

Equations [22] show that, in general, the harmonics with amplitudes A_1 and A_2 affect each other's phase and intensity during propagation. The solution with constant moduli, i.e.:

$$\begin{aligned} A_1 &= \sqrt{I_1}\exp\left[\mathrm{i}\frac{\omega}{c}\Delta n_1 z\right] \\ A_2 &= \sqrt{I_2}\exp\left[\mathrm{i}\frac{2\omega}{c}\Delta n_2 z\right] \end{aligned} \qquad [23]$$

with I_1 and I_2 the respective intensities, provides a 'pure' phase variation and 'apparent' index changes Δn_1 and Δn_2 on either wave. The latter satisfy:

$$\Delta n_1 = d_{\text{eff}}^2 \frac{I_2 - I_1/2}{n_1 - n_2}$$
$$\Delta n_2 = d_{\text{eff}}^2 \frac{I_1}{2I_2} \frac{I_2 - I_1/2}{n_1 - n_2} \qquad [24]$$

with the constraint (for plane waves to exist):

$$I_1 = 2I_2 - 2\frac{n_1 - n_2}{d_{\text{eff}}}\sqrt{I_2} \qquad [25]$$

Equations [24] encompass a great deal of the physics in self-trapping via SHG. When eqn [25] is satisfied (at least approximately), the effect of SHG is exclusively on the phase. Assuming eqn [24] is valid for (wide) beams, the propagation in the presence of diffraction can be described by two coupled equations of the type [1] with relevant perturbations (eqn [24]). This shows that the sign of the term $(2I_2 - I_1)/(n_1 - n_2)$ is crucial, because focusing at both harmonics ($\Delta n_{1,2} > 0$) requires $(2I_2 - I_1)/(n_1 - n_2) > 0$.

When just a fundamental frequency (ω) is launched, solitons are generated for $\Delta k > 0$: a small 2ω component is provided via SHG and $I_1 > 2I_2$. In fact, since SHG ($\omega + \omega = 2\omega$) is followed by difference frequency generation ($2\omega - \omega = \omega$, DFG), for any finite mismatch $\Delta k > 0$, a portion of the FF travels for a while at 2ω before it is back-converted: this is commonly named 'cascading' and, as a consequence, part of the FF undergoes a positive phase shift. The converse happens in the DFG + SHG process. Finally, if an SH beam with a small FF seed is input, solitons form for $\Delta k < 0$ being $2I_2 > I_1$. From a physical point of view, these mechanisms create the phase-front distortion required to counteract diffraction. Figure 7 is an artist's sketch on the formation of an SHG soliton.

Equations [24] are approximately valid and rely on a plane-wave insight, but they point out the wealth of possibilities offered by three-wave interactions for light transverse localization via parametric processes. The full set of coupled equations describing parametric solitons from SHG differs from eqn [1]:

$$\mathrm{i}\frac{\mathrm{d}A_1}{\mathrm{d}z} + \frac{1}{2k_0 n_1}\nabla_\perp^2 A_1 + \frac{\omega d_{\text{eff}}}{c} A_1^* A_2 \mathrm{e}^{-\mathrm{i}\Delta k z} = 0$$
$$\mathrm{i}\frac{\mathrm{d}A_2}{\mathrm{d}z} + \frac{1}{4k_0 n_2}\nabla_\perp^2 A_2 + \frac{\omega d_{\text{eff}}}{c} A_1^2 \mathrm{e}^{\mathrm{i}\Delta k z} = 0 \qquad [26]$$

as it includes diffraction terms added to eqn [22].

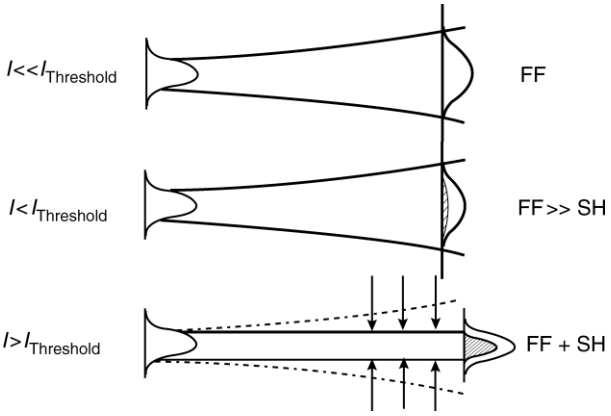


Figure 7 Parametric solitons via SHG. Intuitive description of soliton formation with an FF Gaussian input. At low powers (top) and for a small amount of generated SH (center) linear diffraction dominates. Above threshold parametric gain efficiently counteracts diffraction, originating a spatial soliton via quadratic cascading.

The self-trapped solutions are found as in eqn [23], with

$$A_1 = a_1(x, y)\exp(\mathrm{i}\beta z)$$
$$A_2 = a_2(x, y)\exp(\mathrm{i}\Delta k z)\exp(2\mathrm{i}\beta z), \quad \text{and} \qquad [27]$$
$$a_{1,2}(x, y) = \sqrt{I_{1,2}(x, y)}$$

These show that, even in the presence of diffraction, SHG can give rise to pure phase effect if the profiles satisfy:

$$\frac{1}{2k_0 n_1}\nabla_\perp^2 a_1 + \frac{\omega d_{\text{eff}}}{c} a_1 a_2 = \beta a_1$$
$$\frac{1}{8k_0 n_2}\nabla_\perp^2 a_2 + \frac{\omega d_{\text{eff}}}{2c} a_1^2 = \beta + \frac{\Delta k}{2} a_2 \qquad [28]$$

which are the space-dependent counterpart of eqn [25]. Again, β is a nonlinear contribution to the FF wavevector, and determines the corresponding correction $\beta_2 = 2\beta + \Delta k$ at SH. β is univocally determined by the power of the two harmonics in the solutions of eqn [28].

In the particular case:

$$\Delta k = \beta\frac{n_1 - 4n_2}{2n_2} \qquad [29]$$

1D parametric solitons exhibit profiles

$$a_1 = \sqrt{\frac{n_1}{n_2}}\frac{3c}{4\omega d_{\text{eff}}}\beta\,\mathrm{sech}\left(\sqrt{\frac{\beta k_0 n_1}{2}}x\right)^2$$
$$a_2 = \frac{3c}{4\omega d_{\text{eff}}}\beta\,\mathrm{sech}\left(\sqrt{\frac{\beta k_0 n_1}{2}}x\right)^2 \qquad [30]$$

and the powers per unit of transverse wavefront along y at the two components are:

$$P_1 = \int_{-\infty}^{\infty} a_1^2(x)\mathrm{d}x = \frac{n_1}{n_2}\frac{3\sqrt{2}c^2}{4\omega^2 d_{\mathrm{eff}}^2\sqrt{k_0 n_1}}\beta^{3/2}$$
$$P_2 = \int_{-\infty}^{\infty} a_2^2(x)\mathrm{d}x = \frac{3\sqrt{2}c^2}{4\omega^2 d_{\mathrm{eff}}^2\sqrt{k_0 n_1}}\beta^{3/2} \quad [31]$$

with the total soliton power $P_1 + P_2$, directly related to β.

Equation [29] implies that, for $n_1 > 4n_2$ and therefore $\Delta k > 0$, 1D parametric solitons exhibit the sech^2 profile. Otherwise, no analytical solutions are known even in 1D, and numerical techniques need to be employed. The same applies to 2D soliton profiles in all instances.

Notably, eqns [28] are essentially identical to eqns [21] for solitons in NLC. Therefore, Figure 4 provides the features of 2D parametric solitons as well, with 'a' replaced by 'a_1', and 'Ψ' by 'a_2'. For parametric solitons $\alpha = 4n_2(1 + \Delta k/2\beta)/n_1$, showing that at phase-matching ($n_2 = n_1$, $\Delta k = 0$) $\alpha = 4$, while large α's correspond to large Δk or small β's, e.g., low power and very wide solitons. The sech^2 profile for 1D solitons corresponds to $\alpha = 1$. 1D and 2D soliton solutions exist for any $\alpha > 0$.

The model for SHG solitons is more complicated than the previously discussed Kerr or saturable cases. The stability question can be formulated as the constrained-minimum problem of a specific functional, the Hamiltonian H. In general, for a given soliton power (the sum of the powers at the two harmonics) two branches exist for H. They correspond to two solitary solutions, differing in the ratio between the two powers. The solution with $P_2 > P_1$ corresponds to higher H, and is unstable via parametric amplification. When a harmonic field propagates in the presence of energy at a subharmonic (e.g., a small amount of FF), it tends to break up and transfer energy to it. Looking at Figure 4, this corresponds to small α's, only attainable when $\Delta k < 0$. As a result, quadratic solitons with a relevant SH content are unstable and thus, more difficult to observe.

The generation of a parametric soliton with a Gaussian input at FF, has a threshold (see Figure 7). This is because the input has to generate sufficient SH in order to produce and 'feel' the nonlinear phase-shift. A numerical simulation of eqn [26] shows the formation of a parametric spatial soliton, as displayed in Figure 8. As shown in (a), if only a small second-harmonic is generated by an FF excitation, diffraction dominates. Above threshold, as in (b), a stable self-trapped two-color beam is generated.

Two significant experimental results are reproduced in Figures 9 and 10. The photograph in Figure 10 shows a soliton propagating in periodically poled KTP (PPKTP) over five Rayleigh ranges, obtained by imaging the out-of-plane scattered light. Parametric solitons are typically observed with the FF in the near-infrared and the SH in the visible.

Figure 10 shows the output spots at FF and SH of a beam propagating in PPKTP for various FF excitations, until a soliton is generated (right end).

We close this section by observing that more general models must be considered to describe the whole physics of solitons in media with a quadratic

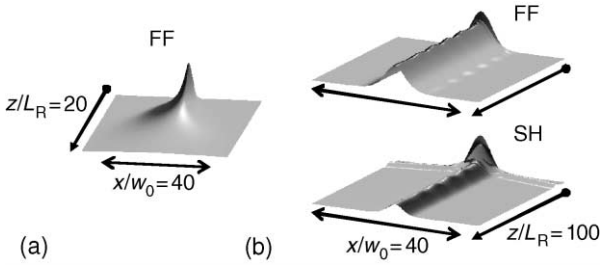


Figure 8 Parametric solitons via SHG. 1D numerical simulations using the beam propagation method. (a) An FF Gaussian excitation diffracts at low powers; (b) at high FF powers, after a transient both harmonics propagate undiffracted as a parametric spatial soliton. (Adapted with permission from Assanto G and Stegeman GI (2002) Simple physics of quadratic spatial solitons. *Optics Express* 10: 388–396).

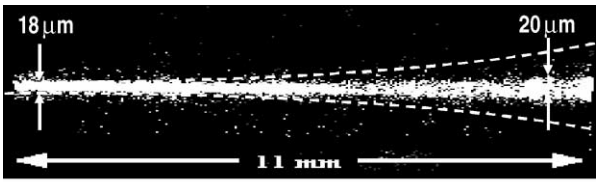


Figure 9 2D parametric solitons via SHG. Top view photograph of a soliton in PPKTP (20 ps pulses at $\lambda \approx 983$ nm and pulse energy of about 0.3 μJ). The dashed lines illustrate the diffractive behavior in the absence of nonlinear effects. (Adapted with permission from Malendevich R, Jankovic L, Polyakov S, Fuerst R, Stegeman G, Bosshard C and Gunter P (2002) Two-dimensional type I quadratic spatial solitons in KNbO3 near non-critical phase-matching. *Optics Letters* 27: 631–673.)

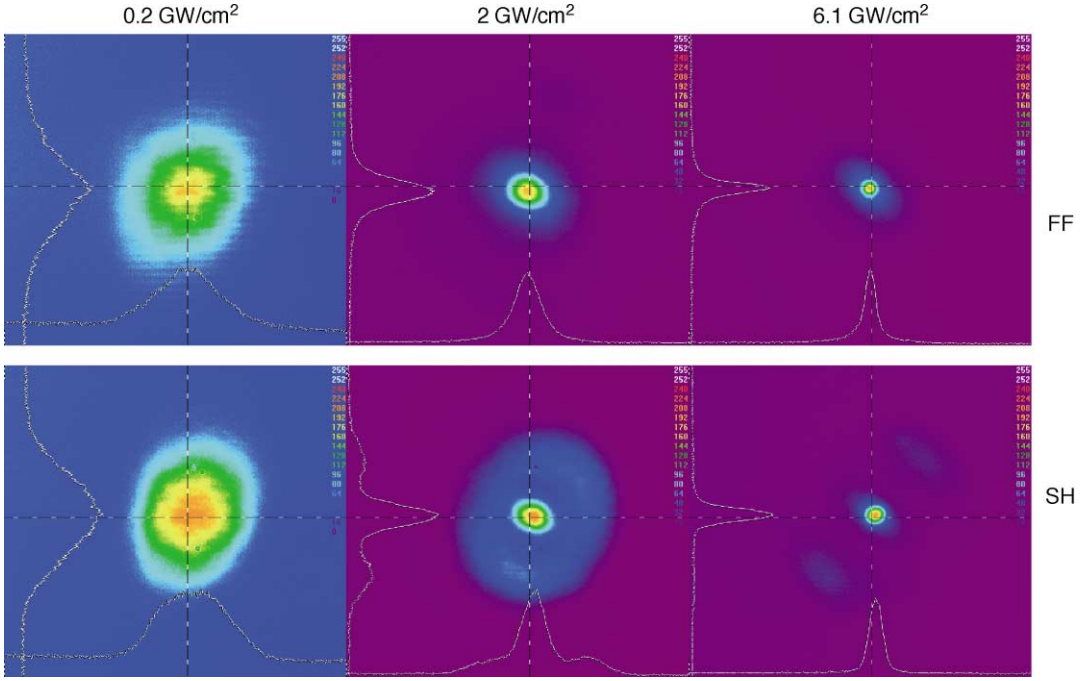


Figure 10 2D parametric solitons via SHG. Beam spots at FF (top row) and SH (bottom) for various input intensities in a PPKTP crystal. (Adapted with permission from Stegeman GI, Jankovic L, Kim H, *et al.* (2003) Generation of, and interactions between quadratic spatial solitons in non-critically-phase-matched crystals. *Journal of Nonlinear Optical Physics and Materials*. 12: 1–20).

response. We mention, for example, the subtleties associated to anisotropy (walk-off), nondegenerate three wave interactions, additional Kerr and higher-order effects.

Conclusions

Bright spatial solitons are widespread, and have been reported in several materials using various wavelengths and geometries. The brief overview presented above was intended to introduce the variety of basic mechanisms underlying their generation and propagation, but could not even touch upon the wealth of additional properties and potential applications of individual and multiple spatial solitons, their interactions and collisional properties, their stability and their control with other linear and nonlinear light waves, as well as with additional nonoptical effects (temperature, voltage, acoustic, and magnetic fields, etc.). Fascinating phenomena involve bright spatial solitons and are the subject of current research. Among them we like to mention all-optical logic and signal readdressing, beam shaping and filtering, limiting, sub-wavelength spatial compression, and digital imaging. To learn more on the subject, we encourage the interested readers to look up the vast literature covering both fundamental and applied aspects of optical spatial solitons.

List of Units and Nomenclature

Angular frequency [s^{-1}]	ω
Beam waist [m]	w
Boltzmann constant [$J\,K^{-1}$]	k_B
Current density vector [$A\,m^{-2}$]	$\mathbf{J}$
Density of ionized donors [m^{-3}]	N_d^i
Dielectric constant of vacuum [$F\,m^{-1}$]	ε_0
Director angle perturbation [rad]	ψ
Director orientation angle [rad]	θ_0
Electric field amplitude, normalized [$W^{1/2}\,m^{-1}$]	a, A
Electric field amplitude, optical frequency [Vm^{-1}]	E_{opt}
Electric field vector, low frequency [$V\,m^{-1}$]	$\mathbf{E}$
Electron charge [C]	q
Electron density [m^{-3}]	ρ
Electron mobility [$m^2\,V^{-1}\,s^{-1}$]	μ
Electro-optic second-order tensor [$m\,V^{-1}$]	r, r_{ijk}

Electro-optic third-order tensor $[m^4\ C^{-2}]$	g_{eff}, g_{ijkl}
Element of the photovoltaic tensor $[C\ m^3\ J^{-1}]$	β_{ph}
Kerr nonlinear coefficient $[m^2 W^{-1}]$	n_2
Low frequency permittivity $[F\ m^{-1}]$	ε_S
Medium elastic constant [N]	K
Medium forcing constant $[N\ m^{-2}]$	G
Non local decay distance [m]	R_c
Nonlinear length [m]	L_N
Nonlinear wavevector correction $[m^{-1}]$	β, β_2
Optic axis versor	$\mathbf{c}_{\text{opt}}$
Optical background intensity $[Wm^{-2}]$	I_b
Optical dark intensity $[Wm^{-2}]$	I_{dark}
Optical index anisotropy	n_a
Optical intensity $[Wm^{-2}]$	I
Optical saturation intensity $[Wm^{-2}]$	I_S
Photo-ionization cross-section $[m^2\ J^{-1}]$	s
Photovoltaic index perturbation constant $[V\ m^{-1}]$	E_{ph}
Power [W]	P
Power per wavefront unit $[W\ m^{-1}]$	P_1, P_2
Propagation coordinate [m]	z
Rayleigh length [m]	L_R
Recombination rate $[m^3\ s^{-1}]$	γ_r
Refractive index	n
Refractive index perturbation	Δn
Relative dielectric constant	ε_r
Sample width [m]	L
Second order effective susceptibility $[m\ V^{-1}]$	$d^{(2)}_{\text{eff}}$
Second order nonlinear coefficient $[m\ W^{-1/2}]$	d_{eff}
Self-focusing power [W]	P_c
Soliton power [W]	P_S
Speed of light in vacuum $[ms^{-1}]$	c
Static dielectric constant $[F\ m^{-1}]$	ε_s
Temperature [K]	T
Total density of negatively charged sites $[m^{-3}]$	N_A
Total donor density $[m^{-3}]$	N_d
Transverse coordinates [m]	x, y, r
Transverse wave-vector $[m^{-1}]$	k_x
Vacuum impedance $[\Omega]$	Z_0
Wavelength [m]	$\lambda = 2\pi n/k = 2\pi/k_0$
Wavevector in medium $[m^{-1}]$	$k = n\omega/c$
Wavevector in vacuum $[m^{-1}]$	$k_0 = \omega/c$

See also

Babinet's Principle. **Fiber and Guided Wave Optics:** Nonlinear Optics. **Nonlinear Optics, Applications:** Self-focusing and Related Effects (Solitons and Multiphoton Absorption). **Nonlinear Optics, Basics:** Cascading; Photorefraction.

Further Reading

Assanto G and Peccianti M (2003) Spatial solitons in nematic liquid crystals. *IEEE Journal of Quantum Electronics* 39: 13–21.

Assanto G and Stegeman GI (2002) Simple physics of quadratic spatial solitons. *Optics Express* 10: 388–396.

Assanto G, Peccianti M and Conti C (2003) Nematicons: optical spatial solitons in nematic liquid crystals. *Optics and Photonics News* 14: 44–48.

Buriak AV, Di Trapani P, Skryabin DV and Trillo S (2002) Optical solitons due to quadratic nonlinearities: from basic physics to futuristic applications. *Physics Reports* 370: 63–235.

Chiao RY, Garmire E and Townes CH (1964) Self-trapping of optical beams. *Physics Review Letters* 13: 479–482.

Conti C, Peccianti M and Assanto G (2003) Route to nonlocality and observation of accessible solitons. *Physics Review Letters* 91: 073901.

Conti C, Peccianti M and Assanto G (2004) Observation of optical spatial solitons in a highly nonlocal medium. *Physics Review Letters* 92: 113902.

Crosignani B and Salamo G (2002) Photorefractive solitons. *Optics and Photonics News* 13: 38–41.

Kivshar Y and Agrawal G (2003) *Optical Solitons: From Fibers to Photonic Crystals*. New York: Academic Press.

Stegeman GI, Christodoulides DN and Segev M (2000) Optical spatial solitons: historical perspectives. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1419–1426.

Torner L and Sukhorukov AP (2002) Quadratic solitons. *Optics and Photonics News* 13: 42–47.

Various Authors (1998) In: Kajzar F and Reinisch R (eds) *Beam Shaping and Control with Nonlinear Optics*. New York: Plenum Press.

Various Authors (2001) In: Boardman AD and Sukhorukov AP (eds). *Soliton-driven Photonics*. London: Kluwer Academic.

Various Authors (2002) In: Trillo S and Torruellas WE (eds) *Spatial Solitons*. Berlin: Springer-Verlag.

Optical Fiber Solitons, Physical Origin and Properties

G Millot and P Tchofo-Dinda, Université de Bourgogne, Dijon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In the last few decades there has been considerable effort made in understanding energy-localization phenomena in many areas of physics. Optical fibers constitute excellent media where such phenomena are routinely observed by means of relatively simple experimental setups. A well-known example of those energy-localization phenomena is the optical soliton, which corresponds to a light pulse that can propagate in optical fibers over relatively long distance without experiencing significant degradations. From a fundamental point of view, the high stability of the optical soliton results from a delicate balance between the fiber chromatic dispersion and the nonlinear refractive (or Kerr) index. The idea of using optical solitons as information bits in high-speed communication systems was first proposed in 1973, and then demonstrated experimentally in 1980. After these discoveries, interest in optical soliton transmission started to increase. In an ideal fiber, optical solitons can be modeled approximately by the nonlinear Schrödinger equation, whose solutions are well known. In this article we present the main types of soliton behavior and the fundamental physical processes that are involved in soliton stability.

The Nonlinear Schrödinger Equation

In optical fibers, there are many nonlinear phenomena that are difficult to understand by direct analysis of the nonlinear polarization (e.g., the soliton effect) but that can be readily understood by use of the nonlinear Schrödinger equation (NLSE). Let us consider an optical pulse that propagates in the z-direction, in a dielectric Kerr medium such as an optical fiber, with frequency ω_0, wavenumber k_0, and an envelope $\mathcal{A}$:

$$E = \mathcal{A}\exp[\mathrm{i}(\mathbf{k}_0\cdot\mathbf{r} - \omega_0 t)] \qquad [1]$$

We assume that the wave envelope $\mathcal{A}$ varies slowly in time and space compared with the fast variations of the carrier wave ω_0. In this situation, the Fourier spectrum of the wave is restricted to frequencies and wavenumbers close to $\omega = \omega_0$ and $k = k_0$, respectively. Then, the wavenumber k can be expanded in a Taylor series around k_0, as follows:

$$k(\omega) = k_0 + \left(\frac{\partial k}{\partial \omega}\right)_{\omega_0}(\omega-\omega_0) + \frac{1}{2}\left(\frac{\partial^2 k}{\partial \omega^2}\right)_{\omega_0}(\omega-\omega_0)^2 + \left(\frac{\partial k}{\partial |\mathcal{A}|^2}\right)_{\mathcal{A}_0=0}|\mathcal{A}|^2 + \cdots \qquad [2]$$

The nonlinear term $Q = \left(\partial k/\partial|\mathcal{A}|^2\right)_{\mathcal{A}_0=0}$ arises from the optical Kerr effect, as $k = n\omega/c = k(\omega, |\mathcal{A}|^2)$.

Setting $K = k - k_0$, $\Omega = \omega - \omega_0$, $\beta_1 = (\partial k/\partial\omega)_{\omega_0}$ and $\beta_2 = \left(\partial^2 k/\partial\omega^2\right)_{\omega_0}$, one can rewrite eqn [2] as

$$K = \beta_1\Omega + \frac{1}{2}\beta_2\Omega^2 + Q|\mathcal{A}|^2 \qquad [3]$$

where β_1 is the inverse of the group velocity and β_2 is the group velocity dispersion (GVD) coefficient. Equation [3] represents the nonlinear dispersion relation for the envelope. Using a Fourier transform method in which K is replaced by a spatial operator $-\mathrm{i}\,\partial/\partial z$, and Ω by a temporal operator $\mathrm{i}\,\partial/\partial T$, one can apply eqn [3] on the envelope $\mathcal{A}$ to obtain the following equation:

$$\mathrm{i}\frac{\partial\mathcal{A}}{\partial z} + \mathrm{i}\beta_1\frac{\partial\mathcal{A}}{\partial T} - \frac{1}{2}\beta_2\frac{\partial^2\mathcal{A}}{\partial T^2} + Q|\mathcal{A}|^2\mathcal{A} = 0 \qquad [4]$$

In practice it is more convenient to make use of the quantity $A = (\alpha A_{\mathrm{eff}})^{1/2}\mathcal{A}$, for which $|A|^2$ represents the wave power. The parameter A_{eff} is the effective core area, which takes into account the variation of the light intensity across the transverse section of the fiber. The value of A_{eff} strongly depends on the mode distribution. Using a system of coordinates that moves at the group velocity v_{g}, i.e., replacing T by $t = T - z/v_{\mathrm{g}}$, eqn [4] reduces to the basic NLSE:

$$\mathrm{i}\frac{\partial A}{\partial z} - \frac{1}{2}\beta_2\frac{\partial^2 A}{\partial t^2} + \gamma|A|^2 A = 0 \qquad [5]$$

where γ is the nonlinear coefficient defined by $\gamma = Q/(\alpha A_{\mathrm{eff}}) = \omega_0 n_2/(cA_{\mathrm{eff}})$ and expressed in $\mathrm{W}^{-1}\mathrm{m}^{-1}$. This equation is formally analogous to the Schrödinger equation in quantum mechanics but with a nonlinear potential. Furthermore, in this basic NLSE, all higher-order terms that describe effects such as the fiber losses, third-order dispersion, or stimulated Raman scattering, are neglected. The NLSE for pulse propagation in optical fibers was

first derived in 1973 by Hasegawa and Tappert. It has also found applications in other areas of physics, such as in plasma physics, fluid dynamics and electrical transmission lines. From a fundamental point of view, the basic NLSE involves the combined effects of the group velocity dispersion (GVD) and self-phase modulation (SPM), which are described by the second and third terms in eqn [5], respectively. Each of these two effects is analyzed below.

Group Velocity Dispersion

The individual effect of the GVD on pulse propagation can be obtained by setting the nonlinear coefficient to zero ($\gamma = 0$). In this case, eqn [5] reduces to

$$\mathrm{i}\frac{\partial A}{\partial z} = \frac{1}{2}\beta_2 \frac{\partial^2 A}{\partial t^2} \qquad [6]$$

In the frequency domain, eqn [6] becomes

$$\mathrm{i}\frac{\partial \tilde{A}}{\partial z} = -\frac{1}{2}\beta_2 \omega^2 \tilde{A} \qquad [7]$$

where $\tilde{A}$ represents the Fourier transform of A. The solution of eqn [7] is given by

$$\tilde{A}(z,\omega) = \tilde{A}(0,\omega)\exp(\mathrm{i}\beta_2\omega^2 z/2) \qquad [8]$$

As eqn [8] shows, the GVD changes the phase of each spectral component of the pulse by an amount that depends on the propagated distance and the frequency of the spectral component. But this effect does not modify the power spectrum: $(|\tilde{A}(z,\omega)|^2 = |\tilde{A}(0,\omega)|^2)$. This phase change can modify the shape of the pulse envelope. For example, a Gaussian pulse with a width of δ_0 at $z = 0$, $A(z=0,t) = \exp\{(-t^2/(2\delta_0^2)\}$, will transform into

$$A(z,t) = [\delta_0^2/(\delta_0^2 - \mathrm{i}\beta_2 z)]^{1/2}\exp[-t^2/2(\delta_0^2 - \mathrm{i}\beta_2 z)] \qquad [9]$$

at a distance z. In other words, the pulse width increases from δ_0 to $\delta_0[1+(z/L_\mathrm{D})^2]^{1/2}$. The parameter $L_\mathrm{D} = \delta_0^2/|\beta_2|$ defines a dispersion length, that is, a length scale over which the dispersive effects become important. At $z = L_\mathrm{D}$ the Gaussian pulse broadens by a factor of $\sqrt{2}$. Thus, the GVD causes the spectral components of the pulse to move at different velocities, which broadens the pulse, as shown by Figure 1.

Another important effect of the GVD comes from the time dependence of the pulse phase, which one can easily obtain by rewriting eqn [9] as $A(z,t) = |A(z,t)|\exp[\mathrm{i}\phi(z,t)]$. It turns out that the GVD induces an instantaneous variation of the frequency across the pulse, called a frequency chirp. For the initially unchirped Gaussian pulse, the GVD induces a chirp given by

$$\delta\omega = -\frac{\partial\phi}{\partial t} = \frac{t\beta_2(z/L_\mathrm{D})}{|\beta_2|\delta_0^2[1+(z/L_\mathrm{D})^2]} \qquad [10]$$

As eqn [10] illustrates, the GVD-induced frequency chirp depends on the sign of β_2, and varies linearly across the pulse.

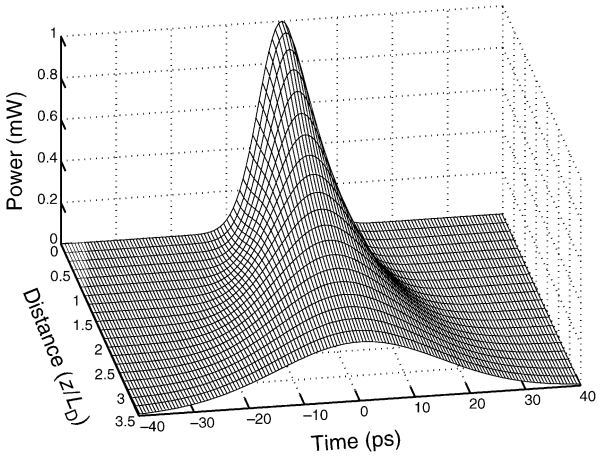


Figure 1 Temporal evolution of the power of a Gaussian pulse, as a function of the propagation distance. The simulation parameters are the following: input peak power $P_0 = 1$ mW, $\beta_2 = -20$ ps^2 km^{-1}, $\gamma = 3\times10^{-3}$ m^{-1} W^{-1} and $\delta_0 = 10$ ps. The dispersion length is $L_\mathrm{D} = 1.8$ km.

Self-Phase Modulation

Similarly to dispersion effects, nonlinear effects in optical fibers require a minimum propagation distance to develop significantly. The distance over which nonlinear effects become important is defined as the nonlinear length L_NL, given by $L_\mathrm{NL} = 1/(\gamma P_0)$, where P_0 is the peak power.

Let us consider the propagation of a light pulse in a fiber of length L, such that $L_\mathrm{NL} < L \ll L_\mathrm{D}$. In this particular case, the effects of dispersion can be neglected, and then the NLS equation becomes

$$\mathrm{i}\frac{\partial A}{\partial z} + \gamma|A|^2 A = 0 \qquad [11]$$

It results from this equation that $|A(z,t)|^2 = |A(z=0,t)|^2$. This means that the pulse shape remains unchanged during propagation. The solution of eqn [11] is given by

$$A(z,t) = A(0,t)\exp[\mathrm{i}\Phi_\mathrm{NL}(z,t)] \qquad [12]$$

where $\Phi_{NL}(z,t) = \gamma|A(0,t)|^2 z$. Thus, the Kerr effect causes the phase of the wave envelope to vary during the propagation. The nonlinear phase shift Φ_{NL} increases with the propagated distance z. The maximum phase shift Φ_{NL}^{max} occurs at the pulse center and is given by $\Phi_{NL}^{max} = \gamma|A(0,0)|^2 z = \gamma P_0 z$. One can see that the nonlinear length represents the propagation distance at which $\Phi_{NL}^{max} = 1$. Since the nonlinear phase shift Φ_{NL} depends on the initial pulse shape, it varies nonlinearly with time t. This implies that the instantaneous optical frequency differs across the pulse from its central value ω_0. The frequency chirp is given by

$$\delta\omega(t) = \omega(t) - \omega_0 = -\frac{\partial\Phi_{NL}}{\partial t} = -\frac{\partial|A(0,t)|^2}{\partial t}\gamma z \qquad [13]$$

The modification of the phase of the wave envelope by itself is called self-phase modulation. The frequency chirp increases with the propagated distance, thus implying that new frequencies are continuously generated as the pulse propagates through the fiber. This spectral broadening depends on the initial pulse shape. Figure 2 shows the variation of Φ_{NL} and the corresponding induced frequency chirp across the pulse at $z = L_{NL}$ for the case of an initial Gaussian pulse.

The SPM-induced chirp exhibits several interesting features. $\delta\omega$ is negative near the leading edge (red shift) and is positive near the trailing edge (blue shift). The chirp is linear (and increases) in the central region of the pulse. In fact, the chirp $\delta\omega$ at a given propagation distance z depends on the initial chirp. As Figure 3 shows, at a sufficiently long propagated distance, the SPM-induced spectrum exhibits several peaks due to interference effects across the pulse. Indeed as Figure 2 shows, the same frequency chirp occurs at two distinct times associated with two distinct nonlinear phase shifts.

The spectrum width can be estimated by calculating the maximum value $\delta\omega_{max}$ of the frequency chirp. For an initial unchirped Gaussian pulse one obtains

$$\delta\omega_{max} \approx \frac{0.86\gamma P_0 z}{\delta_0} \qquad [14]$$

Thus, the spectral width increases linearly with the propagated distance z and the input peak power P_0, and is inversely proportional to the initial pulse width δ_0. The SPM can be detrimental for optical communication systems. The SPM broadens the spectrum of the signal and makes it then more susceptible to GVD effects. Note that in wavelength division multiplexing (WDM) systems, the phase modulation of a given channel can come not only from the power in the considered channel itself (SPM) but also from the power in a different channel of the WDM system (CPM, i.e., cross-phase modulation). By appropriate dispersion compensation, the impact of SPM and CPM can be significantly reduced. On the other hand, the SPM, combined with dispersive effects, leads to very useful effects such as pulse compression, or the generation of optical solitons (described below).

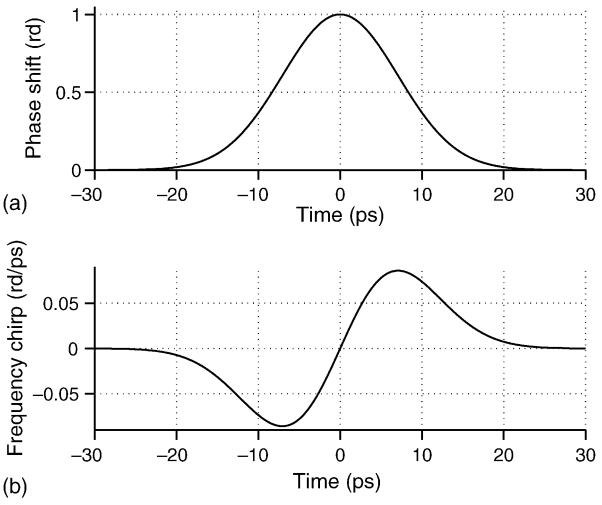


Figure 2 Temporal variation of the nonlinear phase shift (a) and the corresponding frequency chirp induced by SPM (b), for a Gaussian pulse.

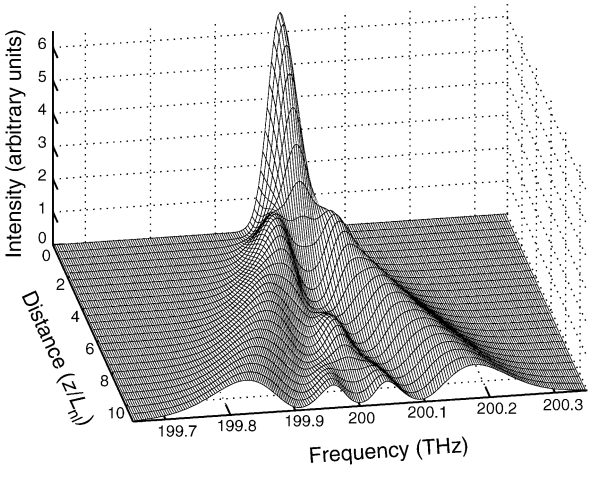


Figure 3 Evolution of the SPM-induced spectral broadening of an initially unchirped Gaussian pulse, as a function of the propagation distance. Here the normalized distance is z/L_{NL}.

Optical Solitons

Bright Solitons

Soliton propagation results fundamentally from a delicate balance between two phenomena: GVD and SPM. As we mentioned above, the GVD causes the spectral components of the pump to move at different velocities, thus leading to a temporal broadening of the pulse as shown in Figure 1 in the case of a Gaussian input pulse. Moreover, eqn [10] shows that the effect of GVD on a propagating Gaussian pulse is associated with a linear frequency chirp $\delta\omega$ and a parabolic temporal phase, proportional to β_2. On the other hand, eqn [12] shows that the SPM induces a frequency chirp across the pulse proportional to the temporal derivative of the pulse intensity. The chirp induced by SPM is then linear only in the central region of the pulse as long as the initial pulse shape is nonparabolic. As a consequence a Gaussian input pulse cannot propagate in a distortion-free manner since the GVD for anomalous dispersion ($\beta_2 < 0$) is able to completely cancel the intensity-dependent phase shift from SPM only over the central region of the pulse. On the other hand we show below that, in the anomalous dispersion regime, the exact cancellation of the nonlinear and dispersive temporal phase components (and corresponding chirps) is obtained in the case of the propagation of a hyperbolic secant pulse, which corresponds to the fundamental bright optical fiber soliton.

We consider the propagation of an unchirped hyperbolic secant pulse at a wavelength of 1.55 μm in a standard telecommunications fiber with $\beta_2 = -21 \times 10^{-3}$ ps^2 m^{-1} and $\gamma = 1.2 \times 10^{-3}$ W^{-1} m^{-1}. The input pulse at $z = 0$ is given by $A(0,t) = \sqrt{P_0}$ sech (t/δ_0) with $\delta_0 = 10$ ps and $P_0 = 175$ mW. With this choice of parameters the dispersion length is equal to the nonlinear length, i.e. $L_D = L_{NL} = 4.76$ km. In fact, these conditions correspond to those which would support a propagating soliton. However, let us consider first how SPM and GVD modify the incident pulse when acting independently. Figure 4 shows numerically computed results from the solutions of eqns [12] and [8] to illustrate the effects of nonlinear and dispersive evolution over a range of propagation distances. The figure shows the pulse intensity as well as the time-dependent phase and the corresponding frequency chirp. The evolution is clearly very different in the presence of only nonlinear or dispersive effects. Nonlinear SPM alone induces no change in the temporal intensity of the pulse, but the magnitudes of the SPM-induced phase and chirp increase with propagation distance. In contrast, GVD induces changes in both the intensity and the phase of the pulse as it propagates, and the temporal broadening with propagation is associated with a significant change in the functional forms of the phase and frequency chirp. In particular, whilst for large propagation distances it can be seen that GVD induces a parabolic temporal phase and corresponding linear chirp, this is clearly not the case for small propagation distances. This result strongly differs from that obtained for a propagating Gaussian pulse for which the temporal phase is always parabolic. In the context of soliton propagation, the results in Figure 4b are of particular significance, since they clearly show that at $z = L_D/10$ where neither SPM nor GVD noticeably modify the pulse intensity profile, the SPM and GVD-induced phases and frequency chirps have qualitatively similar forms, yet opposite sign. The split-step Fourier method demonstrates that over a sufficiently small propagation distance (e.g., $\delta z = L_D/10$) the effects of nonlinearity and dispersion can be considered to act independently on a propagating pulse. This suggests that, over the propagation segment ($0 < z < \Delta z$), the phases are capable of cancelling each other across the entire pulse profile. It is obvious that, if the effects cancel over this distance, then the propagation in the following segment will begin from the same initial condition of an unchirped pulse which was the case at $z = 0$. The nonlinear and dispersive phase components will therefore continue to cancel over successive iterations for an arbitrary length of fiber, and we thus obtain a simple physical interpretation to explain the distortion-free propagation of an optical soliton. Thus, the combined effects of GVD and SPM can prevent the occurrence of a frequency chirp across the pulse. In this situation the pulse propagates with a constant temporal profile. Such light pulses are called optical solitons.

Mathematically, the basic NLSE, eqn [5], belongs to a remarkable class of integrable nonlinear equations, which can be solved exactly for an arbitrary initial condition by means of the inverse scattering method. The solutions depend strongly on the sign of dispersion, and the relative importance of dispersive and nonlinear effects, measured by a dimensionless parameter N:

$$N^2 = \frac{L_D}{L_{NL}} = \frac{\gamma P_0 \delta_0^2}{|\beta_2|} \qquad [15]$$

In the anomalous dispersion regime ($\beta_2 < 0$) the NLS equation admits envelope-soliton solutions. The soliton order is given by the parameter N. Only the fundamental soliton, defined by $N = 1$, can propagate in the fiber with a constant profile, as schematically represented in Figure 5.

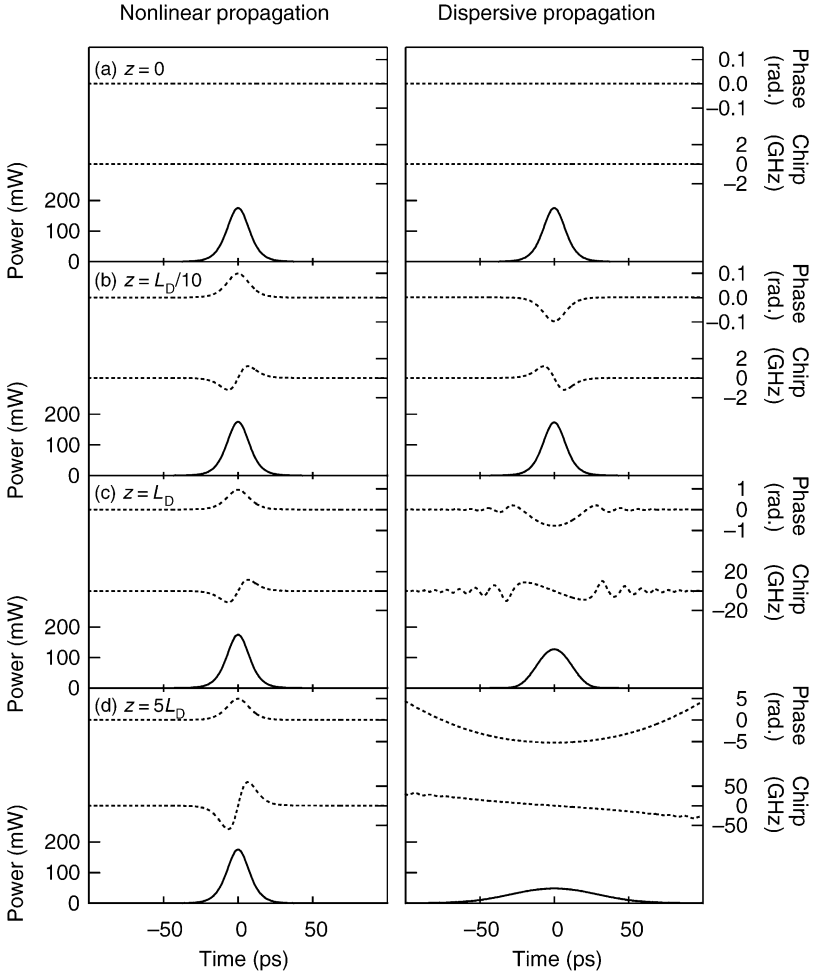


Figure 4 Numerical results showing nonlinear (left column) and dispersive (right column) propagation in an optical fiber, with parameters as described in the text, and for propagation distances as indicated. The solid lines show the pulse intensity (left axis) whilst the dashed lines show the calculated phase (right axis, top) and chirp (right axis, bottom). Constant phase offsets have been removed for clarity.

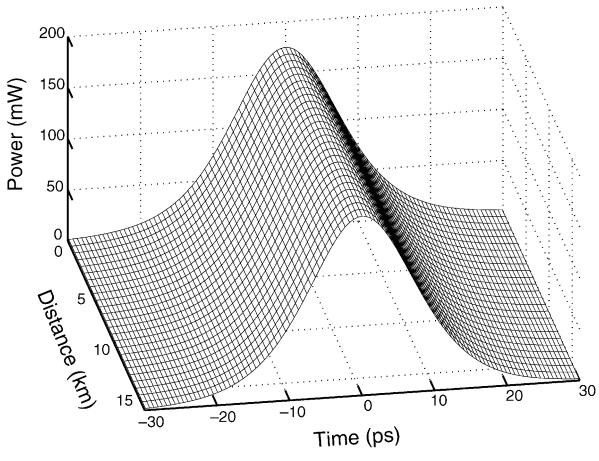


Figure 5 Temporal evolution of the power of a fundamental bright soliton, as a function of the propagation distance. The simulation parameters are the following: input peak power $P_0 = 67$ mW, $\beta_2 = -20$ ps^2 km^{-1}, $\gamma = 3 \times 10^{-3}$ m^{-1} W^{-1} and $\delta_0 = 10$ ps. The nonlinear length is $L_{NL} = 5$ km.

For the fundamental soliton, dispersion and nonlinearity exactly balance each other in a way such that neither the pulse shape nor the pulse spectrum change during the pulse propagation. The fundamental soliton ($N = 1$) is given by

$$A(z,t) = \sqrt{P_0}\,\mathrm{sech}(t/\delta_0)\,\exp(\mathrm{i}\gamma P_0 z/2) \qquad [16]$$

where the peak power required to generate this soliton is determined by $L_D = L_{NL}$, and given by

$$P_0 = \frac{|\beta_2|}{\gamma\delta_0^2} \qquad [17]$$

Note that the phase of this bright soliton remains constant across the entire pulse. For a standard single-mode fiber (SMF), at 1.55 μm wavelength, typical values of the dispersion and nonlinear parameters are $\beta_2 = -21$ ps^2 km^{-1} and $\gamma = 1.2 \times 10^{-3}$ W^{-1} m^{-1}, respectively.

So, the peak power required to propagate a fundamental soliton with pulse width $\delta_0 = 10$ ps, is $P_0 = 175$ mW. This soliton is represented in Figure 5. A further reduction of P_0 can be obtained by using dispersion-shifted fibers (DSF), for which the dispersion can be significantly reduced to values of the order of $1\ \text{ps}^2\,\text{km}^{-1}$. In this situation, the power required to create a soliton reduces to $P_0 = 8.3$ mW. Thus, fundamental solitons can be generated at power levels that are easily obtainable from conventional laser diodes. The first experimental observation of soliton propagation in optical fibers was reported in 1980 by Mollenauer, Stolen and Gordon.

Note that all the higher-order solitons ($N > 1$) execute a recurrent deformation during their propagation, as illustrated in Figure 6 in the case of the second-order soliton ($N = 2$).

Dark Solitons

In the normal dispersion regime ($\beta_2 > 0$) the chirps induced by GVD and SPM reinforce each other and thus make the hyperbolic secant pulse more susceptible to temporal broadening. Thus, in the normal dispersion regime bright solitons cannot propagate. However, in this case the NLS equation, eqn [5], admits a hole-soliton solution. In other words, the intensity profile of this soliton exhibits a dip in a uniform background, as represented in Figure 7.

Mathematically, this hole-soliton, commonly called a dark soliton, takes the following form:

$$A(z,t) = \pm\sqrt{P_0}\,\tanh(t/\delta_0)\,\exp(\mathrm{i}\gamma P_0 z) \qquad [18]$$

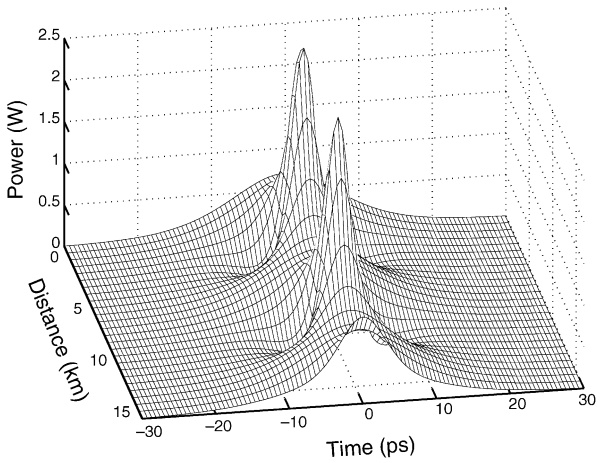


Figure 6 Temporal evolution of the power of a second-order soliton, as a function of the propagation distance. The simulation parameters are the following: input peak power $P_0 = 267$ mW. The fiber parameters are the same as in Figure 5. The nonlinear length is $L_{\text{NL}} = 1.25$ km.

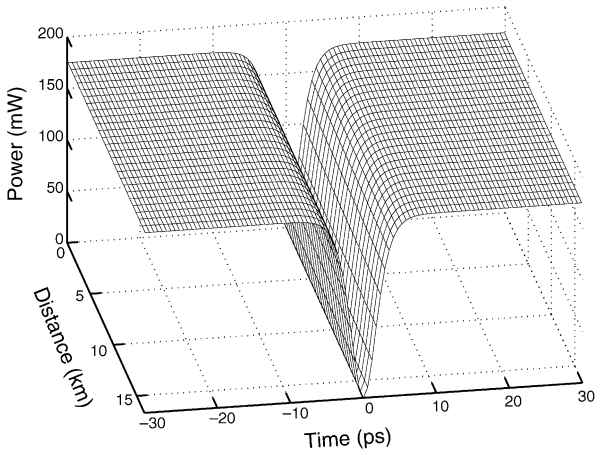


Figure 7 Temporal evolution of the power of a dark soliton, as a function of the propagation distance, for an input peak power $P_0 = 67$ mW.

As eqn [18] shows, the amplitude of the dark soliton is given by an odd function of time. Thus its phase exhibits an abrupt π jump at the dip center, thus implying that dark solitons are chirped. The phase between each jump is constant. Note that there exist other soliton solutions that are similar to the dark soliton, but for which the intensity at the dip center does not fall to zero and the phase variation is more gradual and smaller. These solutions are often called gray solitons.

A common feature of bright and dark solitons is their robustness. This property is very important for ensuring practical applications in optical communications. Moreover, optical solitons emerge unchanged from collision processes. However, dissipative perturbations such as material loss or Raman scattering can destroy such solitons.

Dispersion-Managed Solitons

As we mentioned above, the fundamental soliton is based on a delicate balance between dispersion and nonlinearity, through the relation $L_{\text{D}} = L_{\text{NL}}$. In addition, the existence of this soliton requires that only these two effects are present in the fiber. But in practice, additional linear and nonlinear effects such as fiber losses, Raman scattering, or third-order dispersion, are always present in the fiber. In particular, the fiber losses cause the soliton power to continually decrease during the propagation. The use of lumped amplifiers along the transmission line allows one to compensate for the power loss, but does not prevent the occurrence of chirp. As a consequence, the fundamental soliton does not survive over long distances (a few tens of kilometers), without in-line control. Moreover, this soliton is not stable

against phenomena such as CPM and FWM processes that occur unavoidably in WDM systems. The CPM between the pulses in a given channel and the pulses in other channels induce a frequency chirp (in the considered channel), which is converted into amplitude variations by the fiber dispersion. The FWM induces energy-exchange processes between different channels, which lead to pulse distortions.

A new transmission scheme, called *dispersion management*, was recently proposed to resolve the above-mentioned limitations. Basically, the dispersion-management technique utilizes a transmission line with a periodic dispersion map, such that each period is built up by two types of fibers, generally with different lengths and opposite group-velocity dispersion. Whereas the fundamental soliton requires total cancellation of the chirp at each propagation distance, the dispersion-managed (DM) soliton induces a small variation of the chirp, with a zero mean value over each period of dispersion management.

The DM soliton exhibits major differences with respect to the fundamental bright soliton. First, whereas the fundamental bright soliton possesses a smooth profile (as Figure 5 shows), the profile of the DM soliton exhibits an internal oscillatory structure, with several side-lobes on each side of the soliton's center. A typical representation of a DM-soliton profile is shown in Figure 8, at the free-chirp point of the dispersion map. Another outstanding difference is the following: whereas the fundamental soliton propagates with a constant profile, the DM soliton executes a periodic deformation, with a period that coincides with that of the dispersion map. The minimum value of the soliton's width occurs at the free-chirp point. From this point, the soliton's width can execute several oscillations within one dispersion map, with a peak value that can be quite large compared with the width at the free-chirp point. Figure 9 illustrates the propagation of a DM soliton within one dispersion map of a lossless DM line.

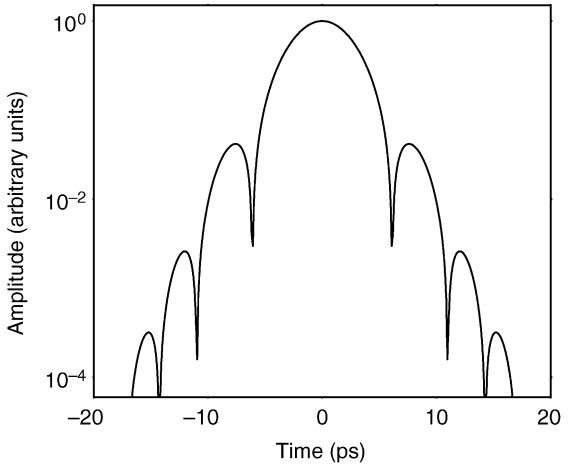


Figure 8 Plot showing the profile of a DM soliton at the free-chirp point of a typical DM line made up of 20.5 km of anomalous-dispersion fiber [$\beta_2 = -1.5 \times 10^{-3}$ ps^2 m^{-1}, $\gamma = 1.4 \times 10^{-3}$ W^{-1} m^{-1}], and 19.5 km of normal-dispersion fiber [$\beta_2 = 1.5 \times 10^{-3}$ ps^2 m^{-1}, $\gamma = 1.4 \times 10^{-3}$ W^{-1} m^{-1}].

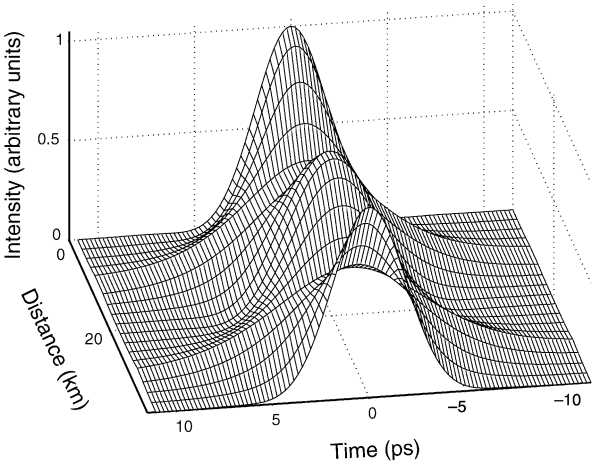


Figure 9 Plot showing the evolution of the DM soliton's profile as a function of the propagation coordinate z.

Recent experiments have demonstrated that in this new transmission scheme the pulse propagation becomes much more stable than fundamental solitons against linear and nonlinear effects such that SPM, FWM, CPM, or third-order dispersion. The DM transmission lines are currently subject to intense research with a view both to upgrade the capacity of existing terrestrial networks, and to design submarine fiber systems with ultrahigh capacity.

Modulational Instability

In optical fibers, the interaction between nonlinear and dispersive effects may lead to modulational instability (MI), a phenomenon in which a continuous or quasicontinuous wave undergoes a modulation of its amplitude or phase in the presence of noise or any other small perturbation. The perturbation can originate from quantum noise (spontaneous-MI) or from a frequency shifted signal wave (induced-MI). MI was observed for the first time by Tai, Hasegawa and Tomita in 1986 for a single pump wave propagating in a standard nonbirefringent fiber (scalar MI). Scalar MI occurs only when the group velocity dispersion (GVD) is negative (anomalous dispersion regime). Modulational instability in eqn [5] is examined through a linear stability analysis (LSA)

of its steady state solution

$$A = \sqrt{P_0}\,\exp[\mathrm{i}\gamma P_0 z] \qquad [19]$$

Equation [19] shows that the power remains constant through propagation along the fiber. The linear stability of the steady state solution is examined by looking into the system in the presence of small amplitude and phase perturbations *a;* that is, we consider

$$A = (\sqrt{P_0} + a)\,\exp[\mathrm{i}\gamma P_0 z] \qquad [20]$$

Then, the linearization of the NLSE with respect to a yields

$$\frac{\partial a}{\partial z} + \frac{\mathrm{i}}{2}\beta_2\frac{\partial^2 a}{\partial t^2} = \mathrm{i}\gamma P_0(a + a^*) \qquad [21]$$

Then one can assume for the perturbation a modulational ansatz with wavenumber K and frequency Ω, of the form

$$a(z,t) = u\cos(Kz - \Omega t) + \mathrm{i}v\sin(Kz - \Omega t) \qquad [22]$$

Substitution of eqn [22] into eqn [21] leads to the following dispersion relation for the perturbation

$$K^2 = \frac{\beta_2^2\Omega^2}{4}\left[\Omega^2 + \frac{4\gamma P_0}{\beta_2}\right] \qquad [23]$$

The MI phenomenon manifests itself by an exponential growth of the amplitude of the perturbation. This behavior occurs when the wavenumber K possesses a nonzero imaginary part. The dispersion relation [23] clearly shows that MI strongly depends on the sign of the dispersion. For normal dispersion ($\beta_2 > 0$), K is always real and the steady state is stable, whereas in the anomalous dispersion regime, MI appears for $\Omega < \Omega_c = \sqrt{4\gamma P_0/|\beta_2|}$. The importance of the phenomenon is measured by a power gain defined by

$$g(\Omega) = 2|\mathrm{Im}(K)| = |\beta_2|\Omega\sqrt{\Omega_c^2 - \Omega^2} \qquad [24]$$

Figure 10 shows the MI-gain spectra at three power levels with fiber parameters appropriate for standard silica fibers at $\lambda_0 = 1.319\ \mu\mathrm{m}$.

The peak value of the gain, $g_{\max} = 2\gamma P_0$, occurs at a frequency

$$\Omega_{\mathrm{opt}} = \sqrt{\frac{2\gamma P_0}{|\beta_2|}} \qquad [25]$$

called the optimal modulational frequency. In practice, when the perturbation corresponds to quantum noise, the spectral components that are amplified by the MI-process are close to Ω_{opt}. Here, as shown by Figure 11, the MI phenomenon leads to two sidebands that are situated symmetrically with respect to the pump, at frequency detunings of $\approx \pm\Omega_{\mathrm{opt}}$ (see Figure 10 for a pump peak power $P_0 = 3$ W).

In the time domain, induced-MI leads to the break-up of the quasi-cw pump wave into a train of ultrashort pulses. The repetition rate of the pulses is determined by the frequency detuning between the signal and the pump (i.e. the modulational frequency). The temporal shape of these ultrashort pulses depends not only on the powers

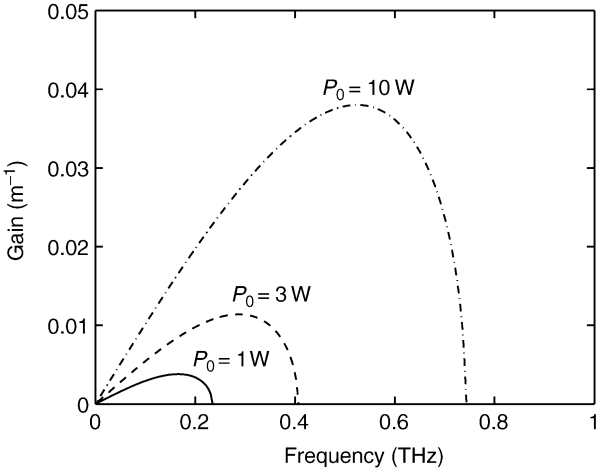


Figure 10 Spectral profiles of MI gain g for a pump peak power $P_0 = 1$ W (solid), 3 W (dashed) and 10 W (dot-dashed). The fiber parameters are $\beta_2 = -3.5\ \mathrm{ps}^2\ \mathrm{km}^{-1}$ and $\gamma = 1.9 \times 10^{-3}\ \mathrm{m}^{-1}\mathrm{W}^{-1}$.

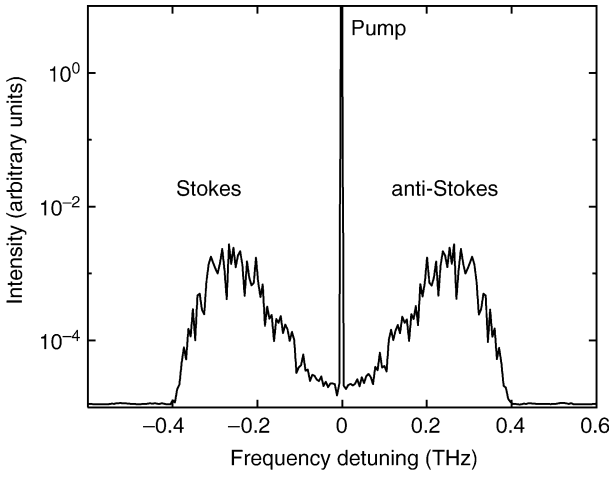


Figure 11 Spontaneous modulational instability spectrum (i.e., parametrically amplified spontaneous noise), as obtained for a peak pump power $P_0 = 3$ W and a fiber length $L = 1$ km. The fiber parameters are the same as in Figure 10.

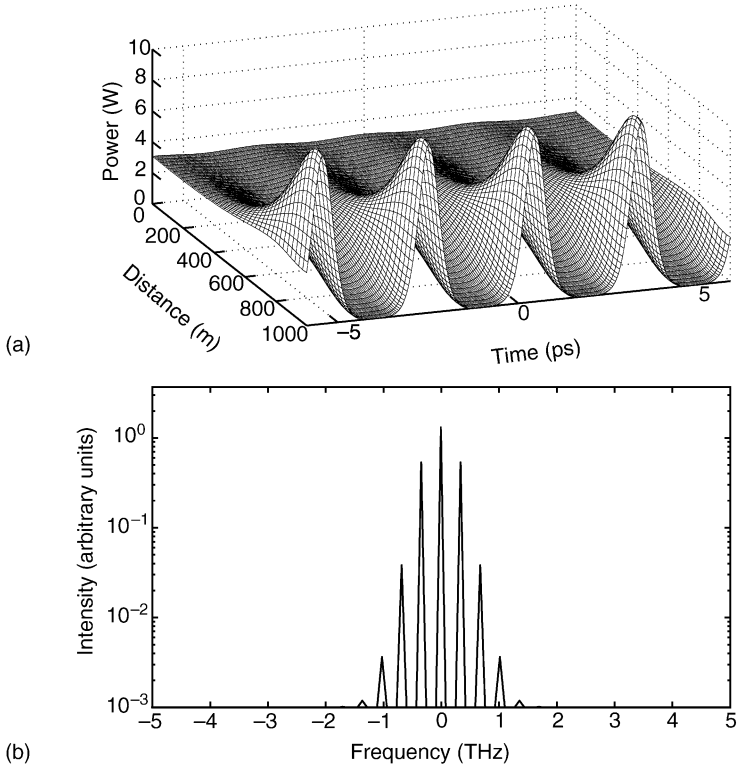


Figure 12 (a) Theoretical temporal evolution of the power for induced-MI, as a function of the propagation distance. (b) Output power spectrum in logarithmic intensity scale. The pump (signal) power is 3 W (0.5 mW), the pump-signal frequency detuning is 0.34 THz, and the fiber length is 1 km. The fiber parameters are the same as in **Figure 10**.

of the input waves but also on the modulational frequency. In particular, trains of bright and dark solitons with a terahertz repetition rate have been generated by MI in optical fibers. **Figure 12a** shows theoretically the evolution of a small initial modulation.

As **Figure 12a** shows, MI causes a small initial perturbation to evolve into a periodic pulse train, whose period is inversely proportional to the frequency spacing Δf between the pump and signal waves. Thus the technique of induced-MI can be exploited for generating high-repetition-rate ultrashort pulse trains. As shown in **Figure 12b**, such pulse train generation manifests itself in the frequency domain by the growth of a cascade of sidebands. The number of harmonics ($n\Delta f$) is determined by the temporal shape of the generated pulses and the initial conditions at the fiber input.

Cross-phase modulation (CPM) between two modes can be used to extend the instability domain to the normal dispersion regime (positive GVD). This is achieved by using orthogonally polarized pump waves. This CPM-induced MI is called vector modulational instability. The optical field evolutions are described by two coupled NLS equations. Several experiments involving two polarization modes, or two spatial modes, have been done in order to observe MI in the normal GVD regime.

Conclusion

The invention of soliton transmission, predicted in 1973 and experimentally demonstrated in 1980, is one of the major discoveries made in the area of optical telecommunication. Although from the period of invention of optical solitons up to very recent years the main line of research activities was based on the concept of the 'classical soliton', which represents an exact balance between the fiber group-velocity dispersion and its intensity-dependent refractive index, most recent experimental and theoretical developments focus on dispersion-managed solitons. It is now widely recognized that DM solitons in fiber-optic links are the most promising way both to upgrade the capacity of existing terrestrial telecommunication networks and to design submarine fiber systems. Most of the current research and development activities in this context aim to develop a non-empirical methodology of designing ultrahigh-speed DM fiber systems.

List of Units and Nomenclature

Dispersion length L_D [m]
Group velocity dispersion coefficient β_2 [$ps^2\,km^{-1} = 10^{-27}\,s^2\,m^{-1}$]
Nonlinear length L_{NL} [m]
Nonlinear parameter γ [$W^{-1}\,m^{-1}$]
Nonlinear refractive index n_2 [$m^2\,W^{-1}$]
Optical intensity I [$GW\,cm^{-2} = 10^9\,W\,cm^{-2}$]

CPM: cross-phase modulation
DM Soliton: dispersion-managed soliton
FWM: four-wave mixing
GVD: group velocity dispersion
LSA: linear stability analysis
MI: modulational instability
NLS (E): nonlinear Schrödinger (equation)
SPM: self-phase modulation
WDM: wavelength division multiplexing

See also

Fiber and Guided Wave Optics: Dispersion; Light Propagation; Nonlinear Effects (Basics). **Lasers:** Optical Fiber Lasers. **Scattering:** Scattering Theory. **Solitons:** Soliton Communication Systems.

Further Reading

Ablowitz MJ and Clarkson PA (1991) *Solitons, Evolution Equations and Inverse Scattering*. Cambridge, UK: Cambridge University Press.
Agrawal GP (1997) *Fiber-Optic Communication Systems*. New-York: Wiley-Interscience.
Agrawal GP (2001) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.
Agrawal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego, CA: Academic Press.
Hasegawa A (1989) *Solitons in Optical Fibers*. New-York: Springer Verlag.
Hasegawa A and Kodama Y (1995) *Solitons in Optical Communications*. Oxford, UK: Clarendon Press.
Remoissenet M (1999) *Waves Called Solitons, Concepts and Experiments*, 3rd edn. Berlin: Springer.
Taylor JR (ed.) (1992) *Optical Solitons – Theory and Experiment*. Cambridge, UK: Cambridge University Press.
Zakharov VE and Wabnitz S (eds) (1999) *Optical Solitons: Theoretical Challenges and Industrial Perspectives*. Les Ulis (France) and Berlin (Germany): EDP Sciences and Springer.

Soliton Communication Systems

M Karlsson and P Andrekson, Chalmers University of Technology, Gothenburg, Sweden

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In 1973 Hasegawa and Tappert proposed that optical soliton pulses in fibers could exist. A soliton is a propagating wave packet that is localized in the sense that it does not spread its energy during propagation, and with the additional property that it is so stable that it can collide with other solitons and emerge unaffected with respect to energy, shape, and momentum after the collision. Solitons are based on some kind of nonlinearity in the system, and for optical fibers the weak Kerr nonlinearity (which makes the refractive index increase in proportion to the optical intensity) can counteract the pulse broadening induced by group-velocity dispersion (GVD). The two effects can form a stable balance in the form of a soliton pulse, which then propagates without changing shape.

Due to the lack of short-pulse laser sources at wavelengths above 1.3 μm, and low-loss silica fibers, it took another seven years for optical soliton pulses to be experimentally verified. In an experiment by Mollenauer, Stolen, and Gordon in 1980, soliton pulse transmission over 700 m fiber was demonstrated. During the 1980s soliton research aimed towards the use of solitons as information carriers in optical communications, and in 1990 the first data transmission experiment using solitons (2.8 Gb/s over 23 km of fiber) were reported by Iwatsuki *et al*.

During the 1990s, soliton-based communication systems have matured, an important reason being the development of the erbium-doped fiber amplifier (EDFA), which made high power levels commercially available. The most recent developments have been towards the use of solitons in alternating dispersion maps (so-called dispersion management), and such systems have reached performance levels near commercialization.

It should be stressed that fiber soliton research has been almost exclusively devoted to single-mode fibers at a wavelength of 1550 nm. In principle, solitons can also be used to counteract mode dispersion in multimode fibres, but that generally requires unrealistically high peak powers, and is therefore regarded to be of less interest. Similarly, solitons with wavelengths around 1300 nm in single-mode fibers have been

considered, but found to be of less interest due to the lack of low-noise optical amplifiers in this spectral range. Another kind of soliton, the so-called dark soliton, can also be generated in single-mode fibers. However, dark solitons have only been used as information carriers in a very limited number of experiments, and will also be omitted from this discussion.

In the present review we will discuss both theoretical and experimental aspects of soliton transmission. We will distinguish between conventional solitons, which have constant dispersion during transmission, and dispersion-managed solitons (although solitary waves would be the proper mathematical name) for which the dispersion varies periodically during transmission.

Conventional Fiber Solitons

Group-velocity dispersion means that the group velocity of the light changes with the optical frequency, and it is usually defined as the second derivative of the propagation constant β with respect to the angular frequency ω, and denoted β_0'' [$\mathrm{ps^2\,km^{-1}}$]. Standard single-mode fibers have a $\beta_0'' \approx -20$ [$\mathrm{ps^2\,km^{-1}}$] around a wavelength of 1550 nm, whereas dispersion-shifted fibers (DSFs) have β_0'' close to zero in the 1550 nm region. The dispersion is defined with respect to sign, so that the GVD is said to be normal if the group velocity decreases with frequency (i.e., $\beta_0'' > 0$), and anomalous if the group velocity increases with frequency (i.e., $\beta_0'' < 0$). A pulse propagating in a dispersive fiber will broaden with propagation distance at a rate proportional to the product of β_0'' and the spectral width of the pulse. In a communication system such a broadening is devastating, since neighboring pulses will interfere and disperse power into adjacent bit slots. The problem is increased for high data rates (as this requires short pulses with broad spectra) and/or long transmission distances.

The nonlinearity in optical fibers manifests itself as a dependence of the refractive index with power, and is known as the optical Kerr effect. For pulse propagation the Kerr effect will give rise to self-phase modulation (SPM), which means that the pulse will get a phase modulation proportional to its own intensity. Mathematically speaking, the amplitude of an optical pulse $u(0, t)$ [$\mathrm{W^{1/2}}$] launched at $z = 0$ will evolve according to $u(z,t) = u(0,t)\exp(-\mathrm{i}\gamma z|u(0,t)|^2)$, where γ [$\mathrm{W^{-1}\,km^{-1}}$] is the nonlinear coefficient of the fiber, z [km] is the propagation distance along the fiber, t [ps] is the time coordinate, and $|u(z, t)|^2$ [W] is the power level of the light. The value of γ varies between different fiber types due to the dependence on the core area of the fiber, but for standard single-mode fibers, in the 1550 nm region, it is approximately 2.2 $\mathrm{km^{-1}\,W^{-1}}$.

Taking both self-phase modulation and dispersion into account, one can derive a propagation equation for light, known as the nonlinear Schrödinger (NLS) equation. Of particular importance for the solution to the NLS equation is solitons, which is a solution that propagates without dispersive broadening. The soliton solution arises as a result of a cancellation of the dispersion and the self-phase modulation effects, and it can only be formed when the dispersion is anomalous. For normal dispersion the nonlinearity and the dispersion will instead act together to increase the spectral broadening of the pulse. The soliton solution of duration t_0 and dispersion β_0'' reads

$$u_{\mathrm{sol}}(z,t) = \sqrt{\frac{1}{\gamma L_{\mathrm{D}}}}\,\mathrm{sech}\left(\frac{t}{t_0}\right)\exp\left(-\mathrm{i}\frac{z}{2L_{\mathrm{D}}}\right) \qquad [1]$$

where the dispersive length is $L_{\mathrm{D}} = t_0^2/|\beta_0''|$ [km]. The soliton peak power P_{s} [W] and energy E_{s} [J] can be expressed as $P_{\mathrm{s}} = |u_{\mathrm{sol}}(z,0)|^2 = |\beta_0''|/\gamma t_0^2$ and $E_{\mathrm{s}} = 2P_{\mathrm{s}}t_0 = 2|\beta_0''|/\gamma t_0$.

The fact that solitons are not broadened by dispersion will make them ideally suited as information carriers in a communication system. However, the solitons are affected by various perturbations that will affect the balance by dispersion and nonlinearity. Unless proper care is taken in the design of the system, those perturbations can destroy the performance of the communication system. Next, we will review the most important perturbations and design issues of soliton-based communication systems.

Solitons in the Presence of Third-Order Dispersion

The third-order dispersion (TOD), or the dispersion slope, is modeled with the derivative of the GVD, β_0''' [$\mathrm{ps^3\,km^{-1}}$], which is typically 0.1 $\mathrm{ps^3\ km^{-1}}$ in standard and dispersion-shifted fibers. For solitons in dispersion-shifted fibers the effect of the dispersion slope is increasingly important the closer the solitons are to the zero dispersion wavelength.

When a soliton is placed close to the zero-dispersion wavelength, it will emit radiation in the normal dispersion regime of the spectrum, lose energy, and recoil further into the anomalous dispersion region so that the radiation stops and a stable soliton is formed. As a design criterion $\beta_0'''/(|\beta_0''|t_0) < 0.24$ have been suggested for this phenomenon to provide negligible energy loss.

Solitons in the Presence of Amplification and Loss

The most important property that has been neglected in the derivation of the NLS equation for optical

pulses is the effect of loss. The loss will decrease the power and hence the nonlinearity so that the balance sustaining the soliton is no longer maintained. As a result of the loss (which is of the order of 0.2 dBkm^{-1} in standard transmission fibers) the soliton pulse will be broadened. Partly, this can be circumvented by requiring the launched peak power of the soliton after each amplifier to be such that the path-average power between the amplifiers equals the soliton power, P_s. The peak power P_{peak} falls off as $\exp(-\alpha z)$, where $\alpha = 0.042$ km^{-1} is the fiber loss coefficient, and the average peak power over one amplifier span is then $P_{peak}\ (1 - \exp(-\alpha L_a))/\alpha L_a = P_{peak}(G - 1)/G\ \ln(G) = P_s$, where $G = \exp(\alpha L_a)$ is the gain of each amplifier and L_a [km] is the amplifier spacing.

Nevertheless the periodic fluctuation of power and width over each amplifier length may cause the solitons to become unstable and lose power to dispersive wave radiation. To minimize this loss a useful design criterion is that the amplifier length L_a should be smaller than the dispersive length, L_D.

Sources of Timing Jitter

Another transmission obstacle is the various sources of random movement of the pulse in the bit slot, i.e., timing jitter of the pulses in the data transmission link. There are various sources of timing jitter, e.g., soliton interaction, Gordon–Haus, acoustic, and wavelength division multiplexing collision induced jitter.

Soliton interactions

Solitons, being nonlinear pulses, will interact with adjacent pulses in the pulse train. The interaction between solitons of the same polarization state and wavelength is phase sensitive so that in-phase solitons will attract each other whereas out-of-phase solitons will repel each other. To reduce this effect the solitons must be sufficiently spaced in the data stream. A typical design condition is that the pulse width of the solitons are one fifth of the bit slot, or one fifth of the inverse bit rate. A straightforward way of reducing the interaction is to have alternating amplitudes of the soliton in the pulse train, and the solitons can then be packed more densely.

Orthogonally polarized solitons interact substantially less, since it is the intensity overlap that causes the interaction, rather than the amplitude overlap as for copolarized solitons. In polarization multiplexed systems where adjacent pulse have orthogonal polarization it is usually enough with half the above separation, i.e., a pulse duration of 2.5 less than the bit separation.

Gordon–Haus jitter

Noise from the in-line amplifiers will give rise to a small jitter in the carrier frequency of each soliton, which, via the dispersion, will alter the group velocity and hence affect the arrival time of each pulse. This is known as the Gordon–Haus effect, and it has to be accounted for in long-distance systems. The variance of the timing jitter is proportional to the ratio of the power spectral density of the noise to the energy of the soliton, and also proportional to the transmission distance cubed. The fact that Gordon–Haus jitter grows cubically with distance makes it particularly important at transoceanic lengths, typically exceeding 1 Mm.

Acoustic jitter

The electrostriction nonlinearity in the fiber gives rise to a mechanical pressure proportional to the optical intensity, which in turn modifies the refractive index of the fiber. In particular, an intense optical pulse like a soliton will give rise to a pressure (acoustic) wave moving radially outwards from the fiber center. The pulses in the wake of this wave will experience a randomly changing local refractive index, and hence (just as for Gordon–Haus jitter) a randomly changing carrier frequency that transforms into a timing jitter. The variance of this jitter will grow as the transmission distance to the power of four.

WDM Considerations

Wavelength division multiplexing (WDM) is another way of increasing the bit-rate of soliton systems. That is, several frequency channels are used for solitons transmission. The problem with WDM transmission using solitons stems mainly from collisions between solitons in the different wavelength channels. In a perfectly ideal NLS equation solitons would collide elastically without changing carrier wavelength. However, the presence of losses and amplification may cause the collisions to be asymmetric if they occur around an amplifier, and the result will be a frequency displacement and a concomitant timing jitter.

Soliton Control

Soliton control is the common name for methods to control the soliton parameters such as wavelength and position, mainly in order to reduce the deleterious timing jitter effects. There are two different approaches: passive and active control.

Passive soliton control has been suggested in the form of filters that are inserted along the transmission path. This helps to keep the soliton wavelength fixed. In this way not only Gordon–Haus and acoustic jitter can be remedied, but also interaction jitter and

WDM-collision-induced jitter. A problem with this kind of filtering is that it defines a spectral region with excess gain, in which amplifier noise will grow excessively. A way around that problem is to slightly shift the center wavelength of the filters along the transmission path. In that way the solitons will follow the frequency shift, but the linear noise will not. Such sliding filter experiments have demonstrated 8×10 Gb/s WDM soliton transmission over 10 Mm.

Active control usually acts in the time domain by using phase or amplitude modulators to retime and reshape the solitons. Using this technique, 10 Gbit/s over unlimited distances has been demonstrated. However, this kind of active reshaping of the pulses suffers from the same drawbacks as conventional electronic regeneration, i.e., incompatibility with WDM, complexity, and high cost.

Dispersion-Managed Solitons

Introduction

In the last years of the 1990s, solitons have become substantially more attractive through the rapidly emerging strategy of improving the performance of soliton transmission with dispersion management (DM). While the DM strategy, which involves altering the local dispersion between a large positive and a large negative GVD such that the average GVD is small, has long been used in linear systems, it was only relatively recently appreciated that the same technique, if properly implemented, gives rise to several very striking improvements over conventional soliton transmission systems. While DM-solitons are clearly nonlinear pulses they are by no means classical solitons.

From a commercial viewpoint, however, the most important benefit with using DM-solitons is that they, in principle, can use the already installed conventional fibers (with zero dispersion at 1300 nm) allowing a much more cost-effective upgrade together with dispersion-compensating fibers (DCF) or chirped fiber gratings, which can be located together with in-line amplifiers at certain intervals in the link.

In Figure 1, an example of breathing in spectral and temporal widths of a DM-soliton can be seen, together with its dispersion map at the bottom.

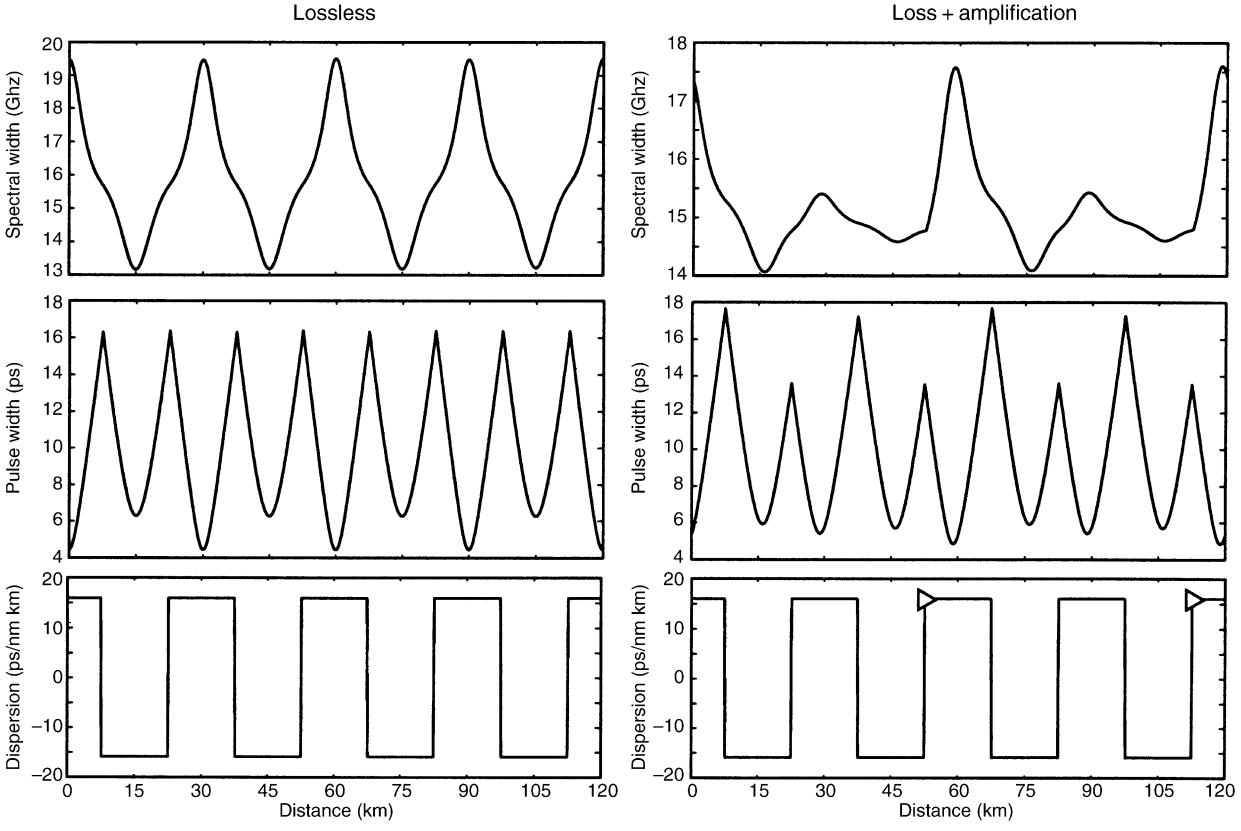


Figure 1 Plot of the spectral width (top), temporal width (middle), dispersion map (bottom) of a dispersion-managed soliton in the lossless case (left column) and the lossy case that is periodically amplified (right column). The positions of the amplifiers are denoted by the small triangles in the plot of the dispersion map.

Two cases are shown; one without fiber attenuation (left column) and one including fiber losses and periodic, lumped amplification every 60 km. One may note the strictly periodic behavior of both the spectral and temporal widths in both cases, and that the evolution of the spectral width is significantly different in the lossy case. This is because a change of the pulse spectrum requires nonlinearities and high power, and the spectral broadening is therefore highest directly after the amplifiers.

Properties

DM-solitons emerged from extensive simulation work, and it was quite surprising to many researchers that the simulations revealed such stable and strictly periodic pulses. The periodicity follows that of the dispersion variation of the system. Usually the system under study has a dispersion map, i.e., a plot of how the GVD varies along the transmission path. The map has a characteristic period (typically 10–100 km) and an average GVD which is significantly lower than the local GVD in the map.

As a result of the massive simulation work done by many groups the following properties of DM-solitons have been found:

- The pulse width, chirp, and spectral width oscillates periodically in the dispersion map. There are two points within the period at which the pulse is chirp-free, and those correspond to local minima of the pulse width. One of those points is the global minimum width, referred to as the 'shortest pulse' below.
- A central parameter that is useful for the characterization of DM-solitons is the map strength, $S = (L_1|\beta_1''| + L_2|\beta_2''|)T_{\mathrm{FWHM}}^2$, where L is the length, β'' is the dispersion, T_{FWHM} is the minimum pulse width in the full width half maximum sense, and the subscripts 1 and 2 refer to the two fibers in the dispersion map. In this example we thus assume the map to consist of two fibers, one with normal and one with anomalous dispersion. Physically the map strength is the number of dispersive lengths the pulse propagates in one period. DM-solitons have been found for map strengths ranging from $S = 0$ (which is the same as conventional solitons) to $S \approx 12$, although this upper limit is a transition regime in which the pulses radiate and perfect periodic evolution never arises.
- DM-solitons have been found for anomalous, normal, and zero average dispersion β'', defined as $\beta_0'' = (L_1\beta_1'' + L_2\beta_2'')/(L_1 + L_2)$. Normal average dispersion is only possible for map strengths above 3.9.
- The shortest pulse shape ranges from hyperbolic secant at $S = 0$ to Gaussian for higher map strengths, and this is also evident from the time-bandwidth product, which increases with S from 0.32 (sech-shape) at $S = 0$ to 0.44 (Gaussian shape) and even higher for large values of S. In addition, the shortest pulse has oscillating tails in the wings.
- The energy of a DM-soliton pulse is enhanced relative to a soliton with the same average dispersion and pulse width.
- The interaction between DM-solitons is less than that of conventional solitons, and an optimum map strength exists that minimizes the interaction.

The fact that DM solitons can work for an average net zero GVD and even for normal dispersion, was a striking and unexpected difference from conventional solitons. This can be understood by the fact that the spectrum is widest in the anomalous dispersion regime (i.e. positive values of the dispersion parameter D, cf. Figure 1), and it has been conjectured that the average dispersion, weighted by the spectral width of the DM-soliton, should always be anomalous. This is then a generalization of the fact that conventional solitons must always have a constant anomalous dispersion.

The technical improvements with DM solitons over conventional ones are numerous. The signal-to-noise ratio is improved since the DM-solitons have a larger peak power than the corresponding conventional solitons. DM-solitons have less Gordon–Haus and acoustic timing jitter, since the system average GVD is much smaller in these systems. A very important added benefit appears in WDM systems. Due to the alteration between large positive and negative GVD along the path, the jitter induced from WDM-soliton collisions is greatly reduced. This, in turn, allows for very dense WDM which will improve the spectral efficiency substantially.

An important practical consequence of using DM-solitons is that they reduce (or eliminate) the need for in-line soliton control such as synchronous modulation or sliding filters. Yet soliton control methods are still applicable and will give improvement in terms of signal-to-noise ratio for DM-solitons as well.

Quite impressive circulating loop experiments including WDM have been reported. For example, 51 densely packed WDM channels each operating at 20 Gb/s were transmitted over 1000 km with 100 km sections of standard fiber, clearly demonstrating the strength of the DM soliton technique.

Quasi-linear Transmission Intrachannel Impairments

As a rule, map strengths in the range 1–8 make best use of the unique features of DM-solitons. This means that the use of installed standard fiber becomes difficult at very high bit-rates (say beyond 30–40 Gb/s) as shorter pulses require a more rapidly (sub-km) varying dispersion map to maintain a proper S-value. If this cannot be maintained, i.e. for S values significantly above 10, one reaches the quasi-linear regime, in which the dispersion significantly dominates the transmission.

In the quasi-linear regime, there are a couple of nonlinear transmission impairments that need to be accounted for and analyzed in more detail. They are the so-called intra-channel effects; intra-channel four-wave mixing (ICFWM), and intra-channel cross-phase modulation (ICXPM), and arise due to the nonlinear interaction between two neighboring pulses. Four-wave mixing (FWM) and cross-phase modulation (XPM) are usually effects associated with WDM transmission. However, the fact that the pulses are chirped and broadened, will cause different frequency components from neighboring pulses within the same wavelength channel to overlap in time, thereby causing FWM and/or XPM within the same channel.

ICFWM arises for two neighboring pulses that, via four-wave mixing (FWM), creates new frequency components that in the time domain will give rise to a new pulse (commonly referred to as a ghost pulse), next to the two. The ghost pulse will then give rise to intersymbol interference and reduction of the eye opening. ICFWM is most prominent for large map strengths and high power.

ICXPM can be viewed as pulse-to-pulse interaction, and physically, it manifests as the frequency shift of one pulse induced by the presence of a neighboring pulse, which, by the dispersion transforms into a timing jitter. The effect can be minimized by selecting proper map strength and pre-chirp of the pulses.

Finally it should be emphasized that these intra-channel impairments will affect also non-soliton systems (e.g., non-return-to-zero (NRZ) modulated systems) if the nonlinearities are significant.

Experiments and Field Trials

Soliton Pulse Sources

When doing soliton experiments, be it conventional or DM-solitons, particular importance is placed on the properties of the pulse source, as it sets the lower limit of the system performance. A high bit-rate soliton pulse source needs to produce low-chirp, low timing-jitter pulses with proper duration (in the ps regime), repetition rate (10–40 GHz), and shape.

One possible choice is gain-switched (GS) laser diodes, possibly with an external cavity for tunability. However, they suffer the drawback of producing pulses that are strongly chirped, asymmetric and often too wide.

For laboratory experiments, fiber ring lasers (FRL) are very attractive, as they provide wavelength and pulse width tunability, besides meeting the above demands. Their drawback is that they are bulky, need active stabilization, and sometimes also temperature control to achieve long-term stability.

Finally it appears quite clear that electro-absorption modulators (EAM), which can be integrated or not with a distributed feedback (DFB) laser, are very useful and simple sources for soliton transmission. While such sources were developed for linear NRZ systems, they have now proven to be near ideal in soliton systems as well. Although EAMs are not commercially available at 40 GHz yet, they are likely to be so soon.

Special considerations need to be taken in DM-soliton systems, however, as the launch condition is different from that for conventional solitons. The pulses should have a linear chirp such that it fits seamlessly in the periodically induced chirp variation along the link. This can be achieved by incorporating a proper length of fiber (or chirped fiber grating) in the transmitter once the overall dispersion map is known.

Loop Experiments

In order to investigate really long distances (megameters) of transmission, loop experiments were developed in the early 1990s. This means that the data pulses are injected in a loop consisting of transmission fiber and amplifier, and then left to propagate a number of laps corresponding to a certain transmission distance. Acousto-optic switches are used to switch the pulse train in and out from the loop at proper time intervals. The drawback of loop experiments is that they may be poor models of reality when it comes to things like dispersion variation along the fiber, polarization mode dispersion (PMD), or various kinds of drift that may arise. In addition, a real system has more options to fine tune, e.g., amplifiers along the transmission line. However, as long as these drawbacks are recognized, loop experiments are very powerful indeed, and invaluable in laboratory evaluations of long-distance transmission.

Field Trials

In the field, there are many transmission link design restrictions and fiber properties which make the systems far from optimal. The actual fiber parameters are nonperiodic with propagation distance as the systems are straight lines rather than relatively short loops. Both loss (in particular when including many contacts and splices along the link) and dispersion might vary significantly along a fiber span. In addition, it may not be possible to tailor the dispersion map and amplifier locations to reach an optimal state. All these examples of nonidealities justify the need for field experiments.

Several soliton field experiments have been conducted over the past few years in Japan by NTT, in the US by MCI/Pirelli, as well as in Europe by ACTS projects, and the results from those experiments are discussed below. It is a good indication that solitons are indeed foreseen as very interesting candidates in commercial systems. Again, it is not very easy to compare the results as the situation in each case differs. All of the systems operated in the 1550 nm range, used optical time division demultiplexing with a 10 Gb/s electronic base rate, and the average loss/km ranged from 0.24 to 0.33 dB/km. Dispersion-shifted fiber was always used, apart from in two cases, where standard fiber was used. One field trial has used DCF for dispersion compensation, which makes this the only DM-soliton field experiment to date.

Polarization multiplexing, for which adjacent pulses have orthogonal polarization, have been used in a few experiments and this serves mainly to allow the use of relatively wide pulses, which in turn allows for larger amplifier spans. Polarization-multiplexing, however, is not as useful if the PMD of the system is high, as then the orthogonally polarized pulses would start to drift statistically in time relative to each other, thereby creating intersymbol interference and increasing the soliton interaction. In the 40 Gb/s cases and above, PMD was found to be the main capacity limiting factor.

Most of the recent experiments used a mode-locked fiber ring laser as a source, mainly because these provide excellent pulse quality as well as tunability in terms of wavelength and pulse width. Other experiments used either gain-switched lasers or electro-absorption modulators.

Future soliton field trials are expected to (1) take advantage of the now well-known strategy of improving soliton transmission performance with dispersion-management, this being very attractive for upgrading existing fiber plants, (2) implement dense WDM (in fiber lines that do not contain dispersion-shifted fibers) to boost aggregate capacity, (3) utilize different forms of in-line control, particularly at high bit rates, and (4) further address the implications of PMD and techniques to combat it. The interesting trade-off between wavelength and time division multiplexing (WDM–TDM) for optimization of overall aggregate capacity will depend on the details of the fiber line parameters.

Future Outlook

To conclude, we note that the motivation for using solitons as information carriers has changed over the years. The property of being resistant to dispersive broadening was originally the main feature, but this was considered less important when dispersion-compensating fibers became commercially available. Instead, this led to the development of the dispersion-managed soliton. The advantages of the DM-soliton over linear transmission are features like the large power which gives a high signal-to-noise ratio. On the other hand, the difference between linear and nonlinear pulses is becoming increasingly fuzzy, and perhaps the distinction should be made between return-to-zero (RZ) and NRZ modulation rather than between linear and nonlinear transmission.

It is nevertheless interesting to note that solitons are now not only considered for oceanic systems but also for shorter terrestrial systems. There are still several challenges and opportunities remaining in order to take full advantage of solitons and to reach a better understanding. Further work is also needed on WDM soliton- and very high-speed TDM soliton-systems. The use of DM-solitons is a very recently established technique and there are thus many issues to consider. These include studies of robustness to deviations of optimum conditions, e.g., improper pulse launch conditions, the impact of nonperiodic dispersion maps and of PMD (both of which are difficult to study in loop experiments), and intrachannel effects. Nevertheless, solitons have now reached a level of maturity such that commercialization seems very near.

List of Units and Nomenclature

Amplifier gain	G	$[-]$
Amplifier separation	L_a	[km]
Average dispersion	β_0''	$[\mathrm{ps}^2\,\mathrm{km}^{-1}]$
Dispersive length	L_D	[km]
Fiber loss coefficient	α	$[\mathrm{km}^{-1}]$
Fiber nonlinear coefficient	γ	$[\mathrm{W}^{-1}\,\mathrm{km} - 1]$
Group velocity dispersion (GVD)	β_0''	$[\mathrm{ps}^2\,\mathrm{km}^{-1}]$

Map strength	S	[−]
Optical field amplitude	$u(z, t)$	[$W^{1/2}$]
Pulse duration	t_0	[ps]
Soliton energy	E_s	[J]
Soliton peak power	P_s	[W]
Third-order dispersion	β_0''	[$ps^3\ km^{-1}$]
Time	t	[ps]
Transmission distance	z	[km]

See also

Fiber and Guided Wave Optics: Nonlinear Effects (Basics). **Solitons:** Temporal Solitons.

Further Reading

Agarwal GP (1995) *Nonlinear Fiber Optics*, 2nd edn. San Diego: Academic Press.

Agarwal GP (2001) *Nonlinear Fiber Optics*, 3rd edn. San Diego: Academic Press.

Agarwal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego: Academic Press.

Hasegawa A and Kodama Y (1995) *Solitons in Optical Communications*. Oxford: Clarendon Press.

Iannone E, Matera F, Mecozzi A and Settembre M (1998) *Nonlinear Optical Communication Networks*. New York: Wiley.

Mollenauer LF, Gordon JP and Mamyshev PV (1997) Solitons in high bit-rate long-distance transmission. In: Kaminow IP and Koch TL (eds). *Optical Fiber Telecommunications III A*. San Diego: Academic Press.

Smith NJ and Doran NJ (1995) Evaluating the capacity of phase modular-controlled long-haul soliton transmission. *Optical Fiber Technology* 1: 218–235.

Suzuki K and Nakazawa M (1995) Recent progress in optical soliton communication. *Optical Fiber Technology* 1: 289–308.

Taylor JR (ed) (1992) *Optical Solitons–Theory and Experiment*. Cambridge: Cambridge University Press.

The final report of the ACTS-MIDAS project, Jan 1999, can be downloaded at *www.infowin.org/ACTS/RUS/PROJECTS/FINAL-REPORTS/fr-053.pdf*.

Turitsyn SK, Mezentsev VK and Shapiro EG (1998) Dispersion-managed solitons and optimization of the dispersion management. *Optical Fiber Technology* 4: 384–452.

Temporal Solitons

S Trillo, University of Ferrara, Ferrara, Italy

A Tonello, Université de Bourgogne, Dijon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The fact that wave propagation phenomena are intrinsically dispersive, that is localized disturbances tend to spread, is part of our experience. Less well known is the fact that the spreading of wavepackets can be counteracted by the nonlinear response of the host medium. Solitons are wavepackets that exploit this principle, realizing a perfect balance of dispersion and nonlinearity that allows them to self-trap and travel undistorded, or at least to recover their shape periodically, for long (virtually unlimited) distances. More specifically, here we deal with optical temporal solitons constituted by envelopes, or pulses with relatively narrow bandwidths around a carrier optical frequency, whose temporal profile remains unchanged (either strictly or on average) upon propagation. In contrast with spatial solitons predicted earlier and such that nonlinearities are exploited to compensate for diffractive beam spreading even in the two transverse dimensions (i.e., in bulk samples, for details the reader is referred to **Solitons:** Bright Spatial Solitons), temporal soliton trapping is intrinsically a one-dimensional phenomenon typically observed in optical fibers where diffraction is compensated by the standard fiber guidance mechanism. Temporal solitons were first predicted and observed in single-mode fibers (SMFs). Their applications in digital optical communications have increased due to their capability to behave as nearly ideal bits. Meanwhile, however, the concept of temporal soliton has been considerably generalized by investigating new settings and effects such as polarization, parametric amplification (*see* **Nonlinear Optics, Basics:** $\chi^{(2)}$-Harmonic Generation), gratings (*see* **Fiber Gratings**), stimulated scattering (*see* **Scattering:** Scattering Phenomena in Optical Fibers), and fiber lasers (*see* **Lasers**: Optical Fiber Lasers), etc. Here we briefly review the essential features of these different families of temporal solitons. In particular, in the section on Fiber Temporal Solitons below, we introduce the basic model for fiber solitons, whereas we discuss their application to optical communications in the section on Solitons in Optical Communications below. In the sections on Multi-component Temporal Solitons, and Temporal Solitons in Bragg Gratings below, we introduce significant generalizations of the concept involving

multicomponent and slow-light grating solitons, respectively. Finally, in the section on Solitons on Dissipative nonlinear optics we discuss issues related to solitons in strongly dissipative systems. Note that, throughout the paper, the word soliton is used in the aforementioned general sense, which does not necessarily require the underlying model to be integrable (*see* the section on Fiber Temporal Solitons below,), as implied by the more restrictive and rigorous mathematical meaning of soliton. We further focus our discussion to bell-shaped solitons (for dark solitons, holes on a continuous-wave background, *see* **Solitons:** Optical Fiber Solitons, Physical Origin and Properties), which can be excited by means of standard laser pulses.

Fiber Temporal Solitons

An optical pulse is described by the complex electric-field envelope $E(z,t)$ that modulates the carrier frequency ω_0. Its evolution along the z axis of a silica SMF is described with good approximation by the following universal model for dispersive and weakly nonlinear systems, known as the nonlinear Schrödinger equation (NLSE),

$$\mathrm{i}\frac{\partial E}{\partial z} - \frac{k''}{2}\frac{\partial^2 E}{\partial t^2} + \gamma|E|^2E = 0 \qquad [1]$$

In eqn [1] the second term is responsible for dispersive spreading of pulses ruled by the group-velocity dispersion

$$k' = \left.\frac{\mathrm{d}^2 k}{\mathrm{d}\omega}\right|_{\omega_0}$$

(GVD, known also as chromatic or second-order dispersion), while the first-order dispersion

$$k' = \left.\frac{\mathrm{d}k}{\mathrm{d}\omega}\right|_{\omega_0}$$

has been removed in eqn [1] by introducing the retarded time $t = T - k'z$ measured in the pulse frame traveling at the natural group-velocity $V \equiv k'^{-1}$ of light, T being the physical time in the laboratory frame. The cubic term in eqn [1] is responsible for self-phase modulation (SPM) that stems from the instantaneous optical Kerr effect characterized by the effective coefficient $\gamma = k_0 n_{2I}/A_{\mathrm{eff}} = (\omega_0/cA_{\mathrm{eff}})n_{2I}$, where n_{2I} is the nonlinear index of silica, i.e., the nonlinear refractive index change divided by the intensity which has produced it. The real-world total electric field can be recovered as $\mathcal{E}(z,r,t) = \mathrm{Re}[f(r)E(z,t)\exp(\mathrm{i}k_0 z - \mathrm{i}\omega_0 T)]$, where the transverse modal profile $f(r)$ has been averaged out in eqn [1] to yield an effective fiber area A_{eff}.

The mechanism behind soliton formation can be understood as due to the instantaneous frequency changes along the pulse caused by the Kerr SPM, compensating those due to the group velocity dispersion (GVD) in the anomalous case ($k'' < 0$, at carrier wavelengths longer than 1.3 μm in standard SMFs) where the GVD induces a blue-shift on the pulse leading edge and a red-shift on the pulse trailing edge (in the absence of such compensation the pulse broadens because anomalous GVD means that blue-shifted frequencies travel faster than red-shifted frequencies). Such mechanism of compensation becomes ideal in the case of a soliton waveform. This concept finds its mathematical support in the fact that eqn [1] is a conservative (Hamiltonian) model with the remarkable property to possess infinite conserved quantities, a rather exceptional feature which makes it exactly integrable. Among the solutions, the following fundamental (so-called $N = 1$, see below) bright temporal soliton exists whenever $k'' < 0$:

$$E(z,t) = \frac{1}{\sqrt{\gamma L_{\mathrm{d}}}}\,\mathrm{sech}[\eta(t - \mathrm{d}k'z - t_0)] \times \exp\left[-\mathrm{i}\delta\omega t + \frac{\mathrm{i}}{2}\left(\frac{1}{L_{\mathrm{d}}} + k''\delta\omega^2\right)z + \mathrm{i}\phi_0\right] \qquad [2]$$

where t_0 and ϕ_0 are arbitrary initial position and phase, and $L_{\mathrm{d}} = (k''|\eta^2)^{-1}$ is the characteristic distance after which the GVD causes a significant pulse broadening, the so-called dispersion length. Temporal solitons [2] are characterized by two independent parameters, namely the duration $1/\eta$ and the frequency detuning $\delta\omega$, and possess several remarkable features:

(i) they travel with invariant shape by realizing the condition $N \equiv L_{\mathrm{d}}/L_{\mathrm{nl}} = 1$, i.e., exact balance between the two length scales associated with eqn [1], namely L_{d} and the nonlinear length $L_{\mathrm{nl}} = (\gamma P_1)^{-1}$ associated with the peak power $P_1 = (\gamma L_{\mathrm{d}})^{-1} = |k''|\eta^2/\gamma$ needed for the input pulse [2] to be a soliton;

(ii) in any given fiber with fixed GVD k'' and nonlinearity γ, a larger soliton power P_1 is required for a shorter duration, i.e., narrower solitons are taller;

(iii) solitons experience a phase retardation as a whole (particle), which increases with P_1, i.e., taller solitons have larger shifts;

(iv) as a consequence of Galileian invariance, a frequency shift $\delta\omega$ from ω_0 induces the soliton velocity to change from k'^{-1} to $(k' + \mathrm{d}k')^{-1}$, with $\mathrm{d}k' = \mathrm{d}k''\delta\omega$;

(v) unlike other waves in nonlinear media, collisions between solitons with different velocities are strictly elastic, as shown in the example of Figure 1a; and

(vi) solitons are extremely robust: for instance they can be excited with input peak powers differing from P_1, whereas a weak attenuation α ($\alpha L_\mathrm{d} \ll 1$) induces the soliton to reshape adiabatically, i.e., the peak power decreases, yet the pulse remains locally a soliton by adjusting its width, as displayed in Figure 1b.

More general (and cumbersome) solutions of eqn [1] describe periodic soliton evolutions, such as the interaction of in-phase adjacent input pulses at the same frequency (see Figure 1c), or a single input pulse with sech-shape but peak powers corresponding to higher integer values of N, so-called breathers because of their ability to recover periodically the initial shape after 'breathing' (see Figure 1d for $N = 3$) over the distance $\pi L_\mathrm{d}/2$.

Solitons [2] describe stable propagation-invariant pulses under ideal conditions (lossless case, relatively long pulses and short SMFs). In practical settings, the importance of solitons stems from their robustness against real-world perturbing phenomena described by terms that in turn break the integrability of the underlying model [1]. Situations of this kind arise, for instance, when one accounts for stimulated Raman scattering whose effect is to down-shift continuously the central frequency of ultra-short solitons, or in optical communication systems treated below.

Solitons in Optical Communications

Solitons [2] represent ideal bits for RZ (return-to-zero) optical data sequences. However, fibers have unavoidable losses of about 0.2 dB/km at the usual operating wavelength $\lambda_0 = 1550$ nm. To ensure long-haul connections, the optical signal must be periodically amplified, commonly using erbium-doped fiber amplifiers (EDFAs). The pulse peak power hence

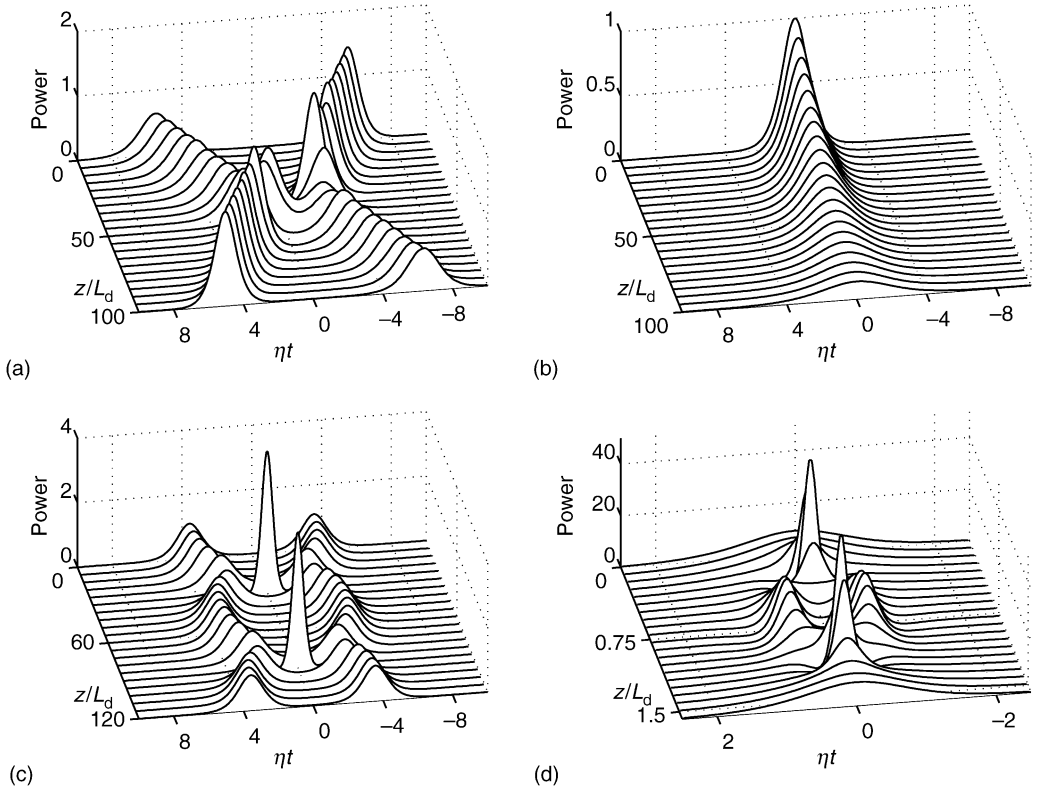


Figure 1 Fundamental features of fiber (NLSE) solitons: (a) elastic collision of two $N = 1$ solitons with different values of parameter pairs (frequency–velocity and amplitude–width); (b) adiabatic soliton damping in the presence of losses; (c) two-soliton bound state (in-phase interacting solitons of equal amplitudes); (d) evolution of a $N = 3$ soliton breather along one period of its evolution. Here the power $|E|^2$ is scaled in units of fundamental soliton power P_1, while distance z and time t are reported in units of L_d and initial pulse-width $1/\eta$, respectively.

undergoes both large exponential decay, because of distributed losses and abrupt lumped amplification under the action of EDFAs. Therefore, the stable local balance between GVD and nonlinearity cannot be rigorously sustained in the fiber. If both losses and amplifiers are accounted for, the mathematical model for pulse envelope propagation along M spans becomes:

$$\mathrm{i}\frac{\partial E}{\partial z} - \frac{k''}{2}\frac{\partial^2 E}{\partial t^2} + \chi|E|^2 E = -\mathrm{i}\alpha E + \mathrm{i}[\exp(\alpha Z_\mathrm{A}) - 1] \times \sum_{n=1}^{M} \delta(z - nZ_\mathrm{A})E(nZ_\mathrm{A}, t) \quad [3]$$

where the RHS contains the perturbing terms arising form distributed losses (α) and lumped amplifications (described by the comb with spacing Z_A).

The analysis of nonintegrable systems, such as eqn [3], can be carried out by appropriate transformations that allow the recovery of a reduced integrable model. The key result is that, under the limitations set by an appropriate length scale, the solitons of the reduced system can be close to those of the original problem. For temporal solitons, such transformation was first called 'the guiding center soliton', envisaging a similarity with the guiding center motion of a charged particle in a magnetic field. The main outcome is that eqn [3] can be reduced to the standard NLSE [1], clearly showing that soliton-like propagation still occurs in the presence of losses (even large, $\alpha L_\mathrm{d} \gg 1$) and rapid periodic amplification. The appropriate scale for these systems is the amplifier spacing Z_A, that must be much shorter than L_d. To reproduce an average behavior similar to the lossless case, it is sufficient to multiply the input condition for an enhancement factor that counteracts the effect of fiber losses. Therefore soliton-like solutions do exist even in fiber systems with loss management, though on average (over several fiber spans) rather than locally.

Another issue that makes solitons suitable as optical bits, lies in their intrinsic aptness to be controlled either through passive or through active systems. Solitons tend to follow their relative center of mass in the temporal or spectral domain. For instance, a powerful technique to overcome one of the major problems, i.e., the growth of noise due to amplified spontaneous emission, may be obtained by deploying periodically along the line optical passband filters. Mollenauer and co-workers have demonstrated that, by sliding the filters progressively (i.e., changing slightly the central peak from filter to filter), one can guide a shape-invariant soliton which adapts its carrier frequency while effectively suppressing the noise. Similar control techniques involve amplitude or phase modulators and a consistent amount of work has focused on the problem of all optical soliton regeneration.

To exploit the idea of soliton in optical communications, however, we should consider the fact that ultra-short optical pulses, suitable for high transmission rates, require extremely low fiber GVD to accomplish the stable balance between dispersion and nonlinearity for a fixed pulse-width and available power. For instance, in order to transmit at 40 Gbit/s, we should achieve a fixed GVD coefficient $D = -(2\pi c/\lambda_0^2)k'' \simeq 0.25$ ps/(km nm) for the whole link.

On the other hand, in many optical systems the fiber nonlinearity is not exploited at all but rather considered as a perturbing effect and a source of penalty that limits the maximum transmission distance. In such an approach the desired cumulated GVD should be as close as possible to zero to avoid errors coming from inter-symbol interference caused by dispersive broadening. Dispersion-shifted fibers may ensure low GVD only in a narrow spectral region, thus limiting the transmission bandwidth. Conversely GVD can be reduced to zero on average, simply combining fiber spans with opposite GVD. The periodic alternance of fiber types may also carry several benefits, e.g., limiting the impact of resonant nonlinear interactions such as four-wave mixing (*see* **Fiber and Guided Wave Optics:** Nonlinear Effects (Basics) for its definition). The different architectures of dispersion management proposed with the twofold scope of limiting nonlinear effects and reducing the cumulated GVD, have immediately stimulated the investigation of their impact on soliton propagation. In the early schemes, the attention was focused on the so-called adiabatic dispersion management, where the GVD was changed proportionally to the soliton power to improve soliton system performances. In contrast with previous works, nonadiabatic map profiles were introduced in the experiments by Suzuki and co-workers to compensate the cumulated dispersion by locating periodically dispersion compensating fibers (DCFs). In net contrast, Doran and co-workers have suggested the use of a combination of fibers with alternate normal and anomalous GVD (i.e., a periodic stepwise variation of GVD), predicting the existence of solitons with a peak power larger than the equivalent classical soliton power, i.e., dispersion-managed solitons (DMSs) The periodic stepwise GVD map is not only conceptually different from earlier designs (with exponentially decaying GVD), but has opened the field to real application due to its simplicity of implementation. The innovative discovery that special pulse waveforms may

propagate in periodically dispersion managed fiber links has triggered the interest of several groups, leading to a complete characterization of the problem, both theoretically and experimentally.

Similar to the case of guiding center solitons, a DMS usually shows a solitonic behavior. However, the DMS has a Gaussian shape surrounded by oscillating tails (see Figure 8 in **Solitons:** Optical Fiber Solitons, Physical Origin and Properties), thus being more suitable, due to its faster decaying tails, to be closely packed in a bit sequence in comparison with a sech-soliton (eqn [2]) of the same FWHM. We show an example of DMS transmission line in the top frame of Figure 2, where we sketched the variation of the GVD of a standard SMF periodically balanced by shorter segments of DCFs with opposite GVD. We report, in the middle frame, the DMS pulsewidth and chirp and in the bottom frame the pulse peak power. As shown, the macroscopic soliton parameters are characterized by large fluctuations: the local minima of pulsewidth correspond to chirp-free points and at such points one observes either a minimum or a maximum of the pulse spectral bandwidth. Nonetheless, all these pulse parameters have periodical evolutions that are fingerprints of an invariant and soliton-like pulse when observed at integer multiples of Z_A. The exact periodicity preserves the pulse waveform and hence the carried information. In DMS transmission we assist at the threefold compensation of energy, GVD and nonlinearity. Similarly to the guiding center solitons, the threefold compensation has an average, rather than pointwise, connotation. However, unlike guiding center solitons, for DMSs the definition of average should take into account the different relevance of pulse peak power in the map. Consequently the nonlinearity is compensated in a weighted manner, and it is not surprising in this context (though astonishing if placed back in the context of the NLSE [1] which requires anomalous GVD), that DMSs have been observed even when the uniform average GVD is weakly normal.

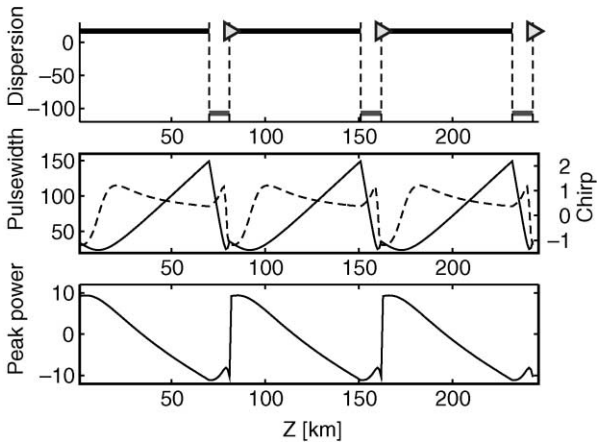


Figure 2 Evolution of soliton parameters along three periods of a fiber link with lumped amplifiers and dispersion management. Top frame: GVD map D(z) [ps/(nm km)] and EDFAs (triangles) location; Middle frame: soliton chirp (in units of 1000 ps^{-2}, dashed line), and pulse-width (in ps, solid line) vs. distance. Bottom frame: DMS peak power (dBm) vs. distance.

The net difference between the local and average GVD allows simultaneous control of different detrimental phenomena: for instance four-wave mixing is reduced by the large local GVD, whereas a small average GVD limits the impact of timing jitter due to amplified spontaneous emission (Gordon–Haus effect). As a result, DMSs perform better than conventional solitons.

Figure 3 shows a typical high-speed (20 Gb/s) single-channel experiment, and the achieved performance in terms of quality factor Q versus propagation distance, with the output after 8100 km. We recall that the Q factor measures the bit-error-rate (an error rate lower than 10^{-9} corresponds to a Q higher than $\sim$ 16 dB). DMSs also allow the control of collisions between solitons that pertain to different channels (we recall that solitons at different carrier frequencies go at different velocities, and hence collide), thus enhancing performances of standard or dense WDM systems.

Multi-Component Temporal Solitons

So far we have discussed temporal solitons described by a single envelope. However, the same mechanism of mutual balance between nonlinearity and dispersion can also be effective for multiple pulses, leading to trap simultaneously multicomponent (or so-called vector) solitons. Solitons of this type are, for instance, those due to the interaction of polarization modes in isotropic or birefringent Kerr media, higher-order modes in large-core (multimode) fibers, or mixing of pulses at different carrier frequencies that interact parametrically (i.e., without net exchange of energy with the medium). The latter case entails a broad class of solitons since virtually any parametric four-photon mixing process in cubic media (e.g., nondegenerate or partially degenerate four-wave mixing, third harmonic generation, ...), or three-photon mixing process in quadratic media (sum- and difference-frequency generation, second-harmonic generation (SHG), optical rectification, etc.) is allowed to sustain them. This opens up the field to an entirely new class of materials, namely noncentrosymmetric (quadratic) materials which usually have stronger nonlinear responses if compared with silica fibers. SHG can be taken as a paradigm to

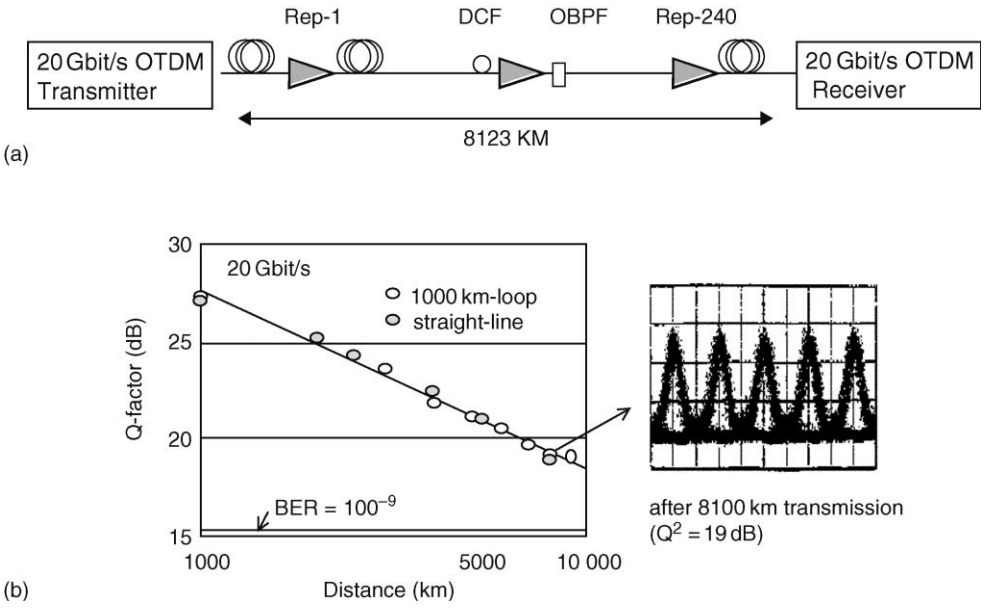


Figure 3 (a) Schematic diagram of the straight-line experiment with optical time-division multiplexing (OTDM) of soliton at 20 Gbit/s; (b) Measured *Q* factor vs. distance comparing experimental results obtained in the straight-line and loop configuration (loop length of 1000 km), respectively. The waveform received after 8100 km is also shown. With permission from Suzuki M and Edagawa N (2003) Dispersion-managed high-capacity ultra-long-haul transmission. *Journal of Lightwave Technology* 21: 916–929. © 2003 IEEE.

understand the main features of these multicomponent solitons. In this case, the evolution of the interacting envelopes E_j at carrier frequencies $j\omega_0$, $j = 1, 2$, is ruled by the coupled equations:

$$\left(\mathrm{i}\frac{\partial}{\partial z} - \frac{k_1''}{2}\frac{\partial^2}{\partial t^2}\right)E_1 + \chi E_2 E_1^* \exp(\mathrm{i}\Delta k z) = 0,$$

$$\left(\mathrm{i}\frac{\partial}{\partial z} + \mathrm{i}\delta k'\frac{\partial}{\partial t} - \frac{k_2''}{2}\frac{\partial^2}{\partial t^2}\right)E_2 + \chi E_1^2 \exp(-\mathrm{i}\Delta k z) = 0, \quad [4]$$

where the subscript $j = 1, 2$ indicates that the relative quantities evaluated at frequency $j\omega_0$, $t = T - k'_1 z$, $k''_{1,2}$ are GVD coefficients, $\delta k' \equiv k'_2 - k'_1$ is the group-delay, χ is the effective nonlinear coefficient, and $\Delta k = k_2 - 2k_1$ is the wave-number (or phase) mismatch. In spite of the fact that eqns [4] are not integrable, their stable soliton solutions can be found (numerically) in a wide range of parameters as symbiotic phase-locked pulses $E_{1,2}(z,t)$ that travel undistorted. For instance, SHG solitons exist either in phase-matching ($\Delta k = 0$) or far-off phase-matching (with some limitations for $\Delta k > 0$), and, unlike fiber solitons, also in normally dispersive media ($k''_{1,2} > 0$). As for fiber solitons, their width is related to their peak power, since the basic mechanism is the mutual balance between the dispersive broadening (that depend on the pulsewidth) and nonlinear phase-shifts (that depend on power), which occur due to repeated up-conversion and down-conversion processes, often referred to as cascading. A SHG soliton is robust enough to permit the observation of trapping, even from an input that differs considerably from the soliton (e.g., from the fundamental only), or in the presence of a relatively strong group-delay which tends to rip apart the two-component soliton, inducing walk-off between the two envelopes.

In contrast with fibers, in this case, experiments are necessarily made in relatively short bulk samples, thus resulting in a limited number of dispersion lengths. Usually broad (quasi-planewave) beams are conveniently employed in order to make diffraction ineffective.

The main experimental problem is that, in crystals commonly employed for SHG, the dynamics of typical laser pulses (duration above 100 fs) is fully dominated by the group-delay $\delta k'$. In other words, the material GVD becomes relevant only for pulses as short as a few fs. Therefore, the observation of solitons with longer pulses has been required to tailor the dispersion, which can be accomplished by means of pulses with tilted phase front. As displayed in **Figure 4**, successful compression of 200 fs (FWHM) tilted pulses was observed at relatively large negative Δk in a BBO bulk crystal with effective zero group-delay and enhanced (anomalous) GVD.

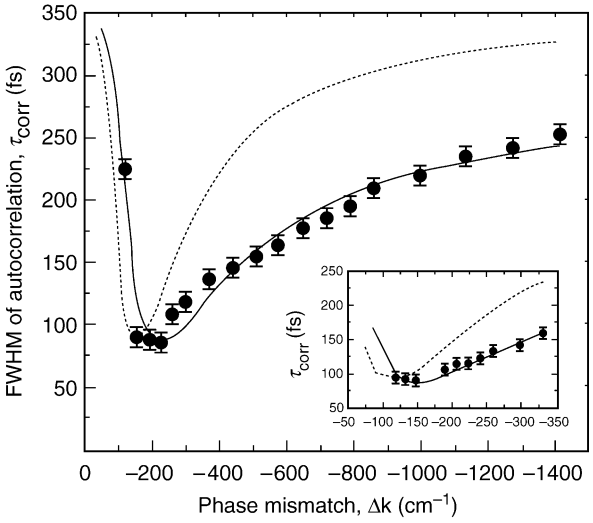


Figure 4 Soliton fomation via SHG: the circles show the output pulse duration (full width at half maximum (FWHM) of autocorrelation τ_{corr}) vs. mismatch Δk, measured after propagation of an input 200 fs tilted pulse (input $\tau_{corr} = 284$ fs) with 13.4 GW/cm^2 peak intensity, in a 7 mm BBO crystal (inset: same with 10.4 GW/cm^2). Superimposed (dotted line) is the corresponding theoretical value obtained from eqn [4]. The solid curve that fits better the data accounts for the additional effect of the material Kerr effect. With permission from Di Trapani P, Caironi D, Valiulis G, Dubietis A, Danielius R and Piskarskas A (1998) Observation of temporal solitons in second-harmonic generation with tilted pulses. *Physical Review Letters* 81: 570–573. © 1998 APS.

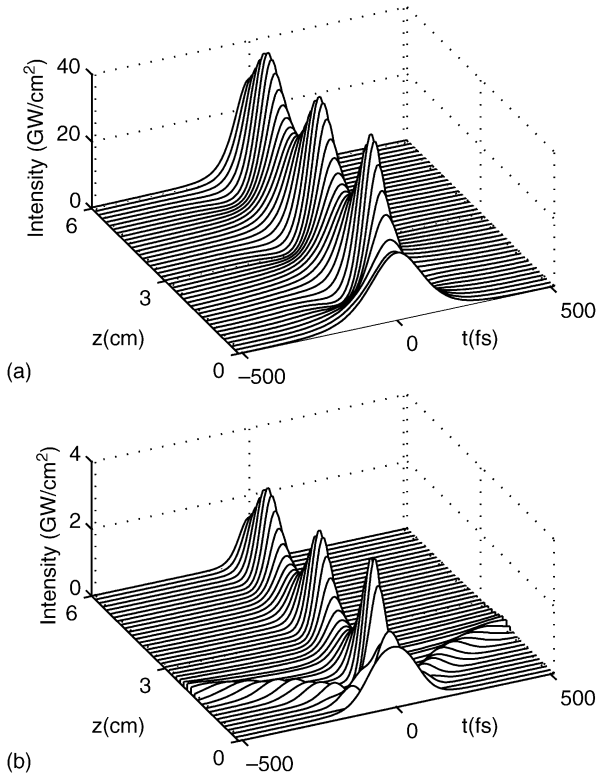


Figure 5 Dynamics of temporal soliton formation ruled by eqn [4]. A compressed soliton is formed via strongly mismatched ($\Delta k = -100$ cm^{-1}) SHG from a 200 fs (FWHM) input pulse at fundamental carrier frequency, in a crystal with $\chi = 1.4 \times 10^{-4}$ W$^{-1/2}$, effective dispersion $\delta k' = 0$ (group–velocity matching) and $k_1'' = -1.8$ ps^2/m, $k_2'' = -0.75$ ps^2/m.

The measured compression is a signature of soliton formation, as clearly indicated by the example of simulated pulse dynamics reported in Figure 5: the pulse at second-harmonic is generated and then trapped with the input one as a two-component bound state. Radiation and oscillations occur, in this case, because the launching conditions (single envelope) are far enough from the two-component soliton.

Temporal Solitons in Bragg Gratings

A linear periodic structure, such as a Bragg grating with pitch Λ, couples counter-propagating waves when their frequency lies close to the Bragg frequency ω_B (vacuum wavelength $\lambda_B = 2n_0\Lambda$) defined by the resonance condition $k(\omega_B) = \pi/\Lambda$. Such a type of structure supports a different type of temporal solitons, so-called gap solitons (GSs) since they originate from a linear dispersion relationship of the type shown in Figure 6a, characterized by a stop-band (gap) centered around ωB. The gap is characterized by a unitary grating reflectivity, as shown in Figure 6b, and the forward linear propagation is therefore forbidden. However, if the incoming wave is intense enough to induce the refractive index to increase through the optical Kerr effect, the Bragg resonance (and the whole gap) is shifted downwards (in frequency) allowing for self-transparency at the operating frequency. GSs can be understood as bell-shaped envelopes which are allowed to travel along the grating due to a high-intensity core that bleaches the reflectivity, while the high reflectivity seen by the tails takes the pulse together. GSs are expected to be slow because the slope of the dispersion curve tends to zero when approaching the gap, while its curvature results in the grating GVD (huge if compared with material GVD in real-world units), which is exactly balanced by the nonlinearity when the GS is formed. This naive portrait is supported by a more thorough analysis starting from the coupled-mode equations obeyed by the forward $E_+(z,T)$ and backward $E_-(z,T)$ envelopes at frequency ω_B, propagating along a (e.g., fiber) Bragg grating with Kerr effect (inducing SPM and

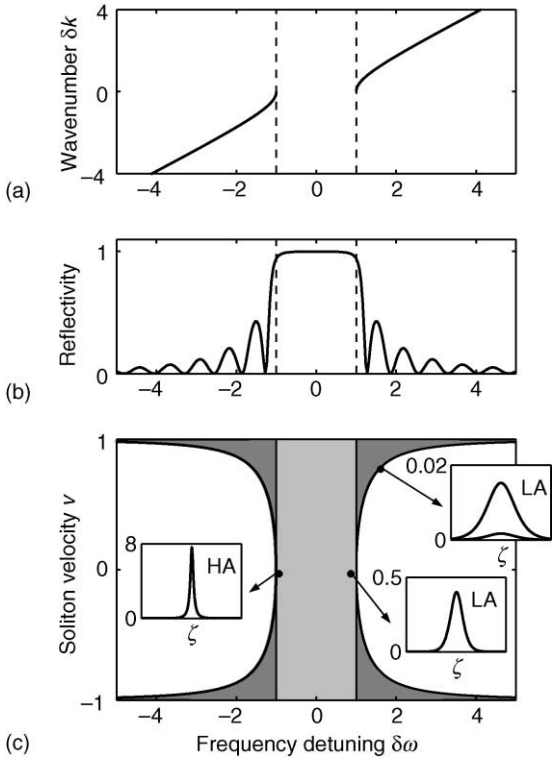


Figure 6 (a) Bragg linear dispersion relationship in normalized (dimensionless) units: wave-number shift $\delta k = (k(\omega) - \pi/\Lambda)/\Gamma$ vs. frequency detuning $\delta\omega = k'(\omega - \omega_B)/\Gamma$; (b) linear grating reflectivity vs. $\delta\omega$ for a grating figure of merit $\Gamma L = 4$; (c) existence domain (whole shaded area) of GSs in the normalized detuning–velocity plane ($\delta\omega, \nu \equiv k V_{sol}$). The light shaded area corresponds to the stop-band $|\delta\omega| < 1$ (inner region between dashed lines in a–b). The insets show samples of GS intensity profiles vs. the moving frame coordinate $\zeta = (1 - v^2)(z - V_{sol} T)$. Zero-velocity GSs have symmetric envelopes ($|E_+| = |E_-|$) and range from high-amplitude (HA) to low-amplitude (LA) across the gap. The third inset shows a moving GS in the experimentally accessible region, where it retains LA but has a stronger component in the direction of motion.

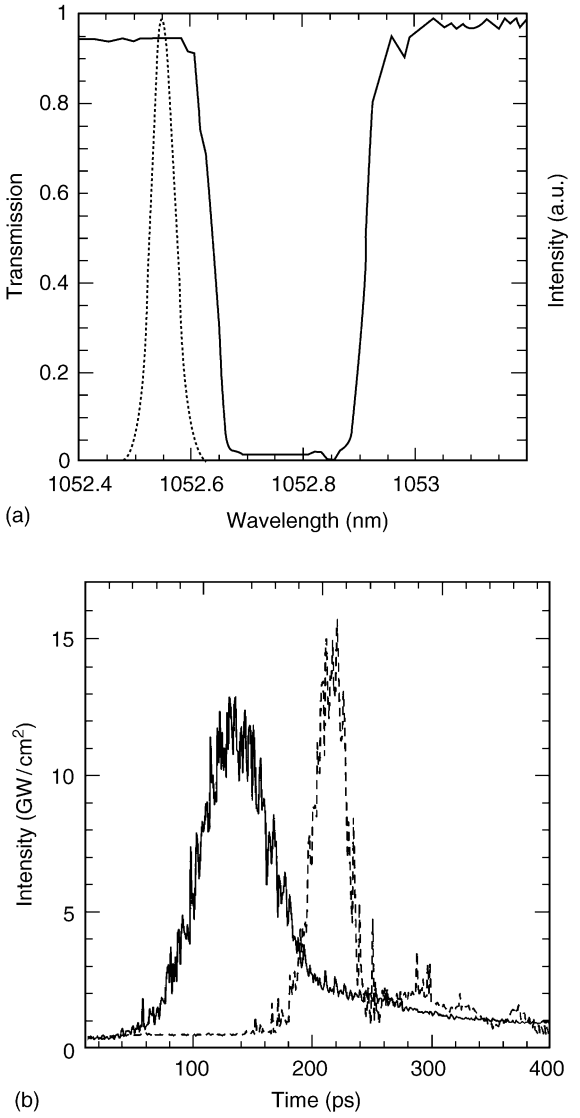


Figure 7 Gap solitons in a fiber Bragg grating: (a) measured pulse spectrum compared with the linear grating transmission, (b) compressed (15 ps) and delayed pulse transmitted at high intensity (dashed line), compared with the pulse transmitted at frequency far from the Bragg resonance (solid line) where it retains its input width of 80 ps. With permission from Eggleton BJ, Slusher RE, de Sterke CM, Krug PA and Sipe JE (1996) Bragg grating solitons. *Physical Review Letters* 76: 1627–1630. ©1996 APS.

cross-phase modulation):

$$i\left(k'\frac{\partial E_\pm}{\partial T} \pm \frac{\partial E_\pm}{\partial z}\right) + \Gamma E_\mp + \gamma(|E_\pm|^2 + 2|E_\mp|^2)E_\pm = 0 \quad [5]$$

The linear dispersion associated with eqn [5] is as in Figure 6a, with a stop-band $|\omega - \omega_B| \le \Gamma/k'$ proportional to the index corrugation depth through the Bragg coupling coefficient Γ. In the nonlinear regime, GS solutions of eqn [5] differ qualitatively from conventional fiber solitons [2]:

(i) first, GSs are indeed slow since the two envelopes can travel locked with any soliton velocity V_{sol} such that $|v| \equiv |k' V_{sol}| \le 1$, i.e., lower than the natural group velocity k'^{-1};

(ii) GSs exist only in a narrow bandwidth, i.e., the shaded area in Figure 6c, which corresponds to the gap seen in the soliton frame (it reduces to the conventional stop-band for still solitons and gets wider for faster GSs); and

(iii) the gap is strongly asymmetric since lower intensity is required to induce transparency close to the high-frequency band-edge where, as a consequence, GSs have low amplitudes and

are stable (the amplitude grows moving towards the low-frequency bandedge, where GSs become unstable).

Experimentally, GSs have been typically observed with pulses at frequency close to the blue bandedge of fibers (see **Figure 7a**) and AlGaAs waveguides with built-in Bragg gratings of a few centimeters. The typical signature of GSs, illustrated in **Figure 7b**, is the enhanced pulse transmission observed under the operating conditions of **Figure 7a** when the input intensity is sufficiently high. The transmitted pulse experiences pulse compression and a significant group-delay (measured with respect to the transit-time at frequency far off the stop-band), witnessing its nature of slow-light solitons.

Finally, we emphasize that grating solitons can also be supported in the case where the two modal envelopes propagate in the forward direction along a grating whose period Λ is long enough to be resonant with the modal phase slippage, thus inducing efficient linear coupling between them. In this case, the structure has a gap in wave-number, and solitons are slow in the frame traveling at the average modal group-velocity. Soliton formation is expected to lead to grating-induced compensation of the group-delay between the modes. Experiments on such types of structures have been carried out only recently.

Solitons in Dissipative Nonlinear Optics

Temporal solitons can also be observed in nonlinear optical interactions that involve changes of the state (energy level) of matter as in inverted media or scattering phenomena. For instance, early studies in the 1960s have demonstrated that coherent pulses with energy above a critical level can pass through resonant (two-level) atoms without being absorbed, a phenomenon denoted as self-induced transparency. Such pulses are temporal solitons solutions of the Maxwell–Bloch model which describes the atom–light interaction.

Also scattering phenomena occurring in optical fibers, such as Raman and Brilluoin scattering, possess three-wave solitons, in which the pump and scattered (Stokes) light waves are coupled to a phonon field of the acoustic or optical branch (for Brillouin or Raman, respectively).

Perhaps more important from the point of view of the modern applications, are fiber lasers composed by spans of active and/or passive fiber closed to form a recirculating loop. These types of structures can often be described by distributed (averaged over several round-trips) models, which take the form of a dissipative perturbed NLSE, known as the Ginzburg–Landau equation (GLE):

$$i\frac{\partial E}{\partial z} - \frac{k''}{2}\frac{\partial^2 E}{\partial t^2} + \chi|E|^2E = igE + ig_{nl}|E|^2E + iq|E|^4E + \frac{i}{\Omega_f^2}\frac{\partial^2 E}{\partial t^2} \quad [6]$$

where g is the excess (net) gain ($g > 0$) or loss ($g < 0$), g_{nl} and q describe the nonlinear gain mechanism and its saturation (usually introduced to control the mode-locking), and Ω_f is linked to spectral filtering. Several soliton (pulse) solutions of the GLE [6] can be found even when the parameters (g, g_{nl}, q, Ω_f) are such as to constitute a strong perturbation of the NLSE. Unlike NLSE solitons [2], pulse solutions of the GLE [6] are usually chirped (i.e., possess a nonlinear phase profile), and have fixed amplitude for a given value of the parameters. The set of soliton waveforms is generally richer and includes, for instance, flat-top stationary pulses, two-pulse bound-states, solitons moving at fixed velocity, etc. Although instabilities of the background (for $g > 0$) or of the pulse itself can affect their observability, large islands of stability in the parameter space can be found. The relevant role of stable solitons in mode-locking and intra-cavity pulse formation mechanisms has been demonstrated in many different fiber laser geometries, for which we refer the reader to the more specialized papers listed as Further Reading at the end of this article. We point out that some schemes, for instance the stretched pulse-fiber laser developed at MIT, involve a basic configuration (alternating GVD within the cavity length) which is quite similar to that seen in the framework of DMSs, thus exhibiting the same basic physical mechanism (strong compression and stretching of the pulse along each round-trip which play the role of the map period for a DMS).

List of Units and Nomenclature

BBO	β-barium borate
DCF	dispersion compensating fiber
DMS	disperion managed soliton
EDFA	Erbium doped fiber amplifier
GLE	Ginzburg–Landau equation
GS	gap soliton
GVD	group-velocity dispersion
NLSE	nonlinear Schrödinger equation
OTDM	optical time-division multiplexing
SHG	second-harmonic generation

SPM self-phase modulation
WDM wavelength division multiplexing
Amplifier spacing L_A [m]
Bragg frequency ω_B [Hz]
Bragg period Λ [m]
Central wavelength λ_0 [nm]
Coupling coefficient Γ [m^{-1}]
Dispersion coefficient D [ps nm^{-1} km^{-1}]
Dispersion length L_d [m]
Effective area A_{eff} [m^2]
First-order dispersion k' [s/m]
Fundamental soliton power P_1 [W]
Gain (linear) g [m^{-1}]
Gain (nonlinear) g_{nl} [m^{-1}]
Group-delay δ [s/m]
Group-velocity $V = k'^{-1}$ [m/s]
Group velocity dispersion k'' [ps^2 m^{-1}]
Inverse pulse-width η [s^{-1}]
Light velocity in vacuum c [m/s]
Loss coefficient α [m^{-1}]
Nonlinear coefficient (fiber) γ [m^{-1} W^{-1}]
Nonlinear coefficient (quadratic bulk materials) χ [$W^{-1/2}$]
Nonlinear refractive index n_{2I} [m^2 W^{-1}]
Nonlinear length L_{nl} [m]
Refractive index (linear) n_0
Retarded time (in the moving frame) t [s]
Spectral filtering coefficient Ω_f [Hz $m^{1/2}$]
Time in the laboratory frame T [s]
Wave-number mismatch Δk [m^{-1}]

See also

Fiber and Guided Wave Optics: Nonlinear Effects (Basics); Nonlinear Optics. **Fiber Gratings**. **Lasers:** Optical Fiber Lasers. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation. **Optical Communication Systems:** Wavelength Division Multiplexing. **Scattering:** Scattering Phenomena in Optical Fibers. **Solitons:** Bright Spatial Solitons; Optical Fiber Solitons, Physical Origin and Properties.

Further Reading

Ablowitz MJ, Biondini G and Ostrovsky LA (eds) (2000) Focus issue: optical solitons: perspectives and applications. *Chaos* 10: 471–640.

Agrawal GP (1995) *Nonlinear Fiber Optics*, 2nd edn. London: Academic Press.

Agrawal GP and Kivshar YS (2003) *Optical Solitons*. London: Academic Press.

Akhmediev NN and Ankiewicz A (1997) *Solitons: Nonlinear Pulses and Beams*. London: Chapman and Hall.

Buryak AV, Di Trapani P, Skryabin D and Trillo S (2002) Optical solitons due to quadratic nonlinearities: from basic physics to futuristic applications. *Physics Reports* 370: 63–235.

De Sterke CM, Eggleton BJ and Sipe JE (2001) In: Trillo S and Torruellas W (eds) *Spatial Solitons*. Berlin: Springer-Verlag.

Di Trapani P, Caironi D, Valiulis G, Dubietis A, Danielius R and Piskarskas A (1998) Observation of temporal solitons in second-harmonic generation with tilted pulses. *Physics Review Letters* 81: 570–573.

Eggleton BJ, Slusher RE, de Sterke CM, Krug PA and Sipe JE (1996) Bragg grating solitons. *Physics Review Letters* 76: 1627–1630.

Forysiak W, Knox FM and Doran NJ (1994) Average soliton propagation in periodically amplified systems with stepwise dispersion-profiled fiber. *Optics Letters* 19: 174–176.

Hasegawa A (2000) Soliton-based optical communications: An overview. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1161–1172.

Hasegawa A and Kodama Y (1995) *Solitons in Optical Communications*. London: Oxford University Press.

Hasegawa A and Matsumoto M (2002) *Optical Solitons in Fibers*. Berlin: Springer-Verlag.

Haus HA (2000) Mode-locking of lasers. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1173–1185.

Mollenauer LF and Mamyshev PV (1998) Massive wavelength-division multiplexing with solitons. *IEEE Journal of Quantum Electronics* 34: 2102.

Nakazawa N, Kubota H, Sahara A, Suzuki K and Yamada E (2000) Ultrahigh-speed long-distance TDM and WDM soliton transmission technologies. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 363.

Porsezian K and Kuriakose VC (2003) *Optical Solitons: Theoretical and Experimental Challenges*. Berlin: Springer-Verlag.

Slusher RE and Eggleton BJ (eds) (2003) *Nonlinear Photonics Crystals*. Berlin: Springer-Verlag.

Suzuki M and Edagawa N (2003) Dispersion-managed high-capacity ultra-long-haul transmission. *Journal of Lightwave Technology* 21: 916–929.

Suzuki M, Morita I, Edagawa N, Yamamoto S, Toga H and Akiba S (1995) Reduction of Gordon-Haus jitter by periodic dispersion compensation in soliton transmission. *Electronic Letters* 31: 2027–2028.

Turitsyn SK, Aceves AB, Jones CKRT, Zharnitsky V and Mezentsev VK (1999) Hamiltonian averaging in soliton-bearing systems with a periodically varying dispersion. *Physics Review E* 59: 3843–3846.

Wabnitz S, Kodama Y and Aceves AB (1995) Control of optical soliton interactions. *Optics and Fiber Technology* 1: 187–217.

Zakharov VE and Wabnitz S (1999) *Optical Solitons: Theoretical Challenges and Industrial Perspectives*. Berlin: Springer-Verlag.

SPECTROSCOPY

Contents

Absolute Optical Frequency Metrology
Fourier Transform Sepctroscopy
Hadamard Spectroscopy and Imaging
Nonlinear Laser Spectroscopy
Raman Spectroscopy
Second-Harmonic Spectroscopy
Single Molecule Spectroscopy

Absolute Optical Frequency Metrology

S T Cundiff, National Institute of Standards and Technology and University of Colorado, Boulder, CO, USA

L Hollberg, National Institute of Standards and Technology, Boulder, CO, USA

Published by Elsevier Ltd.

Introduction

Visible light occupies a portion of the electromagnetic spectrum between frequencies of approximately 400 terahertz (THz) to 750 THz, where 1 THz is 10^{12} oscillations per second. Optical radiation, which includes adjacent infrared and ultraviolet portions of the spectrum, covers a frequency range of approximately 15–1500 THz. The techniques required for measuring frequencies in this range are qualitatively different from those currently employed in the radio frequency (RF) and microwave portions of the spectrum, where high-speed electronic counters make a direct measurement possible.

The frequency of an oscillatory phenomenon is the number of oscillations that occurs in unit time, or alternatively, the inverse of the period of one oscillation. Thus, to understand frequency, we must understand time, which is intrinsically related to phase. Only a time interval, the time elapsed between two events, has fundamental physical meaning. The 'time' that we are most familiar with, the time of day (including the date), is based on an arbitrarily chosen starting point. The starting point chosen in current international agreements can be traced to noon on December 31st, 1899. Furthermore, the time defined by those agreements, Universal Coordinated Time (UTC), is periodically adjusted relative to atomic time (discussed below) to compensate for irregularities in the rotation of the Earth.

Knowledge of the transition frequencies (energies) in simple atoms gives detailed information on the structure of the atoms. This can be used to test our understanding of the fundamental interactions in nature by comparing the measured frequencies with predictions of fundamental physical theories, in this case quantum electrodynamics (QED). Beautiful results along these lines are exemplified by precision measurements of transition frequencies in hydrogen and helium. In addition, the ability to measure and precisely control laser frequencies promises tremendous advances in the performance of the next generation of atomic clocks and frequency standards. These have potential applications to navigation and communication systems.

The measurement of frequency and time intervals also involves a choice, that of the unit of time, i.e., the definition of the second. The original definition of the second was based on the rotation of the Earth. However, not only is it irregular, but it is slowing down. Both of these facts limit its utility as a frequency standard. Once atomic clocks were highly developed, the second was redefined in 1967 to be the time it takes for the $F = 4$, $m_F = 0 \rightarrow F = 3$, $m_F = 0$ transition in the hyperfine structure of the ground state of ^{133}Cs to undergo 9192 631 770 oscillations. Thus an 'absolute' measurement of time interval, and hence frequency, must be directly connected to this defining frequency. The basic structure of an atomic clock is shown in Figure 1.

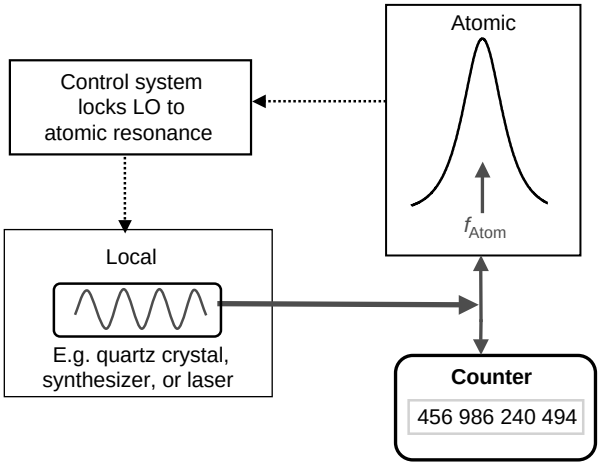


Figure 1 An atomic clock has three basic components: an atomic resonance, an oscillator to probe and lock to the atomic resonance, and a counter that records the number of oscillations and thus displays the time interval since some chosen starting time.

Frequency standards (or clocks) are characterized by three properties: accuracy, stability, and reproducibility. Accuracy describes how well the natural and fundamental atomic frequency is produced by the atomic frequency standard. Stability describes the frequency fluctuations on short timescales (an accurate but unstable standard produces a frequency that fluctuates around a constant mean that is the atomic frequency). Reproducibility between standards addresses whether or not two implementations of the standard produce exactly the same frequency. Atomic clocks are intrinsically accurate because the frequency of the transitions they are based on is determined by identical atoms in nature. The largest inaccuracies are usually due to environmental perturbations such as tiny variations in the local magnetic field.

A careful analysis of all of these issues, together with technical considerations, yielded the choice of cesium for the current definition of the second. Other common frequency standards include quartz crystal oscillators, hydrogen masers, and rubidium vapor cell standards. Quartz oscillators are inexpensive and have good stability, but in comparison to atomic standards they suffer from accuracy and reproducibility issues. Hydrogen masers have the best stability, but also poor accuracy and reproducibility compared to cesium. Vapor cell rubidium standards can be small and inexpensive but have poorer accuracy than cesium beam standards due to cell and buffer gas effects.

There are several motivations for using optical sources rather than RF or microwaves to measure time intervals and frequencies. Obviously, a higher oscillation frequency can divide time into smaller units and this can provide higher measurement precision. It is thus natural to use the highest oscillation frequency that can be precisely counted to measure time and frequency. Not long after lasers were invented, suggestions were made that it would be possible to make optical frequency standards using lasers locked to narrow atomic resonances.

A basic feature of optical transitions is that they have a very high Q factor (the ratio of the center frequency to the linewidth). This allows them to be measured to very high precision (current measurements are good to a few parts in 10^{15}). It may be possible to use this high precision to build improved atomic clocks that use optical transitions instead of the microwave transition currently being used. Candidates for optical standards include both trapped ions (Hg^+, Yb^+, Sr^+, and In^+ are under investigation) and laser-cooled neutral atoms (Ca, Sr, Mg, Ba, and Ag). The high precision is also utilized for fundamental tests of quantum electrodynamics by comparing transition frequencies (usually of hydrogen or helium atoms) to first-principle calculations. For comparison Figure 2 shows the relevant transitions and energy levels in the Cs microwave standard and the Hg^+ optical frequency. It is even possible to use the precision of atomic clocks for high-sensitivity tests of variation in the fundamental constants with time. Finally, because the speed of light is constant length measurements based on the wavelength of light with a known frequency, the precision of length metrology ultimately rests on optical frequency metrology.

An obvious question, which might occur to many in the field of optics, is why a standard spectrometer does not provide absolute frequency measurements. All spectrometers measure wavelength, not frequency, based on interference between light traveling two different path lengths. Thus, knowledge of the path length is needed to determine the wavelength. However, since length is now defined using the speed of light, the definitions become circular unless a source of light with known absolute frequency is available. In addition, frequency measurement is intrinsically more precise and reliable than mechanical length measurement.

Recently, there has been a significant breakthrough in optical-frequency metrology and optical clocks by using mode-locked lasers that generate optical pulses with durations of a few femtoseconds. Before discussing optical frequency metrology based on mode-locked lasers, we will review earlier techniques to provide the necessary background. After describing

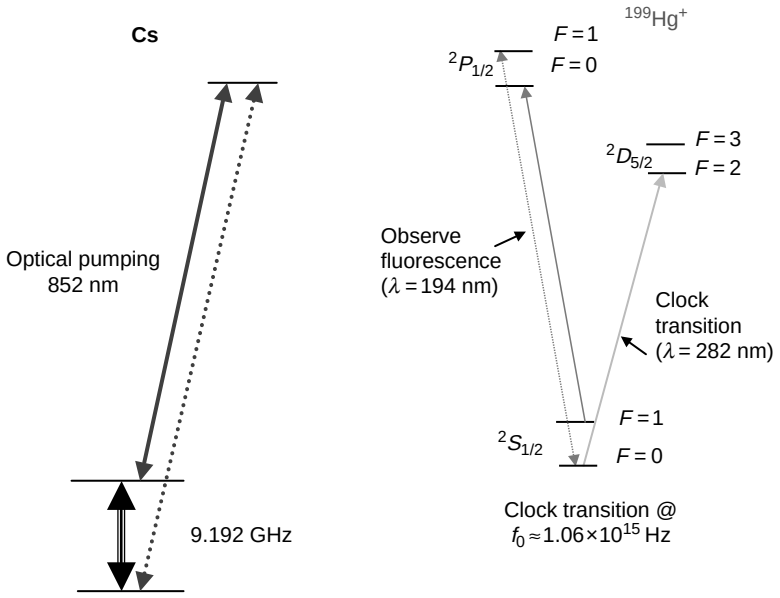


Figure 2 Simplified energy level diagram of Hg^+ and Cs.

methods that use mode-locked lasers, we will then provide a summary of current measurements and standards. Finally we will briefly describe the outlook for the future of absolute optical frequency metrology.

Background

To imagine being able to measure the frequency of light we need to have laser sources with high coherence, like that of radio waves, where the phase is stable enough to be measured. Even though the first beat-note between two lasers was demonstrated in the 1960s, most lasers were not coherent (stable) enough that the optical phase could be tracked for much longer than about 10 to 100 ns. After many years of research and inspiring technological achievements, Hall, Chebotayev, Bergquist and others were able to stabilize the frequency of tunable laser sources to the point that lasers could indeed be considered coherent sources, with the phase continuously measurable and even predictable for times as long as seconds. Once the lasers were stable enough to be counted, there was still the essential problem that no counter was fast enough to actually count optical frequencies.

With spectrally pure laser sources it was possible to conceive of measuring optical frequencies by extending the nonlinear methods that are used to generate and measure RF and microwave frequencies. By using simple nonlinear components such as diodes, it is easy to generate high harmonics of an input frequency in the RF and microwave regions of the spectrum. At least in principle, and by direct analogy, it seemed that it should be possible to multiply coherent sine waves from radio frequencies up through the microwave region to reach the THz frequencies of far-infrared lasers, and then to use nonlinear optical methods to multiply on up to the infrared (IR) and eventually to the visible region of the spectrum. The concept is simple enough, $f \rightarrow 100 \times f \rightarrow 1000 \times f \rightarrow 2000 \times f \rightarrow 4000 \times f \ldots$ until you reach an optical frequency. Multiplication up to the microwave region works quite well, but unfortunately the nonlinear optical processes in the far-infrared, IR, and optical regions are typically quite inefficient; the power at the second harmonic is typically only 10^{-5} times the square of the fundamental power in watts. Just multiplying a single-frequency laser by a factor of two can be a technological challenge. After reaching a microwave frequency of about 50 GHz we still require a multiplication factor of 10^4 to reach the visible at 500 THz. Since we are usually forced to multiply in the IR and optical by factors of 2, we have $2^N = 10^4$, which means we require $N \approx 13$ multiplication stages. This is approximately the number of stabilized laser oscillators that are required to span the frequency gap from the microwave to the visible. Notwithstanding these technological limitations, Evenson and collaborators did a groundbreaking demonstration of a harmonic frequency chain in 1972 that connected a microwave atomic frequency standard and a stabilized HeNe laser in the IR at 88 THz. With a concerted multiperson effort, more lasers, and 10 years of research, Jennings and co-workers finally succeeded in extending the harmonic optical frequency chain to the visible. That original optical frequency chain from the RF to the

visible is diagramed in Figure 3. In the succeeding 18 years, only three other harmonic frequency chains were built world-wide that connected the microwave standards to optical frequencies.

An alternative scheme for measuring optical frequencies is the optical bisection method, a conceptually simple technique proposed by Telle, Meschede and Hänsch, for dividing down from

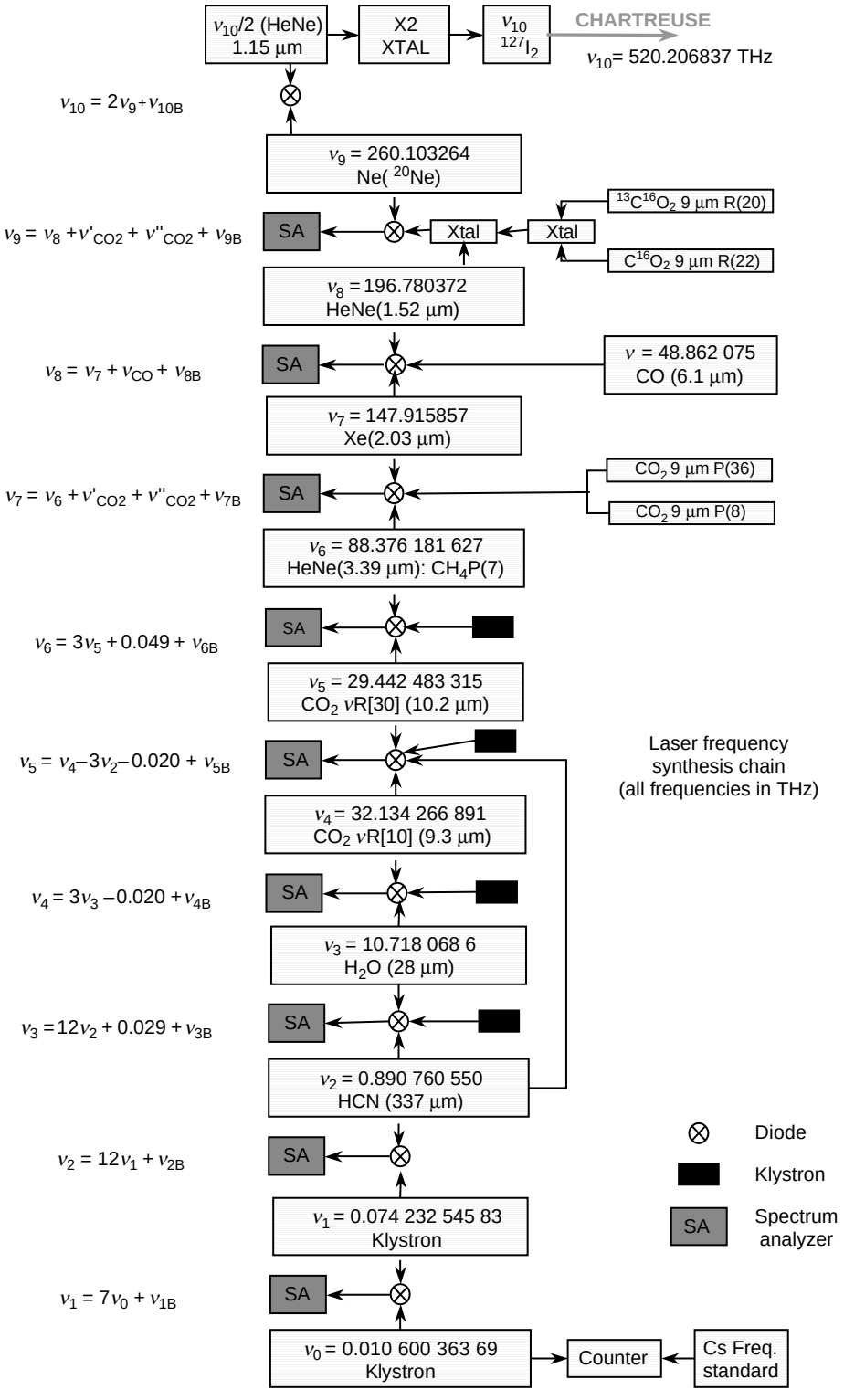


Figure 3 Schematic of the first frequency chain used to measure optical frequencies relative to the Cs standard.

optical frequencies to microwaves. In fact, the bisection method divides an arbitrary optical frequency interval in half by utilizing second-harmonic generation and sum-frequency-mixing, and forcing the condition that $2f_3 = f_1 + f_2$. Thus, the laser at frequency $f_3 = (f_1 + f_2)/2$ bisects the frequency interval between the lasers at f_1 and f_2. The optical bisection method has the following two significant advantages over the multiplication method: the phase noise is decreased in successive divisions (rather than increased by multiplications), and the system could be constructed out of one convenient and reproducible laser technology, say diode lasers in the near-visible. In principle, any optical frequency could be measured by successive operations of the optical bisector. Optical frequency intervals as large as 8 THz have been measured with this method, but even though it is feasible, no bisection system has been built to date that connects the optical region all the way to countable microwave frequencies. The main limitation is the same as that facing the multiplication scheme: it still requires lots of stabilized laser sources to divide by factors of 2 from 10^{15} to 5×10^{10} Hz. The problem can be reduced somewhat by taking advantage of an optical-frequency comb generator based on an electro-optic modulator inside a resonant optical cavity. These systems, developed by Korougi and others, generate many coherent microwave sidebands on a laser so that optical frequency intervals on the order of 5 THz can be measured.

Even combining all of these ideas and techniques, a convenient system to count optical frequencies remained elusive. This has now changed with the revolutionary new optical frequency combs based on mode-locked femtosecond lasers.

Optical Frequency Metrology with Femtosecond Combs

Recently, the introduction of mode-locked lasers into the field of absolute optical frequency metrology has resulted in an important advance. The extremely complex frequency chains described above can be replaced by a single mode-locked laser if it produces pulses with sufficiently large bandwidth. Although the underlying ideas are not new (they were originally discussed by Hänsch and Chebotayaev in the 1970s) it is only recently that mode-locked lasers with the required characteristics were developed and the concept carefully tested.

The dramatic simplification of a frequency chain that uses a mode-locked laser means that absolute optical-frequency measurements can now be made by a single person as compared to the team of approximately 10 highly trained scientists required to run the previous chains. This has resulted in a large number of measurements being reported in the last 2 years. The simplicity also allows longer averaging times and a greater number of measurements to be used in determining the final reported frequency. In addition, since a simpler system has fewer sources of error, it has improved the quality of measurements made in a given amount of time.

Mode-locked lasers produce ultrashort pulses of light. For a typical modern high-quality mode-locked laser, the pulse duration is around 10 fs (1 fs $= 10^{-15}$ second), corresponding to three or four optical cycles in the near-infrared, where these lasers typically operate. The current best mode-locked lasers can produce pulses that are shorter than two optical cycles in duration (5–6 fs). Since the spectrum of these lasers is complicated, the bandwidth is significantly larger than it would be for a smooth pulse such as one with a Gaussian or hyperbolic-secant temporal intensity profile.

At first sight, using mode-locked lasers for optical-frequency metrology seems counter-intuitive, because optical-frequency metrology clearly requires very well-defined frequencies in contrast to the broad frequency spectrum of ultrashort pulses. This apparent conundrum is resolved by the fact that a mode-locked laser actually produces a very regularly spaced train of ultrashort pulses. The ultrashort pulses emitted by a mode-locked laser are replicas of a steady-state pulse that circulates inside the cavity of the laser. Every time the internal pulse impinges on the output coupler, which is a partially reflective mirror, a portion of it is transmitted, producing the output. The timing between these output pulses is determined purely by the time, τ_c, for the intracavity pulse to make one round trip, which is typically between 1 and 20 ns. The spectrum of a train of pulses is a regularly spaced 'comb' of frequencies spaced by the repetition rate of the train, $f_{rep} = 1/\tau_c$. If all of the pulses are identical, then these comb lines are just integer multiples of f_{rep}. However, due to dispersion, the group and phase velocities are different inside the cavity of a mode-locked laser. This leads to a pulse-to-pulse shift in the phase between the carrier and envelope (which we call the carrier–envelope phase; ϕ_{ce}, the pulse-to-pulse shift of ϕ_{ce} is designated by $\Delta\phi_{ce}$). The presence of $\Delta\phi_{ce}$ results in a rigid shift of all the comb lines by an amount $\delta = 2\pi\Delta\phi_{ce}/f_{rep}$. Thus the frequencies of the comb lines are given by $\nu_n = nf_{rep} + \delta$, where n is a large integer of order 10^6. The essential point of this equation is that it gives optical frequencies in terms of RF frequencies (f_{rep} and δ) that can easily be measured with conventional electronics and

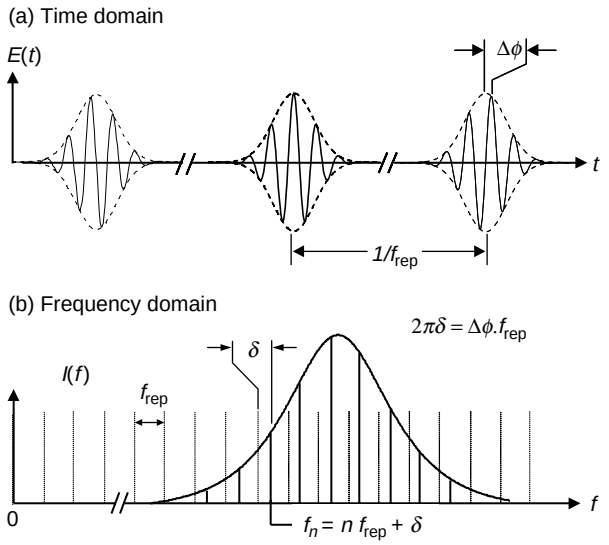


Figure 4 Time–frequency correspondence and relationship between $\Delta\phi$ and δ. (a) In the time domain, the relative phase between the carrier and the envelope evolves from pulse to pulse by the amount $\Delta\phi$. (b) In the frequency domain, the elements of the frequency comb of a mode-locked pulse train are spaced by f_{rep}. The entire comb is offset from integer multiples of f_{rep} by an offset frequency $\delta = \Delta\phi f_{rep}/2\pi$.

compared to the microwave cesium frequency standard. The relationship between the time and frequency domains is shown in Figure 4.

Given such a comb of frequencies, a heterodyne measurement readily yields the frequency difference between the unknown optical frequency of a narrow-band source (typically a single frequency laser) and the optical frequency of a nearby comb line. Thus, the absolute unknown frequency can be determined if the absolute frequencies of the comb lines are known. This requires measurement of f_{rep} and δ. The repetition rate, f_{rep}, is easily measured with a fast photodiode. Measurement of δ requires somewhat greater effort.

The development of a 'self-referencing' method to easily measure δ is the key enabling breakthrough that has brought about the recent revolution in optical-frequency metrology. It is called self-referencing because it provides a direct measure of δ with no other optical frequencies as input. It works by comparing frequencies that differ by a factor of two (typically these lie in the wings of the spectrum) using second-harmonic generation. Specifically, the frequency difference between the second harmonic of a comb line n and the comb line $2n$ is given by $2f_n - f_{2n} = 2(nf_{rep} - \delta) - (2nf_{rep} - \delta) = \delta$. Such a frequency difference is easily measured using a heterodyne beat. This technique is shown schematically in Figure 5. Other variations on this basic idea are possible; for example, rather than frequency doubling

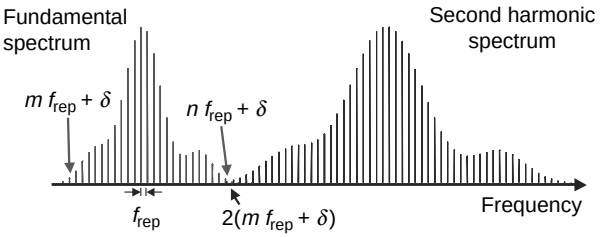


Figure 5 Schematic of the self-referencing technique for determining the offset frequency of the frequency comb produced by a femtosecond mode-locked laser. The frequency difference between the fundamental comb and its second harmonic is the offset frequency. In the region where overlap occurs, this is easily detected as a heterodyne beat.

the comb itself, it is also possible to extract the same information by comparing the comb lines to a single-frequency laser and its second harmonic.

Implementation of the self-referencing technique requires a pulse spectrum that spans a factor of two in frequency (an optical octave). Although a laser has recently been demonstrated that produces such a broad spectrum, they are not commonly available. However, the development of microstructured fiber has made it easy to broaden the spectrum of ordinary mode-locked lasers so that it spans an octave, as was first demonstrated by Ranka and co-workers at Bell Labs. The broadening occurs because optical nonlinearity in the fiber results in self-phase modulation of the laser pulse. Microstructure fiber achieves guiding by surrounding the core region with microscopic air holes, as compared to regular fiber that uses doping to produce a difference in the index of refraction. The much larger difference in the index of refraction in microstructure fiber allows a smaller core region to be used, which increases the effective nonlinearity. It also makes it possible to modify the dispersion such that the group velocity dispersion goes through zero for light with a wavelength close to 800 nm. This is crucial because group-velocity dispersion in ordinary fiber causes ultrashort pulses to temporally spread, which lowers the peak power and hence nonlinearity. It is possible to implement self-referencing techniques using less bandwidth if higher orders of nonlinearity are used; for example, if the difference between the second harmonic of the high-frequency end and third harmonic of the low-frequency end of the comb are used, only a half-octave bandwidth is required.

Given the ability to measure both f_{rep} and δ, and as long as its frequency lies within the comb spectrum, it is possible to measure the optical frequency of any optical source by measuring the heterodyne beat

with a nearby comb line. Thus, an absolute optical frequency measurement comes down to measuring three RF frequencies. It is often advantageous to actively control either one or both of f_{rep} and δ to hold them to a fixed frequency. This requires real-time adjustment of some parameters of the mode-locked laser to affect the necessary feedback. Controlling f_{rep} is straightforward; it is achieved by simply changing the cavity length of the laser, typically by mounting a mirror on a piezoelectric transducer. Controlling δ is more challenging, as it requires a differential change between the cavity group delay and phase delay. This has been achieved by adjusting the pump power, which in turn controls the energy of the intracavity pulse, and/or by tilting the end mirror of the cavity in lasers that employ prism dispersion compensators. Since the optical spectrum is spread out across this mirror, tilting it can be thought of as a frequency-dependent change in the cavity length, with the optical frequency corresponding to the pivot point of the tilt experiencing no length change, while those on opposite sides of the pivot point see changes with opposite signs.

In addition to the three RF frequencies, it is necessary to know the large integer n that describes the comb line that is beat against the unknown source. If a priori knowledge about the unknown source is available with precision somewhat better than $f_{rep}/2$ or better, n can readily be determined. For typical mode-locked lasers, standard commercial wavemeters can provide this level of precision. In the absence of additional outside information, n can also be determined with some effort and precision measurements by changing f_{rep} and δ in a systematic fashion. In the end, the simple formula $f_n = nf_{rep} + \delta$ gives the optical frequency of any arbitrary nth mode of the femtosecond optical comb. With this established, the frequency of any single-mode laser within the octave span of the comb can be determined simply by measuring the frequency of the beat-note generated in a photodiode between the cw laser and the known frequencies of the comb.

The diagram in Figure 6 shows how a femtosecond comb acts as an optical frequency synthesizer, which provides frequencies that are known relative to the input frequency (provided by a hydrogen maser in the figure). A typical series of measurements is shown in Figure 7 for the NIST Hg^+ single ion standard. The inset shows the results of numerous measurements for the frequency of Hg^+ relative to the NIST primary cesium standard.

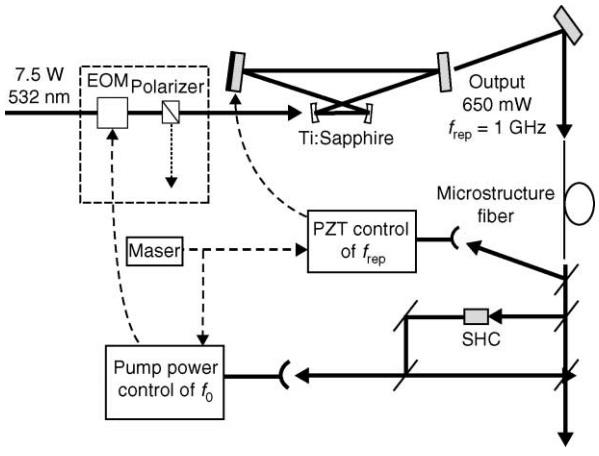


Figure 6 Schematic example of how the femtosecond comb is configured to make optical frequency measurements relative to the reference frequency provided by a hydrogen maser (Maser) that is in turn calibrated to a cesium atomic frequency standard. The repetition rate of the pulses from the Ti:Sapphire laser are detected in a photodiode and are controlled using a piezoelectric transducer (PZT) that changes the length of the laser cavity. The microstructure fiber and second harmonic crystal (SHC), provide the 'self-referencing' beat-note signal that is used to control the offset frequency, δ, by changing the power of the pump laser with an electro-optic modulator (EOM). This 'self-referenced' system then provides a set of modes with known frequencies spanning the visible. An unknown laser frequency can then be measured by making a beat-note with one of the modes of the comb.

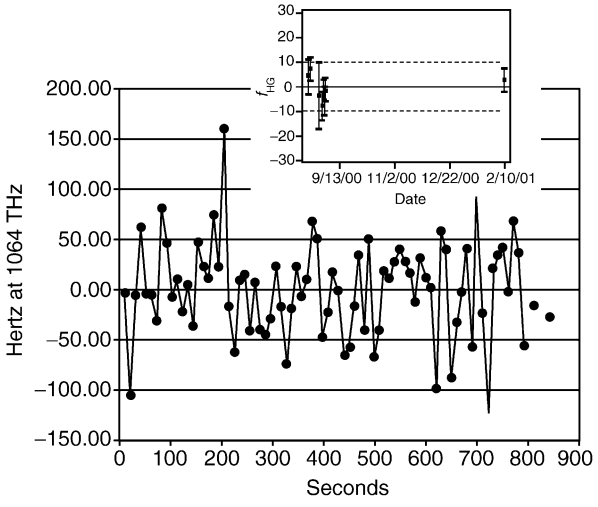


Figure 7 Actual frequency measurements of the beat-note between a hydrogen maser stabilized optical frequency comb (as shown in Figure 6) and the NIST Hg^+ optical frequency standard. The scatter in the data for counter gate time of 10 seconds corresponds to a fractional frequency uncertainty of about 4×10^{-14}, consistent within the instability in the maser. The inset shows the results from numerous datasets of measurement of the frequency of the Hg^+ optical standard relative to the cesium primary standard at NIST. The frequency data are plotted relative to 1064 721 609 899 142.6 Hz. The dotted lines give estimates of systematic uncertainties in lieu of a complete accuracy evaluation.

Table 1 Summary of optical frequency measurements of a selection of molecular, atomic, and ionic transitions. Recent measurements listed for all and additional measurements are given for a few to indicate the level of agreement

Atom/molecule/ion	*Wavelength (nm)*	*Frequency (THz)*	*Uncertainty*		*Year*	*First author and institution*
			± Hz	*Frac.*		
OsO_4	10318.0	29.054 057 446 660	50	2×10^{-12}	1985	Clarion, LPTF
		29.054 057 446 579	10	3×10^{-13}	1999	Ducos, LPTF
CH_4	3392.2	88.376 181 627 000	50 000	6×10^{-10}	1972	Evenson, NBS
		88.373 149 028 553	200	2×10^{-11}	1998	Ering, PTB
Rb−2 photon	778.11	385.285 142 367 000	8000	2×10^{-11}	1993	Nez, ENS
		385.285 142 374 800	3000	8×10^{-12}	2000	Diddams, JILA
Sr^+	674.03	444.779 904 409 540	200	4×10^{-13}	1999	Bernard, NRC
Ca	657.45	455.986 240 495 150	8	10×10^{-14}	2003	Helmcke, PTB
		455.986 240 494 158	26	6×10^{-14}	2000	Udem, NIST
I_2 a_{15} of $R(127)$ 11-5	632.99	473.612 340 492 000	74 000	2×10^{-10}	1983	Jennings, NBS
I_2 a_{16} of $R(127)$ 11-5	632.99	473.612 353 604 800	1200	3×10^{-12}	2000	Ye, JILA
I_2 a_{10} of $R(56)$ 32-0	532.24	563.260 223 480 000	70 000	1×10^{-10}	1995	Jungner, JILA
		563.260 223 514 000	5000	9×10^{-12}	2000	Diddams, JILA
Yb^+	435.51	688.358 979 230 931	6	9×10^{-15}	2001	Stenger, PTB
H 2*S*-8*S*	389.01	770.649 561 581 100	5900	8×10^{-12}	1997	de Bouroir, ENS
Hg^+	281.57	1064.721 609 899 140	10	9×10^{-15}	2000	Udem, NIST
In^+	236.54	1267.402 452 899 920	230	2×10^{-13}	2000	von Zanthier, MPQ
H 1*S*-2*S*	121.57	2466.061 413 187 100	46	2×10^{-14}	2000	Holzwarth, MPQ

Summary of Measurements/Standards

Starting with the first measurement of the frequency of the methane-stabilized HeNe laser in 1972, a new laser frequency measurement was reported every year or so using harmonic frequency chains. These are summarized in Table 1. An indication of the history of frequency measurements of stable laser references is given in Figure 8. The accuracy, in terms of fractional frequency uncertainty, improved from 10^{-10} in 1972 to 10^{-12} in 1999, at which time the femtosecond optical frequency combs came on line and there was a dramatic improvement in the precision as well as the number of measurements that were completed. It is impossible to predict with certainty but some have projected that optical standards can ultimately reach uncertainties as small as 10^{-18}.

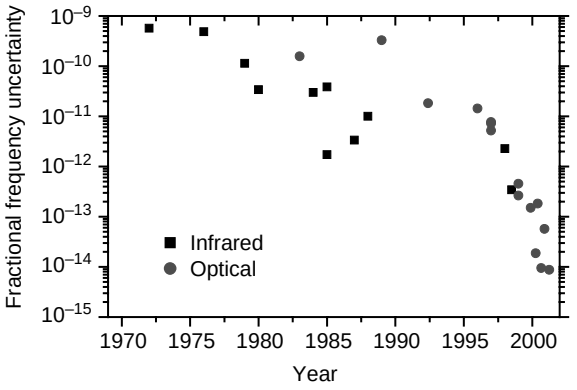


Figure 8 Progress in the accuracy of optical frequency measurements.

Outlook

Three separate technologies have now reached a level of maturity that it is possible to build high-performance optical-frequency standards and clocks. The essential achievements are: laser cooling and trapping of atoms (first proposed by Wineland and Dehmelt, and Hänsch and Schawlow), highly stabilized narrow-linewidth cw lasers, and femtosecond optical frequency combs. Combining these key ingredients, we can construct an optical atomic clock as shown schematically in Figure 9.

Here, for comparison with Figure 1, the cold atoms or single ion provide the narrow atomic resonance, the cw laser serves as the local oscillator to probe the resonance, and the femto-comb serves as the counter. Optical frequency standards of the future are expected to provide orders of magnitude better stability and improved accuracy over the existing atomic frequency standards that now use microwave transitions in atoms. As described in the previous section, frequency combs produced by femtosecond lasers can directly measure the frequency of a stable laser locked to an atomic transition relative to a known microwave frequency standard. This gives fundamental information about the atomic energy levels, and structure, and allows comparisons between different elements. However, to take

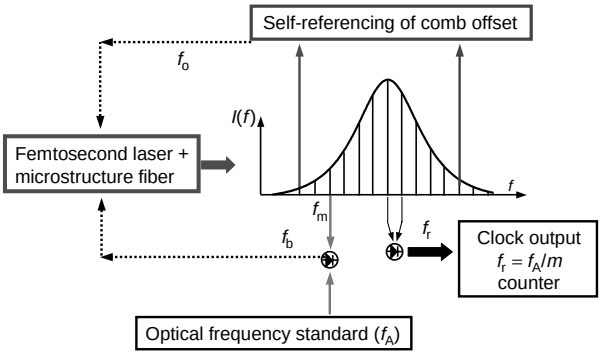


Figure 9 Simplified schematic of an optical clock.

advantage of the high stability of the optical references we run the system as an optical clock (Figure 9) where the stable laser and the femto-comb are locked to the atomic resonance, and the clock output comes as pulses at the repetition rate (e.g., 1 GHz) of the femtosecond mode-locked laser. With a judicious choice of control parameters it is possible to have the pulse repetition frequency, i.e., the clock output, at an exact subharmonic of the optical transition frequency. It is intriguing to note that a portable optical clock could measure time and length at the same time.

See also

Instrumentation: Spectrometers. **Quantum Electrodynamics:** Quantum Theory of the Electromagnetic Field.

Further Reading

Barnett JE (1998) Time's Pendulum. New York: Plenum.

Bergquist JC, Jefferts SR and Wineland DJ (2001) Time measurement at the millennium. *Physics Today* March: 37–42.

Cundiff ST, Ye J and Hall JL (2001) Optical frequency synthesis based on mode-locked lasers. *Reviews of Scientific Instruments* 72: 3749–3771.

Diddams SA, Jones DJ, Ye J, *et al.* (2000) Towards the ultimate control of light. *Optics and Photonics News* October: 16–22.

Diddams SA, Udem T, Bergquist JC, *et al.* (2001) An optical clock based on a single trapped $^{199}Hg^{+}$ ion. *Science* 293: 825–828.

Diels JC and Rudolph W (1996) Ultrashort Laser Pulse Phenomena. San Diego, CA: Academic Press.

Hall JL (2000) Optical frequency measurement: 40 years of technology revolutions. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1136–1144.

Hall JL and Ye J (2001) A new era of frequency standards and optical frequency measurements. *Optics and Photonics News* February: 45–50.

Hänsch TW and Walther H (1999) Laser spectroscopy and quantum optics. *Reviews of Modern Physics* 71: S242–S252.

Hollberg L, Diddams SA, Curtis EA, Oates CW and Fox RW (2001) Optical frequency standards for clocks of the future. In: The International Society for Optical Engineering. *Harnessing Light: Optical Science and Metrology at NIST*. Bellingham, DC. *Proceedings of SPIE* 4450: 54–59.

Jennings DA, Evenson KE and Knight DJE (1986) Optical frequency measurements. *Proceedings of the IEEE* 74: 168–179.

Jespersen J and Fitz-Randolph J (1977) From Sundials to Atomic Clocks. New York: Dover.

Jones T (2000) Splitting the Second: The Story of Atomic Time. Bristol, UK: Institute of Physics.

Luiten AN (2001) Frequency Measurement and Control: Advanced Techniques and Future Trends. Berlin: Springer-Verlag.

Menoni CS (ed.) (2001) Feature section on optical frequency synthesis: a new tool for precision optical frequency metrology. *IEEE Journal of Quantum Electronics* 37: 1481–1513.

Quinn TJ (1999) Practical realization of the definition of the metre. *Metrologia* 36: 211–244.

Fourier Transform Spectroscopy

T Fromherz, Johannes Kepler Universität, Linz, Austria

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Fourier transform spectroscopy (FTS) has emerged as one of the most powerful spectroscopic techniques since the first instruments became commercially available in the late 1960s. With this method, high spectral resolution, high wavenumber accuracy, broad spectral range, high optical throughput, and high signal to noise ratio can be achieved simultaneously. Because of these combined advantages that will be discussed in more detail throughout this article, especially in the far- (10–500 cm^{-1}) and mid-infrared (FIR and MIR) spectral region (500–5000 cm^{-1}), FTS has become the method of choice for the most sensitive spectral investigations. In this article, an overview of the setup, the working principle, and the computation of spectra from the

measured interferograms will be given. Finally, by comparing FTS to dispersive spectroscopic methods, the conditions under which FTS can be exploited to advantage, will be discussed.

Spectrometer Setup

Almost all commercially available Fourier transform (FT) spectrometers are based on a Michelson interferometer. A typical layout of this interferometer is sketched in Figure 1. A collimated beam with spectral density $S(\sigma)$ entering the spectrometer is divided into two beams by a beamsplitter. Here, σ denotes the wavenumber that is related to the wavelength (λ) of the radiation by $\sigma = 1/\lambda$. As shown in Figure 1, one part of the radiation is reflected by the beamsplitter and propagates through the compensator plate to a fixed mirror. The mirror reflects the beam through the compensator plate back onto the beamsplitter. The other part of the radiation beam is transmitted through the beamsplitter and its substrate, and hits a moveable mirror from which it is reflected back onto the beamsplitter. At the beamsplitter, both beams are divided again. As shown in Figure 1, two beams, one propagating to the source, the other propagating to the detector, finally result. Both beams contain a contribution that was reflected by the fixed mirror and a second contribution that was reflected by the moveable mirror. In Figure 1, lines indicate the paths of the beams schematically. In the case of a reflection on the beamsplitter, the line type is changed in Figure 1, whereas for a transmission, the line type remains unchanged. The widths of the lines in Figure 1 indicate the size of the field amplitudes and, therefore, are reduced for each reflection/transmission on the beamsplitter. For clarity, the lines indicating the beams reflected into themselves by the fixed and moveable mirrors in Figure 1 are offset laterally.

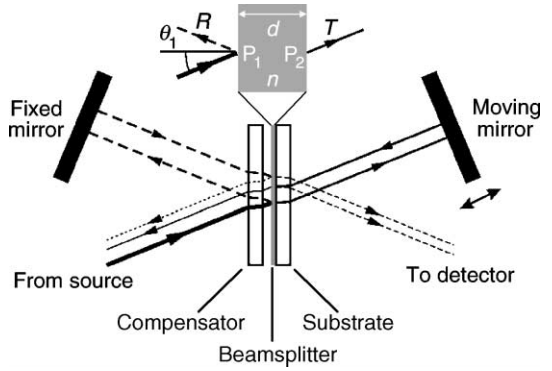


Figure 1 Schematic sketch of the layout of a Michelson interferometer. The paths of the rays are indicated by lines and arrows. At a reflection on the beamsplitter, the line type is changed, whereas for a transmission it remains unchanged. The widths of the lines indicate the size of the field amplitudes. For clarity, the lines indicating the beams reflected into themselves by the fixed and moveable mirrors, are offset laterally. The inset shows a magnified cross-section of the beamsplitter. In the inset, the geometrical details used for calculating the phase difference between R and T are defined.

By changing the position of the moveable mirror along the beam axis, a difference between the optical paths from the beamsplitter to the fixed and moveable mirror is generated. The compensator consists of the same material and has the same thickness as the beamsplitter substrate and, therefore, minimizes the influence of the dispersion of the beamsplitter substrate on the optical path difference.

Fourier Analysis and Interferometry

To calculate the intensities of the combined beams falling on the detector and on the source, we start by considering the phase difference between the reflected and transmitted beams leaving the beamsplitter. For simplicity, we ignore the compensator and beamsplitter substrate, since ideally, they generate the same phases in both arms of the interferometer, and therefore, do not contribute to the phase difference. Thus the beamsplitter is regarded as a freestanding, dielectric slab with thickness d. The complex reflection (R) and transmission (T) coefficients, including the effect of multiple internal reflections within the beamsplitter layer, are given by:

$$\begin{aligned} R &= r(1 - e^{i2\gamma})/(1 - r^2 e^{i2\gamma}) \\ T &= e^{i\gamma}(1 - r^2)/(1 - r^2 e^{i2\gamma}) \end{aligned} \quad [1]$$

Here, r is the Fresnel reflection coefficient that describes the ratio of the moduli and the phase difference between incoming and reflected electrical field for a electromagnetic wave incident under a angle Θ_1 on the surface of a dielectric with dielectric constant n, and $\gamma = 2\pi\sigma nd\cos(\Theta_2)$ is the phase difference evaluated at P_1 and P_2 in Figure 1 for a single pass of the electromagnetic wave through the dielectric slab under the angle Θ_2 (by Snellius law, $\sin(\Theta_1) = n\sin(\Theta_2)$). Since n is real, so is r and, therefore, the phase difference $\Delta\varphi$ between R and T is calculated from eqn [1] according to:

$$\begin{aligned} \frac{R}{T} &= \frac{|R|}{|T|} e^{i\Delta\varphi_{RT}} = \frac{r}{1 - r^2}(e^{-i\gamma} - e^{i\gamma}) \\ &= -2i\frac{r}{1 - r^2}\sin(\gamma) \end{aligned} \quad [2]$$

i.e., $\Delta\varphi_{RT} = -\pi/2$. With this result, we are able to calculate the phase difference between the two rays falling on the detector and the two rays falling back

on the source. The geometry considered in the following is sketched in Figure 1. The coefficients R and T connect the phases and magnitudes of reflected and transmitted electric field amplitudes at the points P_1 and P_2. From these points, the radiation beams travel back and forth to the fixed and moveable mirrors, thereby accumulating phases according to: $\varphi_f = 2\pi 2x_f\sigma$ and $\varphi_m = 2\pi 2(x_f + \Delta x)\sigma$, where we have defined $\Delta x = x_m - x_f$ and x_m, x_f denote the distances of moveable and fixed mirrors from the points P_1 and P_2. The two planewaves falling on the detector interact twice with the beamsplitter. Each planewave is once reflected at and once transmitted through the beamsplitter. Therefore, the beamsplitter adds no additional phase difference to the two planewaves and the total phase difference is given by $\Delta\varphi = \varphi_m - \varphi_f$. The situation is different for the two beams falling back on the source. Here, one beam is reflected at both interactions with the beamsplitter, whereas the other is transmitted at both interactions. According to eqn [2], between these two beams an additional phase shift of π results. Quantitatively, the total field amplitude at the detector, neglecting an arbitrary phase, is given by $E_D = E_0RT(e^{i\varphi_m} + e^{i\varphi_f})$, and the corresponding intensity is:

$$\begin{aligned} I_D \propto E_D E_D^* &= 2|E_0|^2|R|^2|T|^2\{1 + \mathrm{Re}[e^{i(\varphi_m - \varphi_f)}]\} \\ &= 2|E_0|^2|R|^2|T|^2\{1 + \cos(4\pi\sigma\Delta x)\} \end{aligned} \quad [3]$$

The sum of the field amplitudes falling back on the source is $E_S = E_0(R^2e^{i\varphi_m} + T^2e^{i\varphi_f})$. For the intensity one gets:

$$I_S \propto E_S E_S^* = |E_0|^2\{|R|^4 + |T|^4 + 2\mathrm{Re}[R^2T^{*2}e^{i(\varphi_m - \varphi_f)}]\} \quad [4]$$

Denoting the complex coefficients R and T by their modulus and their phase $R = |R|e^{i\varphi_R}$, $T = |T|e^{i\varphi_T}$ and using eqn [2], one gets:

$$I_S \propto |E_0|^2\{|R|^4 + |T|^4 + 2|R|^2|T|^2\cos(4\pi\sigma\Delta x + \pi)\} \quad [5]$$

Both intensities, I_D and I_S contain a constant component and a component that varies with $\cos(4\pi\sigma\Delta x)$. The varying parts have the same amplitude $2|E_0|^2|R|^2|T^2|$ but are out of phase by π. Consequently, a maximum in I_D corresponds to a minimum in I_S and the sum of both intensities is independent of the mirror displacement Δx.

In most of the commercially available spectrometers, only I_D is used for spectral measurements. Usually, the detector signal is amplified by an AC coupled electronics. Therefore, in a measurement of $I_D(\Delta x)$, the constant part of the signal is suppressed. Assuming a spectral density $S(\sigma)$ of the source and an ideal detector with a responsivity R_D linear in the intensity and independent of the wavenumber σ, the detector signal $D(\Delta x)$ is given by:

$$D(\Delta x) = R_D\int_0^\infty S(\sigma)\cos(4\pi\sigma\Delta x)\mathrm{d}\sigma \quad [6]$$

The information about the spectrum is contained in $S(\sigma)$ and can be extracted from the interferogram $D(\Delta x)$ by a cosine Fourier transformation. The transformation has to be done numerically and, therefore, it is more convenient to use a complex, fast Fourier transform algorithm. For that purpose, eqn [6] is modified by using the complex expansion of the cosine $\cos(x) = (e^{ix} + e^{-ix})/2$:

$$\begin{aligned} D(\Delta x) &= \frac{R_D}{2}\left(\int_0^\infty S(\sigma)e^{i4\pi\sigma\Delta x}\mathrm{d}\sigma + \int_0^\infty S(\sigma)e^{-i4\pi\sigma\Delta x}\mathrm{d}\sigma\right) \\ &= \frac{R_D}{2}\left(\int_0^\infty S(\sigma)e^{i4\pi\sigma\Delta x}\mathrm{d}\sigma + \int_{-\infty}^0 S(-\sigma)e^{i4\pi\sigma\Delta x}\mathrm{d}\sigma\right) \end{aligned} \quad [7]$$

Extending $S(\sigma)$ to negative wavenumbers according to:

$$S'(\sigma) = \begin{cases} S(\sigma) & \text{for } \sigma \geq 0 \\ S(-\sigma) & \text{for } \sigma < 0 \end{cases} \quad [8]$$

leads to the following expression for the interferogram:

$$D(\Delta x) = \frac{R_D}{2}\int_{-\infty}^\infty S'(\sigma)e^{i4\pi\sigma\Delta x}\mathrm{d}\sigma \quad [9]$$

and the spectral density $S'(s)$ can be obtained by the complex Fourier transform according to:

$$S'(\sigma) = \frac{4}{R_D}\int_{-\infty}^\infty D(\Delta x)e^{-i4\pi\sigma\Delta x}\mathrm{d}(\Delta x) \quad [10]$$

In the following, we will skip the prime and S will refer to the spectral density extended to negative values of σ.

To calculate the spectral density according to eqn [10], one has to use an interferogram extending from $-\infty$ to $+\infty$ measured at infinitely small sampling distances $\mathrm{d}(\Delta x)$. In practise, of course, the interferogram is measured only between finite mirror displacements $-\Delta x_{max}$ and Δx_{max}, at sampling points equally spaced by ξ.

The effect of the finite mirror displacement on the resulting spectral density S can be included into eqn [10] by multiplying the interferogram $D(\Delta x)$ with a function that vanishes for $|\Delta x| > \Delta x_{max}$. As an example, we use the boxcar function

$\theta(\Delta x + \Delta x_{max})\theta(\Delta x_{max} - \Delta x)$, where $\theta(x)$ is the Heaviside step function with values 0 and 1 for $x < 0$ and $x > 0$, respectively, and calculate the output spectrum for a spectral density $S_1(\sigma)$ consisting of a single line at wavenumber $\sigma_0 : S_1(\sigma) = (\delta(\sigma - \sigma_0) + \delta(\sigma + \sigma_0))/2$ (note that the spectral density is extended to negative wavenumbers according to the definitions given above). According to eqn [9], the interferogram is $D_1(\Delta x) = (R_D/2)\cos(4\pi\sigma_0\Delta x)$. Multiplying the interferogram with the boxcar function and Fourier transforming the product back using eqn [10], the resulting spectral density S_1^{box} is:

$$S_1^{box} = 2\Delta x_{max}\left\{\frac{\sin[4\pi(\sigma_0 - \sigma)\Delta x_{max}]}{4\pi(\sigma_0 - \sigma)\Delta x_{max}} + \frac{\sin[4\pi(\sigma_0 + \sigma)\Delta x_{max}]}{4\pi(\sigma_0 + \sigma)\Delta x_{max}}\right\} \quad [11]$$

Therefore, the approximation to a monochromatic spectrum is a sum of $\mathrm{sinc}(z_\pm)$ $(= \sin(z_\pm)/z_\pm)$ functions where $z_\pm = 4\pi(\sigma_0 \pm \sigma)\Delta x_{max}$. The lineshape function S_1^{box} is plotted around σ_0 in Figure 2 by the full line. Compared to the monochromatic spectrum S_1, the lines of S_1^{box} have a finite linewidth with a fullwidth at half maximum (FWHM) of approximately $0.3/\Delta x_{max}$. As a general rule in spectroscopy, two spectral lines of equal intensity and linewidth have to be separated at least by their FWHM, in order to be clearly distinguishable. According to this definition, the limit of resolution of a FT spectrometer $\Delta\sigma_{min}$ is inversely proportional to the maximum mirror displacement and given by $\Delta\sigma_{min} = 0.3/\Delta x_{max}$.

A less precise, but more intuitive definition of the resolution refers to the difference of the phases of the interferograms corresponding to two monochromatic lines separated by $\Delta\sigma$. The two lines are defined to be resolvable if the phases of their interferograms differ by more than 2π between the mirror positions $-\Delta x_{max}$ and Δx_{max}, i.e., the beating in the superposition of the interferograms shows at least one period. It is easily seen that, according to this definition $\Delta\sigma_{min} = 1/(2\Delta x_{max})$, a result in good agreement but slightly larger than that obtained by the more precise definition given above.

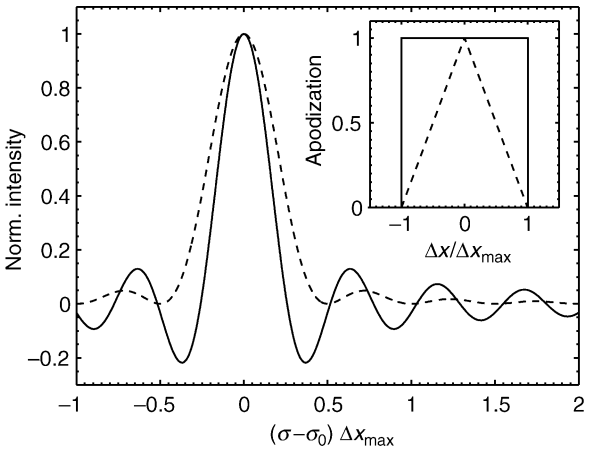


Figure 2 Normalized lineshape functions for boxcar (full line) and triangular apodization (broken line). In the inset, the corresponding apodization functions are shown.

As shown in Figure 2 by the full line, the sinc function has negative side-lobes with magnitudes of up to 22% of the maximum value. For some spectroscopic applications this side lobes are unwanted since they lead to a suppression of smaller peaks separated by approximately the resolution limit of the spectrometer. Thus, in these cases, it appears to be reasonable to multiply the interferogram by a (apodization) function different from the natural boxcar function. Usually, triangular or trapezoidal functions vanishing for $|\Delta x| > \Delta x_{max}$ are used. By Fourier transformation of these functions, it can be shown that no negative side lobes are generated. However, these functions have a larger (up to ~50%) peak FWHM than the sinc function. For comparison, in Figure 2, also the lineshape of a monochromatic spectrum using a triangular apodization function $A(\Delta x) = \theta(\Delta x + \Delta x_{max})\theta(\Delta x_{max} - \Delta x)(1 - |\Delta x/\Delta x_{max}|)$ is shown by the broken line. In this case, the negative sidelobes are strongly suppressed but the FWHM is increased.

Generally, the output spectrum $S^A(\sigma)$ corresponding to an arbitrary spectral density $S(\sigma)$ and an apodization function $A(\Delta x)$ is given by the convolution $S^A(\sigma) = 2[S \otimes A^{FT}](\sigma)$, where the lineshape function A^{FT} is the Fourier transform of $A(\Delta x)$:

$$A^{FT}(\sigma) = \int_{-\infty}^{\infty} A(\Delta x)e^{-4\pi\sigma\Delta x}d(\Delta x) \quad [12]$$

and

$$[f \otimes g](t) \equiv \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau \quad [13]$$

Discrete Sampling

In order to discuss the effect of the discrete, equally spaced sampling points with a mutual distance of ξ on the output spectrum, Δx is replaced by $x_n = n\xi$ in eqn [10], where n is an integer and runs from $-N_{max}$ to N_{max} and $N_{max}\xi = \Delta x_{max}$. Instead of the continuous interferogram $D(\Delta x)$, a discrete interferogram:

$$D_n = \frac{R_D}{2}\int_{-\infty}^{\infty} S(\sigma)e^{i4\pi n\xi\sigma}d\sigma \quad [14]$$

results. From eqn [14], it is readily seen (by shifting the integration variable) that replacing $S(\sigma)$ with $S[\sigma + m/(2\xi)]$, where m is an integer, does not change the discrete interferogram D_n. Keeping in mind that according to our definition $S(-\sigma) = S(\sigma)$, it is clear that $S[-\sigma + m/(2\xi)]$ also leads to the same interferogram as $S(\sigma)$. Therefore, to be able to transform the interferogram to the spectral density $S(\sigma)$ for a wavenumber band $[0, \sigma_{max}]$ unambiguously, one has to make sure, by electronically and/or optical filtering, that no wavenumbers outside this range contribute to the interferogram and that the sampling interval ξ is small enough so that $\sigma_{max} < -\sigma_{max} + 1/(2\xi)$, i.e., $\sigma_{max} < 1/(4\xi)$. In Figure 3, a schematic sketch of a spectral density $S(\sigma)$, consisting of two spectral bands shown by the full and dashed lines, is plotted. For the spectral band shown by the full line, the sampling interval ξ is small enough so that in the interval $[0, 1/(4\xi)]$ no ambiguity arises. For the spectral band shown by the dashed line, the interferogram appears to be under-sampled: ξ is too large, and, therefore, the part of the spectral band at negative wavenumbers shifted by $+1/(2\xi)$ falls into the interval $[0, 1/(4\xi)]$ (shown by the dotted line in Figure 3). Thus, for the sampling interval ξ, the interferogram for the sum of spectral bands, shown by the full and the dashed lines in Figure 3, is the same as for the sum of the spectral bands shown by the full and dotted lines. Since the interferograms are the same, the two spectra are indistinguishable at sampling intervals ξ. This ambiguity of spectra for an under-sampled interferogram is called aliasing. Comparing the correct sampling interval $\xi < 1/(4\sigma_{max})$ with the period of the interferogram D_n corresponding to a spectrum consisting of a single line at σ_{max} ($D_n = R_D \cos(4\pi n\xi\sigma_{max})$), it follows that the interferogram has to be sampled more than two times per period in order to obtain an unambiguous result (Nyquist Theorem).

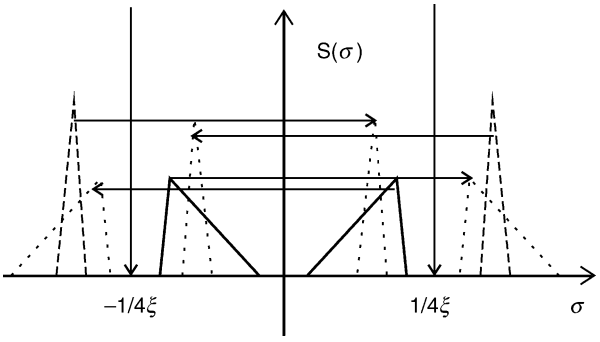


Figure 3 Schematic sketch of aliasing for under-sampled interferograms. For spectral density shown by the full line, the sampling interval ξ is chosen correctly and the $1/(2\xi)$ wavenumber ambiguity indicated by the arrows in the plot does not result in erroneous features in the spectral range $[0, 1/(4\xi)]$. The situation is different for the spectral band sketched by the dashed line: For this band, the sampling interval ξ is too large, and consequently, the $1/(2\xi)$ wavenumber ambiguity maps the line from the negative wavenumber region into the range $[0, 1/(4\xi)]$. The mapped spectral band is indicated by the dotted line.

Mathematically, the spectral density is calculated from the discrete interferogram D_n by a discrete Fourier transformation:

$$S_m = \frac{2}{R_D N} \sum_n^N D_n e^{-2\pi i n m/N} \qquad [15]$$

where N is the number of sampling points ($N = 2N_{max} + 1$) and the wavenumbers corresponding to S_m are given by $\sigma_m = m/2\xi N$.

In order to obtain a high accuracy in the sampling intervals ξ, usually a reference interferogram of a monochromatic laser line (for example from a HeNe laser) is measured simultaneously with the interferogram of the actual experiment. By counting, for example, the zeros with positive slope in the interferogram of the laser line, uniformly spaced sampling points can be achieved with high accuracy. If, in an experiment many interferograms have to be averaged, the absolute position of the sampling points is important, in order to ensure that only points of equal mirror displacement are averaged. Therefore, in some spectrometers, the interferogram of a broadband source is measured simultaneously with the interferograms of the laser and the experiment and the maximum of the interferogram of the broadband source is taken as the origin for counting the zeros of the laser interferogram.

Phase Correction

Ideally, the interferogram is expected to be symmetrically around the position of equal optical path in the two arms of the interferometer. However, in practise this is seldom the case. Many reasons can lead to an asymmetric interferogram, the most obvious being that none of the discrete sampling points of the interferogram coincides with the mirror position for equal optical path in the two interferometer arms. Let Δx_m be the measured mirror displacement that differs from the real mirror displacement Δx by x_0, i.e., $\Delta x = \Delta x_m + x_0$. In that case the measured interferogram is:

$$D(\Delta x_m) = \frac{R_D}{2} \int_{-\infty}^{\infty} S(\sigma) e^{i[4\pi\sigma(\Delta x_m + x_0)]} d\sigma \qquad [16]$$

and the Fourier transformation, according to eqn [10] with Δx_m as the integration variable, results in

$S(\sigma)e^{i4\pi x_0\sigma}$, that is, in a spectrum with both real and imaginary parts. Also, mirror or compensator misalignment can result in a wavenumber-dependent optical path at a fixed mirror position. Finally, for rapid-scan interferometers, in that the movable mirror is scanned at a constant velocity v, the electronic filtering and amplification of the detector signal will also add different phases, depending on the signal frequency ν that is correlated to the wavenumber σ by $\nu = 2\sigma v$. In general, as long as the phase φ depends only on the wavenumber and not on the mirror displacement, the measured interferogram is:

$$D(\Delta x_m) = \frac{R_D}{2}\int_{-\infty}^{\infty} S(\sigma)e^{i[4\pi\sigma\Delta x_m+\varphi(\sigma)]}d\sigma \qquad [17]$$

and by the Fourier transformation, according to eqn [10], one obtains $S(\sigma)e^{i\varphi(\sigma)}$. There are several methods implemented in the software of commercially available Fourier spectrometers to eliminate the phase from the Fourier transform. In most experimental cases, in that the measured interferograms correspond to spectra with $S(\sigma) \geq 0$ for all $\sigma \in [0, \sigma_{max}]$, these methods usually give the correct results. However, care has to be taken if spectra contain positive and negative components. This can be the case, for example, if the transmission of a sample is periodically modulated by externally applied voltage or a laser light and the detector signal is amplified by a lock-in technique before it is fed into the electronics of the Fourier transform spectrometer. If the modulation increases the sample absorbance in one part of the spectral range and decreases it in a different part, then the measured interferogram corresponds to a spectrum with both negative and positive regions. In the following, first the phase correction methods for spectra with positive values only are discussed before it is argued, why these methods might fail for spectra with positive and negative parts.

For spectra with only positive values, the most simple method to eliminate the unknown phase is to measure the interferogram symmetrically around $\Delta x_m = 0$ and calculate $S(\sigma)e^{i\varphi(\sigma)}$ according to eqn [10]. Including the effect of finite scan length, the output spectrum becomes $S^A = 2[S(\sigma)e^{i\varphi(\sigma)}]\otimes A(\sigma)^{FT}$. Under normal experimental conditions, the phase $\varphi(\sigma)$ does not vary over the peak FWHM of $A(\sigma)^{FT}$ and, therefore, the phase term can be put outside the integral of the convolution and $S^A = 2e^{i\varphi(\sigma)}[S(\sigma)\otimes A(\sigma)^{FT}]$ results. Since $A(\Delta x_m) \geq 0$, and chosen to be symmetric around $\Delta x_m = 0$, the convolution is real and positive and the true spectrum with instrumental resolution $2[S(\sigma)\otimes A(\sigma)^{FT}]$ can be obtained by calculating the modulus of S^A without any knowledge about $\varphi(\sigma)$. However, for several reasons this method for phase correction is not ideal. First of all, since the maximum spectral resolution of a Fourier transform spectrometer is inversely proportional to the maximum mirror displacement, for a given spectrometer layout with a fixed maximal mirror scan length L, the maximal achievable spectral resolution is reduced by a factor of two if scans symmetric around $\Delta x = 0$ have to be performed. Secondly, for signals closer to or smaller than the noise level of the spectrometer, the random noise will be entirely positive by taking the modulus. Hence, for small signals, the noise level will be increased relative to the signal level by taking the modulus.

Therefore, usually the 'Mertz phase correction method' is used. This method is based on the assumption that the phase $\varphi(\sigma)$ is a slowly varying function of σ. In this case, $\varphi(\sigma)$ can be determined with low spectral resolution, i.e., by a short, symmetric scan of length $2L_1$ around $\Delta x = 0$. It can be shown that from this scan the phase can be obtained according to:

$$\varphi(\sigma) = \arctan\left\{\frac{\mathrm{Im}[S(\sigma)e^{i\varphi(\sigma)}\otimes A^{FT}(\sigma)]}{\mathrm{Re}[S(\sigma)e^{i\varphi(\sigma)}\otimes A^{FT}(\sigma)]}\right\} + \pi \text{ (if denominator is negative)} \qquad [18]$$

Here, $S(\sigma)e^{i\varphi(\sigma)}\otimes A(\sigma)^{FT}$ results from the Fourier transformation of the short interferogram measured between $-L_1$ and L_1 and the apodization function $A(\Delta x_m)$ is adjusted to $-L_1$ and L_1. The addition of π for negative denominators is a consequence of the assumption that $S(\sigma)$ is positive everywhere. Under this assumption, a negative denominator in eqn [18] corresponds to a phase angle $|\varphi(\sigma)| > \pi/2$, that is mapped into $[-\pi/2, \pi/2]$ by the arctan-function. To get the correct phase angle, π has to be added to the phase angle, thereby extending the range of $\varphi(\sigma)$ to $[0, 2\pi]$.

With $\varphi(\sigma)$ known, the interferogram has to be recorded only for mirror displacements $\Delta x_m \geq 0$, since the information about the interferogram for $\Delta x_m < 0$ is contained in the phase spectrum and the interferogram for $\Delta x_m \geq 0$. Therefore, with this method, the totally available mirror displacement length can be used for increasing the spectral resolution.

However, for spectra containing positive and negative spectral bands, the Mertz phase correction usually fails, due to the addition of π if the denominator in eqn [18] becomes negative. For spectra with positive and negative bands, the sign of the denominator is determined by both $S(\sigma)$ and $\varphi(\sigma)$. Applying the Mertz algorithm in a negative region of

the spectrum with a vanishing physical phase angle, results in a calculated phase angle of $\varphi = \pi$. As a consequence, the negative spectral band appears as a positive band in the output spectrum calculated according to the Mertz algorithm.

To avoid erroneous phase-flips, the addition of π in eqn [18] for negative denominators can be omitted and the allowed values for $\varphi(\sigma)$ are reduced to $[-\pi/2, \pi/2]$. The corresponding phase correction method is usually referred to as the 'Mertz signed method'. However, it is immediately clear from the discussion above, that in this case any physical phase angle outside the range $[-\pi/2, \pi/2]$ results to an erroneous π-jump of the phase. According to eqn [16], a physical phase shift larger than $\pi/2$ can, for example, be the consequence of a large offset x_0, i.e., of an inaccurate alignment of Δx_m and Δx. At $\Delta x = 0$, all contributions to the interferogram add constructively and the height of the interferogram is proportional to the area under the spectral curve $S(\sigma)$, as can be readily seen by setting $\Delta x = 0$ in eqn [10]. Therefore, for $S(\sigma) > 0$ the position $\Delta x_m = 0$ is usually estimated by the global maximum of the measured interferogram. However, for spectra with both positive and negative bands, in most cases the interferogram does not exhibit a maximum at $\Delta x = 0$ (for positive and negative bands of equal height and width, the interferogram vanishes at $\Delta x = 0$), and estimating $\Delta x_m = 0$ by an extremum of the interferogram results, therefore, in a large offset x_0. Thus, in order to be able to use the 'Mertz signed' algorithm, $\Delta x_m = 0$ has to be determined separately in a calibration experiment in that the condition $S(\sigma) > 0$ is fulfilled. However, even for accurately determined $\Delta x_m = 0$, erroneous phase jumps can appear if lines of different intensity and opposite sign are spaced approximately equal to the resolution used for determining the phase spectrum $\varphi(\sigma)$.

In conclusion, there are no general strategies to completely exclude phase jumps in the Fourier transformation of interferograms corresponding to spectra with positive and negative bands. The experimentalist has to check from case to case if lines with strange shapes appear in the output spectra, and modify the phase correction accordingly. More elaborate strategies for phase correction, that are beyond the scope of this article, are reported in literature.

Example of Experimental Results

To illustrate the features of FTS discussed in the previous paragraphs by a practical example, in Figure 4 the results of a typical measurement in the MIR between 500 cm^{-1} and 5000 cm^{-1} are shown. The experiments were performed with a BRUKER IFS113 spectrometer using a silicon carbide glowbar as a source for the infrared radiation, a potassium bromide beamsplitter, and a liquid-nitrogen cooled mercury cadmium telluride detector. The spectrometer used for the experiments is set up in a vacuum chamber. The two interferograms shown in Figure 4a were measured at two different pressures (broken line: 10 mbar; full line: 300 mbar), i.e., at two different levels of absorption of the atmospheric gas in the interferometer chamber (for clarity, the interferograms are offset vertically). The abscissa axis is broken in Figure 4a and shows portions of the interferograms around zero mirror displacement and around the maximum mirror displacement of 2.5 cm (chosen for 0.2 cm^{-1} spectral resolution). For higher chamber pressures, the increased absorption of the atmospheric gas results in an increased amplitude of the wiggles of the interferogram apparent for large mirror displacements in Figure 4a. The amplitude of these wiggles is approximately three orders of magnitude smaller than the detector signal observed at the center-burst of the interferogram around zero mirror displacement. However, in the corresponding spectra (Figure 4b) calculated from the interferograms after phase correction, the absorption of the atmospheric gas is clearly evident by a strong increase of the dense and narrow absorption lines in the spectral regions 1200 cm^{-1}–2200 cm^{-1}, 2300 cm^{-1}–2400 cm^{-1}, and 3200 cm^{-1}–4000 cm^{-1} (note that in Figure 4b the spectra are offset vertically by 0.1 for clarity). The lines observed between 2300 cm^{-1}–2400 cm^{-1} are due to absorption of the CO_2 molecule and are shown in more detail in Figure 4c from 2335 cm^{-1} to 2345 cm^{-1}. In order to simulate the contribution of the CO_2 lines to the interferogram shown in Figure 4a, two neighboring lines in Figure 4c were fitted to a doublet of Lorentz-absorption lines (indicated by the symbols in Figure 4c). The corresponding interferogram was calculated by Fourier transformation of the Lorentz lines and is shown in Figure 4d. It consists of a rapid oscillating, cosine-like signal (see inset for an expanded plot between 0.6 and 0.602 cm) with an amplitude modulation corresponding to the difference of the resonance wavenumbers of the two Lorentz lines. The ordinate axes of Figures 4a,d have the same arbitrary units and, therefore, the amplitudes of the interferograms shown in these two panels can be directly compared. From Figure 4a, d we conclude that the two CO_2 lines contribute approximately with only one in 10^5 parts to the center-burst of the measured interferogram, i.e., theses lines absorb only a fraction of 10^{-5} of the integrated (over the wavenumbers) intensity of

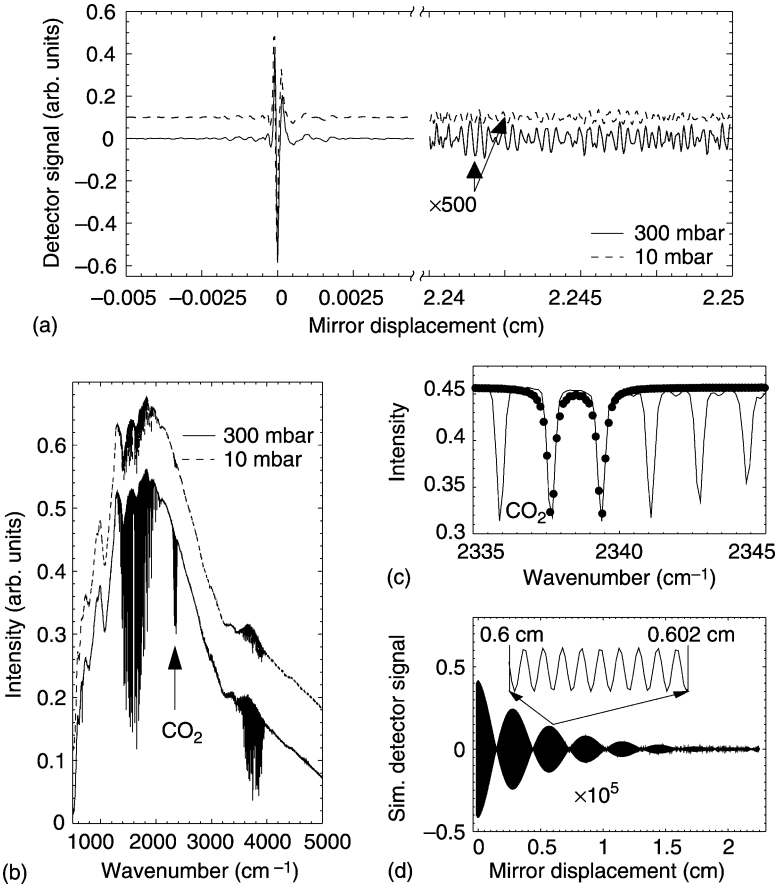


Figure 4 Measurement of the absorption of atmospheric gas for different pressures in the interferometer chamber (broken lines: 10 mbar, full lines: 300 mbar). (a) Interferograms (note the break in the axis of the mirror displacement). For clarity, the interferograms are offset vertically by 0.1 and 0.0002 on the left and right side of the axis break. On the right side of the axis break, the interferograms are amplified by a factor 500. (b) Spectra corresponding to the interferograms shown in (a). The spectra are offset vertically by 0.1 for clarity. (c) Part of the CO_2 absorption lines on an expanded wavenumber axis. The symbols indicate a fit of two Lorentz lines to neighbouring CO_2 absorption lines. The fit is used for calculating the interferogram shown in (d). (d) Simulated interferogram corresponding to the two CO_2 lines shown by the symbols in (c). The arbitrary units of the ordinate axis are the same as those used in (a) but the interferogram is amplified by a factor 10^5. In the inset, the interferogram is shown on an enlarged scale between 0.6 and 0.602 cm mirror displacement.

the incident radiation. Despite this small fraction, they are extremely well resolved with a large signal-to-noise ratio in Figures 4b,c, thus demonstrating the ability of FTS to measure spectra simultaneously with high resolution and over a large spectral range. This ability is one of the advantages (the so-called multiplex advantage) inherent in FTS and will be discussed in more detail in the following paragraph.

Advantage of FTS Over Dispersive Spectrometer

The two main advantages of FTS over dispersive spectroscopic methods, using, for example, a grating for the spectral decomposition of an incoming radiation, are directly connected with the working principle of FT spectrometers; high optical throughput and multiplexed measurement over the whole spectral range of interest. These advantages are discussed in the following, in more detail.

For a fair comparison of the optical throughput, spectrometers with equal spectral resolving power ρ have to be compared. Here, $\rho = \sigma_{max}/\Delta\sigma$, where σ_{max} is the maximum wavenumber analyzed and $\Delta\sigma$ is the spectral resolution. For an ideal Fourier transform spectrometer, the limit for σ_{max} depends on the divergence angle of the beams propagating in the spectrometer. As sketched in Figure 5, the divergence angle α is determined by the radius r_a of the entrance aperture and the focal length f of the collimating lens by $\alpha \approx r_a/f$. This angle causes a phase shift between the rays propagating parallel to and those propagating under the angle of α inclined to the optical axis

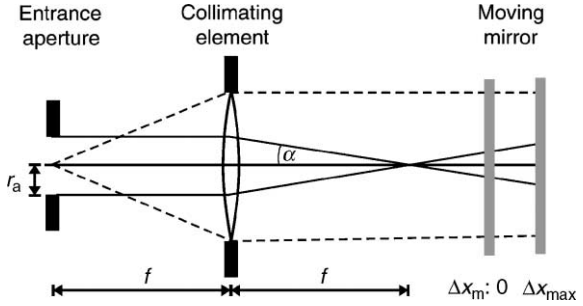


Figure 5 Entrance aperture and collimating element for an interferometer. For simplicity, the beamsplitter and fixed mirror are omitted in the sketch. The extreme rays entering the interferometer are indicated by the full and broken lines. Due to divergence α, a monochromatic light with wavenumber σ_0 produces all periods in the range $[4\pi\sigma_0, 4\pi\sigma_0/\cos(\alpha)]$ in the interferogram.

through the interferometer. The phase difference φ_{div} depends on the mirror displacement according to $\varphi_{div} = 4\pi\sigma\Delta x((1/\cos(\alpha)) - 1)$ and can be approximated for small α by $\varphi_{div} \approx 2\pi\sigma\Delta x\alpha^2$. It can be shown, that as long as φ_{div} does not exceed approximately π, the influence of the divergence on the interferogram can be tolerated, i.e., $\sigma_{max} \approx 1/(2\Delta x_{max}\alpha^2)$. Using $\Delta\sigma \approx 1/(2\Delta x_{max})$ for the resolution, one obtains for the resolving power of a FT spectrometer, $\rho = (f/r_a)^2$.

The optical throughput Θ describes the allowable energy per unit time that the system can let through and is related to area πr_a^2 of the aperture and the solid angle Ω of the collimating (or focusing) optics by $\Theta = \pi r_a^2\Omega$. Using $\Omega = A_{FTS}/f^2$, where A_{FTS} is the clear area of the collimating lens (for a collimating mirror, A_{FTS} is the area of the mirror projected on a plane perpendicular to the beam direction), results in $\Theta_{FTS} = \pi A_{FTS}r_a^2/f^2 = \pi A_{FTS}/\rho$. For a grating spectrometer, $\Theta_{grating} = hA_G/f\rho$. (see respective chapter of this encyclopedia) where h is the height of the slit, f the focal length of the collimating optics, and A_G the projected area of the grating. Generally, h/f is smaller than 1/20 in up-to-date grating spectrometers. Thus, for the same resolving power ρ and similar instrument size, FT spectrometer can offer a more than $\pi f/h \approx$ 60 times larger energy gathering capability. Therefore, for measurements with weak signals in that the detector noise is the dominant noise source, the large throughput of a FTS greatly increases the spectral signal to noise ratio that can be achieved for a fixed measuring time.

The other major principal advantage compared to dispersive methods relies on the simultaneous (multiplexed) gathering of information about all spectral bands $\Delta\sigma$ over a broad spectral range from σ_{min} and σ_{max}. For a grating spectrometer using a single detector element, the measurement time T_m available for recording the spectrum in that range has to be distributed between the $M = (\sigma_{max} - \sigma_{min})/\Delta\sigma$ spectral bands so that for a single spectral band the available integration time is T_m/M and the integrated signal increases proportional to T_m/M. If the dominant noise source is the detector noise which is independent of the signal level, then the noise will be proportional to $(T_m/M)^{1/2}$ and, therefore, the signal to noise ratio for the grating spectrometer $(S/N)_G$ is:

$$(S/N)_G \propto (T_m/M)^{1/2} \qquad [19]$$

For a FTS the situation is different, since it detects in the band $[\sigma_{min}, \sigma_{max}]$ all M small bands over the whole measurement time T_m via their contribution to the interferogram. So the integrated signal in a small band $\Delta\sigma$ is proportional to T_m. If the noise is again assumed to be random and independent of the signal level, for the FTS:

$$(S/N)_{FTS} \propto T_m^{1/2} \qquad [20]$$

results. For the same detector, assumed in the grating spectrometer and in the FTS, the proportionality constants are the same in eqns [19] and [20] and, therefore

$$\frac{(S/N)_{FTS}}{(S/N)_G} = M^{1/2} \qquad [21]$$

Equation [21] shows that the multiplex advantage becomes increasingly important for measurements of broad spectral bands with high resolution. In this case, an enhancement of the $(S/N)_{FTS}$ over $(S/N)_G$ by 2–3 orders of magnitude can be achieved. However, it has to be noted that the multiplex advantage of a FTS can be exploited only in cases where the detector noise is the dominant source of noise. In the visible and near infrared spectral region, low noise detectors, that allow single photon detection, are available. With these detectors, the spectral noise is dominated by the photon noise that is proportional to the square root of the intensity. It is evident that in this case the multiplexed detection is not an advantage, since for M spectral bands of width $\Delta\sigma$ contributing to the interferogram at a given mirror position, the noise is enhanced by $M^{1/2}$ compared to the noise that would be measured for a single band (here, it is assumed that all M bands contribute with the same intensity to the interferogram). Consequently, in this case, the noise enhancement by $M^{1/2}$ cancels the beneficial factor $M^{1/2}$ in eqn [21] and with single-detector grating and FT spectrometers, the same S/N ratio is achieved for a given measurement time T_m. In addition, using linear detector arrays with dispersive spectrometers, the multiplex advantage increases the S/N ratio for the dispersive spectrometer. However, detector arrays

and low-noise detectors are mainly available for spectral regions with $\sigma > 6000\ \mathrm{cm}^{-1}$, i.e., in the near-infrared and visible spectral region. In the FIR and MIR region ($\sigma < 6000\ \mathrm{cm}^{-1}$) in that such detectors are absent, the multiplexing inherent in FTS makes it the method of choice if low-noise, broadband, and high-resolution spectroscopic data are required.

List of Units and Nomenclature

Angle of incidence for radiation on beamsplitter	Θ_1
Angle of propagation of radiation in the beamsplitter layer	Θ_2
Beam divergence angle	α
Beam splitter thickness (cm)	d
Complex reflection coefficient	R, r
Complex transmission coefficient	T, t
Detector Signal normalized to detector area ($\mathrm{A\ cm}^{-2}$)	D
Difference between real and measured mirror displacement (cm)	x_0
Dirac delta-function	δ
Discrete interferogram ($\mathrm{A\ cm}^{-2}$)	D_n
Discrete spectral density ($\mathrm{W\ cm}^{-1}$)	S_m
Distance from beamsplitter to fixed mirror (cm)	x_f
Distance from beamsplitter to movable mirror (cm)	x_m
Electric field amplitude of incident radiation (V/m)	E_0
Electric field amplitude of radiation falling back on source (V/m)	E_S
Electric field amplitude of radiation reaching detector (V/m)	E_D
Focal length (cm)	f
Fourier transform of apodization function A (cm)	A^FT
General apodization function	A
Heavyside step function	θ
Intensity falling back on source ($\mathrm{W\ cm}^{-2}$)	I_S
Intensity on detector ($\mathrm{W\ cm}^{-2}$)	I_D
Maximum mirror displacement (cm)	Δx_max
Maximum wavenumber (cm^{-1})	σ_max
Measured mirror displacement (cm)	Δx_m
Measured single line spectral density for a finite scan length ($\mathrm{W\ cm}^{-1}$)	S_1^box
Measured single line spectral density for a finite scan length using the apodization function A ($\mathrm{W\ cm}^{-1}$)	S^A
Measurement time for complete spectrum (s)	T_m
Minimum wavenumber (cm^{-1})	σ_min
Mirror displacement	Δx
Mirror velocity ($\mathrm{cm\ s}^{-1}$)	v
Modulation frequency (s^{-1})	ν
Number of sampling points of the interferogram	N
Number of spectral bands	M
Optical throughput (cm^2)	Θ
Optical throughput for FTS (cm^2)	Θ_FTS
Optical throughput for grating spectrometer (cm^2)	Θ_grating
Phase accumulation due to optical path to fixed mirror	φ_f
Phase accumulation due to optical path to moveable mirror	φ_m
Phase angle	γ
Phase angle	φ
Phase difference	$\Delta\varphi$
Phase difference between beams reflected at and transmitted through beamsplitter	$\Delta\varphi_\mathrm{RT}$
Phase difference due to beam divergence	φ_div
Phase jump for reflection at beamsplitter	φ_R
Phase jump for transmission through beamsplitter	φ_T
Projected area of collimating element of FTS (cm^2)	A_FTS
Projected area of grating (cm^2)	A_G
Radius of entrance aperture (cm)	r_a
Refractive index	n
Responsivity ($\mathrm{A\ W}^{-1}$)	R_D
Sampling interval (cm)	ξ
Signal to noise ratio for FTS	$(\mathrm{S/N})_\mathrm{FTS}$
Signal to noise ratio for grating spectrometer	$(\mathrm{S/N})_\mathrm{G}$
Single line spectral density ($\mathrm{W\ cm}^{-1}$)	S_1
Solid angle	Ω
Spectral bandwidth (cm^{-1})	$\Delta\sigma$
Spectral density ($\mathrm{W\ cm}^{-1}$)	S
Spectral resolving power	ρ
Wave length (μm)	λ
Wave number (cm^{-1})	σ, σ_0
Wave number resolution (cm^{-1})	$\Delta\sigma_\mathrm{min}$

See also

Interferometry: Overview; Phase Measurement Interferometry.

Further Reading

Bell RJ (1972) *Introductory Fourier Transform Spectroscopy*. New York, London: Academic Press.

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.

Griffiths PR (1975) *Chemical Infrared Fourier Transform Spectroscopy*. New York, London, Sydney, Toronto: John Wiley & Sons.
Hecht E (1987) *Optics*. Reading, MA: Addison-Wesley.
Mertz L (1965) *Transformations in Optics*. New York, London, Sydney: John Wiley & Sons.
Saptari V (2004) *Fourier-Transform Spectroscopy Instrumentation Engineering*. Bellingham: SPIE Press.

Hadamard Spectroscopy and Imaging

R A DeVerse, Plain Sight Systems, Inc., Kailua-Kona, HI, USA

R M Hammaker, Kansas State University, Manhattan, KS, USA

W G Fateley, Plain Sight Systems, Inc., Manhattan, KS, USA

F B Geshwind and A C Coppi, Plain Sight Systems, Inc., Hamden, CT, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Consider the task of measuring the spectra of a source, the spectra of an array of contiguous sources, or simply the broadband irradiance of an array of sources. These measures are commonly known as spectrometry, spectral imaging, and imaging, respectively. The sensitivity at which spectrometric, spectral imaging, and imaging systems can be operated is often dominated by noise associated with the detection system and the ability of the optics employed to deliver sufficient photonic flux to the detector subsystem to overcome this noise limitation. The ideal case is where the optical system can be fashioned in such a way as to deliver greater signal to the detector in order to maximize the signal-to-noise ratio (SNR) for a given finite source irradiance and detector noise.

Let us consider the number of discrete spectral resolution elements of spectrometers, spatio-spectral resolution elements of spectral imagers, and spatial resolution elements of imaging systems as being the N quantities of interest we would like to measure. We wish to measure these N quantities in time T with the greatest accuracy and precision as determined by the maximum SNR possible given a finite source flux and a fixed detector noise floor. One option might be to use a multi-element detection system (multi-channel detection system) with N detectors, so each detector is measuring one of the N quantities for the full time T in order to maximize the available integration time. However, in practice, these detectors can be problematic or simply not available for many applications when compared to single-element detectors. When a single-element detector is employed in spectrometry, spectral imaging, or imaging, the measurement methodology required becomes N measurements, each for a time T/N, so that each of the N quantities is measured separately or alternatively, one can record N measurements, each for time T/N, for various groups of the N quantities. The measurement procedure for various groups of the N quantities requires an encoding procedure to choose which of the N quantities to include in each of the N measurements and a decoding procedure to recover the values for each of the N individual quantities. When combinations of N quantities are measured, the signal at the single detector element is increased and the procedure is called multiplexing. Multiplexing may be defined as measuring the sum of groups of quantities rather than measuring each quantity separately with the primary goal of increasing SNR. Multiplexing can and does increase SNR when the detection system of the measuring device is operating in the detector noise limit. The detector noise limit is defined as the noise of the detector output that is independent of the amplitude of the signal. Photon-noise-limited detector systems can benefit from multiplexing strategies, however, one must be careful not to realize a multiplex disadvantage when the source of noise is not detector noise limited.

Fourier Transform Multiplexing

In optical spectrometry a common multiplexing method employs the Michelson interferometer in a Fourier transform (FT) spectrometer where, in the best case, as much as half of the spectral energy of the source entering the aperture is incident upon the single detector element for the duration of the measurement. The output of the detector is recorded as an interferogram, which is the detector response as a function of optical retardation (twice the path difference between the arms of the Michelson interferometer). The recorded interferogram is decoded using a Fourier transform algorithm to yield the desired spectrum. FT spectrometers are routinely employed in near-infrared, mid-infrared, and

far-infrared spectrometry. In Raman spectrometry, FT spectrometers are used in the near-infrared spectral region typically using an Nd:YAG laser providing an excitation wavelength of 1064 nm. FT techniques are also employed in nuclear magnetic resonance (NMR) spectrometry and mass spectrometry.

Hadamard Transform Multiplexing

Hadamard transform multiplexing in dispersive optical spectrometry can be accomplished by use of an encoding mask at the focal plane of a dispersive spectrometer, that is operable to select different combinations of spectral resolution elements that are allowed to pass to the single detector element for each of the N required measurements. The specification of which ones of the N quantities are included in each of the N measurements of groups of the N quantities is called a weighing design. The term weighing design originally referred to a procedure for accurately determining the mass of a number of objects by weighing them in groups rather than one at a time. The optimal weighing designs and corresponding optimal encoding masks are specified by Hadamard mathematics. The Hadamard matrix or $\mathbf{H}$ matrix has all elements of $+1$ and -1 and provides the optimal weighing design for a two-pan balance or an optical instrument having two detector elements for the $+1$ and -1 measures. The Hadamard matrices specify which quantities go to each balance pan or detector element.

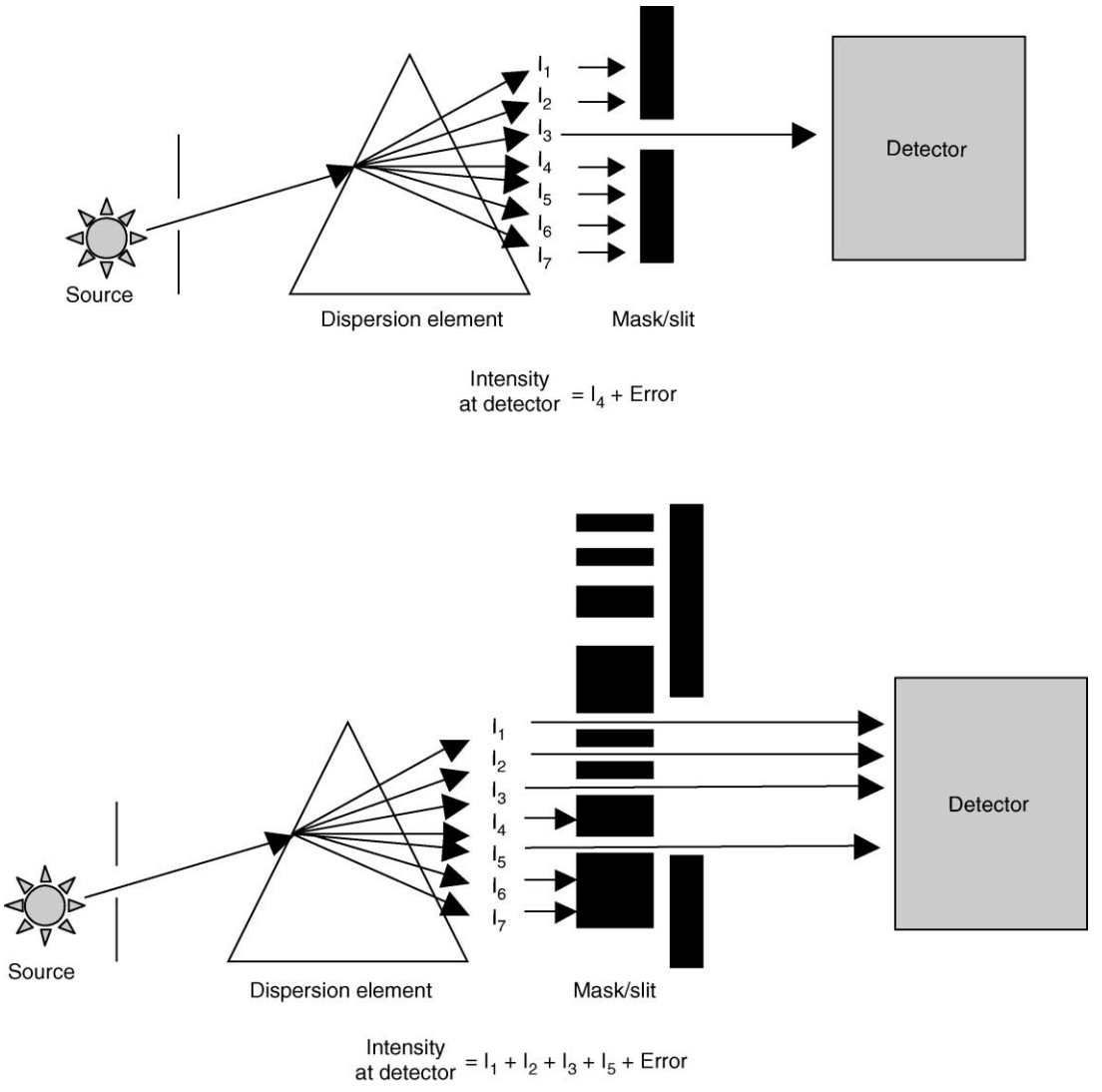


Figure 1 Simple schematic of a conventional and a Hadamard transform spectrometer. The top shows a conventional spectrometer where only one of seven spectral resolution elements ($N = 7$) impinge upon the single element detector for each of N measurements. The bottom shows a Hadamard transform spectrometer where $N = 7$. One more than half of the spectral resolution elements impinge upon the single element detector for each of the N measurements.

One and Two Detector Hadamard Encodement Matrices

In the case of a dispersive spectrometer with a single detector element, the optimal weighing designs and corresponding optimal encoding masks are provided by the simplex matrix or **S** matrix derived from the **H** matrix. The type of **S** matrix that is most convenient for optical applications is the cyclic or left circulant **S** matrix where only the first row needs to be known to generate the complete cyclic **S** matrix of N columns and N rows. Each row is obtained by shifting the previous row by a left-shift registration in which the element in the far left column moves to the far right column of the next row and each of the other $N-1$ elements moves one column to the left. Each of the N rows of the cyclic **S** matrix specifies one of the N encodements and each of the N columns specifies one of the N spectral resolution elements. The elements of the cyclic **S** matrix are 1 and 0 instead of the $+1$ and -1 found in the **H** matrix implementation. Spectral resolution elements allowed and not allowed to reach the single detector element are given as 1's and 0's, respectively. Each row and each column of the cyclic **S** matrix is a different sequence of $(N+1)/2$ elements of 1's and $(N-1)/2$ elements 0's so that each encodement in the row of the cyclic **S** matrix has a different combination of $(N+1)/2$ spectral resolution elements that are incident upon the single detector element. Just as the positions of the 1's in a row show which spectral resolution elements are included in a particular encodement, the positions of the 1's in a column show which encodements include that particular spectral resolution element. The result is that provided the detector noise is independent of its output signal, the SNR is increased by a factor of $(N+1)/(2\sqrt{N})$, approximately $(\sqrt{N})/2$, when compared to the method of measuring each spectral resolution element separately as in a raster scan performed by a classical scanning dispersive spectrometer.

Multiplexing in dispersive optical spectrometry as introduced in the previous paragraph is called Hadamard transform spectrometry (HTS) (see Figure 1). It is a unique combination of dispersive and multiplexing spectrometries. The choice of transparent or opaque for the elements of a multislit array or Hadamard encoding mask provides an encoding of spectral information that may be decoded into a conventional spectrum by Hadamard mathematics. In a typical dispersive spectrometer, the source energy is collected and separated into its individual spectral resolution elements by a spectral separator and then collected and focused for spatial presentation on a focal plane. Unlike a scanning dispersive spectrometer, which possesses a single exit slit, the Hadamard transform (HT) spectrometer employs a multislit array as a Hadamard encoding mask. This arrangement allows for the simultaneous measurement of a multitude of spectral resolution elements at a single-element detector and produces a multiplexing spectrometer.

Theory of Hadamard Multiplexing

The operation of the HT spectrometer can be described using matrix algebra with column vectors [**D**] and [**I**] and the N-order square matrices $\mathbf{S}_N$, $\mathbf{S}_N^{-1}$, $\mathbf{S}_N^{\mathrm{T}}$, and $\mathbf{T}_N$. [**D**] is a column vector of N rows containing the detector responses for the N encodements. [**I**] is a column vector of N rows containing the N spectral resolution elements to be determined. $\mathbf{S}_N$ is the cyclic **S** matrix of order N that is the basis for construction and design of the Hadamard encoding mask. $\mathbf{S}_N^{-1}$ is the inverse of $\mathbf{S}_N$ and is given by $(2/(N+1))\mathbf{T}_N$. $\mathbf{T}_N$ is a matrix of $+1$ and -1

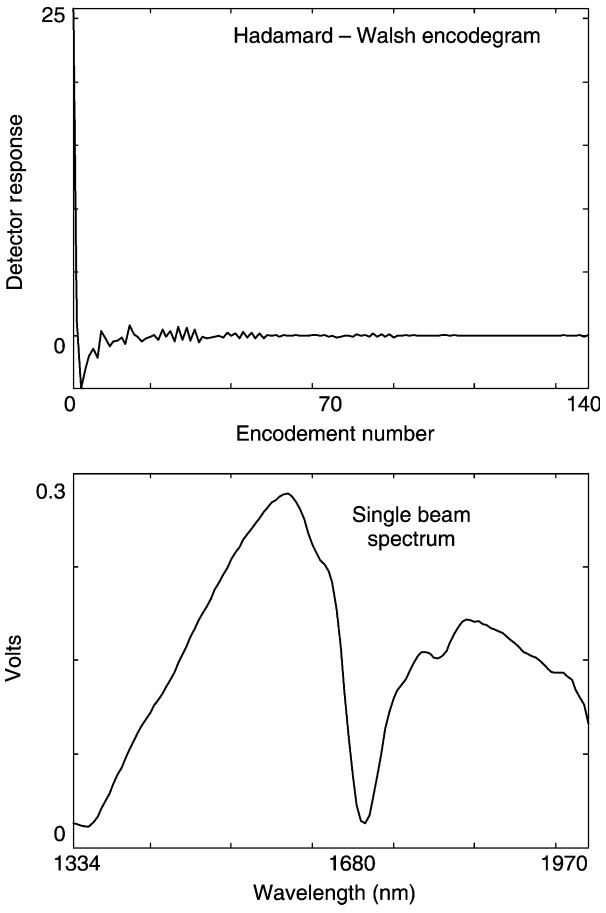


Figure 2 An encodegram (top plot) and the corresponding single-beam spectrum (bottom plot) as a result of decoding the encodegram using Hadamard mathematics.

elements obtained from $\mathbf{S}_N^T$, the transpose of $\mathbf{S}_N$, with $\mathbf{T}_N$ having +1's and −1's where $\mathbf{S}_N^T$ has 1's and 0's, respectively. A cyclic matrix is required to be symmetric so that an interchange of columns and rows can recover the original matrix and $\mathbf{S}_N^T$ then becomes equal to $\mathbf{S}_N$. The encoding of the radiation by the HT spectrometer is given by $[\mathbf{D}] = \mathbf{S}_N[\mathbf{I}]$ and the resulting primary data are a record of detector response versus encodement number. The plot of detector response versus encodement number is called an encodegram (analogous to the interferogram produced by an FT spectrometer). The decoding of the encodegram to recover the conventional spectrum using a Hadamard transform can be described by $[\mathbf{I}] = \mathbf{S}_N^{-1}[\mathbf{D}]$, obtained by left multiplying $[\mathbf{D}] = \mathbf{S}_N[\mathbf{I}]$ by $\mathbf{S}_N^{-1}$. In practice, it is not necessary to invert $\mathbf{S}_N$ because there is a fast Hadamard transform (FHT) algorithm available when N is of the form $N = 2^w - 1$, w being an integer. There are three known methods for generating the first row of cyclic S matrices of order $N = 2^w - 1$. Consider the example of $N = 7$ for $w = 3$ where the first row of $\mathbf{S}_7$ is 1 1 1 0 1 0 0. Figure 2 shows an encodegram and the result of decoding using Hadamard mathematics. Figure 3 contains the matrices and matrix equations for $N = 7$.

Hadamard Encoding Masks

The key component of any HT spectrometer is the Hadamard encoding mask (Figure 4). Two choices for the Hadamard encoding mask are the moving mechanical mask and the stationary electro-optical mask. The moving mask for N spectral resolution elements requires $2N - 1$ mask elements that are either completely open with a transmittance of 1 or completely closed with a transmittance of 0. The stationary mask for N spectral resolution elements requires only N mask elements that are selected to be either transparent: transmittance $T_t = 1$ or <1 or opaque: transmittance $T_o = 0$ or >0. The resulting SNR is dependent upon $\Delta T = T_t - T_o$, which has its optimal value of unity only for an ideal mask which is $T_t = 1$, $T_o = 0$. The advantage of the moving mask is

$$
[D] = \begin{matrix} D_1 \\ D_2 \\ D_3 \\ D_4 = \\ D_5 \\ D_6 \\ D_7 \end{matrix}
\begin{matrix} I_1 + I_2 + I_3 + I_5 \\ I_1 + I_2 + I_4 + I_7 \\ I_1 + I_3 + I_6 + I_7 \\ I_2 + I_5 + I_6 + I_7 \\ I_1 + I_4 + I_5 + I_6 \\ I_3 + I_4 + I_5 + I_7 \\ I_2 + I_3 + I_4 + I_6 \end{matrix}
\qquad
S_7 = \begin{matrix} 1\,1\,1\,0\,1\,0\,0 \\ 1\,1\,0\,1\,0\,0\,1 \\ 1\,0\,1\,0\,0\,1\,1 \\ 0\,1\,0\,0\,1\,1\,1 \\ 1\,0\,0\,1\,1\,1\,0 \\ 0\,0\,1\,1\,1\,0\,1 \\ 0\,1\,1\,1\,0\,1\,0 \end{matrix}
\qquad
[I] = \begin{matrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \end{matrix}
$$

$$
T_7 = \begin{matrix} 1\,1\,1\,{-1}\,1\,{-1}\,{-1} \\ 1\,1\,{-1}\,1\,{-1}\,{-1}\,1 \\ 1\,{-1}\,1\,{-1}\,{-1}\,1\,1 \\ {-1}\,1\,{-1}\,{-1}\,1\,1\,1 \\ 1\,{-1}\,{-1}\,1\,1\,1\,{-1} \\ {-1}\,{-1}\,1\,1\,1\,{-1}\,1 \\ {-1}\,1\,1\,1\,{-1}\,1\,{-1} \end{matrix}
\qquad
S_7^{-1} = (1/4)T_7 = \begin{matrix} (1/4)(1\,1\,1\,{-1}\,1\,{-1}\,{-1}) \\ (1/4)(1\,1\,{-1}\,1\,{-1}\,{-1}\,1) \\ (1/4)(1\,{-1}\,1\,{-1}\,{-1}\,1\,1) \\ (1/4)({-1}\,1\,{-1}\,{-1}\,1\,1\,1) \\ (1/4)(1\,{-1}\,{-1}\,1\,1\,1\,{-1}) \\ (1/4)({-1}\,{-1}\,1\,1\,1\,{-1}\,1) \\ (1/4)({-1}\,1\,1\,1\,{-1}\,1\,{-1}) \end{matrix}
$$

$$
[I] = \begin{matrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \end{matrix}
= S_7^{-1}[D] =
\begin{matrix} (1/4)(\,D_1 + D_2 + D_3 - D_4 + D_5 - D_6 - D_7\,) \\ (1/4)(\,D_1 + D_2 - D_3 + D_4 - D_5 - D_6 + D_7\,) \\ (1/4)(\,D_1 - D_2 + D_3 - D_4 - D_5 + D_6 + D_7\,) \\ (1/4)(-D_1 + D_2 - D_3 - D_4 + D_5 + D_6 + D_7\,) \\ (1/4)(\,D_1 - D_2 - D_3 + D_4 + D_5 + D_6 - D_7\,) \\ (1/4)(-D_1 - D_2 + D_3 + D_4 + D_5 - D_6 + D_7\,) \\ (1/4)(-D_1 + D_2 + D_3 + D_4 - D_5 + D_6 - D_7\,) \end{matrix}
$$

For example: Note $I_1 = (1/4)\,[\,(D_1 + D_2 + D_3 + D_5\,) - (D_4 + D_6 + D_7\,)\,]$ so

$$
I_1 = (1/4)\,\{\,[\,(I_1 + I_2 + I_3 + I_5\,) + (I_1 + I_2 + I_4 + I_7\,) + (I_1 + I_3 + I_6 + I_7\,) + (I_1 + I_4 + I_5 + I_6\,)\,] - [\,(I_2 + I_5 + I_6 + I_7\,) + (I_3 + I_4 + I_5 + I_7\,) + (I_2 + I_3 + I_4 + I_6\,)\,]\,\} = (1/4)\,\{\,4I_1\,\} = I_1
$$

Figure 3 Matrices and matrix equations (note since $\mathbf{S}_7$ is symmetric $\mathbf{S}_7^T = \mathbf{S}_7$).



Figure 4 A linear encodement of length $N = 43$ generated by using a quadratic residue construction.

no optical transmission problems. However, it suffers the disadvantage of having moving parts and position repeatability problems. For the case of the stationary mask, it has the advantage that it has no moving parts and no positioning repeatability problems. The disadvantage of the stationary mask is the optical transmission problem where $T_t < 1$ and $T_o > 0$. The moving Hadamard encoding mask of $2N - 1$ mask elements is changed to the next encodement by translation of the Hadamard encoding mask to the left past the aperture by a single mask element. For $N = 7$ and $2N - 1 = 13$, the assignment of 1 and 0 to positions 1 through 13 is [1 1 1 0 1 0 0 1 1 1 0 1 0]. For example, the first encodement corresponding to the first row of S_7 in Figure 3 is from positions 1 through 7 as [1 1 1 0 1 0 0] and the fourth encodement corresponding to the fourth row of S_7 in Figure 3 is from positions 4 through 10 as [0 1 0 0 1 1 1].

2D Encoding Masks for Multiplexing in Spectral Imaging and Imaging

If the conventional one-dimensional (1D) Hadamard encoding mask is folded in some manner, the result is the generation of a two-dimensional (2D) Hadamard encoding mask (Figure 5). Similar to the 1D Hadamard encoding masks that are used for spectral mutiplexing, 2D Hadamard encoding masks can be used for the spatial multiplexing done in Hadamard transform spectral imaging (HTSI) and Hadamard transform imaging (HTI). It is important to note that spectrometry as HTS, spectral imaging as HTSI, and imaging as HTI all realize the same theoretical SNR improvement based on the number of spectral, spatio-spectral, or spatial resolution elements (N), respectively. Consider the folding of a 1D Hadamard encoding mask for $N = 15$ into a 2D Hadamard encoding mask of three rows and five columns (Figure 6). A cyclic S_{15} has a first row of [0 0 0 1 0 0 1 1 0 1 × 0 1 1 1 1] so the 2D Hadamard encoding mask has a first row of [0 0 0 1 0], a second row of [0 1 1 0 1], and a third row of [0 1 1 1 1] from columns 1 through 5, 6 through 10 and 11 through 15, respectively, from S_{15}.

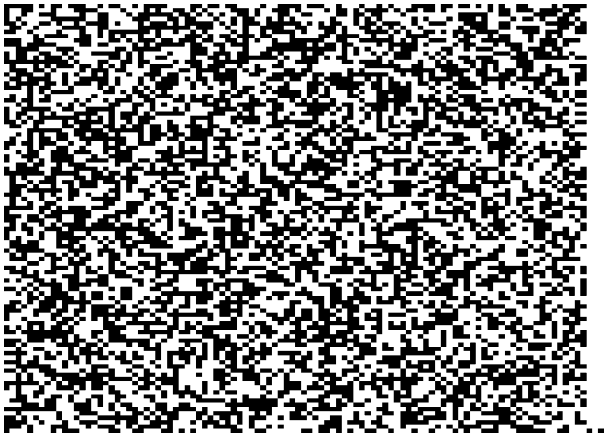


Figure 5 A two dimensional encodement of length $N = 10{,}007$ in a 100×100 format generated by quadratic residue construction. The remaining 7 elements are shown as the last 7 elements in the bottom row on the right. In practice these 'extra' data points can be masked off and used as a measure of system noise characteristics.

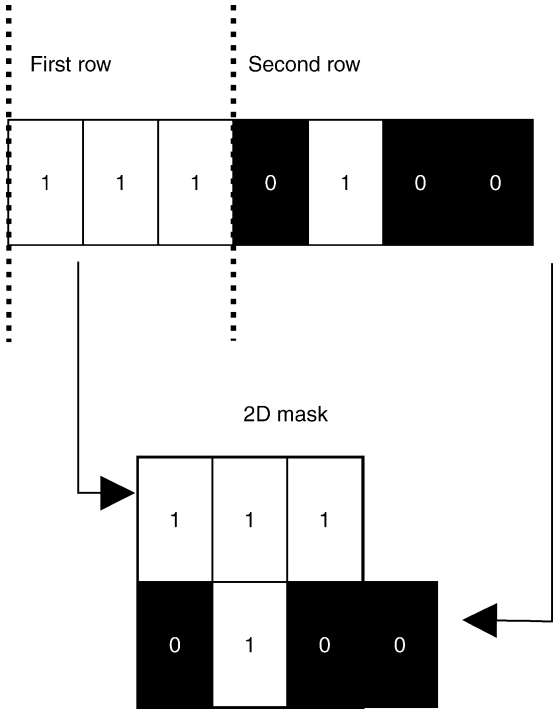


Figure 6 Spectrometry to imagery: The folding of a ID encoding mask that can be used for spectrometry is used to generate a 2D encoding mask that can be used for imaging. Note that using a symmetric aperture window where only 3×2 elements are used, the mask element that is not included in the windowed aperture can be used as a measure of the dark noise of the system, however, a more conventional encodement matrix might be one that can be divided into an integer number of rows and columns such as $N = 15$ where one can construct a 5×3 2D encodement mask for imagery and use a symmetric aperture that uses the entire length of the encodement matrix.

Hadamard Mathematics

Consider a general multiplexing problem where one seeks to measure N quantities, $X = \{x_0, x_1, \ldots, x_{N-1}\}$, in the presence of noise, and wishes to determine a vector function $F(X)$, such that the measurement of $F(X)$ yields the best estimate of X. Conservation of energy and superposition principles dictate a restriction to linear functions. The principle of rank suggests that we restrict our measurements to N-dimensional linear combinations. Determining the best linear combination is the challenge. In the case of constant error for each measurement, an answer is provided by Hotelling's theorem, which proves Hadamard matrices are optimal for such measurements. Hadamard matrices uniquely achieve the optimal bound for root-mean-square measurement error. A Hadamard matrix is a matrix $\mathbf{H}$ for which $\mathbf{H}_{i,j} = \pm 1$, and $\mathbf{H}^{-1} = \frac{1}{N}\mathbf{H}^{\mathrm{T}}$. The definitions of matrix inverse, transpose, and multiplication dictate that the rows of the matrix $\mathbf{H}$ form an orthogonal basis with each basis element having norm $\sqrt{N}$. The norms of the rows of $\mathbf{H}$ follow from the entries of the matrix that are ± 1, and place no additional restriction on $\mathbf{H}$. The columns of $\mathbf{H}$ are also an orthogonal basis with these same norms. The rows (columns) of the matrix $\frac{1}{\sqrt{N}}\mathbf{H}$ form an orthonormal basis. From these few statements and definitions, one can already derive estimates of the multiplex advantage. For example, one seeks to measure X and in one case measures $Y_1 = X + \varepsilon$, where ε is a random noise or error that has a Gaussian distribution with a mean of 0 and variance σ^2. In another case, suppose that one measures $Y_2 = HX + \varepsilon$. In this second case, one would estimate

$$X_2 = \frac{1}{N}\mathbf{H}^{\mathrm{T}}(\mathbf{H}X + \varepsilon) = X + \left(\frac{1}{\sqrt{N}}\right)^2 \mathbf{H}^{\mathrm{T}}\varepsilon$$
$$= X + \frac{1}{\sqrt{N}}\varepsilon_1$$

where ε_1 is random noise with the same characteristics as ε. It follows that the measurement error in the second case has a standard deviation σ that is less than in the former case, with an improvement by the factor $\frac{1}{\sqrt{N}}$.

Hadamard Computations

The ± 1 Hadamard matrices can be applied to two detector optical systems, and 0,1 matrices can be applied to single detector systems. In a two detector system $\mathbf{H}^{\mathrm{T}}$ is proportional to $\mathbf{H}^{-1}$ and one can invert these transforms. Suppose that one has chosen a particular Hadamard matrix $\mathbf{H}$, with which to conduct a multiplexed measurement. One can show that multiplication of any row or any column by -1 does not affect the efficacy of $\mathbf{H}$ in improving SNR. Hence one may assume that the first row and first column of $\mathbf{H}$ are all $+1$. From $\mathbf{H}$, we can form a new $(N-1)\times(N-1)$ matrix $\mathbf{S}$ by deleting the first row and column of $\mathbf{H}$, and then changing each 1 to a 0, and each -1 to a 1. This new matrix $\mathbf{S}$, an $\mathbf{S}$-matrix, is the one corresponding to $\mathbf{H}$ that is to be used for single-element detector optical systems. It is known that $\mathbf{S}$ matrices are optimal for 0,1 matrices just as Hadamard matrices are optimal for ± 1 matrices, with signal to noise improvement by a factor $\frac{\sqrt{N-1}}{2}$, where here N is the dimension of $\mathbf{H}$.

Letting $\mathbf{I}$ be the identity matrix, and $\mathbf{J}$ be the matrix whose entries are all 1, it follows that $\mathbf{S}^{\mathrm{T}}\mathbf{S} = \frac{N}{4}(\mathbf{I} + \mathbf{J})$, where N is the dimension of $\mathbf{H}$. From this we can ascertain that a multiplexed system using the matrix $\mathbf{S}$, multiplied by the encoded parameters of $\mathbf{S}^{\mathrm{T}}$ recovers the original signal to within the addition of an offset. In practice this is what is done to recover the original X. In certain cases, there is also a fast algorithm for computing $\mathbf{S}^{\mathrm{T}}Y$, without ever explicitly computing $\mathbf{S}^{\mathrm{T}}$.

Constructions for Hadamard Matrices

It is known that if $\mathbf{H}$ is an $N \times N$ Hadamard matrix, then N must be 1, 2, or a multiple of 4 and so nontrivial $\mathbf{S}$-matrices must be of size one less than a multiple of 4. It is conjectured that Hadamard matrices exist for each N which is a multiple of 4. Some constructions of Hadamard matrices are enumerated below.

Walsh–Hadamard Construction

A Hadamard matrix of size $N = 2^K$ can be constructed by sampling the first N Walsh functions at $0, 1/N, 2/N, 3/N, \ldots, (N-1)/N$, where the Walsh functions are defined by $W_0(x) = 1$, for $0 \le x < 1$, $W_0(x) = 0$ otherwise, and the recurrence relations, for each non-negative integer n:

$$\begin{aligned} W_{2n}(x) &= W_n(2x) + W_n(2x-1), \text{ and} \\ W_{2n+1}(x) &= W_n(2x) - W_n(2x-1) \end{aligned} \qquad [1]$$

Tensor Product Construction

Walsh is a special case of iterating this construction, starting with the matrix $\mathbf{H}_1 = [1]$. In general, if $\mathbf{H}_N$ is

a Hadamard matrix of size $N \times N$, then:

$$\mathbf{H}_{2N} = \begin{bmatrix} H_N & H_N \\ H_N & -H_N \end{bmatrix} \qquad [2]$$

is a Hadamard matrix of size $2N \times 2N$.

Quadratic Residue Construction

This construction is an example of a cyclic **S**-matrix: one for which each row is a cyclic shift of the previous row. Given a prime number $p = 4n + 3$, define, for $i = 0, \ldots, p - 1$, the values $\mathbf{S}_{1,i} = 1$ if there is some integer k, such that the remainder of dividing k^2 by p is i, and $\mathbf{S}_{1,i} = 0$ otherwise. Taking the cyclic matrix with the first row given by $\mathbf{S}_{1,i}$, yields a Hadamard **S**-matrix.

Other constructions are possible. Maximal length shift register codes, and a construction involving quadratic residues of twin primes are two examples.

Fast Hadamard Transforms

The fast Walsh transform can be deduced from eqn [1]. The basic idea is that, in eqn [1], the function W_n is specified by $\log_2 n$ recurrence relations. For this reason, a Walsh transform (or its transpose) can be computed in $O(n \log_2 n)$ steps. The algorithm goes as follows, for a sequence of length $N = 2^n$:

(i) begin with $Y_{0,0}, Y_{0,1}, \ldots, Y_{0,N-1} = Y_0, Y_1, \ldots, Y_{N-1}$
(ii) for $i = 1, \ldots, n - 1$, let $Y_{i,k} = Y_{i-1,2k} + Y_{i-1,2k+1}$, and $Y_{i,n/2+k} = Y_{i-1,2k} - Y_{i-1,2k+1}$, for $k = 0, \ldots, N/2$
(iii) output $X_i = Y_{n-1,i}$

Then the resulting X_i is the inverse Walsh transform of Y_i.

Note that cyclic **S** matrices that come from maximal length shift registers are known to be permutations of the Wzalsh–Hadamard **S** matrices, and hence there is also a fast algorithm for inverting these, albeit involving some tedious permutations.

Optics for Hadamard Encoded Apertures

Hadamard optical systems utilize spatially encoded apertures that can be employed either at the entrance aperture of the optical system, the exit aperture or both. They have the common attribute that they encode the available aperture spatially where the spatial resolution elements that make up the encodement dictate the spectral, spatio-spectral, or spatial resolution elements that propagate through the optical system, including diffractive optical elements and on to the sensor or exit aperture. These masks have some spatial extent that places special requirements on the optics of the system. As the encodement mask grows either by longer length encodements with fixed subapertures or as the subaperture dimension grows for a fixed encodement length, the spatial resolution elements making up the subapertures in the encodement mask depart from the optical axis. When the resolution elements depart from the optical axis or paraxial condition, it becomes important to employ optics that can image the off-axis resolution elements without inducing excessive aberrations that degrade the performance and cripple the advantages gained by HT multiplexing.

Typically the optical path for conventional monochromators begins with a source that is focused onto an aperture plane that has a large aspect ratio aperture know as a slit. This slit is often very small in extent in the dispersion plane compared to the other extent in the spatial plane. However, it is not required that this aspect ratio is large. If the aspect ratio is close to 1 then simple spherical optics can be employed that perform well as long as the departure from the optical axis is kept to a minimum. However, most monochromators have a large aspect ratio in order to increase the opportunity to maximize throughput, and detectors must be able to 'see' the large extent of the slit aperture. The light entering the slit aperture is then dispersed and focused onto an exit slit aperture. Monochromators are only required to perform well on the optical axis and do not typically employ optics that can manage rays that depart from the optical axis in the plane of dispersion as required by HT multiplexing instruments. To employ encoding techniques the optical system is required to utilize optical performance attributes normally found only in imaging and spectral imaging systems. This requirement is driven by the extent of the encoding mask. The extent of the encoding mask is governed by the diffraction limit of the wavelengths within the bandpass, the encodement length N, and the attributes of the optical system.

In a conventional dispersive spectrometer the radiation from a source is collected and separated into its individual spectral resolution elements by a spectral separator such as a diffraction grating or prism and then is collected and focused for spatial

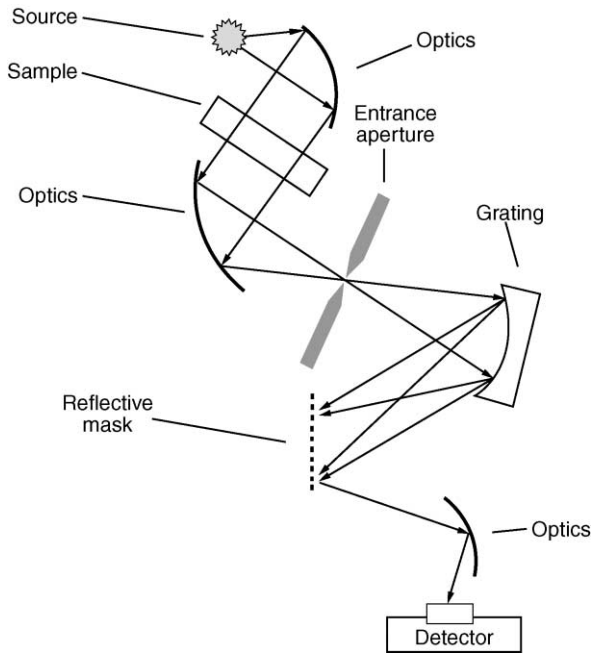


Figure 7 Example implementation of a Hadamard transform spectrometer with reflective optics. Light from a source passes through a sample, is transferred into an entrance aperture, bounced off of a concave grating, and then a reflective mask, and finally transferred to a detector.

presentation on a focal plane. The dispersive spectrometer uses a single exit slit to select one spectral resolution element at any given time, out of N possible spectral resolution elements for measurement by the detector. The Hadamard transform spectrometer (HTS, Figure 7) uses an array of slits (i.e., a mask) at the focal plane to select one more than half, $(N + 1)/2$, of the spectral resolution elements at the focal plane for measurement by the detection system. The optical challenge to effect an HT multiplexing spectrometer is to collect all of the spatially distributed individual bandpass images of the entrance slit and transfer them to as small a detector as possible. One desires to keep the area of the detector at a minimum as the noise of many detectors increases with the square of the area. If the optics are able to illuminate a single detector element with all of the available light impinging upon the focal plane containing the spatially distributed images of the slit for each of the N bandpass resolution elements, a multitude of spectral resolution elements can be measured simultaneously using a single detector element. This arrangement results in a multiplexing spectrometer. The recovery of N spectral resolution elements requires measuring the detector response for N different encodements of $(N + 1)/2$ open mask elements. The raw data are recorded as the detector response versus encodement number and is called an encodegram. Hadamard transformation of the encodegram yields the spectrum.

The History of Applied Hadamard Multiplexing

The Hadamard transform instruments developed in the 1960s and 1970s employed moving masks. Significant problems such as misalignment and jamming associated with a moving mask led to a reputation of poor reliability and contributed to a dormant period in the development of HTS and HTI. Interest was rekindled in the 1980s using stationary Hadamard encoding masks based on liquid crystal (LC) technology. The first generation 1D stationary Hadamard encoding mask was a cholesteric LC with $N = 127$ mask elements and used polarization as its operating phenomenon. Two parallel polarizers and rotation or lack of rotation of the polarized radiation generated the opaque and transparent states, respectively. The second generation 1D stationary Hadamard encoding mask was fabricated using a polymer dispersed liquid crystal (PDLC) material with $N = 255$ mask elements and used light scattering as its operating phenomenon. The PDLC contained LC droplets dispersed in a polymer matrix whose index of refraction matched the index of refraction in one direction in the birefringent LC droplet. Alignment of the LC droplets optical axis under an applied voltage removed discontinuities in index of refraction at the polymer matrix/LC interface to generate a good transparent state while random orientation of LC droplets in the polymer matrix generated the opaque state from light scattering by the discontinuities in index of refraction at the polymer matrix/LC droplet interface. A 2D stationary Hadamard encoding mask was also based on LC technology. A ferro-electric liquid crystal positioned between a pair of polarizers with perpendicular orientation operated as an electro-optic half-wave plate when a +value of applied voltage rotated the plane of polarization by 90 degrees to produce the transparent state and a −value of applied voltage left the plane of polarization unaltered to produce the opaque state.

Development based on stationary Hadamard encoding masks continued in the 1990s and a 2D moving Hadamard encoding mask was also fabricated and used to perform imaging in the near-infrared and mid-infrared spectral regions. Note that the mid-infrared spectral region is not generally accessible via Hadamard encoding masks based on

LC technology, since any LC material is expected to have strong absorption bands in the mid-infrared spectral region. From the mid-1990s to the present, the stationary Hadamard encoding mask of choice for the visible and near-infrared spectral regions has been the digital micromirror device (DMD), a device based on micro-optoelectromechanical systems (MOEMS) technology and developed by Texas Instruments for projector display applications. One DMD format incorporates 508 800 micromirrors in an 848 column by 600 row array that is 14.4 mm wide by 10.2 mm high. Each individual micromirror is 16 μm square and adjacent micromirrors are separated by a 1 μm gap. The micromirrors are individually addressable and rotate by +10 or −10 degrees about the diagonal axis to produce binary 'on' and 'off' states. The on state has T_t determined by the mirror reflectivity and approaches 1 while the off state approaches $T_o = 0$. The ideal condition of on and off is not realized due to diffraction of the light off of the small and periodic features of the micromirror device. The DMD is an array of spatial resolution elements that may be selected as groups of super-resolution elements or as individual resolution elements consisting of a single micromirror. The DMD resolution elements are realized as spectral resolution elements in the spectrometer with the columns attributed to the frequency or wavelength dimension and the rows attributed to the slit height dimension. The DMD resolution elements are utilized as spatio-spectral resolution elements in the imaging spectrograph with the columns as the frequency or wavelength dimension and the rows as a vertical spatial dimension with the horizontal spatial dimension being accessed, if desired, by translating the sample relative to the imaging spectrograph. The DMD resolution elements are spatial resolution elements in the imager with the columns for the horizontal dimension and the rows for the vertical dimension and the frequency or wavelength dimension provided by other instrumentation. If a photo-acoustic detection system is present then the depth dimension of the sample may also be accessed by changing the modulation frequency used in the photo-acoustic detection system.

Some important features of HTS to keep in mind are:

(i) it is a dispersive technique using a spectral separator;
(ii) it is a multiplexing technique using a single-element detector;
(iii) it uses a Hadamard encoding mask (multislit array) in a focal plane;
(iv) it sends one more than half the resolution elements to the single-element detector in each encodement;
(v) it uses a number of encodements equal to the number of resolution elements desired and the number of mask elements (pixels) in the stationary Hadamard encoding mask (a moving Hadamard encoding mask has $2N - 1$ mask elements);
(vi) it has each encodement containing a different combination of one more than half the resolution elements;
(vii) it has as its primary data the encodegram, a record of detector response versus encodement number; and
(viii) it uses a FHT of the encodegram to decode the encodegram and generate the spectrum or image.

Note that all of these, except for item (i), also apply to HTSI and HTI. The DMD promises to be the best Hadamard encoding mask yet developed for the visible and near-infrared spectral regions. However, its potential applications in spectrometry and imaging are by no means limited to HT techniques since the information corresponding to any micromirror in the DMD may be included in or excluded from any measurement as desired by the investigator. An instrument with no moving parts, other than the micromirrors in the DMD, promises to provide a compact and robust instrument for operation in potentially hostile environments ranging from process control to outer space. It is our belief that the combination of a DMD with a single-element detector may provide an important advance in spectroscopic instrumentation and that instrumentation based on the DMD may lead to a host of environmental, industrial, medical, and military applications.

See also

Fourier Optics. **Imaging:** Multiplex Imaging. **Optical Communication Systems:** Optical Time Division Multiplexing. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Bellamy MK, Mortensen AN, Hammaker RM and Fateley WG (1997) Chemical mapping in the mid- and near-IR spectral regions by Hadamard transform/FT-IR spectrometry. *Applied Spectroscopy* 51: 477–486.

Fateley WG, Hammaker RM, Paukstelis JV, *et al.* (1993) Application of a two-dimensional Hadamard encoding

mask for the imaging of thin-layer chromatography plates by laser-induced fluorescence or surface-enhanced Raman scattering and for use with a photoacoustic detector to generate three-dimensional photoacoustic images. *Applied Spectroscopy* 47: 1464–1470.

Fateley WG, Hammaker RM, DeVerse RA, Coifman RR and Geshwind FB (2002) The other spectroscopy: demonstration of a new de-dispersion imaging spectrograph. *Vibrational Spectroscopy* 29: 163–170.

Hammaker RM, Bohlke AP, Freeman RD, *et al.* (1996) Hadamard transform Raman spectrometry. In: Laserna JJ (ed.) *Modern Techniques in Raman Spectroscopy*, pp. 143–226. Chichester: Wiley.

Hammaker RM, DeVerse RA, Asunskis DJ and Fateley WG (2002) Hadamard transform near-infrared spectrometers. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, vol. 1, pp. 453–460. Chichester: Wiley.

Harwit M and Sloane NJA (1979) *Hadamard Transform Optics*. New York: Academic Press.

Nonlinear Laser Spectroscopy

P Ewart, University of Oxford, Oxford, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction: Linear and Nonlinear Spectroscopy

Light propagates in transparent or weakly absorbing media by inducing the atoms or molecules in the material to oscillate in response to the electromagnetic field of the light. The induced dipole oscillations set up a polarization field **P** which radiates the wave in the propagation direction. The induced polarization **P** is proportional to the amplitude of the electric field **E** in the incident light:

$$\mathbf{P} = \varepsilon_0 \chi^{(1)}(\omega)\mathbf{E} \qquad [1]$$

where ε_0 is the permittivity of free space and $\chi^{(1)}(\omega)$ is the first-order, frequency-dependent, susceptibility of the material. This interaction of the light with the medium generally results in some absorption of energy from the field and a shift in the phase of the propagating light. Both the phase delay and the absorption depend on the frequency ω (and wavelength λ) of the light. The phase shifting is responsible for dispersion – the variation, as a function of frequency, of the speed (phase velocity) of light in the medium. The dispersion and absorption effects are described by the real and imaginary parts, respectively, of the complex susceptibility:

$$\chi(\omega) = \chi_{\mathrm{Re}}(\omega) + \mathrm{i}\chi_{\mathrm{Im}}(\omega) \qquad [2]$$

This susceptibility parameter describes the optical properties of the medium. Linear spectroscopy uses the frequency-dependent property of the $\chi(\omega)$ to create spectra by absorption or dispersion of specific frequencies that are resonant with transitions between internal energy states of the molecules. The electric field in the incident light is considered to be weak when it leaves the atomic and molecular properties essentially unaffected. Nonlinear optical spectroscopy, however, is based on the ability of intense laser light to modify the optical properties of the medium through which it propagates. These effects are also frequency dependent and so can be used to generate spectra. When the strength of the electric field becomes comparable with the electric fields within the atoms or molecules (typically $\sim 10^{11}\,\mathrm{V\,m^{-1}}$) then the induced polarization becomes nonlinear and is described by a power series expansion:

$$\mathbf{P} = \varepsilon_0[\chi^{(1)}(\omega)\mathbf{E}_1 + \chi^{(2)}(\omega)\mathbf{E}_1\mathbf{E}_2 + \chi^{(3)}(\omega)\mathbf{E}_1\mathbf{E}_2\mathbf{E}_3 + \text{higher-order terms}] \qquad [3]$$

$\chi^{(2)}(\omega)$ and $\chi^{(3)}(\omega)$ are the second- and third-order susceptibilities respectively. The real and imaginary parts of these susceptibilities (and of the higher-order terms) describe observable effects arising from absorption and dispersion. The associated interactions involve two and three photon processes, respectively. In general, nonlinear interactions involve multiphoton processes and will be described by the appropriate term in the expansion of the induced polarization.

Broadly speaking, nonlinear spectroscopy is based on detection of these multiphoton interactions by either incoherent or coherent processes. Incoherent processes include fluorescence and ionization where, following the multiphoton interaction, photons or electrons respectively are emitted spontaneously in all directions. Coherent processes, on the other hand, arise from the coupling of the input driving fields via the nonlinear polarization of the medium.

The waveform of the polarization oscillations is a distorted form of the driving field. Fourier analysis of these nonlinear oscillations would reveal extra frequencies not present in the incident field. When a suitably phased array of dipoles is established, the nonlinear optical process is detected by emission of coherent directional light. The efficiency with which this coherent signal beam is generated is crucially dependent upon the correct phasing of the radiating molecules. This process is known as phase matching and is achieved by geometrical arrangements of the input beams or by modifying the refractive index to counteract dispersion between the generating polarization wave and the radiated signal wave.

Apart from the general division of nonlinear spectroscopy according to whether the detection uses incoherent or coherent emission, the techniques are characterized as frequency or time domain methods, according to how the signal is generated and detected. In most cases the signals recorded in either domain may be related by a Fourier transformation to an equivalent signal in the other.

The advantages of nonlinear spectroscopic techniques lie in their ability to provide high spectral, temporal, and spatial resolution. In particular, it is possible, using a variety of nonlinear optical processes, to eliminate inhomogeneous broadening of the spectral lines recorded. Second, when coherent signal beams are generated they can be readily distinguished from incoherent background emissions, thus permitting high signal-to-noise ratios. Third, the use of short laser pulses, down to the order of 10^{-15} s, allows high temporal resolution. High spatial resolution is achieved by generating signals from interaction regions defined by crossed and focused beams. Although the coefficients of successive terms in the power series expansion of the nonlinear polarization decrease by many orders of magnitude with increasing order, signal strengths may be made comparable to those of linear spectroscopy. Strong signals are obtained by the use of resonant enhancement of the nonlinear susceptibility, high field strengths from laser sources, and the use of sensitive photodetectors.

Second-Order Nonlinear Effects

Weak light fields leave unaffected the optical properties of the medium through which they propagate. When, however, the second-order susceptibility $\chi^{(2)}$ contributes significantly to the induced polarization, the complex refractive index includes a field-dependent term:

$$\mathbf{P} = \varepsilon_0[\chi^{(1)}(\omega) + \chi^{(2)}(\omega)\mathbf{E}_1]\mathbf{E}_2 \qquad [4]$$

Owing to the symmetry properties of $\chi^{(2)}$, second-order nonlinear effects are absent in isotropic media or those possessing a center of inversion symmetry. They may, however, be observed in anisotropic materials such as crystalline solids where they are the basis of frequency mixing schemes leading to second-harmonic generation, sum- and difference-frequency generation, and optical parametric amplification and oscillation. The symmetry of gas phase media, such as atomic vapors, may be broken by application of an external electric or magnetic field. Resonantly enhanced second-harmonic generation has been achieved in alkali metal vapors in the presence of a magnetic field giving potential for spectroscopy. Symmetry breaking is achieved also on solid surfaces where sum-frequency generation has been used to study surface morphology, monatomic layers, and catalytic effects of adsorbed materials.

Third-Order Nonlinear Spectroscopy

In gas phase or isotropic solid and liquid media, the first nonlinear term is of third order and leads to a refractive index that is dependent on the field intensity $|\mathbf{E}|^2$:

$$\mathbf{P} = \varepsilon_0[\chi^{(1)}(\omega) + \chi^{(3)}(\omega)\mathbf{E}^2]\mathbf{E} \qquad [5]$$

Intensity-dependent refractive index effects lead to a wide range of nonlinear optical phenomena. In general, the polarization contains terms that are the product of three input fields generating Fourier components at sum and difference frequencies. Phase matching constraints will generally determine which of the possible interactions will be observed. A wide range of physical effects contribute to $\chi^{(3)}(\omega)$. The simplest of these nonlinear effects is saturation of the absorption.

Linear absorption is defined by the situation where the intensity decays exponentially with propagation distance x, according to Beer's Law:

$$I(x) = I(0)\exp\{-\alpha_0(\omega)x\} \qquad [6]$$

where $I(0)$ and $I(x)$ are the intensity values at $x = 0$ and x, respectively, and $\alpha_0(\omega)$ is the weak field absorption coefficient. The fractional power absorbed in this case is, therefore:

$$[I(0) - I(x)]/I(0) = \alpha_0(\omega)x \qquad [7]$$

and is independent of the intensity. When the incident intensity increases, until the rate of stimulated absorption from the lower level equals the spontaneous emission rate out of the upper level, the medium is said to be saturated. The intensity is

then I_{sat}. A simple rate-equation analysis shows that the absorption coefficient in the presence of a strong light field is:

$$\alpha(\omega) = \frac{\alpha_0(\omega)}{[1 + I/I_{sat}]} \qquad [8]$$

The absorption is then no longer independent of the light intensity.

Saturation Spectroscopy

In the case of a homogeneously broadened transition the effect of a saturating monochromatic light field is to reduce the absorption across the entire line profile. This is shown in Figure 1a where the frequency of the incident light w_L is detuned from line center ω_0. The degree of saturation and the effect on the absorption coefficient depends on the intensity I and the detuning $\Delta\omega = \omega_L - \omega_0$.

For a transition dominated by inhomogeneous broadening a monochromatic light field interacts strongly only with a sub-group of atoms. This sub-group is determined by those having a resonance frequency within the spectral range covered by the homogeneous width and centered on the incident frequency. Saturation in this case results in 'hole burning' as shown in Figure 1b. A dip in the absorption coefficient $\alpha(\omega)$ is produced centered on ω_L of width given by the homogeneous width.

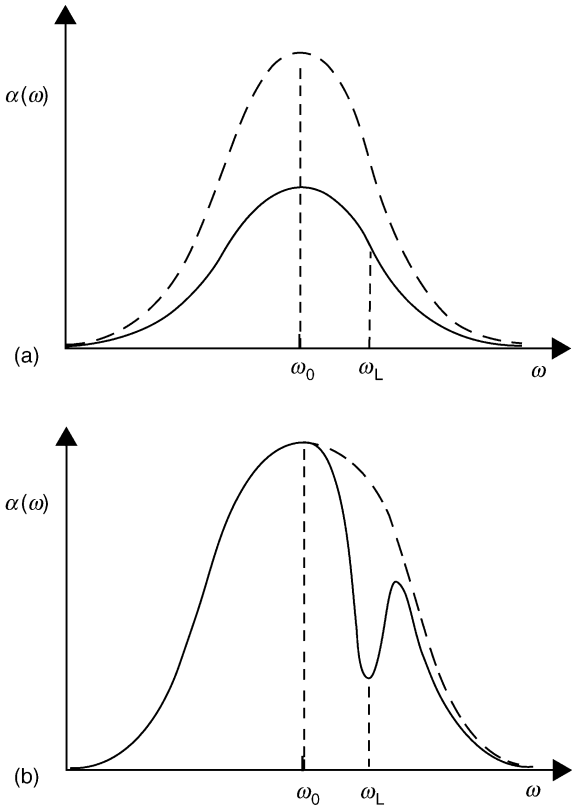


Figure 1 Absorption lineshapes for (a) homogeneously broadened and (b) inhomogeneously broadened lines centered at ω_0. Dashed lines in both figures show lineshape measured by weak probe. Solid lines show lineshape measured by weak probe in presence of a monochromatic, saturating beam at frequency $\omega_L \neq \omega_0$.

Saturation affects the population of both upper and lower levels involved in an optical transition and so affects both absorption and amplification (gain). The phenomenon shows itself as a narrow dip in the spectral profile of lasers operating on an inhomogeneously broadened gain line (e.g., a HeNe gas laser). These 'Lamb dips', as they are called, are the result of saturation by the laser mode which has a very narrow spectral width. Only those atoms in the Maxwell–Boltzmann velocity distribution are saturated, whose velocity results in a Doppler shift of the light that brings it into resonance, i.e., to within the atom's homogeneous linewidth of the resonance frequency. When the laser mode frequency is detuned from atomic line center two velocity subgroups of atoms, within the distribution, are saturated corresponding to positive and negative Doppler shifts for counter-propagating fields of the mode. When the mode frequency is tuned to atomic line center, the counter-propagating fields interact with the same (zero) velocity subgroup. The saturation effect is thereby enhanced, reducing the gain by depletion of the population inversion, leading to a dip in output intensity at line center. The width of the dip is the homogeneous width of the transition (Figure 2).

The variation of output intensity around the Lamb dip is used to provide an error signal in one scheme for laser frequency stabilization. A feedback system is employed to adjust the cavity length, L, in order to maintain the mode frequency, ν, exactly on atomic line-center ($\nu = mc/2L$ where m is an integer). This scheme uses gain saturation of the amplifying medium, whereas an alternative uses saturated absorption. In this scheme, a cell of gas having a narrow absorption line, within the gain profile of the laser, is placed inside the cavity. When the laser mode is tuned to absorption, line-center saturation reduces the absorption loss, resulting in a narrow peak in the laser emission. Drifting off this peak can be used as an error signal, as before, to stabilize the mode frequency.

Saturated absorption spectroscopy uses the same mechanism by placing the absorbing specimen outside the laser cavity. The effect is detected by monitoring the transmission of a weak probe beam in the presence of a strong counter-propagating 'pump' beam. Both probe and pump beams are

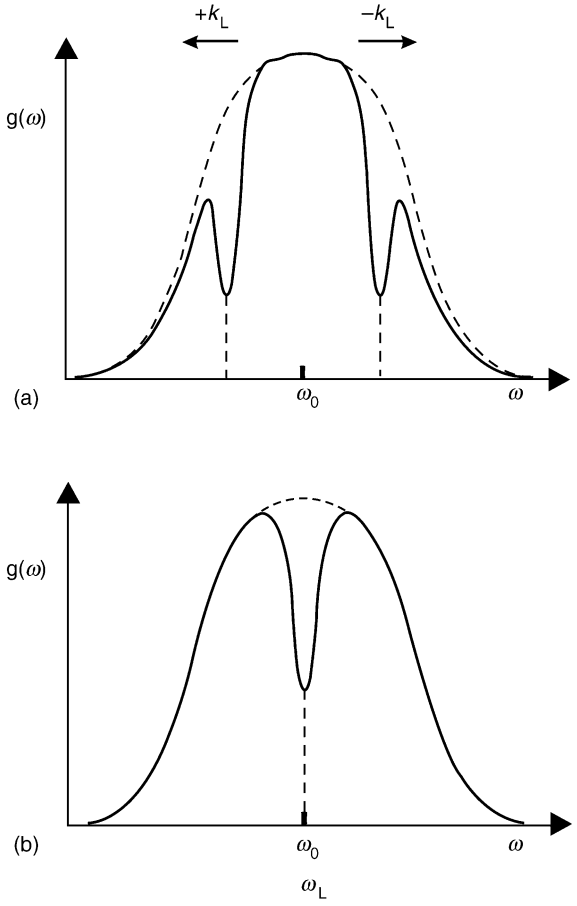


Figure 2 (a) Holes burned in gain profile by two counter-propagating monochromatic waves at $\omega_L \neq \omega_0$. (b) 'Lamb dip' in gain profile $g(\omega)$ (or absorption profile $\alpha(\omega)$) when $\omega_L = \omega_0$.

derived from the same laser using a suitable beamsplitter and, thus, have identical frequency. When tuned to line-center of the absorbing sample, the transmission of the probe increases owing to saturation by the strong pump beam interacting with the same subgroup of the inhomogeneously broadened medium (Figure 3).

The elimination of the first-order Doppler effect, in gaseous media, using this technique results in very high-resolution spectra. The residual linewidth is determined by homogeneous processes such as lifetime broadening, collisional (or pressure) broadening and power (or saturation) broadening. The latter effect is inevitable, as saturation is necessary, but may be controlled so as to contribute very little to the overall broadening. Power broadening becomes significant only when the laser intensity is appreciably in excess of the saturation intensity. Another limiting effect is transit time broadening: moving atoms spend only a limited time interacting with the narrow beams as they pass through. The consequent limit on the interaction time leads to conjugate broadening of the frequency of the transition as a result of the uncertainty principle.

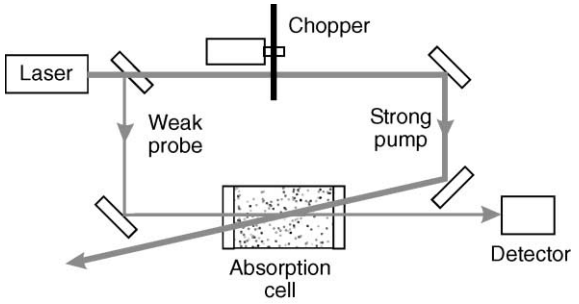


Figure 3 Schematic diagram of saturated absorption spectroscopy experiment. The saturating pump beam is chopped or modulated and the signal is processed using phase-sensitive detection to improve the signal-to-noise ratio.

The general problem of the spectral lineshape of signals in saturation spectroscopy is complex. Apart from the effects mentioned above, the lineshape will be affected by relativistic effects (second-order Doppler shifts), velocity-changing collisions, recoil effects from photon absorption and emission, velocity-dependent transit time broadening, optical pumping and laser lineshape effects.

Saturation spectroscopy is widely used in high-resolution spectroscopy for measurement of atomic and molecular transition frequencies, time and frequency standards using stabilized lasers, measurement of fundamental constants such as the Rydberg constant and the fine structure constant. Dynamical effects such as atomic collisions are also studied using saturation, whereby the relaxation of hole burning is measured in the time or frequency domain.

Polarization Spectroscopy

An important variation of saturation spectroscopy uses the change induced by a saturating beam in the real part of the complex refractive index, i.e., dispersion rather than absorption. Specifically, the technique of polarization spectroscopy employs a saturating beam that is circularly polarized to effect optical pumping on the degenerate lower level of a transition. The degenerate states may be identified, for example, with magnetic substates based on orientation of the intrinsic atomic or molecular angular momentum and labeled by the quantum numbers m_J. In equilibrium, the populations of all these degenerate substates will be equal and the optical properties of the system will then be isotropic. In particular, the polarization state of a linearly polarized probe beam will be unaffected by passage

through the medium. This probe is totally rejected by an analyzer (a linear polarizer set to transmit only light in an orthogonal plane). The linear polarized beam is a linear combination of two oppositely circular polarization states, σ^+ and σ^-, and both components experience the same complex refractive index. In the presence of a circularly polarized saturating pump beam, the distribution of population amongst the degenerate substates is disturbed from its equilibrium by optical pumping. The result will be an excess of population in higher or lower m_J states (a condition known as orientation) and this leads to optical anisotropy in the medium. The refractive index for σ^+ light is now different from that for σ^- light and a relative phase shift is introduced between these components of the linearly polarized probe resulting in elliptical polarization. A component of this elliptically polarized light will then be transmitted by the analyzer and is detected. When the strong pump beam is closely counter-propagating to the probe, then both beams interact with the same velocity subgroup as in Doppler-free saturation spectroscopy. The inhomogeneous (Doppler) broadening is thus significantly reduced. The technique provides a signal above an essentially 'zero' background and is thus capable of a higher signal-to-noise ratio than saturated absorption spectroscopy. How close to 'zero' the background is depends crucially on the degree of extinction by the analyzer. High-quality polarizer/analyzer devices can achieve extinction ratios of $1{:}10^{-8}$. Stress-induced birefringence in the windows of the cell containing the sample is a common limiting factor (Figure 4).

A probe laser of intensity I_{probe} is transmitted through a sample of length l with an intensity $I_{\text{PS}}(\omega)$ given by the following expression, that takes account of different contributions to the detected signal:

$$I_{\text{PS}}(\omega) = I_{\text{probe}}\left[\xi + \varphi^2 + b^2 + \frac{1}{2}\frac{b\Delta\alpha_{\pm}l}{(1+x^2)} + \frac{1}{2}\varphi\Delta\alpha_{\pm}l\frac{x}{(1+x^2)} + \left(\frac{1}{4}\Delta\alpha_{\pm}l\right)^2\frac{1}{(1+x)^2}\right] \quad [9]$$

where $x = 2(\omega_0 - \omega)/\gamma$ is the detuning from the resonance frequency ω_0 normalized by the halfwidth γ of the collision broadened line. The frequency-dependent terms provide absorptive (Lorentzian) and dispersive contributions to the total lineshape. The first three, frequency independent, terms in eqn [9] are responsible for a constant background signal. They originate from residual transmission of the polarizers (ξ), their accidental uncrossing (φ), and a possible residual birefringence (b) from optical elements between both polarizers in the probe beam path. ($\xi + b^2$), normally is in the range 10^{-6}. The differential absorption coefficient for left and right circularly polarized light, $\Delta\alpha_{\pm}$, at resonance ($x = 0$), is given by:

$$\Delta\alpha_{\pm}(x=0) = \Delta\alpha_{\pm}^0 = \alpha^+ - \alpha^- = \alpha_0 S_0 \Delta C_{JJ_1} \quad [10]$$

where α_0 is the unsaturated absorption coefficient, S_0 is the saturation parameter for the pump wave ($I_{\text{pump}}/I_{\text{sat}}$), and ΔC_{JJ_1} are Clebsch–Gordan coefficients for the respective transition and coupling case.

The dispersion-shaped component may be made to dominate by deliberately uncrossing the polarizer/analyzer combination to a small extent. This technique enhances the resolution of the transition frequency. Linewidths are limited by residual Doppler broadening owing to finite crossing angles of pump and probe beams, collisional broadening, transit time broadening, etc.

Applications of the technique have included Doppler-free spectroscopy of atomic transitions for precision measurements of fundamental constants, molecular spectroscopy and detection of atomic, molecular, and radical species in flames. By employing a different laser to accomplish the induced optical anisotropy from that used to probe the medium, molecular spectra can be simplified by the technique of 'polarization labeling'. Signals are obtained only when the probe and pump excite transitions sharing a common level and the particular transitions observed are sensitive to the selection rules for ΔJ, governing both pump and probe transitions. In this way P- and R-branch transitions can be distinguished.

Advantages of polarization spectroscopy include high sensitivity (10^2–10^3 times that of saturation spectroscopy), high signal-to-noise, simplification of

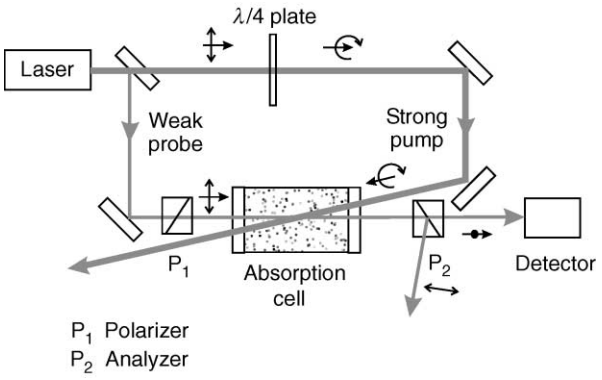


Figure 4 Schematic diagram of polarization spectroscopy experiment. Circularly polarized pump beam may be modulated for phase-sensitive detection of transmitted probe beam.

molecular spectra by polarization labeling, dispersion profiles for enhanced resolution, and for frequency stabilization techniques.

Multiphoton Absorption Spectroscopy

Multiphoton absorption is the term used to describe a process in which an atom or molecule makes a single transition between two of its allowed energy levels by absorbing the energy from more than a single photon. Sequential transitions, by which successive photons are absorbed via single-photon transitions between stationary energy levels, take place provided photons arrive within the lifetime of the intermediate state. Multiphoton transitions take place via 'virtual' energy levels. The lifetime of a virtual level is determined by the uncertainty principle to be the inverse frequency detuning from the nearest stationary energy level. Thus a high photon flux is usually required for multiphoton effects to be observed (Figure 5).

The simplest multiphoton effect, two-photon absorption, may be observed using strong atomic or molecular electronic transitions, even with modest laser power. The absorption is detected by observing the fluorescence from the spontaneous, radiative decay of the excited state. Two-photon absorption spectroscopy may be made Doppler-free by arranging absorption of one photon from each of two counter-propagating beams. The moving atoms experience opposite Doppler shifts $\mathbf{k}\cdot\mathbf{v}$ from each beam. The transition energy is thus:

$$E = (\hbar\omega + \mathbf{k}\cdot\mathbf{v}) + (\hbar\omega - \mathbf{k}\cdot\mathbf{v}) = 2\hbar\omega \qquad [11]$$

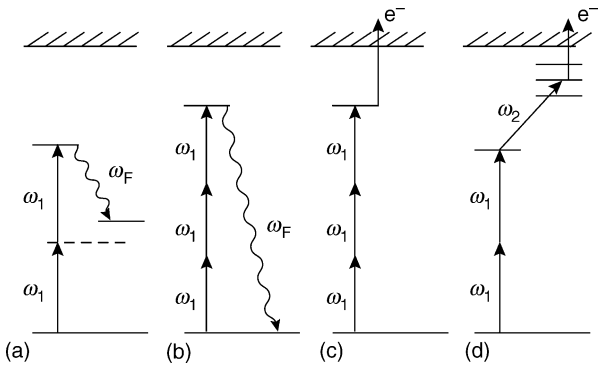


Figure 5 Multiphoton absorption spectroscopy. (a) Two-photon absorption using fluorescence detection of radiative decay at ω_F. (b)Three-photon absorption spectroscopy using fluorescence detection of radiative decay at ω_F. (c) Three-photon absorption spectroscopy using ionization detection following photoionization, collisional or field ionization. (d) Two-photon resonant multiphoton ionization spectroscopy; laser at frequency ω_1 is fixed and laser at ω_2 is tuned over transitions to high-lying levels.

Absorption of two photons from a single beam is prevented by choosing a transition between levels of the same parity and angular momentum quantum number, i.e., $\Delta J = 0$ and using circularly polarized light. Absorption of two σ^+ or σ^- photons gives $\Delta J = \pm 2$, which would be forbidden, e.g., on a n^2S–n'^2S transition in an alkali metal. This method of Doppler-free spectroscopy has the advantage that it involves all the atoms and not just the zero-velocity subgroup, as in saturation or polarization spectroscopy. In principle, it is also possible to use three-photon absorption, with the momentum of the wavevectors arranged to sum to zero, to achieve Doppler-free absorption on a transition between opposite parity states.

Instead of observing fluorescence following multiphoton absorption, the excited atom or molecule may be ionized by atomic collisions or by applying a d.c. electric field. Excitation of Rydberg levels (i.e., having high value of the principal quantum number n) by two- or three-photon absorption is readily accomplished. The sensitivity of this technique is enhanced by using two lasers. One laser at frequency ω_1 is tuned to a two-photon transition (energy $= 2\hbar\omega_1$) and the second laser ω_2 is tunable to excite transitions from the two-photon intermediate state to higher-lying levels (energy $= 2\hbar\omega_1 + \hbar\omega_2$). Thus, energy levels normally requiring excitation in the vacuum UV may be reached using visible frequency lasers. In such cases ionization detection may be more efficient than fluorescence since the nonradiative decay rate of these highly excited levels may exceed the fluorescence decay rate. The ion signal produced may be significantly amplified by the process of space charge amplification. A weak discharge established in the atomic vapor creates a space charge around the cathode. A single ion moving slowly through the space charge cloud acts to cancel the local field allowing many (up to 10^6) electrons to escape to the anode with a resulting amplification of 10^6 in the 'photoionization current'.

Multiphoton ionization spectroscopy is a sensitive method of recording highly excited energy levels. The sensitivity is enhanced in the technique of resonance ionization spectroscopy (RIS): two lasers are used such that the frequency of one laser is fixed close to a two-photon allowed transition and the second is tuned to excite a range of higher excited states. Absorption of an even number of photons allows excitation of levels having the same parity as the ground state and thus not accessible to normal absorption spectroscopy. In combination with a mass spectrometer to detect the ions produced by the multiphoton absorption resonance ionization mass spectrometry (RIMS), affords additional species

selectivity and sensitivity. RIMS finds application in high-resolution spectroscopy of rare isotopic species and is the basis for a scheme for isotope separation.

Coherent Third-Order Processes

An important class of spectroscopic techniques exploits the coherent interaction of three input waves via the nonlinear response of the medium. Specifically, the third-order term has the form:

$$\mathbf{P}_i^{(3)}(\omega_4, \vec{r}) = \chi_{ijkl}^{(3)}(\omega_1, \omega_2, \omega_3, \omega_4)\mathbf{E}_j(\omega_1, \vec{r}) \times \mathbf{E}_k(\omega_2, \vec{r})\mathbf{E}_l(\omega_3, \vec{r}) \quad [12]$$

where $\mathbf{E}_m(\omega_n)$ are complex field amplitudes and the subscript $m = i, j, k, l$ refers to each of the Cartesian coordinates x, y, z of the electric vector. $\chi_{ijkl}^{(3)}$ is a fourth-rank tensor. In condensed phase media, the third-order nonlinearity is readily observed, even with modest laser intensities. In the gas phase, although $\chi_{ijkl}^{(3)}(\omega)$ is small relative to $\chi^{(1)}$, its effect is enhanced when the incident fields are resonant and this nonlinear response may be used for spectroscopy. Equation [12] shows that the induced polarization couples three waves to radiate a fourth and describes a class of interactions known collectively as four-wave mixing.

The physics of four-wave mixing interactions may be visualized as the formation and read-out of a volume hologram. The three input beams are denoted as two pumps and one probe. The process involves the interference of one pump with the probe to form a spatially periodic grating pattern in the electric field distribution. The nonlinear response of the medium via $\chi^{(3)}$ leads to a spatial modulation of the refractive index – a laser-induced grating. The second pump beam generates a signal by scattering off this grating in a manner analogous to Bragg scattering of X-rays from regular crystal planes. The condition for generating a coherent scattered beam is equivalent to matching the momentum (or wavevectors) of the interacting photons, i.e., phase matching.

Degenerate Four-Wave Mixing (DFWM) Spectroscopy

The simplest case of four-wave mixing spectroscopy involves pump and probe beams of identical (i.e., degenerate) frequency to generate a signal at the same frequency. Spectra are generated as the frequency is tuned through atomic or molecular resonances where the interaction is strongly enhanced. In this degenerate case, the pump and probe interference pattern is stationary and the remaining pump beam scatters off this grating with the same frequency. As with all parametric processes, the interaction is phase matched by a suitable geometrical arrangement of the beams. A probe intersecting with counter-propagating pump beams leads to a signal that counter-propagates with the probe. This arrangement provides 'Doppler-free' lineshapes, since only those atoms with essentially zero velocity along the line of intersection of the beams contribute to the signal. Alternatively, pump and probe beams may be arranged to cross at an angle (usually small) in the forward direction – the 'forward geometry', with the resulting signal propagating at the same small angle to the incoming beams. Signals in this case are stronger, since all velocity classes may contribute but the lineshape is Doppler broadened (Figure 6).

An additional feature of the counter-propagating pump geometry is that the phase-matched signal arises from terms that involve products of fields with the complex conjugate of the incident

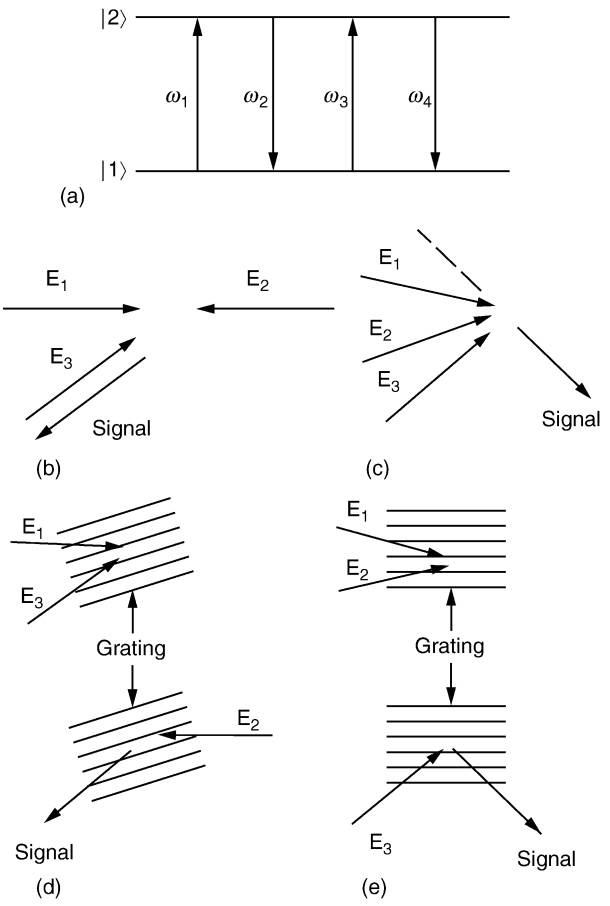


Figure 6 Degenerate four-wave mixing: (a) all input frequencies are degenerate, $\omega_1 = \omega_2 = \omega_3$ and give signal at the same frequency, ω_4. Phase matching geometries are shown in (b) counter-propagating pump (phase conjugate) and (c) forward geometry. (d) and (e) show the grating formation and read-out respectively.

probe field. The induced polarization radiating the signal wave has a component given by:

$$\mathbf{P}^{(3)}_{\text{signal}}(\omega) = \chi^{(3)}(\omega)\mathbf{E}_{\text{pump}}(\omega)\mathbf{E}_{\text{pump}}(\omega)\mathbf{E}^{*}_{\text{probe}}(\omega) \quad [13]$$

The backward propagating signal appears as a phase conjugated reflection of the incident probe wave. A probe wave that has propagated through a phase-distorting medium thus generates a signal with an initial phase distortion that will be undone as it retraces the path through the distorting medium. This phase conjugating property finds application in adaptive optics but also lends an advantage when DFWM is used as a spectroscopic probe in hostile or distorting environments, such as flames or plasmas. The counter-propagating pump geometry is often referred to as the phase conjugating geometry.

The intensity of the DFWM signal beam may be calculated using perturbation theory to derive $\chi^{(3)}(\omega)$ arising from saturated absorption. In the perturbation approximation, the probe beam is assumed to be weak and interacts with stationary atoms in the presence of counter-propagating, saturating pump fields. The signal is found from the phase-matched terms in the wave radiated from the induced polarization. When integrated over the atomic resonance line, the signal intensity, $I^{\text{int}}_{\text{sig}}$, is given by analytic expressions in the two limiting regimes of weak and strong saturation by the pump:

$$I^{\text{int}}_{\text{sig}} = \frac{|\mu_{12}|^8 N^2 k^2 T_1^2 T_2^3}{[1 + 4I_{\text{pump}}/I_{\text{sat}}(0)]^{5/2}} I^2_{\text{pump}} I_{\text{probe}} \quad [14]$$

when $I_{\text{pump}} \ll I_{\text{sat}}(0)$ and

$$I^{\text{int}}_{\text{sig}} \propto |\mu_{12}|^3 N^2 k^2 T_1^{-1/2} T_2^{1/2} I^{-1/2}_{\text{pump}} I_{\text{probe}} \quad [15]$$

when $I_{\text{pump}} \gg I_{\text{sat}}(0)$.

In these relations μ_{12} is the atomic dipole moment, N the atomic number density, $I_S(0) = \hbar^2/T_1 T_2 |\mu_{12}|^2$ is the line-center saturation intensity, and T_1 and T_2 are the longitudinal and transverse relaxation times, respectively. Extensions of the basic theory accommodate the effects of moving absorbers (Doppler effect) in both the phase-conjugate and forward phase matching geometry.

These results are derived using perturbation theory, by assuming that the probe field is weak. In practice the signal is optimized when pump and probe are of equal intensity and approximately equal to the saturation intensity. This situation violates the perturbation theory assumption but the signal intensity may be derived using nonperturbative analytical methods, although the resulting expression is more complicated. Alternatively, the quantum mechanical equations for the nonlinear susceptibility can be solved numerically to find the signal level for arbitrary intensity of the incident fields. Optimum signal-to-noise ratio in the spectra is achieved using approximately saturating intensities.

The sensitivity of resonant DFWM and the coherent nature of the signal beam suits the process to spectroscopy of nonfluorescing or trace molecular species, such as radicals created in chemical reactions. Equation [15] shows that the DFWM signal becomes relatively insensitive to collisional effects when the incident lasers saturate the transition. This feature of DFWM gives it an additional advantage over laser-induced fluorescence in optical diagnostic applications where quantitative evaluation of fluorescence intensity is hampered by collisional quenching. The state and species selectivity afforded by resonant DFWM is exploited in diagnostics of combustion and plasma processes, where the reduction of Doppler broadening allows high-resolution spectra from high-temperature gases. The relative intensity of spectral lines is used to derive temperature and concentration information with time and space resolution.

Additional selectivity is provided by using two input frequencies in two-color four-wave mixing (TC-FWM), to probe molecular excited state dynamics. The TC-FWM process uses resonant enhancement on two coupled transitions, i.e., the pump beams excite one transition and the probe and signal are generated on another transition having a common level with that excited by the pumps.

Coherent Raman Spectroscopy

The classical theory of spontaneous Raman scattering shows that a nonresonant pump of frequency ω_p wave can excite a resonance ω_m in a molecule with a nonlinear polarizability $\beta(x)$. (x is the spatial coordinate of the internuclear potential.) The wave resulting from this inelastic scattering is a Stokes wave of frequency $\omega_S = \omega_p - \omega_m$. The propagation of the Stokes wave is governed by the complex Raman susceptibility (**Figure 7a**):

$$\chi_{\text{Raman}}(\omega_s) = \chi'_{\text{Raman}}(\omega_s) + i\chi''_{\text{Raman}}(\omega_s) \quad [16]$$

This generated Stokes wave may then couple with the pump to drive the molecular resonance coherently leading to amplification of the Stokes wave at the expense of the pump. This process of stimulated Raman scattering (SRS), provides gain if the amplification exceeds the loss due to absorption, scattering, etc. There is, therefore, a definite threshold intensity that the pump must satisfy to observe SRS. The SRS

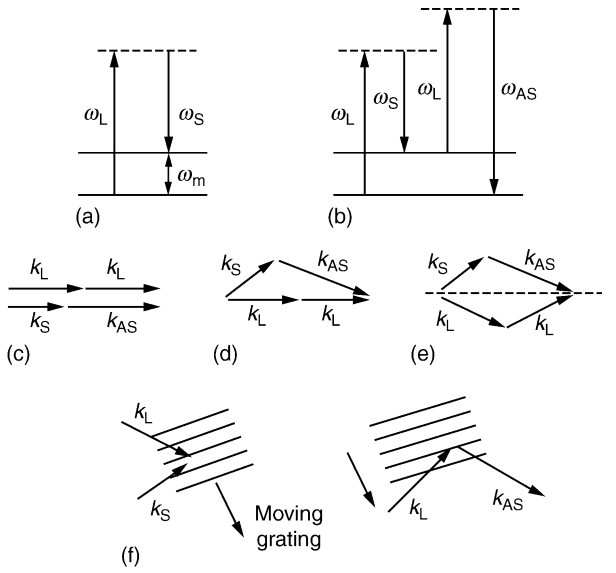


Figure 7 Coherent Raman spectroscopy. (a) Spontaneous Raman scattering of input pump ω_L producing Stokes wave ω_S in inelastic scattering. (b) Coherent anti-Stokes Raman scattering (CARS). Phase matching geometries for CARS is shown for a nondispersive medium in (c) and dispersive media in (d) and (e). The planar boxcars geometry in (e) may be folded along the dotted line (folded boxcars geometry) to separate the signal into an orthogonal plane from the pump beams to reduce scattered pump light reaching the detector.

gain $g(\omega_s)$ at the Stokes frequency ω_s is given by:

$$g(\omega_s) = -\frac{k_s}{2n_s^2}|E(\omega_L)|^2\chi''_{\text{Raman}}(\omega_s) \qquad [17]$$

where k_s and n_s are the wavevector and refractive index for the Stokes wave, $E(\omega_L)$ is the field amplitude at the pump laser frequency ω_L, and $\chi''_{\text{Raman}}(\omega_s)$ is the imaginary part of the complex Raman susceptibility:

$$\chi''_{\text{Raman}}(\omega_s) = \frac{-\varepsilon_0 N\left(\dfrac{\partial\beta}{\partial x}\right)^2\gamma}{32m_e\omega_v\left\{(\omega_m-(\omega_L-\omega_S))^2+\dfrac{\gamma^2}{4}\right\}} \qquad [18]$$

In this expression $\partial\beta/\partial x$ is the variation of polarizability β with internuclear displacement x, N is the molecular population density, and γ is the linewidth due to damping. Spectroscopy based on SRS provides signals that are many orders of magnitude stronger than those from spontaneous Raman scattering, and is widely used to study spectra of complex molecules having no absorption spectra in the visible region.

A four-wave mixing interaction, involving a pump and probe beam of different frequencies, generates a nonstationary grating. The signal generated by scattering of a second pump beam off this moving grating is Doppler-shifted to lower or higher frequency. These signal beams correspond respectively to Stokes or anti-Stokes waves in Raman scattering. A probe beam at the Stokes frequency will thus interact with a pump beam to generate a coherent signal at the anti-Stokes frequency; a process known as coherent anti-Stokes Raman scattering (**Figure 7b**). A probe at the anti-Stokes frequency correspondingly gives the process of coherent Stokes Raman scattering (CSRS). Phase matching is necessary for both CARS and CSRS and is satisfied by an appropriate arrangement of the incident wavevectors (**Figure 7c,d** and **e**).

CARS has proven to be a valuable spectroscopic technique in a wide variety of applications. A CARS spectrum is generated by scanning the frequency of the probe, i.e., the Stokes laser. The signal strength is given by

$$I_4 = I_{\text{CARS}} = \frac{\omega_4^2}{n_1^2n_2n_4c^4\varepsilon_0^2}I_1^2I_2|\chi_{\text{CARS}}|^2l^2 \times\left(\frac{\sin(\Delta kl/2)}{\Delta kl/2}\right)^2 \qquad [19]$$

where I_1 and I_2 are the intensity of the pump and Stokes beams, respectively, l is the interaction length, and n_i are the refractive indices at frequency ω_i. The maximum signal is generated when the phase mismatch $\Delta k = 0$. The third-order susceptibility, $\chi^{(3)} = \chi_{\text{CARS}}$, in general consists of a nonresonant part $\chi^{(3)}_{\text{nr}}$, which is independent of the frequency of the exciting beams, and a Raman-resonant contribution $\chi^{(3)}_{\text{r}}$:

$$\chi^{(3)} = \chi^{(3)}_{\text{nr}} + \chi^{(3)}_{\text{r}} \qquad [20]$$

The lineshape in coherent Raman spectroscopy is affected by interference between resonant and nonresonant contributions. This effect is exploited to derive the concentration of the resonant (Raman active) species relative to that of the background of known density. A more common application of CARS, however, is for temperature measurement. The dependence of the signal intensity on $|\chi_{\text{CARS}}|^2$, and therefore on the square of the molecular population of the initial state, yields a spectrum from which the temperature may be derived by assuming a Maxwell–Boltzmann distribution over the state populations.

A particularly useful application of CARS employs a narrow linewidth pump laser and a

broad bandwidth Stokes laser. CARS signals are generated simultaneously on all frequencies such that the corresponding Stokes frequency is available from the broadband probe. The spectrum recorded with a single laser pulse can thus provide population information on a wide range of molecular states, giving time- and space-resolved measurement of temperature and concentration.

CARS and CSRS spectroscopy is applied to both gas and condensed phase media. Owing to the higher density in the latter, signals may be strong in spite of the two-photon nature of the resonant interaction.

Time Domain Four-Wave Mixing Spectroscopy

Four-wave mixing spectroscopy may also be executed in the time domain. The three incident waves may arrive as pulses separated in time by variable delays. The interaction is, therefore, mediated by the atom or molecule's induced polarization or coherence. DFWM, CARS, and CSRS processes may thus be used to study ultrafast relaxation since the signal intensity gives a measure of the coherence decay in the intervening delay between the pulses. These processes are intimately related to coherent transient spectroscopy using measurements of free induction decay, photon echoes, and stimulated photon echoes.

Laser-Induced Grating Spectroscopy

The four-wave mixing processes discussed above involve a spatial modulation of the molecular population and a consequent modulation of the refractive index for resonantly interacting waves. The signals are radiated by the coherent oscillations of the molecular populations that decay rapidly. Relaxation processes result in transfer of energy from the excited molecular states, leading to a temperature and density grating in the bulk medium with a period Λ, given by:

$$\Lambda = \lambda_{\text{pump}}/(2 \sin \theta_1) \qquad [21]$$

where θ_1 is the angle between the interfering beams of wavelength λ_{pump}. A nonresonant wave of wavelength λ_{probe} may be scattered by this laser-induced grating provided it is incident on the grating at the angle θ_2 satisfying the Bragg condition:

$$\theta_2 = \sin^{-1}(\lambda_{\text{probe}}/2\Lambda) \qquad [22]$$

The periodic temperature/density grating, when induced by a short duration laser pulse, is accompanied by two oppositely propagating acoustic waves. The reflectivity of Bragg scattering is temporally modulated as the acoustic waves traverse the stationary grating at the local sound speed. Measurement of the period of these oscillations yields the sound speed from which the temperature may be derived. The stationary grating and acoustic waves decay by diffusive and damping processes, dependent upon the bulk medium properties. Measurement of the decay time of the grating induced in gas phase media yields the gas pressure if the gas dynamic parameters are known (Figure 8).

Laser-induced grating spectroscopy (LIGS), has been widely employed for studies of condensed phase relaxation processes. Applications have also been developed for gas phase studies using resonant absorption to enhance the thermal grating component. A grating may, however, also be induced in response to the electrostrictive effect of the high electric field in the interference pattern of two intense pump waves. Such electrostrictive gratings do not require resonant absorption and their spectroscopic utility centers on the study of relaxation effects in the molecular medium.

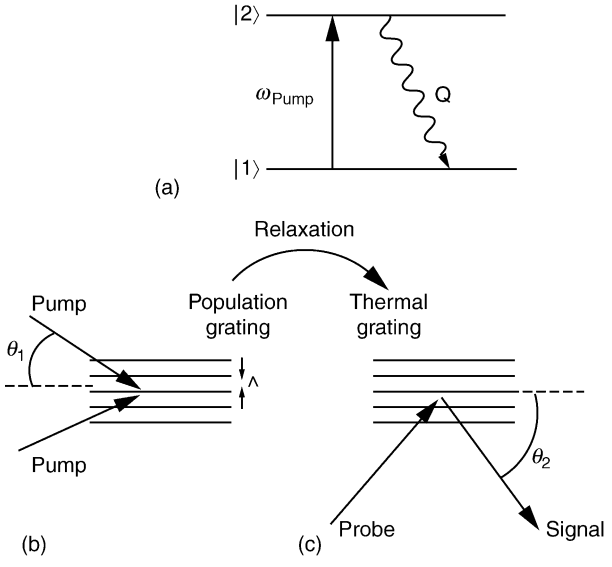


Figure 8 Laser-induced thermal grating spectroscopy. (a) Absorption of energy from pump lasers creates excited molecular population in grating pattern shown in (b). Spacing of laser induced grating Λ is set by angle of intersection of pumps. Quenching, Q, of excited population by collisions transfers heat to bulk medium and forms thermal grating. (c) Probe incident at Bragg angle θ_2 is scattered to form signal.

List of Units and Nomenclature

Formulae are expressed in SI units

c	Speed of electromagnetic wave *in vacuo*
$\mathbf{E}$	Electric field amplitude
$g_0(\omega)$, $g(\omega)$	Gain coefficient at line center, at frequency ω
h	Planck's constant
I	Intensity of electromagnetic wave
I_{sat}	Saturation intensity of medium
$\mathbf{k}$, k	Wave vector, scalar value of wave vector ω/c
m_e	Mass of electron
n_m	Refractive index of medium for mth beam
N	Atomic or molecular population density
$\mathbf{P}$	Polarization field amplitude
T_1	Longitudinal relaxation time
T_2	Transverse relaxation time
α_0, $\alpha(\omega)$	Absorption coefficient at line center, at frequency ω
$\beta(x)$	Molecular polarizability as function of internuclear separation x
ε_0	Permittivity of free space
μ_{12}	Dipole moment for transition between energy levels 1 and 2
ν	Frequency of electromagnetic wave
$\chi^{(n)}$	nth order susceptibility
ω	Angular frequency of electromagnetic wave

See also

Coherent Transients: Coherent Transient Spectroscopy in Atomic and Molecular Vapours; Foundations of Coherent Transients in Semiconductors; Ultrafast Studies of Semiconductors. **Fiber and Guided Wave Optics:** Nonlinear Optics. **Imaging:** Adaptive Optics. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation; Four-Wave Mixing; Kramers–Kronig Relations in Nonlinear Optics. **Phase Control:** Phase Conjugation and Image Correction. **Physical Applications of Lasers:** Sum-Frequency Generation at Surfaces. **Scattering:** Raman Scattering; Stimulated Scattering. **Spectroscopy:** Raman Spectroscopy; Second-Harmonic Spectroscopy.

Further Reading

Clark RJH and Hester RE (eds) (1988) Quantitative CARS spectroscopy. In: *Advances in Non-Linear Spectroscopy*, pp. 193–251. New York: John Wiley.

Corney A (1986) *Atomic and Laser Spectroscopy*. Oxford, UK: Oxford University Press.

Demtröder W (1996) *Laser Spectroscopy*, 2nd edn. Berlin: Springer-Verlag.

Dreier T and Ewart P (2002) Coherent techniques for measurements with intermediate concentrations. In: Kohse-Höinghaus K and Jeffries JB (eds) *Applied Combustion Diagnostics*. London: Taylor and Francis.

Eesley GL (1981) *Coherent Raman Spectroscopy*. London: Pergamon.

Eichler HJ, Günter P and Pohl DW (1986) *Laser Induced Dynamic Gratings*. Berlin: Springer-Verlag.

Feld MS and Letokhov VS (eds) (1980) *Coherent Nonlinear Optics, Recent Advances*. Berlin: Springer-Verlag.

Fisher RA (ed.) (1983) *Optical Phase Conjugation*. San Diego, CA: Academic Press.

Levenson MD and Kano SS (1987) *Introduction to Nonlinear Laser Spectroscopy*. Revised edn. San Diego, CA: Academic Press.

Raman Spectroscopy

R Withnall, University of Greenwich, Chatham, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Inelastic light scattering, the optical analogue of Compton scattering, had been predicted to occur by Smekal in 1923, but it was Chandresekar Venkataraman Raman who provided the first experimental demonstration of the phenomenon in February 1928. Only a few months later in 1928 the Russian scientists, Landsberg and Mandelstam observed inelastic light scattering from a quartz crystal. In spite of initial claims to the contrary, this was the same effect that had been observed by Raman, although the Russian scientists did not acknowledge this, preferring to call the effect 'combinatorial scattering'. It was not until the 1970s that the effect became known as Raman scattering everywhere, including Russia.

As typically practised, Raman spectroscopy involves laser excitation of a sample and measurement of the wavelength and intensity distribution of the scattered Raman light. When a microscope is used for delivering laser excitation and/or collecting the inelastically scattered light, the technique is referred to as Raman microscopy. However, it is important to

recognize that, apart from the differences in sampling configuration, there are no fundamental differences between Raman microscopy and Raman spectroscopy; the terms merely identify the different sampling techniques.

Since its invention in 1960, major advances in the technology of the laser have occurred, resulting in the wide variety of laser light sources that are currently available for Raman spectroscopy. Characteristic properties of laser systems for Raman microscopy will be described in this chapter, following an introduction to the technique and its applications in the next section below. Then, the commonly used laser sources for Raman microscopy are categorized according to the wavelength regions of their emissions, along with their merits and drawbacks for each application. The chapter will consider mainly continuous wave laser sources, because pulsed laser beams that are tightly focused with microscope optics give very high irradiance that would destroy most samples.

Raman Microscopy

The first experimental Raman microscopes were developed in the mid-1970s by two independent groups, and it was not long before the first generation of Raman microscopes were subsequently commercialized. Many of these early instruments simply consisted of commercial optical microscopes coupled to Raman spectrometers, as shown in Figure 1. It was recognized that the use of a microscope for sampling in this way can facilitate the Raman examination of tiny particles of micrometer dimensions.

The optical lay-out in Figure 1 shows an infinity corrected microscope. The advantage of using this, rather than a standard tube length microscope, is that its length can be extended in order to incorporate additional optical elements that are required for coupling to the Raman excitation and collection optics. In the optical layout, the incident laser beam is spatially filtered by a pinhole, D_1, in order to remove the diffraction rings and give a point source. The laser light is then partially reflected by beamsplitter, B_1, and focused on to the sample by the microscope objective. The back-scattered Raman radiation is collected by the same microscope objective, partially transmitted by the beamsplitter, and directed into the entrance slit of the Raman spectrometer by means of coupling optics. The accurate location of an aperture, D_2, at the focal point improves the spatial resolution and allows depth profiling of transparent samples. The two pinholes, D_1 and D_2, act as spatial filters and are referred to as confocal diaphragms. The Raman microscope is described as being confocal, because out-of-focus light, collected from outside the focal volume, is rejected.

The use of the beamsplitter, in the optical path of early Raman microscope designs, gives rise to significant losses of the precious Raman

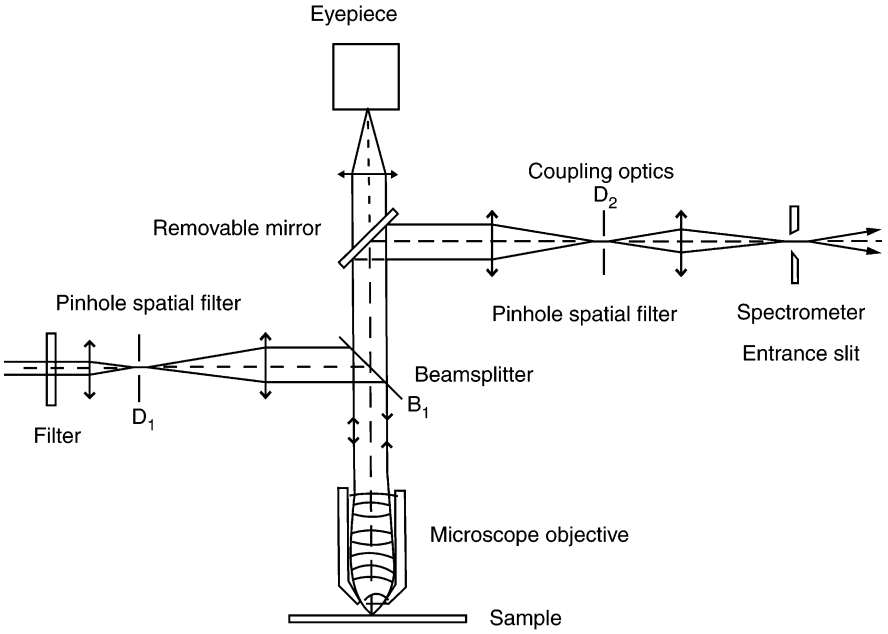


Figure 1 Diagram of the optical layout of a visible Raman microscope (Reproduced from Turrell G, Delhaye M and Dhamelincourt P (1996) Characteristics of Raman microscopy. In: Turrell G and Corset J (eds) *Raman Microscopy. Development and Applications*, 2. London: Academic Press).

scattered light. If a 50:50 beamsplitter is used then half of the incoming laser light is lost by transmission at B_1 and, more significantly, half of the Raman scattered light is lost by reflection at B_1. As long as a higher laser power can be employed, it is better to use a 90:10 beamsplitter that transmits 90% of the light and reflects 10%; then, only 10% of the incoming laser light is directed by B_1 on to the sample, but 90% of the Raman scattered light is transmitted by B_1 towards the detector and only 10% of the Raman light is lost by reflection at B_1. The efficiency of a beamsplitter can be defined as the product of the reflectivity at the laser wavelength and the transmission at the Raman-shifted wavelength. Thus, conventional 50:50 and 90:10 beamsplitters would have efficiencies of 25% and 9%, respectively. Recently, however, holographic notch filters have been developed which can be used as higher efficiency beamsplitters. These typically have a laser rejection contrast ratio of better than 10^4, reflecting ca. 90% of the laser excitation and transmitting ca. 90% of the Raman scattered light, thus giving an efficiency of ca. 81%. Due to these considerations, higher throughput, commercial Raman microscope designs usually incorporate a holographic notch filter.

The importance of Raman microscopy stems from the fact that it is the only microanalytic method available today, by use of which it is possible to identify, or characterize, small particles of micrometer dimensions *in situ*. It is also advantageous that no sample preparation is required when performing Raman microscopy.

The fundamental limit of the lateral spatial resolution (i.e., for a diffraction limited focus) is the separation at which the maximum of one Airy disc function just touches the first minimum of an adjacent Airy disc, given by:

$$\text{Lateral spatial resolution} = \frac{1.22\lambda}{2\text{NA}} \qquad [1]$$

where λ is the wavelength of the light and NA is the numerical aperture of the microscope objective.

The axial Airy disc function determines the resolution in the longitudinal direction. A good estimate of the axial resolution limit for low numerical aperture is given by:

$$\text{Axial spatial resolution} = \frac{2\lambda}{(\text{NA})^2} \qquad [2]$$

When the sample is heterogeneous and exhibits fluorescence that is not evenly distributed within the sample, a region of the sample can be selected with the microscope that shows the minimum amount of fluorescence. If the fluorescence is intrinsic to the sample itself, then it is possible to use the shift subtract technique and/or temporal discrimination between the fluorescence and the Raman scattering. The latter can be achieved by Kerr gate fluorescence rejection, in which a pulse of light is used to close a Kerr gate shutter. The problem with this approach, however, is that it uses a laser pulse, and even pulses of modest energy will have extremely high irradiance when focused to a tiny spot by a microscope objective. Such laser pulses would inevitably destroy the sample under a microscope.

Another approach for reducing fluorescence is to use near infrared excitation, which is low enough in energy so that absorption cannot occur to promote electronic transitions. The technique of Fourier Transform Raman (FTR) spectroscopy often offers the best chance of obtaining Raman spectra from fluorescent samples for this reason. Low energy excitation of wavelength equal to 1064 nm, provided by a Nd:YAG laser, is normally used for FTR. Disadvantages are that the Raman scattering efficiency is low relative to that obtained with visible excitation, due to the ν^4 dependency (ν^4 refers to Rayleigh scattering of the Raman light). This is compensated to some extent by employing a high throughput (Jacquinot advantage) interferometer. However, this works better for macro rather than micro samples. This is because the coupling of a microscope to an FT Raman spectrometer has a fundamental drawback; the microscope is throughput limited, so the high throughput advantage of the FTR spectrometer cannot be realized in FTR microscopy. The image of the Jacquinot stop (the large circular aperture which is the entrance to the interferometer) at the sample plane typically has a diameter of a few hundred μm, which is much larger than particles with diameters of ca. 1 μm having similar dimensions to the waist diameter of the 1064 nm light excitation at its diffraction-limited focus. Consequently the throughput advantage of the interferometer is not fully exploited and there is a trade-off of spatial resolution with signal-to-noise at the detector. For this reason, the typical spatial resolution that is achieved is in the range of 15–100 μm rather than 4 μm, as claimed in the literature. Bruker has commercialized an FTR microscope by coupling an optical microscope to an FTR spectrometer with near infrared optical fibers.

An approach that offers more promise for reducing fluorescence and achieving spatial resolution close to the diffraction limit is to use near infrared excitation in conjunction with dispersive Raman microscopy. For example, semiconductor lasers operating from 785 to 852 nm can be used in conjunction with sensitive multichannel silicon-based CCD arrays. If longer excitation wavelengths are necessary, in order to

overcome the problems with fluorescence, then diode lasers can be used that emit at longer wavelengths in conjunction with dispersive Raman spectrometers equipped with multichannel near infrared (NIR) detectors, e.g., Ge- or InGaAs-array detectors.

The applications of Raman microscopy cover many areas, including material science, the earth sciences, environmental science, biology and medicine, forensic science, and even the analysis of artworks. These areas are too wide ranging to be described in detail here, but the interested reader is referred to the Further Reading section at the end of this article.

Characteristics of Laser Sources

The laser is an excellent light source for Raman spectroscopy to such an extent that the terms 'Raman spectroscopy' and 'laser Raman spectroscopy' are synonymous for all but the most esoteric experiments, for example, with synchotron sources. Indeed, the advent of the laser was the stimulation for the renaissance of Raman spectroscopy in the 1960s, given its special properties, such as monochromaticity, high intensity, beam collimation, and coherence.

The characteristics of laser light and the advantages it offers to Raman microscopy are now considered.

Beam Quality

The light inside a laser tube is formed from a number of standing waves having distinct vibrational modes. There are a limited number of these modes transverse to the beam and these are characterized by the TEM_{pq} number (where p and q can be 0, 1, 2,...) where TEM is an acronym for 'transverse electromagnetic'. When a laser is operating in its fundamental TEM_{00} mode, light rays are reflected on axes between the end mirrors of the laser cavity. The '00' indicates that there are no nodes in the beam profile (Figure 2a), and the laser beam has a Gaussian intensity profile in the radial direction:

$$I(r) = I_0 \exp\left(-\frac{2r^2}{w^2}\right) \qquad [3]$$

where $I(r)$ is the irradiance at a radial distance r from the axis of the beam, I_0 is the axial irradiance, and w is the beam radius.

For higher-order modes, a number of nodes are observed in the beam profile, which arise from off-axis light rays being reflected between the end mirrors, for example, the TEM_{11} mode has two nodes which are mutually perpendicular (Figure 2b).

The propagation characteristics of a Gaussian beam can be fully defined, either by the diameter of the beam waist or by the far-field divergence.

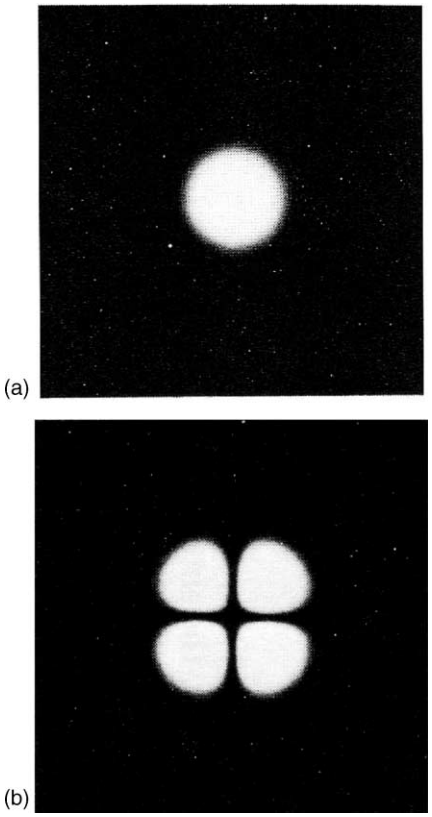


Figure 2 Transverse electromagnetic modes formed with confocal concave mirrors, (a) TEM_{00}, and (b) TEM_{11}. (Reproduced with permission from Young M (1977) *Optics and Lasers: an Engineering Approach.* Berlin, Heidelberg: Springer-Verlag

Consequently, it is only necessary to know the diameter of the beam waist ($2w_0$), or the diameter of the beam ($2w_z$) at a longitudinal distance z from the waist, in order to determine the propagation characteristics of the beam:

$$w(z) = w_0\sqrt{1 + \left(\frac{z}{z_R}\right)^2} \qquad [4]$$

$$R(z) = z\left[1 + \left(\frac{z_R}{z}\right)^2\right] \qquad [5]$$

where the quantity $z_R = \pi w_0^2/\lambda$ is known as the Rayleigh range of the beam, λ is the wavelength of the laser radiation, and $R(z)$ is the radius of curvature of the wavefront at a distance z from the beam waist.

The wavefront is planar at the minimum beam waist and the Rayleigh range is the distance from the beam waist to the location at which the wavefront is most curved (Figure 3), the region from the waist to

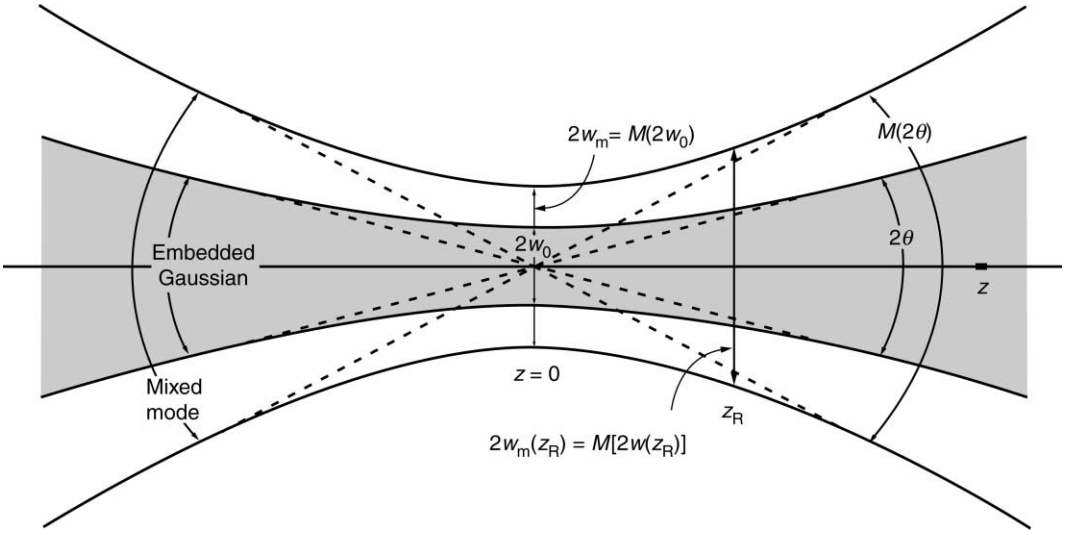


Figure 3 The Rayleigh range and embedded Gaussian (adapted from http://beammeasurement.mellesgriot.com/tut_m2.asp).

the Rayleigh range being the near field. Beyond approximately ten times the Rayleigh range, in the far field, the beam diverges as a cone with approximately straight sides. It can be seen by substituting $z = z_R$ into eqn [4] that the beam diameter at the Rayleigh range is $\sqrt{2}$ times the waist diameter.

Unfortunately laser beams do not conform to pure Gaussian functions and therefore they do not propagate according to the above equations. As a result, a dimensionless beam propagation parameter was devised in the early 1970s. This parameter, known as the M^2 factor or the 'times diffraction limit', is based on the brightness theorem, which states that for any laser beam the product of the beam diameter and the far-field divergence is a constant. Thus, M^2 is defined as the ratio of the laser beam's multimode diameter-divergence product to the ideal diffraction-limited (TEM_{00}) beam diameter-divergence product:

$$M^2 = \left(\frac{2w_m \, \Theta_m}{2w_0 \, \theta_0} \right) \qquad [6]$$

where Θ_m is the laser beam's multimode divergence, θ_0 is the theoretical diffraction-limited divergence, $2w_m$ is the laser beam's multimode waist diameter, and $2w_0$ is the ideal diffraction-limited beam waist diameter.

Alternatively, M^2 can be defined as the square of the ratio of the multimode beam diameter ($2w_m$) to the diffraction-limited beam diameter ($2w_0$):

$$M^2 = \left(\frac{2w_m}{2w_0} \right)^2 \qquad [7]$$

The intensity of the beam has a Gaussian distribution in the radial direction, but the accepted definition of beam diameter is the distance at which the intensity has fallen to $1/e^2$ (i.e., ca. 13.5%) of its axial intensity, as can be seen by setting $r = w$ in eqn [3] above. This definition only applies properly to Gaussian beams; for other beam profiles the diameter can be calculated using a second moment measurement of the entire beam.

The real beam, then, can be treated as Gaussian by substituting an artificial wavelength $M^2\lambda$ into the equations that apply to an ideal diffraction-limited TEM_{00} beam. The ideal TEM_{00} Gaussian can be superimposed as an 'embedded Gaussian' on the optical diagram of the real multimode beam (Figure 3). It can be seen that the multimode beam has a waist diameter and a divergence that are both M times larger than those of the embedded Gaussian.

The M^2 factor has a value of unity for an ideal diffraction-limited TEM_{00} beam, and values of greater than unity for all other beams; it can be as high as several hundred for a distorted beam of poor quality. This is an important parameter where a tightly focused laser spot (e.g., confocal Raman microscopy) or low divergence (e.g., remote sensing) are required.

Polarization

Although polarization is not an inherent property of laser light, lasers provide plane polarized light when the end windows of the laser tube are mounted at Brewster's angle (θ_B):

$$\theta_B = \tan^{-1} n \qquad [8]$$

where n is the refractive index of the window material for the appropriate wavelength of laser light. For this angle, reflected light is completely plane polarized and consequently the laser light that is generated is also plane polarized. Brewster windows are often used inside the laser cavity in order to reduce reflection losses. This is important in low-gain lasers such as HeNe lasers where laser action could be prevented by reflection losses. Air-cooled ionic gas lasers typically deliver highly linearly polarised output with ratios exceeding 1000:1.

The polarization property of the exciting light is used a great deal in Raman spectroscopy. In combination with the polarization properties of the Raman scattered light, it can be used to determine the depolarization ratios, ρ, of Raman bands in the solution state. In solid state studies it can be used to determine the degree of orientation, for example, in polymeric fibers or films, and in single crystal studies it can be used to determine the symmetry classes of the Raman active vibrations. In such studies a linear polarizer is used to improve the degree of plane polarization of the laser light, and a half-wave plate is used to rotate the plane of polarization by 90°.

Plane polarized laser light can be converted to circularly polarized laser light which is required for Raman optical activity (ROA) studies. For example, plane polarized 514.5 nm laser light provided by an argon ion laser can be converted to circularly polarized light by an electro-optic modulator. As CCD detectors are usually employed in present-day ROA experiments and they have relatively long sampling times, slow polarization modulation is required which can be achieved by periodically rotating a quarter-wave plate or by applying a square quarter-wave voltage to a Pockels cell.

Whereas optical isomers or enantiomers give identical Raman spectra, they are distinguished in the ROA experiment by the sign of the ROA signal. Indeed, the sign can be used to determine the absolute configuration of a chiral molecule. Also the enantiomeric excess of mixtures of enantiomers can be determined in such experiments, with obvious applications in the pharmaceutical industry, and the conformations of biological molecules, for example, proteins, nucleic acids, sugars, and viruses, can be determined.

Up until now, the ROA technique has been practised by no more than a handful of research groups world-wide, but it is expected that this situation will change in the near future, since the first commercial ROA spectrometer (the ChiralRAMAN spectrometer) has recently been launched by Biotools Inc. This spectrometer makes use of a solid state laser as the source of the polarized laser light.

Beam Divergence

Referring to Figure 3, the full divergence angle, Θ, for the fundamental TEM_{00} Gaussian beam is given by:

$$\Theta = \lim_{z \to \infty} \frac{2w(z)}{z} = \frac{2\lambda}{\pi w_0} \qquad [9]$$

From eqn [4] it can be seen that the spot size increases linearly with the longitudinal distance z and, from eqn [9], that it diverges at a constant cone angle, Θ. It can also be seen from eqn [9] that the smaller the spot size, w_0, the greater the beam divergence angle, Θ.

Indeed, a highly divergent beam is a problem with edge-emitting laser diodes due to the small dimensions of the light source. Furthermore, it is much smaller in the vertical direction than in the horizontal, giving rise to an elliptical output beam that diverges much more rapidly in the vertical direction than in the horizontal. The highly divergent, elliptical beam can be corrected, to an extent, with a cylindrical lens, but the inherent problem of a small, elliptical source can never be completely corrected.

In contrast, vertical-cavity surface-emitting semiconductor lasers (VCSELs) do not have the same limitations on beam divergence because the cavity is in the vertical direction and the light emission occurs from the surface, which has a much larger area than the light source of an edge emitter. Indeed, frequency-doubled green and blue semiconductor lasers are now available, having beams with M^2 values of less than 1.1 and beam divergences of a few milliradians. In terms of these beam characteristics, it appears that these lasers are beginning to challenge the argon ion laser.

Emission Linewidth

The laser linewidth limit, $\Delta\nu_L$, is given by the Schawlow–Townes expression:

$$\Delta\nu_L = \frac{2\pi h\nu(\Delta\nu_c)^2}{P_{out}} \qquad [10]$$

where $\Delta\nu_c$ is the linewidth of the passive resonator, $h\nu$ is the photon energy, and P_{out} is the laser power output.

The power output, P_{out}, is equal to the number of photons, N_p, in the resonator times the energy per photon, $h\nu$, divided by the photon lifetime, τ_c:

$$P_{out} = \frac{N_p h\nu}{\tau_c} \qquad [11]$$

Furthermore, the width $\Delta\nu$ of the resonance curve, at which the intensity has fallen off to half the maximum

value, is given by:

$$\Delta\nu_c = (2\pi\tau_c)^{-1} \quad [12]$$

Substituting for P_{out} and $\Delta\nu_c$ from eqns [8] and [9] into eqn [10] gives the limit for the linewidth, $\Delta\nu_L$:

$$\Delta\nu_L = \frac{\Delta\nu_c}{N_p} \quad [13]$$

The linewidth of a laser is characterized by its Q factor:

$$Q = \frac{\nu}{\Delta\nu} \quad [14]$$

where ν is the frequency of the laser line and $\Delta\nu$ is the linewidth.

Other useful relationships involving linewidth parameters are:

$$Q = \frac{\lambda}{\Delta\lambda} \quad [15]$$

and

$$\Delta\tilde{\nu} = -\frac{\Delta\lambda}{\lambda^2} \quad [16]$$

where λ and $\Delta\lambda$ are the wavelength and linewidth (in units of metres), respectively, and $\Delta\tilde{\nu}$ is the linewidth expressed as a wavenumber (in units of cm^{-1}).

The Q factor can be as high as 10^8 which is of great use in high resolution spectroscopy.

Visible Lasers

Helium-Neon Laser

The helium-neon laser is continuously pumped electrically using a high dc voltage of up to 1 kV. The gain medium consists of a gas mixture of about 10 parts helium to each part of neon at a pressure of about 3 Torr. The gain of this medium is extremely low, hence the requirement for Brewster-angle windows to eliminate reflection loss of the light polarized in the plane that includes the axis of the laser and the normal to the Brewster window. Due to this requirement, the light output is necessarily plane polarized.

The pumping excites helium atoms by electron impact, and resonant energy transfer to neon atoms then occurs via collisions of the gaseous atoms (Figure 4). This creates a population inversion in the neon atoms, enabling the laser transition to occur at 632.8 nm. Following emission, the neon atoms decay nonradiatively down to their metastable $2p^53s^1$ level from which they decay back down to the ground state via collisional de-excitation with the walls of the tube. This mechanism restricts the maximum achievable power output to ca. 50 mW, because of the need to depopulate the metastable neon atoms by wall collisions.

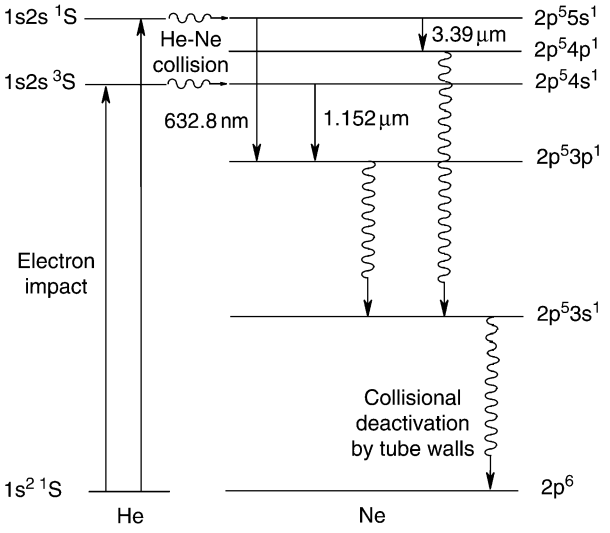


Figure 4 Energy levels of the He-Ne laser.

An advantage of using helium-neon lasers for Raman spectroscopy is that they generally operate in the fundamental TEM_{00} mode, which is critically dependent on the ratio of the length of the tube to its diameter. For this reason they are designed to have small tube diameters of around a few millimetres and lengths of 0.15–0.50 m, the small diameter also aiding collisional de-activation with the walls due to the relatively large surface to volume ratio of the tube. Another advantage of this laser for Raman spectroscopy is that the linewidth of the 632.8 nm emission line is ca. 1.5 GHz.

A disadvantage of the helium-neon laser is that it emits a large number of spontaneous emission lines originating from the excited neon atoms. These plasma emissions are observed in the Raman spectrum as sharp lines unless they are filtered out, for example, by a pre-monochromator or an interference filter. It should be mentioned, in passing, that these plasma lines are not always unwanted, because they can be very useful for wavenumber calibration of the Raman spectrometer.

The low output power of <50 mW is not generally a disadvantage for Raman microscopy since higher laser powers can often cause photolytic or thermal degradation of the sample. This is due to the high irradiance at the sample when the laser light is focused to a tight spot by the microscope objective. The excitation wavelength of 632.8 nm is suitable for combined use of the laser with a silicon-based CCD

detector in Raman spectroscopy, because large Stokes shifts in the 3000–4000 cm^{-1} region lie well within the quantum efficiency curve of this detector. It is also possible, though less common, to operate the helium-neon laser on weaker transitions that include the green 543.5 nm line.

Argon Ion Laser

The argon ion laser was one of the first lasers to be discovered following the invention of the laser and up until now it has been used extensively for Raman spectroscopy, among many other applications. Excitation is provided by a continuous electrical discharge, and because of the high energy required to ionize the argon atoms and then promote the ions to an excited state, the efficiency of the laser is very small (ca. 0.1%). In spite of this, once population inversion has been achieved, the gain of the laser is very high and output powers of up to 25 W can be obtained for the strong lines at 488.0 and 514.5 nm and up to 50 W for multiline operation.

An advantage of the argon ion laser is that it can provide emission at more than 35 discrete wavelengths, the strongest of which are listed in Table 1. These lie in the green, blue, and near ultraviolet regions of the spectrum, a number of the ultraviolet lines only being obtained on the larger frame argon ion lasers. Discrete laser emission lines are selected by tuning a prism or grating inside the cavity. The argon ion laser can also be operated in multiline mode, for example, for pumping dye or Ti:sapphire lasers.

Table 1 Continuous wave argon ion laser wavelengths and output powers for the Coherent Inc. Innova 300 argon ion laser system

Wavelength (nm)	*Power (mW)*
228.9	30
238.3	100
244.0	400
248.3	180
257.3	750
275.4	5
300–305.5	20
333.4	40
333.8	30
335.8	20
351.1	200
351.4	60
363.8	240
454.5	140
457.9	420
465.8	180
472.7	240
476.5	720
488.0	1800
496.5	720
501.7	480
514.5	2400
528.7	420

The five lines below 260 nm are frequency doubled.

The linewidths of the argon ion emission lines at 488.0 and 514.5 nm are around 4.0 GHz. The high temperature laser tube has a diameter of approximately 12 mm and a length in the range of 0.1 to 1.8 m. A 240 V three phase power supply and ca. 35 A are required for pumping a medium power 5 W argon ion laser. The tube also requires water cooling at a flow rate of ca. 10 L/min and a pressure of 25 psi because of the large amount of heat dissipation.

For Raman microscopy smaller, air-cooled argon ion lasers, which only require a 240 V single phase power supply, can be used to provide output of a few hundred milliwatts on the 488.0 and 514.5 nm lines. For this application, the excitation is provided by argon ion lasers, which are designed to operate in the TEM_{00} mode.

Until recently, the argon ion laser has been a workhorse as the most common excitation source for Raman spectroscopy, but it is now losing ground to solid state lasers. The principle reasons for this are that the latter are much more efficient and consequently do not in general have three phase power and water-cooling requirements. Nevertheless, argon ion lasers still have a niche when high excitation powers on the order of watts are required in conjunction with a good beam quality ($M^2 < 1.1$), for example, for Raman spectroscopy of gaseous samples. Another advantage of argon ion lasers over solid state lasers is their multiline capability.

Krypton Ion Laser

The krypton ion laser has useful discrete emission lines in the near ultraviolet, blue, yellow, red, and near infrared regions of the electromagnetic spectrum; the strongest lines of a Coherent Inc. Innova 400 krypton ion laser are listed in Table 2.

The argon ion and krypton ion lasers are close relatives, thus the large frame krypton ion laser has similar power and water-cooling requirements to those mentioned above for the argon ion laser. Also, like the argon ion laser, air-cooled models with lower power output are available. The characteristics of the two lasers are also similar; for example, the linewidths of the krypton ion transitions at 530.9, 568.2, and 647.1 nm are about 4.0 GHz, the laser tube has similar dimensions (0.1–2.0 m in length) and the laser can be operated in either TEM_{00} or multimode.

The krypton ion laser is even less efficient than the argon ion laser, consequently lower-power outputs of

Table 2 Continuous wave krypton ion laser wavelengths and output powers for the Coherent Inc. Innova 400 krypton ion laser system

Wavelength (nm)	*Power (mW)*
206.5	4
234	8
337.5–356.4	2000
406.7	900
413.1	1800
415.4	280
468.0	500
476.2	400
482.5	400
520.8	700
530.9	1500
568.2	1100
647.1	3500
676.4	900
720.8	45
752.5	1200
793.1–799.3	300

The 206.5 and 234 nm lines are frequency doubled.

up to about 20 W can be achieved when operated in multiline mode.

Mixed argon and krypton ion laser tubes are also commonly used that provide laser lines originating from both argon ion and krypton ion transitions. As with the individual argon ion and krypton ion lasers, sharp plasma lines, due to spontaneous emission from the rare gas ions, are observed in the Raman spectrum unless they are filtered out. If interference filters are used for this, each laser line requires its own filter that is tailored to the specific emission wavelength.

HeCd Laser

Laser lines at 325 and 442 nm, both capable of providing milliwatts of output power, can be obtained from the helium-cadmium laser. These emission lines result from electronic transitions in free cadmium atoms.

It is obviously advantageous to use blue, rather than longer wavelength, excitation (e.g., provided by the 442 nm line of the HeCd laser) for off-resonance Raman spectroscopy, due to the ν^4 dependence of the Raman light scattering efficiency, provided that the efficiencies of the illumination/collection optics and Raman spectrometer throughput are optimized in the blue region of the spectrum. Unfortunately, far more samples fluoresce when excited with blue light than with red or near infrared radiation, which explains why red or near infrared lasers (e.g., HeNe, semiconductor lasers) are more commonly used for general Raman applications.

Near Infrared Lasers

Although the traditional laser systems for Raman spectroscopy have been the argon ion, krypton ion, and helium-neon lasers for discrete excitation wavelengths, the diode laser has gained popularity over recent years.

The principal advantage of near infrared semiconductor lasers for Raman microscopy is that they generally excite less fluorescence in the Raman spectrum than visible lasers due to their longer wavelength. Commonly available wavelengths are 670, 785, 830, and 852 nm and, of these, the first two can be used in conjunction with silicon-based CCD detectors to give Stokes Raman shifts over the whole range of fundamental vibrations, up to 4000 cm^{-1}. However, for excitation wavelengths of 830 and 852 nm, only Stokes Raman bands in the fingerprint region can be detected with a silicon-based CCD detector because higher wavenumber Raman shifts approach the bandgap of the silicon semiconductor ($\lambda >$ ca. 1050 nm).

The disadvantages of diode lasers are that they cannot supply high power of narrow linewidth in single-mode and they are susceptible to mode hopping which gives rise to a shift in the excitation wavelength. This latter drawback is particularly disadvantageous for Raman spectroscopy because it results in a shift in the wavenumber positions of the Raman bands. When good beam quality is not required, broad stripe, high power laser diodes can be used; these can have output powers that are greater than 1 W but their emission linewidths are equal to ca. 2 nm, which is far too wide for Raman spectroscopic applications. These laser diodes find applications as pumps of solid state lasers, for example, Nd:YAG, Nd:YVO_4 lasers, however. Additionally, amplified stimulated emission (ASE) gives an unwanted background signal that can be very broad (ca. 20–30 nm) and 0.1–1% of the intensity of the laser line. This necessitates the use of bandpass filters in order to reduce this unwanted background.

Although laser diodes having an output of less than 200 mW can operate in TEM_{00} and in single longitudinal modes, mode hopping can occur due to optical feedback or fluctuations in environmental factors and this causes severe broadening of the linewidth. The sensitivity to optical feedback can be greatly reduced, hence the laser linewidth can be narrowed appreciably, by confining the frequency of the photons either in an internal cavity containing a small diffraction grating or in an external cavity. The former types of laser are sometimes called 'distributed feedback' (DFB) lasers because the feedback is distributed over the length of the grating, rather

than occurring all at once at a mirror. The wavelength that is fed back is determined by the period of the grating. Usually, a DFB laser has a grating fabricated into the entire length of the laser. A variation referred to as a distributed Bragg reflector (DBR) laser has a distinct grating fabricated into the substrate on each side of the active area. The external cavity semiconductor laser (ECSL) is fabricated by placing the laser diode in a separate resonator like a conventional gas or solid-state laser. The DBR and ECSL lasers are now discussed.

Distributed Bragg Reflector Lasers

The DBR laser consists of a grating on each side of the active region (Figure 5); these gratings act as mirrors having a reflectivity that is optimized at one particular wavelength, in addition to narrowing the laser linewidth. At present, DBR lasers are only commercially available having an excitation wavelength of 852 nm.

The emission linewidth is less than 4 MHz, which is suitable for the vast majority of Raman spectroscopic applications and the laser operates in single TEM_{00} mode. The disadvantages of DBR lasers are that the maximum output power is restricted to around 150 mW, so an optical isolator is usually necessary in order to prevent external facet damage caused by external optical feedback, and only an excitation wavelength of 852 nm is currently available.

External-Cavity Semiconductor Lasers (ECSL)

ECSLs provide higher output power (up to ca. 1 W) and a wider range of excitation wavelengths (630 nm to around 850 nm), as well as being less expensive than DBR lasers. They use the semiconductor chip only as the gain medium and employ an external grating, both as the frequency selector and the reflective mirror. Specific excitation wavelengths are becoming widely adopted in Raman microscopy employing single grating spectrograph in conjunction with CCD detection, such as 670, 785, 830, and 852 nm. The ECSLs have emission linewidths as low as a few MHz, but in general they span the range 2 MHz–30 GHz, and can operate in a nearly diffraction-limited transverse mode.

Three different cavity designs employing a diffraction grating have been discussed in the literature: Littrow type-I, Littrow type-II, and Littman configurations. In the Littrow type-I design, the laser diode light is incident on the grating at an angle of incidence equal to θ_{Littrow} (Figure 6a). The diffracted light in first order is re-directed back to the laser diode to provide feedback and the output is the zeroth order (i.e., specularly reflected) light. A disadvantage of this design is that the bandwidth of the grating is relatively large because a single pass geometry is used. In the Littrow type-II design, light from the rear facet of the laser diode is collimated by a lens and directed on to a grating at an angle of incidence equal to θ_{Littrow} (Figure 6b). The diffracted light in first order is re-directed back to the laser diode, in a similar fashion to the Littrow type-I design, in order to provide optical feedback, and the output is the laser light emitted from the front facet. Disadvantages of the Littrow type-II design are the requirement for

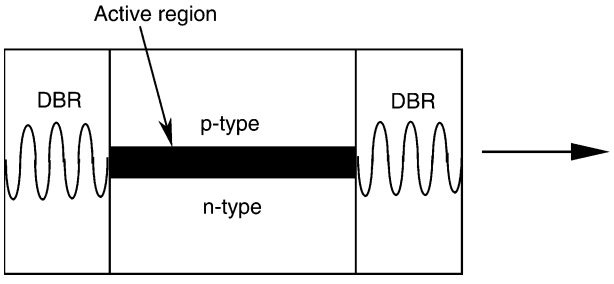


Figure 5 Diagram of the cavity of the DBR laser (adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley).

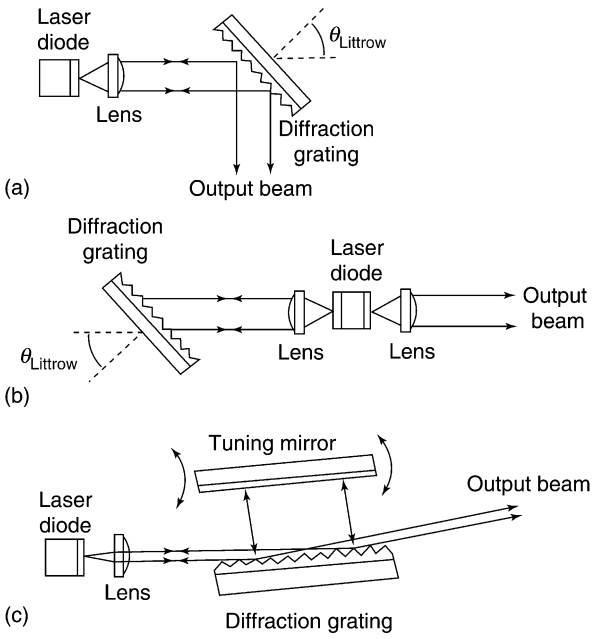


Figure 6 Diagrams of the cavities of (a) Littrow type-I, (b) Littrow type-II, and (c) Littman external cavity diode lasers (adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley).

custom fabrication of antireflection coatings on both facets and the sensitivity to external optical feedback. In the Littman design, the laser diode light is collimated by a lens and directed on to a grating at a grazing angle of incidence. The diffracted light is reflected by a mirror and diffracted back to the laser diode by the grating in order to provide optical feedback (Figure 6c). An advantage of this design is that the grating bandwidth is less than half that of the Littrow designs due to the double pass geometry, but a disadvantage is that the external cavity length is longer, resulting in a narrower spacing of the cavity modes. Another advantage of this design is that the tuning is accomplished by rotating the mirror instead of the grating, thus the alignment of the output beam is not altered when tuning. For Littrow type I and II and Littman ECSL designs, filtering of the ASE is required, as it can have an intensity of ca. 0.1 to 1% of the intensity of the laser line. Thus, these designs necessitate the incorporation of a bandpass filter (having a rejection of better than 10^{-4} at the ASE).

Nd:YAG Laser

Under normal conditions, the Nd:YAG laser oscillates on a transition in the near infrared at 1064 nm, and this is the excitation line that is used in most commercial Fourier Transform Raman spectrometers. The gain medium is a crystal of yttrium aluminium garnet ($Y_3Al_5O_{12}$, YAG) doped with ca. 1.0 mol% Nd^{3+} cations that substitute Y^{3+} ions in the cubic YAG lattice.

The Nd:YAG laser is a four-level system (Figure 7), which has high gain and low threshold due to the narrow fluorescent linewidth of the laser transition. Absorption bands of the Nd^{3+}ions around 808 nm conveniently match the energy of commercially available high-power multimode diode lasers that serve as the pump. The Nd^{3+} ions decay nonradiatively to the upper laser level, the $^4F_{3/2}$ state, thereby creating a population inversion. This is because the lower laser level, the $^4I_{11/2}$ state, has no appreciable thermal population at room temperature, since it is $>2000\ cm^{-1}$ above the $^4I_{9/2}$ ground state, and the $^4F_{3/2}$ excited state has a relatively long lifetime of 230 μs.

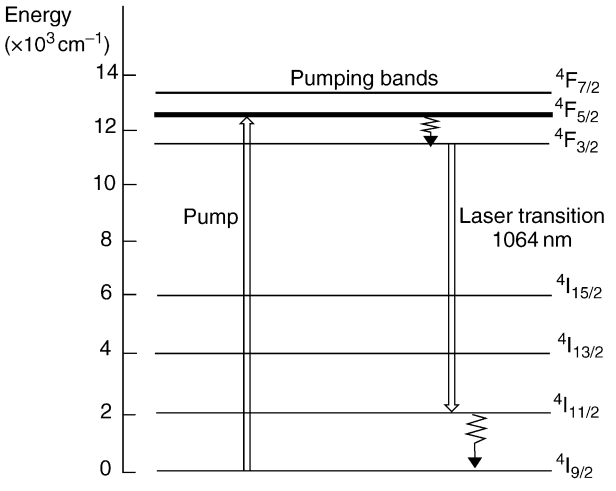


Figure 7 Energy levels of the Nd:YAG laser.

The high thermal conductivity of Nd:YAG enables the laser to operate at high power in either continuous wave (CW) or *Q*-switched modes, and the diode pumped solid state (DPSS) variety of the cw Nd:YAG laser can currently achieve an output power in excess of 20 W on the 1064 nm line. The laser can be designed to operate in the fundamental TEM_{00} mode and the full width at half maximum (FWHM) linewidth of the spontaneous emission of the 1064 nm laser transition is 120 GHz (ca. 0.45 nm). Advantages of the diode pumped Nd:YAG laser are that it is air-cooled and can be operated at 240 V single phase. Also, being all solid state and having a small footprint, it is robust and portable. The lifetime of the laser is dependent on the laser diodes used for pumping, but some Nd:YAG laser designs enable these to be replaced in the field.

The frequency-doubled Nd:YAG laser emitting green light, having a wavelength of 532 nm, is also commonly used nowadays for dispersive Raman spectroscopy, and an output power of 10 W on the 532 nm laser line is provided by commercially available frequency-doubled diode-pumped Nd:YAG lasers.

It is worth mentioning that Nd:YAG lasers have also been fabricated that emit either at 946 or 1330 nm, by suppressing the strong emission at 1064 nm and optimising the optics for the desired wavelength.

Nd:YVO_4 and Nd:YLF Lasers

An intracavity, frequency-doubled DPSS Nd:YVO_4 laser has recently become commercially available, and it has some advantages over the Nd:YAG laser including a larger stimulated emission cross-section and a higher absorption coefficient (along the extraordinary direction of the birefringent crystal). It has the same emission wavelength as the frequency-doubled Nd:YAG laser, i.e., 532 nm, and Nd:YVO_4 is the material of choice for cw end-pumped lasers having around 5 W output power. This is because the diode laser pump beam is tightly focused in the end-pumped system, but a small waist diameter cannot be retained over a distance of more than a

few millimeters, consequently a high absorption coefficient and gain are very beneficial.

The gain medium of the Nd:YLF laser consists of a crystal of yttrium lithium fluoride (YLF) that is doped with Nd^{3+} ions on the Y^{3+} cation sites. Unlike the Nd:YAG and Nd:YVO_4 lasers, the emission does not occur at 1064 nm; instead the $^4F_{3/2}-{}^4I_{11/2}$ emission occurs at wavelengths of either 1047 or 1053 nm, depending on the polarization that is selected. The former is due to extraordinary polarized light, whereas the latter is due to ordinary polarization, and either of these emission wavelengths can be selected using an intracavity polarizer. The Nd:YLF laser can offer benefits in *Q*-switched operation when the longer fluorescence lifetime (480 μs) of Nd^{3+} ions in the $^4F_{3/2}$ state enables a higher energy to be stored for the same number of pump laser diodes.

Ti:Sapphire Laser

The Ti:sapphire laser is tunable over the approximate range of 670–1070 nm with the peak of the gain curve at ca. 800 nm. In the gain medium, ca. 0.1% by weight Ti^{3+} is doped into a crystal of sapphire grown by the Czochralski method. The Ti^{3+} ions substitute for Al^{3+} ions in the Al_2O_3 of the sapphire and the laser emission is due to the $^2E-{}^2T_2$ transition of the Ti^{3+} cation, which has a $3d^1$ valence electronic configuration (Figure 8). The laser has a large stimulated emission cross-section but the fluorescence lifetime of the upper laser level (^2E state) is quite short (3.2 μs), thus the laser is usually laser (e.g., by argon ion or frequency-doubled Nd:YAG or Nd:YVO_4 lasers) rather than flashlamp pumped because a very high pump flux is required.

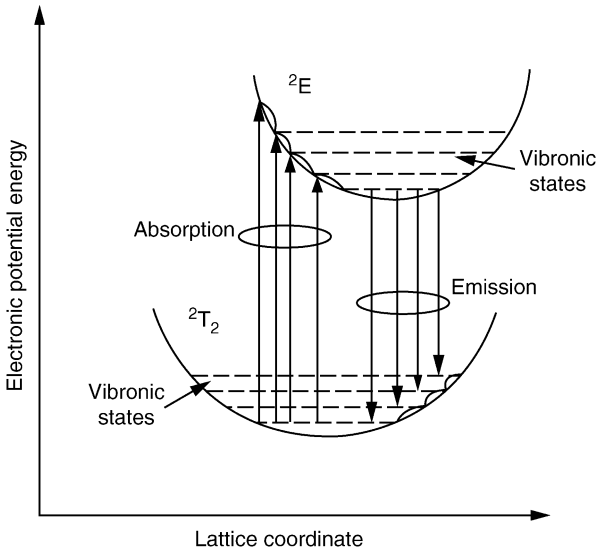


Figure 8 Energy levels of the Ti:sapphire laser (adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman Spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley).

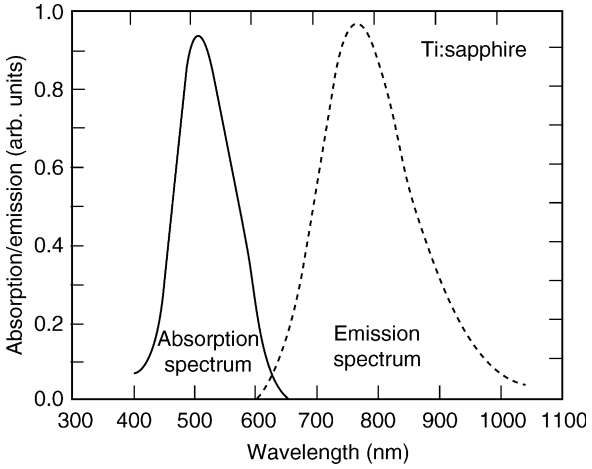


Figure 9 Absorption and emission spectra of the Ti^{3+} ion in sapphire. Al_2O_3 (adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman Spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley).

The absorption and emission bands are broad and widely separated, due to the vibronic coupling between the Ti^{3+} host and the Al_2O_3 lattice (Figure 9). The lower laser level is any one of the vibronic levels of the $^2T_2$ state. Following the laser transition, the Ti^{3+} ions decay from the upper vibronic levels of the $^2T_2$ state down to the lower vibronic levels. Hence, this is a four-level laser.

The broad spontaneous emission linewidth of the $^2E-{}^2T_2$ laser transition is around 100 THz, the ouput power can be as high as 50 W, and the laser can be operated in either TEM_{00} or multimodes.

UV Lasers

UV laser sources offer numerous advantages over visible laser sources for Raman spectroscopy. A major consideration is that many analytes have absorptions in the near UV, making them amenable to resonance Raman spectroscopic studies. The signal enhancement (up to 10^6), that can be achieved in resonance Raman spectra, results in a large increase in detection sensitivity. Furthermore, some vibrational modes, which normally give rise to weak bands in the off resonance Raman spectrum, can show strong enhancement in the UV excited resonance Raman spectrum. A good example of this is the amide II band of peptides and proteins, which is due to a combination of N–H in plane bending and C–N stretching modes of the peptide linkage. This band is normally weak in the Raman spectrum but strong in the infrared spectrum; however, the amide II band

can show strong enhancement in the UV excited resonance Raman spectrum and this can be useful for determining secondary structure in proteins. A further benefit of UV excited Raman spectroscopy is that fluorescence can often be avoided as it tends to occur at a lower energy, outside the Stokes Raman spectral window.

It can be advantageous to use cw rather than pulsed excitation for UV resonance Raman spectroscopy, if the analyte has a long excited state lifetime relative to the pulsewidth of the pulsed laser. This is because the concentration of analyte molecules in the ground state is significantly depleted during pulsed excitation, due to Raman saturation giving a lower signal to noise ratio in the Raman spectrum than for the case of cw excitation. It has been found that pulse energy flux densities should be less than 1 mJ/cm^2 and flowing or spinning samples should be used, in order to ensure that nonlinear phenomena and saturation effects do not occur.

For UV Raman microscopy, where the laser beam is focused into a small spot of micrometer dimensions on the sample, it is essential to avoid using pulsed lasers delivering high peak powers. This is because such pulsed lasers can cause dielectric breakdown, and even at lower peak powers they can cause nonlinear effects and Raman saturation phenomena. Consequently, with UV Raman microscopy, one is restricted to the use of cw or quasicontinuous laser excitation.

It is only recently that high thoughput UV Raman microspectrometers have been developed. It has been demonstrated for visible Raman spectroscopy that a significantly higher throughput can be achieved by employing a single stage spectrograph with a notch filter instead of a double or triple monochromator. In the UV region, blocking the Rayleigh scattering is a problem because notch filters are not currently available. However, it has been discovered that this can be overcome by introducing two modifications to the design of the optical layout of a visible Raman microscope. First, two novel dielectric longpass filters were used instead of a notch filter in order to reject the elastically scattered light. Second, an all-reflecting Cassegrain microscope objective was used, instead of a lens, in order to block the specular reflection, and thereby further reduce the Rayleigh scattering background. These design modifications have enabled a single stage spectrograph with a holographic grating to be used to disperse the Stokes Raman radiation in a UV Raman microspectrometer (Figure 10).

In the optical layout of Figure 10 it can be seen that the laser excitation is not focused by the Cassegrain microscope objective because an epi-illumination configuration is not employed. Instead the laser is

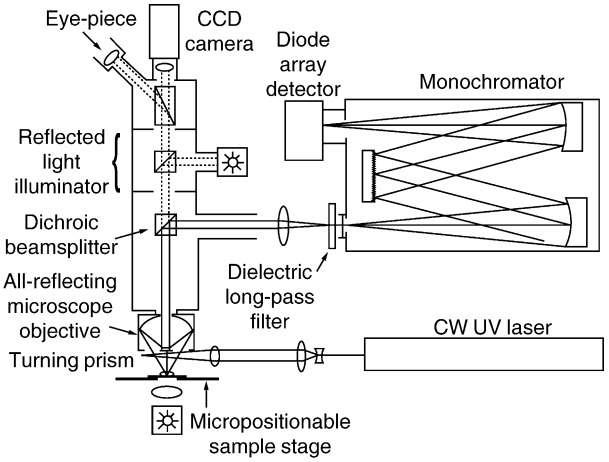


Figure 10 Diagram of the optical layout of a UV Raman microscope (adapted from Pajcini V, Munro CH, Bormett RW, Witkowski RE and Asher SA (1997) UV Raman microspectroscopy: Spectral and spatial selectivity and simplicity. *Applied Spectroscopy* 51: 81–86).

focused by a separate lens and directed on to the sample with a turning prism located directly underneath the Cassegrain objective. This minimizes loss of laser beam throughput or scattered light throughput to the spectrometer at the expense of spatial resolution, since the laser light is focused by a longer focal length lens than the microscope objective. An additional difference between the optical layouts of the visible and UV Raman microscopes shown in Figures 1 and 10, respectively, is that pinhole spatial filters are not shown in the latter. However, better axial spatial resolution could be achieved if this UV Raman microscope was made confocal by introducing a pinhole and lens.

Frequency-Doubled Argon Ion Laser

The frequency doubled argon ion laser is a popular choice among UV laser sources. The cw UV laser contains a nonlinear optical beta-barium borate (BBO) crystal, which is located within the laser cavity; this crystal frequency doubles the strong Ar^+ lines to give five UV lines below 260 nm which are listed in Table 1 for the Coherent Inc. Innova 300 Ar^+ ion laser system. Of these UV lines, the 228.9 nm line is almost ideal for resonance Raman enhancement of tyrosine and trpytophan residues in proteins and the 244.0 nm line is well suited for studying tyrosinate groups. The 244.0 nm line has also been used to excite selectively UV resonance Raman spectra from spatially resolved areas of biological samples within the nucleus of a single cell, from DNA in particular, using low laser powers and short acquisition times to keep sample damage to a minimum under the microscope.

Frequency-Doubled Krypton Ion Laser

Like the cw frequency-doubled argon ion laser mentioned above, the cw frequency-doubled krypton ion laser contains a nonlinear optical BBO crystal, which is located within the laser cavity; this crystal frequency doubles the strong Kr^+ lines to give 206.5 and 234 nm UV lines which are listed in Table 2 for the Coherent Inc. Innova 400 Kr^+ ion laser system. The 206.5 nm laser line is useful for exciting resonance Raman spectra within the $\pi-\pi^*$ amide transition of the peptide backbone of proteins, and the enhanced protein amide vibrational bands can be used to determine the protein secondary structure.

HeCd Laser

The 325 nm HeCd excitation line, in the near ultraviolet, can be used for combined micro-Raman/photoluminescence studies of, for example, semiconductors. These lasers are suitable for low-power applications, typically having output powers in the 1–100 mW range, and they can be designed to operate in TEM_{00} single-mode or multimode.

Quasi-CW Mode-Locked Ti:Sapphire Laser

The second harmonic of the mode-locked Ti:sapphire laser (ca. 350–500 nm), third harmonic (ca. 233–333 nm), and fourth harmonic (ca. 200–250 nm), are all available with conventional nonlinear crystals.

Typically, the Ti:sapphire crystal is pumped by a cw argon ion laser, for example, as the excitation source for a UV Raman microscope. This quasicontinuous laser system produces 2–3 ps pulses at a 76 MHz repetition rate and is continuously tunable over the 200–300 nm range. The typical power output is ca. 50 mW at 250 nm, ca. 10 mW at 240 nm, and 1 mW at 230 nm. In future, it is anticipated that a frequency-doubled Nd:YAG laser will be increasingly favored as the pump, creating an all-solid-state laser source.

Conclusions

Laser systems for Raman microscopy have been described in this article. Although the differences in the terms 'Raman spectroscopy' and 'Raman microscopy' only refer to whether a macro or micro sampling configuration is used, this does have a bearing on the types of laser systems that are employed. This is because the latter imposes a restriction on using low duty pulsed lasers with high pulse peak power, as both spatial and temporal confinement of the laser excitation can lead to dielectric breakdown of the sample, nonlinear phenomena, or Raman saturation. Thus, for this reason, one is limited to cw lasers or quasi-cw lasers for Raman microscopy whereas there is no such restriction for Raman spectroscopy.

The quality of the laser beam, measured by the M^2 value or 'times diffraction limit' influences the ultimate spatial resolution that can be achieved in Raman microscopy.

In the visible region, argon ion, krypton ion, and frequency-doubled Nd:YAG lasers have good beam quality and high power, with helium-neon and helium-cadmium lasers also having good beam quality but lower power; this makes these lasers good light sources for visible Raman microscopy.

For FT Raman microscopy using Nd:YAG excitation, there is a lack of fluorescence interference from a wide variety of samples, but sensitivity and spatial resolution are compromised in comparison to visible dispersive Raman microscopy.

Red and near infrared semiconductor lasers operating in TEM_{00} single mode can have good beam quality, approaching that of gas lasers, but they are power limited. Obviously, when good beam quality is not necessary, for example, for optically pumping other laser sources, high-power diode lasers can be used in multimode operation. Solid state red and near infrared laser sources are becoming very popular for routine Raman microscopic analysis due to their compactness, robustness, and economical power consumption.

Another reason for the popularity of the red and near infrared diode laser sources for Raman microscopy is that fluorescence is less of a problem due to the lower energy of the light, compared to conventional green 514.5 nm excitation. Although there is a penalty, due to the dependency of the scattering efficiency on the ν^4 term, present day silicon-based CCD arrays can exhibit excellent detection quantum efficiency of Stokes shifted Raman radiation, across the whole vibrational spectrum, when using standard 670 and 785 nm diode lasers and, at least across the fingerprint region, when using 830 and 852 nm diode lasers.

For ultimate spatial resolution in the Raman microscopic experiment, short wavelength excitation is advantageous because it is fundamentally possible to focus to a smaller diffraction-limited laser spot. In this regard, it would be preferable to use UV Raman microscopy where fluorescence is also not as problematic as it is in the visible region. However, in spite of recent improvements in UV laser sources such as the intracavity frequency-doubled argon and krypton ion cw lasers, there are current instrumental limitations on the throughput efficiency of the UV Raman microscope imposed by the lack of availability of

suitable notch filters and low stray light and high throughput optics.

It is anticipated that in the future all-solid-state, quasi-cw, diode pumped, mode-locked Ti:sapphire lasers operating on their third or fourth harmonics will become increasingly popular for providing tunable UV excitation affording good beam quality. It is now also possible to fabricate hollow cathode metal ion deep UV lasers, 10–15 cm in length, 2–4 cm in diameter, weighing 50–100 g and requiring only 2–3 W of electrical power. These lasers have low beam quality (M^2 equal to ca. 18), but they make possible the development of portable UV fluorescence imaging and Raman microprobes for geobiological exploration of terrestrial and extraterrestrial environments.

See also

Fourier Optics. **Imaging:** Infrared Imaging. **Introductory Article:** Early Development of the He-Ne Laser (Title TBC). **Lasers:** Noble Gas Ion Lasers; Semiconductor Lasers. **Nonlinear Optics, Applications:** Raman Lasers. **Scattering:** Raman Scattering.

Further Reading

Andrews DL (1990) *Lasers in Chemistry* (2nd edition). Berlin, Heidelberg: Springer-Verlag.

Asher SA, Bormett RW, Chen XG, *et al.* (1993) UV resonance Raman spectroscopy using a new cw laser source: Convenience and experimental simplicity. *Applied Spectroscopy* 47: 628–633.

Bell SEJ, Bourguignon ESO, O'Grady A, Villaumie J and Dennis AC (2002) Extracting Raman spectra from highly fluorescent samples with "Scissors" (SSRS, Shifted-Subtracted Raman Spectroscopy). *Spectroscopy Europe* 14: 17.

Best SP, Clark RJH and Withnall R (1992) Non-destructive pigment analysis of artefacts by Raman microscopy. *Endeavour* 16: 66.

Boustany NN, Manoharan R, Dasari RR and Feld MS (2000) Ultraviolet resonance Raman spectroscopy of bulk and microscopic human colon tissue. *Applied Spectroscopy* 54: 24–30.

Delhaye M, Barbillat J, Aubard J, Bridoux M and Da Silva E (1996) Instrumentation. In: Turrell G and Corset J (eds) *Raman Microscopy: Developments and Applications*, chap. 3. London: Academic Press.

Delhaye M and Dhamelincourt P (1975) Raman microprobe and microscope with laser excitation. *Journal of Raman Spectroscopy* 3: 33.

Derbyshire A and Withnall R (1999) Pigment analysis of portrait miniatures using Raman microscopy. *Journal of Raman Sprectroscopy* 30: 185.

Gordeyev SA, Nikolaeva GYu, Prokhorov KA, Withnall R, Dunkin IR and Shilton SJ (2001) Super-selective polysulfone hollow fiber membranes for gas separation: assessment of molecular orientation by Raman spectroscopy. *Laser Physics* 11: 82–85.

Hayward CL, Best SP, Clark RJH, Ross NL and Withnall R (1994) Polarised single crystal Raman spectroscopy of sinhalite, $MgAlBO_4$. *Spectrochimica Acta* 50A: 1287–1294.

Hendra P, Jones C and Warnes G (1991) *Fourier Transform Raman Spectroscopy*. London: Ellis Horwood.

Hitz B (2003) Solid state lasers are gunning for argon ion's place. *Photonics Spectra* 54–58.

Holtz JSW, Bormett RW, Chi Z, *et al.* (1996) Applications of a new 206.5 nm continuous wave laser source: UV Raman determination of protein secondary structure and CVD diamond material properties. *Applied Spectroscopy* 50: 1459–1468.

http://www.btools.com

Koechner W and Bass M (2003) *Solid-State Lasers*. New York: Springer-Verlag.

Landsberg G and Mandelstam L (1928) Eine neue Erschienung bei der Lichtzerstreaung in Krystallen. *Naturwiss* 16: 557.

Matousek P, Towrie M, Stanley A and Parker AW (1999) Efficient rejection of fluorescence from Raman spectra using picosecond Kerr gating. *Applied Spectroscopy* 53: 1485.

Nakashima S, Okumura H, Yamamoto T and Shimidzu R (2004) Deep-ultraviolet Raman microspectroscopy: Characterization of wide-gap semiconductors. *Applied Spectroscopy* 58: 224–229.

Pajcini V, Munro CH, Bormett RW, Witkowski RE and Asher SA (1997) UV Raman microspectroscopy: Spectral and spatial selectivity and simplicity. *Applied Spectroscopy* 51: 81–86.

Pallister DM and Morris MD (1993) In: Morris MD (ed.) *Microscopic and Spectroscopic Imaging of the Chemical State*, chap. 1. New York: Marcel Dekker.

Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy,* 1. Chichester, UK: Wiley.

Pan M-W, Benner RE, Johnson CW and Smith LM (2000) Near-IR lasers rejuvenate Raman spectroscopy. *Optoelectronics World* S5–S10.

Raman CV and Krishnan KS (1928) A new type of secondary radiation. *Nature* 121: 501.

Rosasco GJ (1980) Raman microprobe spectroscopy. In: Clark RJH and Hester RE (eds) *Advances in Infrared and Raman Spectroscopy*, vol. 7. Chichester: Wiley.

Rosasco J, Etz ES and Cassatt WA (1975) The analysis of discrete fine particles by Raman spectroscopy. *Applied Spectroscopy* 29: 396.

Schawlow AL and Townes CH (1958) Infrared and optical masers. *Physics Review* 112: 1940.

Smekal A (1923) Sur Quantentheorie der Dispersion. *Naturwiss* 11: 873.

Storrie-Lombardi MC, Hug WF, McDonald GD, Tsapin AI and Nealson KH (2001) Hollow cathode ion lasers for deep ultraviolet Raman spectroscopy and fluorescence

imaging. *Review of Scientific Instruments* 72: 4452–4459.
Turrell G, Delhaye M and Dhamelincourt P (1996) Characteristics of Raman microscopy. In: Turrell G and Corset J (eds) *Raman Microscopy: Development and Applications*, 2. London: Academic Press.
Young M (1977) *Optics and Laser: An Engineering Approach*. Berlin, Heidelberg: Springer-Verlag.
Yu Y-M, Nam S, Byungsung O, *et al.* (2002) Resonant Raman scattering in ZnS epilayers. *Materials Chemistry and Physics* 78: 149–153.

Second-Harmonic Spectroscopy

J I Dadap and T F Heinz, Columbia University, New York, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The demonstration of second-harmonic generation (SHG), one of the early experiments with pulsed lasers, is considered to have marked the birth of the field of nonlinear optics. The importance of second-harmonic generation, and the related phenomena of sum- and difference-frequency generation, as methods for producing new frequencies of coherent radiation was recognized immediately and has in no way diminished over the years. Associated with these applications came an interest in the fundamental nature of the second-order nonlinear response, and some spectroscopic studies were undertaken with this motivation. These investigations were of rather limited scope and were not developed as a general probe of materials. In contrast, the third-order nonlinear optical interactions have given rise to a panoply of significant spectroscopic measurement techniques, including, to name a few, coherent Raman spectroscopy, pump-probe and other four-wave mixing measurements, two-photon absorption and hole burning measurements, photon echoes, and Doppler-free spectroscopy.

A central factor in the lack of spectroscopic applications of the second-order nonlinear response lies in a fundamental symmetry constraint. Unlike the third-order nonlinearity, which is present in all materials, only materials lacking a center of inversion give rise to an (electric-dipole) allowed second-order nonlinear response. As we demonstrate in this article, however, it is this property that makes SHG, as well as the sum-frequency generation process discussed elsewhere in this encyclopedia, such a remarkable probe of surfaces and interfaces. For centrosymmetric materials, the second-order nonlinear response is strong only at interfaces, where only a monolayer or so of material contributes to the response. This yields a purely optical probe (with incident and emitted photons) that gives an inherent surface sensitivity comparable to the best electron spectroscopies, such as electron diffraction, Auger-electron spectroscopy, photoemission spectroscopy, and scanning tunneling microscopy, the workhorses of surface science. Neither linear nor third-order nonlinear optical measurements provide this inherent interface sensitivity. The possibility of probing interfaces by purely optical means permits one to take advantage of several important features of optical radiation. Perhaps most noteworthy is the possibility of probing buried interfaces of every sort, from solid/solid to liquid/liquid, as well as the more conventional cases of solid/vapor, liquid/vapor, and liquid/solid. In addition, laser-based techniques offer unequalled capabilities for spectral or temporal resolution.

In this article, we present a brief overview of SH spectroscopy for the study of surfaces and interfaces. The article is divided into two principal sections: an introductory section that presents some of the fundamentals of SHG and the behavior at interfaces; and a discussion of the different types of information that the method can yield, as illustrated by some representative examples. A brief section on the experimental technique links these two parts.

Fundamentals of Surface SHG

Basic Notions

Within the domain of classical optics, the response of a material to electromagnetic radiation is described by an induced polarization **P** that varies linearly with the electric field strength, **E**. This situation reflects the fact that the strength of electric fields for light encountered under conventional conditions is minute compared to that of the electric fields binding atoms and solids together. The electric field binding an atom may be estimated as $E_a \sim 1\ \text{V/Å} = 10^{10}\ \text{V/m}$. To reproduce this field strength by an electromagnetic wave requires an irradiance of $I_a \sim 10^{17}\ \text{W/m}^2$, so the validity of this linear approximation in classical optics is clear. The advent of the laser, with its capability for producing light beams with very high optical power and irradiance, has permitted us to

probe and exploit the nonlinear response of materials, giving rise to the field of nonlinear optics.

The next-order material response beyond the linear approximation consists of an induced nonlinear source polarization $\mathbf{P}^{(2)} = \varepsilon_0 \boldsymbol{\chi}^{(2)} : \mathbf{EE}$ that depends quadratically on the driving electric field $\mathbf{E}$, described by a second-order nonlinear susceptibility $\boldsymbol{\chi}^{(2)}$. For excitation by a single laser at a frequency ω, this response gives rise to SHG at a frequency of 2ω, as well as optical rectification with a dc induced-polarization. For beams at two distinct frequencies of ω_1 and ω_2, the processes of sum- or difference-frequency generation (SFG or DFG) become operative with the production of the new frequencies of $|\omega_1 \pm \omega_2|$. For a relationship of the given form, which assumes only a spatially local quadratic dependence of the polarization on the driving electric field, one can show that $\boldsymbol{\chi}^{(2)}$ must vanish within any centrosymmetric medium. This is a direct consequence of the fact that $\boldsymbol{\chi}^{(2)}$ as a quantity describing a centrosymmetric medium should not change under an inversion operation, while the quantities $\mathbf{P}^{(2)}$ and $\mathbf{E}$ as polar vectors will change sign. The interface specificity of the SHG technique for centrosymmetric media follows from this general observation: only at interfaces is the bulk symmetry broken and the SHG process allowed.

Anharmonic Oscillator Model

Before we embark on our discussion of surface SHG, we first briefly consider a simple classical model of SHG based on an anharmonic oscillator. Although this is not a realistic treatment of the response of materials, it helps to illustrate the characteristic features of the surface SHG process and its spectroscopic attributes. In the anharmonic oscillator model, the medium is treated as a collection of electrons bound to fixed ion cores. While the usual Lorentz model only includes the effect of a linear restoring force, to describe second-order response, we generalize this description to include a leading-order anharmonic restoring force. For simplicity, we consider motion of the anharmonic oscillator in one spatial dimension, in which case the classical equation of motion for the displacement, x, of the electron can then be written as

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + 2\gamma\frac{\mathrm{d}x}{\mathrm{d}t} + \omega_0^2 x - bx^2 = -\frac{e}{m}E(t) \qquad [1]$$

Here $-e$ and m denote, respectively, the electron charge and mass, and $E(t)$ is the optical driving field. The (angular) frequency ω_0 defines the resonance of the harmonic component of the response, and γ represents a phenomenological damping rate for the oscillator. The nonlinear restoring force corresponds to the term containing the parameter b, a material-dependent constant. We are interested in the solution of this equation for a monochromatic driving field at frequency ω with amplitude $E(\omega)$ which we write as

$$E(t) = E(\omega)\exp(-\mathrm{i}\omega t) + \text{c.c.} \qquad [2]$$

An exact solution to eqn [1] in the presence of this electric field is not possible. Let us, however, consider a regime in which we do not excite the oscillator too strongly, and the effect of the anharmonic term remains relatively small compared to the harmonic one. In this case, we may solve the problem perturbatively in powers of the driving field. We find a linear response at the fundamental frequency and a nonlinear response at the SH frequency, which we can write in terms of the corresponding induced dipole moments $p = -ex$:

$$p(t) = p^{(1)}(\omega)\exp(-\mathrm{i}\omega t) + p^{(2)}(2\omega)\exp(-\mathrm{i}2\omega t) + \text{c.c.} \qquad [3]$$

We can then write the amplitudes for the harmonically varying dipole moments as

$$p^{(1)}(\omega) = \alpha^{(1)}(\omega)E(\omega) \qquad [4]$$

$$p^{(2)}(2\omega) = \alpha^{(2)}(2\omega = \omega + \omega)E(\omega)E(\omega) \qquad [5]$$

where $\alpha^{(1)}$, the linear polarizability, and $\alpha^{(2)}$, the second-order polarizability (or first hyperpolarizability) for SHG are given, respectively, by

$$\alpha^{(1)}(\omega) = \frac{e^2}{m}\frac{1}{D(\omega)} \qquad [6]$$

$$\alpha^{(2)}(2\omega = \omega + \omega) = \frac{(e^3/m^2)b}{D(2\omega)D^2(\omega)} \qquad [7]$$

Here we have introduced the resonant response at the fundamental frequency by $D(\omega) \equiv \omega_0^2 - \omega^2 - 2\mathrm{i}\gamma\omega$, with a corresponding relation for the response $D(2\omega)$ at the SH frequency. A similar derivation can, of course, be applied to obtain the corresponding material response for the SFG or DFG processes.

The spectral dependence of the polarizabilities $\alpha^{(1)}$ and $\alpha^{(2)}$ is illustrated in Figure 1 for the anharmonic oscillator model with a resonance frequency of ω_0. In the linear case, one sees the expected behavior of the loss, associated with the $\mathrm{Im}[\alpha^{(1)}]$, and dispersion associated with $\mathrm{Re}[\alpha^{(1)}]$. A single resonance in the linear response is seen for the driving frequency ω near the oscillator frequency ω_0. For the case of SHG, on the other hand, we observe both resonances when the fundamental frequency $\omega \approx \omega_0$ and when the SH

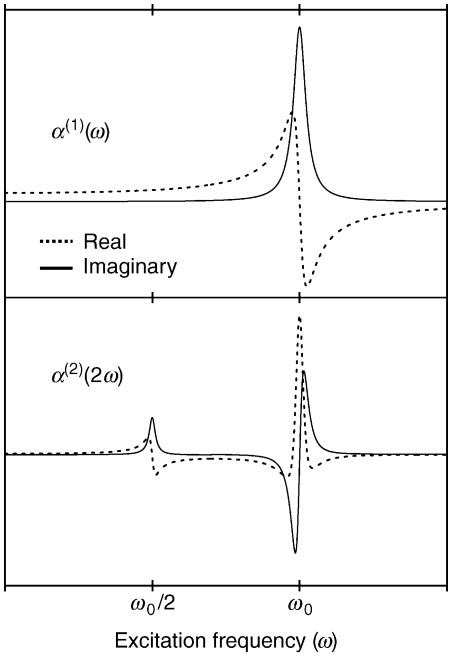


Figure 1 Real and imaginary components of the linear polarizability $\alpha^{(1)}(\omega)$ (top panel) and the second-order nonlinear polarizability $\alpha^{(2)}(2\omega = \omega + \omega)$ (bottom panel) plotted versus frequency according for the anharmonic oscillator model. For the SH response, the resonance features at $\omega = \omega_0$ and $\omega = \omega_0/2$ correspond, respectively, to single-photon and two-photon resonances.

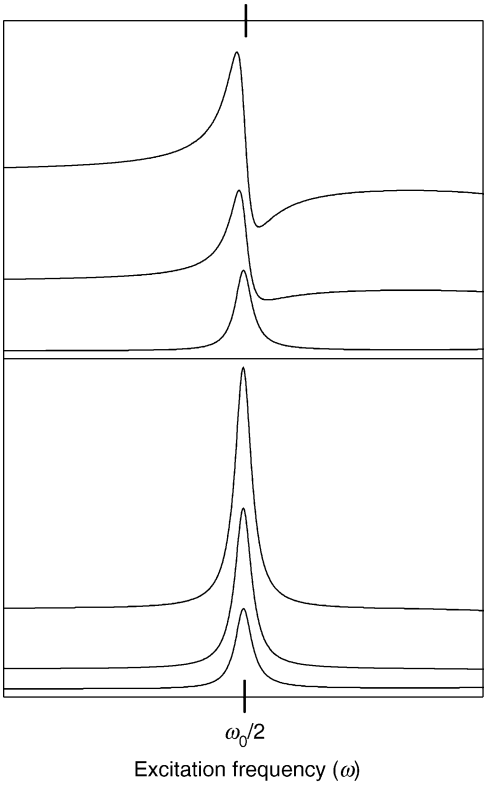


Figure 2 SH spectra $I(2\omega)$ versus pump laser frequency ω corresponding to Figure 1 in the vicinity of the two-photon resonance, with and without the presence of a nonresonant (frequency-independent) background: $I(2\omega) \sim |\alpha^{(2)}(2\omega = \omega + \omega) + k|^2$, where k represents a constant background and parametrized through $k = m|\alpha^{(2)}(2\omega = \omega_0)|$ and m is constant factor. Top panel: $m = 0$, 0.5, 1 (bottom to top curves). Bottom panel: $m = 0$, $0.5i$, i (bottom to top curves). In the top panel, the upper two lineshapes were displaced vertically to distinguish the three curves.

frequency $2\omega \approx \omega_0$. In quantum mechanical language introduced below, these peaks can be considered as arising, respectively, from matching of the fundamental and second-harmonic photon energies with the transition energy in the material.

In SHG spectroscopy, one normally measures the SH power as a function of the laser frequency. The SH field will scale with the nonlinear material response, i.e., with $\alpha^{(2)}(2\omega = \omega + \omega)$, so that the SH power will be proportional to $|\alpha^{(2)}(2\omega = \omega + \omega)|^2$. Figure 2 shows the corresponding spectra in the vicinity of the second-harmonic resonance. An important aspect of this type of spectroscopy, which is also present in other nonlinear spectroscopies, such as coherent Raman measurements, is the role that may be played by a nonresonant background. This may arise either from the off-resonant response of the system under study or from a coherent background from another material. In either case, the effect can be described by adding a spectral flat background response to the nonlinearity. Now when we detect the optical power, we then measure a quantity proportional to $|\alpha^2(2\omega = \omega + \omega) + k|^2$, where k represents the nonresonant background. The presence of k obviously elevates the baseline response near a resonance. In addition, however, depending on the relative phase of k, this extra term can introduce significant changes in the observed lineshapes. Figure 2 illustrates this effect, which must always be borne in mind in the interpretation of experimental spectra.

For the purposes of studies of surfaces and interfaces, the key property of second-order nonlinear processes is the fact that they exhibit, for centrosymmetric bulk media, an inherent interface specificity. This feature can be readily understood in the context of the nonlinear oscillator model in which the SH polarization scales with the material parameter b. Since this term is associated with a potential that varies as x^3, it is clear that the value of the b parameter can be taken as a measure of the material's departure from inversion symmetry. At the interface, such a term can be present, reflecting the inherent

asymmetry at a boundary, while it must necessarily be absent in the bulk of a centrosymmetric material.

Quantum-Mechanical Description of SHG

A correct description of SHG naturally requires a quantum mechanical treatment of the material response to the optical field. For our present purposes, we consider a localized entity with a nonlinear response, such as a non-centrosymmetric molecule at a surface. We consider below how these individual units can be added together to yield the surface nonlinear response of the material. Within this picture, we write the induced nonlinear dipole moment $\mathbf{p}^{(2)}$ of each molecule in terms of the driving electric field $\mathbf{E}$ as

$$\mathbf{p}^{(2)}(2\omega) = \boldsymbol{\alpha}^{(2)}(2\omega = \omega + \omega) : \mathbf{E}(\omega)\mathbf{E}(\omega) \qquad [8]$$

This relation is simply the generalization of eqn [5] to include a full 3-dimensional description, where $\mathbf{p}^{(2)}$ and $\mathbf{E}$ are vectors and the second-order nonlinear polarizability $\boldsymbol{\alpha}^{(2)}$ is a third-rank tensor. The quantum mechanical description enters in how we relate the response function $\boldsymbol{\alpha}^{(2)}(2\omega = \omega + \omega)$ to the underlying properties of the material.

We can obtain an expression for the second-order nonlinear polarizability by application of second-order perturbation theory with the light–matter interaction treated as the perturbation. Within the electric-dipole approximation, the interaction Hamiltonian can be taken as $H_{\text{int}} = -\boldsymbol{\mu}\cdot\mathbf{E}(t)$, where $\boldsymbol{\mu} = -e\mathbf{r}$ denotes the electric-dipole operator and $\mathbf{E}(t)$ the electric field of the driving laser beam. Using the standard density-matrix formalism for second-order perturbation theory, one calculates for the Cartesian components $\alpha^{(2)}_{ijk}$ of the tensor $\boldsymbol{\alpha}^{(2)}$ as

$$\alpha^{(2)}_{ijk}(2\omega = \omega + \omega) = -\frac{1}{\hbar^2}\sum_{g,n,n'}\left[\frac{(\mu_i)_{gn}(\mu_j)_{nn'}(\mu_k)_{n'g}}{(2\omega - \omega_{ng} + i\Gamma_{ng})(\omega - \omega_{n'g} + i\Gamma_{n'g})} + \cdots\right]\rho^{(0)}_g \qquad [9]$$

In this expression, the letters g, n', and n represent energy eigenstates $|g\rangle$, $|n'\rangle$, and $|n\rangle$ of the system, with $\rho^{(0)}_g$ corresponding to the thermal population for differing available ground states g. As illustrated in Figure 3, the SHG process can be regarded as involving a series of three transitions: two transitions associated with the absorption of two pump photons, each of energy $\hbar\omega$, and a transition associated with emission of a second-harmonic photon of energy $2\hbar\omega$. These transitions occur through the electric-dipole operator $\boldsymbol{\mu}$ and are characterized by the matrix elements $(\mu_i)_{gn}$. In eqn [9], the energy denominators involve the energy differences $\hbar\omega_{ng} \equiv E_n - E_g$ and widths $\hbar\Gamma_{ng}$ for transitions between eigenstates $|n\rangle$ and $|g\rangle$, and similarly for other combinations of states. In addition to the term indicated explicitly in eqn [9], there are seven additional terms with different Cartesian coordinates in the matrix elements and/or frequency denominators.

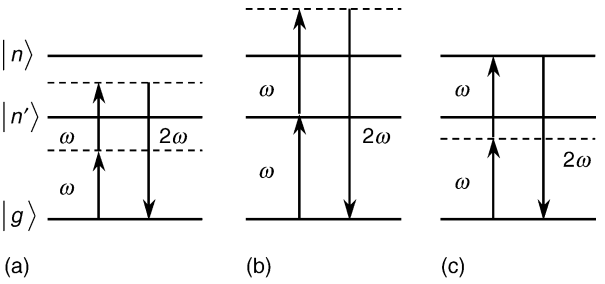


Figure 3 Quantum mechanical description of SHG process for the situations of (a) nonresonant, (b) single-photon resonant, and (c) two-photon resonant interactions. SHG occurs through transition from the initial state electronic state $|g\rangle$ and a first intermediate state $|n'\rangle$, then to a second intermediate state $|n\rangle$, followed by a return to the initial state. The process is just a parametric conversion of two photons at frequency ω into one photon at frequency 2ω.

The frequency denominators in the eight terms of eqn [9] introduce a resonant enhancement in the nonlinearity when either the fundamental frequency ω or the SH frequency 2ω coincides with a transition from a ground state $|g\rangle$ to one of the intermediate states $|n'\rangle$ or $|n\rangle$. Figure 3a shows the situation for a nonresonant nonlinear response, while Figures 3b,c illustrate, respectively, single- and two-photon resonances. The numerator in the perturbation theory expression consists of products of the three dipole matrix elements of the form $(\mu_i)_{gn}(\mu_j)_{nn'}(\mu_k)_{n'g}$. These terms reflect the structure and symmetry of the material that is built into the third-rank tensor $\alpha^{(2)}_{ijk}(2\omega = \omega + \omega)$.

The Surface Nonlinear Response

As a model of the surface nonlinearity, let us consider a monolayer of oriented molecules. The surface nonlinear response accessible to a macroscopic measurement is given, under the neglect of local-field effects, simply by summing the response of the individual molecules. The resulting surface nonlinear susceptibility tensor $\chi^{(2)}_{s,ijk}$ connecting the induced nonlinear sheet polarization to the driving electric field can then be expressed as

$$\chi^{(2)}_{s,ijk} = \frac{N_s}{\varepsilon_0}\langle T_{i\lambda}T_{j\mu}T_{k\nu}\rangle\alpha^{(2)}_{\lambda\mu\nu} \qquad [10]$$

Here N_s denotes the adsorbate surface density, $\alpha^{(2)}_{\lambda\mu\nu}$ the nonlinear polarizability of the molecule expressed in its own coordinate system, $T_{i\lambda}$ the transformation tensor from the molecule's coordinate system to the laboratory frame, $\langle\,\rangle$ an ensemble average over the orientation of the different molecules in the monolayer, and $\varepsilon_0 = 8.85 \times 10^{-12}$ F/m the permittivity of free space. This expression for the surface nonlinear susceptibility has been successful in describing properties of adsorbed molecules, including their coverage, orientation, and spectroscopic features.

For the surfaces of materials such as semiconductors or metals with delocalized electronic states, a somewhat different formulation of the nonlinear response is appropriate. A modified version of eqn [9] can be obtained using the relevant band states. The underlying concepts involved in the surface nonlinear response are, however, similar to those presented for the molecular case.

Regardless of the details of the model describing the surface nonlinear susceptibility, its tensor properties must reflect the symmetry of the interface. In that respect $\chi^{(2)}_{s,ijk}$ is quite analogous to the bulk nonlinear response, $\chi^{(2)}_{ijk}$, in a non-centrosymmetric medium. In the absence of any symmetry constraints, $\boldsymbol{\chi}^{(2)}_s$ will exhibit $3 \times 3 \times 3 = 27$ independent elements for SFG and DFG. For SHG, the order of the last two indices has no significance, and the number of independent elements is reduced to 18. If the surface exhibits a certain in-plane symmetry, then the form of $\boldsymbol{\chi}^{(2)}_s$ will be simplified correspondingly. For the common situation of an isotropic surface, for example, SHG possesses three allowed elements of $\boldsymbol{\chi}^{(2)}_s$ and may be denoted as $\chi^{(2)}_{s,\perp\perp\perp}$, $\chi^{(2)}_{s,\perp\parallel\parallel}$, $\chi^{(2)}_{s,\parallel\perp\parallel} = \chi^{(2)}_{s,\parallel\parallel\perp}$, where $\perp$ corresponds to the direction of the surface normal and $\parallel$ to an in-plane direction. The nonvanishing elements of $\boldsymbol{\chi}^{(2)}_s$ for other commonly encountered surface symmetries are summarized in several textbooks and reviews listed below.

Radiation Properties for Surface SHG

In order to probe and extract information from interfaces through surface SHG, it is necessary to understand how radiation interacts with the relevant media and gives rise to the experimentally observable signals. Here we present the principal results for the usual case of a spatially homogeneous planar interface. Nonplanar and inhomogeneous geometries, which are also of considerable interest, will be considered briefly under the heading of applications of the SHG technique below.

The SHG process is coherent in nature. Thus, for the planar geometry, a pump beam impinging on the interface will give rise to a well-collimated SH radiation emerging in distinct reflected and transmitted directions. To describe this situation in a more detailed fashion and provide formulas for the radiation efficiency, we introduce a general description of SHG by a planar interface excited by a plane wave, as shown in Figure 4. The nonlinear response of the interface described by the surface nonlinear susceptibility tensor $\boldsymbol{\chi}^{(2)}_s$ is incorporated through a nonlinear source polarization $\mathbf{P}^{(2)}$ of

$$\begin{aligned}\mathbf{P}^{(2)}(2\omega) &= \mathbf{P}^{(2)}_s(2\omega)\delta(z) \\ &= \varepsilon_0 \boldsymbol{\chi}^{(2)}_s(2\omega = \omega + \omega) : \mathbf{E}(\omega)\mathbf{E}(\omega)\delta(z) \qquad [11]\end{aligned}$$

localized at the interface ($z = 0$). In addition to the strong nonlinear response at the interface, one must also generally consider the nonlocal nonlinear response of the bulk media. This response is much weaker than the symmetry-allowed interfacial response, but is permitted even in centrosymmetric materials. Since it is present in a much larger volume, however, its cumulative effect, while generally weaker than the surface response, may be of comparable magnitude. For simplicity, however, we neglect this constant background response in our discussion in this article, but a complete discussion can be found in the referenced works. The overall SH response is naturally also influenced by the linear optical properties of the surrounding media. The two bulk media and the interfacial region are characterized, respectively, by frequency-dependent dielectric functions ε_1, ε_2, and ε', as shown in Figure 4. Again for simplicity, and in accordance with most applications, we take the linear response to be isotropic.

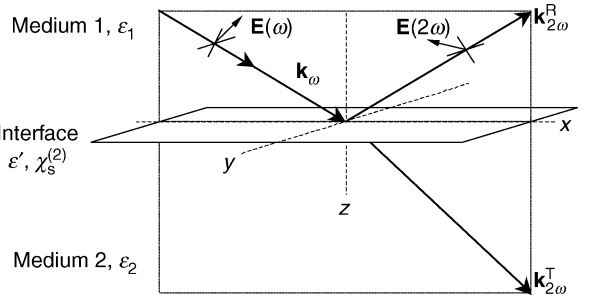


Figure 4 Schematic representation of SHG at the interface between two centrosymmetric media. The pump field $\mathbf{E}(\omega)$ at frequency ω with corresponding wavevector $\mathbf{k}_\omega$ is incident on the interface from medium 1. The SH field $\mathbf{E}(2\omega)$ at frequency 2ω is emitted in directions described by the wavevectors $\mathbf{k}^{R}_{2\omega}$ (reflected in medium 1) and $\mathbf{k}^{T}_{2\omega}$ (transmitted in medium 2). The linear dielectric constants of media 1, 2, and the interface are denoted, respectively, by ε_1, ε_2, and ε'. The nonlinear response of the interface ($z = 0$) is given by the surface nonlinear susceptibility tensor $\boldsymbol{\chi}^{(2)}_s$.

Figure 4 shows the incoming pump radiation and the reflected and transmitted SH beams. The directions of these beams are determined by conservation of the component of the momentum parallel to the interface in the SHG process, i.e., $2\mathbf{k}_{\omega,\parallel} = \mathbf{k}^{\mathrm{R}}_{2\omega,\parallel} = \mathbf{k}^{\mathrm{T}}_{2\omega,\parallel}$. Thus the reflected and transmitted beams both remain in the plane of incidence, with directions governed by the so-called nonlinear Snell's law. For bulk media without dispersion between the frequencies of ω and 2ω, the reflected and transmitted beams simply maintain the same angle of incidence as the pump beam. However, for dispersive bulk media, the directions of the beams are, however, altered, as can be understood immediately from the expression $2k_\omega/k_{2\omega} = n_\omega/n_{2\omega}$, where n denotes the refractive index of the medium. The surface SHG process does, nonetheless, differ in an important respect. In contrast to the situation for bulk media, the interfacial region is of negligible thickness, so phase shifts can obviously be disregarded. The issue of phase matching, which is of great significance for SHG from bulk media, consequently does not enter into the surface SHG problem.

By application of the Maxwell equations with the indicated nonlinear source polarization, one can derive explicit expressions for the SH radiation in terms of the linear and nonlinear response of the materials and the excitation conditions. For the case of the nonlinear reflection, the irradiance for the SH radiation can be written as

$$I(2\omega) = \frac{\omega^2 \sec^2\theta |\mathbf{e}'(2\omega)\cdot\boldsymbol{\chi}^{(2)}_{\mathrm{s}} : \mathbf{e}'(\omega)\mathbf{e}'(\omega)|^2 [I(\omega)]^2}{2\varepsilon_0 c^3 [\varepsilon_1(2\omega)]^{1/2}\varepsilon_1(\omega)} \quad [12]$$

where $I(\omega)$ denotes the irradiance of the pump beam at the fundamental frequency; c is the speed of light in vacuum; θ is the angle of emission of the SH beam with respect to the surface normal. The vectors $\mathbf{e}'(\omega)$ and $\mathbf{e}'(2\omega)$ represent the polarization vectors $\hat{\mathbf{e}}_1(\omega)$, and $\hat{\mathbf{e}}_1(2\omega)$, respectively, after they have been adjusted to account for the linear propagation of the waves to the interface. More specifically, we may write $\mathbf{e}'(\omega) = \mathbf{F}_{1\to 2}\hat{\mathbf{e}}_1(\omega)$. Here the $\mathbf{F}_{1\to 2}$ describes the relationship between the electric field $E\hat{\mathbf{e}}_1$ in medium 1 (propagating towards medium 2), which yields a field $E\mathbf{e}'$ at the interface. For light incident in the x–z plane as shown in Figure 4, $\mathbf{F}_{1\to 2}$ is a diagonal matrix whose elements are $F^{xx}_{1\to 2} = 2\varepsilon_1 k_{2,z}/(\varepsilon_2 k_{1,z} + \varepsilon_1 k_{2,z})$, $F^{yy}_{1\to 2} = 2k_{1,z}/(k_{1,z} + k_{2,z})$, and $F^{zz}_{1\to 2} = 2(\varepsilon_1\varepsilon_2/\varepsilon')k_{1,z}/(\varepsilon_2 k_{1,z} + \varepsilon_1 k_{2,z})$, where the quantity $k_{i,z}$ denotes the magnitude of the z-component of the wavevector in medium i at the relevant wavelength (ω or 2ω) (Figure 4).

A few observations may be made about eqn [12], which describes the relationship between the typical experimental observables of the pump irradiance, $I(\omega)$, and the irradiance of the emitted SH reflection, $I(2\omega)$, from the interface. First, one can immediately recognize the quadratic relationship between the pump irradiance and the SH irradiance that is expected for a second-order nonlinear optical effect. Second, the SH radiation is determined not only by the nonlinear optical response, but by the linear optical properties of the relevant media. Consequently, obtaining the surface nonlinear susceptibility tensor $\boldsymbol{\chi}^{(2)}_{\mathrm{s}}$ from experiment requires knowledge or measurement of the linear response. (It should be noted, however, that the interfacial linear properties always enter into the response in a fixed manner and can be incorporated into an effective surface nonlinear response.) The linear response can significantly enhance or reduce the efficiency of the SHG process. The use of a total-internal reflection geometry, for example, provides a significant enhancement in the SH radiation through the linear optical response of the media. Third, while full knowledge of the linear and nonlinear response will obviously allow one to predict the strength of the SH radiation for any set of conditions, a simple measurement of the SH response does not permit one to determine $\boldsymbol{\chi}^{(2)}_{\mathrm{s}}$, even if the linear optical properties are known. The determination of the tensor $\boldsymbol{\chi}^{(2)}_{\mathrm{s}}$ requires probing the material with different polarization states and/or angles of incidence. Protocols for making these measurements have been discussed extensively in the literature. For the general case of a material with in-plane anisotropy, such as the surface of a crystal, measurements of the response as the crystal is rotated about its surface normal are also very useful in a complete determination of $\boldsymbol{\chi}^{(2)}_{\mathrm{s}}$. Fourth, the simple measurement of the SH intensity as assumed here obviously eliminates information about the overall phase of the $\boldsymbol{\chi}^{(2)}_{\mathrm{s}}$. This phase information is often of considerable interest and importance. For example, if the response of the interface is dominated by an oriented molecular monolayer, the sign of $\boldsymbol{\chi}^{(2)}_{\mathrm{s}}$ will be inverted if the monolayer is reflected through the plane of the surface, i.e., the phase of $\boldsymbol{\chi}^{(2)}_{\mathrm{s}}$ yields information on the polar ordering of the layer. To obtain such phase information, one must employ a scheme using interference with a known SH reference. This approach, which is sketched briefly in the next section, is implemented quite commonly in experiment.

Experimental Considerations

The merits of the SHG technique for selective probing of interfaces are clear. On the other hand, one needs

to be able to detect the optical signal associated with a second-order nonlinear response of a material that arises from an effective thickness of just a monolayer or so. Consequently, it is important to consider various factors that aid in optimizing the strength and detectability of the SH radiation. To discuss how best to achieve this end, let us recast eqn [12] in terms of more convenient experimental parameters. We express the SH signal, S, in units of photons/s and describe the laser excitation by its pulse energy E_p, pulse width τ, and repetition rate R. For irradiation of an area A of the sample, we then obtain

$$S = \frac{\omega \sec^2\theta E_p^2 R |\chi_{eff}^{(2)}|^2}{4\hbar\varepsilon_0 c^3 \tau A} \qquad [13]$$

where $|\chi_{eff}^{(2)}| \equiv |\mathbf{e}'(2\omega)\cdot\boldsymbol{\chi}_s^{(2)} : \mathbf{e}'(\omega)\mathbf{e}'(\omega)| / [\varepsilon_1^{1/2}(2\omega)\,\varepsilon_1(\omega)]^{\frac{1}{2}}$, whose typical value is of the order of 10^{-21} m^2/V. From this relation, we see that increasing either the pulse energy or repetition rate of the laser will enhance our SH count rate. These parameters are, however, often limited by the characteristics of the available laser, as well as possible damage of the sample. For a given pulse energy and repetition rate, the SH signal is also seen to increase as the laser pulse duration or sample area is decreased. In practice, adequate SH signals can generally be obtained from high-power Q-switched lasers (with nanosecond pulses) with a fairly large focal spot. A preferred solution in most cases is, however, provided by modelocked lasers with high repetition rate and pulses of sub-picosecond duration. Surface SH signals obtained, for example, using a femtosecond modelocked Ti:sapphire oscillator, are typically in the range of 10^2–10^6 photons/s. Although these signals are relatively weak, they are easily detectable with photomultiplier tubes (PMT) and appropriate electronics.

The basic arrangement for surface SHG measurements is shown in Figure 5. It makes use of the directionality and frequency selectivity of the SHG process not only to detect the desired signal with high efficiency, but to discriminate strongly against background signals, including the reflected laser beam. In the setup, irradiation of the sample involves a pulsed laser source, LS; a halfwave plate, λ, and polarizer, P, for controlling the input polarization. Spectral control is achieved with a filter F1 that passes the fundamental radiation in the pump beam, but rejects other wavelengths, particularly spurious SH radiation arising from other optical components; and a color filter F2 that passes SH, but rejects the reflected fundamental radiation. In addition, a spectrometer is commonly employed to limit the detection bandwidth further. An analyzer A serves to select the desired SH polarization. The SH signal is then measured by a high-quantum efficiency detector and the signal is monitored by processing electronics, typically a photon counter, a gated integrator, or a lock-in amplifier, depending on the strength of the signal and the repetition rate of the laser. The laser source may be either of fixed wavelength or tunable for spectroscopic measurements.

There are other important experimental configurations for surface SHG that extend beyond the basic SH reflection geometry, shown schematically in Figure 6a. A common improvement (Figure 6b) is the implementation of a SH reference in which a small fraction of the pump radiation is directed into a crystal with a bulk second-order nonlinear response. This approach provides a reference against which to compare the measured signal, thus compensating appropriately for pulse-to-pulse variation in duration,

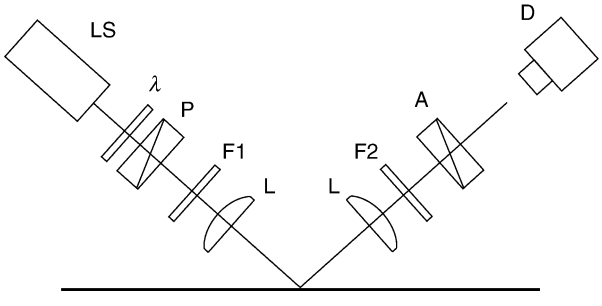


Figure 5 Schematic representation of a basic surface experimental setup for surface SHG. It is comprised of laser source (LS), halfwave plate (λ), polarizer (P), lenses (L), spectral filters (F1, F2), analyzer (A), and detector (D).

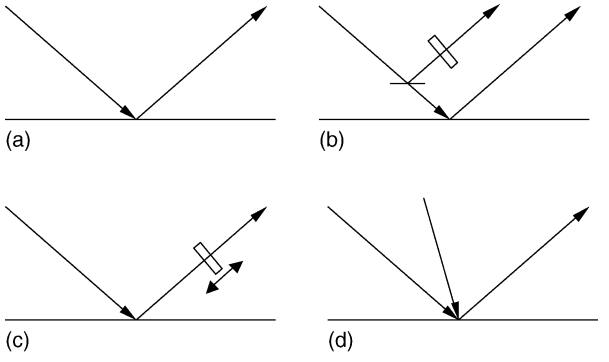


Figure 6 Various schematic representations of different experimental schemes for surface SH measurements. Panel (a) represents the basic scheme as in Figure 5. Panel (b) describes use of a separate SH source as a reference for normalizing the surface SH signals against laser fluctuation and drift. Panel (c) represents the arrangement for a measurement of the phase of the surface SH signal by mixing the fields from the surface and an external source SH radiation. Panel (d) describes the configuration for a time-resolved measurement using the pump-probe scheme.

energy, and spatial profile, as well as for long-term drifts. This type of normalization (against a material with a suitably flat spectral response) is essential for spectroscopy measurements in which the laser frequency is scanned. In addition, use of a calibrated reference is the best way of determining the surface nonlinear response in absolute terms.

It may also be desirable to measure the phase of the signal. For this type of an experiment, a reference field, $\mathbf{E}_{\text{ref}}(2\omega)$, of similar amplitude and known phase response is generated along the path of the surface SH signal $\mathbf{E}(2\omega)$ (Figure 6c). The two collinear SH beams of the same polarization interfere at the detector. The SH power scales as $|\mathbf{E}_{\text{ref}}(2\omega) + \mathbf{E}(2\omega)|^2 = |\mathbf{E}_{\text{ref}}|^2 + 2E_{\text{ref}}E\cos\varphi + |\mathbf{E}|^2$, where φ is the relative phase of the two fields. The relative phase φ may be varied by translating a reference plate along the path of the beam and making use of the dispersion of the ambient air or by inserting another control phase-shifting element. A variant of this phase measurement method is the homodyne scheme in which the reference electric field amplitude $\mathbf{E}_{\text{ref}}(2\omega)$ is significantly stronger than that of the signal amplitude $\mathbf{E}(2\omega)$. The measured SH power then scales as $|\mathbf{E}_{\text{ref}}|^2 + 2E_{\text{ref}}E\cos\varphi$, which yields a signal that varies linearly with the desired SH field strength from the sample.

In the SHG technique, a short-pulse laser is used to interrogate the sample and thus the measurement provides high time-resolution. To make use of this capability, we can add a pump pulse to excite the sample away from equilibrium. This provides the opportunity to examine ultrafast surface dynamics induced by excitation of the sample with a suitable pump laser pulse. For such measurements (Figure 6d), the pump pulse is generally derived from the same laser source as the probe for the SHG measurement, although one or both of these pulses may undergo frequency conversion before being applied to the surface. The critical element is that time synchronism is maintained. The dynamics can then be followed by repeating the SHG measurement at various time delays relative to pump excitation, as controlled by an optical delay line. The ultimate time resolution of this technique is limited only by the duration of the laser pulses and may be as short as a few femtoseconds with state-of-the-art modelocked lasers.

Applications of Surface SHG

The technique of surface SHG is extremely flexible in application, since it relies only on basic symmetry properties for its remarkable surface or interface specificity. SHG measurements can provide quite varied information about the interface, depending on the mode of application of the technique and the material system under study. In this section, we attempt to illustrate the principal types of measurements that can be performed with SHG and the information that can be obtained thereby. For systems with molecular adsorbates, SHG studies can yield information on the density and orientation of molecular species, and can typically do so in real time. For crystalline surfaces, the symmetry and order of the surface can be examined through studies of tensor properties of the surface nonlinear susceptibility. Spectroscopic studies of electronic transitions are also of great value. These can be accomplished for both solids and molecular systems through the same experimental approach of tuning the frequency of the laser source. In all of these investigations, one can achieve very high time resolution, down into the femtosecond range, by the use of pump-probe techniques. In addition, as we discuss below, SHG provides interesting additional capabilities for systems with applied electric and magnetic fields. Finally, the method permits one to probe spatially inhomogeneous systems.

In addition to this diverse range of information that can be provided by surface SHG measurements, the technique is unique in the extremely wide range of material and chemical systems to which it can be – and, indeed, has been – applied. These span the gamut from solids under ultrahigh vacuum to solid/solid, solid/liquid, liquid/vapor, and even liquid/liquid interfaces. Since our examples were drawn to illustrate the type of information obtainable, rather than the range of material systems investigated, the reader is unfortunately obliged to refer to other more detailed treatments of surface SHG measurements to appreciate fully this important facet of the technique.

Adsorbate Density

The remarkably high sensitivity of the SHG technique to the nature of the surface is illustrated by the data in Figure 7, which shows the dramatic change in the SH response from a clean Si(111)-7 × 7 surface upon adsorption of atomic hydrogen. Indeed, it is possible to detect well below 1% of a monolayer (ML) of adsorbed hydrogen. A useful point to note is that the change in $\chi_s^{(2)}$ is linear with adsorbate density for modest values of the coverage, θ. Thus, it is quite easy to obtain a relative gauge of the adsorbate coverage in this regime.

The behavior of the SH response to adsorbates at low coverages can often be described by a simple linearized model of the surface nonlinear susceptibility of $\chi_s^{(2)} = (1 - a\theta)\chi_{s,1}^{(2)} + \theta\chi_{s,2}^{(2)}$, where

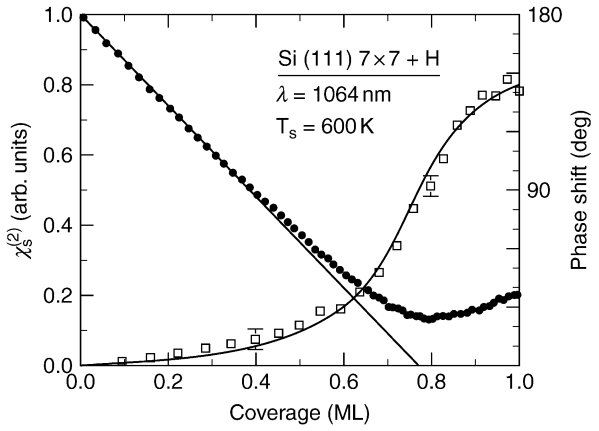


Figure 7 Dependence of the magnitude (full circles) and the phase (open squares) of the nonlinear susceptibility $\chi_s^{(2)}$ for a Si(111)-7 × 7 surface as a function of the coverage of adsorbed atomic hydrogen using a fundamental photon energy 1.17 eV. Reproduced with permission from Höfer U (1996) Nonlinear optical investigations of the dynamics of hydrogen interaction with silicon surfaces. *Applied Physics A* 63: 533.

$\chi_{s,i}^{(2)}$ $(i = 1, 2)$ are complex quantities and a is a constant. This relation describes the surface nonlinear susceptibility as arising from a contribution $\chi_{s,1}^{(2)}$ from the clean surface and a contribution $\chi_{s,2}^{(2)}$ from a saturated monolayer $(\theta = 1)$ of the adsorbate. The term $-a\theta\chi_{s,1}^{(2)}$ then represents the modification of the surface nonlinearity arising from interaction with the adsorbate. For the case of hydrogen adsorption of Figure 7, this is the dominant term, since hydrogen atoms would have no nonlinearity on their own. In other systems, however, the response can be considered as arising essentially from the nonlinear response of aligned adsorbed molecules. The simple linearity of this relation of $\chi_s^{(2)}$ with adsorbate coverage obviously neglects important effects such as adsorbate–adsorbate chemical interactions, as well as local-field effects and the possible influence of differing adsorption sites. All of these factors have been considered in the literature. Nonetheless, the linearity of response with adsorbate coverage is often found to be a good approximation, even for fairly high adsorbate coverages. From the point of view of simply following adsorption dynamics with SHG, it is not, we note, necessary to have any a priori knowledge of the relation between the adsorbate coverage and the SH response. This relation can be established empirically in static measurements and then used advantageously to follow the dynamics of surface processes.

Surface Symmetry and Molecular Orientation

The surface SHG process is described by the third-rank tensor $\boldsymbol{\chi}_s^{(2)}$. The tensorial properties of this response reflect the symmetry of the surface. Other than for a few cases of high symmetry, such a third-rank tensor gives a distinct signature of surfaces with crystalline order compared to surfaces, such as disordered ones, exhibiting effective in-plane isotropy. The form of the $\boldsymbol{\chi}_s^{(2)}$ tensor can be probed both by measurements that vary the polarization of the fundamental and second-harmonic beams and by measurements in which the plane of incidence is changed with respect to the orientation of the surface, generally by rotating the sample about its normal with a fixed experimental geometry.

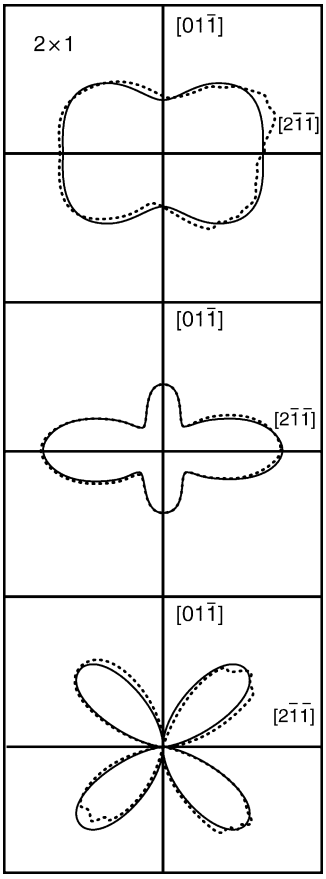


Figure 8 Polarization dependence of SHG for normal incidence excitation of the cleaved Si(111)-2 × 1 surface. The polar plot of the SH signal versus the pump polarization orientation is shown for the case of no polarization analysis (top panel), for the SH analyzer along the [2$\bar{1}\bar{1}$] crystallographic direction of the surface (middle panel), and along the [01$\bar{1}$] direction (bottom panel). The solid curves are fits to theory assuming that the surface has *m* symmetry. Reproduced with permission from Heinz TF, Log MMT and Thomson WA (1985) Study of Si(111) surfaces by optical second-harmonic generation: reconstruction and surface phase-transformation. *Physical Review Letters* 54: 63.

Figure 8 illustrates the capability of surface SHG measurements to probe surface symmetry. The figure shows SHG polarization dependence for a Si(111) surface cleaved in ultrahigh vacuum. While the

underlying bulk symmetry suggests that the surface might exhibit 3m symmetry, the analysis of data indicates that only a single mirror m plane is present. This lowered symmetry corresponds to the metastable 2×1 reconstruction induced by the crystal cleavage. By heating the sample, one can observe a phase transformation to the equilibrium 7×7 reconstruction and the emergence of the corresponding 3m surface symmetry.

The analysis of surface or interface symmetry by SHG relies only on the form of the nonlinear susceptibility tensor, i.e., which elements are independent and nonvanishing. Within the context of a microscopic model of the surface response, however, the relative values of the tensor elements of $\boldsymbol{\chi}_s^{(2)}$ can also yield valuable information about the nature of the surface. This approach has been developed extensively in the context of the analysis of molecular orientation at surfaces. The basic notion behind this treatment relies on the knowledge of the microscopic molecular nonlinear polarizability $\alpha^{(2)}_{\lambda\mu\nu}$ and the measured surface nonlinear susceptibility $\chi^{(2)}_{s,ijk}$ to infer information about the orientational average that relates these quantities. As discussed above in conjunction with eqn [10], under the neglect of local-field corrections, the relationship between these two quantities involves the ensemble average of the coordinate transformation tensor from the molecular to the laboratory coordinate systems $\langle T_{i\lambda} T_{j\mu} T_{k\nu} \rangle$. Hence a knowledge of the quantities $\alpha^{(2)}_{\lambda\mu\nu}$, N_s, and $\chi^{(2)}_{s,ijk}$ allows us to infer information about the orientational distribution of the molecules as described by the quantity $\langle T_{i\lambda} T_{j\mu} T_{k\nu} \rangle$. Even in the absence of knowledge of the magnitudes of the nonlinear polarizabilities and the adsorbate density, one may still obtain useful information for molecules with a simple form of $\alpha^{(2)}_{\lambda\mu\nu}$.

An important example of this analysis of molecular orientation concerns the case of molecules with a dominant nonlinear polarizability along a single axis, i.e., $\boldsymbol{\alpha}^{(2)} = \alpha^{(2)}_{z'z'z'} \hat{\mathbf{z}}' \hat{\mathbf{z}}' \hat{\mathbf{z}}'$. For the usual situation of in-plane isotropy, one can then show that the nonlinear susceptibility elements are given by $\chi^{(2)}_{s,\perp\perp\perp} = N_s \langle \cos^3\theta \rangle \alpha^{(2)}_{z'z'z'}/\varepsilon_0$ and $\chi^{(2)}_{s,\parallel\perp\parallel} = \chi^{(2)}_{s,\parallel\parallel\perp} = \frac{1}{2} N_s \langle \cos\theta \times \sin^2\theta \rangle \ \alpha^{(2)}_{z'z'z'}/\varepsilon_0$, where θ denotes the angle between the surface normal and the z' molecular axis. It should be noted that these orientational moments, reflecting their origin in a third-rank tensor, give a signature of the polar ordering, since a sign change in the elements of $\boldsymbol{\chi}_s^{(2)}$ follows if the molecular orientation is inverted. Also, when only one element of $\boldsymbol{\alpha}^{(2)}$ is present, we can form ratios of the experimentally measurable ratio of tensor elements that depend only on the molecular orientational factors, such as $R \equiv (2\chi^{(2)}_{s,\parallel\perp\parallel} + \chi^{(2)}_{s,\perp\perp\perp})/\chi^{(2)}_{s,\perp\perp\perp} = \langle \cos\theta \rangle / \langle \cos^3\theta \rangle$. For a narrow distribution of the orientations centered at θ_0, this ratio yields directly the molecular orientation angle through $R \approx \sec^2\theta_0$.

This type of molecular orientation analysis has been applied to a wide range of material systems. The method, while powerful, relies on several assumptions, including an adequate knowledge of the molecular nonlinear polarizability, the lack of significant local field corrections, the availability of information about the effective linear dielectric constant in the interfacial region, and a suitable method of eliminating any substrate nonlinear response in the measurement. The method is, consequently, best suited for adsorbed layer of molecules with strong and well-defined nonlinear response, present at relatively modest coverages. The use of resonant excitation is a valuable adjunct in these studies, since it generally simplifies the form and strengthens the molecular nonlinear response. An analogous approach to the analysis of molecular orientation can be accomplished with infrared–visible sum-frequency generation. These measurements often provide a superior method of analysis, since the molecular response and its symmetry are better defined by resonant excitation of a molecular vibration.

Surface and Interface Spectroscopy

Surface SHG can be readily adapted to yield spectroscopic information on surface transitions. As indicated above, both resonances at the fundamental and second-harmonic frequencies are manifested in the SHG response. The examination of resonances at the SH frequency is often convenient, since it permits the application of a relatively intense pump radiation at a frequency well separated from that of the resonance. Such SHG spectroscopy measurements have been widely applied to examine electronic transitions at surfaces and interfaces of solids, as well as in adsorbed molecules under a wide variety of conditions.

An example of SHG spectroscopy of electronic transitions at a buried solid/solid interface is presented in Figure 9. For the epitaxial CaF_2/Si(111) interface, distinctive new electronic transitions, associated with interface states, are observed that are absent from the response of either of the bulk materials. These excitations occur in a spectral range where the bulk silicon sample is highly absorbing, which would render the identification of the new interface transitions very difficult using conventional linear optics. The buried nature of the interface naturally also constrains application of conventional surface probing techniques involving electron

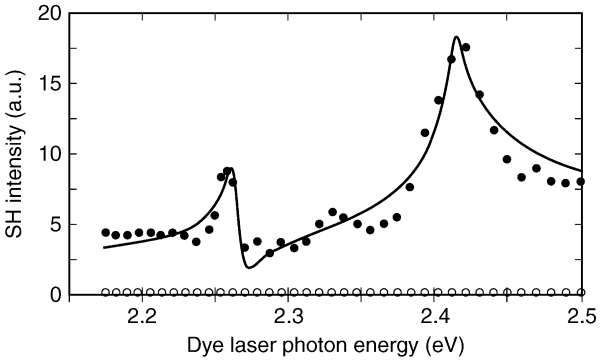


Figure 9 SHG spectrum (solid dots) for the CaF_2/Si(111) interface. For comparison, the open circles represent the corresponding signals for the SiO_2/Si(111) interface, which is seen to give a very weak signal in this frequency range. The solid line is a theoretical fit based on a two-dimensional direct transition of the interface states, together with an excitonic feature slightly below the band gap. Reproduced with permission from Heinz TF, Himpsel FJ, Palange E and Burstein E (1989) Electronic transitions at the CaF_2/Si(111) interface probed by resonant three-wave-mixing spectroscopy. *Physical Review Letters* 63: 644.

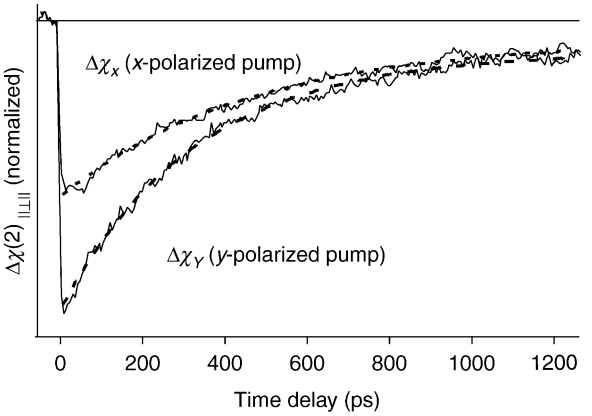


Figure 10 Rotational relaxation of Coumarin 314 molecules at the air–water interface subsequent to photoexcitation. The initial anisotropy in the molecular orientational distribution is created by linearly polarized pump radiation in two orthogonal directions in the surface. Reproduced with permission from Zimdars D, Dadap JI, Eilsenthal KB and Heinz TF (1999) Anisotropic orientational motion of molecular adsorbates at the air–water interface. *Journal of Physical Chemistry B* 103: 3425–3433.

irradiation or emission. Two additional general observations are appropriate. First, the lineshape analysis for such measurements must take into account the fact that the SHG response derives from both the real and imaginary parts of $\boldsymbol{\chi}_s^{(2)}$, as we discussed above. Second, in the absence of any other knowledge, SHG spectroscopic data contain an ambiguity about whether the resonances correspond to the fundamental or SH photon energy. This ambiguity, however, can be resolved in a straightforward fashion by complementing the SHG measurement with a sum-frequency measurement in which one frequency is scanned and one is held fixed, as was demonstrated in the study associated with Figure 9.

Time-Resolved Measurements

One of the fascinating frontiers that can be addressed by SHG measurements is probing the ultrafast dynamics of surfaces and interfaces by the pump-probe method. This possibility is exemplified by a variety of studies ranging from the dynamics of laser-driven phase transitions, charge-transfer, solvation, phonon dynamics, and orientational and torsional dynamics.

An example of applying time-resolved SHG to examine ultrafast molecular dynamics at the liquid/vapor interface of water is presented in Figure 10. In these studies, the rotational motion of dye molecules at the interface, which probes the local environment, is followed in real time. This is accomplished by photo-exciting the dye molecules at the interface. Because polarized pump radiation is used, an anisotropic orientational distribution of excited-state molecules is established. Since the excited molecules have a different SHG response, a change in the signal is observed immediately. Subsequently, the anisotropic orientational distribution relaxes to the equilibrium configuration. The time evolution of the rotational anisotropy is followed by detecting the SH of a probe laser pulse as a function of the delay time. Through a comparison of the results for different initial anisotropic distributions (produced by pump beams with differing polarizations), one may deduce rates for both in-plane and out-of-plane orientational relaxation. This example illustrates the possibilities of time-resolved SHG measurements for probing the detailed dynamics of molecules at an interface on the ultrafast timescale.

Spatially Resolved Measurements

Up to this point, our discussions have centered on spatially homogenous surfaces and interfaces. Surface SHG has also been exploited to probe the lateral variation of inhomogeneous surfaces with the micron to sub-micron resolution characteristic of optical microscopy techniques. Spatial resolution may be achieved simply by detecting the nonlinear response with a focused laser beam that is scanned across the surface. A large area of the surface may also be illuminated and the emitted nonlinear radiation imaged. A wide range of applications of this imaging capability has emerged, including the probing of biological materials, such as cells and tissues, to probing spatially varying electric fields

and magnetic domains. Within the biological context, SH microscopy has recently become an active research direction. In these applications, while probing with SHG typically senses non-centrosymmetric structures in bulk samples as opposed to interfaces, most of the aspects of SHG presented in this article still apply. To further improve the spatial resolution of SHG microscopy, researchers have also employed near-field techniques for spatial resolutions below the diffraction limit.

Complementary approaches to the real-space imaging just described have also been developed based on SH diffraction. In this scheme, periodically modulated surfaces are established holographically on a surface by utilizing two interfering laser beams. In addition to the reflected and transmitted SH beams, additional beams are then generated at well-defined angles as dictated by the grating period. This approach has been used to monitor the relaxation of a monolayer grating of adsorbates in time, which yields precise information on the rate of adsorbate surface diffusion.

Probing Electric Fields at Interfaces

For centrosymmetric media, the SHG process shows a high degree of sensitivity to the presence of electric fields. This property follows from the fact that an applied electric field can break the inversion symmetry, just as an interface does. The resulting process, termed electric-field induced SHG (or EFISH), has been known for many years. From a phenomenological standpoint, it can usually be described as a nonlinear response of the form $\mathbf{P}^{(3)}(2\omega) = \varepsilon_0 \boldsymbol{\chi}^{(3)} : \mathbf{E}(\omega)\mathbf{E}(\omega)\mathbf{E}$, where $\mathbf{E}$ is the electric field being probed and $\mathbf{E}(\omega)$ is the laser pump field. Within the context of surfaces and interfaces, there are many circumstances where probing such electric fields is of interest and importance. The SHG technique provides high sensitivity, good spatial resolution, accessibility to buried interfaces, and, in the pump-probe implementation, extremely high time resolution. In addition, with appropriate measurements of the polarization dependence and homodyne detection, the full vector character of the electric field can be determined.

The behavior of electric-field induced SHG is similar to that arising from the asymmetry of an interface. Indeed, it is formally equivalent to the response of an interface provided that the electric field is present in (or radiates from) a region of less than a wavelength's thickness. The applications of this measurement capability are quite diverse, ranging from studies of charging and acid/base reactions at liquid/solid interfaces to probing fields in metal-oxide-semiconductor structures and Schottky barriers to the measurement of freely-propagating terahertz radiation. In Figure 11,

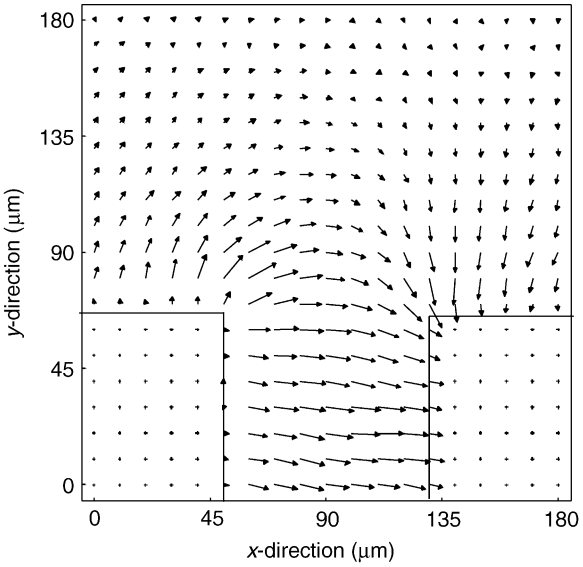


Figure 11 Vector map of the electric field in Si-on-sapphire substrate as deduced by SHG measurements. A voltage is imposed on the rectangular metallic electrodes, which are separated by 80 μm. Reproduced with permission from Dadap JI, Shan J, Weling AS, Misewich JA, Nahata A and Heinz TF (1999) Measurement of the vector character of electric fields by optical second-harmonic generation. *Optics Letters* 24: 1059–1061.

SHG with appropriate polarization control and homodyne detection scheme has been applied to the measurement of the vector character of electric fields present in a dc-biased dipole structure on silicon. By spatially scanning the laser beam, a two-dimensional electric field map can be produced. In a pump/probe scheme, the same measurements can be made with femtosecond time resolution.

Probing Magnetization at Interfaces

In contrast to the application of an external electric field (a polar vector), the presence of an external magnetic field (an axial vector) does not break the inversion symmetry of a material. Thus, SHG retains its interface selectivity for magnetic effects. SHG measurements, as an interface-specific probe, thus provide an excellent complement to the linear magneto-optical Kerr effect (MOKE), which is sensitive to magnetic effects within the penetration depth of the light.

The ability of the SHG technique to sense the magnetization at a surface is illustrated in Figure 12 for a film of epitaxially-grown magnetic Co on nonmagnetic substrate Cu. A hysteresis loop of the magnetization versus the strength of an applied magnetic field is shown, where the magnetization is gauged from the SHG response. Through measurements of the response as a function of the thickness of

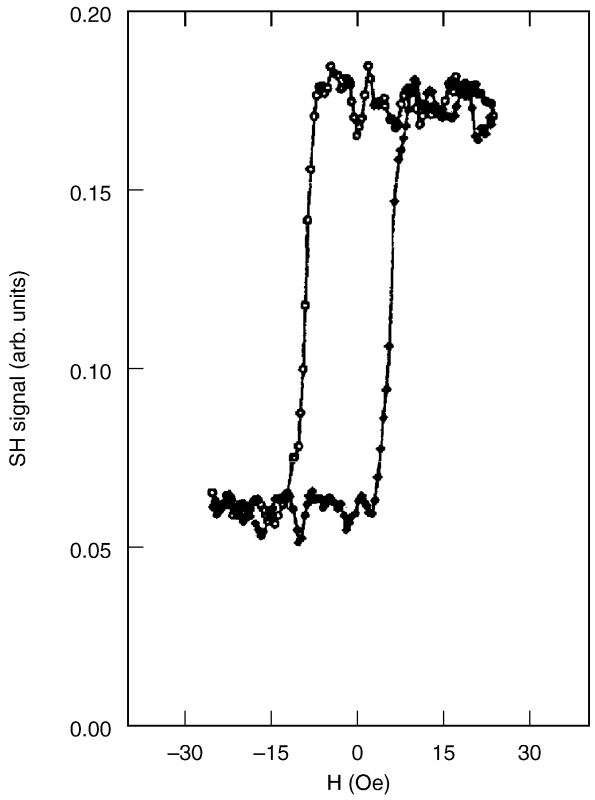


Figure 12 SH response of a Co film on a Cu(001) substrate as a function of the applied magnetic field. A hysteresis loop corresponding to the surface response is observed. Reproduced with permission from Wierenga HA, de Jong W, Prins MWJ *et al.* (1995) Interface magnetism and possible quantum-well oscillations in ultrathin Co/Cu films observed by magnetization induced second harmonic generation. *Physical Review Letters* 74: 1462–1465.

the magnetic film, the researchers demonstrated that the SH response reflected the interface properties, rather than those of the bulk film.

Probing Micro- and Nanoscale Structures

In addition to the microscopy studies described above that probe spatially inhomogeneous surfaces, there has been considerable interest in SHG from a variety of strongly textured surfaces and small particles. From the standpoint of the optical response of materials, particular interest has been attached to studies of metallic surfaces with surface roughness. For appropriate roughness of low-loss materials, such as silver, very significant local-field enhancements can be achieved. Since these field enhancements enter into the strength of the SH emission nonlinearly, overall enhancements in the SH power by several orders of magnitude have been demonstrated. This provides both a useful method of study of such surface enhancement effects and a scheme for amplifying the relatively weak surface nonlinear response.

Beyond these problems associated with roughened surfaces, considerable attention has recently been directed towards the application of the SHG technique for probing micro- and nano-scale particles. The fundamental principle of the interface selectivity of the SHG process for such systems comprised of centrosymmetric materials remains unaltered. The interaction of these particles with the radiation fields is, however, significantly modified. For the case of particles with dimensions of several optical wavelengths, the overall SHG signal strength may be comparable to that from a corresponding area of a planar surface. The radiation pattern will, of course, be strongly modified and will generally be diffuse in character. While this makes detection (and background discrimination) somewhat more difficult than for the planar geometry, many SHG measurements of surface and interface properties have been carried out in this regime. As the particle size becomes progressively smaller, particles with overall inversion symmetry begin to exhibit appreciable cancellation of the SH emission. Indeed, in the limit of vanishing size, full cancellation is expected. This tendency is, however, offset by the increasing surface-to-bulk ratio and the possibility of increasing the number of particles probed. It has been demonstrated experimentally that the surfaces of particles in the nanometer length scale can indeed be probed by SHG.

From the point of view of applications, this extension of the SHG technique to probe micro- and nano-structured materials opens up a rich variety of new possibilities. Results have been reported on metal and semiconductor nanoparticles in suspension, as well as on colloids of minerals. In addition, liquid droplets have been investigated. Self-assembling structures such as liposomes and vesicles have been examined. For these structures, many of the same capabilities of SHG as demonstrated for the planar geometry, such as probing of adsorption and charging effects, have been demonstrated. More uniquely, molecular transport *through* the liposome membrane could also be observed and characterized. The possibility of extension of these measurements to biological systems is of obvious interest and promise.

Concluding Remarks

In this article, we provided a brief review of surface SHG as a technique for interface-specific probing of materials. In addition to discussing some of the fundamental ideas underlying the method, we outlined the wide variety of information that may be obtained from such measurements, ranging from

surface symmetry and molecular ordering to interface spectroscopy and ultrafast dynamics. While several aspects of the method have now become quite well established, developments in widely available modelocked lasers with excellent performance characteristics have greatly enhanced the ease of making surface SHG measurements in a wide class of material systems. In addition, new developments and applications of the method have continued to emerge apace. Of particular note in this regard are applications to probe electric fields and magnetization at interfaces, as well as advances in the use of the method for interface-specific probing of micro- and nanoscale structures. The sharp increase in applications of SHG for probing systems of biological importance constitutes another important development, and one with much promise for the future.

See also

Materials for Nonlinear Optics: Organic Nonlinear Materials. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation. **Physical Applications of Lasers:** Sum-Frequency Generation at Surfaces.

Further Reading

Bennemann KH (ed.) (1998) *Nonlinear Optics in Metals*. Oxford, UK: Clarendon.

Bloembergen N (1965) *Nonlinear Optics*. New York: WA Benjamin.

Boyd RW (1992) *Nonlinear Optics*. New York: Academic.

Brevet PF (1997) *Surface Second Harmonic Generation*. Lausanne, Switzerland: Presses Polytechniques et Universitaires Romandes.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge.

Corn RM and Higgins DA (1994) Optical second-harmonic generation as a probe of surface-chemistry. *Chemical Review* 94: 107–125.

Dadap JI and Heinz TF (2001) Nonlinear optical spectroscopy of surfaces and interfaces. In: Moore JH and Spencer ND (eds) *Encyclopedia of Chemical Physics and Physical Chemistry*, pp. 1089–1125. Bristol, UK: IOP.

Eisenthal KB (1996) Liquid interfaces probed by second-harmonic and sum-frequency spectroscopy. *Chemical Review* 96: 1343–1360.

Heinz TF (1991) Second-order nonlinear optical effects at surfaces and interfaces. In: Ponath H-E and Stegeman GI (eds) *Nonlinear Surface Electromagnetic Phenomena*, pp. 353–416. Amsterdam: North Holland.

Liebsch A (1995) Electronic excitations at metal surfaces. In: Halevi P (ed.) *Photonic Probes of Surfaces*, pp. 479–532. Amsterdam: Elsevier.

Lüpke G (1999) Characterization of semiconductor interfaces by second-harmonic generation. *Surface Science Report* 35: 75–161.

McGilp JF (1995) Optical characterisation of semiconductor surfaces and interfaces. *Progress in Surface Science* 49: 1–106.

McGilp JF (1999) Second-harmonic generation at semiconductor and metal surfaces. *Surface Review Letters* 6: 529–558.

McGilp JF, Patterson CH and Weaire DL (eds) (1995) *Epioptics: Linear and Nonlinear Optical Spectroscopy of Surfaces and Interfaces*. Berlin, New York: Springer.

Prasad PN and Williams DJ (1991) *Introduction to Nonlinear Optical Effects in Molecules and Polymers*. New York: Wiley.

Reider GA and Heinz TF (1995) Second-order nonlinear optical effects at surfaces and interfaces: recent advances. In: Halevi P (ed.) *Photonic Probes of Surfaces*, pp. 413–478. Amsterdam: Elsevier.

Richmond GL (1997) Vibrational spectroscopy of molecules at liquid/liquid interfaces. *Analytical Chemistry* 69: A536–A543.

Richmond GL, Robinson JM and Shannon VL (1988) Second harmonic generation studies of interfacial structure and dynamics. *Progress in Surface Science* 28: 1–70.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: Wiley.

Shen YR (1989) Optical second harmonic-generation at interfaces. *Annual Review of Physical Chemistry* 40: 327–350.

Single Molecule Spectroscopy

X Michalet and S Weiss, University of California, Los Angeles, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Textbooks in physics, chemistry, and biology abound in diagrams depicting single-particle, single-molecule or single-enzyme properties, reactions or interactions with their environment. Such pictures and the numbers associated with them, however, were, until recently, extrapolated from measurements performed on large ensembles of molecules, which only yield quantities averaged out over time, space, conformations and/or the local environment. Advances in detector sensitivity and improvements in instrument design do now provide scientists

with tools to probe single molecules with light, and monitor their photophysical properties with exquisite sensitivity and time resolution. This spectroscopic information is used to explore the molecular structure, conformational dynamics, local environment, and intermolecular interactions. The collected data generate a wealth of new information on single-molecule physical chemistry or biomechanics, but also in polymer gel physics or glass dynamics and other domains of solid state physics.

This article reviews the field of single-molecule spectroscopy. First, the principles underlying single-molecule spectroscopy approaches are introduced, and this is followed by a discussion of the different spectroscopic techniques available to the experimentalist and what kind of information can be extracted with each of them. It also discusses important photophysical properties of molecules that influence the outcome of single-molecule spectroscopic measurements. The next section presents some applications of these techniques, focusing on physical properties of materials and biological applications of single-molecule spectroscopy. The last part briefly reviews recent developments and outlines the future prospects of single-molecule spectroscopy.

Principles of Single-Molecule Spectroscopy

Motivation

Measurements performed on ensembles of molecules only yield averaged-out information. This averaging process is of several kinds: average over molecular heterogeneity (e.g., chemical composition, oxidation state, or spatial conformation) and average over different or changing local environments. Moreover, dynamic processes studied by ensemble measurements cannot sort out molecules exhibiting fast or slow reaction times, although such differences are expected from the stochastic nature of any chemical kinetics. The interest in single-molecule analysis is for these reasons manifold. At low temperature, being able to study a single molecule and its evolution under the influence of various external parameters allows for precise tests of molecular quantum electrodynamics. At higher temperature, it gives access to the temporal evolution of different nanodomains of a bulk material in which single molecules are embedded as reporters of their environment. In biology and biochemistry, the building blocks of any process are individual molecules (genes, enzymes, motors, filaments, etc.); the ability to fully understand their kinetic behavior at the single-molecule level is thus of utmost importance. Finally, reliable observation of single molecules is a prerequisite of the full development of nanotechnology and quantum computing.

Historical Outlook

Single-molecule experiments were first performed on isolated ion channels by the patch clamp technique, which allows the recording of single-ion translocation through a pore-like protein embedded in a fluid membrane. Scanning tunneling and atomic force microscope observations later opened the way to investigations of nanometer-scale objects on surfaces. Only recently have optical methods reached the sensitivity required to detect, image, manipulate, and follow the spectroscopic evolution of single molecules on surfaces, in solids, liquids, and biological membranes as well as inside live cells.

Single-molecule spectroscopy (SMS) was initially demonstrated in a system of pentacene molecules embedded in a *p*-terphenyl host crystal at liquid helium temperature. Absorption (leading to fluorescence excitation) spectroscopy in this system takes advantage of the very narrow zero-phonon absorption line (ZPL) of the rigid pentacene molecule (Figure 1). The position in laser-frequency space of each molecule's ZPL indeed depends on the local fields, allowing for precise molecular selectivity. However this approach is limited to a few molecule–host pairs and requires cryogenic techniques, due to broadening of the ZPL above 10 K.

Room temperature studies started with total internal reflection (TIR) microscopy and standard flow-cytometry methods in the mid-1980s. In the latter case, single molecules randomly passing through the excitation volume gave rise to sudden bursts of photons emitted during their brief transit. Initially developed for rapid DNA sequencing, this simple approach is now widely used for more sophisticated studies in its confocal version (Figure 2), which will be detailed in the following.

The first image of a single molecule at room temperature was later obtained using scanning near-field optical microscopy (Figure 3), rapidly followed by similar achievements using far-field confocal microscopy, at the expense of a lower spatial resolution, but with much more ease of use. TIR and wide-field fluorescence imaging are now used to image and study the two-dimensional and three-dimensional diffusion trajectories of single fluorescent molecules (Figure 4), making a come-back after the pioneering studies of the early 1980s.

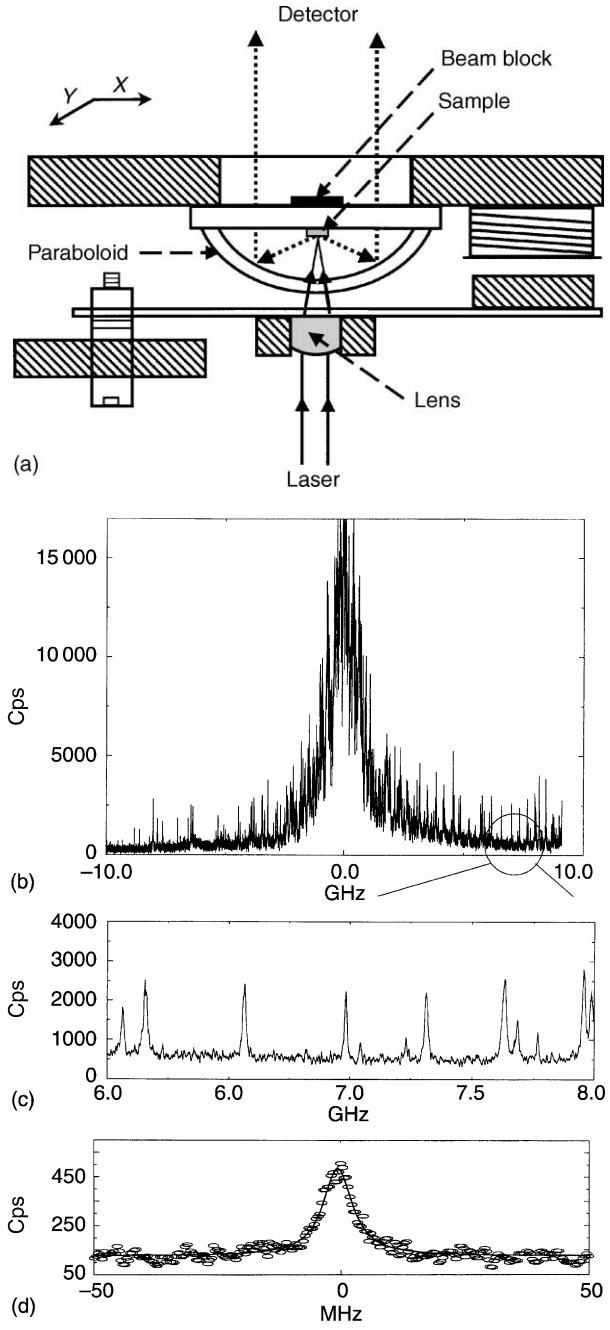


Figure 1 Fluorescence excitation spectra of pentacene in *p*-terphenyl at 1.5 K. (a) Schematic of the cryogenic setup used. A laser beam is focused onto the sample through a lens into the cryogenic chamber. The sample can be moved in the *XY* plane, with a magnet and coil pair. A beam block rejects the incident laser light (black arrow). A paraboloidal mirror collects the emitted fluorescence (dashed arrows) and redirects it towards the detector (photomultiplier tube). The laser frequency (linewidth 3 MHz) is tuned around the absorption line center at 592.321 nm. (b) Broad scan showing the inhomogeneously broadened line shape of pentacene molecules embedded in a *p*-terphenyl crystal at 1.5 K. The sharp lines correspond to spectra of individual molecules (see panel c) (cps: counts per second). (c) Detail of the previous spectrum showing spectra of several individual molecules. The variability of intensities is due to the different positions of the molecule in the excitation beam. (d) Individual molecule spectrum obtained at lower excitation power exhibiting a Lorentzian shape (solid line) centered at 592.407 nm with a lifetime-limited width of 7.8 MHz (intensity is measured in counts-per-second for all three spectra). (a): adapted from Ambrose WP, Basché T and Moerner WE (1991) Detection and spectroscopy of single pentacene molecules in a *p*-terphenyl crystal by means of fluorescence excitation. *Journal of Chemical Physics* 95: 7150–7163. (b)–(d): after Moerner WE (1994) Examined nanoenvironments in solids on the scale of a single, isolated impurity molecule. *Science* 265: 46–53. Copyright 1994, The American Association for the Advancement of Science.

Requirements for Single-Molecule Sensitivity

Observing single molecules may be as easy as finding a needle in a haystack if background, signal and noise are not carefully optimized. Simple considerations illustrated in Figure 5a help measure the technical challenge bearing on single-molecule optical detection. Whatever the type of detected signal (fluorescence, Raman scattering, or others), it is characterized by an absorption cross-section σ (probability of photon absorption) and a quantum yield Q (number of emitted photons per absorbed photon). The optical setup transmits an excitation power P at a frequency ν, resulting in a signal rate $s = EQ\sigma P/(Ah\nu)$ expressed in counts per second (or Hz). This rate is proportional to P/A where A is the cross-section of the excitation beam in the plane of the molecule. E is the collection efficiency of the setup. The environment adds a background rate b per unit volume and unit power, and the detector contributes a dark count rate D. Neglecting noise from the readout electronics, a measurement of duration τ will yield the following signal-to-noise ratio (SNR):

$$\mathrm{SNR} = \frac{s\tau}{\sqrt{(s + bVP + D)\tau}} \qquad [1]$$

where V is the excited volume. The denominator is the root mean square of all contributions to the measured count rate, considered shot-noise limited. The SNR obviously increases at larger integration time τ. Another important quantity is the signal-to-background ratio (SBR):

$$\mathrm{SBR} = \frac{EQ\sigma}{bVAh\nu} \qquad [2]$$

These ratios can be increased by improving the collection efficiency E or decreasing the excitation volume V. A larger excitation power or a longer integration time will improve the SNR without affecting the SBR. Typical values for the above

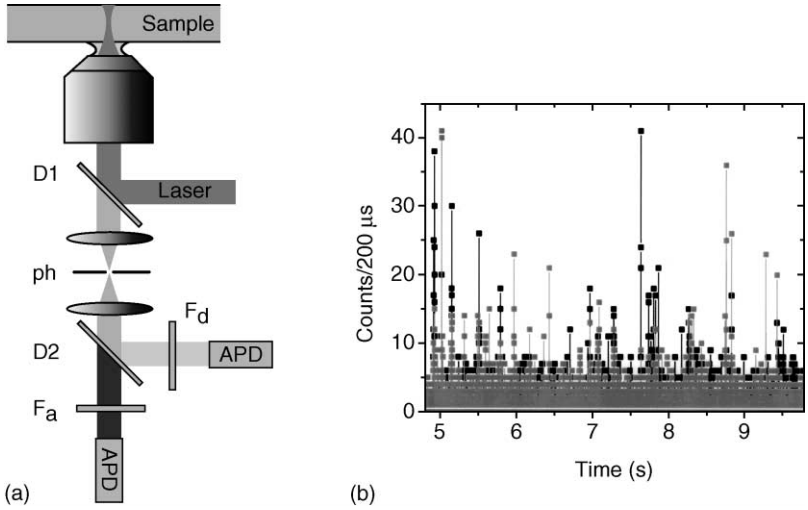


Figure 2 Single-molecule fluorescence burst detection. (a) Schematic of the confocal setup used for single-molecule fluorescence detection. A collimated laser beam is focused through a high numerical aperture (NA) objective in a liquid cell. Molecules diffusing through the diffraction-limited excitation volume (~1 fl) emit photons during their passage, which are collected by the same objective. A dichroic mirror (D1) separates excitation and emission wavelengths. The emitted light is focused on a pinhole (ph) in order to reject out-of-focus background light, and finally recollimated onto two different avalanche photodiodes (APDs) after spectral separation by a dichroic mirror (D2). Filters specific for each channel (F_a: acceptor channel, F_d: donor channel) may be used to improve spectral separation. (b) 5 s time trace of recorded bursts for a 30 pM solution of doubly labeled DNA in 20 mM sodium phosphate buffer. The DNA constructs are used for distance measurement (detailed in **Figure 9**) and have a TMR fluorophore attached to one end of the DNA molecule and a CY5 fluorophore attached to the *n*th base from the end. The time traces correspond to $n = 12$. Gray: TMR fluorescence (donor); black: CY5 fluorescence (acceptor). The laser excitation (0.6 mW, 514 nm) is focused 10 μm within the solution through an oil immersion objective (numerical aperture: 1.4). For each detected photon, the APD generates a pulse, which is recorded by a computer-embedded acquisition board. Counting these events into 200 μs bins, the resulting time trace exhibits distinct bursts distributed randomly in time, with duration and amplitude corresponding to a Brownian diffusion of individual molecules through the excitation volume. (b): is adapted from Deniz AA, Dahan M, Grunwell JR, *et al.* (1999) Single-pair fluorescence resonance energy transfer on freely diffusing molecules: Observation of Förster distance dependence and subpopulations. *Proceedings of the National Academy Sciences USA* 96: 3670–3675. Copyright 1999, The National Academy of Sciences.

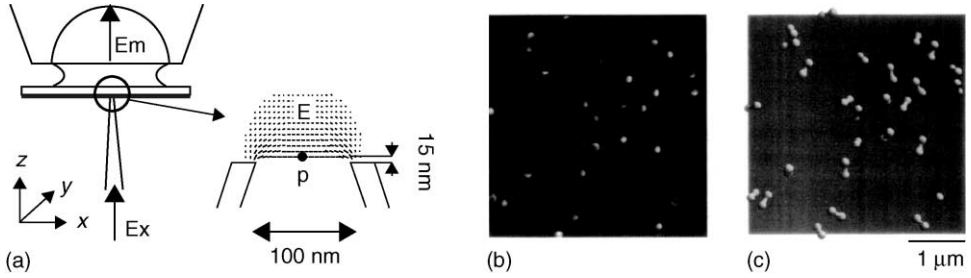


Figure 3 Scanning near-field imaging of single molecules at room temperature. (a) Schematic of a scanning near-field optical microscope setup used for single-molecule imaging. An aluminum-coated tapered optical fiber (raster-scanned at nanometer distance from the sample) serves as a waveguide for laser excitation. Shear-force feedback keeps the tip at a constant distance from the sample, resulting in a signal used for nanometer-resolution topographic reconstruction of the scanned area. The excitation volume and the corresponding local evanescent electric field are detailed in the expanded view. Fluorescence light emitted by individual molecules is collected by an oil immersion, high NA objective and recorded by an avalanche photodiode. (b) $4 \times 4\ \mu m^2$ area of a polymethylmethacrylate (PMMA) coated coverslip on which a diluted methanol solution of a lipophilic carbocyanine dye, $diIC_{12}$, has been deposited. Scanning was performed with randomly polarized light. (c) Orientation of the corresponding individual emitting dipole determined by modeling the field distribution (expanded view in a), simulation of the absorption intensity with polarized excitation and comparison with data obtained at two different excitation polarizations. (a): Adapted and (b), (c) reprinted with permission from Betzig E and Chichester RJ (1993) Single molecules observed by near-field scanning optical microscopy. *Science* 262: 1422–1425. Copyright the American Association for the Advancement of Science.

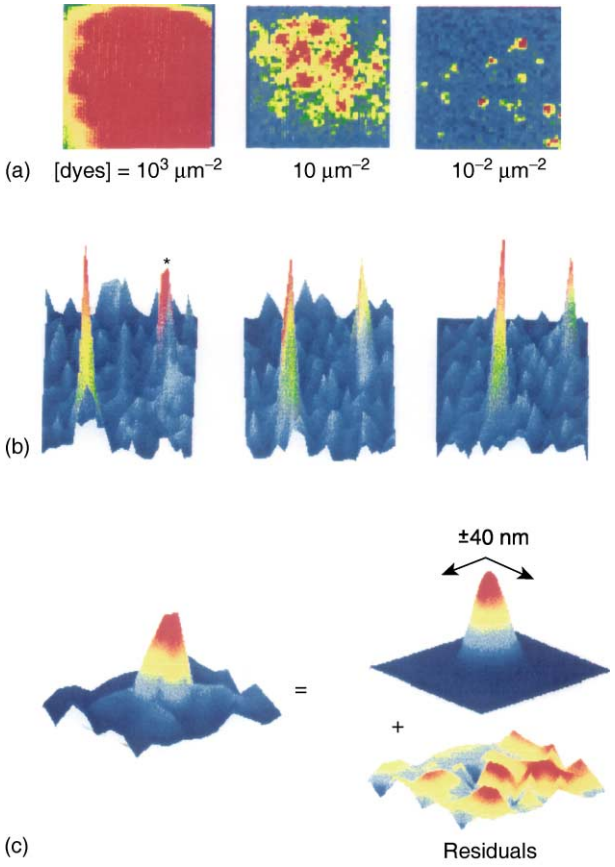


Figure 4 Single molecule diffusion in two dimensions. (a) $6.8 \times 6.8\ \mu m^2$ images of fluorescently labeled lipids (tetramethylrhodamineinthiocarbamoyl, linked to 1,2-dihexadecanoyl-*sn*-glycero-3-phospho-ethanolamine, DHPE) in a lipid double layer membrane (palmitoyloleoylphosphocholine, POPC) at different concentration. From left to right: 10^3 dyes/μm^2, 10 dyes/μm^2, 10^{-2} dyes/μm^2. Each image was obtained with an epifluorescence microscope, illuminated by a defocused 514 nm Ar laser line ($\sim$60 kW/cm^2), and 5 ms integration on a nitrogen cooled CCD camera. Color scale: blue = 0, $d = 60$ counts). Image intensity at 10^3 dyes/μm^2 was divided by 60. (b) Sequence of three $5.4 \times 5.4\ \mu m^2$ images of two labeled lipids observed at 105 ms intervals (actual acquisition rate was one frame every 35 ms, integration time: 5 ms). (c) Subregion of a $2.4 \times 2.4\ \mu m^2$ membrane area showing the peak marked by a star in B. The nonlinear least-square fit of a two-dimensional Gaussian profile to the peak yields values of the fluorescence intensity $I = 168 \pm 25$ counts, width $r = 480 \pm 80$ nm and two-dimensional position with 40 nm resolution. Adapted from Schmidt T, Schutz GJ, Baumgartner W, *et al.* (1996) Imaging of single molecule diffusion. *Proceedings of the National Academy of Sciences USA* 93: 2926–2929. Copyright 1996, The Academy of Sciences.

parameters in the case of a confocal microscopy study of freely diffusing fluorescein isothiocyanate (FITC) in water, using avalanche photodiode detectors (APD) are: $E = 10^{-2}$, $Q = 0.85$, $\sigma = 2.8 \times 10^{-16}\,cm^2$, $A = 0.1\,\mu m^2$, $\lambda = c/\nu = 525\,nm$, $bV = 10$ Hz/μW, $D = 100$ Hz. With an excitation power of 100 μW and a 1 ms integration time, a count rate of 629 counts/ms can be obtained, of which 1 comes from the background sources, leading to a SNR $\sim$25 and SBR $\sim$629. Although these are acceptable figures, they have to be balanced by the fact that single molecules have a finite lifespan. Single molecules in an oxygen-rich environment typically emit on the order of 10^6 photons before irreversible degradation (photobleaching). With the above parameters, this amounts to a total emission of $\sim$3.2 ms. This is enough to observe freely diffusing molecules during their transit time of a few hundred microseconds. For a fixed molecule, however, it sets a stringent limit on the total duration of a single-molecule observation, and all efforts are put on reducing the excitation volume and intensity, as well as background sources. Similar considerations hold on the detector side, where a trade-off has to be found between time-resolution and quantum efficiency.

In addition to SNR and SBR issues, care has to be taken to ensure that the collected signal originates from a single molecule. Two different strategies can be envisioned:

(i) work at low concentration, such that at most one molecule is present in the excitation volume (**Figure 5b**), or equivalently, minimize the excitation volume;
(ii) use a selective excitation or emission detection protocol, such that only one molecule is excited or detected within the sampled volume (**Figure 5c**).

The first case has already been illustrated in **Figure 2**, in which a solution of fluorophores at low concentration yields separate bursts corresponding most of the time to transits of individual molecules. The second strategy can be adopted for the example illustrated in **Figure 1**, in which the ZPL of individual molecules lying in the wings of the ensemble spectrum are well separated. Tuning the laser to the peak absorption frequency of a molecule will excite this molecule only, allowing for a separate study of this molecule.

In summary, a set of six criteria can be defined for SMS:

1. The observed density of emitter is compatible with the known concentration of individual molecules and scales with the original concentration.
2. The observed intensity level is consistent with that of a single emitting molecule.
3. Each immobilized emitter has a well-defined absorption or emission dipole (for organic dyes usually linear, but for other emitters such as a fluorescent semiconductor nanocrystal, possibly circular).

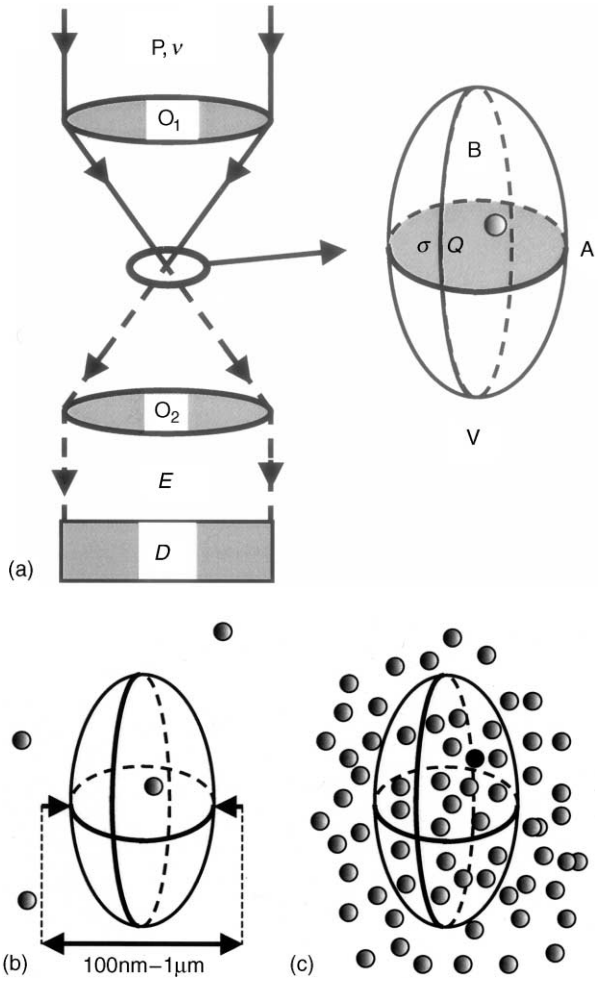


Figure 5 Signal, background and noise issues in single-molecule detection. (a) In a single-molecule spectroscopic experiment, excitation light (power P, frequency ν) is focused on the sample through an optical element O_1 (objective lens, fiber tip), exciting a diffraction-limited volume of section A, represented by an ellipsoid in the expanded view. The absorption process is characterized by an absorption cross-section σ and the emission by a quantum yield Q. The background contribution per watt of incident power is B, whereas the detector will have a dark count rate D. The signal is collected by an optical element O_2 (possibly confounded with O_1), and further guided to the detector, with an overall efficiency E. (b) The signal of a single molecule can be separated from that of other molecules if the probability of finding two or more molecules in the probed volume is negligible. This can be obtained either by a low concentration, or by reducing the excitation volume. (c) If each molecule in the excited volume can be characterized by a different spectral property (which may or may not depend on its exact environment), it is possible to use very selective excitation or detection techniques to probe just one molecule in the sampled volume (the black sphere in the ellipsoid).

4. Fluorescence emission exhibits only two levels (on/off) exemplified in blinking or photobleaching.
5. If there are more emission levels due to spectral jumps (see the section on fluorescence), blinking and spectral jumps are correlated.
6. Antibunching of the emitted photons permits us to test whether the detected signal comes from a single quantum emitter or more (see the section on biological studies below).

Types of Spectroscopy

Fluorophores

A variety of fluorescence emitters (or fluorophores) have been investigated with single-molecule spectroscopy techniques. Fluorescent organic molecules, green fluorescent proteins (GFP) and other biological fluorescent proteins, conjugated polymers (J-aggregates), light-harvesting complexes (e.g., B-phycoerythrin) comprising several fluorophores acting effectively as a single quantum emitter, or semiconductor nanocrystals, are but a few examples of systems which have been extensively studied. Each of these systems exhibits fluorescence based on specific processes, which cannot be described in detail here. A simple picture valid for single fluorescent dyes will be presented instead. It captures, however, the essential ingredients needed to describe the different parameters accessible with single-molecule spectroscopy.

Fluorescence Emission

Most single-molecule experiments use one-photon fluorescence excitation because of its relatively large cross-section ($\sim 10^{-16}$ cm^2). It also allows for a simple rejection of the excitation light: the emitted photons have a lower energy because of losses in intramolecular relaxation processes (the so-called Stokes shift). Excitation and emission processes are conveniently represented by a Jablonski diagram (Figure 6a) depicting the initial, final, and intermediate electronic and vibrational states of the molecule. Upon excitation to the first excited state, the molecule has a certain probability to be shelved in a nonemitting triplet state for up to milliseconds, resulting in 'dark' or off-states (Figure 6b,c). This results in a saturation of the emitted count rate R (Figure 6d) according to:

$$R = R_s \frac{I/I_s}{1 + I/I_s}, \qquad [3]$$

where the saturation intensity $I_s \sim 10-100$ kW/cm^2 and $R_s \sim 1-20$ MHz for typical fluorophores. Incidentally, this puts a limit in SNR enhancement due to saturation of the emitted signal (triplet bottleneck). Other fluorescence emitting 'single molecules' like semiconductor nanocrystals

exhibit similar dark state intervals, although for different reasons (Auger ionization or surface trapping of carriers). Finally, spectral jumps are also observed at room temperature (Figure 6e). This phenomenon results in a shift of absorption and emission maxima due to the changing environment of the molecule, or a sudden conformational change. The corresponding changes in excitation and emission spectra are noticed by sudden fluctuations in the emitted intensity

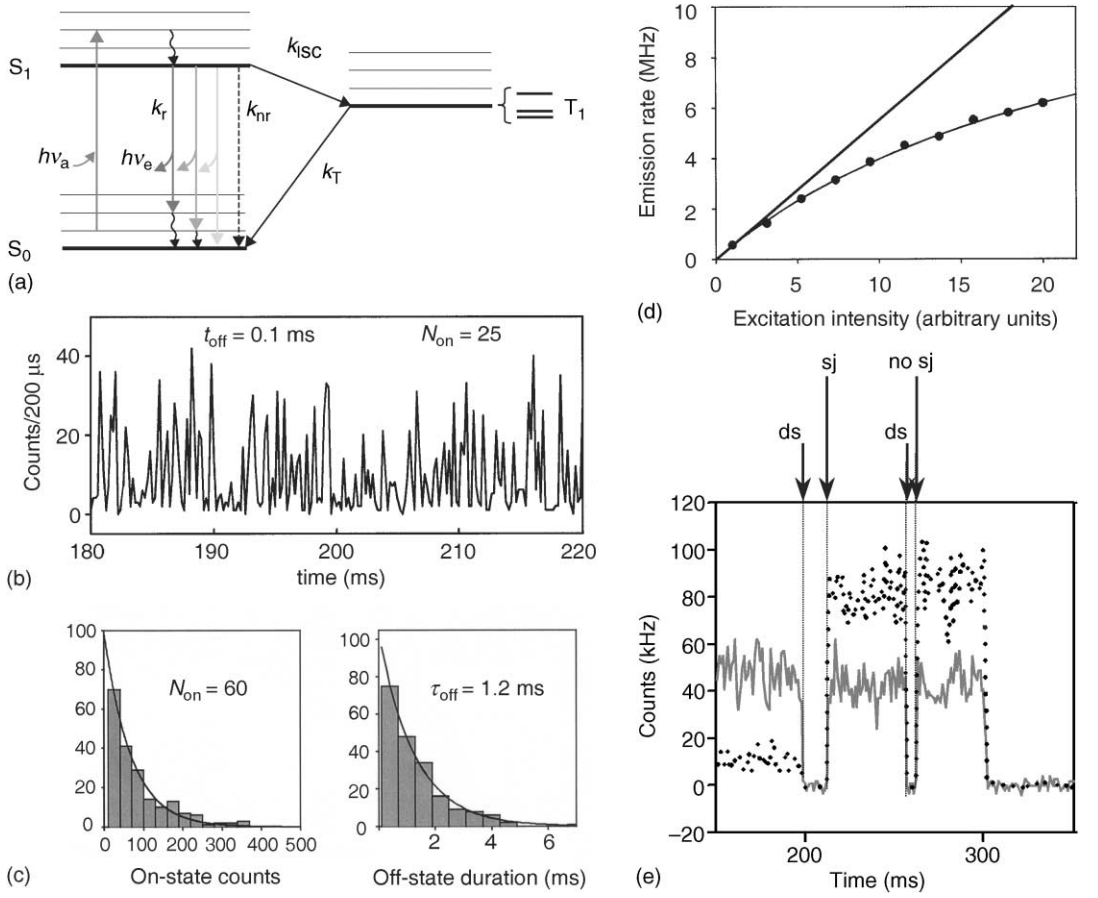


Figure 6 Single-molecule fluorescence. (a) Jablonski diagram for fluorescence (at room temperature). Upon absorption of a photon of energy $h\nu_a$ close to the resonance energy $E_{S_1} - E_{S_0}$, a molecule in a vibronic sublevel of the ground singlet state S_0 is promoted to a vibronic sublevel of the lowest excited singlet state S_1. Nonradiative, fast relaxation brings the molecule down to the lowest sublevel in picoseconds. Emission of a photon of energy $h\nu_e < h\nu_a$ (radiative rate k_r) can take place within nanoseconds and bring the molecule back to one of the vibronic sublevels of the ground state. Alternatively, collisional quenching may bring the molecule back to its ground state without photon emission (nonradiative rate k_{nr}). A third type of process present in organic dye molecules is intersystem crossing to the first excited triplet state T_1 (rate k_{ISC}). Relaxation from this excited state back to the ground state is spin-forbidden and thus the lifetime of this state is in the order of microseconds to milliseconds. Relaxation to the ground state takes place either by photon emission (phosphorescence) or nonradiative relaxation. The fluorescence lifetime is defined by $\tau = 1/\Gamma = (k_r + k_{nr})^{-1}$. (b) Fluorescence time trace (40 ms excerpt) of a Texas Red (TR) molecule attached to a 20-nucleotides-long single-stranded DNA molecule immobilized on a silanized glass coverslip. The confocal setup is similar to that depicted in Figure 2. Excitation: 514 nm, circularly polarized, 6 kW cm^{-2}, binning: 200 µs. On- and off-states (the latter corresponding to the molecule being in the triplet state) alternate stochastically. (c) Histograms of on-state count (left) and off-state duration (right) for the time trace partly shown in (b). The decay parameters of the exponential fits (solid line) are indicated. $t_{off} = 1.2$ ms is the triplet state lifetime. $N_{on}^{-1} = k_{ISC} = 1.7\%$. (d) Intensity dependence of the emission rate of a single Texas Red molecule. The data departs from a linear dependence because of saturation due to intersystem crossing. The fitted curve yields a saturation rate of $R_S = 6.5$ MHz and a triplet lifetime $t_{off} = (k_{ISC}\ R_S)^{-1} = 10$ µs. The difference between this value and the value obtained in (c) illustrates the heterogeneity uncovered by single-molecule measurements. (e) Spectral jumps of a single Texas Red molecule attached to a single-stranded DNA molecule. ds: dark state, sj: spectral jump, no sj: dark state not followed by a spectral jump. Cases where spectral jumps are observed without noticeable dark states may still correspond to dark states with durations shorter than the time resolution of the experiment. (b)–(d) Adapted from Ha T, Enderle T, Chemla DS, *et al.* (1997) Quantum jumps of single molecules at room temperature. *Chemical Physics Letters* 271: 1–5. Copyright 1997, Elsevier Science B. V. (e) Adapted from Ha T, Enderle T, Chemla DS and Weiss S (1996) Dual-molecule spectroscopy: molecular rules for the study of biological macromolecules. *IEEE Journal of Selected Topics in Quantum Electronics* 2: 1115–1127. Copyright 1996 IEEE.

(due to wavelength dependence of the absorption cross-section).

Fluorescence Lifetime

A fluorescence photon is emitted within nanoseconds after absorption of the excitation photon. The decay law is usually mono-exponential, characterized by a lifetime $\tau = \Gamma^{-1}$ (Figure 6a). However, the lifetime depends on the emission spectrum (which can change due to spectral shifts) and the local environment. For instance, the proximity to a dielectric or metallic surface markedly changes it via perturbations of the intramolecular transition matrix elements. Thus, fluorescence lifetime is a sensitive probe of the environment of the molecule (Figure 7). Lifetime measurements require a pulsed laser source and time-correlated single-photon detection and analysis techniques.

Fluorescence Polarization

The efficiency of photon absorption by a fluorescent molecule is proportional to $(\boldsymbol{E}\cdot\boldsymbol{\mu})^2$ where $\boldsymbol{E}$ represents the local electric field and $\boldsymbol{\mu}$ the molecule's absorption dipole moment. For a fixed fluorophore, modulation of the excitation polarization thus allows the determination of the absorption dipole orientation (the emission dipole being generally nearly parallel to it). This information in turn allows the determination of the spatial orientation of the molecule. For a mobile molecule, however, more information is needed, since the emission dipole has time to tumble significantly during the few nanoseconds between absorption and emission. The emission polarization is needed to fully recover the relevant information, as illustrated in Figure 8. Different experimental schemes can be used, besides the near-field optical one described in Figure 3 and the confocal one described in Figure 8. In particular, it is important to recover the projection of the polarization on more than two orthogonal axes. Approaches taking advantage of aberrations induced by index mismatch in high numerical aperture (NA) objectives, or using TIR excitation with alternating polarizations, or using amplitude or phase masks, allow a complete determination of this orientation.

Fluorescence Resonance Energy Transfer

A special case of the local influence of the environment on fluorescence is encountered in fluorescence resonance energy transfer (FRET, see Figure 9a), first theoretically explained by Förster in the late 1940s. In case of the presence of a nearby molecule (acceptor) having an absorption spectrum overlapping that of the donor emission, there is a possibility that the energy absorbed by the donor is transferred nonradiatively to the acceptor. The efficiency E of this transfer is given by:

$$E = \left(1 + \left(\frac{r}{R_0}\right)^6\right)^{-1} \qquad [4]$$

where r is the distance between the two emitting centers and R_0 is the Förster radius (in Å):

$$R_0 = a(\kappa^2 n^{-4} Q_D J(\lambda))^{1/6} \qquad [5]$$

In this complex expression, $a = 9.78 \times 10^3$, n is the medium refractive index, Q_D the donor fluorescence quantum yield in the absence of the acceptor and $J(\lambda)$ is a spectral overlap integral (in M^{-1} cm^3). κ^2 is a geometric factor, which averages out to 2/3 in the case of freely rotating dyes. A typical order of magnitude for R_0 is a few nanometers, which gives the range of the energy transfer. FRET is thus very sensitive to the donor–acceptor distance, which makes it a useful molecular ruler. Experimentally, the efficiency is measured as a function of the recorded donor (F_D) and acceptor (F_A) emissions (dual-channel measurement) via:

$$E = \left(1 + \alpha \frac{F_D}{F_A} \frac{Q_A}{Q_D}\right)^{-1} \qquad [6]$$

where Q_A and Q_D are the fluorescence quantum efficiencies of the donor and acceptor, respectively, and α is the ratio of the acceptor and donor channel detection efficiencies. An illustration of this utilization is given in Figure 9b–d, where different predefined distances between donors and acceptors are set by rigid double-stranded DNA molecules. The corresponding donor–acceptor distances are readily measured on diffusing molecules, with subnanometer precision.

Alternatively, the transfer efficiency can be measured via the fluorescence lifetimes as:

$$E = 1 - \frac{\tau_{D(A)}}{\tau_{D(0)}} \qquad [7]$$

where $\tau_{D(A)}$ and $\tau_{D(0)}$ are the donor lifetimes in the presence or absence of an acceptor, respectively.

Other Spectroscopies

Other types of spectroscopic techniques can be used at the single-molecule level. Fluorescence can for instance be excited via a two-photon absorption process, whereby a laser excitation with half the photon energy needed to attain the excited state

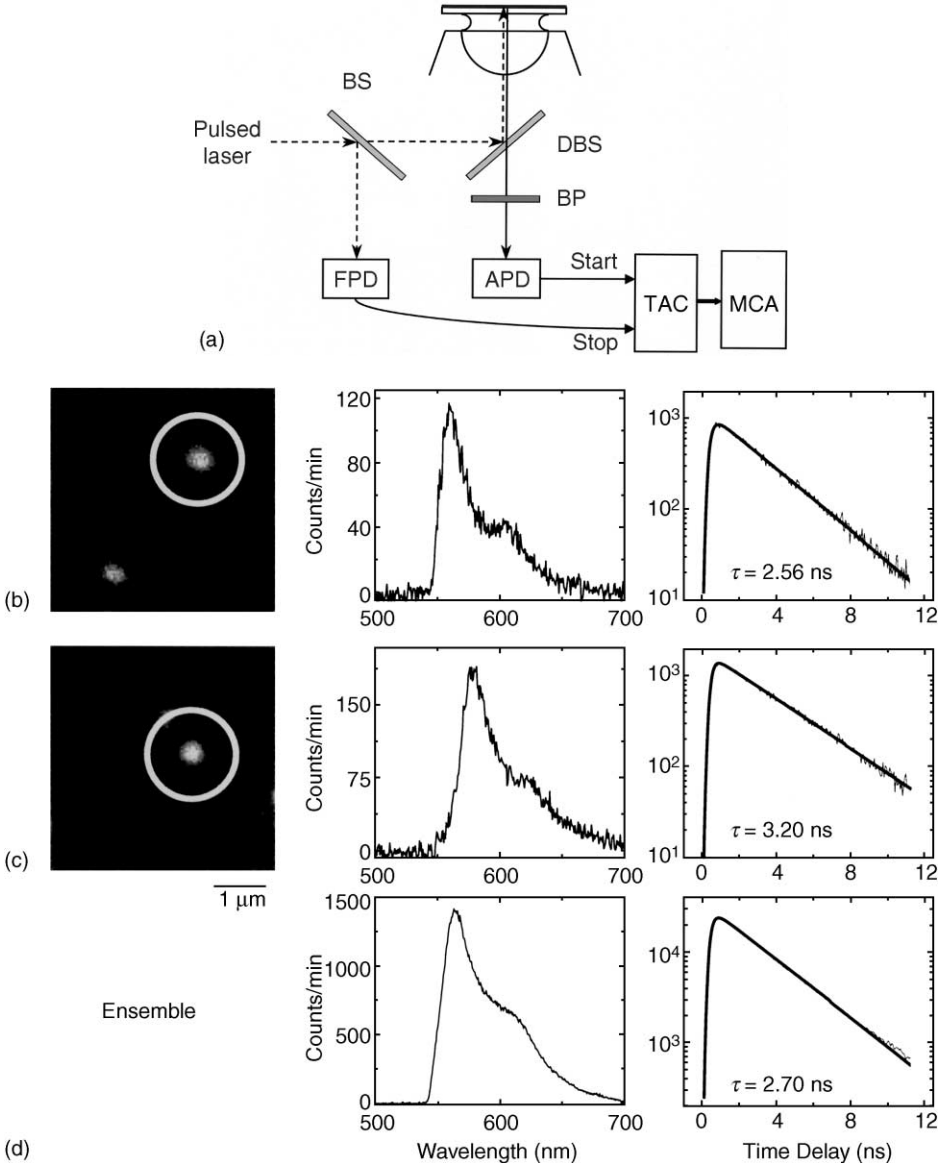


Figure 7 Lifetime measurements. (a) Schematic of a lifetime measurement setup. BS: beamsplitter used to deflect part of the incident pulse towards a fast photodiode (FPD); DBS: dichroic beam splitter used to spectrally separate incident from emitted light; BP: bandpass filter further rejecting stray light; APD: avalanche photodiode. For each detected photon, the APD emits a pulse, which is fed to the start input of a time-to-amplitude converter (TAC). A stop pulse from the laser immediately follows it (regularly emitted at a rate of τ_L^{-1}), allowing the measurement of the separation between both pulses. A multichannel analyzer (MCA) histograms the arrival times, resulting in the curves shown in the right part of Figure 2b–d. (b) and (c) Fluorescence images (left panels), corresponding spectra (middle), and fluorescence decays (right) for two molecules located at a PMMA–air interface. The peak emission wavelength λ_p was 560 nm in (b) and 578 nm in (c). Lifetimes were fitted to a single exponential (dotted curves), with decay times of $\tau = 2.56$ ns (b) and 3.20 ns (c). A ensemble measurement of spectrum and fluorescence decay averaged over several hundred of molecules is shown in (d). The ensemble values are $\lambda_p = 565$ nm and $\tau = 2.70$ ns. (b)–(d): Adapted from Macklin JJ, Trautman JK, Harris TD and Brus LE (1996) Imaging and time-resolved spectroscopy of single molecules at an interface. *Science* 272: 255–258. Copyright 1996, The American Association for the Advancement of Science.

is used. However, due to the very low cross-section and quadratic dependence in the incident energy (two simultaneous photons are needed for excitation), an excitation power several orders of magnitude larger than that for one-photon excitation is needed. In compensation, excitation takes place in a much more reduced volume of the sample, reducing the out-of-focus background contribution. The higher power required increases the probability of photobleaching due to the increased probability of

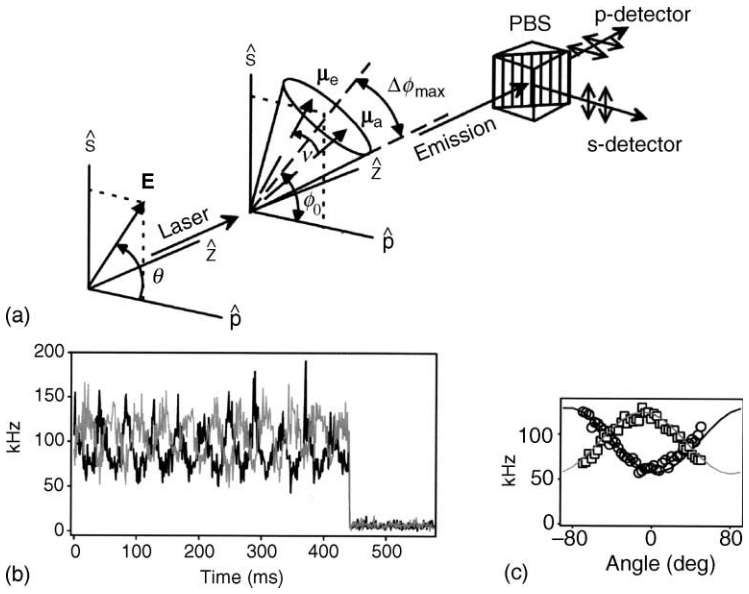


Figure 8 Polarization Measurements. (a) Experiment schematic. **E** is the electric field, making an angle θ with the *p* polarization axis. The excitation propagates along the *z* axis, which is also the collection axis. $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_e$ are the absorption and emission dipole moments, initially aligned. ν represents the rotational diffusion of the emission dipole during the excited lifetime. The dipole is supposed to be confined in a cone positioned at an angle ϕ_0 projected on the (*s*, *p*) plane, and having a half-angle $\Delta\phi_{max}$. A polarizing beam-splitter (PBS) splits the collected emission into two signals I_s and I_p, which are simultaneously recorded by APDs. (b) Simultaneously recorded I_s (black) and I_p (gray) of a molecule rapidly rotating in a liquid. (c) Same data as in (b), but averaged over 11 on-periods. The fit corresponds to a $\Delta\phi_{max}$ close to 90° (freely rotating molecule) and permits the determination of the constrained rotational diffusion parameters. Adapted from Ha T, Glass J, Enderle T, *et al.* (1998) Hindered rotational diffusion and rotational jumps of single molecules. *Physical Review Letters* 80: 2093–2096. Copyright 1998, The American Physical Society.

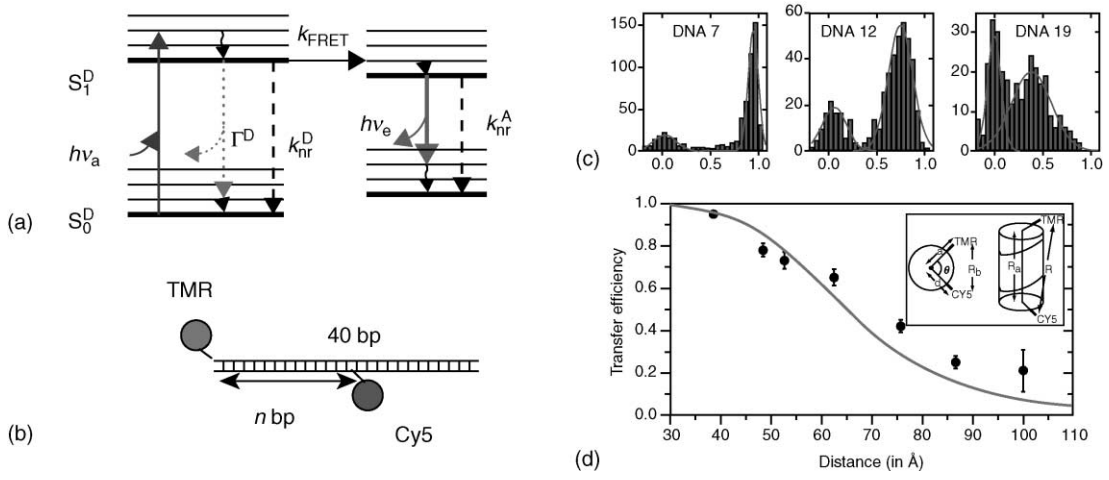


Figure 9 Single pair fluorescence resonance energy transfer. (a) Jablonski diagram for FRET. Fluorescence energy transfer involves two molecules: a donor D and an acceptor A whose absorption spectrum overlaps the emission spectrum of the donor. Excitation of the acceptor to the lowest singlet excited state is a process identical to that previously described for single-molecule fluorescence (**Figure 6a**). In the presence (within a few nm) of a nearby acceptor molecule, donor fluorescence emission is largely prevented by energy transfer to the acceptor by dipole–dipole interaction. The donor exhibits fluorescent emission following the rules outlined in **Figure 6a**. (b) DNA *n* constructs used for the FRET distance study. Tetramethylrhodamine (TMR) is attached to one end of the DNA and CY5 is attached to the *n*th base from the end ($n = 7$, 12, 14, 16, 19, 24 and 27). (c) FRET histograms extracted from time traces for DNAs 7, 12, and 19, as shown in **Figure 2b**. Double Gaussian fits extract numbers for the mean (width) of the higher efficiency peak of 0.95 ± 0.05, 0.75 ± 0.13, and 0.38 ± 0.21, respectively. The peak around 0 efficiency corresponds to bleached or inactive acceptor molecules. (d) Mean FRET efficiencies extracted from FRET histograms plotted as a function of distance for DNA 7, 12, 14, 16, 19, 24, and 27. Distances are calculated using the known B-DNA double helix structure. The solid line corresponds to the expected Förster transfer curve with $R_0 = 65$ Å. (b)–(c) Adapted from Deniz AA, Dahan M, Grunwell JR, *et al.* (1999) Single-pair fluorescence resonance energy transfer on freely diffusing, molecules: Observation of Förster distance dependence and subpopulations. *Proceedings of the National Academy of Sciences USA* 96: 3670–3675. Copyright 1999, The National Academy of Sciences.

photochemical degradation in the long-lived triplet state and the interplay of multiphoton ionization processes.

Although Raman scattering cross-sections ($\sim 10^{-30}$ cm^2) are orders of magnitude lower than the fluorescence cross-section, they can be enormously amplified if the single molecule is trapped in a 'hot spot' of a metallic nanoparticle, giving rise to surface-enhanced Raman scattering (SERS). Highly resolved vibrational information can be recovered, exhibiting much more stable emission than fluorescence. Moreover, this technique can be applied to nonfluorescent molecules. Similarly to fluorescence, the technique can be extended to nonlinear (two-photon) regimes, as in surface-enhanced hyper-Raman scattering and surface-enhanced anti-Stokes Raman scattering. However, since they require the single molecule to be located on rare sites of a metallic particle, these methods have only been applied to fundamental chemical studies of molecular species.

Finally, extending the range of the FRET 'molecular ruler', electron transfer allows measuring distances between an electron donor and an acceptor within a few ångströms (Å), via opening of a nonradiative channel for the relaxation of the electron in the excited state. Lifetime measurements ($\tau = 1/\Gamma = (k_r + k_{nr})^{-1}$ as defined in Figure 6) yield distance information via the electron transfer rate E_{ET}:

$$E_{ET} = k_0 e^{-\beta r}, \qquad [8]$$

where β^{-1} is of the order of one Å and k_0 of the order of 10^{15} s^{-1} are constants characterizing the donor–acceptor pair.

Applications of Single-Molecule Spectroscopy

Single-molecule spectroscopy has already proven to be a useful tool in various domains. The following examples are but a few of an expanding number of applications of SMS.

Molecular Physics

The realization that the fluorescence of a single molecule could be detected with a good SNR over extended periods of time, especially embedded in a crystalline matrix at cryogenic temperature, has led to high-resolution experiments aimed at determining molecular energy levels and transitions. In particular, studies of transitions between sublevels of the triplet state by electron spin resonance (ESR) using superimposed radio-frequency pulses have permitted a precise test of the quantum description of such systems. The dc and ac Stark effect (the shift of a transition energy in the presence of a fixed or alternating electric field) has also been studied in several systems, allowing the symmetry and deformations of the molecule to be determined. The method has been extended to study the photophysics of luminescent polymers (J-aggregates), light-harvesting complexes, green fluorescent proteins (GFP) or semiconductor nanocrystals, and other fundamental systems.

Material Science Studies

Single-molecule photophysical properties are extremely sensitive to their local (nanometer-sized) environment (Figure 1). A precise knowledge of these properties allows using single molecules as sensitive probes of their environment, as illustrated by experiments designed to study the local dynamics of a polymer matrix at the onset of the glass transition (Figure 10).

Molecules of the organic fluorophore Rhodamine 6G (R6G) dispersed within a thin (250 nm) poly(methacrylate) film were observed at temperatures slightly above the melting temperature of the polymer, using a confocal microscope detecting the s and p polarization of the emitted fluorescence (Figure 8a). Observations are performed under nitrogen atmosphere and at low excitation power (20 nW) in order to prevent oxidation, thus allowing long (several hours) observation time. Each observed R6G molecule exhibits a slow rotational diffusion, sometimes accompanied by lateral or transverse diffusion on a much longer time-scale (Figure 10a). Computation of the autocorrelation function of the emission polarization shows a multi-exponential decay, pointing to a distribution of time-scales of rotational diffusion. A closer look at subwindows of the time trace reveals, however, that the decay comes closer to mono-exponential behavior at shorter time-scales (Figure 10b,c). This study demonstrates that an individual molecule probes an increasing number of different environments over time. At long time-scales, the observable characteristics (such as the autocorrelation function) are similar to those measured on an ensemble of molecules, as expected from the ergodicity hypothesis. At short time-scale, the molecule reveals the local homogeneous and stable characteristic of its nanoenvironment.

Biophysical and Biological Studies

SMS has allowed a reassessment of long-standing questions in biophysics, biochemistry, and biology, by giving scientists the possibility to look at individual building blocks of biological processes: DNA, RNA, proteins and other biomolecules or biomolecular

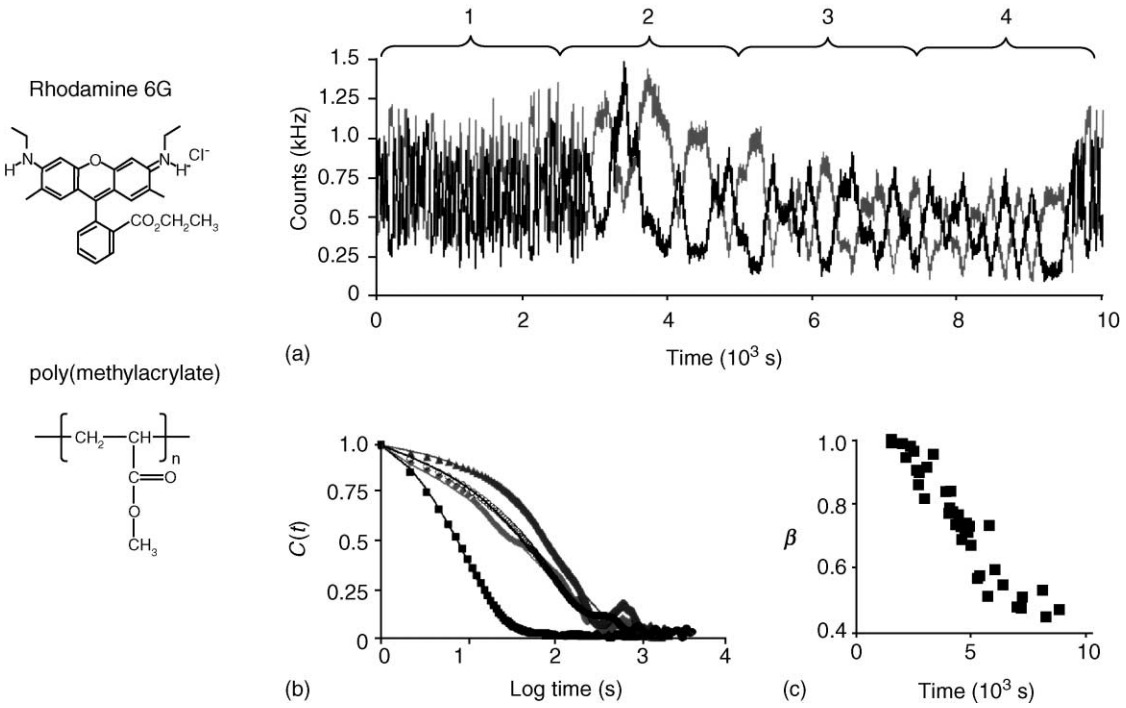


Figure 10 Polymer dynamics near the glass transition. (a) Fluorescence intensity of a single long-lived Rhodamine 6G molecule in poly(methylacrylate) (PMA) at 291 K detected at orthogonal polarizations (gray: *s* and black: *p* polarizations). The melting temperature of PMA is $T_g = 281$ K. The time trace integration is 1 s. This time trace shows the anticorrelation of both polarizations, characteristic of rotational diffusion of the molecule. Different regimes are observed along the time trace (1–4), which could be due to either a modification of the local 'cage' in which the molecule rotates, or escape of the molecule to another nearby nano-environment. The autocorrelation function of polarization computed over the whole trace (not shown) is well fitted by a model adapted to the nonexponential relaxation of polymers, the Kohlrausch–Williams–Watt (KWW) model:

$$\mathrm{KWW}(t) \propto \mathrm{e}^{-(t/\tau)^{\beta}}.$$

with $\tau = 88$ s, $\beta = 0.46$. (b) Correlation functions of subsets 1–4 of the full trace. 1 (squares): $\tau = 15$ s, $\beta = 1.0$. 2 (triangles): $\tau = 112$ s, $\beta = 0.77$. 3 (open circles): $\tau = 90$ s, $\beta = 0.59$. 4 (diamonds): $\tau = 64$ s, $\beta = 0.59$. The β exponent is closer to 1, showing that the rotational diffusion is an unhindered one at short time-scales, converging to the bulk exponent at larger time-scale due to ergodicity. This experiment illustrates changes of environment as the observed molecule diffuses and rotates through the polymer mesh, changes that are hidden in ensemble measurements. (c) Values of β as a function of photochemical survival time in PMA at $T_g + 5$ K. Adapted from Deschenes L and Vanden Bout DA (2001) Single-molecule studies of heterogeneous dynamics in polymer melts near the glass transition. *Science* 292: 255–258. Copyright 2001, The American Association for the Advancement of Science.

complexes formed by association of a few of these molecules.

To simplify the task, some biological molecules are intrinsically fluorescent in their active state, as the cholesterol oxidase (COx) enzyme of Brevibacterium sp., an enzyme that catalyzes the oxidation of cholesterol by oxygen. As a result, it is relatively easy to study the dynamics of this enzyme at the single-molecule level using a confocal setup with enzymes immobilized in an agarose gel (Figure 11). The intervals during which the enzyme is active ('on' state) are followed by intervals during which it is inactive ('off' states). The fluorescence recording of a single enzyme thus appears as a 'blinking' time trace, as shown in Figure 11b. Relevant kinetic parameters can be extracted from the distribution of 'on' and 'off' times. Apart from the usual heterogeneity between enzymes of the same batch characterized by different kinetic parameters for the catalysis reaction, this experiment reveals a dynamic heterogeneity, which shows up for each individually studied enzyme as a tendency for short 'on' times to follow short 'on' times, and for long 'on' times to follow long 'on' times (Figure 11c). This variable behavior of a single enzyme is reminiscent of the variable rotational dynamics of Rhodamine 6G molecules in a polymer matrix (see the previous section), and can be studied only thanks to single-molecule techniques.

Not all biological molecules can be studied in such a convenient way using their intrinsic fluorescence. Several labeling schemes are currently available, which allow quantitative, site-specific labeling of proteins, nucleic acid or lipid molecules with few biochemical side effects.

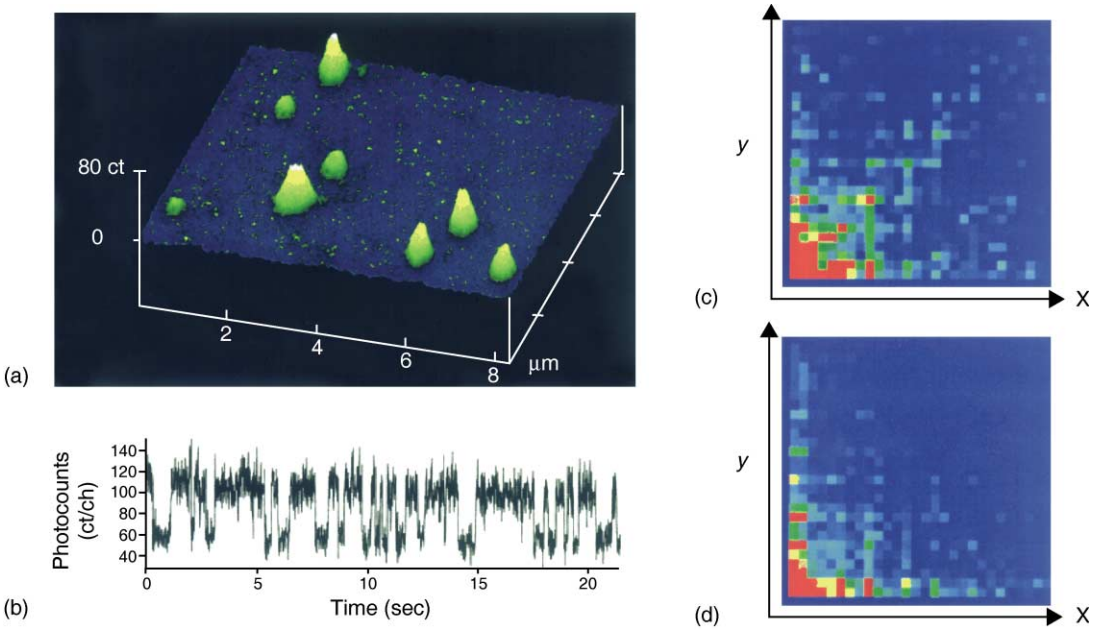


Figure 11 Single-enzyme dynamics. (a) Fluorescence image of single CO_x molecules immobilized in a 10-μm-thick film of agarose gel of 99% pH 7.4 buffer solution (8 μm × 8 μm, 4 min scan, 500 nW excitation at 442 nm). The emission is from the fluorescent active site, FAD, which is tightly bound to the center of CO_x. Each individual peak is attributed to a single CO_x molecule. The intensity variation between the molecules is due to different vertical positions in the excitation volume. (b) Real-time observation of enzymatic turnovers of a single CO_x molecule catalyzing oxidation of cholesterol molecules. This panel shows a portion of an emission intensity trajectory recorded in 13.1 ms per channel. The trajectory was recorded with a cholesterol concentration of 0.2 mM and saturated oxygen concentration of 0.25 mM. The emission exhibits stochastic blinking behavior as the flavin adenine dinucleotide (FAD), the CO_x fluorescent active site toggles between oxidized (fluorescent) and reduced (nonfluorescent) states, each on–off cycle corresponding to an enzymatic turnover. More than 500 on–off cycles are recorded in this trajectory. (c) The 2D conditional probability distribution for on-times (x and y) of two adjacent turnovers as derived from the trajectories of 33 CO_x molecules with 2 mM 5-pregene-3β-20α diol substrate, a derivative of cholesterol. The scales of the x- and y-axes are from 0 to 1 s. A subtle diagonal feature is present, indicative of a memory effect. (d) The 2D conditional histogram for two on-times separated by 10 turnovers for the CO_x molecules in (c). The diagonal feature vanishes because the two on-times become independent of each other at the 10-turnover separation. The color code in (c) and (d) represents the occurrence (z-axis from 350 (red) to 0 (purple). Adapted from Lu HP, Xun L and Xie XS (1998) Single-molecule enzymatic dynamics. *Science* 282: 1877–1882. Copyright 1998, The American Association for the Advancement of Science.

An example of this versatility has been provided by the labeling of the central part of a rotary protein, F_1-ATPase. In this experiment (not shown), the molecular rotor was labeled with a single fluorophore whose orientation, detected by emission polarization measurements, directly reported the angle of the rotor with respect to the shaft. The small size of the fluorophore guaranteed that the protein motion would not be hindered, contrary to previous experiments, which used much larger reporters (micron-size latex beads or fluorescent actin filaments). This experiment reproduced the previous results, showing that the rotor performed 120° steps (recent work hints at substeps of 90° and 30°).

One of the most promising domains where SMS can be of great importance is the study of protein folding. Doubly labeled proteins can be monitored as they undergo conformational changes, taking advantage of the distance dependence of fluorescence energy transfer (demonstrated for a model DNA system in the section on FRET, Figure 9).

Figure 12 shows one of many examples of such a study, performed on diffusing molecules in solution, a configuration that suppresses uncontrolled effects of nearby surfaces. The enzyme chymotrypsin inhibitor 2 (CI2) is a model system for protein folding, believed to have two clearly distinct folding states, which can be controlled by the concentration of denaturant guanidinium chloride. Using single-pair FRET (spFRET) techniques, it is possible to sort out the proportion of folded and unfolded molecules present at different denaturant concentration (Figure 12b). These measurements yield the same titration curve as obtained by ensemble measurements (Figure 12b) and can give access, under some simplifying assumptions, to the energy landscape of the folding reaction. The additional information provided by spFRET is

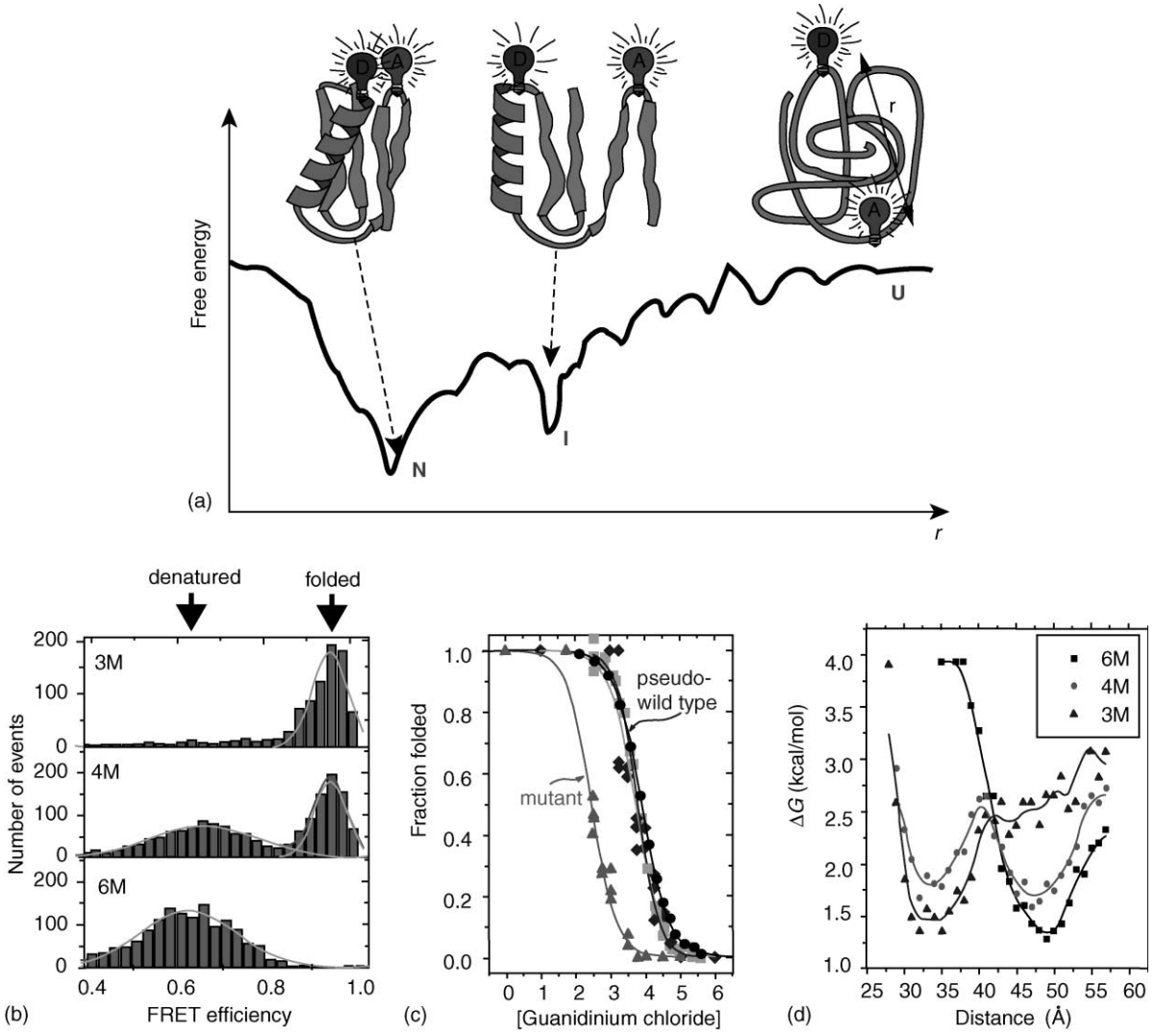


Figure 12 Protein folding. (a) Energy landscape for a protein folding reaction represented as a cartoon of the free energy as a function of distance *r* between donor and acceptor. The rugged energy landscape has a funnel shape with three extrema: U (unfolded state), I (intermediate state) and N (folded or native state). In this labeling scheme, donor and acceptor are close to one another in the native, folded state (N), resulting in a high FRET efficiency. There is reduced energy transfer (larger distance) in the denatured, unfolded state (U). (b) Single-molecule protein folding data. Single-pair FRET (spFRET) histograms of the pseudo-wild type enzyme chymotrypsin inhibitor 2 (PWT CI2) at 3, 4, and 6M denaturant (guanidinium chloride). The FRET efficiency exhibits two peaks, one for folded molecules (high FRET) and one for unfolded molecules (low FRET). Population is transferred from folded to unfolded at higher denaturant concentration, as expected from the protein labeling schematically depicted in (a). (c) Ensemble and single molecule denaturation curves for PWT and K17G mutant CI2. Symbols correspond to data; lines showing sigmoidal fits. PWT CI2: ensemble tryptophan fluorescence experiment (circles), ensemble FRET (squares), spFRET (diamonds); K17G mutant CI2 spFRET (triangles). (d) Free energy functions for PWT CI2 at 3, 4, and 6M denaturant. The solid lines represent a smoothing of the data and are only meant to guide the eye. (a) Adapted from Weiss S (2000) Measuring conformational dynamics of biomolecules by single-molecule fluorescence spectroscopy. *Nature Structural Biology* 7: 724–729. Copyright 2000, Nature Structural Biology. (b)–(d) Adapted from Deniz AA, Laurence TA, Beligere GS, *et al.* (2000) Single-molecule protein folding: diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *Proceedings of the National Academy of Sciences USA* 97: 5179–5184. Copyright 2000, The National Academy of Sciences.

the number of molecules in the two different states, providing direct evidence of the two-states model hypothesized from ensemble measurements.

Other biological systems have been addressed using the spFRET approach. Intramolecular spFRET (used in the CI2 study) allows monitoring conformational changes, whereas by labeling two different molecules, intermolecular spFRET permits the detection of association and dissociation events. For instance, spFRET has revealed transient intermediate states in

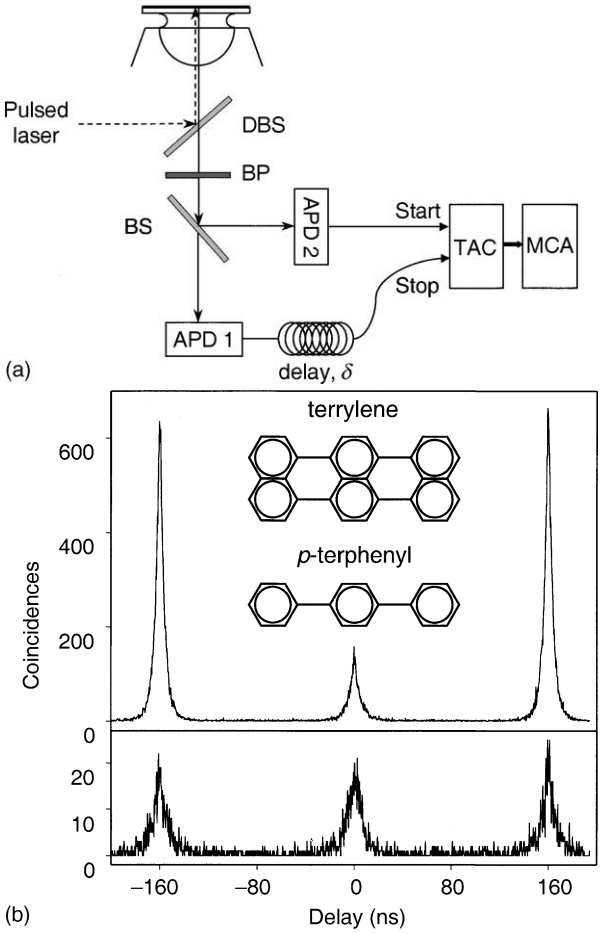


Figure 13 Single-photon source at room temperature. (a) Antibunching measurement setup. The basic elements are similar to those of a lifetime measurement setup (**Figure 7a**) except that the emitted filtered signal (BP: bandpass filter) is split in two with a 50/50 beamsplitter (BS). Two avalanche photodiodes (APD) detect the arrival of photons. One APD pulse is used as the start signal for a time-to-amplitude converter (TAC), which is stopped by the pulse coming from the second APD. Since the TAC response time is finite, it cannot detect simultaneously arriving pulses. A delay time δ is thus added to the stop signal, such that two photons impinging simultaneously on the two APDs will generate pulses with a separation of δ. If the start APD detects a photon from the previous excitation, the measured delay will be 160 ns $+ \delta$, whereas if the opposite happens, the delay will be $\delta - 160$ ns. (b) Results of measurements on a terrylene molecule (upper chemical structure) embedded in a p-terphenyl (lower chemical structure) at room temperature. Upper panel: histogram of delay measurements shifted by an amount δ. The histogram corresponds to 120 s of recording on an isolated emitting spot. The residual central peak corresponding to coincidence of photon arrival at APD 1 and 2 is the result of background contamination. Lower panel: histogram acquired over 300 s on a background spot. In both measurements, the laser pulse width was 35 ps, each pulse being separated by 160 ns. The lifetime of terrylene is measured to be 3.8 ns, which sets the width of the observed peaks. Excitation is performed at 532 nm. (b): Adapted with permission from Lounis B and Moerner WE (2000) Single photons on demand from a single molecule at room temperature. *Nature* 407: 491–493. Copyright 2000, Nature Publishing Group.

the Tetrahymena ribozyme, which had remained unnoticed in ensemble studies.

Quantum Optics

Low-temperature spectroscopic studies of single molecules have turned into a benchmark for testing predictions of quantum mechanics. In particular, one effect expected for a single quantum emitter is the absence of simultaneous emission of two photons: a fluorescent system which is excited at time t_0 has to emit a photon before it can return to the ground state, and be re-excited at a later time t_1. In conditions where the excitation takes place immediately as the electron returns to the ground state, the time difference $t_1 - t_0$ is of the order of the excited state lifetime, of the order of a few nanoseconds. This antibunching phenomenon was first observed in atoms and ions, and rapidly confirmed by low-temperature SMS. This observation is possible either using continuous wave or using pulsed excitation. In fact, nothing prevents this effect from occurring at room temperature, as demonstrated in **Figure 13** for pulsed excitation in which excitation photons are provided as well-separated bunches of a few hundred femtoseconds duration.

In this experiment, a system originally used at low temperature (terrylene in a crystal of p-terphenyl) is excited with a repetition rate of 6.25 MHz. The fluorescence emitted by a single molecule is detected in a confocal setup, split in two and directed to two different detectors (**Figure 13a**). In this way, the time interval between each pair of photons detected by the two detectors can be precisely measured and histogrammed. The result of this time-correlated single-photon counting approach is the absence of coincidence of two photons (**Figure 13b**), besides the residual counts due to background contamination.

Since the quantum efficiency of this system is high, the experimental setup can be seen as a periodic single-photon source, with an extremely low probability for the simultaneous emission of two photons. This is precisely the kind of device that is needed for quantum cryptography, in which information encoded in the polarization of a beam should not be sensitive to a 'beamsplitter' attack. Future efforts will undoubtedly build on such 'turnstile' photon sources to move forward into optical quantum computing applications.

Perspectives

SMS is a mature discipline, which will certainly prove to be most useful in domains beyond those previously mentioned. Its relative ease of implementation will

certainly attract more scientists of other disciplines to this methodology, each of them having specific goals in mind that are difficult to foresee. The principal advantages of SMS outlined at the beginning of this article will remain, but improvements, especially towards simultaneous acquisition of all the fluorescence parameters described earlier are promising. New detectors will permit the combination of the high time resolution of single-photon counting devices with the large field of view and spectral resolution allowed by two-dimensional detectors. SMS will be associated with single molecule manipulation techniques, to correlate applied forces and molecular conformations. New ways of controlling local fields (electric, magnetic, or others) would take advantage of the non-invasiveness, high-temporal and spatial resolution of SMS to get a direct feedback of events at the nanometer scale. Undoubtedly, SMS will be a major tool in the future of nanotechnology.

List of Units and Nomenclature

Å	Ångström
cm^2	square centimeter
fl	femtoliter
K	kelvin
ms	millisecond
MHz	megahertz
μW	microwatt
nm	nanometer

APD	Avalanche photodiode
DNA	Deoxyribonucleic acid
FRET	Fluorescence resonance energy transfer
NA	Numerical aperture
PMMA	Polymethylmethacrylate
SERS	Surface-enhanced Raman scattering
SMS	Single-molecule spectroscopy
SNR	Signal-to-noise ratio
SNE	Signal-to-background ratio
TIR	Total internal reflection
ZPL	Zero phonon line

See also

Scattering: Raman Scattering.

Further Reading

Basché T, Moerner WE, Orrit M and Wild UP (1997) Single-Molecule Optical Detection, Imaging and Spectroscopy. Weinheim: VCH.

Craig DP and Thirunamachandran T (1998) Molecular Quantum Electrondynamics. An Introduction to Radiation Molecule Interactions. Mineola: Dovers.

Forkey JN, Quinlan ME and Goldman YE (2000) Protein structural dynamics by single-molecule fluorescence polarization. *Progress in Biophysics and Molecular Biology* 74: 1–35.

Kneipp K, Katrun H, Kneipp I, *et al*. (1999) Surface-enhanced non-linear Raman scattering at the single-molecule level. *Chemical Physics* 247: 155–162.

Marcus RA and Sutin N (1985) Electron transfers in chemistry and biology. *Biochimica et Biophysica Acta* 811: 265–322.

Mehta AD, Rief M, Spudich JA, Smith DA and Simmons RM (1999) Single-molecule biomechanics with optical method. *Science* 283: 1689–1695.

Moerner WE and Orrit M (1999) Illuminating single molecules in condensed matter. *Science* 283: 1670–1676.

Nie S and Zare RN (1997) Optical detection of single molecules. *Annual Review of Biophysics and Biomolecular Structure* 26: 567–596.

Plakhotnik T, Donley EA and Wild UP (1997) Single-molecule spectroscopy. *Annual Review of Physical Chemistry* 48: 181–212.

Tamarat P, Maali A, Lounis B and Orrit M (2000) Ten years of single-molecule spectroscopy. *Journal of Physical Chemistry A* 104: 1–16.

Weiss S (1999) Fluorescence spectroscopy of single biomolecules. *Science* 283: 1676–1683.

Xie XS and Trautman JK (1988) Optical studies of single molecules at room temperature. *Annual Review of Physical Chemistry* 49: 441–480.

T

TERAHERTZ TECHNOLOGY

Contents

Coherent Terahertz Sources
Terahertz Physics of Semiconductor Heterostructures

Coherent Terahertz Sources

L Wang, Chinese Academy of Sciences, Beijing, China

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Terahertz (THz) radiation is usually referred to as an electromagnetic wave with frequencies ranging from 1 to a few terahertz ($1\,\text{THz} = 10^{12}\,\text{Hz}$). In the electromagnetic spectrum, terahertz radiation lies between the infrared and microwave, as shown in Figure 1. The techniques for generation and detection of THz radiation bridge photonics and electronics in the sense of the concepts and the techniques discussed below.

Physical quantities corresponding to 1 THz are listed as follows;

- Frequency, 1 THz
- Wavelength, 300 μm
- Wavenumber, 33 cm^{-1}
- Energy, 4.1 meV
- Temperature, 48 K

In this spectral region, there exist many rich physical, chemical, and biological phenomena. For example, many phonon resonance and other elementary excitations in condensed matter fall in this region; many molecular vibrations and rotations occur at THz frequency; conformation-related collective vibrational modes in macromolecules, especially in bio-molecules such as proteins and DNA, also lie in THz frequency. However, the development of THz technique lags far behind that of photonics and electronics, due to lack of feasible, reliable, and economic coherent THz sources. The situation has been rapidly changing since the 1990s. Benefiting from advanced material science and the ultrafast laser techniques, various coherent THz sources are being developed and commercialization is soon to follow. We will introduce these coherent THz sources and their main features, and discuss their generation mechanisms. More details can be found in the listed books and articles in the Further Reading section at the end of this article.

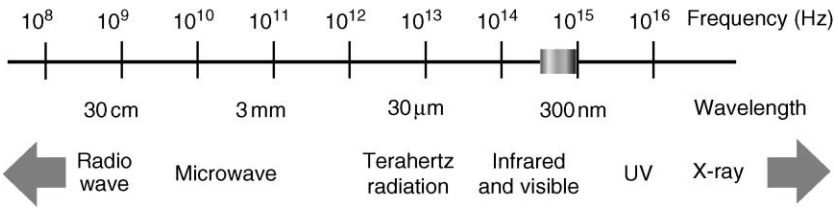


Figure 1 The electromagnetic spectrum with THz radiation sitting between the microwave and the infrared.

Coherent THz Radiation

When a time varying polarization, $\mathbf{P}(\mathbf{r},t)$, is generated in a medium and oscillates at THz frequency, it will radiate a THz wave, $\mathbf{E}(\mathbf{r},t)$, according to Maxwell's equation:

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{r},t) + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{E}(\mathbf{r},t) = -\mu_0 \frac{\partial^2}{\partial t^2}\mathbf{P}(\mathbf{r},t) \quad [1]$$

or

$$\nabla \times \nabla \times \mathbf{E}(\omega) - \frac{\omega^2}{c^2}\mathbf{E}(\omega) = \omega^2 \mu_0 \mathbf{P}(\omega) \quad [2]$$

in frequency domain, for a monochromatic wave or a Fourier component of a wave with a finite bandwidth. The properties of THz radiation are determined by the polarization source $\mathbf{P}(\mathbf{r},t)$. Incoherent THz sources exist in blackbody radiation, and are also detected in astronomy observations. If the phase evolution of the THz wave keeps in step, both spatially and temporally, it is called coherent radiation. A coherent THz source is more powerful and useful in spectroscopy, material characterization, imaging, and many other applications. It is clear that a coherent polarization source is first required for generation of a coherent THz wave. In practice, the working media for THz emitting can be free electrons, quasi-free electrons in solids, or bound electrons and other charge oscillations, such as plasmon oscillation, lattice vibration, etc. Classified by the excitation properties, the emitting process can be of resonance or nonresonance. Coherent polarization sources with different features can be created using various excitation techniques under different principles and device configurations, which lead to different coherent THz sources. Here we focus on coherent THz sources, primarily based on photonic techniques, and working in the spectral range from 1 THz to a few THz.

Gas THz Laser

Molecules have many rotational and vibrational energy levels spaced at THz frequency. For polar molecules, the radiative transition between these levels can radiate THz wave. Following the standard laser technology, molecules pumped either optically or electronically can be used as laser gain media for coherent THz wave generation. For THz gas lasers, the very low THz photon energy poses a problem for lasing, because the coherent radiative transition can be disrupted by thermal depopulation and dephasing. An effective population inversion is difficult to set up and maintain. Therefore, it is important to carefully choose working medium and energy levels, as well as the way to excite the molecules. A number of media has been used for THz laser, such as methanol and HCN vapor, working in both continuous wave (CW) and pulsed modes. For example, methanol lasers operate under excitation of the rocking and asymmetric deformation modes of methanol excited by a CO_2 infrared laser. HCN laser is pumped electrically, which needs a large voltage and current to excite the molecules, a long cavity to obtain sufficient gain, and dedicated control of the temperature in the cavity for stable operation.

In general, THz gas lasers can be tuned discretely in thousands of lasing lines, ranging from a few tens to a few hundreds of micrometers, with power output from μW to mW. THz gas lasers have existed since the early 1970s, and are the only commercialized THz lasers to date. They are still subject to development, for less bulky volume, enhanced tunability and efficiency, and reduced costs.

Free Electron THz Laser

A free electron laser (FEL) uses free electrons as working medium, rather than bound atomic or molecular states in a conventional laser. A typical FEL consists of three parts: (i) an electron source that generates an electron beam with high current; (ii) an accelerator to raise the electron energy; and (iii) a lasing cavity. After the electron beam is generated, electrons are first accelerated to a relativistic velocity, typically with energy of hundreds of MeV, and then enter the cavity consisting of end mirrors and a spatially periodic magnets array called a wiggler (**Figure 2**). Due to the Lorentzean force imposed by the periodic magnetic field, the high-energy electrons move in the cavity along a sinusoidal path, and emit coherent radiation with a wavelength determined by the spacing between magnets and the electron velocity, as well as the magnetic induction. With the feedback provided by the cavity mirrors, electrons are accelerated or decelerated continuously by the optical field, and are bunched via the resonant interaction. The collective motion of the electron bunches radiates powerful coherent synchrotron radiation. In a standard configuration, the wavelength of the radiation can be expressed as:

$$\lambda = \frac{\Lambda}{2\gamma^2}[1 + (k\Lambda B_0)^2/2] \quad [3]$$

where Λ is the spatial period of the wiggler magnets, γ the Lorentz factor of the electron beam, B_0 the peak magnetic induction, and k a constant. With a suitable

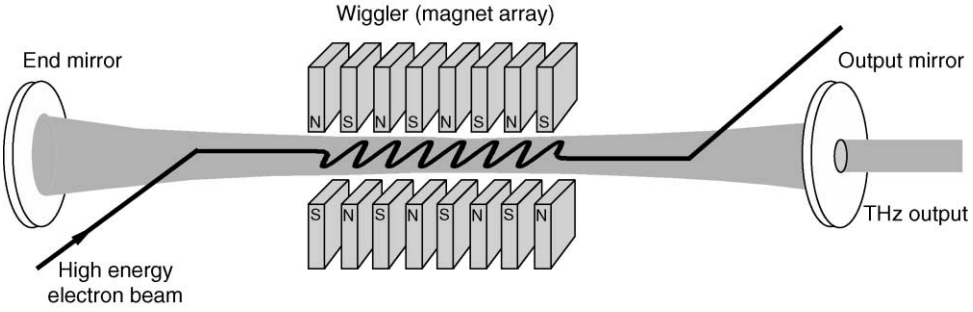


Figure 2 A schematic of THz generation in a free electron laser.

arrangement of the magnet array, as well as other related factors, the facility can work in THz spectral region.

An FEL generates tunable, coherent, high-power THz radiation. It is the most powerful coherent source of THz radiation available. Up to 50 W of average THz power has been generated in the Jefferson Lab recently. The free electron THz laser is becoming an important tools for studying THz radiation and its interactions with condensed matter and biological materials. On the other hand, the FEL is a complicated and expensive facility, and needs dedicated maintenance.

Semiconductor THz Laser

In the twentieth century, semiconductor devices have achieved tremendous success, both in electronic and photonic regime, such as transistors and diode lasers. The semiconductor industry has resulted in revolutionary changes to our world and everyday life. To fill the THz gap between electronics and photonics, great efforts have been made to meet the demand in compact, economic, tunable, and highly efficient THz sources. Encouraging advances have been made since we stepped into the new millennium.

In a semiconductor quantum well, the energy difference between a pair of sublevels can be artificially designed at THz frequency. As early as 1971, the concept of far infrared lasers based on intersub-band transition in semiconductor superlattices was proposed. Although the sublevels can be populated by electrical or optical pumping, the low photon energy at THz frequency causes serious difficulties for lasing, because the two levels sit so close to each other that many thermal processes could destroy the population inversion. Various designs and configurations have been proposed to tackle the problem. An attractive idea is so-called quantum cascade heterostructure. In this kind of device, the electrons are injected into a series of coupled quantum wells electrically. When the electrons are driven by the biased voltage, they can make radiative transition between a pair of sub-bands in a well. Subsequently, electrons enter the next well through resonant tunneling, and so on. However, this kind of device needs the materials of very high quality and suffers severely from fast depopulation of the excited states, even at very low temperatures.

As the advance in material science and unremitting efforts in pursuing the dedicated design of the device structures continued, the real breakthrough came in 2001, when Köhler *et al.* at Pisa demonstrated THz lasing in a quantum cascade heterostructure. Many improvements have been made since then; for example, the idea of using phonon resonance to selectively deplete the population of the lower sub-band has been successful, so that much stabler and robuster population inversion can be set up. The operation temperature has been raised to 136 K, well above the liquid nitrogen temperature, which opens the way for semiconductor THz lasers stepping into many more daily applications. These semiconductor devices are pumped with low voltages and currents at mA levels, and generate narrow bandwidth radiation with mW output at a frequency of a few THz. It is also probable that a four-level lasing system could even working at room temperature.

Another promising device is the p-germanium (Ge) laser developed recently. The Ge laser operates through the electrical excitation of hot holes in p-doped Ge. The laser cavity is formed by polishing the surfaces of the Ge crystal. These lasers could run at 5% duty cycles, with several mW output power. The output wavelength can be continuously tuned from 1 to 4 THz. At the moment, the devices have to operate at 20–30 K and consume about 10–20 W for refrigeration.

It is safe to predict that the feasibility, low cost, and compact semiconductor THz laser will emerge in the near future.

Beside the THz lasers, coherent THz radiation can also be generated by coherently exciting suitable media in other ways. Since the mid-1990s, the

development of ultrafast laser techniques has led to two important pulsed coherent THz sources, i.e., photoconductive antenna and optical rectification THz sources. A unique feature of the pulsed coherent sources is broad bandwidth, which is useful in spectroscopic applications. Figure 3 is a typical setup for pulsed THz generation and spectroscopic study, using photoconductive antenna or optical rectification as THz sources.

THz Photoconductive Antenna (PCA)

Different from a conventional antenna working in radio and microwave frequency that are driven directly by electrical current, a photoconductive antenna for THz wave generation works with ultrafast pulsed lasers. When the photoconductive gap is irradiated by an ultrafast laser pulse, photocarriers are generated in the semiconductor. These photocarriers are driven by a DC voltage bias and form a current transient. The antenna couples the electromagnetic field associated with the time varying current into free space, and produces a THz pulse. The THz pulse usually has a duration of a few picoseconds, with a broad frequency bandwidth centered at THz frequency. The output THz power scales with both the optical pulse power and the DC bias field:

$$E_{\mathrm{THz}} \propto \frac{\partial J}{\partial t} \propto \frac{\partial^2 n_{\mathrm{e}}}{\partial t^2} E_{\mathrm{bias}} \qquad [4]$$

The output power, frequency bandwidth, and polarization of the THz radiation depend on the photoconductive features of the semiconductor layer and the geometric structure of the antenna. For an optimal performance, the substrate should have a suitable bandgap for optical interband excitation, a high electron mobility and short carrier lifetime to support an ultrafast current transient, and high dielectric breakdown threshold and to sustain high bias voltage. High-power PCAs have been demonstrated with semi-insulating GaAs, ion-implanted GaAs, and In-GaAs. In the aspects of antenna structures, coplanar strip lines and large aperture emitters are the most effective. Average output power from a good system can reach 30–40 μW.

A broad bandwidth is a special characteristic of a THz source, which is particularly desirable in the spectroscopic study as a coherent probe source. From the property of Fourier transform, we know that the shorter the THz pulse duration, the broader the spectral band. PCAs made of GaAs typically have a useful bandwidth extending from ~100 GHz to 3 THz, which can be extended to 4 THz by tuning the pumping laser wavelength close to the bandedge. Up to 6 THz bandwidths have been reported for PCAs. The pulse duration and, therefore the achievable bandwidth, are ultimately limited by carrier mobility and TO phonon resonant absorption which is around 8.3 THz in GaAs.

In many applications, THz radiation is coupled to free space from the antenna using a closely attached silicon hemispherical lens. This practice increases the system's output and also provides control of the radiation pattern. The radiation pattern for the common dipole antenna is essentially dipolar, with a weak quadrupole component perpendicular to the bias field. In the far-field, the THz beam has a Gaussian cross-section with high-frequency components concentrated in the center.

As the first practical pulsed coherent THz source, PCA has been widely used for many applications in

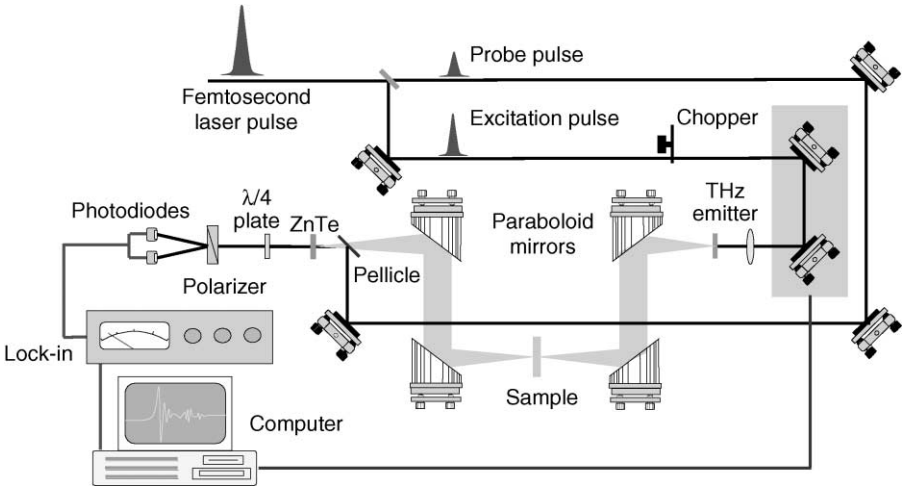


Figure 3 A typical setup for broadband pulsed THz generation and spectroscopic study with a femtosecond laser as excitation source.

scientific research and material characterization. Combined with commercial ultrafast optical fiber lasers, full functional and self-contained THz spectroscopy systems with PCAs as emitters, have been demonstrated and are under commercialization.

Optical Rectification

Optical rectification is a second-order nonlinear optical effect originating from susceptibility $\chi^{(2)}$, and was first observed in 1962 using two intense monochromatic nanosecond laser pulses. In this process, the interaction between laser and nonlinear optical crystal produces a frequency difference polarization:

$$P(0) = \chi^{(2)} E(\omega_1) E(-\omega_1) \qquad [5]$$

When a nonlinear optical crystal is irradiated by an ultrafast laser pulse, the different frequency components in the laser pulse will be coupled with each other by $\chi^{(2)}$, to induce frequency difference polarizations as:

$$P(\Omega) = \chi^{(2)}(\omega_1 - \omega_2 = \Omega) E(\omega_1) E(-\omega_2) \qquad [6]$$

A femtosecond laser pulse usually has a spectral linewidth of at least a few nm, the induced polarization $P(\Omega)$ has a finite frequency distribution centered at THz frequency in the far field, and results in THz radiation $E_{\mathrm{THz}} \propto \partial^2 P(t)/\partial t^2$.

THz radiation in free space by optical rectification was obtained in the early 1990s. One of the differences from a PCA device is that the THz output by optical rectification scales with both the optical pulse power and the DC bias field, while the THz radiation generated by optical rectification solely results from the incident laser. The optimal performance depends on several factors. The generation efficiency first depends on the magnitude of the $\chi^{(2)}$ and phase-matching condition between the optical and THz pulses. The effective magnitude of the $\chi^{(2)}$ coefficient varies with the crystal cut and orientation. Up to now, the most popular material is ZnTe, for its physical durability and excellent phase matching, when a Ti:sapphire femtosecond laser is used as the excitation source. With a moderately focused optical pump beam and a ZnTe crystal, nW T-ray average power can be generated by optical rectification. In general, the generation efficiency may be increased by increasing the laser power, but there are two factors limiting the effect. First, the incident laser power cannot exceed the damage threshold. Second, other second-order nonlinear processes may compete with the optical rectification, such as second-harmonic generation of the pump beam at high optical flux.

For THz radiation generated by optical rectification, the ultimate bandwidth is primarily determined only by the linewidth of the pump laser pulse. Because it is a nonresonance process in most cases, optical rectification can reach broader bandwidths than PCA. Using ultrashort optical pulses and thin nonlinear crystal, THz pulses have been generated with spectra extending to the mid-infrared, well beyond the phonon band of the materials. Like the photoconductive antenna, optical rectification has become a popular technique to generate coherent broadband THz radiation in various applications. Beside the conventional bulk inorganic crystals, many different media have been used for THz generation, such as organic crystal DAST, biased quantum wells, periodically poled LiNbO3 (PPLN), poled polymers, super-conducting thin films, etc.

Other Coherent THz Sources

When two light sources with slightly different wavelengths interact together with a nonlinear optical crystal, a beating polarization will be generated and will radiate a THz wave. It is called difference frequency generation, or three-wave mixing. Optical parametric processing is another way to generate a THz wave. In this process, a near-infrared pump beam generates a second NIR idler beam in a nonlinear crystal, and THz radiation can be generated from the beating of the pump and the idler. The merit of this technique is the continuous frequency tunability of the THz output, which is valuable in spectroscopic applications, and relatively cheaper nanosecond lasers can be used. If the nonlinear crystal is further placed in a cavity and the idler beam is amplified by feedback from end mirrors, a THz optical parametric oscillator is formed. A number of improvements have been made since the mid-1990s with this kind of device. More recently, THz parametric generation with an injection seeded idler beam has shown a reduced linewidth ($\Delta\nu/\nu \approx 10^{-4}$) and a peak THz power of over 100 mW for 3.4 ns pulses. A narrow linewidth combined with reasonable tunability make this kind of THz source attractive in spectroscopic studies.

Besides using photonic techniques, coherent THz radiation can also be generated by electronic techniques through increasing the output frequency of the microwave devices. For example, an electrically driven microwave generator, backward wave oscillator, can generate CW THz radiation up to 2 THz. A backward wave oscillator, running under optimal conditions, can provide up to 300 mW of polarized, narrowband radiation with a limited tunability about 30% of its central frequency.

See also

Imaging: Infrared Imaging. **Lasers:** Free Electron Lasers. **Optical Parametric Devices:** Overview. **Semiconductor Physics:** Infrared Lattice Properties. **Terahertz Technology:** Terahertz Physics of Semiconductor Heterostructures.

Further Reading

Grüner G (1998) *Millimeter and Submillimeter Wave Spectroscopy of Solids*, *Vol. 74 of Topics in Applied Physics*. Berlin: Springer-Verlag.

Kaiser W (1988) *Ultrashort Laser Pulses and Applications*, *Vol. 60 of Topics in Applied Physics*. Berlin: Springer-Verlag.

Khurgin JB (1994) Optical rectification and terahertz emission in semiconductors excited above the bandgap. *Journal of the Optical Society of America B* 11: 2492–2501.

Rullière C (1998) *Femtosecond Laser Pulses*. Berlin: Springer-Verlag.

Saldin EL, Schneidmiller EA and Yurkov MV (2000) *The Physics of Free Electron Lasers*. *Advanced Texts in Physics*. Berlin: Springer-Verlag.

Siegel PH (2002) Terahertz technology. *IEEE Transactions on Microwave Theory and Techniques* 50: 910–928.

Tredicucci A, Gmachl C, Capasso F, *et al.* (2001) Novel quantum cascade devices for long wavelength IR emission. *Optics Materials* 17: 211–217.

Tsen KT (2001) *Ultrafast Physical Processes in Semiconductors*, *Vol. 67 of Semiconductors and Semimetals*. New York: Academic Press.

Terahertz Physics of Semiconductor Heterostructures

R Bratschitsch and K Unterrainer, Technische Universität Wien, Vienna, Austria

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The terahertz (THz) regime of the electromagnetic spectrum holds promise for a large number of important applications in areas like environmental sensing, quality control, medicine, military applications, astronomy, etc. The development of THz devices has been mainly hindered by the lack of compact sources. In this article we overview the types of tabletop semiconductor sources that are available right now. We focus on the use of pulsed THz sources that are optically excited by femtosecond laser pulses. The generated bursts of THz radiation are extremely useful for time-resolved THz spectroscopy of semiconductor heterostructures.

THz Generation from Semiconductors

The first freely propagating THz pulses using femtosecond lasers were generated in 1984 when Auston *et al.* fabricated a Hertzian dipole antenna. It mainly consists of an electrically biased transmission line structure mounted on a semiconductor, i.e. a photoconductive switch that is triggered by a femtosecond laser pulse (Figure 1). The generated transient photocurrent $J(t)$ emits radiation according to $E(t) \propto \partial J(t)/\partial t$. Since 1989, sapphire half-spheres or refractive silicon lenses have been attached to these dipole antennas, which made it possible to produce collimated THz beams for use in everyday spectroscopic experiments. Today generation and detection of THz radiation by semiconductor photoconductive antennas is widely used. However, most of the antennas are limited to the low-frequency regime of the THz spectrum since they usually have a high-frequency cutoff of about 1–2 THz.

THz generation from semiconductor surfaces was introduced by Zhang in 1990. There are three different physical mechanisms that are responsible for the surface emission of THz radiation. One is the transient current, i.e., the acceleration of photogenerated carriers in the semiconductor surface electric field. This process is accompanied by field induced optical rectification which is often called instantaneous polarization, where the electron–hole pairs are polarized in the dc electric field. Finally, bulk optical rectification as a source of THz radiation can be observed in noncentrosymmetric crystals.

The surface emission intensity can be enhanced by using an externally applied electric field instead of the intrinsic electric field. Until recently, the use of laser amplifiers that optically pump these so-called large-aperture antennas has led to the brightest ultrafast tabletop THz emitters.

Recently, Sarukura and co-workers found that the application of a magnetic field dramatically increases the THz emission from bulk InAs. This led to a re-examination of Zhang's work, in which he reported the magnetic field enhancement in bulk GaAs,

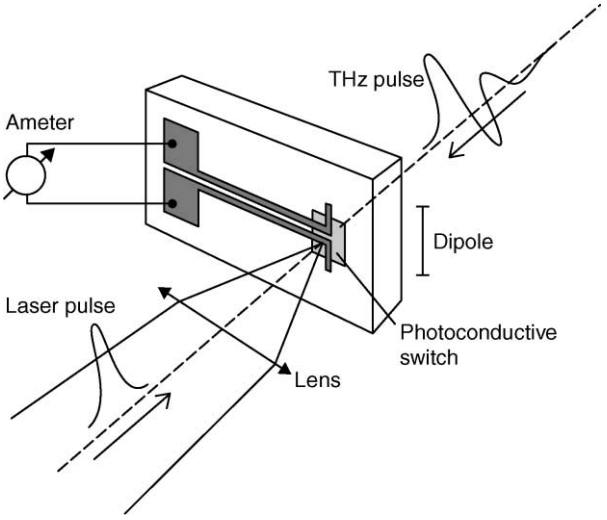


Figure 1 Photoconducting dipole antenna. Reproduced with permission from Gornik E and Kersting R (2001) Coherent THz emission in semiconductors. In: Tsen KT (ed.) *Ultrafast Physical Processes in Semiconductors*, vol. 25, pp. 396. New York: Academic Press.

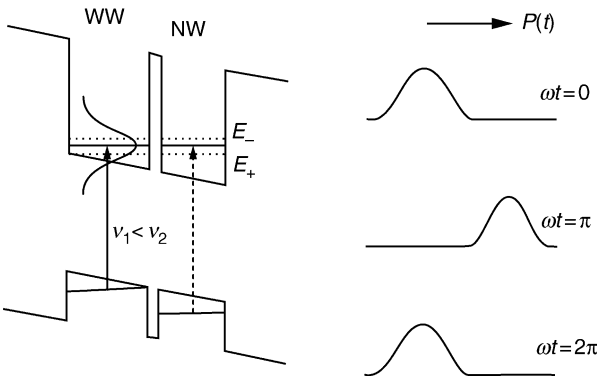


Figure 2 Schematic band diagram of an asymmetric double quantum well structure. A short laser pulse with a spectral width larger than the intersub-band splitting can excite an electronic wavepacket which tunnels back and forth between both wells. Reproduced with permission from Roskos H, Nuss M, Shah J, Leo K, Miller D, Fox A, Schmitt-Rink S and Köhler K (1992) Coherent submillimeter-wave emission from charge oscillations in a double-well potential. *Physical Review Letters* 68: 2216. Copyright (1992) by the American Physical Society.

and to a large number of studies of different types of other semiconductors in various magnetic field geometries. Currently, there is no unified model of the enhancement mechanism(s).

The use of a semiconductor heterostructure by Roskos *et al.* in 1992 (a coupled double quantum well) marked the advent of a new class of semiconductor THz emitters that can be deliberately designed using bandgap engineering. In these devices the quantum nature of electrons in coupled potential wells is used to generate electron wavepackets that tunnel back and forth between different potential wells (Figure 2). The coherent charge oscillation is triggered by a femtosecond laser pulse that has a spectral width that is larger than the separation between the quantized states in the well. In this way a coherent superposition of these levels is created. THz emission from these quantum beats has also been observed from charge oscillations due to light and heavy hole excitons in a quantum well by Planken and co-workers. The concept of quantum beats can even be transferred to the high-THz regime (tens of THz) which was successfully demonstrated by Bonvalet and co-workers. By using an InAs step quantum well they generated THz radiation at a frequency of 30 THz.

Bloch oscillations, one of the fundamental concepts of solid state physics, were predicted in 1934 by Zener. He stated that electrons in a periodic potential under an applied electric field should perform oscillations. Using a quantum mechanical picture, Bloch oscillations can be described as a quantum beat of Wannier–Stark states and in 1993 Waschke *et al.* detected THz emission from these oscillations.

These heterostructure THz sources are inherently frequency tunable since the energy level spacing can be altered by an applied electric field that tilts the quantum wells. The Bloch frequency is given by:

$$\omega_{\text{Bloch}} = \frac{eFd}{\hbar} \qquad [1]$$

where F is the applied electric field and d is the superlattice period. Unfortunately, these devices suffer from their low emission power.

An alternative concept is the use of collective excitations in semiconductor bulk material or heterostructures. Dekorsy and co-workers showed in 1995 that coherent phonons emit THz radiation when excited by femtosecond laser pulses. In the same year Sha and co-workers studied the absorption changes within an electrically biased GaAs p-i-n structure and found distinct oscillations which they ascribed to another collective phenomenon, namely coherent plasma oscillations of the photogenerated carriers (Figure 3). This explanation was supported by the square-root dependence of the observed frequencies on the photogenerated carrier density which is predicted by a simple oscillator model:

$$\omega_{\text{p}} = \sqrt{\frac{ne^2}{\varepsilon\varepsilon_0 m^*}} \qquad [2]$$

where n is the carrier density, ε the dielectric constant, and m^* the effective mass of the carriers. Fischler and co-workers confirmed this finding and in particular investigated the damping behavior of the plasmons. In 1997, Kersting *et al.* detected THz radiation from

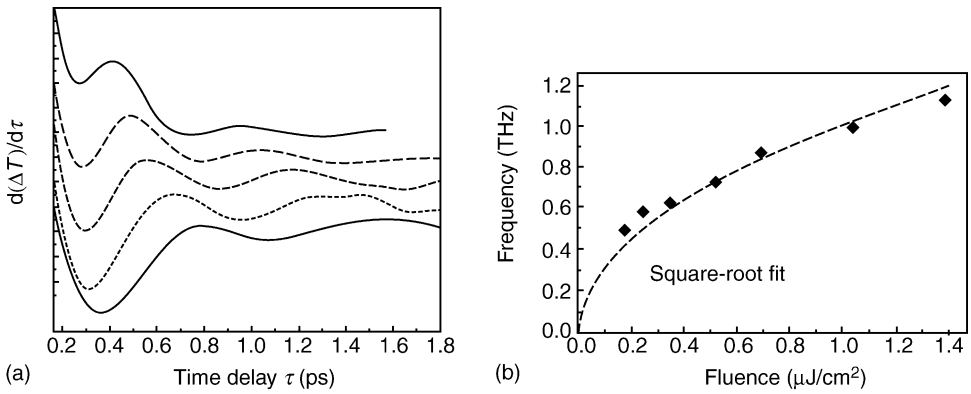


Figure 3 (a) Time-derivative of the charge in electro-absorption versus pump-probe time delay for decreasing fluences (top to bottom respectively); (b) the frequency of the oscillations as a function of fluence. The dashed curve is a square-root function fit. Reproduced with permission from Sha W, Smirl A and Tseng W (1995) Coherent plasma oscillations in bulk semiconductors. *Physical Review Letters* 74: 4273. Copyright (1995) by the American Physical Society.

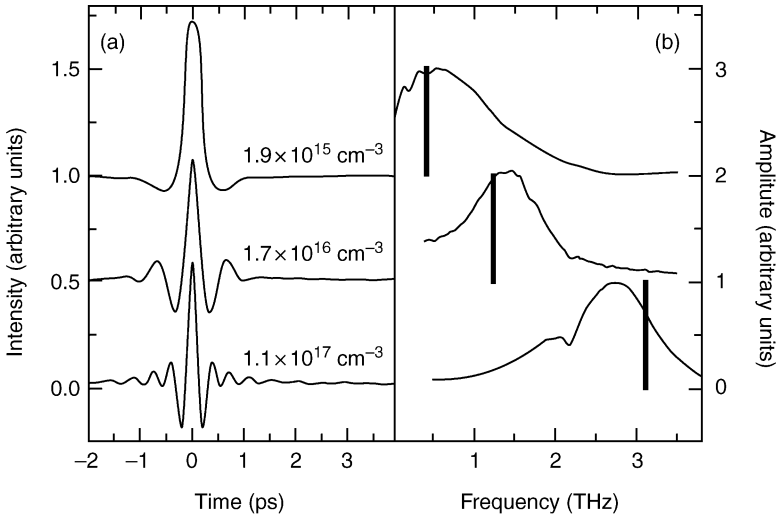


Figure 4 (a) THz autocorrelation traces for three different doping densities; (b) THz spectra; lines denote the frequencies calculated by formula [2]. Reproduced with permission from Kersting R, Heyman J, Strasser G and Unterrainer K (1998) Coherent plasmas in n-doped GaAs. *Physical Review B* 58: 4553. Copyright (1998) by the American Physical Society.

these p-i-n structures and identified heating of the photogenerated carriers as the origin for the strong damping of the plasma oscillations resulting in a low emission intensity.

This problem can be circumvented by using coherent plasma oscillations of extrinsic charge carriers. These electrons are cold compared to the photo-excited carriers and should undergo less scattering. Kersting *et al.* showed that THz pulses are indeed emitted from coherent three-dimensional (3D) plasma oscillations in n-doped GaAs epilayers. The THz pulses are temporally and spatially coherent with intensities up to 100 nW.

Figure 4 shows the THz emission from three GaAs epilayer samples which have been doped with densities 1.9×10^{15}, 1.7×10^{16}, and $1.1 \times 10^{17}\,\mathrm{cm}^{-3}$. The curves are recorded using an autocorrelation technique: 13 nJ, 100 fs laser pulses at a repetition rate of 80 MHz and a central wavelength of 800 nm are split into two pulses with a Michelson interferometer. Every visible pulse generates a THz pulse in the transmission or pseudo-reflection direction. After collimating and focusing of the THz radiation by parabolic mirrors, the interference of the two THz pulses can be recorded by changing the time delay between the visible pulses (Figure 5). Assuming negligible nonlinear interactions between the two excitation processes, the linear autocorrelation function can be recorded. The spectrum is calculated by simply Fourier transforming (Figure 4b). The experiments are performed at room temperature, and the whole setup is purged with dry nitrogen gas to prevent

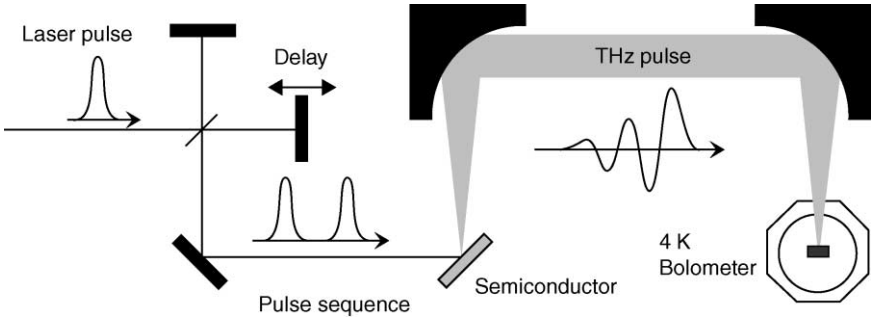


Figure 5 THz autocorrelation setup. Reproduced with permission from Kersting R, Heyman J, Strasser G and Unterrainer K (1998) Coherent plasmons in n-doped GaAs. *Physical Reviews B* 58: 4553. Copyright (1998) by the American Physical Society.

water vapor absorption. Although the autocorrelation traces are recorded at the same optical excitation density of $n_{ex} = 2 \times 10^{16}$ cm^{-3}, the emission frequency changes significantly. In addition, the autocorrelation trace of the highly doped structure (1.1×10^{17} cm^{-3}) shows multiple oscillations while in the case of the sample with the lowest doping (1.9×10^{15} cm^{-3}) the slow oscillation is damped out rather quickly. The frequency behavior of the three different samples can be analyzed from the spectra shown in Figure 4b. The bars indicate the frequency calculated with [2] and setting n equal to the doping density determined by standard Hall measurements. Obviously, the calculated frequencies match very well the experimentally determined ones. This indicates that the THz emission results from a collective oscillation of the extrinsic electrons within the epilayer. This is also supported by the observation that a variation of the excitation density of the photogenerated carriers over two orders of magnitude does not change the THz emission frequency. This is in clear contrast to the experiments with undoped semiconductors where the emission frequency shows a square-root dependence on the optically excited carrier density.

Excitation of the coherent plasma oscillation results from the ultrafast screening of the surface electric field by the femtosecond laser pulse. Before excitation, the doping electrons are confined between the surface electric field and the undoped GaAs substrate (Figure 6). A short laser pulse with an energy above the bandgap of the semiconductor generates electron–hole pairs that screen this surface electric field for a short time. The doping electrons inside the material respond to the change in potential at the surface of the sample leading to the oscillation behavior. A detailed model calculation of the excitation mechanism can be found in Kersting *et al.* (1998). The oscillation frequency – the eigenfrequency of the electron system – is simply the plasma frequency given by the simple formula [2].

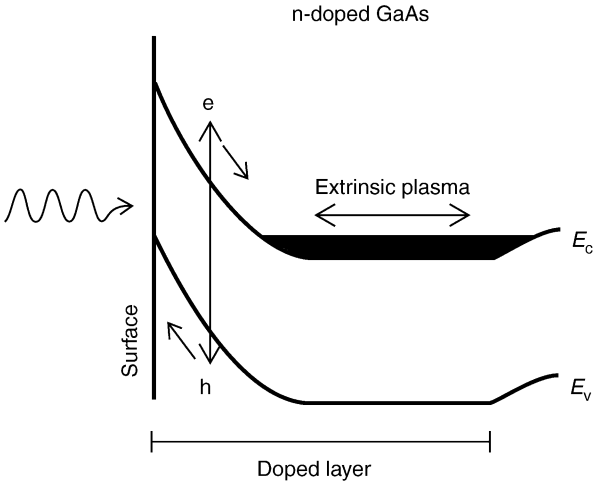


Figure 6 Schematic drawing of the 3D plasmon excitation. After Kersting R, Bratschitsch R, Heyman JN and Unterrainer K (1999) Unpublished.

There are two mechanisms that are responsible for the damping of the plasma oscillations. If the doping density is lower than 10^{16} cm^{-3} the coherence time of the 3D plasmon is mainly limited by optical phonon scattering. This can be verified by measuring the plasmon damping times at different temperatures and comparing the results with mobility data which show the same tendency. Since optical phonon scattering is suppressed at low temperatures, the THz emission intensity can be substantially increased by cooling the sample. In the high-doping-density regime $>10^{16}$ cm^{-3} the observed damping times are shorter than those expected due to the optical phonon scattering mechanism alone. This is because in this regime impurity scattering is the dominant scattering mechanism.

One way around this problem is the use of a remotely doped structure. In this case the scatterers are not in the same region as the electrons that perform the plasma oscillation (Figure 7). The higher mobility should lead to a longer decoherence time.

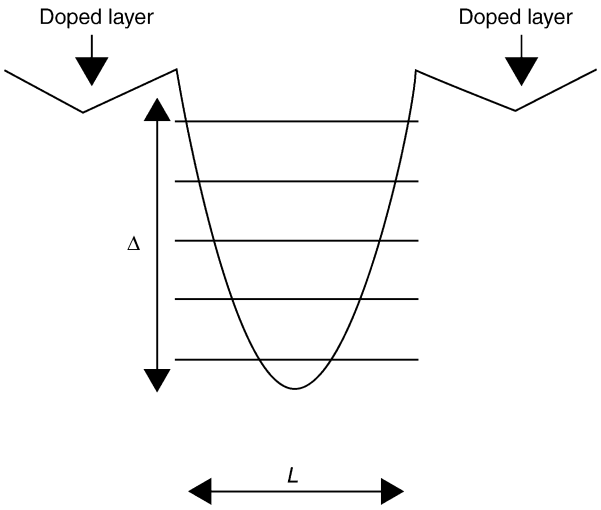


Figure 7 Schematic drawing of a modulation-doped PQW. The doped layers are outside the quantum well.

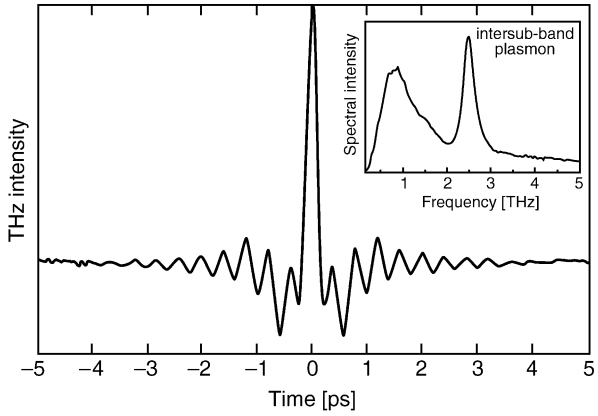


Figure 8 THz autocorrelation signal of a 1400 Å PQW excited by 780 nm laser pulses ($T = 5$ K). Inset: Fourier transform of the recorded autocorrelation. Reproduced with permission from Bratschitsch R, Müller T, Kersting R, Strasser G and Unterrainer K (2000) Coherent THz emission from optically pumped intersubband plasmons in parabolic quantum wells. In: *Ultrafast phenomena XII*. Springer series in Chemical Physics, vol. 66, pp. 204. New York: Springer-Verlag.

For this purpose Bratschitsch *et al.* used modulation doped GaAs/AlGaAs parabolic quantum wells (PQWs), with widths L in the range of 1200–2000 Å and carrier sheet densities n_s of $1.7 \times 10^{11} - 5 \times 10^{11}$ cm^{-2}. The electrons confined within the parabolic potential form a so-called quasi-3D electron system. Figure 8 shows the THz autocorrelation trace of a modulation doped PQW ($L = 1400$ Å, $n_s = 5 \times 10^{11}$ cm^{-2}) excited by 780 nm laser pulses. The density of the optically generated carriers is kept well below the carrier density due to the modulation doping inside the PQW.

The spectrum of the emitted THz radiation (inset of Figure 8) consists of two components: a broad background and a narrow line (FWHM: 0.3 THz) with a center frequency of 2.55 THz. The low-frequency broadband component is found for all the different PQW samples and is independent of the PQW sample structure. It can also be observed at room temperature while the narrowband component completely vanishes. The origin of the broadband component is due to THz generation at the surface of the sample.

The THz emission at 2.55 THz can be observed within a wide range of excitation wavelengths (815–760 nm) above the 819 nm GaAs bandgap ($T = 5$ K). When varying the excitation density over one order of magnitude ($5.6 \times 10^{10} - 6.2 \times 10^{11}$ cm^{-2}), the emission frequency does not change. The narrowband emission results from the oscillation of the carriers inside the PQW. The observed frequency corresponds to the intersub-band plasmon of the PQW. The expected frequency of the intersub-band plasmon is given by

$$\omega_0 = \sqrt{\frac{8\Delta}{L^2 m^*}} \qquad [3]$$

where Δ is the energetic depth of the well, L is the well width, and m^* the effective mass. For the 1400 Å wide PQW the formula yields 2.2 THz for the intersubband plasmon frequency.

The excitation mechanism for the intersub-band plasmon is due to screening of the surface depletion field by the electron–hole pairs injected by the ultrafast laser pulse. In this way the electrons inside the quantum well experience a kick and begin to oscillate with their eigenfrequency (Figure 9). This is supported by the fact that the oscillation can be excited over a large wavelength range of the femtosecond pulses. The onset of intersub-band plasmon emission appears at 815 nm near the 819 nm GaAs bandgap ($T = 5$ K) and can be observed to up to 760 nm. This large range implies that the excitation mechanism is clearly a nonresonant phenomenon, in contrast to the THz quantum beat experiments. Due to ultrafast field screening, a symmetric nanostructure can emit optically driven THz radiation.

The width of the intersub-band plasmon emission line is 0.3 THz (full width at half maximum; FWHM) for the 1400 Å PQW and even 0.2 THz for the 2000 Å PQW sample. The linewidth is a factor of 5 smaller than the smallest linewidth measured for a 3D plasmon emitter. The width of the intersub-band emission is nearly unaffected when increasing the excitation wavelength from near bandgap to approximately 60 nm above bandgap. However, it shows a linear dependence ($5.6 \times 10^{10} - 6.2 \times 10^{11}$ cm^{-2}) on

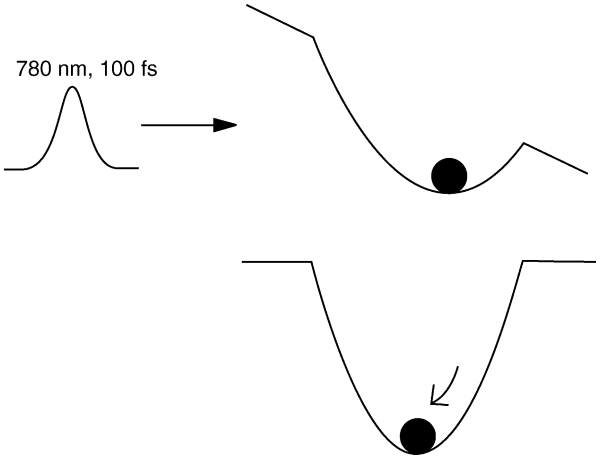


Figure 9 Symbolic drawing of the excitation mechanism: an oscillation of the carriers in the parabolic potential is initiated by ultrafast field screening. Reproduced from Bratschitsch R, Müller T, Kersting R, Strasser G and Unterrainer K (2000) Coherent terahertz emission from optically pumped intersubband plasmons in parabolic quantum wells. *Applied Physics Letters* 76: 3501, with permission from the American Institute of Physics.

the number of optically excited carriers. These optically injected carriers are added to those from the modulation doping and cause a faster dephasing of the intersubband plasmon oscillation. This linear dependence due to carrier–carrier scattering has also been observed in Wolter's experiments dealing with the dephasing of Bloch oscillations in semiconductor superlattices. The linewidth stays constant up to about 50 K (0.33 THz = 1.36 meV) and then starts to broaden until the sample temperature reaches 150 K where the PQW emission is buried by the surface-generated THz emission. The broadening of the line between $T = 50$ K and $T = 150$ K can be explained by the onset of optical phonon scattering in addition to interface roughness scattering which is the dominant scattering mechanism at low temperatures. The onset and linear increase of the line broadening has also been observed with Fourier Transform Infrared (FTIR) absorption and THz time-domain cross-correlation spectroscopy measurements by Kersting and co-workers.

The combination of the designability of the transition frequency, the narrowband emission, and the absence of any processing of the sample make modulation-doped PQWs attractive and easy-to-use THz emitters.

Other THz emission experiments of heterostructures include the optically pumped emission of THz radiation from grating-coupled intrasub-band (2D) plasmons in a doped single quantum well observed by Sekine and co-workers or optically excited THz emission from coherent plasmons in doped superlattices detected by Bratschitsch and co-workers.

Few-Cycle THz Spectroscopy of Semiconductor Quantum Structures

There are a large number of studies that use few-cycle THz radiation for the investigation of semiconductor bulk materials. The complex index of refraction in the THz regime gives insight into the conductivity (i.e., carrier mobility) of the material. The free carriers that change the index of refraction by Drude absorption can be provided either by doping (n-doped silicon or GaAs) or by photo-excitation. An optical pump-THz probe experiment with undoped GaAs reveals the different intervalley scattering times.

There are compelling reasons to use THz radiation to excite and probe carriers in semiconductor quantum structures. The photon energies are comparable to the sub-band spacings, carrier kinetic energies, and phonon energies. In addition, THz radiation does not generate minority carriers so that the experiments directly probe free carriers rather than excitons.

Heyman *et al.* investigated a modulation-doped GaAs/AlGaAs multiple quantum well ($d = 510$ Å, 10×) with a transition energy between the two lowest sub-bands of 1.5 THz and a carrier density of $n_s = 2.75 \times 10^{10}$ cm^{-2}. On top of the sample an aluminum Schottky gate is evaporated and the well is contacted with AuGe alloyed ohmic contacts. The wells can be depleted of carriers by applying a negative gate voltage to the Schottky gate (−10 V gate bias). By modulating the voltage between 0 V and −10 V the response of the 2D electron system can be measured. Otherwise, the recorded signal is dominated by the contribution of the bulk GaAs substrate that dramatically chirps the THz pulse due to its frequency-dependent index of refraction. The THz pulses are coupled into the cleaved edge of the quantum well sample so that the electric field is parallel to the growth direction in order to excite the intersub-band transition. The electrons' response is measured with a cross-correlation technique. Here the THz radiation transmitted through the quantum well is mixed with a short THz pulse emitted from low-temperature grown GaAs to get the desired time resolution in the experiment. This allows the detection of the transmitted signal in amplitude and phase. Figure 10 shows the cross-correlation signal of the QW. The signal from the carriers rises during the first 2 ps in response to the THz field and then continues to oscillate at a constant frequency of 1.5 THz. Due to the femtosecond time resolution in this experiment it is possible to observe the phase of the electrons even during the process of excitation. Initially, the incident THz pulse generates a coherent superposition of states in the first and second sub-band. After the THz driving pulse is over

(>2 ps) the carriers continue to radiate at the intersub-band transition frequency, and the signal is damped out by the free induction decay. Since the amplitude and phase of the radiated electric field is recorded, not only the absorption due to the QW electrons but also the phase shift associated with the absorption can be extracted (Figure 11).

This experiment showed the response of the QW electrons to radiation which is resonant to the intersub-band transition. Using the same experimental technique Kersting *et al.* investigated the response of electrons to a non-resonant excitation of the intersub-band transition. Instead of a square QW, a modulation-doped parabolic quantum well is used in this experiment which has the advantage that the transition frequency is independent of the carrier concentration (Kohn's theorem) and the applied electric field. Furthermore the THz radiation is coupled to the intersub-band transition via a metallic Schottky grating. Figure 12 shows the response of the QW electrons to resonant THz radiation. The electrons follow the driving pulse and continue to radiate at the intersub-band transition frequency. The response to a nonresonant THz pulse is shown in Figure 13. Initially, the electrons inside the QW follow the driving pulse but when the driving pulse has ended a phase jump occurs and the electrons lock to their eigenfrequency and slowly lose their coherence. This behavior can be understood by using time-dependent perturbation theory to solve Schrödinger's

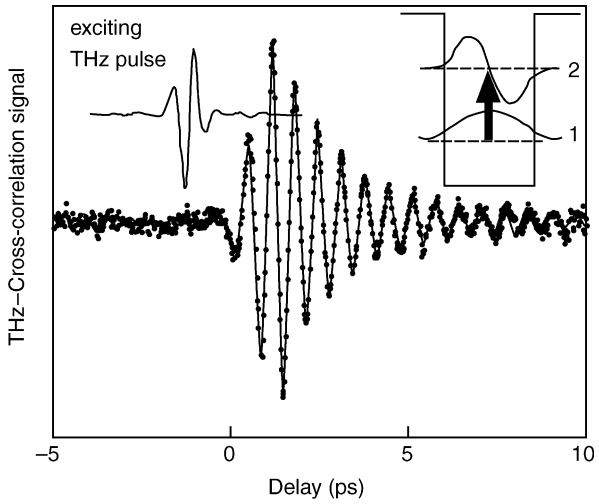


Figure 10 Cross-correlation signal obtained by electrically modulating the charge density in a 510 Å wide quantum well. The THz pulse generates a coherent superposition of the $n = 1$ and $n = 2$ sub-band states. The resulting oscillating polarization radiates and is damped out by the free induction decay. The upper trace shows the incident THz pulse. Reprinted from Unterrainer K, Kersting R, Bratschitsch R, *et al.* (2001) Few-cycle THz spectroscopy of semiconductor quantum structures. *Physica E* 9: 76, with permission from Elsevier.

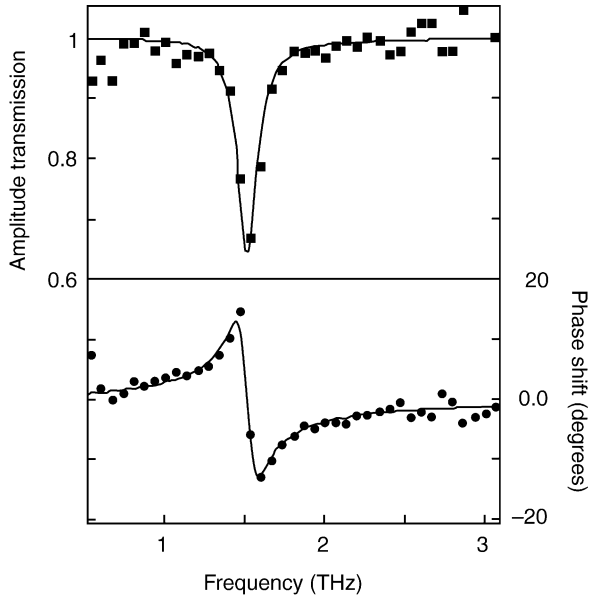


Figure 11 Change in the absorption coefficient and index of refraction in the quantum well due to the quantum well electrons. Solid points are calculated from the time-domain data. Reproduced from Heyman J, Kersting R and Unterrainer K (1998) Time-domain measurement of intersubband oscillations in a quantum well. *Applied Physics Letters* 72: 644, with permission from the American Institute of Physics.

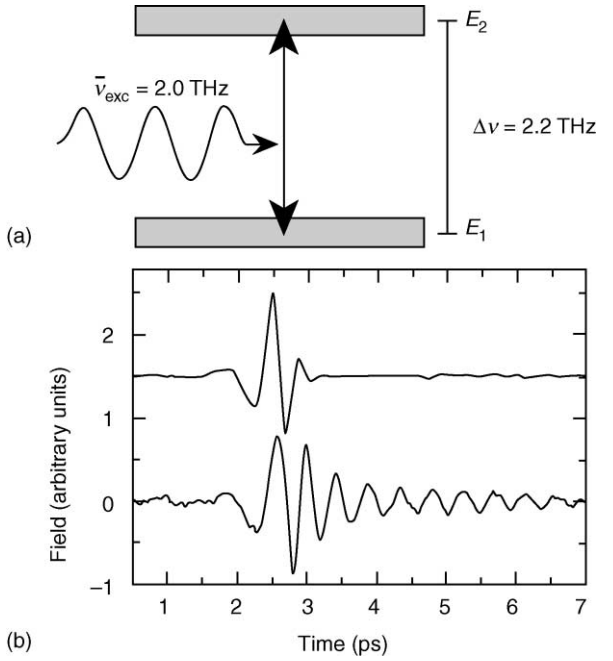


Figure 12 Resonant excitation: (a) band diagram of an intersub-band transition of 2.2 THz and driving THz field at 2.0 THz; (b) exciting THz pulse (upper curve) and cross-correlation signal (lower curve). Reproduced from Kersting R, Bratschitsch R, Strasser G, Unterrainer K and Heyman J (2000) Sampling a THz dipole transition with sub-cycle time-resolution. *Optics Letters* 25: 272, with permission from the Optical Society of America.

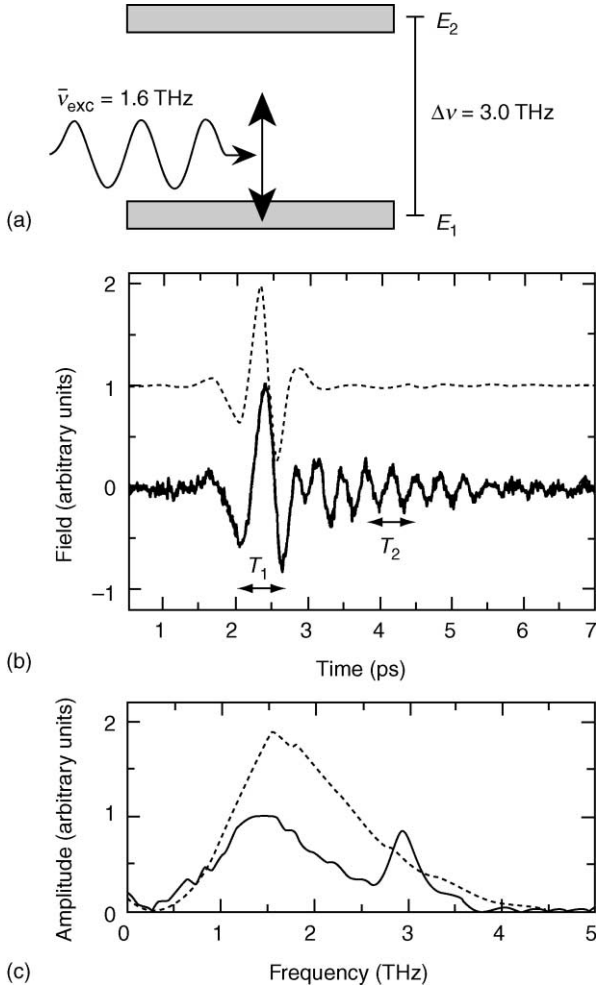


Figure 13 Nonresonant excitation: (a) band diagram of an intersub-band transition of 3.0 THz and driving THz field at 1.6 THz; the oscillation frequency changes from T_1 to T_2; (b) exciting THz pulse (upper curve) and cross-correlation signal (lower curve); (c) spectra of the exciting THz pulse (dotted curve) and cross-correlation signal (solid curve). Reproduced from Kersting R, Bratschitsch R, Strasser G, Unterrainer K and Heyman J (2000) Sampling a THz dipole transition with sub-cycle time-resolution. *Optics Letters* 25: 272, with permission from the Optical Society of America.

equation for a two-level system with an electromagnetic driving field. While the two-level system is treated quantummechanically, the dipole Hamiltonian is used for the incoming THz radiation: $H = H^0 + H'(t)$ and $H'(t) = \mu E(t)$. Making the ansatz $\Psi(t) = a_1(t)\Psi_1 + a_2(t)\Psi_2$ one gets a system of two coupled equations for $a_1(t)$ and $a_2(t)$. They can be decoupled by the condition that the population in the lower level is close to one ($\left\|a_1^2\right\| \approx 1$) since the THz radiation field is not very strong. Figure 14 shows the result of the model calculation that reproduces the change in frequency as well as the observed phase jump after the driving pulse is over.

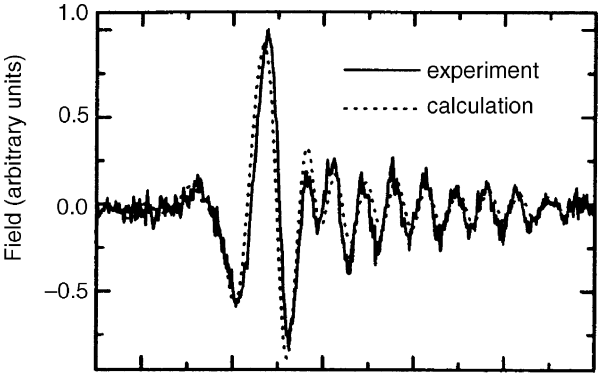


Figure 14 Model calculation of the polarization (dotted curve) and experimental data (solid line). Reproduced from Kersting R, Bratschitsch R, Strasser G, Unterrainer K and Heyman J (2000) Sampling a THz dipole transition with sub-cycle time-resolution. *Optics Letters* 25: 272, with permission from the Optical Society of America.

These experiments are all performed in the linear regime but with nonadiabatic excitation conditions (half-cycle THz pulses). The excellent time resolution together with sufficiently high THz intensities can open up a previously unexplored field of quantum optics beyond the slowly varying envelope approximation.

Conclusions

The recent progress in the development of semiconductor sources together with room temperature THz detection techniques (free-space electro-optic sampling) paves the way for the commercialization of THz technology. In particular, spectrally sensitive THz imaging seems to have great potential in medicine and quality control of products.

See also

Semiconductor Materials: GaAs Based Compounds. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-based Structures.

Further Reading

Auston DH, Cheung KP and Smith PR (1984) Picosecond photoconducting Hertzian dipoles. *Applied Physics Letters* 45: 284.

Bonvalet A, Nagle J, Berger V, Migus A, Martin J-L and Joffre M (1996) Femtosecond infrared emission resulting from coherent charge oscillations in quantum wells. *Physical Review Letters* 76: 4392.

Bratschitsch R, Kersting R, Müller T, Strasser G, Unterrainer K, Fischler W and Höpfel RA (1999) Coherent THz plasmons in GaAs/AlGaAs superlattices. *Physica B* 272/1-4: 375.

Bratschitsch R, Müller T, Kersting R, Strasser G and Unterrainer K (2000) Coherent terahertz emission from optically pumped intersubband plasmons in parabolic quantum wells. *Applied Physics Letters* 76: 3501.

Chamberlain M (ed.) (1999) *Terahertz Spectroscopy and Applications*, SPIE Proceedings (Europto Series), vol. 3828, pp. 1–454.

Dekorsy T, Auer H, Waschke C, Bakker H, Roskos H, Kurz H, Wagner V and Grosse P (1995) Emission of submillimeter electromagnetic waves by coherent phonons. *Physical Review Letters* 74: 738.

Gornik E and Kersting R (2001) Coherent THz emission in semiconductors. In: Tsen KT (ed.) *Ultrafast Physical Processes in Semiconductors*, vol. 25, pp. 396. New York: Academic Press.

Heyman J, Kersting R and Unterrainer K (1998) Time-domain measurement of intersubband oscillations in a quantum well. *Applied Physics Letters* 72: 644.

Kersting R, Unterrainer K, Strasser G, Kauffmann H and Gornik E (1997) Few-cycle thz emission from cold plasma oscillations. *Physical Review Letters* 79: 3038.

Kersting R, Heyman J, Strasser G and Unterrainer K (1998) Coherent plasmons in n-doped GaAs. *Physical Reviews B* 58: 4553.

Kersting R, Bratschitsch R, Strasser G, Unterrainer K and Heyman J (2000) Sampling a THz dipole transition with sub-cycle time-resolution. *Optics Letters* 25: 272.

Nun MC and Orenstein J (1997) Terahertz time-domain spectroscopy. In: Gruener G (ed.) *Millimeter Wave Spectroscopy of Solids*. Berlin: Springer.

Roskos H, Nuss M, Shah J, Leo K, Miller D, Fox A, Schmitt-Rink S and Köhler K (1992) Coherent submillimeter-wave emission from charge oscillations in a double-well potential. *Physical Review Letters* 68: 2216.

Sarukura N, Ohtake H, Izumida S and Liu Z (1998) High average power radiation from femtosecond laser-irradiated InAs in a magnetic field and its elliptical polarization characteristics. *Journal of Applied Physics* 84: 654.

Sekine N, Yamanaka K, Hirakawa K, Voßebürger M, Haring-Bolivar P and Kurz H (1999) Observation of terahertz radiation from higher-order two-dimensional plasmon models in GaAs/AlGaAs single quantum wells. *Applied Physics Letters* 74: 1006.

Sha W, Smirl A and Tseng W (1995) Coherent plasma oscillations in bulk semiconductors. *Physical Review Letters* 74: 4273.

Shah J (1999) *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*. Springer Series in Solid-State Sciences, vol. 115. Berlin: Springer.

Waschke C, Roskos H, Schwedler R, Leo K, Kurz H and Köhler K (1993) Coherent submillimeter-wave emission from bloch oscillations in a semiconductor superlattice. *Physical Review Letters* 70: 3319.

Wolter F, Martini R, Tolk S, *et al.* (1999) Influence of carrier–carrier scattering on intraband dephasing. *Superlattice and Microstructure* 26: 93.

Zhang X-C, Hu B, Darrow J and Auston D (1990) Generation of femtosecond electromagnetic pulses from semiconductor surfaces. *Applied Physics Letters* 56: 1011.

TIME-RESOLVED FLUORESCENCE

Contents

Laser Applications
Measurements in Polymer Science

Laser Applications

B Valeur, Conservatoire National des Arts et Métiers, Paris, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Fluorescence is extensively used in physical, chemical, material, biological, and medical sciences, for investigating the structure and dynamics of matter or living systems. After excitation of a fluorescent sample by a short pulse of light, the fluorescence decay is, in the simplest case, a single exponential whose time constant is the lifetime of the excited state of the fluorescent species (i.e., the average time during which these species stay in the excited state). However, interactions of a fluorescent species with its micro-environment often perturb the fluorescence decay so that information on this micro-environment can be obtained. More specifically, when intermolecular photophysical processes

(such as electron transfer, proton transfer, energy transfer, excimer, or exciplex formation) occur during the lifetime of the excited state, competition with normal de-excitation results in a fluorescence decay that contains the rate constants for those processes. Consequently, the fluorescence decay is generally a sum of discrete exponentials, or a more complicated function; for example, sometimes the system is characterized by a distribution of decay times. Thus it is of major interest to recover the parameters characterizing the δ-pulse response of a fluorescent sample (i.e., the response to an iznfinitely short pulse of light expressed as the Dirac function δ). With this aim, two time-resolved techniques, pulse fluorometry and phase-modulation fluorometry, are usually employed.

In pulse fluorometry, the sample is excited by a short pulse of light, and the fluorescence response is the δ-pulse response of the sample convoluted by the instrument response. In phase-modulation fluorometry, the sample is excited by modulated light at variable frequency, and the harmonic response of the sample obtained is the Fourier transform of the δ-pulse response. The first technique works in the time domain, and the second technique in the frequency domain. Pulse fluorometry and phase-modulation fluorometry are theoretically equivalent, but the principles of instrumentation are different. In both techniques, lasers are currently used as sources of pulsed or modulated light.

It is the aim of this article to explain the principles of pulse and phase fluorometries using lasers as light sources, and to compare these techniques.

General Principles of Time-Resolved Fluorometry

The principles of time-resolved fluorometry are illustrated in Figure 1. For any excitation function $E(t)$, the fluorescence response $R(t)$ of the sample is the convolution product of this function by the δ-pulse response $I(t)$:

$$R(t) = E(t) \otimes I(t) = \int_{-\infty}^{t} E(t')I(t - t')\,dt' \qquad [1]$$

In pulse fluorometry, the sample is excited by a short pulse of light and the fluorescence response is recorded as a function of time. If the duration of the pulse is not short with respect to the time constants of the fluorescence decay, the fluorescence response is the convolution product given by eqn [1]: the fluorescence intensity increases, goes through a maximum, and becomes identical to the true δ-pulse response $i(t)$ as soon as the intensity of the light pulse becomes negligible. In this case, data analysis for the determination of the parameters characterizing the δ-pulse response requires deconvolution of the fluorescence response.

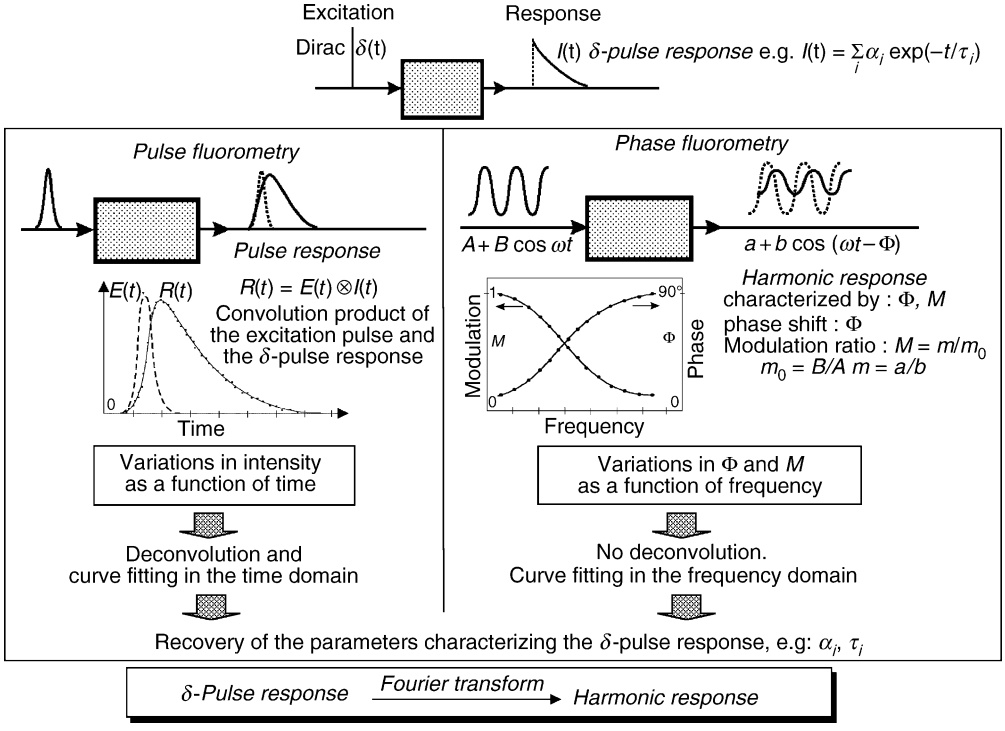


Figure 1 Principles of time-resolved fluorometry.

In phase fluorometry, the sample is excited by a sinusoidally modulated light at high frequency. The fluorescence response, which is the convolution product (eqn [1]) of the δ-pulse response by the excitation function, is sinusoidally modulated at the same frequency, but delayed in phase and partially demodulated with respect to the excitation. The modulation ratio M is defined as the ratio of the modulation depth m (AC/DC ratio) of the fluorescence and the modulation depth of the excitation m_0 (see **Figure 1**). The phase shift Φ and the modulation ratio M characterize the harmonic response of the system. These parameters are measured as a function of the modulation frequency. No deconvolution is necessary because the data are directly analyzed in the frequency domain.

The relationship between the harmonic response and the δ-pulse response can be obtained as follows. When the excitation function is sinusoidal, i.e., of the following form:

$$E(t) = E_0[1 + m_0 \exp(\mathrm{i}\omega t)] \qquad [2]$$

where ω is the angular frequency ($= 2\pi f$), the response of the system can be calculated using eqn [1]. It can then be shown that

$$M \exp(-\mathrm{j}\Phi) = \int_0^\infty i(t) \exp(-\mathrm{i}\omega t)\, \mathrm{d}t \qquad [3]$$

where $i(t)$ is the normalized δ-pulse response according to

$$\int_0^\infty i(t)\, \mathrm{d}t = 1 \qquad [4]$$

This important expression shows that the harmonic response expressed as $M \exp(-\mathrm{j}\Phi)$ is the Fourier transform of the δ-pulse response.

It is convenient to introduce the sine and cosine transforms P and Q of the δ-pulse response:

$$P = \int_0^\infty i(t) \sin(\omega t)\, \mathrm{d}t \qquad [5]$$

$$Q = \int_0^\infty i(t) \cos(\omega t)\, \mathrm{d}t \qquad [6]$$

If the δ-pulse response is not normalized according to eqn [4], then eqns [5] and [6] should be replaced by

$$P = \frac{\int_0^\infty I(t) \sin(\omega t)\, \mathrm{d}t}{\int_0^\infty I(t)\, \mathrm{d}t} \qquad [7]$$

$$Q = \frac{\int_0^\infty I(t) \cos(\omega t)\, \mathrm{d}t}{\int_0^\infty I(t)\, \mathrm{d}t} \qquad [8]$$

Since eqn [3] can be rewritten as $M \cos \Phi - iM \times \sin \Phi = Q - iP$, it is easy to show that the phase shift and the modulation ratio are given by

$$\Phi = \tan^{-1}\left(\frac{P}{Q}\right) \qquad [9]$$

$$M = [P^2 + Q^2]^{1/2} \qquad [10]$$

For a single exponential decay, the δ-pulse response is

$$I(t) = \alpha \exp(-t/\tau) \qquad [11]$$

where τ is the decay time and α is the pre-exponential factor or amplitude. The phase shift and relative modulation are related to the decay time by

$$\tan \Phi = \omega\tau \qquad [12]$$

$$M = \frac{1}{(1 + \omega^2\tau^2)^{1/2}} \qquad [13]$$

For a multi-exponential decay with n components, the δ-pulse response is

$$I(t) = \sum_{i=1}^{n} \alpha_i \exp(-t/\tau_i) \qquad [14]$$

Note that the fractional intensity of component i, i.e., the fractional contribution of component i to the total steady-state intensity, is

$$f_i = \frac{\int_0^\infty I_i(t)\, \mathrm{d}t}{\int_0^\infty I(t)\, \mathrm{d}t} = \frac{\alpha_i\tau_i}{\sum_{i=1}^{n} \alpha_i\tau_i} \qquad [15]$$

with, of course, $\sum_{i=1}^{n} f_i = 1$.

Using eqns [7] and [8], the sine and cosine Fourier transforms, P and Q, are given by

$$P = \frac{\omega \sum_{i=1}^{n} \dfrac{\alpha_i\tau_i^2}{1 + \omega^2\tau_i^2}}{\sum_{i=1}^{n} \alpha_i\tau_i} = \omega \sum_{i=1}^{n} \frac{f_i\tau_i}{1 + \omega^2\tau_i^2} \qquad [16]$$

$$Q = \frac{\sum_{i=1}^{n} \dfrac{\alpha_i\tau_i}{1 + \omega^2\tau_i^2}}{\sum_{i=1}^{n} \alpha_i\tau_i} = \sum_{i=1}^{n} \frac{f_i}{1 + \omega^2\tau_i^2} \qquad [17]$$

It should be noted that the optimum frequency for decay time measurements using either the phase shift or the modulation ratio is, according to eqns [12] and [13], such that $\omega\tau$ is close to 1, i.e., $f \approx 1/(2\pi\tau)$. Therefore, for decay times of 10 ps, 1 ns and 100 ns, the optimum frequencies are about 16 GHz, 160 MHz and 1.6 MHz, respectively.

In the case of a single exponential decay, eqns [12] and [13] provide two independent ways of measuring the decay time:

– by phase measurements:

$$\tau_\Phi = \frac{1}{\omega}\tan\Phi \qquad [18]$$

– by modulation measurements:

$$\tau_M = \frac{1}{\omega}\left(\frac{1}{M^2} - 1\right)^{1/2} \qquad [19]$$

The values measured in these two ways should, of course, be identical and independent of the modulation frequency. This provides two criteria to check whether an instrument is correctly calibrated, by using a lifetime standard whose fluorescence decay is known to be a single exponential.

It should be noted that the measurement of a decay time is fast (a fraction of a second) for a single exponential decay since a single frequency suffices. It is also of interest that a significant difference between values obtained by means of eqns [18] and [19] is compelling evidence for nonexponentiality of the fluorescence decay.

Pulse Fluorometers

The most widely used technique in time domain is the time-correlated single-photon counting technique, preferably called single-photon timing technique (SPT). Time-gated systems are less popular. Streak cameras offer a very good time resolution (a few picoseconds or less) but the dynamic range is smaller than that of the single photon-timing technique. The instruments that provide the best time resolution (a few tens of femtoseconds) are based on fluorescence up-conversion, but they are very expensive and not commercially available. Because of space limitations, only the single-photon timing technique will be presented below.

Principle of the Single-Photon Timing Technique

The basic principle of this technique relies on the fact that the probability to detect a single photon at time t, after an exciting pulse, is proportional to the fluorescence intensity at that time. After timing and recording the single photons following a large number of exciting pulses, the decay of the fluorescence intensity is reconstructed.

Figure 2 shows a schematic diagram of a conventional single-photon counting instrument using a pulsed laser as an excitation source. An electrical pulse associated to the optical pulse is generated (e.g. by a photodiode or the electronics associated to the excitation source) and routed – through a constant-fraction discriminator – to the start input of the time-to-amplitude converter (TAC). Meanwhile, the sample is excited by the optical pulse and emits fluorescence. The optics is tuned (e.g., by means of a neutral density filter) so that the photomultiplier detects no more than one fluorescence photon for each exciting pulse. The corresponding electrical pulse is routed – through a constant-fraction discriminator – to the stop input of the TAC. The latter generates an output pulse whose amplitude is directly proportional to the delay time between the start and the stop pulses. The height analysis of this pulse is achieved by an analog-to-digital converter and a multichannel

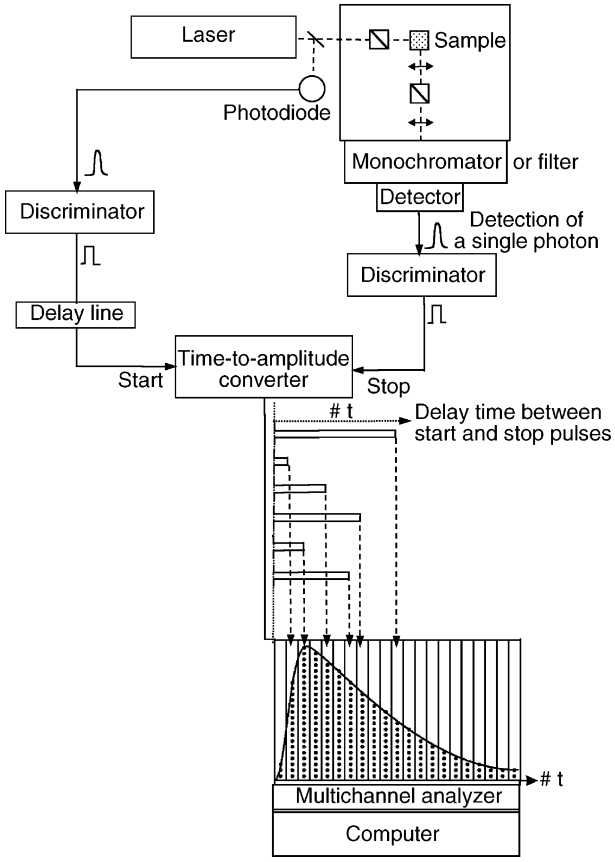


Figure 2 Schematic diagram of a single-photon timing fluorometer.

analyzer (MCA), which increases by one the contents of the memory channel corresponding to the digital value of the pulse. After a large number of excitation and detection events, the histogram of pulse heights represents the fluorescence decay curve.

Obviously, the larger the number of events, the better the accuracy of the decay curve. The required accuracy depends on the complexity of the δ-pulse response of the system; for instance, a high accuracy is, of course, necessary for recovering a distribution of decay times.

When deconvolution is required, the time profile of the exciting pulse is recorded under the same conditions by replacing the sample with a scattering solution (e.g., a suspension of colloidal silica (Ludox)).

It is important to note that the number of detected fluorescence photons must be kept much smaller than the number of exciting pulses ($<0.01-0.05$ stops per pulse), so that the probability of detecting two fluorescence photons per exciting photon is negligible. Otherwise, the TAC will take into account only the first fluorescence pulse and the counting statistics will be distorted: the decay will appear shorter than it is in reality. This effect is called the 'pile-up effect'.

Laser Sources for the Single-Photon Timing Technique

Mode-locked dye lasers and titanium-doped sapphire lasers are well suited to the single-photon timing technique: in fact, they can generate pulses over broad wavelength ranges, and the pulse widths are in the femtosecond and picosecond ranges with a high repetition rate. Flash lamps running in air, or filled with gas, can also be used, but the range of excitation wavelengths is limited; they deliver nanosecond pulses, and the repetition rate is not high (10^4-10^5 Hz). Dye lasers require the handling of large volumes of dye solutions which must be replaced when the dye is bleached. In contrast, the Ti:sapphire laser is a solid-state laser that does not have this drawback and delivers shorter light pulses. Moreover, long operational lifetime, operational simplicity, reliability, and broad tunability are distinct advantages. For these reasons, Ti:sapphire lasers have become more popular, and are described below.

Ti:sapphire lasers

Titanium-doped sapphire crystals are obtained by replacing a small fraction of the Al atoms of the host (Al_2O_3) with Ti atoms. When the concentration of Ti is low, the optical quality and the mechanical properties of Ti:sapphire crystals are comparable to those of ruby crystals in ruby lasers. The absorption spectrum of Ti:sapphire crystals exhibits a maximum at 490 nm and a shoulder at 540 nm. Therefore, pumping can be achieved by an Argon ion laser or a green solid-state laser. The emission spectrum ranges from 600 to 1,100 nm. Such a broad emission spectrum results from electronic coupling between the electronic levels of the 3d electrons with the lattice vibrations. Since excited-state absorption is not possible in the spectral region of the pump and the fluorescence spectrum, the tuning range is broad: it extends from 680 to 1,100 nm, i.e., almost over the entire fluorescence spectrum.

The width of the pulses that can be generated by Ti:sapphire lasers ranges from a few tens of femtoseconds to about 100 picoseconds. Commercially available Ti:sapphire lasers can operate either in the femtosecond configuration or in the picosecond configuration. In time-resolved fluorescence experiments using the single-photon timing technique, the instrument response is limited by the time response of the photomultiplier; the shortest width of this response that can be measured with the faster detectors (i.e., microchannel plate photomultipliers) is about 30–40 ps (see below). Therefore, there is no reason to operate the Ti:sapphire laser in the femtosecond configuration. In the picosecond configuration, the pulse width is $<2-100$ ps.

Figure 3 shows the six-mirror folded cavity design for the picosecond configuration in a Tsunami

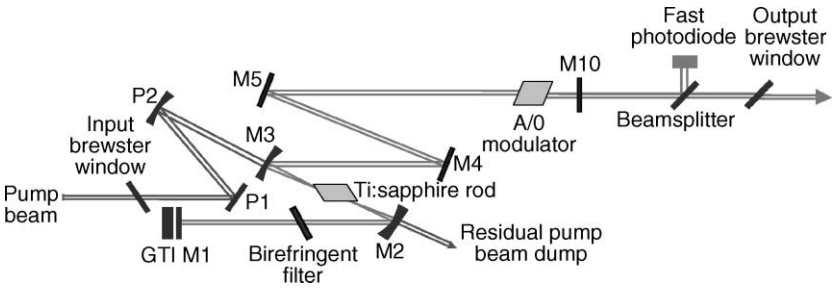


Figure 3 Beam path of the laser cavity of the Tsunami Ti:sapphire laser (Spectra Physics) operating in the picosecond configuration (Adapted from the Tsunami brochure with permission of Spectra Physics).

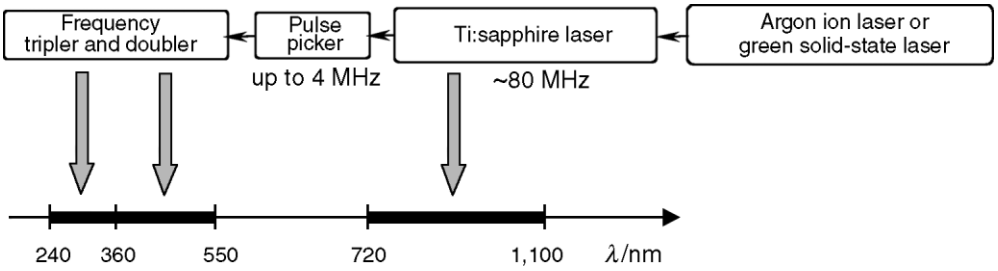


Figure 4 Schematics of the excitation system in time-resolved fluorescence experiments.

Ti:sapphire laser (Spectra Physics). A Brewster's angle acousto-optic modulator, driven by a low-power rf source, insures an 80 MHz repetition rate. Consequently, in contrast to standard passive mode-locking systems, the laser can operate for long periods without dropouts or shut-downs. Gires–Tournois Interferometer (GTI) mirror spacing is used for tuning the laser cavity. Wavelength tuning is achieved by means of a two-plate birefringent filter placed within the laser cavity at Brewster's angle. The tuning range depends on the mirror sets that are used. With broadband cavity mirrors, the entire tuning range of Ti:sapphire is covered, i.e., from below 700 nm to greater than 1,000 nm, and changes in wavelength can be made in a matter of minutes.

When pumping with a 10 W argon ion laser, the average output power is about 1.5 W at 800 nm and the pulse energy is about 15 nJ. Argon ion lasers can now be advantageously replaced by cw green solid-state lasers (532 nm) in which the gain medium ($Nd:YVO_4$) is pumped by diodes, and a LBO crystal placed inside the cavity ensures frequency doubling. These lasers offer several advantages: the head is compact and does not require external cooling water; it is insensitive to temperature, and thermal equilibration is very rapid.

Since the excited-state lifetime of fluorescent compounds can be as long as several hundreds of nanoseconds, the repetition rate of 80 MHz is too high to let the fluorescence of long lifetime samples vanish before a new exciting pulse is generated. Therefore, it must be limited to a few MHz, thanks to a pulse picker using a Bragg-angle acousto-optic modulator synchronized to the mode-locked 80 MHz pulse train. Selection of pulses from the train can thus be achieved at a maximum pulse selection rate of 8 MHz.

For excitation of fluorescent samples, wavelengths in the visible and UV regions are desirable. Frequency doublers and triplers using LBO and BBO crystals, respectively, have been designed with this aim. The possible tuning ranges are shown in Figure 4. When pumping with a 10 W argon ion laser, the average output powers are about 300 mW and 120 mW for frequency doubling and tripling, respectively.

With Ti-sapphire lasers and microchannel plate photomultipliers, decay times as short as 10 ps can be measured.

The results of an experiment are shown in Figure 5.

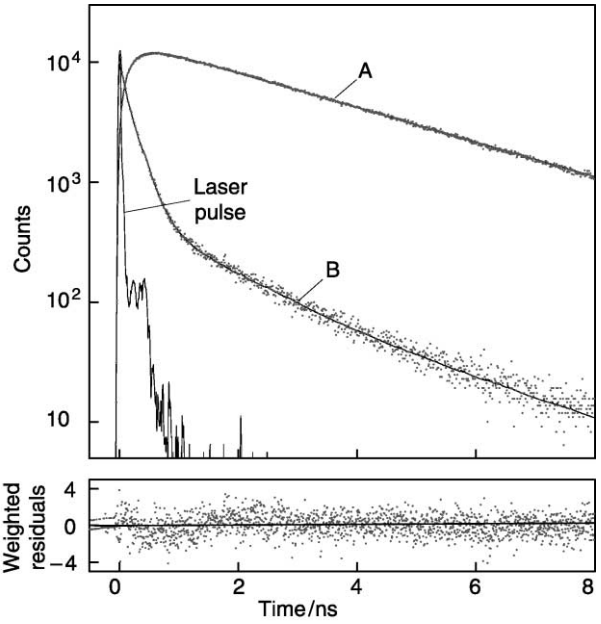


Figure 5 Example of data obtained by the single-photon timing technique using a Spectra-Physics setup composed of a titanium-sapphire Tsunami laser pumped by an argon ion laser, a pulse selector, and doubling (LBO) and tripling (BBO) crystals. The detector is a Hamamatsu MCP photomultiplier R3809U. The sample is a solution of DANS1 (a derivative of dansylamide) in a mixture CH_3CN/H_2O (60:40 v/v) at pH 0.8. Excitation wavelength: 286 nm. Channel width: 9.8 ps. A: Observation through a bandpass filter (300–370 nm) of the protonated form that undergoes photoinduced deprotonation. B: Observation through a cut-off filter ($\lambda > 400$ nm) of the neutral form; the observed risetime corresponds to the formation of the neutral form from the protonated form. Global analysis of the two curves leads to two time constants: 183 ± 9 ps and 3.01 ± 0.03 ns (Global $\chi^2_R = 1.26$) (Métivier R, Leray I and Valeur B, unpublished results. Métivier R, doctoral thesis, Ecole Normale Supérieure de Cachan, 2003).

Pulsed diode lasers

Picosecond diode laser heads offer an interesting alternative to mode-locked lasers and are much less expensive. They can produce light pulses as short as 50–90 ps with repetition rates from single shot to 40–80 MHz. Peak powers up to 500 mW can be obtained. However, the main disadvantage is the absence of tunability, and the number of wavelengths is limited: 375, 400, 440, 635 to 1,550 nm. No wavelength below 375 nm is presently available.

Phase-Modulation Fluorometers

In the frequency domain, the phase and modulation measurements can be taken by using either a cw laser and an optical modulator, or the harmonic content of a pulsed laser.

Phase Fluorometers Using a cw Laser and an Optical Modulator

The optical modulator is generally a Pockels cell. Low-power cw lasers are sufficient and not as expensive as pulsed lasers, so that one can afford several lasers of this type in order to get various excitation wavelengths. Available wavelengths from cw lasers are given in Table 1. A Xenon lamp can also be used; however, the optical modulator works better with a cw laser.

Figure 6 shows the schematic diagram of a multifrequency phase-modulation fluorometer using a cw laser and a Pockels cell. A beamsplitter reflects a few percent of the incident light towards a reference photomultiplier (via, or not, a cuvette containing a reference scattering solution). The fluorescent sample and a reference solution (containing either a scatter or a reference fluorescent compound) are placed in a rotating turret. The emitted fluorescence or scattered light is detected by a photomultiplier through a monochromator or an optical filter. The Pockels cell is driven by a frequency synthesizer and the photomultiplier response is modulated by varying the voltage at the second dynode by means of another frequency

Table 1 Examples of available wavelengths from low-power cw lasers

Laser	*Wavelength (nm)*
He–Cd	325; 442.5
He–Ne	543.3; 594; 611.9; 632.8
Ar^{+a}	457.9; 476.5; 488.0; 514.5

[a]High-power argon ion lasers can also provide wavelengths of 351.1 and 363.8 nm.

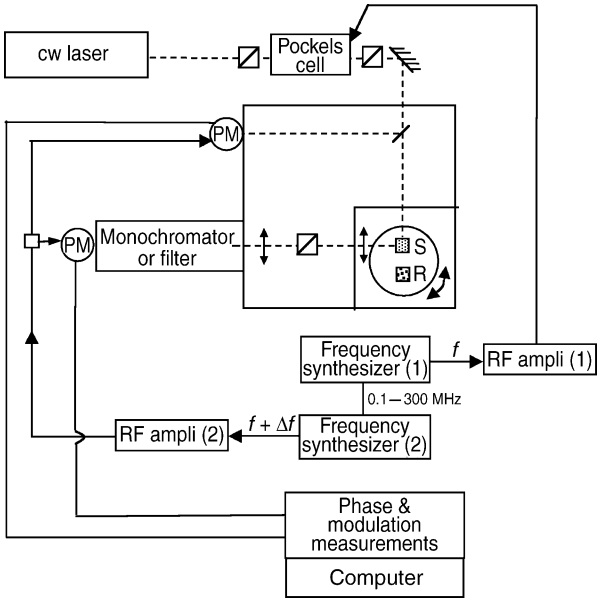


Figure 6 Schematic diagram of a multifrequency phase-modulation fluorometer using a cw laser and a Pockels cell (S: sample; R: reference; PM: photomultiplier).

synthesizer locked in phase with the first one. The two synthesizers provide modulated signals that differ in frequency by a few tens of Hz in order to achieve cross-correlation (heterodyne detection). This procedure offers excellent accuracy because the phase and modulation information contained in the signal is transposed to the low-frequency domain, where phase shifts and modulation depths can be measured with a much better accuracy than in the high-frequency domain.

Practically, the phase delay ϕ_R and the modulation ratio m_R of the light emitted by the scattering solution are measured with respect to the signal detected by the reference photomultiplier. Then, after rotation of the turret, the phase delay ϕ_F and the modulation ratio m_F for the sample fluorescence are measured with respect to the signal detected by the reference photomultiplier. The absolute phase shift and modulation ratio of the sample are then $\Phi = \phi_F - \phi_R$ and $M = m_F/m_R$, respectively.

Figure 7 shows an example of an experiment.

Phase Fluorometers Using the Harmonic Content of a Pulsed Laser

The type of laser source that can be used is exactly the same as that for single-photon timing pulse fluorometry (see above). Such a laser system that delivers pulses in the picosecond range, with a repetition rate of a few MHz, can be considered as an intrinsically modulated source. The harmonic content of the pulse train – which depends on the

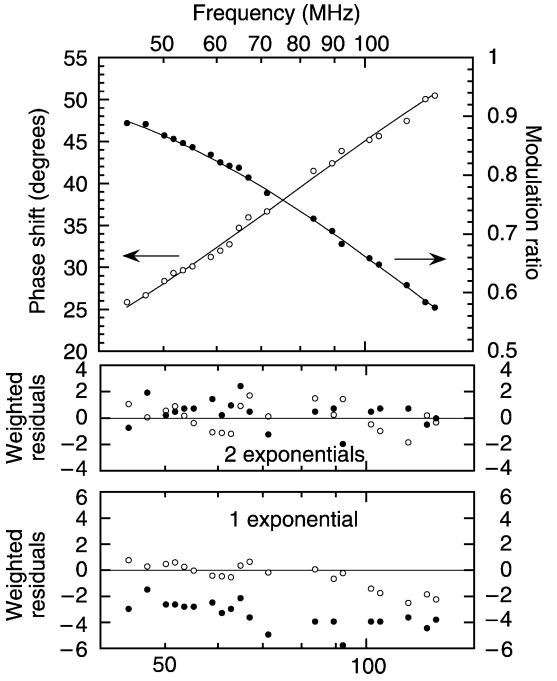


Figure 7 Example of data obtained by the phase and modulation technique using a fluorometer is equipped with a cw Omnichrome He–Cd laser (442 nm) and a Pockels cell operating at frequencies ranging from 0.1 to 200 MHz. The sample is a solution of a calcium complex of a crowned merocyanine (DCM-crown) in acetonitrile (observation through a Balzers bandpass filter centered at 581 nm). The solid line corresponds to the best fit with a biexponential decay. Weighted residuals are shown for the best fit with one and two exponentials. The time constants are 1.87 ± 0.04 and 0.35 ± 0.09 ns with fractional intensities of 0.85 and 0.15, respectively. The value of χ_r^2 is 1.01. (Redrawn from Martin MM, Plaza P, Meyer YH, *et al.* (1996) Steady-state and picosecond spectroscopy of Li^+ or Ca^{2+} complexes with a crowned merocyanine. Reversible photorelease of cations. *Journal of Physical Chemistry* 100: 6879–6888, with permission from the American Chemical Society).

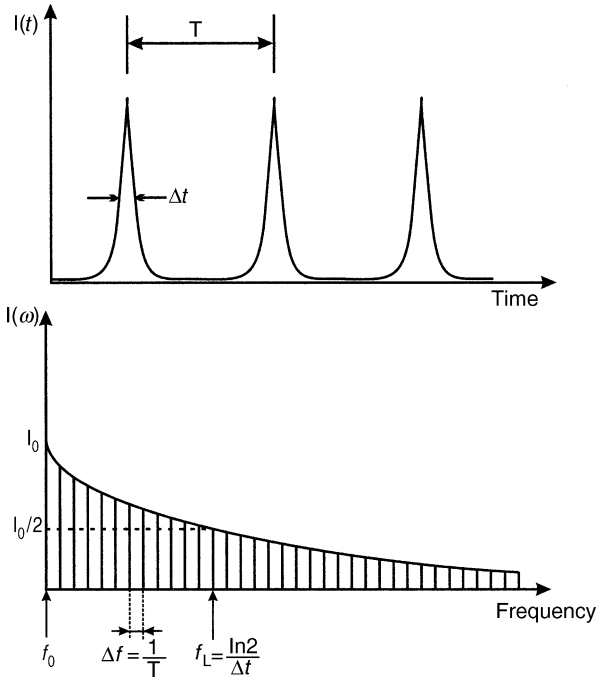


Figure 8 Harmonic content of a pulse train. Example of numerical values: $\Delta f = 4$ MHz; $\Delta t = 4$ Ps; $f_L = 44$ GHz.

pulse width (as illustrated in Figure 8) – extends to several gigahertz.

For high-frequency measurements, normal photomultipliers are too slow, and microchannel plate photomultipliers are required. However, internal cross-correlation is not possible with the latter and an external mixing circuit must be used.

The time resolution of a phase fluorometer, using the harmonic content of a pulsed laser and a microchannel plate photomultiplier, is comparable to that of a single-photon timing instrument using the same kind of laser and detector.

Data Analysis

In both pulse and phase fluorometries, the most widely used method of data analysis is based on a nonlinear least-squares method. The basic principle of this method is to minimize a quantity which expresses the mismatch between data and fitted function. This quantity is the reduced chi-square χ_r^2, defined as the weighted sum of the squares of the deviations of the experimental response $R(t_i)$ from the calculated responses $R_c(t_i)$:

$$\chi_r^2 = \frac{1}{\nu}\sum_{i=1}^{N}\left[\frac{R(t_i) - R_c(t_i)}{\sigma(i)}\right]^2 \qquad [20]$$

where N is the total number of data points and $\sigma(i)$ is the standard deviation of the ith data point, i.e., the uncertainty expected from statistical considerations (noise). ν is the number of degrees of freedom ($\nu = N - p$, where p is the number of fitted parameters). The value of χ_r^2 should be close to 1 for a good fit.

In the single-photon timing technique, the statistics obey the Poisson distribution and the expected deviation $\sigma(i)$ is approximated to $[R(t_i)]^{1/2}$ so that eqn [20] becomes

$$\chi_r^2 = \frac{1}{\nu}\sum_{i=1}^{N}\frac{[R(t_i) - R_c(t_i)]^2}{R(t_i)} \qquad [21]$$

In phase fluorometry, no deconvolution is required: curve fitting is performed in the frequency domain, directly using the variations of the phase shift Φ and the modulation ratio M as functions of the modulation frequency. Phase data and

modulation data can be analyzed separately, or simultaneously. In the latter case the reduced chi-squared is given by

$$\chi_r^2 = \frac{1}{\nu}\left[\sum_{i=1}^{N}\left[\frac{\Phi(\omega_i) - \Phi_c(\omega_i)}{\sigma_\Phi(\omega_i)}\right]^2 + \sum_{i=1}^{N}\left[\frac{M(\omega_i) - M_c(\omega_i)}{\sigma_M(\omega_i)}\right]^2\right] \quad [22]$$

where N is the total number of frequencies. In this case, the number of data points is twice the number of frequencies, so that the number of degrees of freedom is $\nu = 2N - p$.

In addition to the value of χ_r^2, it is useful to display graphical tests. The most important of them is the plot of the weighted residuals defined as

$$W(t_i) = \frac{R(t_i) - R_c(t_i)}{\sigma(i)} \quad [23]$$

where $\sigma(i) = [R(t_i)]^{1/2}$ for single-photon counting data. The fit is satisfactory when the weighted residuals are randomly distributed around zero.

Concluding Remarks

From the instrumental point of view, the latest generations of pulse and phase fluorometers both use pulsed lasers (usually Ti:sapphire lasers) and microchannel plate detectors; only the electronics is different. Since the instrument response is mainly limited by the time response of the detector, the time resolution is the same for both techniques. Moreover, the optical module is identical so that the total cost of instrumentation is similar.

There are, however, significant differences between pulse and phase fluorometers from the methodological point of view. In pulse fluorometry, the single-photon timing technique offers various advantages: direct visualization of the fluorescence decay; high sensitivity; outstanding dynamic range and linearity; and well-defined statistics (Poisson distribution) allowing the accurate weighing of each point in data analysis. In phase-modulation fluorometry, the main advantage is that no deconvolution is necessary, while this operation is often necessary in pulse fluorometry and requires great care in recording the instrument response, especially for very short decay times. Moreover, the short acquisition time for phase shift and modulation ratio measurements at a given frequency is a distinct advantage.

See also

Ultrafast Laser Techniques: Pulse Characterization Techniques.

Further Reading

Alcala JR, Gratton E and Jameson DM (1985) A multifrequency phase fluorometer using the harmonic content of a mode-locked laser. *Analytical Instrumentation* 14: 225–250.

Demas JN (1983) *Excited-State Lifetime Measurements*. New York: Academic Press.

Jameson DM, Gratton E and Hall RD (1984) The measurement and analysis of heterogeneous emissions by multifrequency phase and modulation fluorometry. *Applied Spectroscopy Reviews* 20: 55–106.

Lakowicz JR (ed.) (1991) *Topics in Fluorescence Spectroscopy, Vol. 1: Techniques*. New York: Plenum Press.

Lakowicz JR, Laczko G and Gryczynski I (1986) 2 GHz frequency-domain fluorometer. *Review of Scientific Instruments* 57: 2499–2506.

O'Connor DV and Phillips D (1984) *Time-Correlated Single Photon Counting*. London: Academic Press.

Valeur B (2002) *Molecular Fluorescence. Principles and Applications*, chap. 6. Weinheim, Germany: Wiley-VCH.

Measurements in Polymer Science

T A Smith, University of Melbourne, Melbourne, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In this article we explore the use of ultrashort pulsed lasers in time-resolved fluorescence measurements for the study of the dynamics of excimer formation and energy transfer/migration in synthetic aromatic polymers. The temporal profile of monomer and excimer emission, and the use of time-resolved fluorescence polarization approaches are discussed.

The use of lasers in chemistry has most often exploited the high degree of spectral purity achievable with these devices, in both photochemical and

high-resolution spectroscopic applications. Through the use of lasers, we also have the ability to generate pulses of light of ultrashort duration (a few femtoseconds) and of incredibly high peak powers (petawatts), characteristics significantly beyond those achievable with conventional (incoherent) light sources. This enables the kinetics of the many photophysical and photochemical processes that occur on ultrafast time-scales with high efficiency within polymers, to be determined through various time-resolved measurements.

These processes are dealt with in detail in specialist photochemistry textbooks and so knowledge of the terms associated with photochemistry and photophysics is assumed (e.g., electronic configurations, orbitals, spin, multiplicity, etc.).

Following the absorption of light, an isolated chromophore will be in one of its excited electronic and vibrational states. Once formed, the excited state of a molecule will only survive for a limited time, with the molecule losing its excess energy to return to a preferred lower energy state (electronic configuration). It can do this through a number of spontaneous, radiative, and nonradiative deactivation processes, each with its own characteristic rate constant k_{ij}. Nonradiative pathways include internal conversion (IC), such as intramolecular vibrational redistribution, and vibrational coupling to the bath, and inter-system crossing (ISC) between singlet and triplet manifolds. Radiative pathways involve the emission of a photon of light, in the form of fluorescence or phosphorescence.

The time the molecule spends in the excited state is determined by the rates of all the pathways that contribute to the overall deactivation of that state. The lifetime, τ, of a singlet state is therefore defined as:

$$\tau = \frac{1}{k_{\mathrm{ISC}} + k_{\mathrm{IC}} + k_{\mathrm{Rad}} + \cdots} = \frac{1}{\sum k_{ij}} \qquad [1]$$

In systems of higher chromophore concentration, such as aromatic-containing polymers, additional, bimolecular pathways are available through which the excess excited state energy may be dissipated, making it rare that a single fluorescence 'lifetime' is observed from polymer systems. The observation of complex (multi- or nonexponential) fluorescence decay profiles sometimes invokes the use of the terms 'average lifetimes' or 'lifetime distributions'.

The complex nonexponential fluorescence decay behavior observed, particularly from polymer systems, has been variously attributed to one or more of the following factors: molecular aggregation (e.g., excimer/exciplex equilibria); acid–base equilibria (e.g., proton transfer); one-, two-, or three-dimensional energy transfer and energy migration (and variations in the singlet energy migration length); electron transfer; molecular motion (translational diffusion, rotational relaxation); vibrational or solvent relaxation; concentration quenching; fluorescent impurities, etc. The kinetics of excimer formation are further complicated by heterogeneity of kinetically distinct monomer and/or excimer species, configurational and conformational influences upon excimer formation, and the time dependence of energy trapping (excimer formation) through nonequilibrium diffusion-controlled processes. Here, we will concentrate on excimer formation and energy migration in synthetic aromatic polymers, as shown schematically in Figure 1.

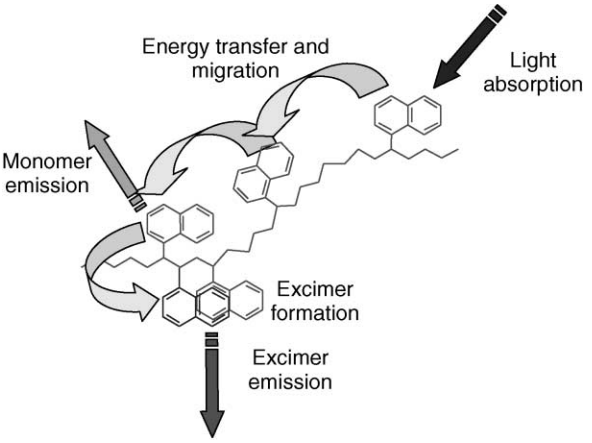


Figure 1 Schematic representation of some of the photophysical processes in polymers.

Ultrafast Laser Spectroscopic Techniques

A wide range of experimental techniques and instrumentation is available for the study of time-resolved processes in polymers. The most widely used methods rely on the detection of emission from the polymer. Fluorescence-based techniques are popular for investigating time-dependent processes in polymers, since the time-scales involved in many of the photochemical and photophysical processes discussed correspond with the time-scales on which fluorescence occurs (normally on the subpicosecond to nanosecond time-scales). Phosphorescence usually occurs on significantly longer time-scales and therefore is less likely to be competitive with fluorescence or other radiative or nonradiative deactivation pathways under normal conditions. Fluorescence-based techniques are attractive for reporting on a number of time-dependent processes since fluorescence is inherently multidimensional, it can be readily resolved temporally, and fluorescence detection can

be extremely sensitive, achieving single photon and single molecule detection levels.

Amongst the most commonly used fluorescence-based techniques are: time-correlated single photon counting (TCSPC, or single photon timing), fluorescence transient digitization, and ultrafast streak cameras. Methods such as transient absorption (flash photolysis) can also provide some details regarding the identity of short-lived species formed following photo-excitation. Other ultrafast laser techniques have been applied to the study of time-dependent processes in polymers in order to achieve higher temporal resolution. So-called pump–probe methods provide the ultimate in time-resolution and can be combined with fluorescence detection through the use of Kerr shutters or ultrafast fluorescence upconversion techniques. Other, more complex pump–probe methods, such as transient grating and multiple pulse photon echo measurements, are potentially very powerful tools in the study of time-resolved processes in polymers.

The generation of short pulses of light can be achieved through numerous mechanisms, including active or passive mode-locking and Q-switching techniques, and superradiance processes (e.g., in nitrogen lasers). Lasers capable of producing short (~ns) pulses of light by Q-switching are usually characterized by high energies per pulse (from μJ up to many hundreds of mJ/pulse) at low (~Hz–kHz) pulse repetition rates. Lasers producing pulses of ultrashort duration (often in the picosecond or femtosecond region) through mode-locking techniques, characteristically operate at high pulse repetition rates (~MHz) and with lower energies per pulse (~nJ/pulse). It should be noted that even pulses of a few nJ energy delivered in ~100 fs, still have peak powers of tens of kW. These high peak powers enable the generation of a range of excitation wavelengths suitable for the study of polymer systems through nonlinear optical wavelength conversion processes, such as second-harmonic generation (SHG) and optical parametric oscillation (OPO) devices.

Excimer Formation

Excimers (or *excited-state dimers*) are formed when an aromatic molecule, while in its excited electronic state, $^1M^*$, interacts with a second molecule of the same chemical type in its ground electronic state, ^1M. Excimer formation is often observed in concentrated molecular systems such as concentrated solutions and molecular crystals; however, it can also occur in dilute solutions of polymers containing aromatic chromophores, or in the confined environment of solid polymer systems such as homopolymer films.

The structure of excimers in solution is generally thought to be a face-to-face, coplanar, sandwiched, or in some cases, partially eclipsed arrangement of the two constituent aromatic moieties, which attempt to adopt the lowest possible energy configuration, with a separation of 3–4 Å, as depicted in Figure 2. A range of possible excimer configurations may exist rather than just the two extremes.

A feature of excimers is that they have a dissociative ground state (Figure 3), i.e., they do not exist prior to the absorption of light by one monomer chromophore and thus cannot be formed directly without the coming together of the two constituent chromophores within the lifetime of the $^1M^*$. Excimers, being of lower energy than the corresponding excited monomer species and with this dissociative ground state, usually exhibit characteristically

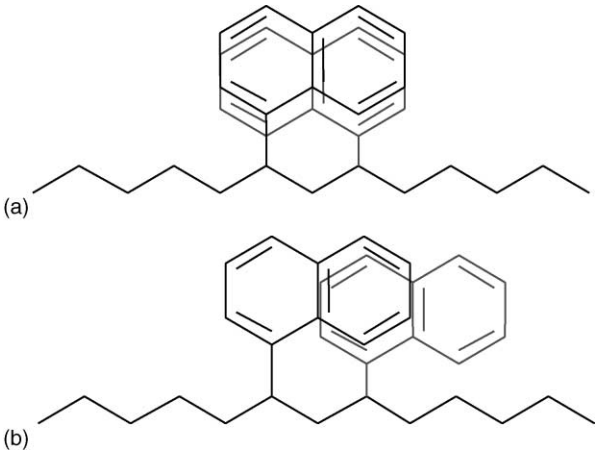


Figure 2 (a) Sandwiched and (b) partially eclipsed excimer configurations.

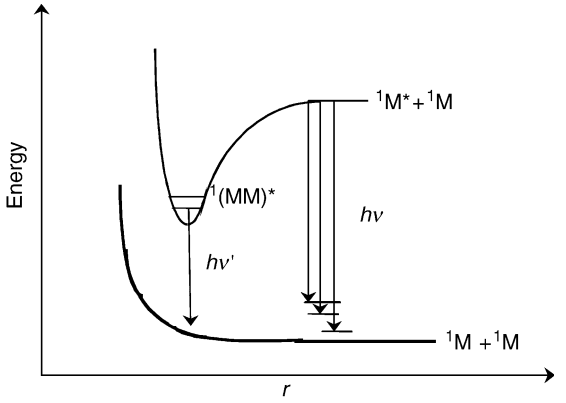


Figure 3 Schematic potential energy diagram for excimer formation and decay. r represents the monomer–monomer separation.

broad, structureless emission to the red of the monomer emission (Figure 4). Upon the emission of fluorescence from an excimer, the complex loses its excess energy and dissociates back to two ground state monomer units (i.e., 2 ^{1}M). Excimers may also dissociate back to a noninteracting pair of $^1M^*$ and ^{1}M.

Excimer formation of simple aromatic molecules in solution is often modeled on the kinetic scheme proposed by Birks and co-workers (Figure 5). From Figure 5, if a δ-pulse excitation, represented by $h\nu_{abs}$, is used to produce an initial concentration of excited singlet monomer molecules, $[^1M^*]_o$, then such a scheme results in the following rate equations:

$$\frac{d[^1M^*]}{dt} = k_{MD}[^1(MM)^*] - X[^1M^*] \quad [2]$$

$$\frac{d[^1(MM)^*]}{dt} = k_{DM}[^1M][^1M^*] - Y[^1(MM)^*] \quad [3]$$

where $X = k_{FM} + k_{NM} + k_{DM}[^1M^*] = k_M + k_{DM}[^1M]$, $Y = k_{FD} + k_{ND} + k_{MD} = k_D + k_{MD}$.

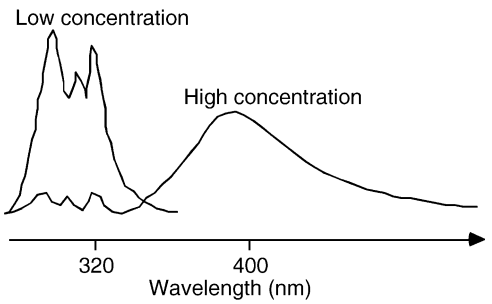


Figure 4 Schematic representation of monomer and excimer emission.

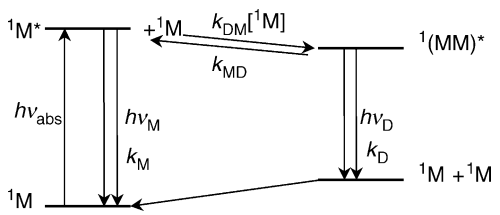


Figure 5 Kinetic scheme for excimer formation where $h\nu_M$ and $h\nu_D$ represent the emission of photons by the monomer and excimer species respectively, k_M and k_D comprise the sum of the various rate constants for all deactivation pathways for the monomer and excimer respectively (excluding k_{MD} and k_{DM}) (e.g., $k_M = k_{FM} + k_{NM}$ where k_{FM} and k_{NM} are the radiative (fluorescence) and nonradiative deactivation rate constants respectively of the monomer).

Solving these rate equations and applying the initial conditions $[^1M^*]_t = [^1M^*]_o$ and $[^1(MM)^*]_o = 0$ at $t = 0$, yields expressions given below for the decay of monomer and excimer fluorescence intensity with time. The emission intensity (monitored in a spectral region in which only the monomer $[^1M^*]$ species emits), will have the temporal form given by eqn [4]:

$$i_M(t) = k_{FM}\frac{[^1M^*]}{[^1M^*]_o} = \left[k_{FM}\frac{(\lambda_2 - X)}{(\lambda_2 - \lambda_1)}\right]\left[\exp(-\lambda_1 t) + A\exp(-\lambda_2 t)\right] \quad [4]$$

where

$$A = \frac{(X - \lambda_1)}{(\lambda_2 - X)},$$

$$\lambda_{1,2} = \frac{1}{2}\left[X + Y \mp \left\{(Y - X)^2 + 4k_{MD}k_{DM}[^1M^*]\right\}^{1/2}\right]$$

The emission intensity collected from a spectral region in which the excimer ($^1[MM]^*$) species only emits, will have the temporal form given by eqn [5]:

$$i_D(t) = k_{FD}k_{DM}\frac{[^1D^*]}{[^1M^*]_o} = \left[\frac{k_{FD}k_{DM}[^1M]}{(\lambda_2 - \lambda_1)}\right]\left[\exp(-\lambda_1 t) - \exp(-\lambda_2 t)\right] \quad [5]$$

If excimer formation follows Birks' scheme, then the decay profile of the monomer fluorescence should therefore be described by the sum of two exponential decay terms, and the excimer emission by an exponential growth term and a decay term (Figure 6). Only two decay constants, λ_1 and λ_2, should be required to fit both the monomer and excimer fluorescence profiles successfully. This is often not observed, which has led to the introduction of more complex interpretations.

In synthetic aromatic polymers, the effective local concentration of the aromatic molecules can be high while the concentration of polymer chains in solution may be very low. This means that excimer formation can be significant in such polymers, even in dilute solution. In fact, in some polymers the emission is dominated by excimers over monomers. The fact that the excimers are of lower energy than the corresponding excited monomer species also means that they can act as effective traps of excitation energy that might be migrating/relocating within a polymer chain as discussed below.

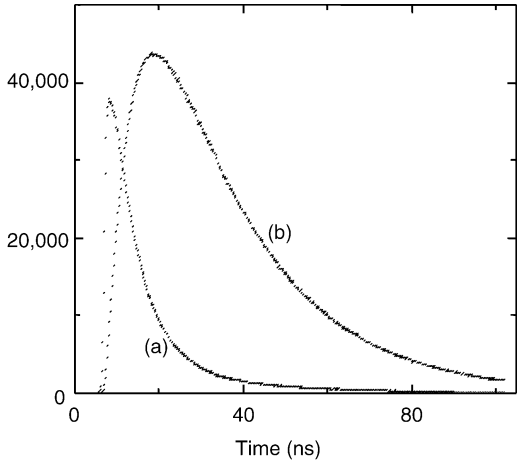


Figure 6 Examples of (a) monomer and (b) excimer fluorescence decays.

Excimer formation in solutions of aromatic polymers is generally thought to proceed via a complicated route that may involve a combination of electronic energy transfer and migration, segmental rotation, and chromophore diffusion. It is believed that energy migrates along a polymer chain until sites suitable for excimer formation are encountered, at which point the energy is trapped. A number of mechanisms for the formation of suitable excimer sites in polymers are then conceivable. These include:

(i) interactions between chromophores belonging to different chains (this *inter*molecular excimer formation is highly concentration dependent);
(ii) interactions between two chromophores far removed from each other, but on the same chain. This necessitates mainchain conformational changes and side group motions during the excited state lifetime, so as to bring both molecules together;
(iii) near neighbor interactions (c.f. Hirayama's '$n = 3$' rule). The configuration required for excimer formation need not exist prior to excitation provided that the two constituent groups can attain the required geometry within the lifetime of the excited chromophore; or
(iv) preformed sites, suitable for excimer formation, present in the polymer before the initial absorption of light occurs. These may be formed by interactions between nearest neighbor chromophores or between two chromophores on loops of the same, or different chains in cases where chain motion is restricted.

The first two of these mechanisms depend upon segmental motion of the polymer chain and the third mechanism may depend upon rotation of the chromophore/s to adopt the close proximity and optimum orientational configuration required for excimer formation. In solution, this may be, at least partially, diffusion-controlled, causing a dominant complicating factor in the kinetics of excimer formation in polymeric systems. One approach is therefore to substitute a time-dependent expression for the rate coefficient of excimer formation, $k_{DM}(t) = A + B/\sqrt{t}$, (based on the Einstein–Smoluchowski diffusion theory), for k_{DM}, into the rate equations. Such an approach produces the following expression for the time dependence of the monomer fluorescence, $I_M(t)$ in the presence of diffusion-controlled excimer formation and the absence of excimer back dissociation:

$$I_M(t) = A \exp\left\{ - at - b\sqrt{t}\right\} \qquad [6]$$

In some cases, such as when excimer back dissociation cannot be neglected, more complex functional forms have been derived, such as:

$$I_M(t) = A \exp\left\{ - at - b\sqrt{t}\right\} + B \exp\left\{ - t/\tau\right\} \qquad [7]$$

Similar expressions can be derived for the time dependence of the excimer fluorescence, $I_E(t)$. More detailed expressions for the term $k(t)_{DM}$, involving the complete form of the diffusion rate coefficient, have also been used in such approaches.

A related approach is the concept of 'convolution kinetics', in which the intimate dependence of excimer formation upon the monomer decay is emphasized. Once formed, the excimer is assumed to decay exponentially. The excimer fluorescence temporal profile is predicted to take the form:

$$I_E(t) = k_1(t) I_M(t) \otimes \exp\left(\frac{-t}{\tau}\right) \qquad [8]$$

where $\otimes$ represents the convolution integral and $I_E(t)$ and $I_M(t)$ are the measured decay profiles in the excimer and monomer emission regions respectively.

Electronic Energy Transfer, Energy Migration and Trapping

In polymers the high local concentration of chromophores makes additional deactivation pathways competitive with the other photophysical processes.The excess energy may transfer from the excited fluorophore of interest to a lower energy chromophore. Three of the more common possible mechanisms, through which energy transfer can occur, are:

(i) *Radiative* ('*trivial*') – involves the reabsorption of donor emission by an acceptor chromophore.

(ii) *Resonance* (*Dipole–Dipole*, *Förster*) – long-range (15–100 Å), nonradiative, single-step energy transfer process arising from Coulombic (dipole–dipole) interactions between an excited-state donor chromophore, D^*, and an initially unexcited acceptor, *A*. Förster derived a statistical expression for the (first-order) rate constant, for such a process, considering three-dimensional, dipole–dipole interactions between stationary molecules:

$$k(R)_{D^*\to A} = \frac{9000\kappa^2\phi_o \ln 10}{128\pi^5\eta^4 N\tau_D R^6}\int_0^\infty \frac{f_D(\bar{\nu})\varepsilon_A(\bar{\nu})}{\bar{\nu}}d\bar{\nu} = \frac{1}{\tau_D}\left(\frac{R_o}{r_D}\right)^6 \quad [9]$$

where R is the donor–acceptor distance, ϕ_o is the quantum yield for emission from the donor in the absence of acceptor molecules, η is the refractive index of the solvent, τ_D is the lifetime of the donor excited state (reduced by the onset of electronic energy transfer), N is Avogadro's number, κ is a constant dependent on the mutual orientation of the donor and acceptor molecules, and the integral term is a measure of the spectral overlap between the donor emission profile, $f_D(\bar{\nu})$ (in quanta cm^{-1} and normalized to unit area) and the acceptor absorption profile, $\varepsilon_A(\bar{\nu})$ (converted to molar extinction coefficients and measured on a wavenumber scale). R_o is the critical transfer distance, at which the probability of resonance transfer, $k(R)_{D^*\to A}$, is equal to the probability of spontaneous decay, $1/\tau_D$.

(iii) *Exchange mechanism* – also a nonradiative, single-step process but involves the exchange of electronic excitation when the donor and acceptor are close enough so that their electron clouds overlap. Consequently it only operates over small distances. Dexter has derived a rate expression for exchange energy transfer, again assuming that the donor and acceptor molecules are stationary:

$$k_{D^*\to A} = \left(\frac{2\pi}{\hbar}\right)Z^2\int f_{D^*}(\bar{\nu})\varepsilon_A(\bar{\nu})d\bar{\nu} \quad [10]$$

where $Z^2 = K^2\exp(-2R/L)$ and K and L are simply constants.

It has been clearly demonstrated that energy transfer, from pendant aromatic monomer units to chemically bound acceptor chromophores, can occur with high efficiency and is influenced significantly by polymer structure and conformation.

In addition to energy transfer between D^* and A (being different chemical species), more than one exchange of excitation energy between chemically and spectroscopically identical chromophores, attached by covalent bonds to a polymer chain, can occur. A single step between identical units is still termed energy transfer, while more than one such step in a sequence would constitute intramolecular energy migration.

Energy migration is a long-range process, thought to occur via a series of energy transfer steps occurring between chromophores, either: (i) along the polymer chain (often the adjacent groups); (ii) across loops in a single polymer chain; or (iii) across different polymer chains. The migration continues until either a trap site is encountered or one of the chromophores excited in the migration process deactivates spontaneously. Each step is believed to occur with a probability defined by the exchange or Förster mechanisms, but since the energy transfer occurs between donor and acceptor groups, which are of the same type, the spectral overlap integral involved in both mechanisms will usually be small. The distances over which such energy transfer can occur therefore cannot be large and typically takes values of $<$10–20 Å. It is worth noting that many authors consider the energy to migrate as an incoherent exciton (where the term exciton is used in a much looser sense than is customary among solid-state physicists).

Long-range energy migration can only occur by energy transfer from the initially excited chromophore to ground state chromophores of the same, or lower, energy. If energy transfer is favorable between one chromophore and another of the same energy, then back transfer to the initially excited chromophore is also highly probable. When there is a change in the emission and absorption dipole orientations, caused by rotation of the chromophores relative to one another on the time-scale of the energy transfer step, or the transfer is to a chromophore of significantly lower energy, then the probability of energy transfer back to the original chromophore might be reduced. The lower energy chromophores on the polymer chain may therefore act as 'traps' of the excitation originally absorbed by a higher energy chromophore, even in the absence of excimer energy trap sites.

Various approaches have been proposed for modeling the kinetics of energy migration. One can consider the time-dependent nature of each energy transfer step of the migration process, and introduce a time-dependent rate coefficient for each step of the energy migration process, $k_{DD}(t)$. The form for $k_{DD}(t)$ by Sienicki *et al.*, is given by eqn [11], which leads to nonexponential functions predicted for the ensemble averaged probability that an originally excited donor

molecule is still excited at time t:

$$k_{DD}(t) = \left(\frac{\Delta}{n}\right)\gamma_\Delta c_D \Gamma(1-\Delta/n)\left(\frac{t}{\tau_{OD}}\right)^{\Delta/n-1} \quad [11]$$

where $\Delta(=1, 2, 3)$ is the dimensionality of the medium, $n(=6, 8, 10,\ldots)$ is the order of the interaction, τ_{OD} is the lifetime of the donor fluorescence, $\Gamma(x)$ is the gamma function, γ_Δ is a constant dependent upon Δ and c_D is the donor concentration.

Energy migration has also been considered as a diffusion-like process, and the Smoluchowski diffusion theory adopted. The rate coefficient for energy migration, $k(t)_{EM}$, is then expressed in a form similar to the time-dependent rate coefficient for partially diffusion-controlled reactions (eqn [12]). In this type of treatment, the rate of change of the concentration of excited monomer chromophores, $^1M^*$, is assumed to be given by eqn [13]:

$$k(t)_{EM} = 4\pi N'_A R D \rho\left[1 + R\rho(\pi D t)^{-1/2}\right] \quad [12]$$

$$\frac{d[^1M^*]}{dt} = I_o - \left(\frac{1}{\tau_o} + k(t)_{EM}\right)[^1M^*] \quad [13]$$

In these expressions, τ_o is the reciprocal of the sum of the rate constants for radiative and nonradiative deactivation, ρ is the reaction probability per collision, R is the interaction distance, and N'_A is the number density of the energy acceptors or traps ($N'_A = NC_A$, with the concentration of acceptors, C_A, in mol cm^{-3}). D is the mutual diffusion coefficient which can be replaced by a migration coefficient, Λ, R can be replaced by the Förster critical transfer distance, R_o, and the reaction probability, ρ, can be set to 0.5 (since R_o is the distance at which the probability of energy transfer is 0.5). This leads to a simplified expression for $k(t)_{EM}$, which can be substituted into eqn [13], to produce an expression (eqn [14]) for the monomer decay:

$$I(t) = A\exp\left(-\left(\frac{1}{\tau_o} + 2\pi N'_A R_o \Lambda[^1M^*]\right)t - 2N'_A R_o^2(\pi\Lambda)^{1/2}[^1M^*]\sqrt{t}\right) \quad [14]$$

Yokota and Tanimoto have also developed a statistical representation of Förster's theory and derived an expression for dipole–dipole transfer in a fluid medium in which the distribution of excited-state donor and acceptor molecules is determined, both by diffusion and by the decay and transfer of the donor excited state. According to the Yokota and Tanimoto approach, the decay law is approximated by the expression:

$$I(t) = A\exp\left\{\frac{-t}{\tau_o} - 2B\delta\left(\frac{-t}{\tau_o}\right)^{1/2}\right\} \quad [15]$$

where $\delta = [C]/[C]_o$ (C_o is the critical molar concentration of acceptor molecules given by: $[C_o] = 3000/2\pi^{3/2}NR_o^3$, and B is a complex, time-dependent function given by: $B = (1 + 10.87x + 15.5x^2 / 1 + 8.743x)^{3/4}$ with $x = D\alpha^{-1/3}t^{2/3}$, $\alpha = R_o^6/\tau_o$ and D is the 'diffusion coefficient' representing energy migration.

Emission depolarization is often exploited in the study of energy migration in polymers. Only molecules with some vector component of their absorption transition dipole lying parallel to the plane of polarization of the incident radiation, will absorb incoming radiation. In the absence of angular motion of the absorbing chromophore within the lifetime of the excited state, emission from the same chromophore will also be highly polarized; however, upon each step in the energy migration process, some loss of polarization will be incurred (**Figure 7**). The extent of depolarization of emission is usually referred to by the term fluorescence anisotropy, $r(t)$ (eqn [16]):

$$r(t) = \frac{I_\parallel(t) - I_\perp(t)}{I_\parallel(t) + 2I_\perp(t)} \quad [16]$$

The rate of energy migration can be estimated directly through observation of the time-dependent decay of $r(t)$ in the absence of other fluorescence depolarizing processes, such as molecule rotation (**Figure 8**).

One of the important properties which characterizes excitation energy transport in a system of n identical molecules, which are distributed randomly in space, is the Green function, $G(r,t)$. The Green function can be thought of as the time-dependent probability of finding an excitation at position r on the polymer chain at time t. It has been shown that the overwhelming contribution to fluorescence polarization is due to fluorescence from the sites that were initially excited. Thus, $G^s(r,t)$ is directly related to the fluorescence anisotropy if the polarization memory of secondary excited molecules can be neglected. The $r(t)$ for donor–donor energy transfer is then given by:

$$r(t) = \alpha r_{rot}(t) G^s(t) \quad [17]$$

in which r_{rot} describes the anisotropy in the absence of energy transfer, i.e., in the presence of chromophore

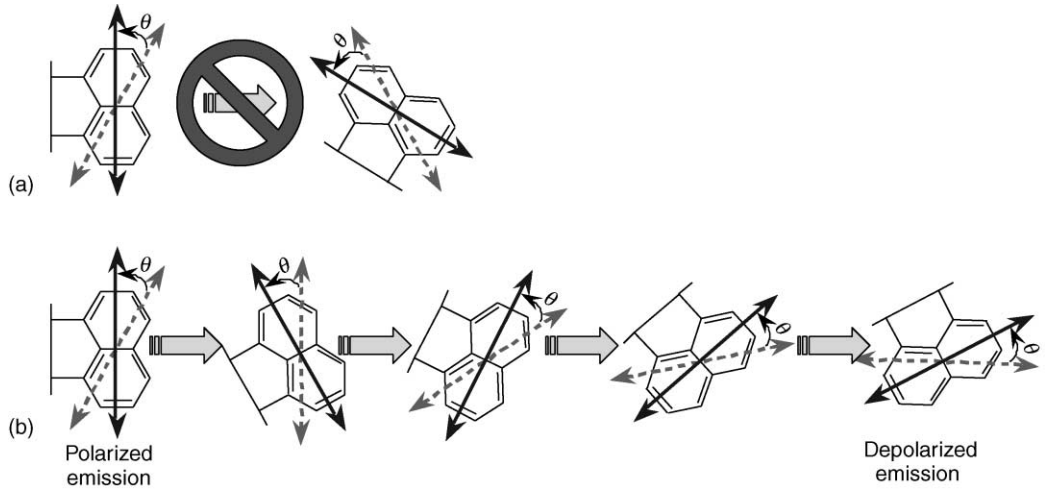


Figure 7 Schematic representation of energy migration. θ is the inherent angle between the donor emission (dotted arrow) and acceptor absorption (solid arrow) transition dipoles of the chromophore (a) orthogonal donor emission and acceptor absorption dipoles – no energy transfer/migration, emission depolarized only due to θ. (b) some vector component of the absorption dipole of the acceptor is parallel with that of the donor emission dipole – this and subsequent energy transfer steps can occur resulting in loss of emission polarization.

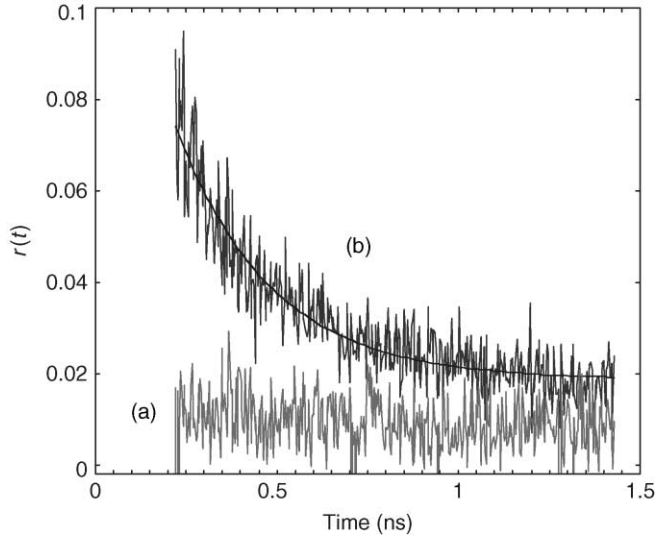


Figure 8 Temporal decays of $r(t)$ from (a) a polymer in which emission depolarization due to energy migration is complete on ultrafast time-scales, and (b) a polymer in which the rate of energy migration has been inhibited. Emission collected from polymers in frozen glass to exclude polymer motion contributions to emission depolarization.

rotation only, and α is the anisotropy from time-independent sources of depolarization.

The determination of the functional form of $G^s(r,t)$, resulting from excitation energy migration in polymer systems, is a challenging theoretical problem to which a number of approaches have been applied. The Gochanour, Anderson and Fayer approach has had some success in arriving at an expression for $G^s(r,t)$, represented as the time decay of emission polarization curves.

Polymeric systems that undergo energy migration, yet contain no excimer forming or other nonmonomeric, energy-trapping sites, are rare. In the presence of singlet energy traps, the problem can be treated as being analogous to energy transfer (via the Förster mechanism), following a diffusive process in a similar manner to that described above.

Fredrickson and Frank have derived expressions for the temporal decays of the monomer and excimer fluorescence when one-dimensional electronic singlet

energy migration is the major excimer-forming path, and segmental rotation and chromophore diffusion are considered. The Fredrickson and Frank model assumes inherently that excimers are formed predominantly via a nearest-neighbor mechanism, where rapid energy migration is at least an order of magnitude faster than any competing rotationally sampled pathways. The expressions for the fluorescence decay of both the monomer, $I_M(t)$, and excimer species, $I_E(t)$, are highly nonexponential, with the expression for $I_M(t)$ being given by eqn [18]:

$$I_M(t) = Q_{fm} k_m (1-q)^2 \times \exp[(4q^2 W - k_m - k_{rot})t]\mathrm{erfc}[2q(Wt)^{1/2}] \quad [18]$$

In eqn [18], Q_{fm} is the quantum efficiency for monomer fluorescence and k_m is a rate constant, which encompasses the total rate of monomer decay by radiative and nonradiative pathways in the absence of energy migration and rotational sampling. Similarly, k_{rot} represents the rate at which an excited monomer is transformed into an excimer by segmental rotation. The parameter W represents the rate constant for nearest-neighbor electronic energy transport between the donor (monomer) chromophores, averaged over both the transition moment orientation and the interchromophore separation associated with each dyad.

Energy transfer from pendant aromatic monomer units to a specific chemically bound acceptor chromophore can occur in some polymers following energy migration. This phenomenon is known as the 'antenna effect', which has commercial applications such as in the sensitization of photo-oxidation–reduction, i.e. reactions, in the stabilization of polymers against photodegradation or in the modeling of photosynthesis by light harvesting. The kinetic behavior of such processes is yet to be modeled definitively, but the approaches discussed above (and others) have been used with varying degrees of success to date.

List of Units and Nomenclature

Ångstrom (Å)	10^{-10} m
chromophore	An aromatic molecule that absorbs (and may emit) light
exciton	An excited state localized on a particular chromophore at any given time
femtoseconds	10^{-15} s
intermolecular	Occurring between molecules or polymer chains
intramolecular	Occurring within a single molecule or polymer chain
petawatts	10^{15} W
δ-pulse excitation	Excitation by an infinitely short laser pulse

See also

Chemical applications of lasers: Non-linear spectroscopies. **Lasers:** Excimer Laser. **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed). **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Andrews DL and Demidov AA (1999) *Resonance Energy Transfer*. New York: Wiley.

Birks JB (1970) *Photophysics of Aromatic Molecules*. London: Wiley-Interscience.

Fleming GR (1986). *Chemical Applications of Ultrafast Spectroscopy*, vol. 13. New York: Oxford University Press.

Fredrickson GH and Frank CW (1983) Nonexponential transient behavior in fluorescent polymeric systems. *Macromolecules* 16: 572–577.

Ghiggino KP and Smith TA (1993) Dynamics of energy migration and trapping in photoirradiated polymers. *Progress in Reaction Kinetics* 18: 375–436.

Gochanour CR, Andersen HC and Fayer MD (1979) Electronic excited state transport in solution. *Journal of Chemical Physics* 70: 4254–4271.

Guillet JE (1985) *Polymer Photophysics and Photochemistry: An Introduction to the Study of Photoprocesses in Macromolecules*. New York: Cambridge University Press.

Lakowicz JR (1999) *Principles of Fluorescence Spectroscopy*, 2nd edn. New York: Plenum.

O'Connor DV and Phillips D (1984) *Time-Correlated Single Photon Counting*. London: Academic Press.

Sienicki K, Itagaki H and Mattice WL (1989) On the theory of concentration depolarization of fluorescence in one- and two-component systems for multipole interactions in one, two, and three dimensional medium. *Journal of Chemical Physics* 91: 4515–4521.

Van der Meer BW, Coker G and Chen S-YS (1994) *Resonance Energy Transfer: Theory and Data*. New York: VCH.

Van Hecke GR and Karukstis KK (1998) *A Guide to Lasers in Chemistry*. Boston, MA: Jones and Bartlett.

Yokota M and Tanimoto O (1967) Effects of diffusion on energy transfer by resonance. *Journal of the Physics Society of Japan* 22: 779–784.

TOMOGRAPHY

Contents

Optical Coherence Tomography
Tomography and Optical Imaging

Optical Coherence Tomography

S A Boppart, University of Illinois, Urbana, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical coherence tomography (OCT) is an emerging imaging technique for a wide range of biological, medical, and material investigations. OCT was initially developed for imaging biological tissue because it permits the imaging of tissue microstructure *in situ*, yielding micron-scale image resolution without the need for excision of a specimen and tissue processing. OCT is analogous to ultrasound B-mode imaging except that it uses low-coherence light rather than sound and performs cross-sectional imaging by measuring the backscattered intensity of light from structures in tissue. The OCT image is a gray-scale or false-color two-dimensional representation of backscattered light intensity in a cross-sectional plane. The OCT image represents the differential backscattering contrast between different tissue types on a micron scale. Because OCT performs imaging using light, it has a one to two order-of-magnitude higher spatial resolution than ultrasound and does not require contact with the specimen or sample.

OCT was originally developed and demonstrated in ophthalmology for high-resolution tomographic imaging of the retina and anterior eye. Because the eye is transparent and is optically accessible, it is well suited for diagnostic OCT imaging. OCT is promising for the diagnosis of retinal disease because it can provide images of retinal pathology with 10 micron resolution, almost one order-of-magnitude higher than previously possible using ultrasound. Recently, OCT has been applied for imaging in a wide range of nontransparent tissues. In tissues other than the eye, the imaging depth is limited by optical attenuation due to scattering and absorption. Ophthalmic imaging is typically performed at 800 nm wavelengths. However, because optical scattering decreases with increasing wavelength, OCT imaging in nontransparent tissues is possible using longer near-infrared wavelengths. In most tissues, imaging depths of 2–3 mm can be achieved using a system detection sensitivity of 100 to 110 dB. Imaging studies have been performed in a wide range of biological, medical, and surgical specialties including developmental biology, cardiology, gastroenterology, urology, and neurosurgery. High-resolution OCT using short coherence length, short pulsed light sources has also been demonstrated and axial resolutions less than 2 μm have been achieved. High-speed real-time OCT at image acquisition rates of 4 to 8 frames per second have also been demonstrated. OCT has been extended to perform Doppler imaging of blood flow and birefringence imaging to investigate tissue injury. Different imaging delivery systems including transverse imaging catheters and forward imaging devices have been developed to enable internal body OCT imaging. Most recently, an OCT catheter has been combined with endoscope-based delivery to perform *in vivo* imaging in animal models and human patients.

This chapter will provide an overview of the OCT technology, beginning with the basic principles of operation. Technological advancements over the last decade have enabled OCT to transition from laboratory- to clinical-based imaging. These enabling advancements will be discussed, followed by representative applications in biology, medicine, surgery, and materials.

Principles of Operation

OCT is based on optical ranging, the high-resolution, high dynamic range detection of backscattered light. In contrast to ultrasound, the velocity of light is extremely high. Therefore, the echo time delay of reflected light cannot be measured directly and interferometric detection techniques must be used. One method for measuring echo time delay is to use low coherence interferometry or optical coherence

domain reflectometry. Low-coherence interferometry was first developed for measuring reflections in fiber optics and optoelectronic devices.

The echo time delay of reflected light is measured by using a Michelson-type interferometer (Figure 1).

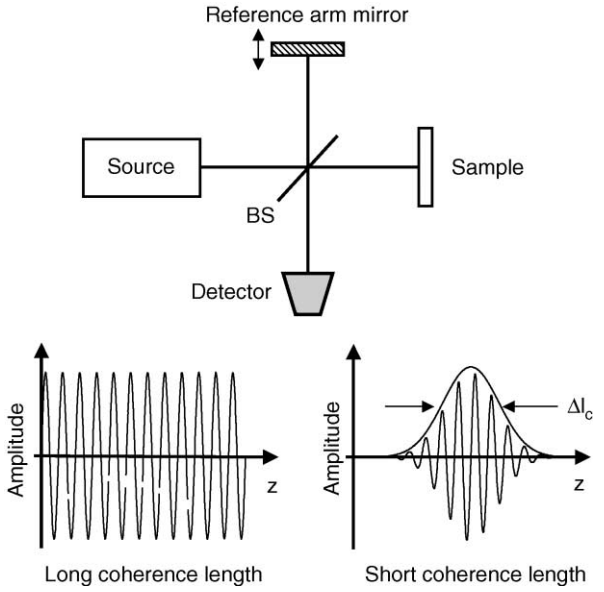


Figure 1 Schematic illustrating the concept of low-coherence interferometry. Using a short coherence-length light source and a Michelson-type interferometer, interference fringes are observed only when the path lengths of the two interferometer arms are matched to within the coherence length of the optical source. Abbreviations: BS, beam splitter; Δl_c, coherence length.

The light reflected from the specimen or sample is interfered with light that is reflected from a reference path of known path length. Interference of the light reflected from the sample arm and reference arm of the interferometer can occur only when the optical path lengths of the two arms match to within the coherence length of the optical source. As the reference arm optical path length is scanned, different echo delays of backscattered light from within the sample are measured. The interference signal is detected at the output port of the interferometer, electronically band-pass filtered, demodulated, digitized, and stored on a computer. The position of the incident beam on the specimen is typically scanned in the transverse direction and multiple axial measurements are performed. This generates a two-dimensional data array that represents the optical backscattering through a cross-sectional plane in the specimen (Figure 2). The logarithm of the backscatter intensity is then mapped to false-color or gray-scale and displayed as an OCT image. The interferometer in an OCT instrument can be implemented using a fiber optic coupler and beam-scanning can be performed with small mechanical galvanometers in order to yield a compact and robust system (Figure 3).

In contrast to conventional microscopy, the axial resolution in OCT images is determined by the coherence length of the light source. The axial point spread function of the OCT measurement as defined by the signal detected at the output of the

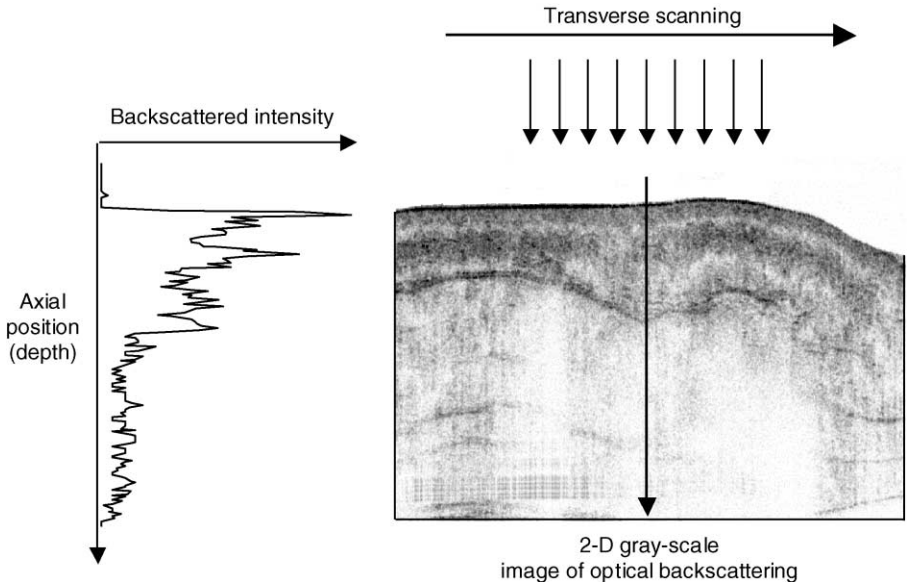


Figure 2 OCT image generation. The OCT image is typically acquired by performing axial measurements of optical backscatter at different transverse positions on the specimen and displaying the resulting two-dimensional data set as a gray-scale (or false-color) image.

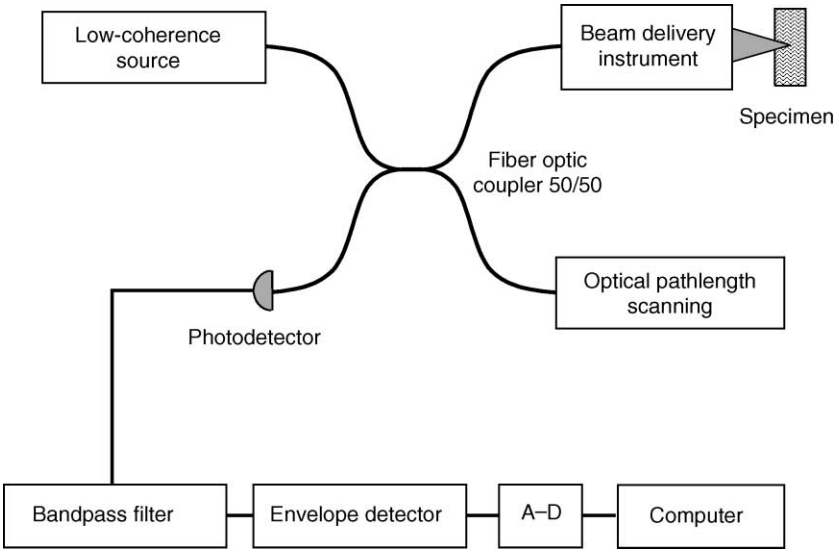


Figure 3 Schematic representation of an OCT system implemented using fiber optics. The Michelson interferometer is implemented using a fiber-optic coupler (beam splitter). The sample arm contains a beam-delivery instrument and the reference arm contains a mechanism for rapidly scanning the optical pathlength. The output of the interferometer is detected with a photodiode, filtered, demodulated, analog-to-digital (A–D) converted, and processed by a computer.

interferometer is the electric-field autocorrelation of the source. The coherence length of the light is the spatial width of the field autocorrelation and the envelope of the field autocorrelation is equivalent to the Fourier transform of its power spectrum. Thus, the width of the autocorrelation function, or the axial resolution, is inversely proportional to the width of the power spectrum. For a source with a Gaussian spectral distribution, the axial resolution Δz is given:

$$\Delta z = \frac{2\ln 2}{\pi} \cdot \frac{\lambda^2}{\Delta\lambda}$$

where Δz and $\Delta\lambda$ are the full-widths-at-half-maximum of the autocorrelation function and power spectrum respectively and λ is the source central wavelength. Figure 4 illustrates the dependence of the coherence length (axial resolution) on the bandwidth of the optical source. To achieve high axial resolution requires broad bandwidth optical sources. Resolution is also improved by using shorter wavelengths, however, shorter wavelengths are scattered and absorbed more in biological tissue.

The transverse resolution in an OCT imaging system is determined by the focused spot size in analogy with conventional microscopy and is given by:

$$\Delta x = \frac{4\lambda}{\pi} \cdot \frac{f}{d}$$

Where d is the spot size on the objective lens and f is its focal length. High transverse resolution can be obtained by using a large numerical aperture and

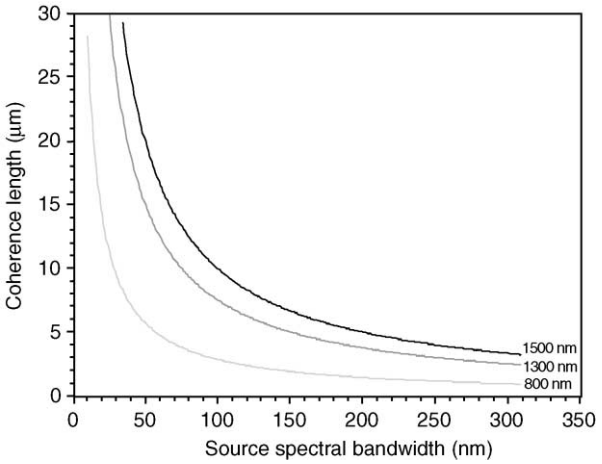


Figure 4 Dependence of coherence length (axial resolution) on the bandwidth of the optical source. Curves are shown for imaging wavelengths of 800 nm, 1300 nm, and 1500 nm. To achieve micron-scale resolution imaging, broad spectral bandwidths are necessary.

focusing the beam to a small spot size. The transverse resolution is also related to the depth of focus or the confocal parameter $2z_R$ (two times the Raleigh range).

$$2z_R = \frac{\pi\Delta x^2}{2\lambda}$$

Thus, increasing the transverse resolution results in a reduced depth of field. Typically, the confocal parameter or depth of focus is chosen to match the desired depth of imaging. Increased resolution may

also be obtained by using a higher numerical aperture lens or objective and spatially tracking the focus through the specimen.

Finally, the detection signal-to-noise is given by the optical power backscattered from the sample divided by the noise equivalent bandwidth:

$$\mathrm{SNR} = 10\log\left(\frac{\eta}{\hbar\omega}\frac{P_{\mathrm{SAM}}}{\mathrm{NEB}}\right)$$

Depending upon the desired signal to noise performance, incident powers of 5–10 mW are typically required for OCT imaging of 250–500 square pixel images at several frames per second. If lower data acquisition speeds or signal-to-noise can be tolerated, power requirements can be reduced accordingly.

Technological Developments

Since the inception of OCT in the early 1990s, there has been rapid technological developments aimed at improving the imaging resolution, acquisition rate, and methods for beam delivery to the tissue or sample. Investigators have also explored other imaging methods using the principles of OCT to extract information from the tissue or sample. Some of these methods have included acquiring optical Doppler signals from moving scatterers or structures, obtaining images based on the polarization state of the returned light, and extracting spectroscopic information based on the local absorption or scattering properties of the tissue.

Optical Sources / High-Resolution Imaging

The majority of OCT imaging systems to date have used superluminescent diodes (SLDs) as low coherence light sources. SLDs are manufactured with a similar structure to that of laser diodes, but commonly end facets are angle-cleaved to suppress lasing. These device structures do not support oscillation modes and generate output based on amplified spontaneous emission, resulting in emission spectra that are broader than laser diodes. SLDs are commercially available at a range of wavelengths including 800 nm, 1.3 μm, and 1.5 μm and are attractive because they are compact, have high efficiency, and low noise. However, output powers are typically limited to only a few milliwatts which limits fast real-time acquisition rates, and the available bandwidths are relatively narrow, permitting imaging with 10–15 micron resolution. Recent advances in short-pulse solid-state laser technology make these sources attractive for OCT imaging in research applications. Femtosecond solid-state lasers can generate tunable, low-coherence light at powers sufficient to permit high-speed OCT imaging. Short pulse generation has been achieved across the full wavelength range in titanium:sapphire ($Ti:Al_2O_3$) from 0.7 μm to 1.1 μm and over more limited tuning ranges near 1.3 μm and 1.5 μm in chromium: forsterite ($Cr^{4+}:Mg_2SiO_4$) and chromium:yttrium-aluminum-garnet (Cr^{4+}:YAG) lasers, respectively. OCT imaging with resolutions of 1 μm and 5 μm has been demonstrated at 800 nm and 1.3 μm respectively using $Ti:Al_2O_3$ and $Cr^{4+}:Mg_2SiO_4$ sources. More compact and convenient sources such as superluminescent fiber sources, are currently under investigation. The titanium:sapphire laser technology is routinely used in multi-photon microscopy applications for its high peak intensities to enable multi-photon absorption and subsequent emission of fluorescence from exogenous fluorescent contrast agents. Combined OCT and multi-photon microscopy has been used to provide complementary image data using a single optical source.

Fast Scanning

The short-pulse solid-state laser technology not only provides broad spectral bandwidths for high-resolution OCT imaging, but also higher output powers to enable fast real-time OCT imaging. Higher incident powers are required to maintain equivalent signal-to-noise ratios when scanning at a faster rate. Linearly translating a reference arm mirror is problematic at high rates and provides axial scan frequencies of approximately 100 Hz, depending on the mirror size and the translating galvanometer. Several investigators have utilized rotating glass cubes, piezoelectric modulators, and multi-pass optical cavities to increase axial scan rates while maintaining scan ranges of 1–2 mm. An optical delay based on the principles used in femtosecond pulse shaping has been demonstrated for OCT. This delay line spectrally disperses the reference arm beam with a grating. The dispersed beam is then focused by a lens on to a rotating mirror mounted on a galvanometer. The mirror, located in the Fourier-transform plane of the lens, imparts a wavelength-dependent phase shift on the light. Subsequently, when re-coupled back into the interferometer, this phase-shift is equivalent to a time-delay in the time domain. The use of high-speed resonant galvanometers has permitted axial scan rates as high as 8 kHz over scan ranges of several millimeters. Depending on the image pixel size (number of axial scans within each image), this scan rate can provide video-rate OCT imaging (30 frames per second). An example of high-speed *functional* OCT imaging is shown in Figure 5. Images of a

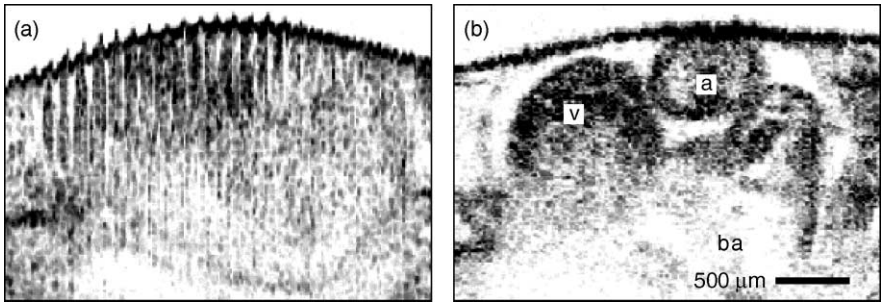


Figure 5 Comparison of slow and fast OCT imaging of a beating *Xenopus* (African frog) tadpole heart. The OCT image in (a) was acquired in 30 seconds and shows oscillations and blurring from the beating heart. The image on the right (b) was acquired in 250 ms (4 frames per second), fast enough to capture the anatomy of the heart with minimal motion artifacts. Abbreviations: a, atrium; ba, bulbous arteriosus; v, ventricle. (Images reprinted with permission from Boppart SA, Tearney GJ, Bouma BE, *et al.* (1997) Noninvasive assessment of the developing *Xenopus* cardiovascular system using optical coherence tomography. *Proceedings of the National Academy of Sciences* 94: 4256–4261.)

beating *Xenopus laevis* (African frog) tadpole were compared at slow (30 s per image) and at fast (250 ms per image) acquisition rates. Artifacts due to cardiac motion were present at slow acquisition rates, but were minimized with acquisition rates of 4 frames per second. Functional cardiac parameters could be measured and high-speed processes such as chamber and valve function could be visualized in real-time.

Depth-priority scanning, as described above, is performed by rapidly varying the optical delay in the reference arm and collecting a single axial scan before translating the beam laterally and repeating this depth-scanning. An alternative method for generating OCT images is with transverse-priority scanning. This is equivalent to optical sectioning in confocal and multi-photon microscopy. Three-dimensional OCT imaging using transverse-priority scanning can be obtained by stepping the position of the reference arm mirror after each *en face* image is acquired. The OCT images produced from transverse-priority scanning can be correlated with confocal or multi-photon microscopy images, which are acquired with the same scanning method. This method can also utilize higher numerical aperture objective lenses to provide high transverse resolutions since a large depth of focus is not needed as in depth-priority OCT scanning. The combination of OCT with high-numerical aperture objectives has been termed optical coherence microscopy.

Doppler Optical Coherence Tomography

Conventional OCT detects the amplitude of the optical backscatter versus position within the sample. Doppler OCT is a technique which has been applied to measuring fluid flow within small capillary tubes and *in vivo* vessels in the skin, retina, and heart of small animal models and humans. This technique is based on digitally sampling the interference fringes that are produced from a moving scatterer within the sample or specimen and determining the Doppler frequency shift in the signal due to the moving scatterer. Figure 6 illustrates two- and three-dimensional optical Doppler data for fluid flow through a cylindrical tube. A two-dimensional cross-sectional profile of fluid flow through a silicon tube with a 600 μm inner-diameter is plotted in Figure 6. The acquired data (points) closely match the predicted profile (line) for these tube dimensions. Below the plot, one method of obtaining a three-dimensional fluid flow profile is illustrated. While imaging the tube in cross-section, the fluid flow velocity was altered with a perfusion pump. The detection filter bandwidth of the OCT electronics served a windowing function, detecting only flow velocities with frequencies within the detection bandwidth. For an unknown flow profile, the center frequency of the detection filter can be shifted to map the flow velocities present within the sample.

Beam Delivery

The OCT imaging technology is modular in design and a variety of optical instruments can be used to deliver the OCT beam to the tissue or sample. Because OCT is fiber-optic based, single optical fibers can be used to deliver the OCT beam and collect the reflected light. The OCT technology can readily be integrated into existing optical instruments such as research and surgical microscopes, ophthalmic slit-lamp biomicroscopes, and hand-held imaging probes.

Imaging penetration is determined by the optical absorption and scattering properties of the tissue or specimen. The imaging penetration for OCT ranges from tens of millimeters for transparent tissues such as the eye to less than 3 mm in highly-scattering tissues such as skin. To image highly-scattering tissues deep within the body, novel beam-delivery

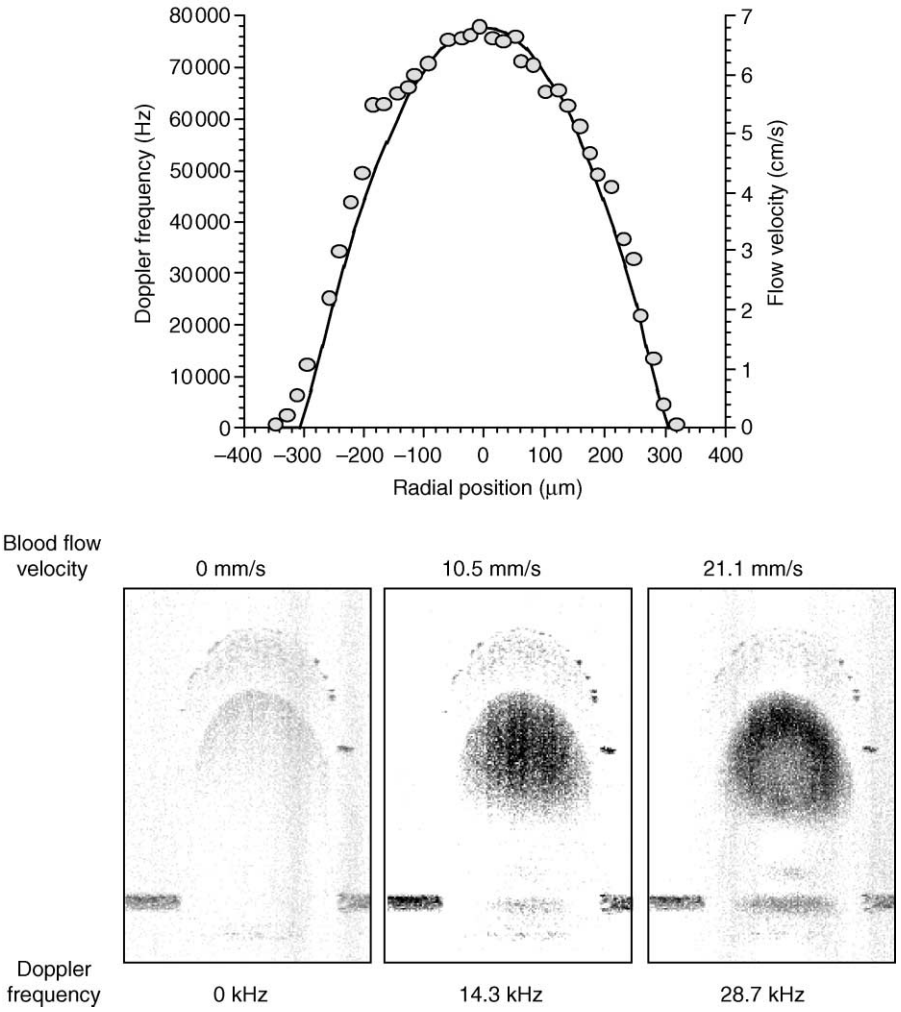


Figure 6 Doppler OCT. A two-dimensional plot of the fluid-flow velocity profile of flowing blood through a 600 μm inner-diameter tube is shown. Data points acquired using OCT closely follow the theoretical laminar flow profile (line). Two-dimensional Doppler OCT images of blood flow through the tube are shown below the plot. As fluid flow increases, only fluid-flow velocities within the bandwidth of the detection electronics are detected. This represents one method by which 3-D fluid flow can be imaged.

instruments have been developed to relay the OCT beam to the site of the tissue to be imaged. An OCT catheter has been developed for insertion into biological lumens such as the gastrointestinal tract. Used in conjunction with endoscopy, the 1 mm diameter catheter can be inserted through the working channel of the endoscope for simultaneous OCT and video imaging. Minimally-invasive surgical procedures utilize laparoscopes, which are long, thin, rigid optical instruments that permit video-based imaging within the abdominal cavity. Laparoscopic OCT imaging has been demonstrated by passing the OCT beam through the optical elements of a laparoscope. Deep solid-tissue imaging is possible with the use of fiber-needle probes. Small (400 μm diameter) needles housing a single optical fiber and micro-optic elements can be inserted into solid tissues and rotated to acquire OCT images. Recently, microfabricated micro-electro-optical-mechanical systems (MEOMS) technology has been used to miniaturize the OCT beam scan mechanism.

Applications

Developmental Biology

OCT has been demonstrated in the field of developmental biology as a method to perform high-resolution, high-speed imaging of developing morphology and function. Cellular-level imaging is possible, providing a non-invasive technique for visualizing cellular processes such as mitosis and migration. Imaging studies have been performed on several standard biological animal models commonly employed in developmental biology investigations including *Rana pipiens* (Leopard frog), *Xenopus laevis*

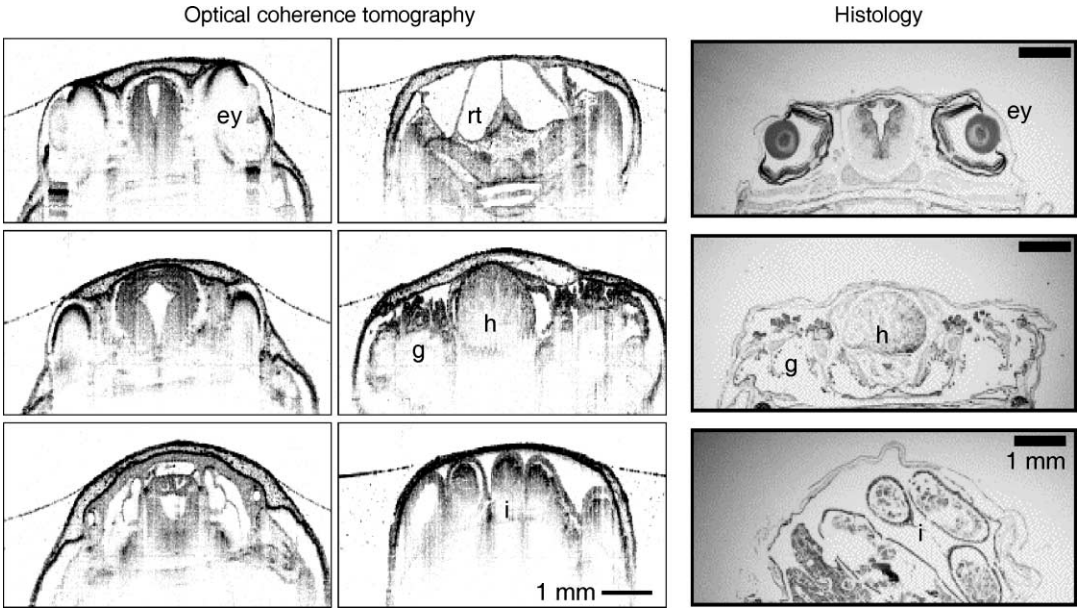


Figure 7 OCT images and corresponding histology of a Stage 49 (12 day) *Rana pipiens* (Leopard frog) tadpole. OCT images in the left and middle columns were acquired with the OCT beam incident from the dorsal and ventral sides of the specimen, respectively. The corresponding histology in the right column illustrates strong correlations between OCT images and the actual tissue morphology. Abbreviations: ey, eye; g, gills; h, heart; i, intestines; rt, respiratory tract. (Images reprinted with permission from Boppart SA, Brezinski ME, Bouma BE, *et al.* (1996) Investigation of developing embryonic morphology using optical coherence tomography. *Dev. Biol.* 177: 54–64.)

(African frog), and *Brachydanio rerio* (zebrafish) embryos and eggs, and the murine (mouse) model.

A series of cross-sectional images acquired *in vitro* from the dorsal and ventral sides of a Stage 49 (12 day) *Rana pipiens* (Leopard frog) tadpole is shown in Figure 7. Features of internal architectural morphology are clearly visible in the images. The image of the eye differentiates structures corresponding to the cornea, lens, and iris. Internal morphology not accessible in one orientation due to the specimen size or shadowing effects can be imaged by reorienting the specimen and scanning in the same cross sectional image plane. With the OCT beam incident on the ventral side, images of the respiratory tract, ventricle of the heart, internal gills, and gastrointestinal tract were acquired.

These images can be compared with corresponding histology (Figure 7). Histological images are acquired by euthanizing the specimen, immersing the specimen in a chemical fixative, and physically sectioning thin (2–5 micron-thick) slices using a microtome. The slices are placed on a microscope slide, selectively stained to highlight particular features, and viewed with light microscopy. The correlations between OCT and histology images are strong, suggesting that OCT images can accurately represent the *in vivo* specimen morphology. The potential exists to repeatedly image specimens to quantify organo- and morphogenesis throughout development. Technologies such as OCT are likely to become increasingly important in functional genomics, relating genetic features to the morphology and function in living specimens.

OCT images represent the optical backscatter intensity from regions within the tissue or sample. Because OCT relies on the inherent optical scattering changes to produce imaging contrast, no exogenous contrast agents or fluorophores are necessary. This permits long-term sequential imaging of development *in vivo* without loss of specimen viability. Repeated images of a developing zebrafish embryo within its egg beginning immediately after fertilization and up until hatching have been demonstrated without loss of specimen viability or without developmental abnormalities. In this example, the zebrafish egg and embryo were semi-transparent and the use of OCT significantly complemented observations made using light microscopy. By imaging subtle differences in backscattering intensity, interfacial structural layers millimeters deep within specimens can be clearly delineated.

Previous OCT images have characterized morphological features within biological specimens. These structures are static even though they may have been acquired from *in vivo* specimens. *In vivo* imaging in living specimens, particularly in larger organisms and

for medical diagnostic applications, must be performed at high speeds to eliminate motion artifacts within the images. Functional OCT imaging is the quantification of *in vivo* images which yield information characterizing the functional properties of the organ system or organism. High speed OCT permits both the positioning and manipulation of specimens as well as imaging in real time and is a powerful technology for functional imaging in developmental biology animal models.

Studies investigating normal and abnormal cardiac development have been frequently limited by an inability to access cardiovascular function within the intact organism. OCT has been demonstrated for the high-resolution assessment of structure and function in the developing *Xenopus laevis* (African frog) cardiovascular system (Figure 5). The morphology of the *in vivo* cardiac chambers is clearly delineated. Image acquisition rates are fast enough to capture the cardiac chambers in mid-cycle. With this capability, images can be acquired at various times during the cardiac cycle. These frames can be displayed in real-time to produce a movie illustrating the dynamic, functional behavior of the developing heart. OCT, unlike technologies such as computed tomography and magnetic resonance imaging, provides high-speed *in vivo* imaging, allowing quantitative dynamic activity, such as ventricular ejection fraction, to be assessed.

Cellular Imaging

Although previous studies have demonstrated *in vivo* OCT imaging of tissue morphology, most have imaged tissue at $\sim 10-15\ \mu m$ resolutions, which does not allow differentiation of cellular structure. The ability of OCT to identify the mitotic activity, the nuclear-to-cytoplasmic ratio, and the migration of cells has the potential to not only impact the fields of cell and developmental biology, but also impact medical and surgical disciplines for the early diagnostics of disease such as cancer.

The *Xenopus laevis* (African frog) tadpole has been used to demonstrate the feasibility of OCT for high-resolution *in vivo* cellular and subcellular imaging. Many of the cells in this common developmental biology animal model are rapidly dividing and migrating during the early growth stages of the tadpole, providing an opportunity to image dynamic cellular processes. Three-dimensional volumes of high-resolution OCT data have been acquired from these specimens throughout development. From this 3-D data set, cells undergoing division were identified and tracked in three dimensions. In a similar manner, 3-D data sets were acquired to track single melanocytes (neural crest cells) as they migrated through the living specimens. The ability of OCT to characterize cellular processes such as mitosis and migration not only are of interest in cell and developmental biology, but also have relevance for cancer diagnostics and tumor metastasis.

An example of cellular-level OCT imaging in these specimens is shown in Figure 8. This composite image (0.83×1 mm, 1800×1000 pixels) was acquired using a titanium:sapphire laser with a broad bandwidth (~ 260 nm). The axial and transverse resolution in this image is 1 μm and 5 μm, respectively (see Figure 4). Because the high transverse resolution reduced the depth-of-focus to 49 μm, separate OCT images were first acquired with the focus at different depths within the specimen. These images were then assembled to produce the composite image shown in Figure 8. This type of image construction is similar to C-mode ultrasound. Cellular features including cell membranes, nuclei, and nuclear morphology are clearly observed.

Medicine – Imaging Barrett's Esophagus

OCT performs *in situ* imaging and has the potential to be used as a screening technique of early pathological changes in patients. OCT imaging has the potential to be a general diagnostic in many organ systems,

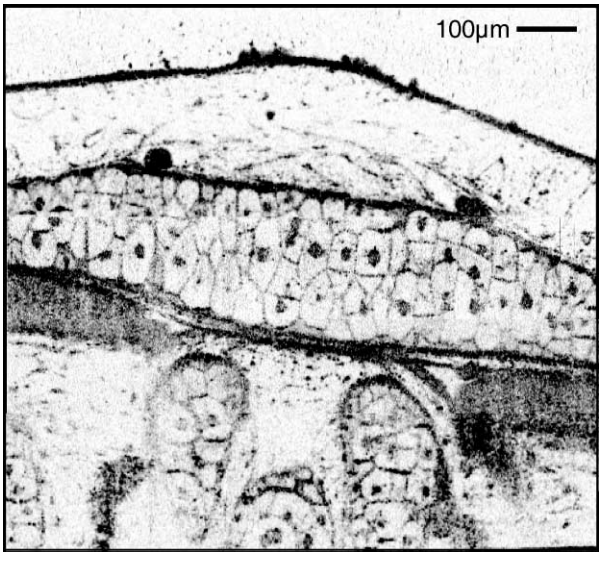


Figure 8 Cellular-level OCT imaging using a broad bandwidth titanium:sapphire laser. Axial and transverse image resolutions are 1 μm and 5 μm, respectively, enabling the visualization of cell membranes, nuclear morphology, and sub-cellular organelles. Cellular processes such as mitosis and migration can be visualized in real-time in living specimens. (Image reprinted with permission from Drexler W, Morgner U, Kartner FX, *et al.* (1999) *In vivo* ultrahigh resolution optical coherence tomography. *Optics Letters* 24: 1221–1223.)

particularly where current methods of screening by excisional biopsy are limited. Barrett's esophagus is a condition where the cells of the distal esophagus undergo a metaplastic change, resembling cells of the lower gastrointestinal tract. This condition is believed to be caused by chronic gastroesophageal reflux. Several studies have demonstrated that Barrett's esophagus is associated with a 30–125 times increased risk of developing adenocarcinoma. For this reason, endoscopic surveillance of Barrett's epithelium every 12–18 months is recommended. Endoscopic screening currently involves random four-quadrant biopsies every 1–2 cm along the length of suspect mucosa. However, excisional biopsy is prone to sampling errors and small foci of carcinoma or dysplasia may be missed. Because of the imprecision and high cost associated with screening, new methods are being developed to assess patients at increased risk. Endoscopic ultrasound catheters have been used for imaging the gastrointestinal tract with 50–100 μm resolution. However, these resolutions are insufficient to resolve early epithelial changes that occur in Barrett's esophagus and the pre-malignant changes that lead to adenocarcinoma of the esophagus.

The ability of OCT to differentiate normal and pathologic tissue is a central question being addressed by many research groups. Comparisons with histology show good correlations between OCT images and histological findings. The image resolution of conventional OCT (10–15 μm) is sufficient to differentiate architectural but not cellular morphology. Still, endoscopic OCT resolution can differentiate normal from Barrett's epithelium in real-time based on differences in epithelial architecture. Crypt- and gland-like structures that disrupt the relatively uniform layers of squamous epithelium can be readily identified, enabling differentiation between normal and Barrett's epithelium (Figure 9). The ability to differentiate normal from Barrett's epithelium suggests that the OCT technology could be used for screening applications.

Oncology – Identifying Tumors and Tumor Margins

OCT has been used to differentiate between the morphological structure of normal and cancerous tissue for a wide-range of tumors. The use of OCT to identify tumors and tumor margins *in situ* will represent a significant advancement for medical or image-guided surgical applications. OCT has been demonstrated for the detection of brain tumors and their margins with normal brain parenchyma, suggesting a role for guiding surgical resection. A hand-held surgical imaging probe was constructed for this application. The compact and portable probe permits OCT imaging within the surgical field while

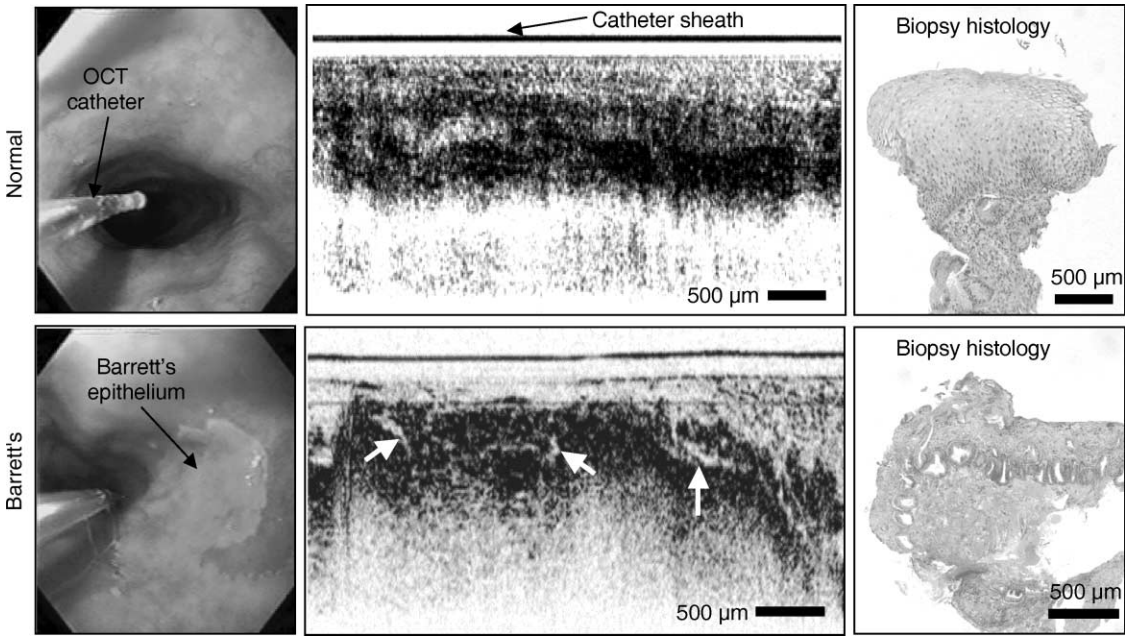


Figure 9 Endoscopic OCT imaging in humans. An OCT catheter was inserted down the working channel of an endoscope. Guided by video imaging, regions of normal and Barrett's esophagus were imaged. Corresponding biopsy histology is shown. OCT can identify Barrett's esophagus by the presence of gland- and crypt-like structures (arrows). (Images reprinted with permission from Li XD, Boppart SA, Van Dam J, *et al.* (2000) Optical coherence tomography: advanced technology for the endoscopic imaging of Barrett's esophagus. *Endoscopy* 32: 921–930.)

the OCT instrument can be remotely located in the surgical suite.

Figure 10 shows a specimen of outer human cerebral cortex with metastatic melanoma. The OCT images in Figure 10a,b were acquired through the tumor. These original images were threshold segmented to identify regions of high backscatter within the tumor. The original images were then overlaid with the segmented data and are shown in Figure 10c,d. The OCT images show increased optical backscattering in the region of the larger tumor (arrows). Smaller tumor lesions also appear within the image (arrows). A shadowing effect is observed below each tumor site due to the increased optical backscatter and the subsequent loss of optical power penetrating beneath the tumor. In Figure 10a,c, the boundary of the tumor can be identified. In Figure 10b,d, the tumor is identified below the surface of normal cortex. The histology in Figure 10e,f confirms the presence and relative size of the tumor.

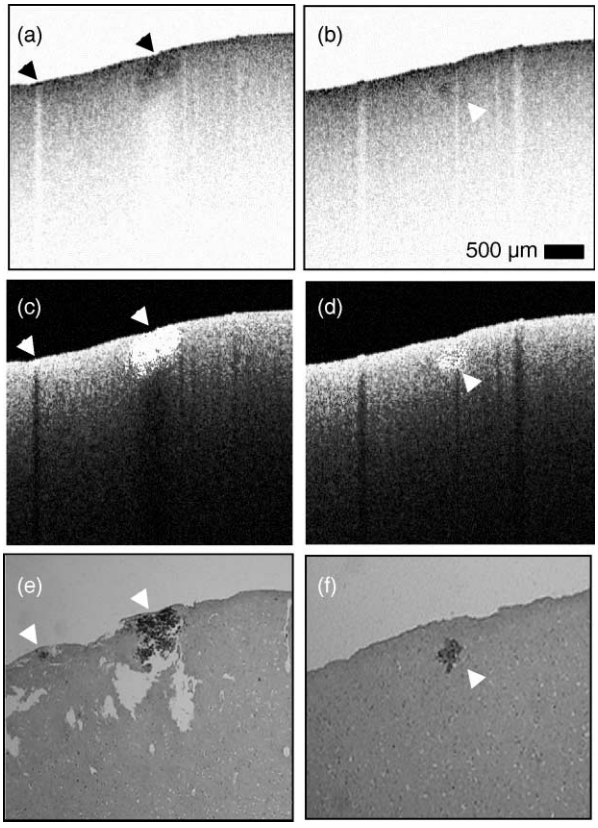


Figure 10 Cancer diagnostics. OCT imaging was used to image metastatic melanoma in an *ex vivo* specimen of human cortex (brain). Raw OCT images (a,b) were threshold-segmented (c,d) to highlight regions of increased optical backscatter and to identify tumor margins. OCT images correlate well with histological findings (e,f). (Images reprinted from Boppart SA, Brezinski ME, Pitris C and Fujimoto JG (1998) Optical coherence tomography for neurosurgical imaging of human intracortical melanoma. *Neurosurgery* 43: 834–841, with permission from Lippincott Williams and Wilkins.)

The image resolutions used to acquire the images in Figure 10 were as high as 16 μm, higher than any current ultrasound, CT, or MRI intraoperative imaging technique. This allowed the tumor-cortex interface and the extent of tumor below the surface to be defined with high resolution. At higher imaging resolutions, it may be possible to image individual tumor cells which have migrated away from the central tumor. OCT represents a new high-resolution optical imaging technology that has the potential for identifying tumors and tumor margins on the micron scale and in real-time. OCT offers imaging performance not achievable with current imaging modalities and may contribute significantly toward the surgical resection of neoplasms.

Image-Guided Surgery

The repair of vessels and nerves is necessary to restore function following traumatic injury. Although the repair of these sensitive structures is performed with the aid of surgical microscopes and loupes to magnify the surgical field, surgeons are limited to the *en face* view that they provide. A technique capable of subsurface, three-dimensional, micron-scale imaging in real-time would permit the intraoperative monitoring of microsurgical procedures. The capabilities of OCT for the intraoperative assessment of microsurgical procedures have been demonstrated. High-speed OCT imaging was integrated with a surgical microscope to performed micron-scale three-dimensional imaging on microsurgical specimens.

OCT has been used to image *in vitro* peripheral nerves and identify individual fascicles. Longitudinal tracking of the spatial orientation of rabbit peripheral nerve fascicles is demonstrated in Figure 11. Representative cross-sectional images of the peripheral nerve are shown in Figure 11a–d. For each slice, one fascicle was manually segmented, colored white, and tracked through the acquired volume of data. Forty images at 100 μm spacing were assembled for the 3-D projections shown in Figure 11e,f. The horizontally- and vertically-rotated projections of the peripheral nerve dramatically reveal the twisted path of the segmented fascicle along the longitudinal axis of the nerve. In addition, a branch in an unsegmented fascicle is observed in Figure 11f. The use of OCT to acquire multiple cross-sectional images and three-dimensionally reconstruct the peripheral nerve offers the opportunity to determine the relative diameters of individual fascicles and to longitudinally track their spatial orientation. OCT may also be useful at identifying and grading the degree of injury in nerves during surgical repair. These results have shown how 2-D OCT images and 3-D OCT projections can

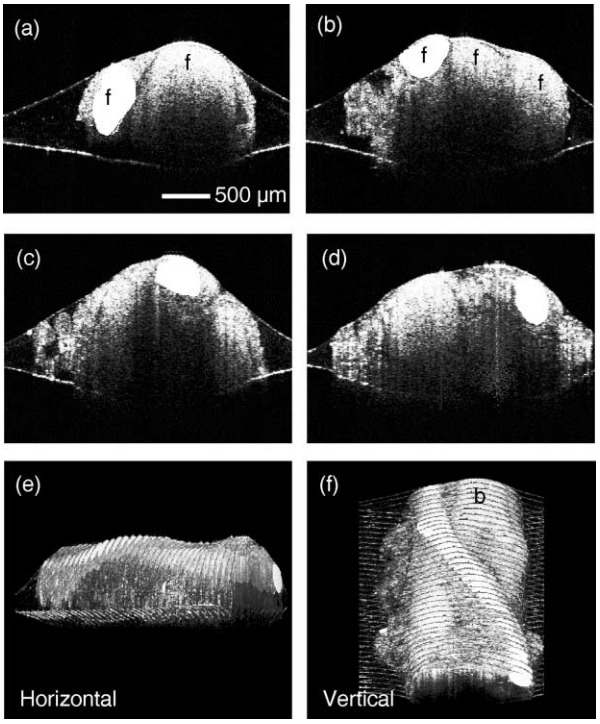


Figure 11 OCT image-guided microsurgery. Multiple cross-sectional OCT images were acquired through a resected rabbit peripheral nerve (a–d). A single nerve fascicle (f) was identified and segmented longitudinally along the length of the nerve. Three-dimensional reconstruction projections of the nerve in (e,f) reveal the tortuous path of the segmented fascicle as well as reveal a bifurcation (b). Intraoperative OCT can guide surgical repair and potentially identify regions of damaged tissue. (Images reprinted from Boppart SA, Bouma BE, Pitris C, *et al.* (1998) Intraoperative assessment of microsurgery with three-dimensional optical coherence tomography. *Radiology* 208: 81–86, with permission from the Radiology Society of North America.)

provide diagnostic feedback to assess microsurgical anastomoses. This previously unavailable diagnostic ability offers the potential to directly impact and improve patient outcome by incorporating high-speed, high-resolution intraoperative image-guidance during microsurgical procedures.

Surgical intervention requires visualization to identify tissue morphology, precision to avoid sensitive tissue structures, and continuous feedback to monitor the extent of the intervention. OCT may provide the technological advancements to improve the operative procedure. The feasibility of OCT to perform image-guided surgical intervention has been investigated. OCT has been used to monitor laser ablation therapy in real-time and may enable more precise control of laser delivery. An argon-laser ablation sequence of fresh *ex vivo* rat rectus abdominis muscle is shown in Figure 12. A pair of blood-filled vessels were located 1.5 mm below the tissue surface and centered within the OCT image. Because blood has a higher absorption coefficient than muscle at the 514 nm argon laser wavelength, coagulation and thermal heating of the blood occurred prior to thermal damage to the overlying tissue. The accompanying histology reveals both the thermally-damaged overlying tissue and coagulated blood present within the vessel pair.

These examples demonstrate the use of OCT for guiding and monitoring surgical intervention. The laser is only a representative interventional surgical technique for a wide range of instruments and techniques including scalpels, electrosurgery, radio-frequency, microwaves, and ultrasound ablation. OCT imaging was performed at 8 frames per second, fast enough to capture dynamic changes in the optical properties of the tissue during thermal ablation. These image sequences provided interesting insight into ablation mechanisms for a variety of tissue types. OCT can monitor the extent of thermal injury below the surface of the tissue by imaging the changes in optical backscatter. OCT imaging can therefore provide empiric information for dosimetry to minimize the extent of collateral injury. The use of OCT for guiding surgical interventions has the potential to improve intraoperative monitoring and more effectively control interventional procedures.

Materials

While the majority of OCT applications have been in the fields of biology and medicine. OCT has also been demonstrated in the non-biological areas of materials investigation, optical data storage, and microfluidic devices. The highly-scattering or reflecting optical properties of many materials prohibits deep imaging penetration using OCT. Many material defects, however, originate near the surface or at interfacial boundaries, making the use of OCT a possibility for inspection and quality control. OCT has been used to identify subsurface defects in ceramics and polymer composites. The optical ranging capabilities of OCT through scattering materials has been utilized for increasing the data storage capacity by assembling multiple layers of optically-accessible data.

The advancement of microfabrication techniques has led to increasingly complex microfluidic and bioMEM (biological micro-electro-mechanical) systems. Microstructures within microfluidic systems range from 10–1000 μm, within both the imaging depth and resolution of OCT. In addition, microfluidic systems are typically fabricated from transparent or semi-transparent substrates, facilitating imaging penetration to deeper three-dimensional features.

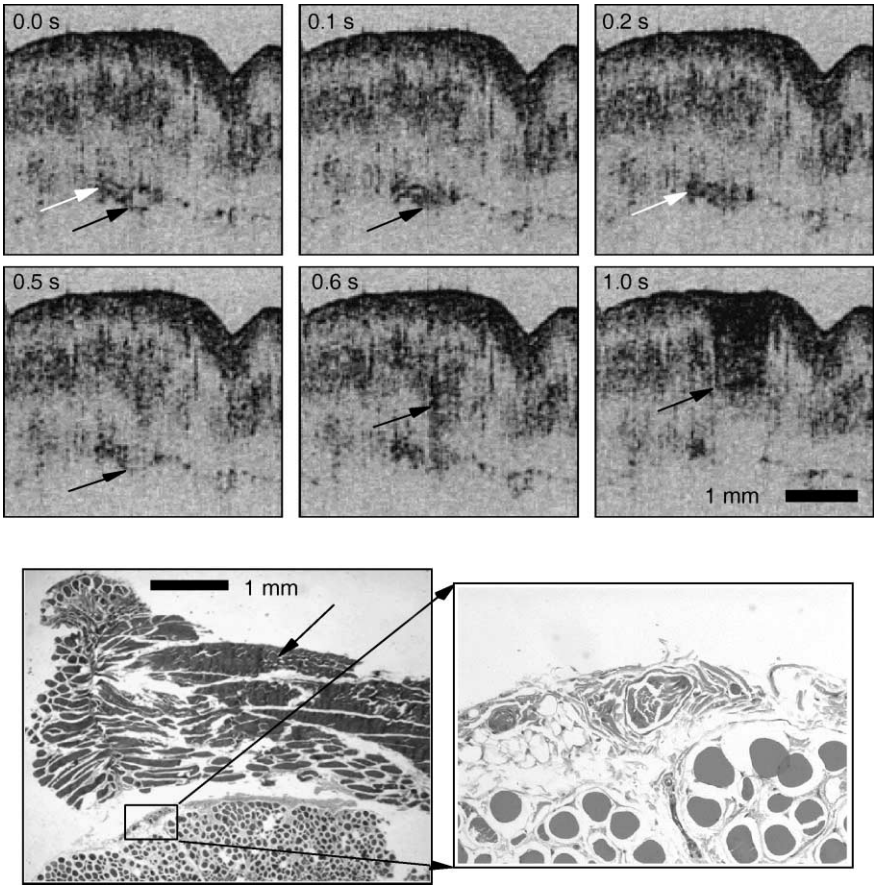


Figure 12 OCT image-guided laser ablation. High-speed OCT image sequence of argon-laser ablation of blood vessels in *ex vivo* rat muscle tissue. Thermal heating within the blood vessels (arrows, 0.0–0.5 s) induces coagulation prior to injury to superficial muscle tissue (0.6–1.0 s). Accompanying histology shows coagulated blood within the vessels.

OCT is not only capable of imaging three-dimensional microstructures at micron-scale resolutions, but is also capable of obtaining dynamic functional data from microfluidic systems. Microfluidic devices made from polymeric materials are attractive since they provide flexibility in the design of the channels, can be manufactured in large volumes, and are mostly bio-compatible. A three-dimensional imaging technique, such as OCT, will be instrumental in the analysis of these devices and the monitoring of device performance.

Conclusions

The capabilities of OCT offer a unique and informative means of imaging biological specimens and non-biological samples. The non-contact nature of OCT and the use of low-power near-infrared radiation for imaging causes few harmful effects on living cells or damage to materials. OCT imaging does not require the addition of fluorophores, dyes, or stains in order to improve contrast in images. Instead, OCT relies on the inherent optical contrast generated from variations in optical scattering and index of refraction. These factors permit the use of OCT for extended imaging over the course of hours, days, or weeks. OCT permits the cross-sectional imaging of tissue and samples and enables *in vivo* structure to be visualized in opaque specimens or in specimens too large for high-resolution confocal or light microscopy.

Imaging at cellular and subcellular resolutions with OCT is an important area of ongoing research. The *Xenopus* (African frog) developmental animal model has been commonly used because its care and handling are relatively simple while allowing cells with a high mitotic index to be assessed. Many of the cells observed were as large as 100 μm in diameter, but ranged in size down to a few microns, below the resolution of most OCT systems. Imaging human cells *in vivo* remains a challenge since differentiated human cells are 10 to 20 microns in size, too small for most OCT systems. With further advances in OCT technology, improved discrimination and

imaging of more detailed structures should be possible. New laser sources at other wavelengths in the near-infrared can enhance tissue contrast as well as potentially provide functional information since tissue scattering and absorbance properties in specimens are wavelength dependent. Short coherence length short pulse laser sources have been used to achieve higher axial resolutions on the order of 1 to 3 microns. Unfortunately, unlike superluminescent diode source, these high-speed and high-resolution systems utilize femtosecond lasers that are relatively complex and costly. Developing compact and portable optical sources at near-infrared wavelengths, broad spectral bandwidths, and high output powers is an area of active research.

Optical coherence tomography provides high-resolution morphological, functional, and cellular information of biological, medical, and material specimens and samples. OCT represents a multifunctional investigative tool that not only complements many of the existing imaging technologies available today, but also is poised to become a major optical imaging modality.

List of Units and Nomenclature

$Cr^{4+}:Mg_2SiO_4$	chromium:forsterite
Cr^{4+}:YAG	chromium:yttrium-aluminum-garnet
CT	computed tomography
d	beam diameter
f	focal length
MRI	magnetic resonance imaging
NEB	noise equivalent bandwidth
OCT	optical coherence tomography
P_{SAM}	power from sample
SNR	signal-to-noise ratio
SLD	superluminescent diode
$Ti:Al_2O_3$	titanium:sapphire
z_R	Rayleigh range
3-D	three-dimensional
Δx	transverse resolution
$\Delta\lambda$	spectral bandwidth
Δz	axial resolution
η	quantum efficiency
λ	center wavelength

See also

Coherence: Coherence and Imaging; Overview. **Coherent Lightwave Systems**. **Interferometry:** Overview. **Fiber and Guided Wave Optics:** Optical Fiber Cables. **Imaging:** Imaging Through Scattering Media; Interferometric Imaging. **Terahertz Technology:** Coherent Terahertz Sources. **Tomography:** Tomography and Optical Imaging. **Microscopy:** Confocal Microscopy; Imaging multiple Photon Fluorescence Microscopy; Overview.

Further Reading

Boppart SA, Tearney GJ, Bouma BE, *et al.* (1997) Noninvasive assessment of the developing *Xenopus* cardiovascular system using optical coherence tomography. *Proceedings of the National Academy of Sciences* 94: 4256–4261.

Boppart SA, Bouma BE, Pitris C, *et al.* (1998) *In vivo* cellular optical coherence tomography imaging. *Nature Medicine* 4: 861–864.

Boppart SA, Deutsch TF and Rattner DW (1999) Optical imaging technology in minimally invasive surgery. *Surgical Endoscopy* 13: 718–722.

Boppart SA, Brezinski ME and Fujimoto JG (2000) Optical Coherence Tomography Imaging. In: Tuan RS and Lo CW (eds) *Methods in Molecular Biology, Vol. 135: Developmental Biology Protocols*, vol. 1, pp. 217–233. Totowa, NJ: Humana Press, Inc.

Bouma BE and Tearney GJ (eds) (2001) *The Optical Coherence Tomography Handbook*. New York: Marcel Dekker Inc.

Chen Z, Zhao Y, Srinivas SM, *et al.* (1999) Optical Doppler tomography. *IEEE Journal of Selected Topics in Quantum Electronics* 5(4): 1134–1142.

Fujimoto JG, Brezinski ME, Tearney GJ, *et al.* (1995) Biomedical imaging and optical biopsy using optical coherence tomography. *Nature Medicine* 1: 970–972.

Fujimoto JG, Pitris C, Boppart SA and Brezinski ME (2000) Optical coherence tomography: An emerging technology for biomedical imaging and optical biopsy. *Neoplasia* 2(1–2): 9–25.

Huang D, Swanson EA, Lin CP, *et al.* (1991) Optical coherence tomography. *Science* 254: 1178–1181.

Masters BR (ed.) (2001) *Selected Papers on Optical Low-Coherence Reflectometry & Tomography.* SPIE Milestone Series, Vol. MS 165, Thompson BJ (general ed.) Bellingham, WA: SPIE Optical Engineering Press.

Puliafito CA, Hee MR, Schuman JS and Fujimoto JG (1995) Optical Coherence Tomography of Ocular Diseases. Thorofare, NJ: Slack, Inc.

Schmitt M (1999) Optical coherence tomography (OCT): A review. *IEEE Journal of Selected Topics in Quantum Electronics* 5(4): 1205–1215.

Tadrous PJ (2000) Methods for imaging the structure and function of living tissues and cells: 1. Optical coherence tomography. *Journal of Pathology* 191(2): 115–119.

Tearney GJ, Brezinski ME, Bouma BE, *et al.* (1997) *In vivo* endoscopic optical biopsy with optical coherence tomography. *Science* 276: 2037–2039.

Welzel J (2001) Optical coherence tomography in dermatology: a review. *Skin Research and Technology* 7(1): 1–9.

Tomography and Optical Imaging

Z Chen, University of California, Irvine, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

'Tomos' means section, in Greek. A tomogram is a sectional image of the internal structure of a three-dimensional object. One of the challenges for tomographic imaging of biological tissue using visible or near-infrared light is tissue optical scattering. When photons propagate inside a tissue, they undergo multiple scattering. The mean free path of photons in the skin, for example, is 5 μm. If one focuses the light 1 mm below the skin surface, most photons undergo multiple scattering and never reach the target point. Less than one out of a million photons will be able to reach the target, undergo a single backscattering, and return to the detector (Figure 1). Most photons that are backscattered and reach the detector are multiple scattered photons that carry little information on the target point. The challenge for tomographic imaging of tissues is to separate the single backscattered photons, that carry information about the target point, from multiple scattered photons.

Several optical tomography techniques have been developed. Each uses different techniques to discriminate the multiple scattering photons from the minimally scattered photons for imaging constructions. The confocal microscope, for example, uses a spatial filtering technique to reject multiple scattering photons because multiple scattered photons are more spread out in their angular distribution than the minimally scattered photons. Time-of-flight imaging uses a time gating technique to discriminate diffuse photons because single backscattered photons arrive earlier than diffuse photons. Optical coherence tomography (OCT) uses coherent gating to discriminate single scattered photons from multiple scattered photons. These gating techniques provide good discrimination of minimally scattered photons from multiple scattered photons and can form high-resolution sectional images, albeit with a limited imaging depth. Diffuse photons can propagate deep into the tissue and these photons can also be used for optical tomography. If a modulated photon density wave is propagated inside the tissue, the scattering and absorption properties of the tissue will modulate the phase and amplitude of the photon density wave. However, the resolution of optical tomography, based on diffuse photons, is usually much lower than that of the tomography constructed from single or minimally scattered photons. Three of the most important parameters for optical tomography are contrast, spatial resolution, and image depth.

Most optical tomography uses absorption, scattering, and fluorescent properties of materials as contrast. OCT, for example, uses variations in tissue scattering as contrast for tomographic imaging. However, in many instances, and especially during the early stages of disease, the change in tissue scattering properties between normal and diseased tissues is small and difficult to measure. One of the great challenges for extending applications of OCT is to find more contrast mechanisms that can provide physiological information in addition to morphological structure. This article will focus on optical tomography technologies that use nonconventional contrast mechanisms to obtain tomographic images. These include optical Doppler tomography (ODT), optical polarization tomography, and optical second-harmonic tomography. These optical tomography technologies are extensions of OCT with contrast mechanisms that can provide physiological information, in addition to morphological structure. ODT, also known as Doppler OCT, combines the Doppler principle with OCT to obtain high-resolution tomographic images of tissue structure and blood flow simultaneously. Optical polarization tomography, also known as polarization sensitive OCT (PS-OCT), combines polarization sensitive

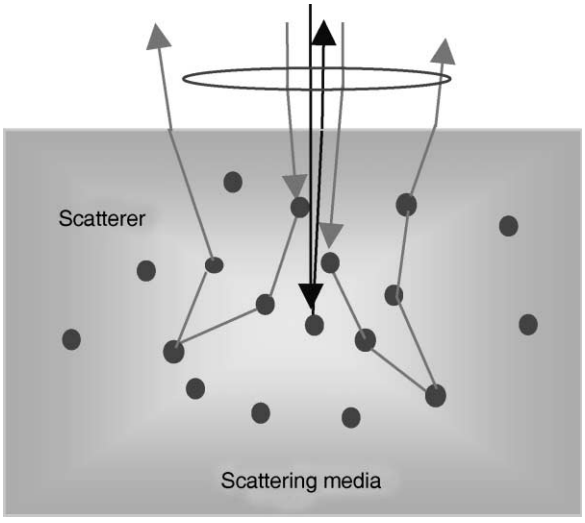


Figure 1 Optical imaging in scattering media.

detection with OCT, to determine tissue birefringence. Optical second-harmonic tomography, also known as second-harmonic optical coherence tomography (SH-OCT), combines second-harmonic generation (SHG) with coherence gating to obtain images with molecular contrast. These functional extensions of OCT provide clinically important information on tissue physiology, such as tissue blood perfusion, oxygen saturation, hemodynamics, and structural remodeling. Each provides several potential clinical applications, such as vasoactive drug screening, tissue viability and burn depth determination, tumor angiogenesis studies, tumor diagnosis, bleeding ulcer management, and ocular pathology evaluation.

Optical Doppler Tomography

ODT combines the Doppler principle with OCT to obtain high-resolution tomographic images of static and moving constituents in high scattering media. When light backscattered from a moving particle interferes with the reference beam, a Doppler frequency shift (f_D) occurs in the interference fringe:

$$f_D = \frac{1}{2\pi}(\mathbf{k}_s - \mathbf{k}_i)\cdot\mathbf{v} \qquad [1]$$

where $\mathbf{k}_i$ and $\mathbf{k}_s$ are wavevectors of incoming and scattered light, respectively, and $\mathbf{v}$ is the velocity vector of the moving particle (Figure 2). Since ODT measures the backscattered light, assuming the angle between flow and sampling beam is θ, the Doppler shift equation is simplified to:

$$f_D = \frac{2V\cos\theta}{\lambda_0} \qquad [2]$$

where λ_0 is the vacuum center wavelength of the light source.

The optical system of ODT is similar to that of OCT. The primary difference is in signal processing. Figure 3 illustrates an ODT instrument that uses a fiber optic Michelson interferometer with a broadband light source. Light from a broadband partial coherence source is coupled into a fiber interferometer by a 2 × 2 fiber coupler and then split equally into reference and target arms of the interferometer. Light backscattered from the turbid sample is coupled back into the fiber and forms interference fringes with the light reflected from the reference arm. A rapid-scanning optical delay line is used for group delay and axial scanning. Because this delay line can decouple the group delay from the phase delay, an electro-optical phase modulator is introduced to produce a stable carrier frequency. The interference fringe intensity signal is amplified, bandpass filtered, and digitized with a high-speed analog-to-digital converter. The signal processing is carried out at the same time as data are transferred to the computer, and real-time display can be accomplished with the use of a digital signal processing board.

To understand the signal processing of ODT, let us look at the fringe signal due to the moving particles.

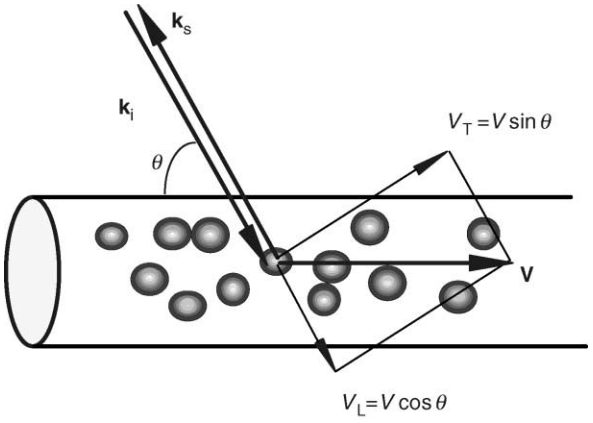


Figure 2 Schematic of flow direction and probe beam angle.

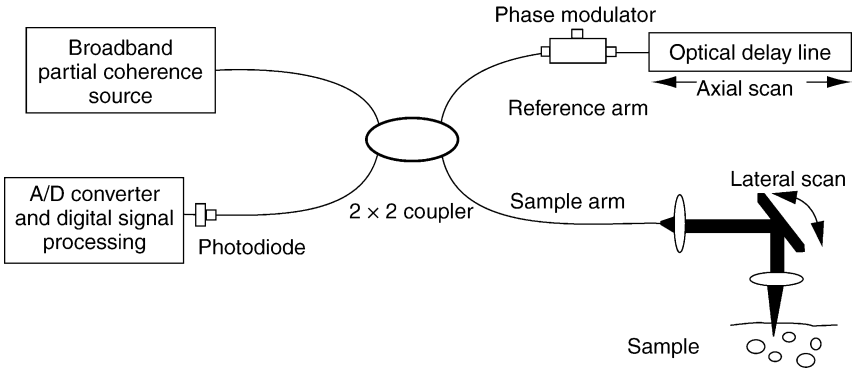


Figure 3 Schematic of ODT system consists of a fiber-based Michelson interferometer with a partially coherent light source.

If we denote $U(t)$ as a complex-valued analytic signal of a stochastic process representing the field amplitude emitted by a low coherent light source and $\bar{U}(\nu)$ as the corresponding spectral amplitude at optical frequency ν, the amplitude of a partially coherent source light coupled into the interferometer at time t, is written as a harmonic superposition:

$$U(t) = \int_0^\infty \bar{U}(\nu) e^{2\pi i \nu t} \, d\nu \quad [3]$$

Because the stochastic process of a partially coherent light source is stationary, the cross spectral density of $\bar{U}(\nu)$ satisfies:

$$\langle \bar{U}^*(\nu)\bar{U}(\nu')\rangle = S_0(\nu)\delta(\nu - \nu') \quad [4]$$

where $S_0(\nu)$ is the source power spectral density, and $\delta(\nu - \nu')$ is the Dirac delta function. Assuming that light couples equally into the reference arm and sample arm with spectral amplitude of $\bar{U}_0(\nu)$, the light coupled back to the detector from the reference, $\bar{U}_r(\nu)$, and sample, $\bar{U}_s(\nu)$, are:

$$\bar{U}_r(\nu) = e^{2\pi i \nu (2L_r + L_d)/c} K_r(\nu) e^{i\alpha_r} \bar{U}_0(\nu) \quad [5]$$

$$\bar{U}_s(\nu) = e^{2\pi i \nu (2L_s + L_d)/c} K_s(\nu) e^{i\alpha_s} \bar{U}_0(\nu) \quad [6]$$

where L_r and L_s are the optical pathlengths from the beamsplitter to the reference mirror and sample, respectively; L_d is the optical pathlength from the beamsplitter to the detector; and $K_r(\nu)e^{i\alpha_r(\nu)}$ and $K_s(\nu)e^{i\alpha_s(\nu)}$ are the amplitude reflection coefficients of light backscattered from the reference mirror and turbid sample, respectively.

The total power detected at the interferometer output is given by a time-average of the squared light amplitude:

$$P_d(\tau) = \langle |U_s(t) + U_r(t+\tau)|^2 \rangle \quad [7]$$

where τ is the depth scan time delay. Combining harmonic expansions for $U_s(t)$ and $U_r(t)$ and applying eqn [4] when calculating the time-average, the total power detected is a sum of three terms representing reference (I_r), sample (I_s), and the interference fringe intensity ($\Gamma_{ODT}(\tau)$):

$$\begin{aligned} P_d(\tau) &= \int_0^\infty \big(P_r(\nu) + P_s(\nu) + P_{ODT}(\nu)\big) d\nu \\ &= I_r + I_s + \Gamma_{ODT}(\tau) \end{aligned} \quad [8]$$

with

$$P_r(\nu) = S_0(\nu)|K_r(\nu)|^2 \quad [9]$$

$$P_s(\nu) = S_0(\nu)|K_s(\nu)|^2 \quad [10]$$

$$\begin{aligned} P_{ODT}(\nu) = 2S_0(\nu)K_r(\nu)K_s(\nu) \\ \times \cos\left[2\pi\nu(\Delta_d/c + \tau) + \alpha_s(\nu) - \alpha_r(\nu)\right] \end{aligned} \quad [11]$$

and

$$I_r = \int_0^\infty P_r(\nu) d\nu \quad [12]$$

$$I_s = \int_0^\infty P_s(\nu) d\nu \quad [13]$$

$$\Gamma_{ODT}(\tau) = \int_0^\infty P_{ODT}(\nu) d\nu \quad [14]$$

When there is a moving particle, light scattered from this particle is equivalent to a moving phase front; therefore, Δ_d in eqn [11] can be written as:

$$\Delta_d = \Delta + 2\bar{n}V_z t \quad [15]$$

where Δ is the optical pathlength difference between light in the sampling and reference arms; V_z is the velocity of a moving particle parallel to the probe beam; and $\bar{n}$ is the refractive index of flow media.

To simplify the computation, we assume α_s and α_r are constants over the source spectrum and can be neglected. The spectral domain fringe signal, $P_{ODT}(\nu)$, is simplified to:

$$\begin{aligned} P_{ODT}(\nu) = 2S_0(\nu)K_r(\nu)K_s(\nu) \\ \times \cos\left[2\pi\nu\left((\Delta + 2\bar{n}V_z t)/c + \tau\right)\right] \end{aligned} \quad [16]$$

The corresponding time domain signal, $\Gamma_{ODT}(\tau)$, is given by:

$$\begin{aligned} \Gamma_{ODT}(\tau) = 2\int_0^\infty S_0(\nu)K_r(\nu)K_s(\nu) \\ \times \cos\left[2\pi\nu\left((\Delta + 2\bar{n}V_z t)/c + \tau\right)\right] d\nu \end{aligned} \quad [17]$$

A comparison of eqns [16] and [17] shows that there is a Fourier transformation relation between spectral domain and time domain signals. Consequently, there are two methods to acquire the ODT signal: the time domain method and the spectral domain method.

In the time domain method, a delay line is incorporated in the reference arm to generate a delay. A spectrogram analysis or phase-resolved algorithm is then used to determine the Doppler frequency shift. In the spectral domain method, the reference mirror is fixed, and there is no depth scan ($\tau = 0$). The spectral domain fringe signal, $P_{ODT}(\nu)$, is obtained either by a spectrometer at the detection

arm, or by a frequency sweeping light source. The time domain signal, $\Gamma_{ODT}(\tau)$, is determined from the spectral domain signal by a Fourier transformation.

Spectrogram Method

The spectrogram method uses time–frequency analysis to calculate the Doppler shift. Signal processing algorithms to obtain structural and velocity images from recorded temporal interference fringe intensity using the spectrogram method, are illustrated in Figure 4.

The spectrogram is an estimate of the power spectrum of the temporal interference fringe intensity ($\Gamma_{ODT}(\tau_i, f_j)$) in the ith time delay window. The power spectrum of the temporal interference fringe at the ith pixel, corresponding to time delay τ_i in the structural and velocity images, is calculated by a short-time fast Fourier transformation (STFFT) or a wavelet transformation:

$$\hat{\Gamma}_{ODT}(\tau_i, f_m) = |\mathrm{STFFT}(f_m; \Gamma_{ODT}(\tau))|^2 \qquad [18]$$

where f_m is the discrete frequency value. A tomographic structural image is obtained by calculating the value of the power spectrum at the phase modulation frequency (f_0). Fluid flow velocity is determined from the Doppler frequency shift (f_D), which is the difference between the carrier frequency established by the optical phase modulation (f_0) and the centroid (f_c) of the measured power spectrum at the ith pixel:

$$V_{ODT}(i) = \frac{\lambda_0 f_D}{2\bar{n}\cos\theta} = \frac{\lambda_0(f_c - f_0)}{2\bar{n}\cos\theta} \qquad [19]$$

where we have assumed, $\mathbf{k}_s = -\mathbf{k}_i$ and θ is the angle between $\mathbf{k}_i$ and $\mathbf{v}$. The centroid of the measured power spectrum is determined by:

$$f_c = \sum_m f_m \hat{\Gamma}_{ODT}(\tau_i, f_m) \Big/ \sum_m \hat{\Gamma}_{ODT}(\tau_i, f_m) \qquad [20]$$

Figure 5 shows the *in vivo* structural and blood flow images from a chick chorioallantoic membrane. The lumen wall, chorion membrane, and yolk sac membrane are observed in structural image (a). In the velocity image (b), static regions ($V = 0$) appear dark, while blood flowing at different velocities appears as different brightnesses on the gray scale. The velocity profile taken from a horizontal cross-section passing through the center of the vessel is shown in (c).

Phase-Resolved ODT Method

Although spectrogram methods allow simultaneous imaging of *in vivo* tissue structure and flow velocity, the velocity sensitivity is limited for high-speed imaging. When STFFT or wavelet transformation is used to calculate flow velocity, the resolution is determined by the window size of the Fourier transformation for each pixel. The minimum detectable Doppler frequency shift, f_D, varies inversely with the STFFT window size. Because pixel acquisition time is proportional to the STFFT window size, the image frame rate is limited by velocity resolution. Furthermore, spatial resolution is also proportional to the STFFT window size. Therefore, a large STFFT window size increases velocity resolution while decreasing imaging speed and spatial resolution. This coupling between velocity sensitivity, spatial resolution, and imaging speed prevents the spectrogram method from achieving simultaneously both high imaging speed and high velocity sensitivity, which are essential for measuring flow in small blood vessels where flow velocity is low.

Phase-resolved ODT overcomes the compromise between velocity sensitivity and imaging speed by using the phase change between sequential scans to construct flow velocity images. The phase information of the fringe signal can be determined from the complex analytical signal $\tilde{\Gamma}(t)$, which is determined through analytic continuation of the measured interference fringe function, $\Gamma(t)$, using a Hilbert transformation:

$$\tilde{\Gamma}(t) = \Gamma(t) + \frac{i}{\pi} P \int_{-\infty}^{\infty} \frac{\Gamma(\tau)}{\tau - t}\, d\tau = A(t) e^{i\phi(t)} \qquad [21]$$

where P denotes the Cauchy principle value, i is the complex number, and $A(t)$ and $\phi(t)$ are amplitude and phase of $\tilde{\Gamma}(t)$, respectively. Because the interference signal $\Gamma(t)$ is quasi-monochromatic, the complex

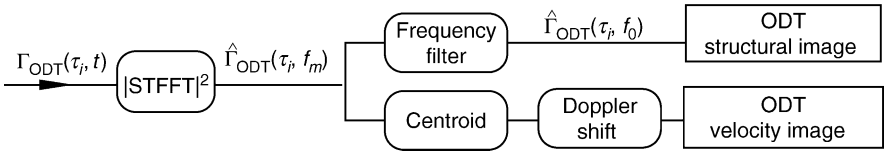


Figure 4 Signal processing algorithms for ODT structural and velocity images.

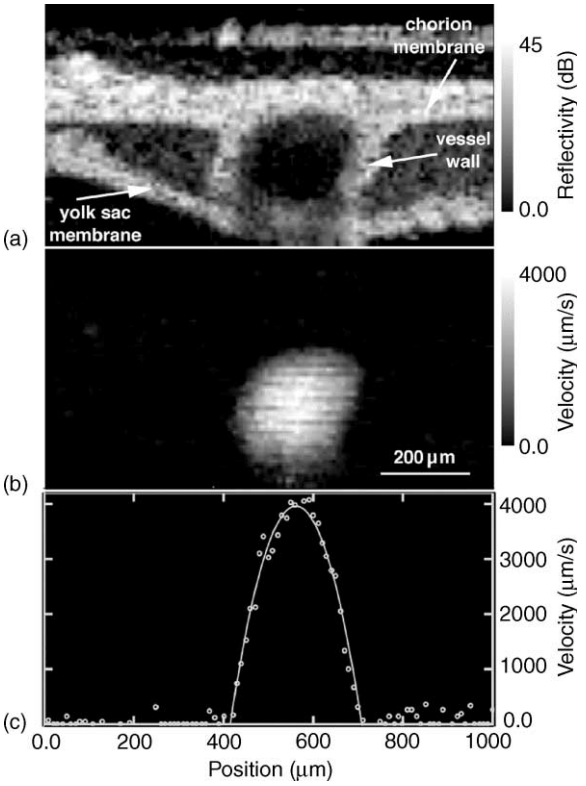


Figure 5 ODT images of *in vivo* blood flow in a CAM vein. (a) Structural image, (b) velocity image, and (c) velocity profile taken from a horizontal cross-section passing through the center of the vein, where the open circles are experimental data and the solid line is a parabolic fit. Reproduced with permission from Chen Z, MilnerTE, Srinivas S, Wang XJ, Malekafzali A, van Gemert MJC and Nelson JS (1997) Noninvasive imaging of *in vivo* blood flow velocity using optical doppler tomography. *Optics Letters* 22: 1119–1121.

analytical signal can be determined by:

$$\tilde{\Gamma}(t) = 2\int_0^{\infty}\int_0^{\tau}\Gamma(t')\exp(-2\pi i\nu t')\mathrm{d}t' \times \exp(2\pi i\nu t)\mathrm{d}\nu \qquad [22]$$

where τ is the time duration of the fringe signal in each axial scan.

The Doppler frequency shift (f_n) at nth pixel in the axial direction is determined from the average phase shift between sequential A-scans. This can be accomplished by calculating the phase change of sequential scans from the individual analytical fringe signal:

$$f_n = \frac{\Delta\phi}{2\pi T} = \frac{1}{2\pi T}\sum_{m=(n-1)M}^{nM}\sum_{j=1}^{N}\left[\tan^{-1}\left(\frac{\mathrm{Im}\tilde{\Gamma}_{j+1}(t_m)}{\mathrm{Re}\tilde{\Gamma}_{j+1}(t_m)}\right) - \tan^{-1}\left(\frac{\mathrm{Im}\tilde{\Gamma}_{j}(t_m)}{\mathrm{Re}\tilde{\Gamma}_{j}(t_m)}\right)\right] \qquad [23]$$

Alternatively, the phase change can also be calculated by the cross-correlation method:

$$f_n = \frac{1}{2\pi T}\tan^{-1}\left(\frac{\mathrm{Im}\left[\sum_{m=(n-1)M}^{nM}\sum_{j=1}^{N}\tilde{\Gamma}_j(t_m)\cdot\tilde{\Gamma}^*_{j+1}(t_m)\right]}{\mathrm{Re}\left[\sum_{m=(n-1)M}^{nM}\sum_{j=1}^{N}\tilde{\Gamma}_j(t_m)\cdot\tilde{\Gamma}^*_{j+1}(t_m)\right]}\right) \qquad [24]$$

where $\tilde{\Gamma}_j(t_m)$ and $\tilde{\Gamma}^*_j(t_m)$ are the complex signals at axial time t_m corresponding to the jth A-scan and its respective conjugate; $\tilde{\Gamma}_{j+1}(t_m)$ and $\tilde{\Gamma}^*_{j+1}(t_m)$ are the complex signals at axial time t_m corresponding to the next A-scan and its respective conjugate; M is an even number that denotes the window size in the axial direction for each pixel; N is the number of sequential scans used to calculate the cross correlation; and T is the time duration between A-scans. Because T is much longer than the pixel time window within each scan used in the spectrogram method, high-velocity sensitivity can be achieved.

Phase-resolved ODT decouples spatial resolution and velocity sensitivity in flow images and increases imaging speed by more than two orders of magnitude, without compromising spatial resolution and velocity sensitivity. In addition, because two sequential A-line scans are compared at the same location, speckle modulations in the fringe signals cancel each other and, therefore, will not affect the phase difference calculation. Consequently, the phase-resolved method reduces the speckle noise in the velocity image. Furthermore, if the phase difference between sequential frames is used, the velocity sensitivity can be further increased. Real-time imaging with velocity sensitivity on the order of 10 μm/s has been demonstrated.

In addition to digital processing of the fringe signal using Hilbert transformation, the complex analytical signal can also be achieved through hardware implementation. Optical Hilbert transformation using polarization optics has been implemented for real-time phase-resolved ODT imaging. Real-time ODT imaging using hardware demodulation of the ODT signal has also been demonstrated.

Spectral Domain Phase-Resolved ODT Method

A schematic diagram of a spectral domain ODT system is shown in Figure 6. The signal from the Michelson interferometer is directly coupled to a

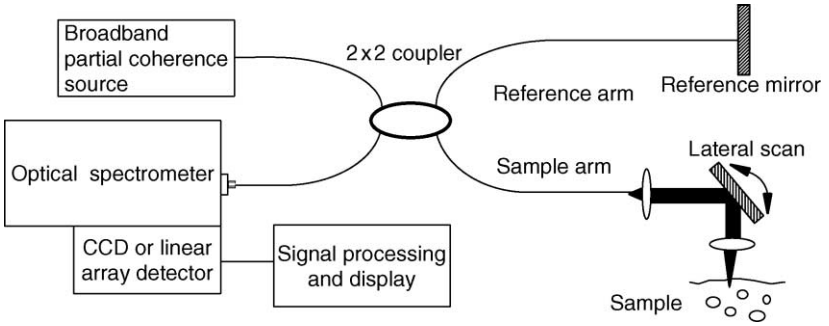


Figure 6 Schematic diagram of a spectral domain ODT instrument.

spectrometer that records the spectral fringe pattern, $P_{ODT}(\nu)$. The temporal interference fringe can be calculated by a Fourier transformation of the spectral fringe pattern (eqn [17]). The Doppler shift can then be determined from the phase shift between sequential scans using the phase-resolved ODT algorithm (eqns [23] and [24]).

Spectral domain phase-resolved ODT has the advantage that no optical delay line is required. In addition, the spectral domain method can achieve high signal to noise ratio over time domain ODT. Since parallel detection can be implemented with a high-speed spectrometer, high-speed imaging acquisition is possible. Finally, because the dynamic range of the phase-resolved ODT depends on the speed of the line scans, spectral domain ODT has the advantage in terms of imaging speed and velocity dynamic range.

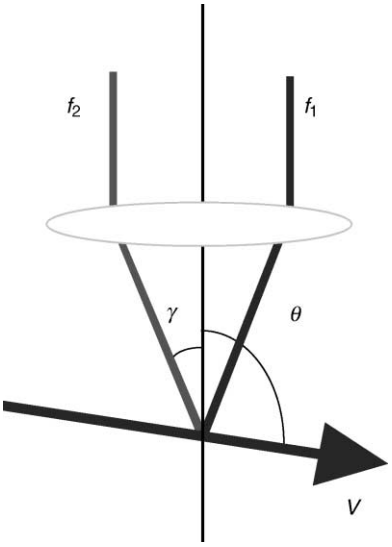


Figure 7 Effect of numerical aperture and transverse flow velocity on Doppler bandwidth.

Transverse Flow Velocity and Doppler Angle Determination

One of the limitations of using the Doppler shift to determine the flow is that the technique is only sensitive to longitudinal flow velocity (Figure 2). If one knows the flow direction, Doppler shift measurement can fully quantify the flow. However, in many biological cases, where flow direction is not known, Doppler shift measurement alone is not enough to fully quantify the flow. Furthermore, there are many clinical cases, such as ocular bloodflow, where vessels are in the plane perpendicular to the probing beam. When flow direction is perpendicular to the probing beam, the Doppler shift is not sensitive to transverse bloodflow (V_T). Therefore, a method to measure transverse flow velocity is essential.

Standard deviation of the Doppler spectra can be used to determine the transverse flow. The technique is based on the fact that ODT imaging uses a relatively large numeric aperture lens in the sampling arm. The beam from different sides of the edges will produce different Doppler shifts, f_1 and f_2, as indicated in Figure 7. Consequently, the Doppler spectra will be broadened by the transverse flow. If the incident beam has a Gaussian spectral profile and contributions from Brownian motion and other sources that are independent of the macroscopic flow velocity are included, a linear relation between standard deviation of the Doppler spectra and transverse flow velocity, $V_T = V \sin\theta$, can be derived:

$$\sigma = \frac{\pi V \sin\theta NA_{\text{eff}}}{8\lambda} + b \qquad [25]$$

where b is a constant, and NA_{eff} is the effective numeric aperture. The standard deviation can be determined from the measured analytical fringe

signal:

$$\sigma^2 = \frac{1}{2\pi^2 T^2} \times \left(1 - \frac{\left| \sum_{m=(n-1)M}^{nM} \sum_{j=1}^{N} \tilde{\Gamma}_j(t_m) \cdot \tilde{\Gamma}^*_{j+1}(t_m) \right|}{\frac{1}{2} \sum_{m=(n-1)M}^{nM} \sum_{j=1}^{N} [\tilde{\Gamma}_j(t_m) \cdot \tilde{\Gamma}^*_j(t_m) + \tilde{\Gamma}_{j+1}(t_m) \cdot \tilde{\Gamma}^*_{j+1}(t_m)]} \right) \quad [26]$$

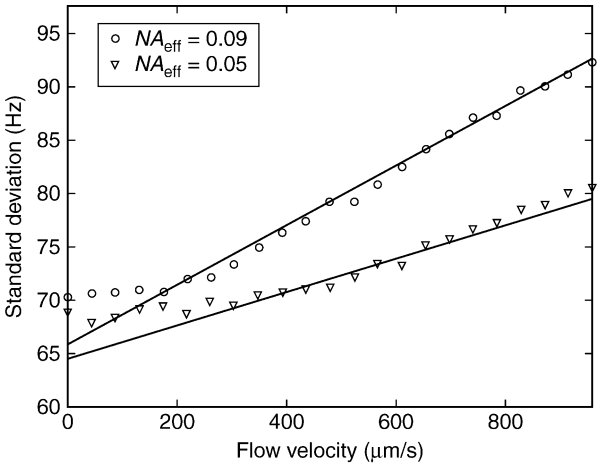


Figure 8 Standard deviation as a function of flow velocity for two different numeric apertures. Reproduced with permission from Ren *et al.* (2002) *Optics Letters* 27: 409–411.

The measured standard deviation as a function of transverse flow velocity is shown in Figure 8. As predicted, the Doppler bandwidth is a linear function of transverse flow velocity above a certain threshold level. The effective numerical aperture of the optical objective in the sample arm determines the slope of this dependence. This result indicates that standard deviation can be used to determine the transverse flow velocity. Since both longitudinal and transverse flow velocity (V_L and V_T) can be measured by the Doppler shift and standard deviation, respectively, flow direction can be determined from a single measurement of the Doppler fringe signal.

Applications

Due to its exceptionally high spatial resolution and velocity sensitivity, several clinical applications of ODT have been demonstrated, including screening vasoactive drugs, monitoring changes in image tissue morphology and hemodynamics following pharmacological intervention and photodynamic therapy, evaluating the efficacy of laser treatment in port wine stain (PWS) patients, assessing the depth of burn wounds, mapping cortical hemodynamics for brain research, imaging ocular bloodflow, and

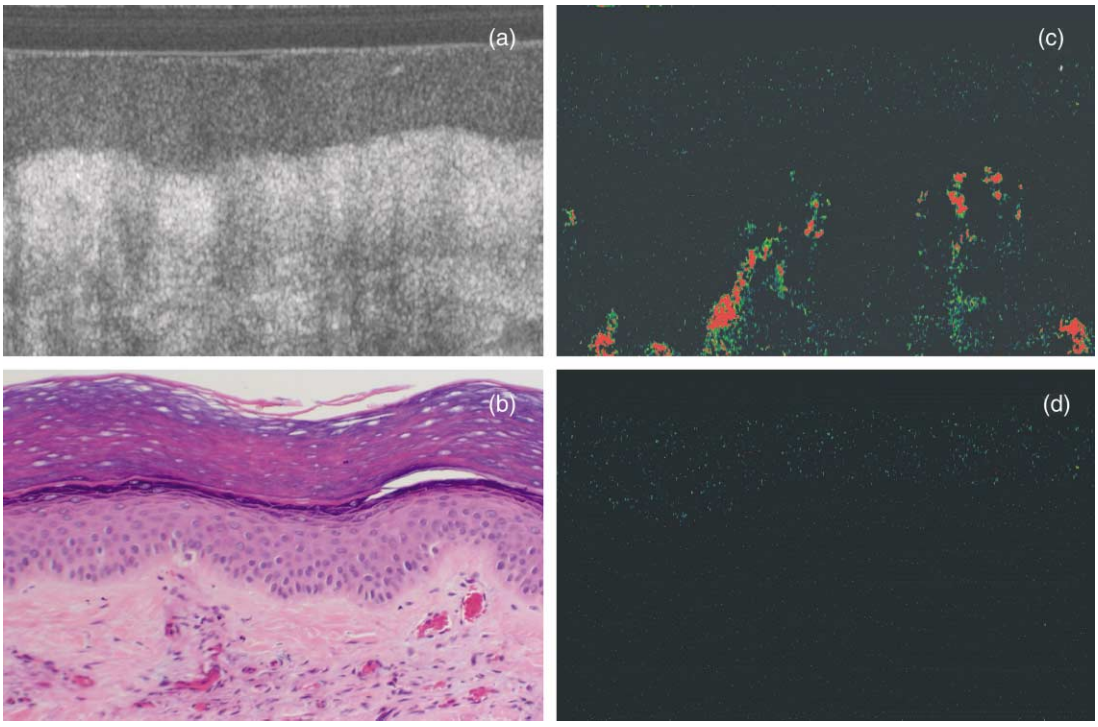


Figure 9 ODT images taken *in situ* from PWS human skin. (a) Structural image, (b) histological section, (c) image before laser treatment, and (d) image after laser treatment (scale bar 100 μm). Reproduced with permission from Nelson *et al.* (2001) *Archives of Dermatology* 137: 741–744. Copyright © 2001, American Medical Association. All rights reserved.

mapping bloodflow in gastrointestinal tracts. Figure 9 shows ODT structural and flow velocity images of a patient with PWS before and after laser treatment, respectively. For comparison, a histology picture taken at the same site is included. The vessel location from the ODT measurement and histology agree very well. Furthermore, the destruction of the vessel by laser can be identified, since no flow appears on the Doppler flow image after laser treatment.

Three-dimensional images of a microvascular network can be obtained by stacking 2D scans together. It is known that the microvasculature of mammary tumors has several distinct differences from normal tissues. Mapping of a 3D cancer microvascular network could provide additional information for cancer diagnosis. Figure 10 shows multiple blood vessels imaged from a patient with an abnormal microvascular network. Different colors represent different signs of the Doppler shift, which depends on the angle between the direction of flow and probing beam.

ODT is also a powerful tool to study flow dynamics in microfluidic channels. In addition to imaging flow velocity profile, ODT can also be used to measure osmotic mobility, quantify size of the scattering particle, and study flow dynamics of microfluids in microchannels of different materials, geometry, and surface treatment.

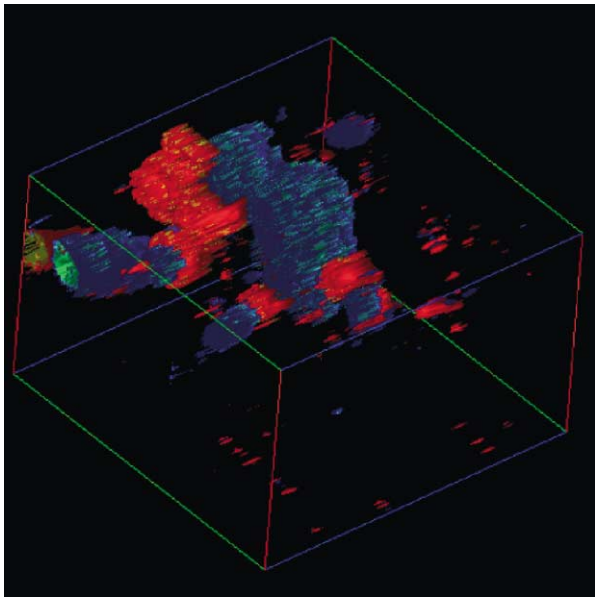


Figure 10 Three-dimensional ODT images of multiple blood vessels in human skin from a patient with a PWS birthmark. Reproduced with permission from Zhao *et al.* (2001) *IEEE J. Select Topics Quantum Electro.* 7: 931–935. © 2001, IEEE.

Polarization Sensitive OCT

PS-OCT combines polarization sensitive detection with OCT to determine tissue birefringence. PS-OCT can obtain enhanced image contrast and additional physiological information by studying polarization properties of biological tissues. Many biological tissues, such as tendon, muscle, nerve, bone, cartilage, and skin, exhibit birefringence properties. Consequently, PS-OCT can provide contrast mechanisms that are directly related to the physiological conditions of these tissues.

The polarization state of a light beam can be described by a Stokes vector **S** as:

$$\mathbf{S} = \begin{bmatrix} I \\ Q \\ U \\ V \end{bmatrix} = \begin{bmatrix} a_x^2 + a_y^2 \\ a_x^2 - a_y^2 \\ 2a_x a_y \cos\varphi \\ 2a_x a_y \sin\varphi \end{bmatrix} \quad [27]$$

where I, Q, U, V are Stokes parameters, a_x and a_y are amplitudes of two orthogonal components of the electric vector, and φ represents the phase difference between the two components.

The effect of an optical device on the polarization of light can be characterized by a 4×4 Mueller matrix. The matrix acts on the input state, $\mathbf{S}_1$, to give the output state, $\mathbf{S}_2$:

$$\mathbf{S}_2 = \begin{bmatrix} I_2 \\ Q_2 \\ U_2 \\ V_2 \end{bmatrix} = \mathbf{M}\mathbf{S}_1$$

$$= \begin{bmatrix} M_{00} & M_{01} & M_{02} & M_{03} \\ M_{10} & M_{11} & M_{12} & M_{13} \\ M_{20} & M_{21} & M_{22} & M_{23} \\ M_{30} & M_{31} & M_{32} & M_{33} \end{bmatrix} \begin{bmatrix} I_1 \\ Q_1 \\ U_1 \\ V_1 \end{bmatrix} \quad [28]$$

In the Poincaré sphere representation, a polarization state, **S**, can be represented by a point (Q, U, V). The Poincaré sphere representation provides a convenient method to evaluate changes of Stokes vectors. Determination of either the Stokes vector or the Mueller matrix allows full quantification of birefringence properties of the biological sample.

A fiber-based, high-speed PS-OCT system is shown in Figure 11. The system is similar to the ODT system, except a polarization modulator and two orthogonal detection channels are included. Due to the coherence detection intrinsic to OCT, backscattered light that

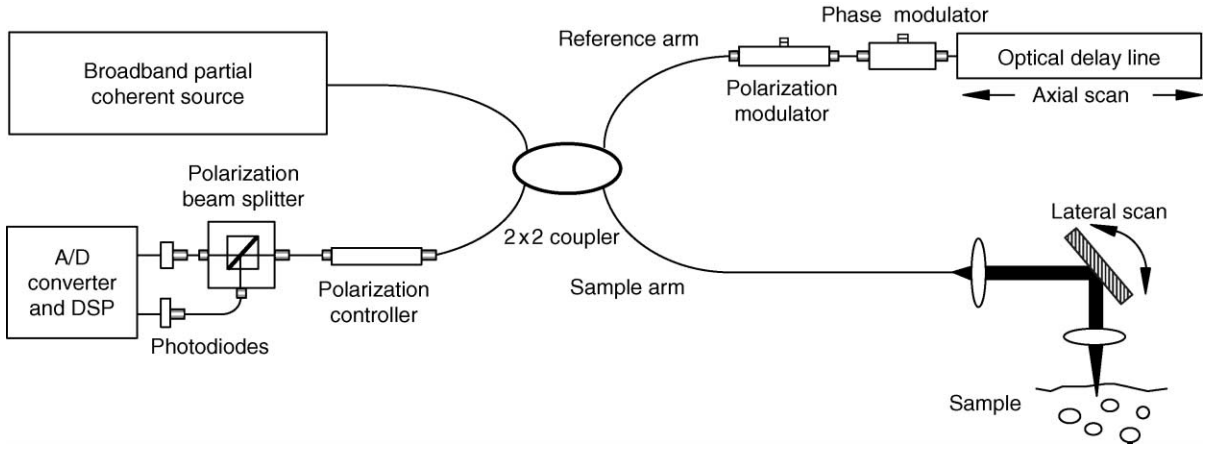


Figure 11 Fiber-based PS-OCT system. DSP: digital signal processing.

originates from incident light in different polarization states can be determined by controlling polarization states in the reference arm. Modulating the reference light in four orthogonal polarization states allows coherent detection of backscattered light from the sample under equivalent illumination in four different polarization states. The choice of orthogonal polarization states in the Poincaré sphere is important because it ensures that the birefringence measurements will be independent of the orientation of the optical axis in the sample.

For every polarization state controlled by the polarization modulator, the A-scan signals corresponding to the two orthogonal polarization diversity channels are digitized. The coherence matrix can be calculated from the complex electrical field vector from these two channels:

$$J = \begin{bmatrix} \langle E_{\mathrm{H}}^{*}(t)E_{\mathrm{H}}(t)\rangle & \langle E_{\mathrm{H}}^{*}(t)E_{\mathrm{V}}(t)\rangle \\ \langle E_{\mathrm{V}}^{*}(t)E_{\mathrm{H}}(t)\rangle & \langle E_{\mathrm{V}}^{*}(t)E_{\mathrm{V}}(t)\rangle \end{bmatrix} = \begin{bmatrix} J_{\mathrm{HH}} & J_{\mathrm{HV}} \\ J_{\mathrm{VH}} & J_{\mathrm{VV}} \end{bmatrix} \quad [29]$$

where E_{H}, E_{V} are the components of the complex electric field vector corresponding to horizontal and vertical polarization channels, respectively, and E_{H}^{*}, E_{V}^{*} are their conjugates, respectively. The Stokes vector can be derived from the coherence matrix:

$$\begin{aligned} S_0 &= J_{\mathrm{HH}} + J_{\mathrm{VV}} & S_2 &= S_0^{-1}(J_{\mathrm{HV}} + J_{\mathrm{VH}}) \\ S_1 &= S_0^{-1}(J_{\mathrm{HH}} - J_{\mathrm{VV}}) & S_3 &= iS_0^{-1}(J_{\mathrm{VH}} - J_{\mathrm{HV}}) \end{aligned} \quad [30]$$

where S_1, S_2, and S_3 are the normalized coordinates of the Stokes vector in the Poincaré sphere, characterizing the polarization state of the backscattered light, and S_0 is the module of the Stokes vector characterizing light intensity. The Stokes vector, corresponding to each of the input polarization states for the nth pixel in the jth A-scan, can be calculated as:

$$\begin{aligned} S_{0,n} &= \sum_{m=(n-1)M}^{nM} \left[\tilde{\Gamma}_j^{\mathrm{H}}(t_m)\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m) + \tilde{\Gamma}_j^{\mathrm{V}}(t_m)\tilde{\Gamma}_j^{\mathrm{V}^*}(t_m)\right] \\ S_{1,n} &= S_{0,n}^{-1} \sum_{m=(n-1)M}^{nM} \left[\tilde{\Gamma}_j^{\mathrm{H}}(t_m)\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m) - \tilde{\Gamma}_j^{\mathrm{V}}(t_m)\tilde{\Gamma}_j^{\mathrm{V}^*}(t_m)\right] \\ S_{2,n} &= S_{0,n}^{-1} \sum_{m=(n-1)M}^{nM} 2\,\mathrm{Re}\left(\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m)\tilde{\Gamma}_j^{\mathrm{V}}(t_m)\right) \\ S_{3,n} &= S_{0,n}^{-1} \sum_{m=(n-1)M}^{nM} 2\,\mathrm{Im}\left(\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m)\tilde{\Gamma}_j^{\mathrm{V}}(t_m)\right) \end{aligned} \quad [31]$$

where M is an integer number that determines the size of the average in the axial direction for each pixel; $\tilde{\Gamma}_j^{\mathrm{H}}(t_m)$ and $\tilde{\Gamma}_j^{\mathrm{V}}(t_m)$ are complex signals detected from the two orthogonal polarization channels at axial time t_m, for the jth A-scan; and $\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m)$ and $\tilde{\Gamma}_j^{\mathrm{V}^*}(t_m)$ are their conjugates, respectively.

In order to measure the birefringence properties of the sample accurately, four states of light polarization are generated for each lateral pixel. For each polarization state, one A-line scan is performed. Therefore, a total of four A-line scans are used to calculate the Stokes vector and phase retardation images simultaneously.

For a sample with an assumed linear birefringence, there exists two eigenwaves that are polarized along the projected fast and slow axes of the sample normal to the propagation direction of incident light. The Stokes vectors of these eigenwaves determine a rotation axis in the equator plane of a Poincaré sphere.

The effect of birefringence is to rotate the Stokes vector about this axis through an angle that is equal to the phase retardation of the sample. Conversely, the rotation axis can be determined from the known polarization states of incident and backscattered light at the sample location. The phase retardation image, which characterizes the accumulated birefringence distribution in the sample, is calculated by the rotation of the Stokes vectors in the Poincaré sphere.

The simultaneous Stokes vectors and phase retardation images of a rat muscle are shown in Figure 12.

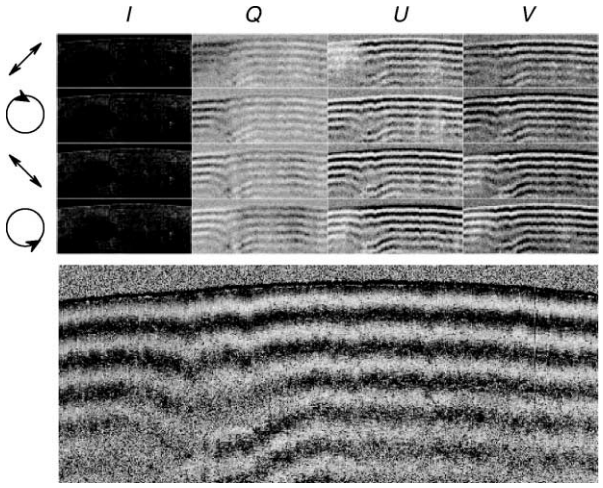


Figure 12 Stokes vectors images (top) and phase retardation image (bottom) in fresh rat muscle. The image area is 2 mm × 5 mm. Reproduced with permission from Ren *et al.* (2003) *SPIE Proceedings* 4956: 320–327.

Stokes vectors, corresponding to four different polarization states, are shown in the top figure. From Stokes vector images, the phase retardation image can be calculated as shown in the bottom figure. The banded structure in the phase retardation image indicates the accumulated tissue birefringence distribution in the rat muscle. In addition, differential optical axis and birefringence images can also be calculated.

There are a number of clinical applications for PS-OCT, including burn depth determination and retinal fiber nerve layer thickness evaluation. Figure 13 shows *in vivo* structural and phase retardation images from rat skin burned at 100°C for 10 seconds. Although the structure image shows very little contrast, the phase retardation image clearly shows the contrast between the normal and burned regions. As light propagates into the tissue, the retardation increases due to tissue birefringence. For normal tissue, it only takes less than 20 μm to cause a phase retardation of 90°. However, for the burned tissue, it takes much thicker tissue to cause the same phase retardation. The decrease of birefringence in thermally damaged tissue can clearly be seen in PS-OCT scans in comparison to normal rat skin.

Second Harmonic OCT

SH-OCT combines SHG with coherence gating for high-resolution tomographic imaging of tissue structures with molecular contrast. SHG is the lowest-order nonlinear optical process where the second-order nonlinear optical susceptibility is

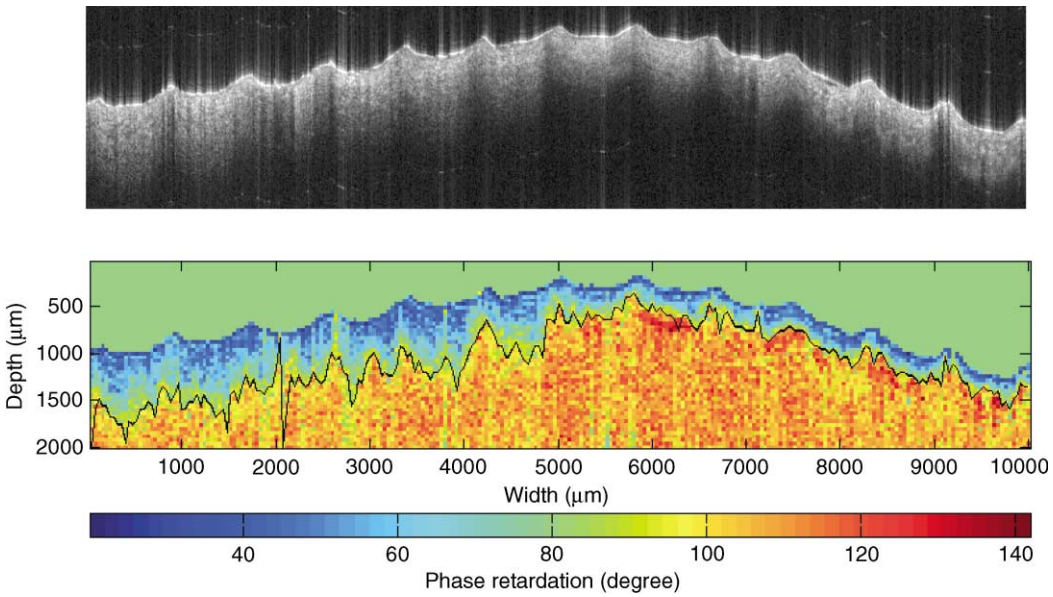


Figure 13 *In vivo* structure (top) and phase retardation image (bottom) from rat skin burned at 100°C for 10 seconds. The black contour line in the phase retardation image demarcates the depth at which 90° phase retardation has been reached with respect to the incident polarization (image size 2 mm × 10 mm).

responsible for the generation of light at second-harmonic frequency. Because second-order nonlinear optical susceptibility is very sensitive to electronic configurations, molecular structure and symmetry, local morphology, and ultrastructures, SHG provides molecular contrast for the coherence image.

The schematic diagram for SH-OCT is shown in Figure 14. A broadband short pulse laser is used as a light source in a Michelson interferometer. The light is split into the reference arm and the sample arm. In the reference arm, a thin nonlinear crystal is used to convert the input radiation to SH photons. Both SH and fundamental waves are then reflected by a metal mirror (M1) mounted on a motorized translation stage, which acts as the delay line in this SH-OCT system. A backscattered SH wave from the sample recombines with the SH wave from the reference arms to form interference fringe. The SH interference fringe signal is detected by a photomultiplier tube (PMT) after passing through a short-pass filter (F1). The fundamental interference fringe signal is detected by a photodiode (PD) after passing through a long-pass filter (F2).

A SH-OCT image of a rat tendon is shown in Figure 15. Compared with conventional OCT performed at fundamental wavelengths, SH-OCT offers enhanced molecular contrast and spatial resolution. It is also an improvement over existing SHG scanning microscopy technology as the intrinsic coherence gating mechanism enables the detection and discrimination of SH signals generated at deeper locations. The enhanced molecular contrast of SH-OCT extends conventional OCT's capability for detecting small changes in molecular structure. SH-OCT is promising for the diagnosis of cancers and other diseases at an early stage when changes in tissue and molecular structure are small.

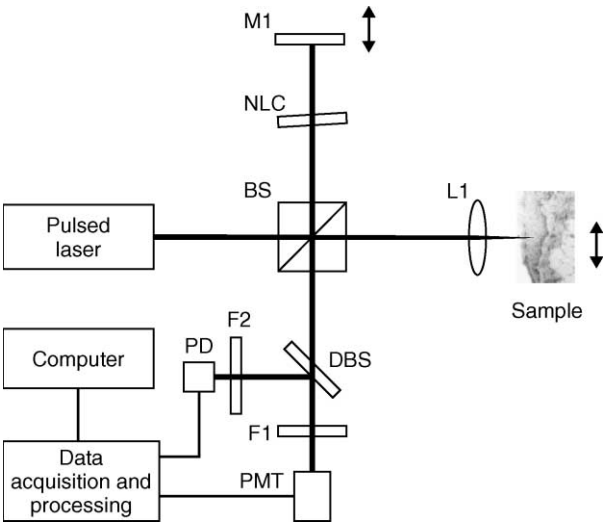


Figure 14 Schematic of experiment set-up for SH-OCT. M1, mirror; NLC, nonlinear crystal; BS, broadband nonpolarization beamsplitter; DBS, dichroic beamsplitter; L1, lens; F1–F2, filters; PD, photo diode; PMT, photomultiplier tube.

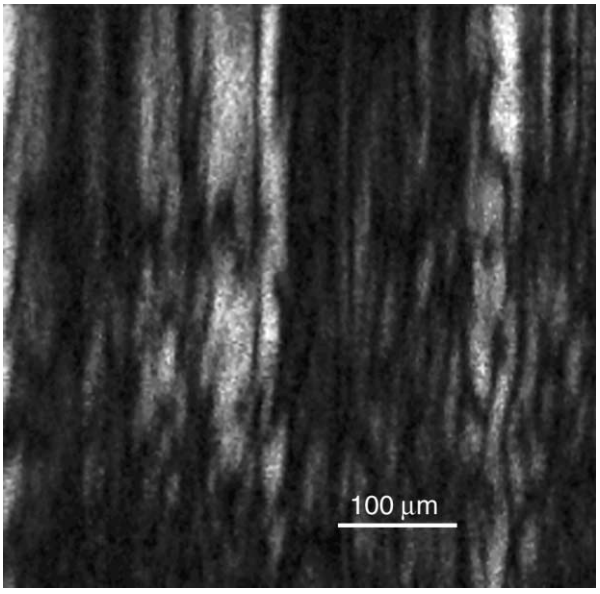


Figure 15 SH-OCT image of a rat tendon; image size 250 μm × 250 μm. Scale bar 100 μm. Image reprinted with permission from Jiang Y *et al.* (2004) Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

Conclusions

Optical tomography is a rapidly developing imaging technology with many potential applications. Through different contrast enhancement mechanisms, ODT, PS-OCT, and SH-OCT provide clinically important physiological information that is not available in the structure image. Given the noninvasive nature and exceptionally high spatial resolution, optical tomography can simultaneously provide tissue structure, blood perfusion, birefringence, and other physiological information and have great potential for basic biomedical research and clinical medicine.

Acknowledgments

I would like to thank many of my colleagues who have contributed to the functional OCT project at the Beckman Laser Institute and Center for Biomedical Engineering at UCI, particularly the students and postdoctoral fellows. I also want to acknowledge the research grants awarded from the National Institutes of Health (EB-00293, NCI-91717, RR-01192 and EB-00255), the National Science Foundation (BES-86924), the Whitaker

Foundation (WF-23281) and the Defense Advanced Research Program Agency (Bioflip program). Institutional support from the Air Force Office of Scientific Research (F49620-00-1-0371), and the Beckman Laser Institute Endowment are also gratefully acknowledged. Dr Chen's e-mail address is zchen@bli.uci.edu.

List of Units and Nomenclature

OCT	Optical coherence tomography
ODT	Optical Doppler tomography
PS-OCT	Polarization sensitive OCT
SH-OCT	Second harmonic OCT
STFFT	Short time fast Fourier transformation

See also

Coherence: Coherence and Imaging; Overview. **Fiber and Guided Wave Optics:** Nonlinear Optics; Optical Fiber Cables. **Imaging:** Imaging Through Scattering Media; Interferometric Imaging; Photon Density Wave Imaging. **Interferometry:** Overview. **Microscopy:** Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy. **Polarization:** Introduction; Matrix Analysis. **Tomography:** Optical Coherence Tomography; Tomography and Optical Imaging.

Further Reading

Bouma BE and Tearney GJ (eds) (2001) *The Optical Coherence Tomography Handbook*. New York: Marcel Dekker, Inc.

Chen Z (2003) Functional optical coherence tomograpy. In: Hwang NHC and Woo LY (eds) *Frontiers in Biomedical Engineering*, pp. 345–364. New York: Kluwer Academic Publishers.

Chen Z, Milner TE, Srinivas S, Wang XJ, Malekafzali A, van Gemert MJC and Nelson JS (1997) Noninvasive imaging of *in vivo* blood flow velocity using optical doppler tomography. *Optics Letters* 22: 1119–1121.

Chen Z, Zhao Y, Srinivas SM, Nelson S, Prakash N and Frostig RD (1999) Optical Doppler tomography. *IEEE Journal of Selected Topics in Quantum Electronics* 5(4): 1134–1142.

de Boer JF and Milner TE (2002) Review of polarization sensitive optical coherence tomography and Stokes vector determination. *Journal of Biomedical Optics* 7: 359–371.

Fercher AF and Hizenberger CK (2002) Optical coherence tomography. *Progress in Optics* 44: 215–296.

Huang D, Swanson EA, Lin CP, *et al.* (1991) Optical coherence tomography. *Science* 254: 1178–1181.

Izatt JA, Kulkarni MD, Yazdanfar S, Barton JK and Welch AJ (1997) *In vivo* bidirectional color Doppler flow imaging of picoliter blood volumes using optical coherence tomography. *Optics Letters* 22: 1439–1441.

Jiang Y, Tomov I, Wang Y and Chen Z (2004) Second harmonic optical coherence tomography. *Optics Letters* 29: 1090–1092.

Jiao S and Wang LV (2002) Jones-matrix imaging of biological tissues with quadruple-channel optical coherence tomography. *Journal of Biomedical Optics* 7: 350–358.

Masters BR and Thompson BJ (eds) (2001) *Selected Papers on Optical Low-Coherence Reflectometry & Tomography*. SPIE Milestone Series, vol. MS 165. Bellingham, WA: SPIE Optical Engineering Press.

Tuchin VV (ed.) (2004) *Coherent-Domain Optical Methods for Biomedical Diagnostics, Environmental and Material Science*, vol. 2. Boston, MA: Kluwer Academic Publishers.

Vo-Dinh T (ed.) (2003) *Biomedical Photonics Handbook*. New York: CRC Press.

U

ULTRAFAST LASER TECHNIQUES

Contents

Generation of Femtosecond Pulses
Pulse Characterization Techniques

Generation of Femtosecond Pulses

D T Reid, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical pulses from femtosecond lasers represent some of the shortest physical events ever observed and their high instantaneous powers allow them to be focused to intensities that can heat and compress matter to resemble the extreme conditions found in a nuclear explosion. The extreme units of time and power needed to describe this area of physics can easily distract from the simple and elegant techniques by which femtosecond pulses can be generated. The review presented here is divided into two main sections. The first explains the theory of pulse generation and shaping and the second concentrates on actual femtosecond laser sources in common use today.

Theory of Femtosecond Pulse Generation

Active Mode-locking

In an inhomogeneously broadened laser system, optical frequencies within the emission bandwidth of the material can experience gain simultaneously at discrete frequencies corresponding to the longitudinal modes of the laser resonator. In the absence of any mechanism to coherently couple energy between these modes, the laser will operate with a continuous-wave (cw) output containing a narrow band of frequencies whose roundtrip cavity gain is highest. An alternative operating regime exists in which each mode is related to its neighbor by a fixed phase relationship and it is this mode-locked condition that is used to generate a periodic sequence of ultrashort optical pulses.

To understand how mode-locking works, consider the frequency separation between adjacent longitudinal resonator modes which can be expressed as

$$\Delta\omega = 2\pi c/l \qquad [1]$$

where l is the optical roundtrip length of the resonator and c is the speed of light. If the nth longitudinal mode has an amplitude E_n then the total optical field can be denoted in complex notation as

$$E(t) = \sum_n E_n \exp \mathrm{i}[(\omega_0 + n\Delta\omega)t + n\Delta\varphi] \qquad [2]$$

with ω_0 being the center frequency of the output and $\Delta\varphi$ being the phase difference between adjacent modes. Figure 1 illustrates how a simple pulse can be produced by coherently adding together adjacent modes. In Figure 1a, three modes with equal amplitudes and the same initial phase ($\Delta\varphi = 0$) are interfered to give a pulse. As further modes are added (Figure 1b) the pulse duration decreases.

Coupling of energy between adjacent cavity modes can be achieved in practice by modulating either the

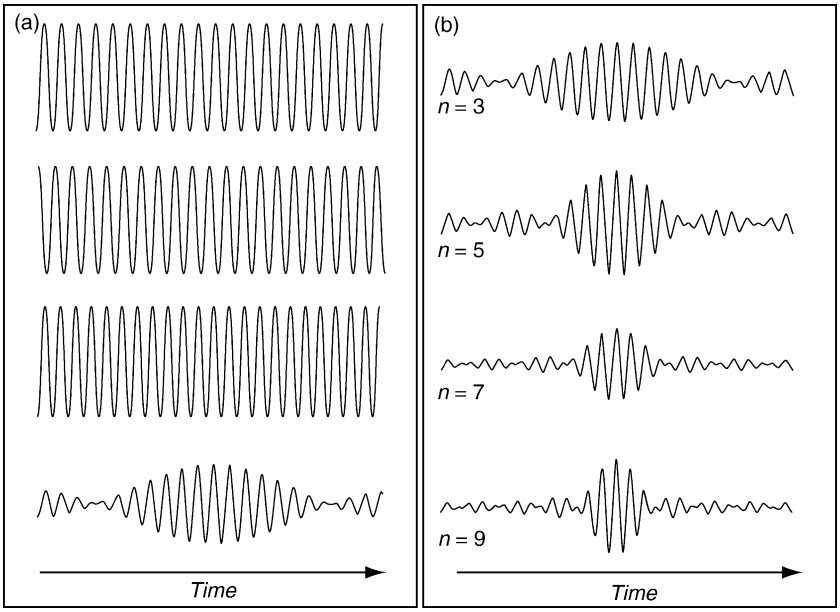


Figure 1 The principles of mode-locking: (a) three adjacent modes add coherently to produce a pulse; (b) shorter pulses require a larger number of *n* coupled modes.

intracavity loss (amplitude modulation (AM) mode-locking) or the intracavity phase (frequency modulation (FM) mode-locking). A common practical amplitude modulator is an acousto-optic modulator and for operation in the visible and near-infrared, fused silica is a common acousto-optic modulator material. Lithium-niobate devices based on the Pockels effect are the most commonly used phase modulators and are attractive for mode-locked fiber-laser applications because fiber-pigtailed waveguide devices are readily available.

Active mode-locking has been employed in a wide variety of laser systems including argon ion, dye, erbium-doped fiber, semiconductor, and solid-state lasers. Unlike passive mode-locking, no optical nonlinearity is required which means that actively mode-locked lasers can be operated with low intracavity powers. The technique is particularly suitable for mode-locked sources which require active synchronization to an external clock source.

A form of 'optical' active mode-locking exists in which gain modulation resulting from optical pumping using another mode-locked laser is the mode-locking mechanism. Known as synchronously pumped mode-locking, this technique has been used to obtain ultrashort pulses from oscillators incorporating gain media of laser or nonlinear optical materials. The technique is most appropriate to laser gain materials which have a short-lived fluorescence lifetime and whose gain therefore exhibits a significant change when the pump light is modulated at high frequencies. Synchronous pumping has been used to generate mode-locked pulses from color-center lasers, dye lasers, and optical parametric oscillators.

Passive Mode-locking

To produce the shortest optical pulses, passive mode-locking, a technique based on exploiting intracavity optical nonlinearities, must be employed. Unlike the active technique, the strength of the mode-locking action increases as the propagating pulses shorten and their peak intensities rise. One consequence of this nonlinear behavior is that the passive technique commonly generates and mode-locks entirely new cavity frequencies which were not originally present during cw operation.

Passive mode-locking has been implemented using a wide range of different nonlinear effects, but in every technique the mode-locking process can be described in terms of dynamic loss saturation combined, in certain systems, with dynamic gain saturation. The principal mode-locking element is a saturable absorber whose loss decreases as the incident pulse intensity increases. A saturable absorber can take the form either of a resonant component such as a solid-state semiconductor device or a dye jet with a fixed absorption wavelength, or a nonresonant component that replicates a saturable-absorber-like action using self-focusing or self-phase-modulation effects. Saturable absorbers can be categorized as either slow saturable absorbers or fast saturable

absorbers, depending on whether their recovery time is longer or shorter than the pulse duration, respectively. The pulse-shaping mechanisms are different in each case and will be described separately.

Slow saturable absorbers are physical devices, commonly dye molecules in solution or semiconductor-doped glass, which rely on resonant excitation to produce an excited electronic state whose transmission is greater than that of the ground state. When an optical pulse is incident on a slow absorber, its leading edge is absorbed and creates an excited state which is relatively transparent to the trailing edge. In a laser medium whose gain is readily saturated, a pulse propagating first through the absorber and then through the gain medium experiences a 'gain window' with a finite duration which strongly shapes the pulse (see Figure 2a,b).

A fast saturable absorber is required to realize mode-locking in laser media which are not easily saturated. The time-dependent gain and loss profiles of the absorber and the gain medium are again depicted in Figure 2c,d and refer to an absorber with both a rapid response and recovery time. In this system, the dynamics of the absorber and gain medium no longer represent the limiting factor influencing the pulse duration, and other effects such as linear and nonlinear dispersion become the dominant pulse-shaping mechanisms.

Pulse Shaping by Material Dispersion

Intense pulses propagating in a mode-locked laser experience strong linear and nonlinear shaping effects associated, respectively, with group-velocity dispersion and self-phase-modulation. Contributions to the net intracavity dispersion arise from the gain medium, the cavity mirrors, other elements such as intracavity prisms, even the air in the beam path. Dispersive effects in materials originate as a result of a frequency-dependent dielectric susceptibility:

$$P(\omega) = \varepsilon_0 \chi^{(1)}(\omega) E(\omega) + \varepsilon_0 \chi^{(2)}(\omega) E(\omega)^2 + \varepsilon_0 \chi^{(3)}(\omega) E(\omega)^3 + \cdots \quad [3]$$

which is due to physical resonances (poles) in the electronic response to an applied E-field.

Linear dispersion

Linear dispersion refers to the variation of the refractive index associated with the real part of the $\chi^{(1)}$ term, namely,

$$n_0 = \sqrt{1 + \mathrm{Re}\{\chi^{(1)}\}} \quad [4]$$

and leads to a wavelength-dependent group velocity which is responsible for pulse broadening and break-up in optical systems. Dispersive effects are best analyzed in terms of optical phase which is related to the refractive index by

$$\varphi(\omega) = \frac{n_0(\omega)\omega L}{c} \quad [5]$$

where L is the medium length and c is the vacuum speed of light. The local variation of the spectral

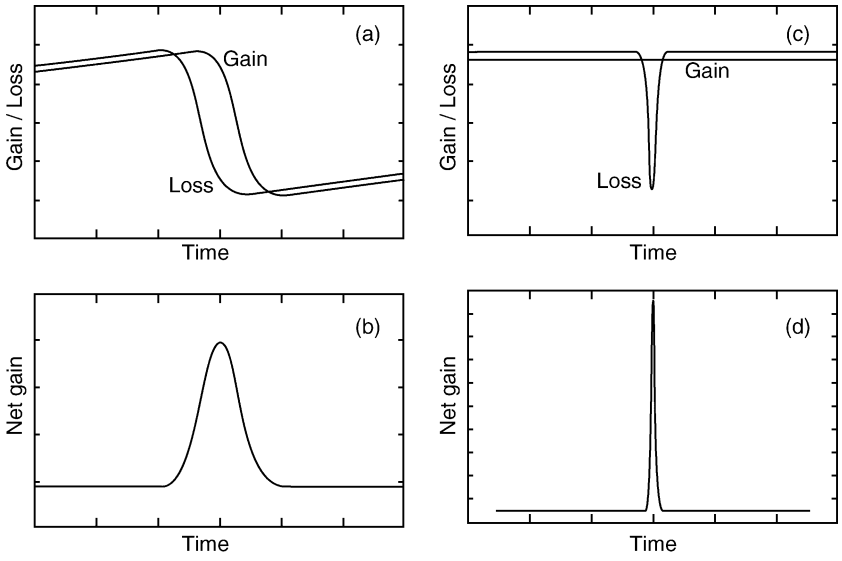


Figure 2 Saturable absorber action: (a) intracavity gain and loss profiles for a laser comprising a slow saturable absorber and a rapid relaxation time gain medium, and (b) the net gain (gain − loss) for this system; (c) gain and loss profiles for a laser comprising a near-instantaneous saturable absorber and a slow relaxation time gain medium, and (d) the net gain for this system.

phase due to an optical medium can be represented as the Taylor series

$$\varphi(\omega) = \varphi(\omega_0) + (\omega - \omega_0)\frac{d\varphi}{d\omega}\bigg|_{\omega_0} + \frac{1}{2}(\omega - \omega_0)^2 \frac{d^2\varphi}{d\omega^2}\bigg|_{\omega_0} + \left|\frac{1}{6}(\omega - \omega_0)^3 \frac{d^3\varphi}{d\omega^3}\right|_{\omega_0} + \cdots \quad [6]$$

where ω_0 is the center frequency of the incident light. When an ultrashort pulse propagates through the medium, only the quadratic and higher-order terms in [6] have a direct influence on the shape of the pulse in time. The $(\omega - \omega_0)^2$ term is responsible for adding linear chirp which refers to a time-varying linear increase or decrease in the instantaneous optical frequency across the pulse. Obtaining the shortest pulses from a mode-locked laser requires the quadratic and higher-order terms in the roundtrip intracavity spectral phase to be zero. In practice this means minimizing the material group-velocity dispersion of the gain medium by using a short but highly doped laser crystal and then identifying additional optical elements whose group-velocity dispersion is equal in magnitude but opposite in sign to that of the crystal. Optical components for compensating group-velocity dispersion fall into two categories: those which use bulk optics and geometry to achieve a wavelength-dependent path length, and others based on interference effects within a dielectric optical coating.

Geometrical methods of dispersion compensation are commonly based on pairs of diffraction gratings or prisms and in both cases the geometry of the systems means that the group delay is wavelength dependent and can be made greater for longer wavelengths than for shorter ones. Analytic expressions describing the quadratic and cubic spectral phase of prism and diffraction grating pairs are summarized in Table 1. Dispersion compensation using prisms has been extremely successful and 11-fs pulses were reported using this approach. The simultaneous compensation of arbitrary second- and third-order spectral phase is, however, difficult and involves identifying a suitable prism material, apex spacing, and tip insertion. The expressions listed in Table 1 indicate that the form of the spectral phase depends on the refractive index of the prism material and its wavelength derivatives, while the apex separation, l, also contributes to a large extent. A further contribution to the spectral phase comes from the glass encountered by the beam passing through the prism tips which contributes a positive dispersion described by the equations for a bulk material given in Table 1. In general, an iterative solution of the equations given in Table 1 is needed to determine the geometry of a prism pair suitable for producing a desired spectral phase profile.

Laser cavity mirrors can be designed with chirped multilayer dielectric coatings which possess a chosen group-velocity dispersion profile. A Bragg mirror with smaller layer periods at the surface and larger periods deeper in the coating provides negative group-velocity dispersion because the longer wavelengths experience larger group delays within the coating than the shorter ones. These chirped mirrors provide constant reflectivity across their stopband and therefore only modify the spectral phase of a pulse, leaving the intensity unchanged. The design of chirped mirrors has been refined in recent years to enhance their operating bandwidth and to reduce undesirable modulations in the group-delay response. The technique has been used to obtain 6.5-fs pulses directly from a Ti:sapphire oscillator and to achieve compression of amplified pulses to below 5 fs.

Nonlinear dispersion

Nonlinear effects arise in all materials due to the $\chi^{(3)}$ term in [3]. In a transparent medium the refractive index due to this net susceptibility is approximately

$$n \approx n_0 + \frac{\chi^{(3)}E^2}{2n_0} = n_0 + n_2 I \quad [7]$$

where I is the optical intensity. The importance of the $\chi^{(3)}$ susceptibility is therefore that it causes the refractive index to become intensity-dependent, and significant changes in the refractive index can be induced by the high peak intensities of ultrashort pulses.

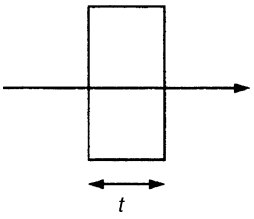
Self-phase-modulation (SPM) resulting from the nonlinear refractive index is essential in generating the shortest pulses from modern ultrafast lasers and is caused by the time-varying intensity profile of the pulse. The higher-intensity components induce a larger nonlinear refractive index change and therefore experience a greater phase shift than the weaker components. The resulting frequency shift across the pulse can be shown to be

$$\Omega(t) = -\frac{d\phi(t)}{dt} = -\frac{\omega n_2 L}{c}\frac{dI(t)}{dt} \quad [8]$$

which describes a redshift of the pulse leading edge and a blueshift of the trailing edge. Only the pulse center frequency remains unchanged and SPM therefore broadens the pulse spectrum by redistributing energy from the center to the wings. Across the center of the pulse, SPM leads to an approximately linear positive chirp that, with appropriate optics, can be

Table 1 Expressions for the second and third-order dispersion of common optical systems

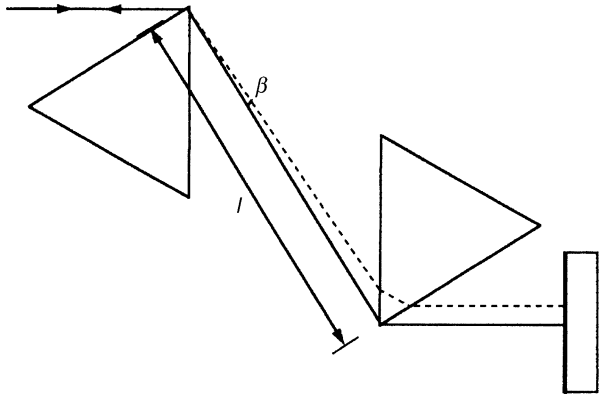
Bulk material, thickness t



$$\frac{d^2\varphi}{d\omega^2} = t\frac{\lambda^3}{2\pi c^2}\frac{d^2n}{d\lambda^2}$$

$$\frac{d^3\varphi}{d\omega^3} = -t\frac{\lambda^4}{4\pi^2c^3}\left(3\frac{d^2n}{d\lambda^2} + \lambda\frac{d^3n}{d\lambda^3}\right)$$

Double-pass prism-pair, apex separation l, angular deviation within ray of β



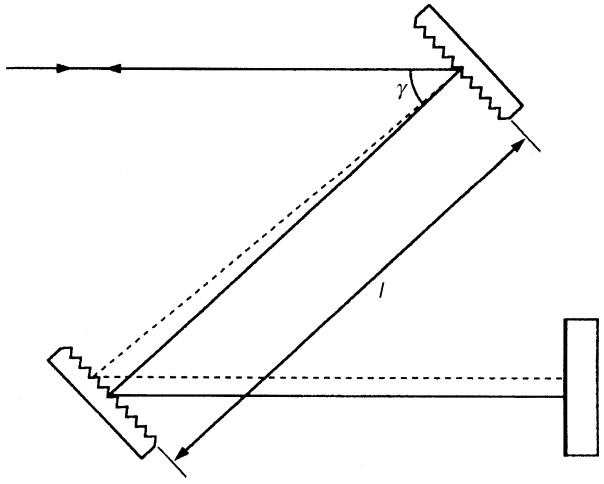
$$\frac{d^2\varphi}{d\omega^2} = \frac{\lambda^3}{2\pi c^2}\frac{d^2P}{d\lambda^2}$$

$$\frac{d^3\varphi}{d\omega^3} = -\frac{\lambda^4}{4\pi^2c^3}\left(3\frac{d^2P}{d\lambda^2} + \lambda\frac{d^3P}{d\lambda^3}\right)$$

$$\frac{d^2P}{d\lambda^2} = 4l\sin\beta\left[\frac{d^2n}{d\lambda^2} + (2n - 1/n^3)\left(\frac{dn}{d\lambda}\right)^2\right] - 8l\cos\beta\left(\frac{dn}{d\lambda}\right)^2$$

$$\frac{d^2P}{d\lambda^3} = l\cos\beta\left[\left(24/n^3 - 48n\right)\left(\frac{dn}{d\lambda}\right)^3 - 24\frac{dn}{d\lambda}\frac{d^2n}{d\lambda^2}\right] + l\sin\beta\left[\begin{array}{l}\left(\frac{dn}{d\lambda}\right)^3\left(12/n^6 + 12/n^4 + 8/n^3 - 16/n^2 + 32n\right) \\ +\left(24n - 12/n^3\right)\frac{dn}{d\lambda}\frac{d^2n}{d\lambda^2} + 4\frac{d^3n}{d\lambda^3}\end{array}\right]$$

Double-pass grating-pair, separation l, angle of incidence γ, line spacing d



$$\frac{d^2\varphi}{d\omega^2} = l\frac{\lambda^3}{\pi c^2 d^2}\left[1 - \left(\frac{\lambda}{d} - \sin\gamma\right)^2\right]^{-3/2}$$

$$\frac{d^3\varphi}{d\omega^3} = -6l\frac{\lambda^4}{c^3d^2}\left(1 + \frac{\lambda}{d}\sin\gamma - \sin^2\gamma\right)\left[1 - \left(\frac{\lambda}{d} - \sin\gamma\right)^2\right]^{-5/2}$$

corrected using an equal but opposite amount of negative second-order dispersion. The combination of spectral broadening using SPM and subsequent chirp-removal using linear dispersion is the basis for optical pulse compression techniques that have been applied effectively to create pulses which exist for only a few carrier-wave periods.

When the pulse propagates in an environment where SPM and linear dispersion are well balanced and evenly distributed, a soliton will evolve which, in the absence of loss, can propagate indefinitely without changing shape. The nonlinear Schrödinger equation predicts that lasers in which the dominant pulse-shaping mechanisms are negative

group-velocity dispersion and self-phase-modulation will produce soliton-like pulses with a $\mathrm{sech}^2(t)$ intensity profile. In experimental measurements of ultrafast lasers the pulse shape is commonly assumed to be $\mathrm{sech}^2(t)$ and the chirp of the pulses is then estimated by measuring their duration–bandwidth product, $\Delta\nu\Delta\tau$, which is a dimensionless quantity that can be treated as a figure-of-merit for ultrashort pulses. For unchirped pulses the duration–bandwidth product is independent of their duration or spectral bandwidth but is sensitive to their shape: chirp-free Gaussian pulses have $\Delta\nu\Delta\tau = 0.441$ while $\mathrm{sech}^2(t)$ pulses have $\Delta\nu\Delta\tau = 0.315$.

Sources of Femtosecond Pulses

Dye Lasers

The first practical passively mode-locked ultrafast lasers were those based on a gain medium of a thin ($\sim$10 μm) organic dye jet and incorporating a slow saturable absorber jet at a separate cavity focus. The most common combination was Rhodamine 6G in ethylene glycol (gain jet) and 3,3-diethyloxadicarbocyanine iodide in ethylene glycol (absorber jet) which produced pulses at a wavelength of 620 nm. The most successful approach was colliding-pulse mode-locking in which two counter-propagating pulses within a ring cavity are synchronized to arrive simultaneously at the absorber jet, so achieving greater saturation of the absorber and allowing the pulses to be transmitted with a loss significantly smaller than in the single-pulse case. The first sub-100-fs optical pulses were reported in 1981 from a colliding-pulse mode-locking dye laser and later work reported pulse durations as low as 27 fs but the configuration was inefficient and produced low average output powers. Other approaches such as synchronous pumping and hybrid mode-locking produced higher output powers but the maintenance-intensive nature of dye lasers and the hazardous nature of many of the chemicals used led to their gradual replacement by more convenient solid-state alternatives, based principally on Ti:sapphire.

Color-Center Lasers

A new source of femtosecond pulses in the near-infrared was reported in 1984. Coined the soliton laser the system was a synchronously pumped KCl:Tl color-center laser tunable from 1.4 to 1.6 μm and modified to allow direct generation of pulses as short as 50 fs, substantially shorter than the 8 ps achieved in synchronously pumped operation alone. Femtosecond operation was achieved by coupling a length of polarization-preserving anomalously dispersive fiber to the main laser cavity using a beamsplitter and retroreflector arrangement. This coupled-cavity arrangement later became known as additive pulse mode-locking (APM) and produced short pulses when the optical length of the fiber was adjusted to be an integral multiple of the main cavity length. The long upper-state lifetime of the laser-active centers in the KCl:Tl gain medium allowed particularly stable APM operation, but mode-locking of this kind was also successfully employed in an actively stabilized $NaCl:OH^-$ color-center laser to produce 110-fs pulses. The need to maintain color-center crystals at liquid-nitrogen temperatures limited the convenience of these lasers but until the advent of Ti:sapphire they remained the only solid-state sources capable of femtosecond operation.

Ti:sapphire Lasers

The creation in 1986 of a new broadband laser material – titanium-doped sapphire or Ti:sapphire – began a revolution in ultrafast laser sources that still continues today. Unlike other contemporary solid-state materials for mode-locked lasers such as Nd:YAG, Ti:sapphire is a vibronic gain medium which exhibits a strong electron-phonon coupling between the titanium ion transition and the host lattice. This vibronic coupling leads to a broad continuum of possible transition energies and results in a fluorescence bandwidth covering 670–1050 nm, making Ti:sapphire capable of supporting extremely short femtosecond pulses. The long fluorescence lifetime of Ti:sapphire means that, unlike dye lasers, femtosecond mode-locking requires a passive method using a fast saturable absorber and two methods have prevailed and are discussed below.

The first passively mode-locked Ti:sapphire lasers were based on the APM technique in which mode-locking is achieved by a fast-saturable absorber action due to interference between the intracavity pulse and a self-phase-modulated replica propagating in an auxiliary cavity. An important breakthrough was made in 1990 when it was reported that an APM Ti:sapphire laser continued to operate even when the auxiliary cavity was blocked. This discovery was attributed to an entirely new mode-locking effect known as Kerr-lens mode-locking (KLM) or self-mode-locking in which the presence of a Kerr lens within the gain medium during mode-locked operation changes the mode focusing within the cavity so that, compared to cw operation, mode-locked pulses experience higher gain. Two configurations of KLM exist and are known as soft aperture mode-locking, where the presence of the Kerr lens increases the gain by improving the overlap between the pump and laser

modes, and hard-aperture mode-locking, in which a physical aperture such as a slit or the edge of a prism is adjusted to introduce greater loss for cw operation. Because KLM is based on a nonresonant nonlinearity, Ti:sapphire lasers can be mode-locked at any wavelength in their gain bandwidth using this technique. Recent advances in controlling the intracavity group-velocity dispersion characteristics have led to the generation of sub-5-fs pulses directly from a Ti:sapphire oscillator. The principal drawback of the KLM method is that it is not self-starting because mode-locked operation must be seeded by an intense noise spike or other short fluctuation. Different starting methods have been applied successfully including mirror tapping, acousto-optic modulation, and mode-dragging, and once mode-locking has been initiated it is generally stable until some external perturbation disturbs the intracavity beam.

An alternative method for obtaining femtosecond pulses has been pioneered largely in parallel with the KLM efforts and is based on using resonant nonlinearities in semiconductors as fast saturable absorbers. The approach is to fabricate a semiconductor multiple-quantum-well (MQW) device whose bandgap has been chosen to match the laser wavelength and whose absorption can be saturated at high fluences. Initially these devices were utilized in an APM geometry in which the MQW device was situated in an auxiliary cavity but it was later realized that an equivalent monolithic configuration could be obtained by sandwiching the MQW layer between two high-reflectivity mirrors (see Figure 3, inset) to form a Fabry–Perot structure and the resulting element was known as an antiresonant Fabry–Perot saturable absorber. The first application of a such a semiconductor saturable absorber mirror (SESAM) for mode-locking Ti:sapphire was reported in 1995 where self-starting operation was demonstrated. A typical modern Ti:sapphire laser design is shown in Figure 3 and illustrates how a saturable absorber mirror can be integrated into the cavity. Further design enhancements such as reducing the finesse of the Fabry–Perot cavity resulted in SESAMs with a broader wavelength response which enabled sub-10-fs self-starting operation.

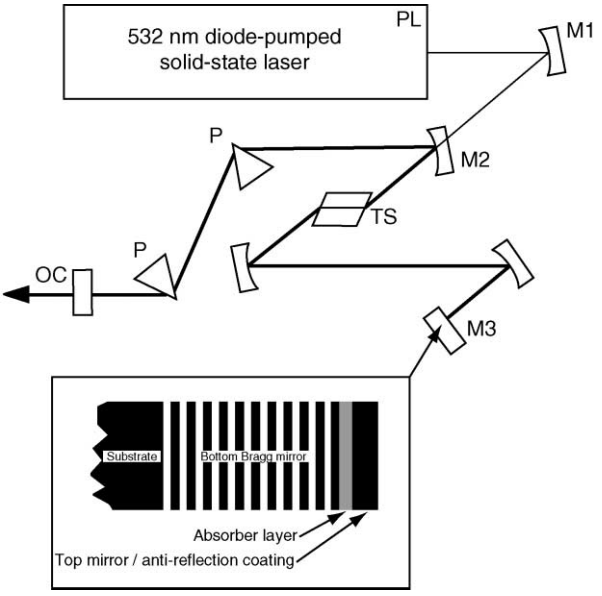


Figure 3 Typical configuration of a modern femtosecond Ti:sapphire laser. PL, pump laser; M1, pump focusing mirror; M2, high-reflectivity cavity mirror (800 nm); TS, Ti:sapphire crystal, typically, 2–10 mm long; M3, SESAM mirror; P, dispersion-compensating prisms; OC, output coupler. Inset: a simplified schematic of a saturable-absorber Bragg mirror.

Fiber Lasers

Ultrafast fiber lasers, particularly those based on silica doped with rare earth Er^{3+}, Nd^{3+}, or Yb^{3+}ions, have enjoyed renewed attention in recent years as efforts to scale their average output powers and pulse energies to practical levels have proved successful. Four main mode-locking strategies have been applied to these systems: active mode-locking which is generally restricted to picosecond generation; intensity-dependent feedback using a nonlinear amplifying loop mirror; polarization APM; and semiconductor saturable absorber mode-locking. The upper-state lifetime of rare-earth-doped fibers is long and therefore a fast saturable absorber mechanism is necessary for passive mode-locking, limiting the choice of mode-locking elements to those based on Kerr or semiconductor nonlinearities.

The development of mode-locked fiber sources has concentrated on Er-doped lasers because of the compatibility of their output wavelengths with the 1.5 μm optical communication window in standard silica fiber. Passively mode-locked designs based on the nonlinear loop mirror work by exploiting the Kerr effect in an intracavity fiber Sagnac interferometer containing a gain section. With suitable control of the intracavity polarization, such a nonlinear amplifying loop mirror (NALM) can be set up to transmit light of high intensities but reflect low intensities because of the differential phase shift induced between the two interferometer arms. By adding the loop mirror to a unidirectional ring cavity a figure-of-eight laser is formed (Figure 4) which is capable of producing femtosecond soliton pulses with average powers of around 1 mW. Drawbacks of the figure-of-eight laser include the presence of spectral sidebands at the shortest pulse durations and a tendency for multiple pulse operation. Mode-locking can also be achieved

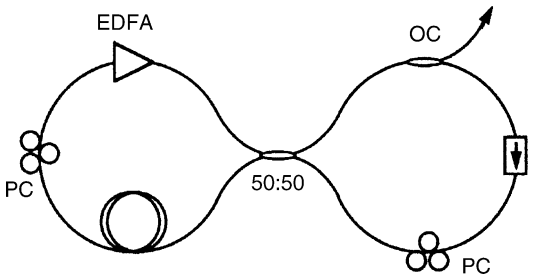


Figure 4 Schematic of a figure-of-eight femtosecond fiber laser. EDFA, erbium-doped fiber amplifier gain section; OC, output coupler; PC, polarization controller.

using a variation on the APM technique and significant average output powers have been demonstrated from stretched-pulse-APM lasers, with up to 90 mW being reported at pulse energies of 2.25 nJ.

Semiconductor saturable absorber mode-locking has been demonstrated in fiber lasers using a variety of configurations including bulk InGaAsP on InP and saturable Bragg reflectors based on InGaAs/InP MQWs on a AlAs/GaAs mirror structure. Fiber lasers mode-locked using semiconductor saturable absorbers are self-starting and exhibit stable mode-locking, normally at the fundamental cavity frequency. The narrow bandwidth of saturable absorber devices at near-infrared wavelengths makes the generation of sub-100-fs pulses difficult.

Semiconductor Sources

Diode lasers and external-cavity lasers based on semiconductor optical amplifiers (SOAs) are of importance because of their ability to produce ultrashort pulses from a compact system at wavelengths of relevance to optical fiber communications. The two principal methods used to generate ultrashort pulses from such devices are mode-locking with an external cavity or gain-switching (equivalently known as Q-switching) in a bow-tie waveguide structure. Subpicosecond pulse durations are now routinely produced using dispersion compensation and/or nonlinear pulse compression techniques and, by combining these strategies with Er:fiber amplification, 20-fs pulses have been produced. The mode-locking and gain-switching techniques each have their individual advantages. Gain-switching provides a simple and compact approach offering high average output power, but significant timing jitter and poor spectral quality can accompany pulses produced in this way. Mode-locked systems are physically larger than gain-switched devices because they require an external cavity and additional optical components, but shorter and higher quality pulses are achievable.

Other Common Solid-State Laser Sources

Certain solid-state gain media offer the potential for direct diode pumping, miniaturization and operation at wavelengths or average powers not accessible with other sources. Systems based on Nd or Cr ion transitions have been shown to be capable of ultrashort picosecond or femtosecond operation. Femtosecond mode-locking (175 fs) has been reported in Nd:glass which has one of the broadest fluorescence spectra of the Nd materials. Crystals with active ions of Cr^{2+}, Cr^{3+}, and Cr^{4+} have broad transitions covering wavelengths in the 2.0 μm, 800 nm and 1.4 μm regions and femtosecond operation has been demonstrated in Cr^{4+}:YAG at 1.52 μm, Cr^{4+}:Mg_2SiO_4 (Cr:forsterite) at 1.25 μm, Cr^{3+}:$LiSrAlF_4$ (Cr:LiSAF) at 850 nm, and Cr^{3+}:$LiSrGaF_6$ (Cr:LiSGAF). Direct diode pumping of Cr-ion-based lasers is attractive because of the opportunities for producing low-noise, compact, and efficient femtosecond sources, but the limited power available from high-quality pump diodes results in low intracavity powers which make KLM operation difficult. As a result, stable operation is better achieved using a physical saturable absorber, and SESAM mode-locking has been applied successfully to many of these lasers.

Sources Based on Nonlinear Frequency Conversion

Laser sources alone are unable to provide ultrashort pulses continuously tunable from the ultraviolet to the mid-infrared, and large gaps exist at wavelengths where no broadband laser materials are available. Harmonic generation using Ti:sapphire can be used to access some parts of the visible and ultraviolet regions, but infrared and certain visible wavelengths cannot be generated in this way and coverage is limited by the ability to tune the source itself. The nonlinear production of infrared wavelengths requires parametric down-conversion using a synchronously pumped optical parametric oscillator. Ti:sapphire pumped synchronously pumped optical parametric oscillators have produced, either directly or using harmonic generation, outputs from the green to beyond 7 μm.

Sources of Amplified Ultrashort Pulses

Modern ultrafast oscillators can produce pulses with peak powers of >1 MW, average powers of several watts, and cavity frequencies from 4 MHz to 2 GHz, but many applications require pulses with higher intensities or lower pulse repetition frequencies. Additional amplification stages can be used to satisfy these requirements and these are commonly

implemented in Nd:glass for picosecond pulses and Ti:sapphire for femtosecond pulses. In order to avoid damage and unwanted nonlinear effects caused by high pulse intensities within the amplifier system, most practical sources now apply the chirped-pulse amplification (CPA) approach in which a low-energy femtosecond seed pulse is stretched to sub-nanosecond durations to reduce its peak power then amplified and finally compressed to its original duration. Using the chirped-pulse amplification technique, pulses with peak powers as high as 1.5 petawatt have been generated.

Two possible amplifier geometries are used which are categorized either as multipass or regenerative amplifiers. In the multipass configuration the injected pulse is refocused many times through the same gain crystal using a system of mirrors. The gain crystal is pumped above the saturation fluence in order to extract the maximum stored energy from the medium per pass. The multipass geometry is more commonly used for high-average-power operation and critical alignment is required to maximize the overlap between the pump mode profile and the multiple beams intersecting the gain medium. The alternative regenerative amplifier is essentially a stable laser resonator which includes a Pockel cell and a polarizing beamsplitter to enable pulses to be switched into and out of the cavity. The gain per pass is considerably lower than in the multipass configuration but regenerative systems produce superior beam quality and their performance can be optimized without affecting the beam pointing of the output. Commercially available Ti:sapphire amplifiers commonly use the regenerative method and are available with pulse energies of up to 1 mJ at repetition frequencies in the 1–300 kHz range.

Further Reading

Agrawal GP (1995) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Duling IN (1995) *Compact Sources of Ultrashort Pulses*. Cambridge University Press.

Fork RL, Martinez OE and Gordon JP (1984) Negative dispersion using pairs of prisms. *Optics Letters* 9: 150–152.

Keller U, Weingarten KJ, Kartner FX, *et al*. (1996) Semiconductor saturable absorber mirrors for femtosecond to nanosecond pulse generation. *Solid-State Lasers, IEEE Selected Topics in Quantum Electronics* 2: 435–453.

Siegman AE (1986) *Lasers*. University Science Books.

Spence DE, Kean PN and Sibbett W (1991) 60-fsec pulse generation from a self-mode-locked Ti-sapphire laser. *Optics Letters* 16: 42–44.

Szipocs R, Ferencz K, Spielmann C and Krausz F (1994) Chirped multilayer coatings for broad-band dispersion control in femtosecond lasers. *Optics Letters* 9: 201–203.

Yariv A (1989) *Quantum Electronics*. New York: Wiley.

Pulse Characterization Techniques

D J Kane, Southwest Sciences, Inc., Santa Fe, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Ultrashort optical pulses have very short time durations, typically less than a few tens of picoseconds. As a result, these pulses are spectrally broad. Because the index of refraction of materials is a function of wavelength, different wavelengths of light travel at different speeds in optical materials, causing the properties of ultrashort optical pulses to change as they propagate. However, the shape of the pulse can influence how the pulse itself interacts with materials. Thus, the goal of ultrafast laser pulse measurement is to obtain not only the intensity profile of the pulse, but also the actual variation of the frequencies that make up the pulse. This is called measuring the intensity and phase of the pulse, respectively.

Even though the development of techniques to characterize ultrashort optical pulses has not been easy, a myriad of pulse measurement techniques have been developed. In this chapter, we will confine our discussion to the most well-known pulse and generally accepted pulse characterization methods in an effort to provide a basic working knowledge of pulse measurement techniques and a fundamental understanding of the principles behind pulse measurement. After a brief discussion of the mathematical representation of ultrashort optical pulses, we will discuss pulse measurement in (roughly) chronological order, starting with autocorrelation methods. The next section introduces a useful method for the measurement of the relative phase between two pulses called spectral interferometry (SI). Following SI, we introduce the notion of the time-frequency

representation of ultrashort optical pulses where techniques such as frequency-resolved optical gating (FROG) are discussed. Following the time-frequency section, a relatively new technique known as spectral phase interferometry for direct electric field reconstruction (SPIDER) is discussed.

Mathematical Representation of an Optical Pulse

The time-dependent variations of the pulse are embodied in the pulse electric field, $A(t)$, which can be written:

$$A(t) = \mathrm{Re}[E(t)\mathrm{e}^{\mathrm{i}\omega_0 t}] \quad [1]$$

where ω_0 is the carrier frequency and Re refers to the real part. While we can use $A(t)$ as it stands for calculations, it is much easier to remove the rapidly varying ω_0 part, $\mathrm{e}^{\mathrm{i}\omega_0 t}$, and use a slowly varying envelope together with a phase term that contains only the frequency variations, not the rapidly varying carrier frequency:

$$E(t) = [I(t)]^{1/2}\mathrm{e}^{\mathrm{i}\varphi(t)} \quad [2]$$

where $I(t)$ and $\varphi(t)$ are the time-dependent intensity and phase of the pulse. (Note that $E(t)$ is complex.) The frequency variation, $\Omega(t)$, is the derivative of $\varphi(t)$ with respect to time:

$$\Omega(t) = -\mathrm{d}\varphi(t)/\mathrm{d}t \quad [3]$$

The pulse field can be written equally well in the frequency domain by taking the Fourier transform of eqn [2]:

$$\tilde{E}(\omega) = [\tilde{I}(\omega)]^{1/2}\mathrm{e}^{-\mathrm{i}\phi(\omega)} \quad [4]$$

where $\tilde{I}(\omega)$ is the spectrum of the pulse, and $\phi(\omega)$ is its phase in the frequency domain. The spectral phase contains time versus frequency information; that is, the derivative of the spectral phase with respect to frequency yields the time arrival of the frequency – the group delay.

Obtaining the intensity and phase, $I(t)$ and $\varphi(t)$ (or $\tilde{I}(\omega)$ and $\phi(\omega)$ is called full characterization of the pulse. Common phase distortions include linear chirp, where the phase (either the time domain or frequency domain) is parabolic. When the frequency is increasing in time, the pulse is said to have positive linear chirp; negative linear chirp is when the high frequencies lead the lower frequencies. Higher-order chirps are common, but for these, differentiation between spectral and temporal chirp is required because spectral phase and temporal phase are not interchangeable.

Autocorrelation

Traditionally, we measure events using shorter events. Unfortunately, for the ultrafast researcher, shorter events do not exist and modern electronics are not fast enough. Therefore, because the shortest event we have is the event we wish to measure, traditional pulse measurement methods use the pulse itself to determine the approximate duration of the pulse in question giving birth to the ubiquitous intensity autocorrelation (see **Figure 1**). While autocorrelations cannot be used to fully characterize ultrashort optical pulses, the methods used in autocorrelations are fundamental to all pulse measurement schemes.

The intensity autocorrelation is measured by combining a pulse and a delayed replica of a pulse in a nonlinear medium such as a second harmonic generation (SHG) crystal. A pulse is sent onto a beamsplitter to produce the two replicas. One pulse is delayed relative to the other and both are focused together into a SHG crystal. As the delay of one pulse relative to the other is varied, the intensity of the second-harmonic signal is recorded. The intensity autocorrelation is not limited to using SHG; any nonlinearity may be utilized. Indeed, two-photon absorption in a semiconductor LED or photodiode often acts as a convenient nonlinearity. Sometimes a third-order nonlinearity is used to provide some direction of time information about the pulse. When the generation of the signal involves phase matching, such as second harmonic generation, care must be taken to use a thin crystal. Typically, 10 μm to 1 mm thick SHG crystals are used, depending on the bandwidth of the pulse to be measured. The exact thickness depends on the SHG crystal, the pulse width, and the requirements of the measurement.

Regardless of the nonlinearity used, an autocorrelation yields only the approximate duration and

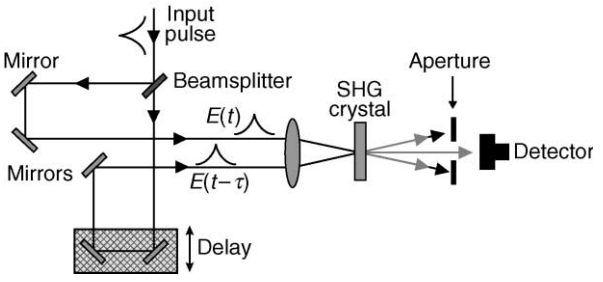


Figure 1 An intensity autocorrelator used to determine the approximate duration of a pulse. The pulse is split into two replicas that are sent into delay lines. The outputs from both delay lines are focused into a doubling crystal. The second harmonic output is monitored as a function of delay between the two pulses. A plot of the signal versus time is the intensity autocorrelation of the pulse.

shape of the pulse. The structure of the pulse is smeared, producing a smooth, featureless profile for even a complex pulse. Thus, the intensity autocorrelation alone does not determine the intensity profile of the pulse, $I(t)$. Only the interaction of the intensity of the pulse is recorded, producing no phase information for the pulse. Some qualitative information can be gleaned from the spectrum and the autocorrelation, but usually only that the pulse is chirped, and higher-order chirps and complex pulse structures elude such an analysis.

Another useful form of intensity autocorrelation is the single-shot autocorrelator. In this case, fairly large beams are used, and these beams are set to intersect at an angle of 2Φ, tilting the pulse fronts, so that delay is mapped onto a spatial coordinate (see **Figure 2**). The beams are then focused into a nonlinear medium using a cylindrical lens. The interaction region in the SHG crystal is imaged onto a linear array or CCD camera. Typically, when second-harmonic generation is used, the SHG crystal is oriented for Type I phase matching. The time window is proportional to $\Delta w(\sin \Phi)/\Delta v_g$, where Δw is the beam waist at the SHG crystal, and v_g is the group velocity of the pulse in the crystal. Like all SHG interactions, care must be exercised to insure that the phase matching is sufficient to mix the entire bandwidth of the pulse.

Jean-Claude Diels improved the intensity autocorrelation by the development of the interferometric autocorrelation. In this configuration, a Michelson interferometer is typically used so that the beams are collinear when arriving at the SHG crystal, which allows the beams to interfere. As a result, fringes appear on the autocorrelation. By examining the shape and extent of the fringes, some information about the pulse chirp can be obtained. While this does add to the information provided by an intensity autocorrelation, the added information is only the spectrum of the second harmonic. Unfortunately the addition of the second harmonic spectrum does not provide enough information for full retrieval of the intensity and phase.

Even though autocorrelation alone does not completely determine pulse intensity and phase, it is a very simple and useful technique to determine approximate pulse duration. Sometimes, cross-correlation can be used to determine the duration of a long pulse, if a shorter one is available, or a pulse too weak (or at an inappropriate wavelength) to measure with autocorrelation. The addition of the spectrum and/or the second harmonic spectrum can provide more,

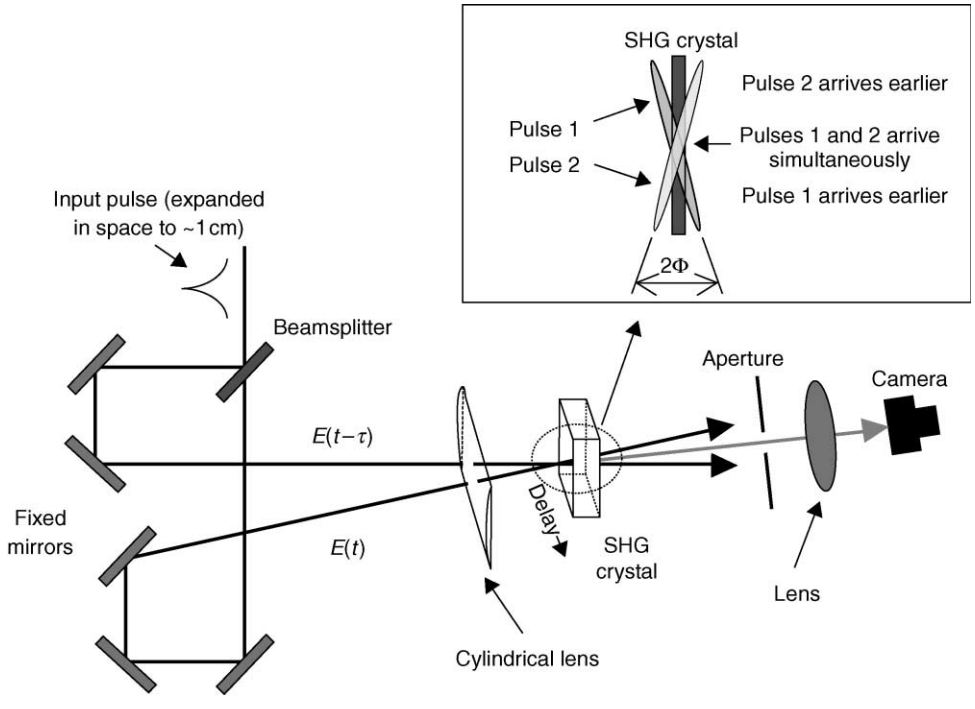


Figure 2 A single shot intensity autocorrelator. For this type of autocorrelator, the beam size is large, on the order of 1 cm in diameter, and the pulse fronts are tilted with respect to each other in order to map relative delay to position. The two beams are focused using a cylindrical lens and the output is recorded using an array detector. The inset shows how delay is mapped into position. Inset: the tilted pulse fronts cause time delay to be mapped spatially.

albeit, incomplete information about pulse chirp. More importantly, the experimental techniques used by correlation methods are fundamental to all pulse measurement methods.

Spectral Interferometry – Relative Phase Measurements

While determination of the absolute intensity and phase of an ultrashort laser pulse is difficult, determination of the relative phase is not. Invented in the late 1800s, spectral interferometry (SI), measures the relative phase between two pulses. SI, a linear technique, can also be modified (albeit, a some-what complex modification) to be self-referencing for full characterization of the ultrashort optical pulse (See the section on SPIDER below). In addition, when combined with a full pulse characterization technique, SI, because it is a linear technique, provides a method by which very weak optical pulses can be characterized (see below).

Typically, an SI apparatus uses a Mach–Zender interferometer (see Figure 3). A pulse is split into two replicas; one is sent through a region or medium to measure and the other through a known path. The two pulses remain separated by some known time and are sent, collinearly, into a spectrometer; no time scanning is required. An analysis of the fringe pattern yields the relative phase between the two pulses. In other words, SI provides $\phi_{\text{unk}}(\omega) - \phi_{\text{ref}}(\omega)$, where $\phi_{\text{unk}}(\omega)$ is the unknown pulse phase versus frequency. The spectrum of the two pulses in the frequency domain is

$$\tilde{I}_{\text{SI}}(\omega) = |\tilde{E}_0(\omega) + \tilde{E}_{\text{unk}}(\omega)|^2$$

$$\tilde{I}_{\text{SI}}(\omega) = \tilde{I}_0(\omega) + \tilde{I}_{\text{unk}}(\omega) + 2\sqrt{\tilde{I}_0(\omega)}\sqrt{\tilde{I}_{\text{unk}}(\omega)} \times \cos(\phi_{\text{unk}}(\omega) - \phi_0(\omega) - \omega\tau) \quad [5]$$

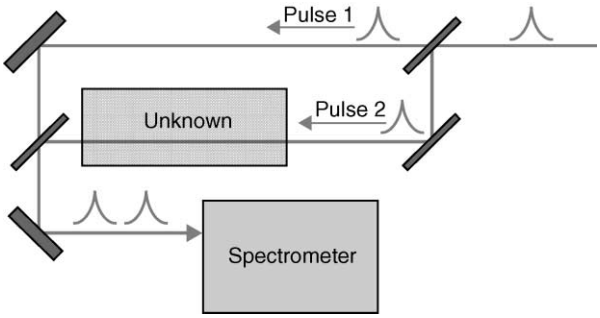


Figure 3 A Mach–Zender interferometer is the typical arrangement for spectral interferometry. One path acts as the reference arm and the sample to be measured is placed in the other arm. Unlike standard interferometry, the delay is fixed in a spectral interferometry experiment.

where $\tilde{I}_{\text{SI}}(\omega)$ is the output spectrum, $\tilde{E}_0(\omega)$ is the reference pulse electric field, $\tilde{E}_{\text{unk}}(\omega)$ is the unknown pulse electric field, $\tilde{I}_0(\omega)$ is the reference pulse spectrum, $\tilde{I}_{\text{unk}}(\omega)$ is the unknown pulse spectrum, $\phi_{\text{unk}}(\omega)$ is the unknown pulse phase, $\phi_0(\omega)$ is the reference pulse phase, and τ is the delay between the two pulses. To facilitate extracting the phase difference from the SI output, τ is chosen to yield fringes in the sum spectrum. The spectral fringes, which have a period inversely proportional to the optical path difference between the two beams (see Figure 4), contain all of the phase difference information.

Figure 4 shows the steps required to obtain the phase difference between the two pulses. The first step is to subtract out the spectra of the individual pulses in order to isolate the spectral interferogram, $S(\omega)$ (Figure 4a), where

$$S(\omega) = 2\sqrt{\tilde{I}_0(\omega)}\sqrt{\tilde{I}_{\text{unk}}(\omega)}\cos(\phi_{\text{unk}}(\omega) - \phi_0(\omega) - \omega\tau) \quad [6]$$

By Fourier transforming $S(\omega)$, we obtain

$$\Im^{-1}[S(\omega)] = f(t-\tau) + f(-t-\tau) \quad [7]$$

where $f(t)$ is the correlation product between the reference and the unknown electric field (see Figure 4b). The Fourier transform of eqn [7] multiplied by a Heaviside function, $\Theta(t)$ (to remove $f(-t-\tau)$), recovers the amplitude and phase of $f(\omega)$, the Fourier transform of $f(t-\tau)$, which contains the relative spectral phase between the unknown pulse and the unmodified pulse. Some care, however, needs to be taken to correctly remove the linear phase due to the time delay between the two pulses, $\omega\tau$. Also, for best results, the spectrometer should be well calibrated and care should be taken to properly window the spectrum before taking the Fourier transform.

Spectral interferometry is a straightforward and simple technique that always provides a relative phase. For this very reason, care must be taken to make sure the beams are mode-matched as well as collinear; also, interferometric stability must be maintained over the course of the measurement. Furthermore, SI is nongating; that is, CW background in the laser can add to the interferogram, masking transient effects. Nevertheless, spectral interferometry is a useful technique that has been applied to the measurement of the linear and nonlinear spectral phase introduced by optical fibers. More recently, SI has been applied to phase-locking and to phase-resolved pump probe experiments.

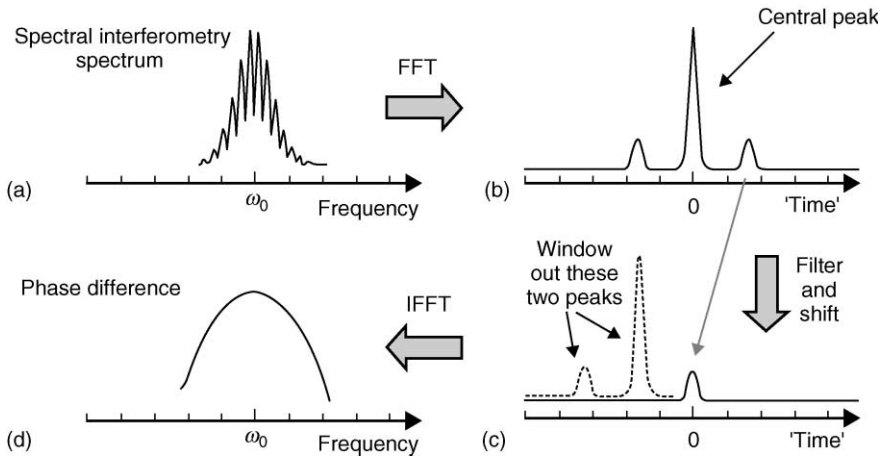


Figure 4 The steps required for the analysis of spectral interferograms. Part (a) shows a sample SI interferogram. The dominant frequency of the fringes is the delay multiplied by the speed of light. The actual phase differences of interest are the perturbations on this frequency. Part (b) is the Fourier transform of the SI interferogram in Part (a). (Note that this is not the true time domain.) The central peak contains only spectral information. All of the phase information is contained in the satellite peaks. The next step is to mask out the central peak and one of the satellite peaks (Part (c)). The phase of the inverse Fourier transform of Part (c) yields the relative phase between the two pulses (Part (d)).

Time-Frequency Representation

In 1971, E.B. Treacy laid the foundations for a revolution in pulse measurement by introducing the idea of measuring the intensity versus time for different spectral splices of a pulse. Because his measurements provided both time and frequency information simultaneously, these measurements could be thought of as applying in a hybrid time-frequency domain. At first, this may seem confusing, but similar methods have been used to visualize sounds patterns (a musical score, for example) and speech. A time-frequency plot of this type, where spectral slices of a pulse are plotted versus time, is called a sonogram. The mathematical formalism of a sonogram is:

$$S_{\mathrm{E}}(\Omega, T) = \left| \int_{-\infty}^{\infty} \tilde{E}(\omega) h(\omega - \Omega) \mathrm{e}^{-\mathrm{i}\omega T} \mathrm{d}\omega \right|^2 \qquad [8]$$

where $\tilde{E}(\omega)$ is the electric field of the pulse to be measured in the frequency domain and $h(\omega - \Omega)$ is a frequency gate that varies with frequency. The magnitude squared of the inverse Fourier transform of spectral slices yields the sonogram.

In 1991, Chilla and Martinez showed that the sonogram could be used to reconstruct the full intensity and phase of the pulse. That is, under certain conditions, the approximate group delay could be determined as a function of frequency from the sonogram by finding the peak time arrival of each spectral slice; integration of the group delay yields the spectral phase, which together, with the pulse spectrum, provides the intensity and phase of the pulse in the frequency domain. They labelled this technique as frequency domain phase measurement (FDPM).

An experimental diagram of an FDPM apparatus is shown in Figure 5. The pulse is split into two replicas, and one of the pulse replicas is spectrally filtered. The spectrally filtered pulse is cross-correlated with the original pulse to find the 'time arrival' of the spectrally filtered pulse – defined as the peak of the cross-correlated pulse. Because the spectrally filtered pulse has a much longer time duration than the original pulse, the small perturbation caused by the finite length of the original pulse is neglected.

The main difficulty of the Chilla–Martinez method is that the frequency gate must be very narrow, reducing the filtered pulse energy significantly, which greatly reduces the measured signal strength. If the spectral phase is not well behaved, the filtered pulse may not be much longer than the original pulse. In addition, for a given frequency, the sonogram may have two or more peaks in time, that is, the group delay may not be a function. However, a method called 2D phase retrieval, discussed in the next section, can alleviate many of these issues.

A spectrogram is a relative of the sonogram. Rather than determining the time arrival of spectral slices of a pulse, a spectrogram is obtained when the spectrum of time slices of the pulse are measured. The spectrogram is experimentally much easier to obtain than the sonogram; however, the measured quantity is not

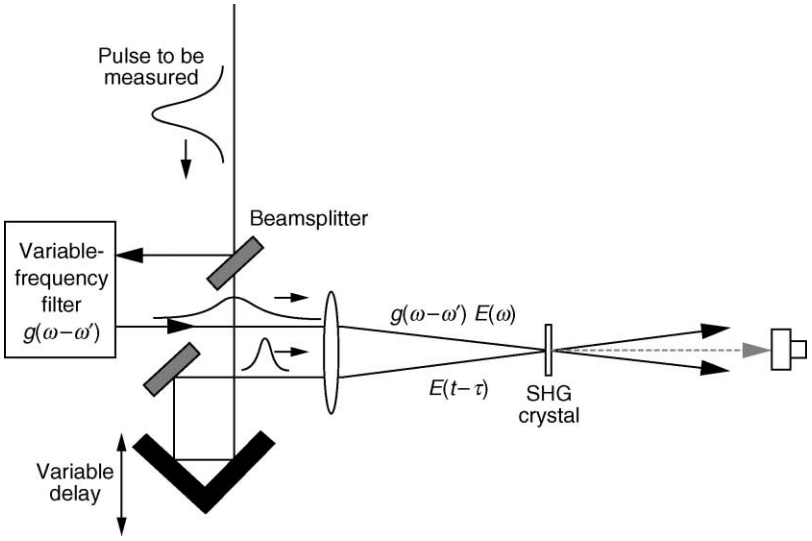


Figure 5 Measuring a sonogram requires determining the time arrival of spectral slices of the pulse. The pulse is split into two replicas using a beamsplitter. One replica is spectrally filtered using a tunable filter. The other pulse is cross-correlated with the spectrally filtered pulse to determine the time arrival.

quite as mathematically useful as the sonogram because the time slicing of the pulse occurs in the time domain. Thus, if a narrow time gate is used, the time domain phase is obtained. The time domain phase together with the intensity of the pulse, $I(t)$, provides the full intensity and phase of the pulse. Unfortunately, the intensity profile is not as readily measurable as the pulse spectrum. Consequently, to obtain the full intensity and phase of a pulse from its spectrogram requires an iterative 2D phase retrieval algorithm. This is the basis of a pulse measurement technique called frequency-resolved optical gating (FROG).

Frequency-Resolved Optical Gating

Frequency-resolved optical gating (FROG), developed by Kane and Trebino, measures the spectrum of a particular temporal component of the pulse (see Figures 6 and 7) by spectrally resolving the signal pulse in an autocorrelation-type experiment using an instantaneously responding nonlinear medium. As shown in Figure 6, FROG involves splitting a pulse and then overlapping the two resulting pulses in an instantaneously responding $\chi^{(3)}$ or $\chi^{(2)}$ medium. Even though any instantaneous nonlinear interaction may be used to implement FROG, perhaps the most intuitive is the polarization-gating configuration. In this case, induced birefringence, due to the electronic Kerr effect, is used as the nonlinear-optical process. In other words, the 'gate' pulse causes the $\chi^{(3)}$ medium, which is placed between two crossed polarizers, to become slightly birefringent. The polarization of the 'gated' probe pulse (which is cleaned up by the first polarizer) is rotated slightly by the induced birefringence allowing some of the 'gated' pulse to leak through the second polarizer. This is referred to as the signal. Because most of the signal emanates from the region of temporal overlap between the gate pulse and the probe pulse, the signal pulse contains the frequencies of the 'gated' probe pulse within this overlap region. The signal is then spectrally resolved, and the signal intensity is measured as a function of wavelength and delay time τ. The resulting trace of intensity versus delay and frequency is a spectrogram, a time- and frequency-resolved transform that intuitively displays time-dependent spectral information of a waveform.

The spectrogram can be expressed as:

$$S_{\mathrm{E}}(\omega,\tau) = \left| \int_{-\infty}^{\infty} E(t) g(t-\tau) \mathrm{e}^{-\mathrm{i}\omega t} \mathrm{d}t \right|^2 \qquad [9]$$

where $E(t)$ is the measured pulse's electric field, $g(t-\tau)$ is the variable-delay gate pulse, and the subscript E on S_{E} indicates the spectrogram's dependence on $E(t)$. The gate pulse $g(t)$ is usually somewhat shorter in length than the pulse to be measured, but not infinitely short. This is an important point: an infinitely short gate pulse yields only the intensity $I(t)$ and conversely, a CW gate yields only the spectrum $I(\omega)$. On the other hand,

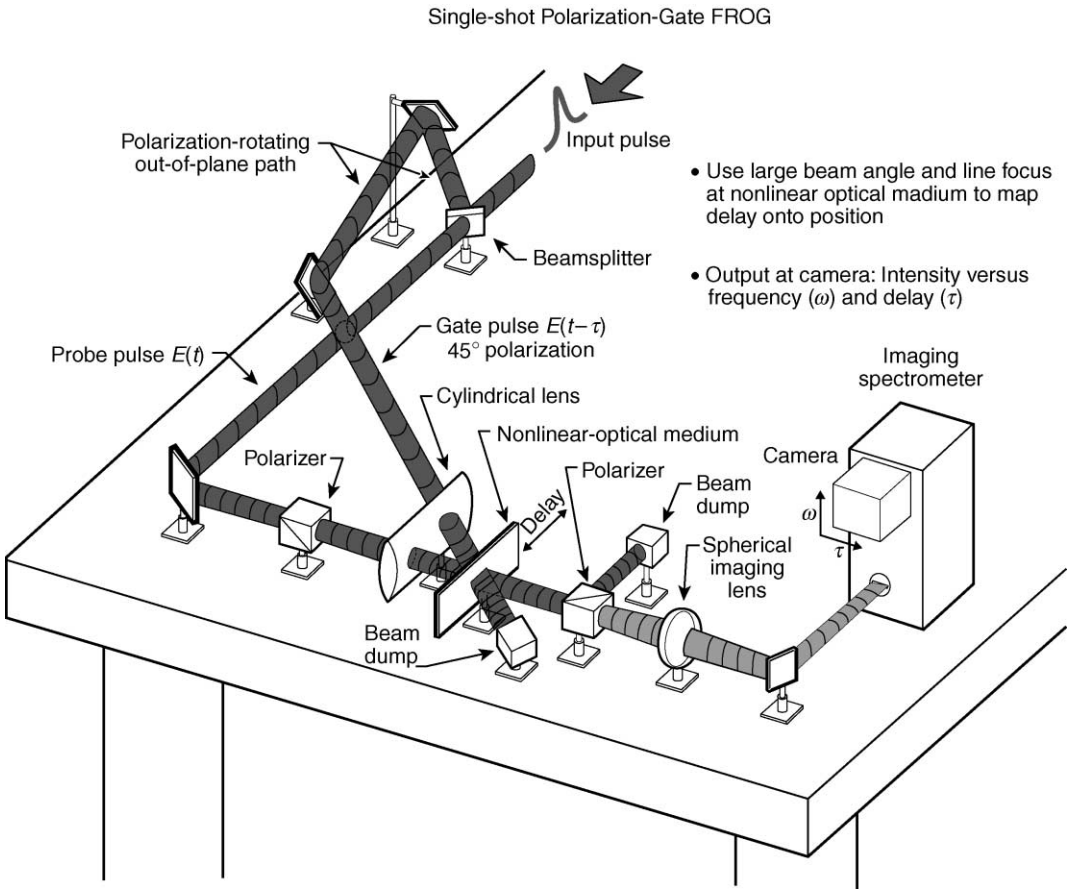


Figure 6 Measuring the spectrogram of a pulse is easier than measuring its sonogram – a spectrogram is a spectrally resolved autocorrelation. In this figure, the optical Kerr–effect is used (polarization-gate) as the nonlinearity. (Adapted with permission from Kane DJ and Trebino R (1993) Single shot measurement of the intensity and phase of an arbitrary ultrashort pulse by using frequency-resolved optical gating. *Optics Letters* 18(10): 823–825.)

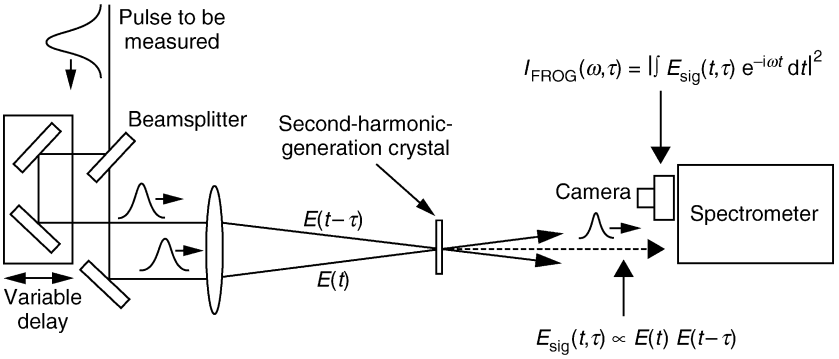


Figure 7 This figure shows a schematic of an SHG FROG device. The SHG signal from the autocorrelation is spectrally resolved. SHG FROG is very simple and sensitive, but it has a direction-of-time ambiguity. For example, if the pulse has chirp, only the magnitude of the chirp is determined – the sign of the chirp remains unknown.

a finite-length gate pulse yields the spectrum of all of the finite pulse segments with duration equal to that of the gate. While the phase information remains lacking in each of these short-time spectra, having spectra of an infinitely large set of pulse segments compensates for this loss. The spectrogram has been shown to nearly uniquely determine both the intensity $I(t)$ and phase $\varphi(t)$ of the pulse, even if the gate pulse is longer

than the pulse to be measured (although if the gate is too long, sensitivity to noise and other practical problems arise).

In FROG, when using optically induced birefringence as the nonlinear effect, the signal pulse is given by:

$$E_{\text{sig}}(t,\tau) \propto E(t)|E(t-\tau)|^2 \qquad [10]$$

So the measured signal intensity $I_{\text{FROG}}(\omega,\tau)$, after the spectrometer is:

$$I_{\text{FROG}}(\omega,\tau) = \left|\int_{-\infty}^{\infty} E(t)|E(t-\tau)|^2 e^{-i\omega t}dt\right|^2 \qquad [11]$$

We see that the FROG trace is a spectrogram of the pulse $E(t)$ although the gate, $|E(t-\tau)|^2$, is a function of the pulse itself.

To see that the FROG trace essentially uniquely determines $E(t)$ for an arbitrary pulse, it is first necessary to observe that $E(t)$ is easily obtained from $E_{\text{sig}}(t,\tau)$. Then it is simply necessary to write eqn [11] in terms of $E_{\text{sig}}(t,\Omega)$, the Fourier transform of the signal field $E_{\text{sig}}(t,\tau)$, with respect to delay variable τ. We then have what appears to be a more complex expression, but one that will give us better insight into the problem:

$$I_{\text{FROG}}(\omega,\tau) = \left|\int_{-\infty}^{\infty} E_{\text{sig}}(t,\Omega)e^{-i\omega t - i\Omega\tau}dt\,d\Omega\right|^2 \qquad [12]$$

Equation [12] indicates that the problem of inverting the FROG trace $I_{\text{FROG}}(\omega,\tau)$, to find the desired quantity $E_{\text{sig}}(t,\Omega)$, is that of inverting the squared magnitude of the two-dimensional (2D) Fourier transform of $E_{\text{sig}}(t,\Omega)$. This problem, which is called the 2D phase-retrieval problem, is well known in many fields, especially in astronomy, where the squared magnitude of the Fourier transform of a 2D image is often measured. At first glance, this problem appears unsolvable; after all, much information is lost when the magnitude is taken. It is well known that the one-dimensional (1D) phase retrieval problem is unsolvable (for example, infinitely many pulse fields give rise to the same spectrum). Intuition fails in this case, however; two- and higher-dimensional phase retrieval essentially always yields unique results.

The Simplified FROG – GRENOUILLE

A simplified version of the FROG device, known as grating eliminated no-nonsense observation of ultrafast incident laser light e-fields (GRENOUILLE), can be constructed using the doubling crystal as both the gating medium and the spectrometer (see Figure 8). In this case, the phase matching condition in certain thick doubling crystals causes a wavelength dependent output from the doubling crystal. A cylindrical lens is used in a Fourier transform configuration to image the wavelength variation onto a CCD camera. A cylindrical lens set at 90 degrees to the Fourier transform lens images the time axis onto the CCD camera. A Fresnel biprism replaces the traditionally used Mach–Zender interferometer to produce two pulses propagating at an angle with respect to each other. Presently, the GRENOUILLE technique works with BBO, in the wavelength region around 500 nm–1200 nm, and with proustite in the wavelength region around 1200 nm–2000 nm.

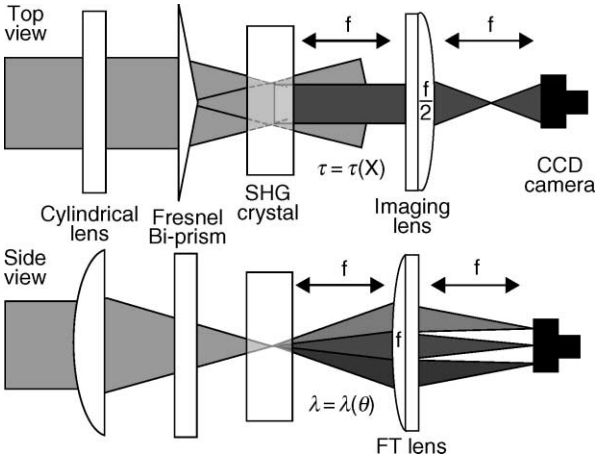


Figure 8 Figure showing a schematic of a GRENOUILLE FROG device which uses a thick doubling crystal as both the gating nonlinearity and the spectrometer. Like single-shot autocorrelators, a cylindrical lens focuses a spatially large input beam into an SHG crystal in only one dimension. Unlike most autocorrelators, the Mach–Zender interferometer is replaced with a Fresnel biprism that forces each half of the input beam to propagate, at an angle, toward the SHG crystal. An imaging cylindrical lens images the spatial dependence of the delay onto a CCD camera. A Fourier transform cylindrical lens maps the angular dependence of wavelength from the SHG crystal to a spatial coordinate on the CCD camera. (Adapted with permission from O'Shea P, Kimmel M, Xun G and Trebino R (2001) Highly simplified device for ultrashort-pulse measurement. *Optics Letters* 26(12): 932–934.)

FROG Inversion Algorithms

An iterative 2D phase retrieval algorithm is required to extract the pulse information from the measured FROG trace (see Figures 9 and 10). This algorithm converges to a pulse that minimizes the difference

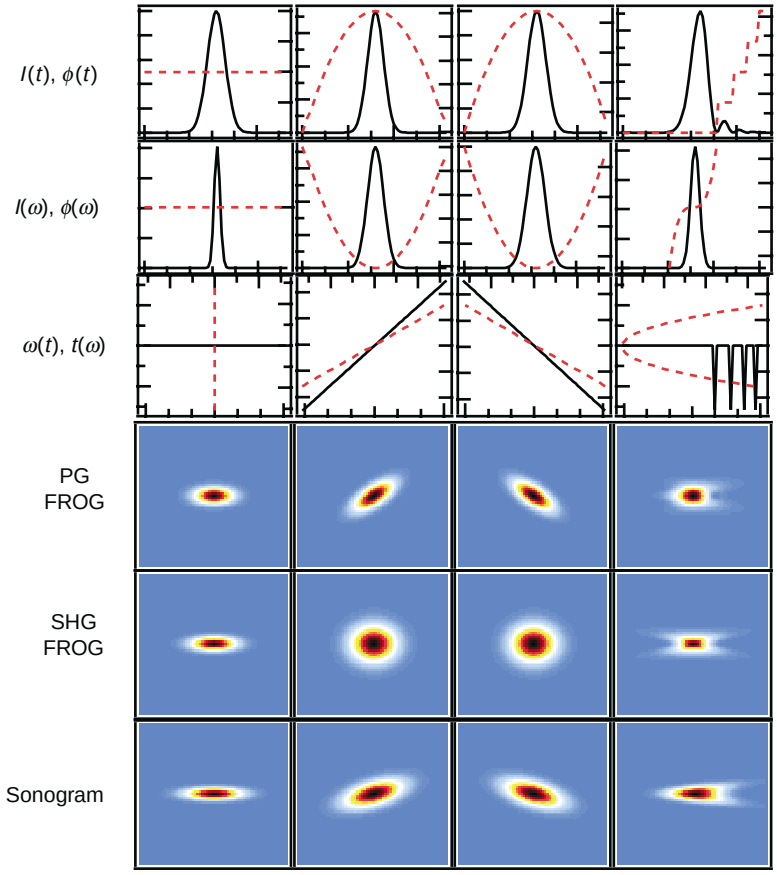


Figure 9 Example FROG traces and sonograms are shown for four different pulses: a transform limited pulse, a positively chirped pulse, and a negatively chirped pulse. The top series of plots are the time domain representation of the pulse intensity (solid line) and phase (dashed line). The second row of four plots is the frequency domain representation of the sample pulse's intensity and phase. The next four plots show the instantaneous frequency and the group delay. The vertical axis is frequency and the horizontal axis is time. The remaining rows show the PG FROG traces, the SHG FROG traces and sonograms of the pulse shown in the first two rows.

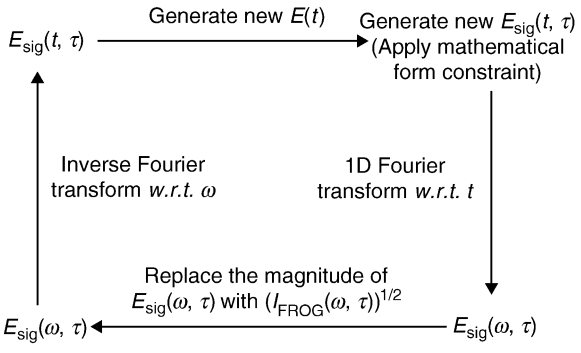


Figure 10 Phase retrieval algorithm for the inversion of FROG spectrograms. Start with an initial guess for the pulse to generate an initial $E_{sig}(t,\tau)$. A 1D Fourier transform generates the FROG trace. The next step is to replace the magnitude of the calculated FROG trace with the square root of the measured FROG trace. Inverse Fourier transform with respect to ω to produce the new signal field and generate a new guess for $E(t)$. Interestingly, it is only the step that produces $E(t)$ from $E_{sig}(t,\tau)$ that differentiates all the FROG algorithms.

between the measured and the calculated FROG trace. While this aspect of FROG has been its Achilles heel in the past, in reality, new generalized projections algorithms (together with faster computers) converge quickly and can track pulse changes at rates up to 30 Hz and beyond, making FROG a true, real-time pulse measurement technique. Indeed, excellent algorithms for the analysis of FROG traces in real-time are commercially available.

The original FROG inversion algorithm, commonly referred to as the vanilla algorithm, is simple and iterates quickly, but tends to stagnate, giving erroneous results, especially for geometries that use a complex gate function such as SHG or self-diffraction. An improved algorithm, incorporating several different algorithms including brute force minimization, was developed to alleviate stagnation programs at the expense of both speed and convergence time. By the mid-1990s, a significant

advance in both speed and stability was made by the addition of a numerical method called generalized projections. This algorithm determines the next guess by constructing a projection – minimizing the error (distance) between the FROG electric field, $E_{\text{sig}}(t, \tau)$, obtained immediately after the application of the intensity constraint, and the FROG electric field calculated from the mathematical form constraint.

The first generalized projections algorithm used a standard minimization procedure to find the electric field for the next iteration, which can still be slow. For the most common FROG geometries, PG and SHG, a different algorithm that determines the next iteration directly has been developed. This algorithm, called the principal components generalized projections (PCGP) algorithm, converts the generalized projections algorithm into an eigenvector problem. Using this algorithm, pulse measurement rates of at least 30 Hz have been achieved.

The goal of the phase retrieval algorithm is to find the $E(t)$ that satisfies two constraints. The first is the FROG trace itself which is the magnitude squared of the 1D Fourier transform of $E_{\text{sig}}(t, \tau)$:

$$I_{\text{FROG}}(\omega, \tau) = \left| \int_{-\infty}^{\infty} E_{\text{sig}}(t, \tau) \mathrm{e}^{-\mathrm{i}\omega t} \mathrm{d}t \right|^2 \qquad [13]$$

The other constraint is the mathematical form of the signal field, $E_{\text{sig}}(t, \tau)$, for the nonlinear interaction used. The signal forms for a variety of FROG beam geometries are:

$$E_{\text{sig}}(t, \tau) \propto \begin{Bmatrix} E(t)|E(t-\tau)|^2 & \text{PG FROG} \\ E(t)^2 E^*(t-\tau) & \text{SD FROG} \\ E(t)E(t-\tau) & \text{SHG FROG} \\ E(t)^2 E(t-\tau) & \text{THG FROG} \end{Bmatrix} \qquad [14]$$

where PG is polarization gate, SD is self-diffraction, SHG is second harmonic generation, and THG is third harmonic generation FROG.

All FROG algorithms work by iterating between two different data sets: the set of all signal fields that satisfy the data constraint, $I_{\text{FROG}}(\omega, \tau)$, and the set of all signal fields that satisfy eqn [14]. The difference between the FROG algorithms is how the iteration between the two sets is completed. In the case of generalized projections, the $E(t)'$s are chosen such that the distance between the $E(t)$ on the magnitude set and the $E(t)$ on the mathematical form set is minimized. This is accomplished by minimizing the following equation:

$$Z = \sum_{i,j=1}^{N} \left| E^{(k)}_{\text{sig(DC)}}(t_i, \tau_j) - E^{(k+1)}_{\text{sig(MF)}}(t_i, \tau_j) \right|^2 \qquad [15]$$

where $E^{(k)}_{\text{sig(DC)}}(t_i, \tau_j)$ is the signal field generated by the data constraint, and $E^{(k+1)}_{\text{sig(MF)}}(t_i, \tau_j)$ is the signal field produced from one of the beam geometry equations in eqn [14]. For the normal generalized projection, the minimization of Z is completed using a standard steepest descent algorithm; the derivative of Z with respect to the signal field is computed to determine the direction of the minimum. The computation of the derivatives are tedious; they are tabulated elsewhere.

An alternative to an algorithm that minimizes eqn [15] is the PCGP algorithm. This algorithm converts the computation of the next guess into an eigenvector problem, reducing the computation of the next guess to simple matrix–vector multiplications. This algorithm works for both the PG and SHG beam geometries, it is robust and the fastest FROG algorithm. Indeed, the PCGP algorithm was used in the original real-time FROG work by D.J. Kane.

Self Checks in FROG Measurements

Unlike any other pulse measurement techniques, FROG can provide a great deal of feedback about both the quality of the measurement (systematic errors) and the quality of the algorithm's performance. The most common check for convergence is the FROG trace error together with a visual comparison between the retrieved FROG trace and the measured FROG trace. The FROG trace error is given by:

$$G = \sqrt{\frac{1}{N^2} \sum_{i,j=1}^{N} \left| I_{\text{FROG}}(\omega_i, \tau_j) - \alpha I^{(k)}_{\text{FROG}}(\omega_i, \tau_j) \right|^2} \qquad [16]$$

where α is a renormalization constant, I_{FROG} is the measured trace, and $I^{(k)}_{\text{FROG}}(\omega, \tau)$ is computed from the retrieved electric field. Typically, the FROG trace error of a PG measurement should be less than 2% for a 64×64 pixel trace, while the FROG trace error of a 64×64 pixel SHG FROG trace should be about 1% or less. Acceptable FROG trace errors decrease as FROG trace size increase for smaller FROG traces. These values are only rules of thumb only;

for example, acceptable retrievals of large and very complicated FROG traces can produce larger FROG trace errors.

In a good FROG measurement, the spectrum of the retrieved pulse should faithfully reproduce the salient features of the pulse's measured spectrum. SHG FROG even provides an additional check called the frequency marginal. The sum of an SHG FROG trace along the time axis yields the autoconvolution of the pulse's spectrum, providing two ways the FROG measurement can be checked. First, the autoconvolution of an independently measured spectrum can be compared to the sum of the FROG trace along the time axis, providing an indication of how well the measurement was made. For example, if the doubling crystal was too thick in the pulse measurement, the FROG trace's frequency marginal will be narrower than the autoconvolution of the measured spectrum. Second, comparing the autoconvolution of the retrieval pulse spectrum with the FROG trace marginal can provide a test of algorithm convergence in addition to a test of the measurement. Phase matching problems appear as a mismatch between the FROG trace frequency marginal and the autoconvolution of the retrieved spectrum.

Because FROG is a spectrally resolved autocorrelation, summing any FROG trace along the frequency axis yields the autocorrelation of the measured pulse. This autocorrelation can be compared to an independently measured autocorrelation, or a comparison can be made between the frequency sum of the FROG trace and the autocorrelation calculated from the retrieved pulse to determine algorithm convergence and the quality of the measurement.

Measuring Pulses Directly – SPIDER

SPIDER is a novel technique that measures the pulse directly – no iterative phase retrieval method is required. To accomplish this feat, spectral interferometry is conducted on two pulse replicas that are shifted in frequency with respect to one another. The original pulse is split into two replicas. The two pulse replicas are sent into a phase modulator that shifts the center frequency of each pulse slightly. The frequency-shifted pulses are sent into a spectrometer and a standard spectral interferometry analysis yields the derivative of the phase. Integration of the phase derivative, together with the spectrum of pulse, provides the full intensity and phase.

Thus, analysis of the spectral interferogram gives:

$$\Delta\phi = (\phi(\omega_1) - \phi(\omega_2))\Delta\omega \qquad [17]$$

where $\Delta\phi$ is the measured phase difference from the spectral interferogram, $\Delta\omega$ is the difference in the center frequency between the two pulses, and $\phi(\omega)$ is the pulse phase as a function of frequency. Consequently, the phase difference must be divided by the frequency difference between the two pulses to determine the true derivative. It is also important to subtract out the linear phase difference resulting from the delay between the two pulses (the delay produces the fringes in the resulting interferogram) as this integrates to a linear chirp.

Because phase modulators are too slow to be used in the femtosecond region, a different method is used to shift the frequency of the two pulses – they are mixed with a highly chirped pulse (see Figure 11).

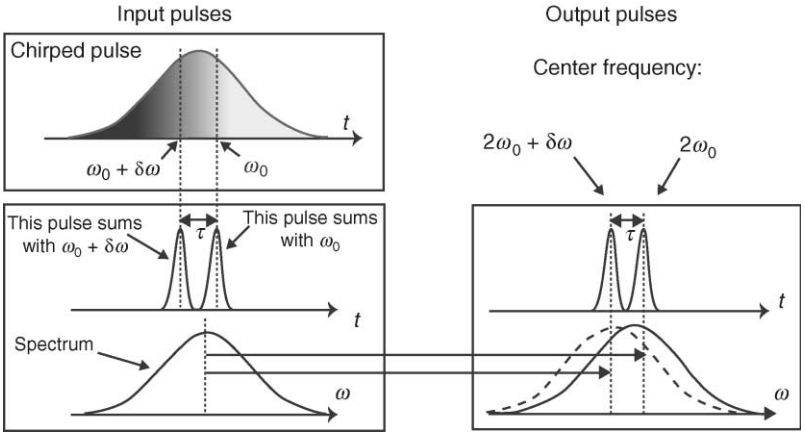


Figure 11 Figure showing how a SPIDER device works. The highly chirped pulse has a frequency that varies with time. Mixing the chirped pulse with two time delayed replicas produces two pulses, separated in time, that have slightly different frequencies, but are otherwise identical. By interfering the two pulses in a spectrometer, the relative phase between $\omega_0 + \delta\omega$ portions of the pulse and ω_0 portions of the pulse are determined.

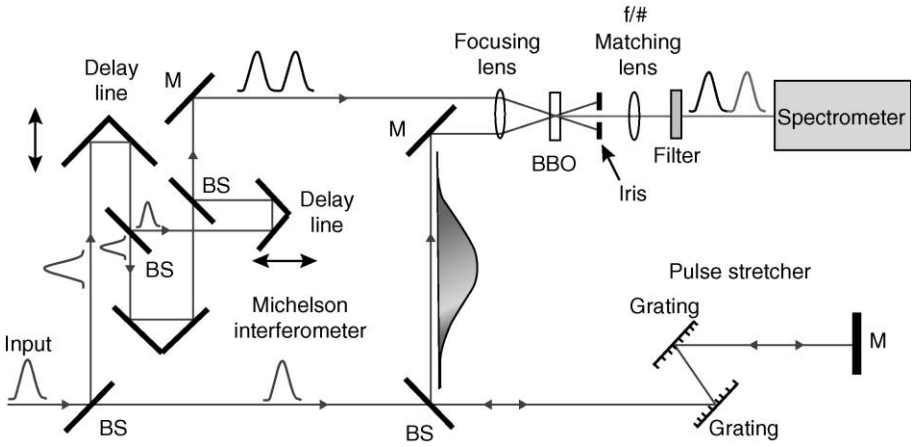


Figure 12 SPIDER apparatus. The input pulse is first split into two pulses. One pulse is stretched in a pulse stretcher. The other pulse is sent into a Michelson interferometer to produce two time-delayed pulses. The two time-delayed pulses are mixed with the stretched pulse to produce two pulses at a slightly different center frequency. The two pulses are sent into a spectrometer to produce a spectral interferogram. Analysis of the spectral interferogram yields the derivative of the pulse phase.

Because the pulses are slightly time delayed, each one aligns with a slightly different frequency in the chirped pulse. The chirped pulse must be so highly chirped that the pulse frequency does not change significantly over the duration of the pulse to be measured.

A schematic of a SPIDER device for the measurement of femtosecond laser pulses is shown in Figure 12. The input pulse is first split into two replicas. The first replica is sent to a pulse stretcher to produce to highly chirped pulse. The other replica is sent into a Michelson interferometer (or a simple étalon may sometimes suffice) to create two time-delayed replicas. The time-delayed replicas are mixed with the chirped pulse. Because each replica is mixed with a slightly different frequency in the chirped pulse, each replica produces a pulse with a slightly different center frequency. When the time-delayed replicas are spectrally resolved, each frequency of the pulse interferes with a frequency-shifted portion of the same pulse. Thus, the measured phase difference is proportional to the derivative of the phase of the original pulse.

Measuring Weak Pulses

Because virtually all the pulse measurement techniques, that do not require a reference pulse, use a nonlinearity, measuring weak ultrafast pulses directly is difficult. Fortunately, because most weak pulses are generated from a stronger pulse, spectral interferometry can be used to determine the relative phase between the weaker pulse and the stronger pulse (for regions where there is spectral overlap between the two pulses). If the stronger pulse is characterized using another pulse measurement technique, then the weaker pulse phase can be determined from the relative phase measurement and the known phase of the stronger pulse. For example, the Trebino group termed the combination of FROG and SI as TADPOLE (Temporal Analysis by Dispersing a Pair Of Light E-fields).

Which Pulse Measurement Method Should I Use?

Choosing a pulse measurement method depends on the required measurement as well as the precision and accuracy required. All the pulse measurement techniques have advantages and disadvantages, but excellent measurements have been made with all of them. Autocorrelation is adequate when all that is required is the approximate pulse duration without any phase information. Spectral interferometry is quite useful for measuring transient changes in a sample in pump-probe experiments. FROG is perhaps the most commonly used and general-purpose pulse measurement technique, providing a great deal of feedback about the quality of the pulse measurement. Inexpensive and convenient FROG devices, together with software, are available commercially, adding to their acceptance. While SPIDER is experimentally complex and does not have the cross checks that FROG has, it provides the quickest answer, requiring no iterative

phase retrieval algorithm to obtain pulse intensity and phase.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant numbers DMI-9801116 and DMI-0091454. The author would also like to acknowledge Rick Trebino for the use of figures from his lectures.

List of Units and Nomenclature

$A(t)$	Full electric field
$E(t)$	Complex electric field without carrier frequency
$E_{\text{sig}}(t, \tau)$	Complex electric field of the FROG trace versus time and time delay
$E_{\text{sig}}(t, \Omega)$	Fourier transform of $E_{\text{sig}}(t, \tau)$ with respect to τ
$\tilde{E}(\omega)$	Complex electric field in the frequency domain
$f(t)$	Correlation product
$f(\omega)$	Fourier transform of the correlation product
$g(t-\tau)$	time gate or filter
G	FROG trace error
$h(\omega-\Omega)$	frequency gate or filter
$I(t)$	Intensity profile of the electric field in the time domain
$\tilde{I}(\omega)$	spectral intensity
$I_{\text{FROG}}(\omega, \tau)$	Intensity of the FROG trace
$\Theta(t)$	Heaviside function $= 0\ t < 0,\ = 1\ t \geqq 0$
$\phi(\omega)$	Frequency domain phase
$S_{\text{E}}(\omega, \tau)$	Spectrogram
$S_{\text{E}}(\Omega, T)$	Sonogram
$S(\omega)$	Spectral interferogram
t, T	variables for time
τ	Fourier transform
v_{g}	group velocity
$\varphi(t)$	Time domain phase
Δw	Beam waist
Z	Distance metric between the FROG signal field generated by the data constraint and the FROG signal field generated by the mathematical form constraint
Φ	Geometric angle between two beams
$\Delta\phi$	Phase difference
$\Pi^{(n)}$	nth order nonlinear susceptibility
ω	Primary variable for frequency
$\Delta\omega$	Difference in center frequency between two pulses
Ω	Secondary variable for frequency
$\Omega(t)$	Instantaneous frequency

See also

Coherent Transients: Ultrafast Studies of Semiconductors. **Nonlinear Optics, Basics:** Ultrafast and Intense-Field Nonlinear Optics. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses. **Ultrafast Technology:** Femtosecond Chemical Dynamics: Gas-Phase; Femtosecond Condensed Phase Spectroscopy: Structural Dynamics; Ultrafast Illumination and Processing.

Further Reading

Diels JC and Rudolph W (1996) *Ultrashort Laser Pulse Phenomena*. San Diego, CA: Academic Press Inc.

Dorrer C, Belabas N, Likforman J-P and Joffre M (2000) Spectral resolution and sampling issues in Fourier-transform spectral interferometry. *Journal of the Optical Society B* 17: 1795.

Kane DJ, Weston J and Chu K-CJ (2003) Real-time inversion of polarization gate frequency-resolved optical gating spectrograms. *Applied Optics* 42: 1140–1144.

Iaconis C and Walmsley IA (1999) Self-referencing spectral interferometry for measuring ultrashort optical pulses. *IEEE Journal of Quantum Electronics* 35: 501–509.

Lepetit L, Cheriaux G and Joffre M (1995) Linear techniques of phase measurement by femtosecond spectral interferometry for applications in spectroscopy. *Journal of the Optical Society of America B* 12: 2467–2474.

Rulliere C (ed.) (2000) *Femtosecond Laser Pulses*. New York: Springer Verlag.

Shuman TM, Anderson ME, Bromage J, Iaconis C, Waxer L and Walmsley IA (1999) Real-time SPIDER: ultrashort pulse characterization at 20 Hz. *Optics Express* 5: 134–143.

Treacy EB (1971) Measurement and interpretation of dynamic spectrograms of picosecond light pulses. *Journal of Applied Physics* 42: 3848–3858.

Trebino R (2000) *Frequency Resolved Optical Gating: The Measurement of Ultrashort Laser Pulses*. Boston, MA: Kluwer Academic Publisher.

Trebino R, DeLong KW, Fittingholf DN, Sweetser JN, Krunbugel MA, Richman BA and Kane DJ (1997) Measuring ultrashort laser pulses in the time-frequency domain using frequency-resolved optical gating. *Review of Scientific Instruments* 68: 3277–3295.

Various authors (1999) Feature issue on ultrashort-laser-pulse measurement and applications. *IEEE Journal of Quantum Electronics* 35: 4.

ULTRAFAST TECHNOLOGY

Contents

Femtosecond Chemical Dynamics: Gas-Phase
Femtosecond Condensed Phase Spectroscopy: Structural Dynamics
Ultrafast Illumination and Processing

Femtosecond Chemical Dynamics: Gas-Phase

M Dantus, Michigan State University, East Lansing, MI, USA

E J Brown, McDaniel College, Westminster, MD, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Chemical reactions are at the heart of chemistry; the quest to understand them has sparked research and study along many different avenues. The fundamental dynamics involved in bond breakage, bond formation, and transition states take place on a very fast time-scale, usually in the tens to hundreds of femtoseconds, where a femtosecond (fs) is equal to 10^{-15} seconds. Therefore, ultrashort laser pulses are required to measure these fundamental processes as well as the rotational and vibrational motion before and after the reaction. The whole set of measurements can then be used for mapping a potential energy surface for the reaction. As a result of these studies, one learns the key details that are required to develop a picture of how chemical reactions occur. The measurement of these fundamental dynamics, using ultrashort laser pulses, is the subject of this article.

To illustrate how long it takes to form or break a chemical bond, one needs to consider the following analogy. The time it takes to break a chemical bond is approximately a tenth of a trillionth of a second (10^{-13} s). During such a short time, light can travel only 0.03 mm in vacuum. This time is so short that it cannot be measured electronically. Its measurement requires the use of ultrashort laser pulses, with pulse durations of 50 femtoseconds or shorter. The development of methods aimed at measuring the fundamental steps involved in chemical reactions culminated in the development of femtosecond transition state spectroscopy (FTS), a method developed in the 1980s by Zewail and co-workers at Caltech. This work allows researchers to observe bond breakage and formation, vibrational and rotational dynamics, and transition state configurations, as the reaction occurs. The Nobel Prize in Chemistry was awarded in 1999 to Ahmed Zewail for this development. Since the 1980s, femtosecond lasers have been used to study molecular dynamics and chemical dynamics on very fast time-scales. The field, known as femtochemistry, has grown quickly and diversified to involve hundreds of scientists around the world. This article discusses the measurement of ultrafast dynamics measured in gas-phase samples. The presentation encompasses some of the measurement techniques and the fundamental values to be measured and concludes with an outlook towards the future.

From Nanosecond to Picosecond to Femtosecond Reaction Kinetics

The quest for shorter light pulses to study chemical reactions predates the invention of the laser. In the early days, electric discharges were used to measure fast dynamics in the millisecond (10^{-3} s) and then the microsecond (10^{-6} s) time-scales. With the invention of the laser, measurements improved drastically with temporal resolution going from the nanosecond (10^{-9} s) to the picosecond (10^{-12} s) time-scales. Since the early days, it was recognized that the best time resolution was obtained when pairs of light pulses were used – one to initiate the reaction and the second to probe it. These pump-probe measurements have progressed with the laser technology almost to the single femtosecond level.

The measurements being considered here go beyond the measurement of fast reaction kinetics. When a chemical reaction occurs in a beaker, one usually measures a statistical rate of reaction that involves diffusion of reagents in a solvent. These kinetic measurements, which involve an ensemble of molecules, are much slower and do not reveal the

fundamental steps of the reaction such as bond formation. The introduction of ultrafast pulses allowed, for the first time, the measurement of the motion of the atoms in real time, as if the motion was frozen by ultrafast flash photography. Capturing this motion helps to reveal the internal forces, described by a potential energy surface, that act on the atoms in the molecule during the chemical reaction. This concept is illustrated in Figure 1, where we see the reagents and the products for the reaction. Typically, the starting and ending point of a chemical reaction are very well known and the compounds are very well characterized. However, how the reagents become the products is not always known. Chemists typically deduce a reaction mechanism based on a large body of experiments where different parameters including the structure of the reagents are modified to evaluate their effect on the outcome of the reaction. Unfortunately in most cases, reaction mechanisms, even when consistent with all available experimental evidence, may not accurately reflect how a particular chemical reaction takes place.

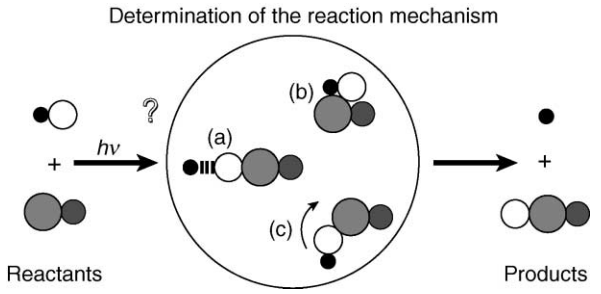


Figure 1 A sketch of a chemical reaction, AB + CD → A + BCD. The reactants and products are well known, but it is not known how they change from reactants to products. This mechanism may involve a transition state which is (a) a linear complex with vibrational motion, (b) a nonlinear complex, (c) a bent complex with rotational motion, or many other possibilities and combinations. The elucidation of the mechanism and transition state dynamics are the goals of femtosecond time-resolved studies.

The ideal method to make the determination involves direct observation. As mentioned earlier, direct observation requires femtosecond laser pulses, just like a fast camera shutter is required to take pictures of fast-moving objects. The concept of a femtosecond pump-probe measurement is illustrated in Figure 2. The chemical reaction is initiated at time $t = 0$ fs by the pump laser. The pump laser provides the energy required for initiating the chemical transformation. The probe laser, delayed in time, probes the formation of the product. Figure 2 illustrates a number of pump-probe measurements

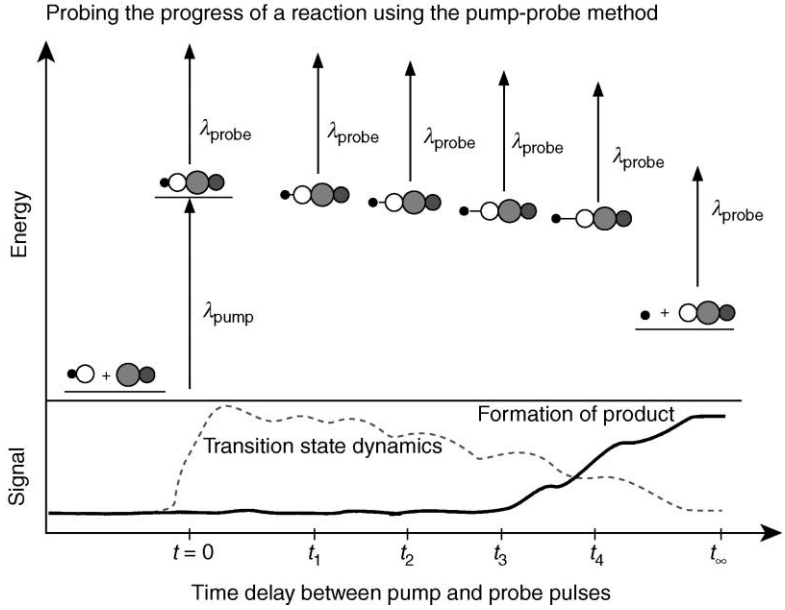


Figure 2 Energy schematic of the reaction of AB + CD → A + BCD assuming vibrational motion in a linear transition state. The pump pulse provides enough energy to excite the reactants to the transition state at $t = 0$. This transition state is probed by the second pulse. This pulse arrives at different time delays, from no delay ($t = 0$), to increasing delays (t_1, t_2,…), to infinite delay (t_∞). The signal varies depending on the time delay between the two pulses and these changes in signal intensity correspond to vibrational and rotational information about the transition state. To obtain a transient as shown at the bottom, the time delay between the pump and probe pulses is scanned. The dashed line shows the transient that would be obtained when the transition state is monitored; the solid line shows the transient that would be obtained when the product is monitored.

each obtained at a different time delay between the pump and probe laser pulses. When a series of pump-probe measurements are gathered as a function of time delay, one obtains a transient that contains a record for the evolution of the reagents to product as a function of time.

The Formation of Coherent Wave Packets and their Motion

The illustration shown in Figure 2 shows atoms in a molecule moving as classical particles. Because of their small size, atoms behave quantum mechanically and this has some implications on the observed motion. A prototypical pump-probe measurement is illustrated for this purpose in Figure 3. For resonant excitation, a photon is absorbed by a molecular system, originally in the ground electronic state (V_0). This absorption causes a vertical transition to an excited electronic state (V_1). Consistent with Heisenberg's uncertainty principle, a pulse that has a short temporal duration has a large spectral bandwidth; for most molecular systems and ultrashort pulses, the frequency spread of the pulse is wider than the vibrational energy spacing in the excited state. Thus, a number of quantum vibrational states are populated on V_1 by the absorption of a short laser pulse. When a short pulse excites a number of states, a coherent superposition of states is formed. One can understand the formation of a wave packet by simply realizing that in the short time the laser pulse excites the system there is no opportunity for the atoms in the molecule to move. For this reason the initial wave packet is very similar to the ground state atomic arrangement.

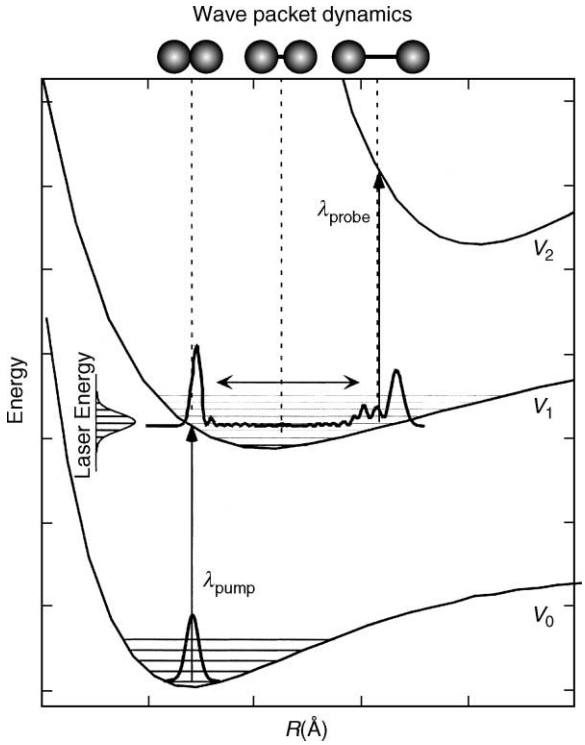


Figure 3 Pump-probe experiment. The pump pulses excite the molecular system from the ground state V_0 to excited state V_1. The large spectral bandwidth of femtosecond pulses allows for a number of vibrational levels in V_1 to be populated. A wave packet is formed that vibrates on V_1; the classical motion of the particle is shown as well at the top. In this case the energy of the probe pulse will excite the molecules to V_2 when the bond is stretched (right side of oscillation); however, when the bond is compressed, the energy is not resonant with any excitation and the probe pulse will not be absorbed.

When a short laser pulse excites a molecule to a bound electronic state, the resulting wave packet, $\Psi(t)$, is a sum of all the possible states φ_n in V_1 with amplitudes, a_n, $\Psi = \sum_n a_n \varphi_n$. The amplitudes, a_n, depend on the overlap between the initial (ground state) and final (excited state) spatial overlap. This overlap indicates that the excitation process is faster than the motion of atoms in a molecule. Based on this approximation for the overlap integral, the amplitudes are replaced by the Franck–Condon overlaps to obtain $\Psi = C\sum_n \langle \varphi_n | \varphi_0^{(0)} \rangle \varphi_n$. The wave packet on V_1 evolves in time according to $\Psi(t) = \sum_n a_n \exp[-iE_n t/\hbar]\varphi_n$, where E_n is the energy that corresponds to the state φ_n and $\hbar$ is Planck's constant divided by 2π. When a pulse is very short in time, the coherent superposition of states closely resembles the initial distribution of atoms in the molecule. The motion from inner to outer turning point of the quantum mechanical wave packet is illustrated in Figure 3. One can imagine this wave packet as a classical particle oscillating on the excited potential energy curve with the periodic nature of the molecular vibrations. When the pulse is very long and only one state is populated, the dynamics can no longer be observed, as is the case with conventional frequency-resolved spectroscopy.

Molecular dynamics can be inferred from high-resolution frequency-resolved spectroscopy; however, the spectra can become congested and the data analysis becomes more complicated for large molecules, for high temperatures, or in solutions. Molecular dynamics are obtained from frequency-resolved spectra by taking the Fourier transform of the data. It seems more intuitive to consider these dynamics in the time domain where the dynamics can be observed directly as 'snapshot' taken by the femtosecond pulses. Ultrafast time-resolved spectroscopy has other advantages – high peak intensities allow for multiphoton excitation and other nonlinear processes, the selection of the detection wavelength discriminates between different species and results in

an increased signal-to-noise ratio for intermediate short-lived species, collisions take place on a longer time-scale and do not affect these measurements, and the characteristics of the ultrafast laser pulse (such as phase, frequency components, and chirp) can be manipulated to control the dynamics and the yield of chemical reactions.

Methods for Measuring Femtosecond Dynamics

Pump-Probe Method

Pump-probe techniques use a pair of pulses separated in time, one to initiate the reaction and the second to determine its progress. The first femtosecond pulse with wavelength λ_{pump} excites a molecule from its (bound) ground electronic state (V_0) to an excited (V_1) state. After a variable time delay, a second femtosecond pulse with wavelength λ_{probe} causes a transition to another excited state (V_2) as shown in Figures 3 and 4.

The progress of the reaction dynamics is followed through the measurement of signal resulting from the absorption of the probe pulse by the system. There are different detection modes depending on the potential energy surfaces reached by the pump and probe pulses. In one case (Figure 4a), the first excited state (V_1) is nonfluorescent but the second one (V_2) fluoresces and this fluorescence is measured. In another case (Figure 4b), state V_1 is fluorescent but state V_2 is not; in this case the depletion of the fluorescence by the probe is measured. Finally (Figure 4c), excited states V_1 and V_2 may be dissociative; therefore, the fluorescence detected is emitted by the products. In any of these cases, as the wave packet on the intermediate state propagates (e.g., vibrates, dissociates), the transition probability varies (i.e., the ability of the probe pulse to be absorbed by the molecule changes). Therefore, the amount of laser-induced fluorescence (LIF) emitted varies with the pump-probe time delay. In some cases, the probe pulse produces ions and there are a number of methods to detect the electrons or the ions with great sensitivity.

Measurement of rotational dynamics

Measurement of rotational dynamics requires a vectorial frame of reference and this is achieved with linearly polarized laser beams. Excitation by the pump pulse selects an initial population distribution described by $\cos^2\theta$ where θ is the angle between the transition dipole and the pump laser polarization vector. As the rotational wave packet dephases due to the distribution of the population in different rotational levels, the probe pulse probes the transient alignment of the molecules in space. As the molecules rotate away from the initial alignment, the signal decreases. In Figure 5 rotational dynamics are depicted schematically. First, a single (classical) molecule is considered (top panel). As the molecule rotates, its dipole changes from being aligned with the polarization vector of the laser field

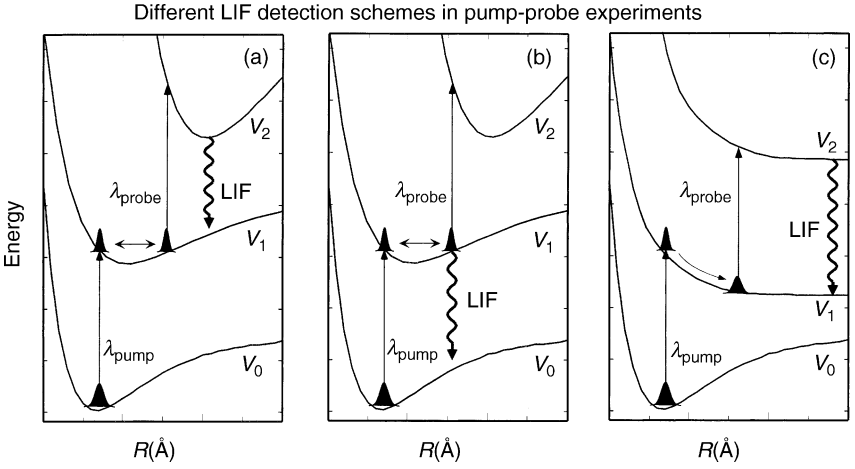


Figure 4 Different pump-probe experiments.(a) Bound-to-bound transitions from $V_0 \rightarrow V_1$ and from $V_1 \rightarrow V_2$. The V_2 state fluoresces and this increase in LIF from V_2 is measured as signal. (b) Bound-to-bound transition from $V_0 \rightarrow V_1$ where V_1 fluoresces. When the probe pulse causes an excitation from $V_1 \rightarrow V_2$, a loss of LIF signal is observed. The depletion of fluorescence is measured in this case. (c) Bound-to-repulsive surface transition from $V_0 \rightarrow V_1$. The energy of the probe causes an excitation to V_2 at a particular molecular distance. Fluorescence of a product fragment is measured.

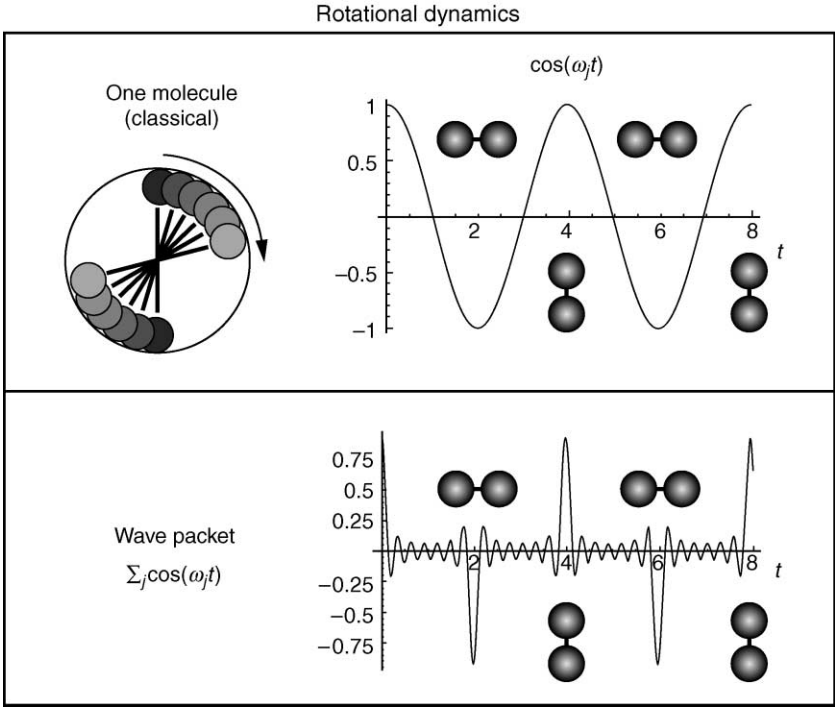


Figure 5 Rotational motion of molecules. Excitation by a linearly polarized laser pulse initially selects molecules that are well aligned ($\cos^2\theta$) with its polarization. Considering only one molecule evolving in time (top), the rotational motion in the molecule causes it to move away from the initial alignments and then eventually return to it. The signal will decrease when the molecule is not aligned with the polarization of the probe laser and will increase when aligned well. The signal looks like a cosine function that depends on the rotational frequency of the molecule (ω_j). For a rotational wave packet, the rotational dynamics are similar except that a number of rotational levels are populated; thus, the signal depends on the summation of many cosine functions and the features in the data become sharper.

(assumed vertical) to being perpendicular. The rotational motion can be described by a cosine function. The rotational frequency ω_j is proportional to the rotational level and the rotational constant of the molecule. Because the rotational constant is inversely proportional to the angular momentum, larger molecules rotate slower than smaller molecules. When several rotational levels are contained in a rotational wave packet the motion produces sharper features (bottom panel). The sharp features are called rotational recurrences or revivals. Note that the opposite trend is observed if the probe pulse is polarized perpendicular to the pump pulse; the signal increases as the molecules rotate away from the initial alignment and decreases as the alignment recurs. The time-resolved data for each polarization arrangement (parallel and perpendicular) contain both an isotropic component (vibrational motion) and an anisotropic component (rotational motion). In rotational anisotropy measurements, data from these two polarization arrangements are manipulated mathematically to yield just the rotational motion; rotational dynamics and rotational information such as the spectroscopic rotational constants and the rotational energy partitioning in the reactants and products can be extracted from these measurements.

Three-Pulse Four-Wave Mixing Method

Some methods are more complicated than pump-probe technique but they are more powerful. Three-pulse four-wave mixing (FWM) is one such method and is briefly presented here. Time-resolved four-wave mixing techniques can be understood based on the formation of transient gratings by two of the incident lasers. When two laser beams cross, the spatial modulation of their electric fields varies due to constructive and destructive interference as shown in Figure 6. The molecules in the interaction region experience varying electric field intensities according to their position and this leads to the formation of a transient grating of polarized molecules in space. The transient grating formation can be probed easily by the detection of Bragg scattering of a third laser beam, also known as homodyne detection; here, the signal beam is scattered in the phase-matching direction $\mathbf{k}_{\text{signal}}$ as indicated in Figure 6a. Alternatively, probing can be achieved by detection of changes in the intensity of a fourth beam that propagates in the

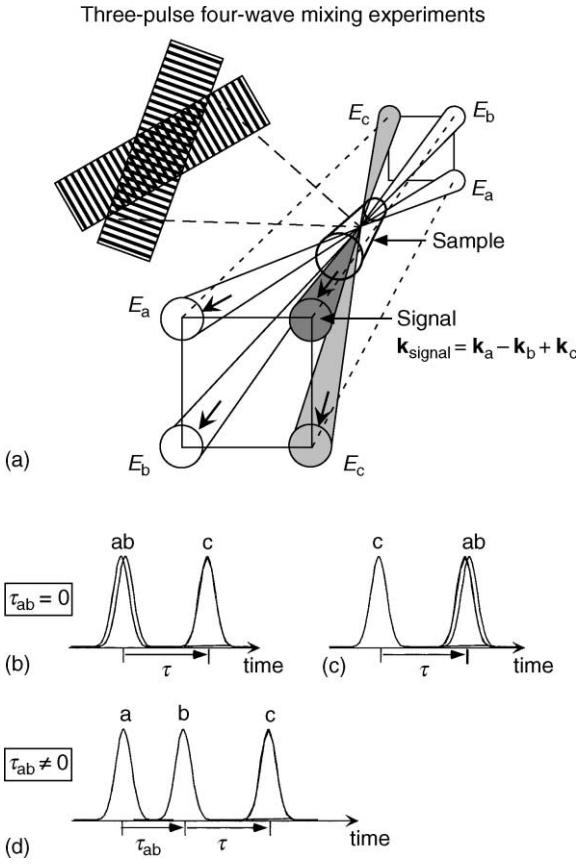


Figure 6 (a) FWM experimental configuration. Two of the beams cross which forms a grating and the third beam scatters off the grating to generate a fourth beam (the signal). The three laser pulses are noted as electric fields E_a, E_b, and E_c. These three laser pulses can be delayed or advanced in time relative to each other giving rise to a number of different pulse sequences. (b) Pulse sequence with fields E_a and E_b overlapped in time ($\tau_{ab} = 0$). Field E_c follows at a time delay τ. (c) Pulse sequence in which E_c precedes E_a and E_b by time delay τ. For both of these cases, τ is scanned in time to obtain the transients. (d) Pulse sequence in which no pulses are overlapped in time and are applied in the order E_a, E_b, and E_c. Time delay τ_{ab} can be set to specific values while τ is scanned in time.

same direction as the signal; this is called heterodyne detection. The medium that constitutes the gratings is composed of the sample molecules, which in the gas phase move freely. Therefore as time evolves, the molecular dynamics (rotation and vibration) modulate the transient grating and hence the signal.

Three-pulse four-wave mixing is a nonlinear spectroscopic method that combines the interaction of three laser pulses in a phase-matched geometry with a well-defined time sequence of the pulses. The signal from these measurements arises from a third-order polarization resulting from the interaction of the three electric fields. Three-pulse FWM is similar to the pump-probe technique in that a preparation step is followed, after some variable time delay, by a probing step. However, three-pulse FWM allows for a greater degree of control over the preparation and probing processes.

When the lasers are in resonance with an electronic state, FWM measurements can provide dynamic information from the ground and excited electronic states. When two pulses are incident on the sample at $t = 0$ and a third pulse probes the dynamics at a later time τ (as in Figure 6b), the measurements provide information that is very similar to that from a pump-probe measurement. However, having three pulses allows for other different setups. For example, when one pulse arrives at the sample first and the other two pulses follow after a time delay (as in Figure 6c), the resulting signal provides dynamic information from the excited state. This type of measurement also provides a measurement of the electronic coherence of the sample, i.e., for how long the molecules remember the phase of the incident laser pulses. This information is important if one wants to take advantage of the phase of laser pulses to cause quantum mechanical interference in the molecular sample. These measurements are known as photon echo and are much more common in condensed-phase samples. There are numerous pulse sequences possible; only three examples are shown in Figure 6.

FWM measurements do not depend on resonance excitation to an electronic excited state. In the absence of such resonance, the grating is formed in the ground state and ground state vibrational and rotational dynamics are observed. These measurements, also known as transient grating (TG), depend on the impulsive Raman scattering process. By changing the time delay (in Figure 6b) between the electric fields that generate the grating and the third (probe) beam, the signal is modulated by the molecular dynamics. Analysis of the time-dependent signal yields the rotational and vibrational spectroscopic values from the sample being studied.

Studies of Reaction Dynamics

Dissociation on a Repulsive Surface

In 1987, Zewail and co-workers studied the photodissociation of cyanogen iodide (ICN) using femtosecond lasers; the first direct observation of a chemical reaction from reactants to products was conducted. The excitation and decay of the reactants (ICN), the formation and decay of the intermediates (I–CN), and the formation of the products (I + CN) were observed in this experiment.

The pump pulse (306 nm) excited ground state cyanogen iodide to a repulsive excited state of ICN which dissociated into I and CN fragments (both in the ground state). The probe pulse when set to 388.5 nm was resonant with an excitation of the CN fragment; therefore, the reaction was monitored by detecting the laser-induced fluorescence (LIF) of this CN^* fragment using a probe pulse at different time delays (as shown in Figure 4c). The wavelength of the probe pulse could be detuned to different wavelengths for making observations at different lengths of the I–C bond. The results of these experimental studies showed that ICN dissociated in about 200 fs and that the transition state lived for only about 50 fs. The data from the wavelength detuning studies were used to map the excited state potential of ICN both in time and in space; this data analysis is called inversion to the potential.

A schematic representation of the ICN experiment is shown in Figure 7. The potential energy curves are shown in Figure 7a. The pump laser initiates the chemical reaction. Three different probe wavelengths are depicted, the first probing early in the reaction, the second at an intermediate distance, and the final one probing the free products. Notice that the wavelength of the probe determines the location in the potential energy curve that is probed. This concept allows one to study reagents as they are transformed into products. The signals obtained from the three different probe wavelengths are shown in Figure 7b.

Reactions Involving Crossing of Potential Energy Surfaces

For many chemical reactions, the progress of the reaction requires changes of the electronic state. This can take place where potential energy surfaces cross. The best-known example of reaction dynamics involves the experiment on the photodissociation of NaI by Zewail and co-workers. This experiment is depicted in Figure 8. The pump pulse excites the wave packet from ground state (ionic) to an excited state (covalent) where it begins to oscillate. The ground and excited states overlap with an avoided crossing in this molecule (ionic versus covalent nature, as shown in Figure 8a); when the wave packet approaches this avoided crossing, it splits into two partial waves. One wave stays on the excited state while the other 'escapes' onto the dissociative surface and separates into Na and I. The partial wave on the excited state again oscillates and then splits at the avoided crossing and the pattern continues. The time delay of the probe pulse shows the oscillations of the wave packet on the excited surface when detecting the activated complex $[NaI]^{*‡}$ (thin line), while the data of the product fragments (free Na atoms detected by using the D-line absorption) (thick line) show a step-wise accumulation in the signal corresponding to the sequential formation of products during each oscillation near the avoided crossing region of the surfaces (see Figure 8b).

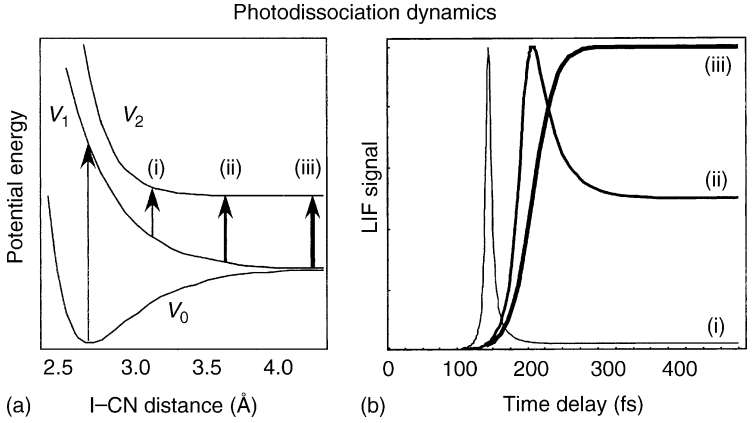


Figure 7 Photodissociation of ICN. (a) Potential energy surface schematic showing the initial excitation by the pump laser to a dissociative surface. The reaction could be monitored at different I–CN distances by using different probe wavelengths as shown by three examples here, (i), (ii) and (iii), from shorter to longer distances. (b) At short distances (long probe wavelengths), the signal is observed at shorter time delays only. The signal decreases at later time delays because the internuclear distance continues to lengthen as the dissociation occurs over time. When the distance increases as monitored by shorter probe wavelengths, the signal rises to a maximum at longer time delays reflecting the dissociation of the reaction in progress. At 'infinite' I–CN distance (i.e., the molecule is dissociated and only free I atoms and CN fragments remain), the signal reaches a maximum and remains there because the reaction is complete.

Chemical Dynamics with Multiple Pathways

The pump-probe concept can be used to probe more complex reactions. To date, a great number of different chemical reactions have been studied (see Further Reading). Among the most important reactions we list ring opening, isomerization, proton transfer, and concerted chemical reactions. We illustrate these measurements with experiments on the photodissociation of dihalogenated alkanes. Upon high-energy excitation (12 eV), the methylene iodide (CH_2I_2) molecule dissociates and forms molecular iodine in excited states D′ and f as well as producing I atoms (main channel). The detection wavelength used in the experiment can distinguish between these products; D′ fluorescence can be collected at 342 nm and f fluorescence is collected at 272 nm. The Dantus group has studied the dynamics of this type of concerted molecular elimination reaction extensively. The photodissociation reaction occurs by a multiphoton excitation (three-photons from the pump beam at 310 nm) reaching a dissociative state where the CH_2 fragment and excited molecular iodine are formed. The 620 nm probe pulse depletes the fluorescence from the excited I_2 molecule (either D′ or f state) as shown in Figure 4b; this fluorescence depletion is measured and analyzed to determine the dynamics of the reaction. The object of this experimental work was to determine if the photodissociation mechanism for forming molecular iodine followed a step-wise, synchronous concerted, or asynchronous concerted mechanism as shown in Figure 9.

The time transients reveal coherent vibrational motion of the I_2 product. Analysis of the transition state dynamics shows that the breaking of the carbon–halogen bond and the formation of the halogen–halogen bond occurs within 50 fs. Further, the iodine molecule is formed with a very hot rotational distribution indicating that the symmetry of the molecule is broken during the process; the torque imparts a large amount of rotational motion. Overall, the results of these studies show that the photodissociation process of CH_2I_2 follows an asynchronous concerted reaction mechanism and corresponds to a symmetry barrier that exists in C_{2v} molecules preventing the formation of the interhalogen bond. The existence of this symmetry barrier was explored by using CH_2ICl, a molecule that does not have C_{2v} symmetry; as expected, the halogen molecule (ICl) was formed with a cold rotational distribution indicating that it was able to follow a synchronous concerted mechanism. The dissociation times were also studied on the family of compounds CX_2Y_2 and $R{-}CHI_2$ to investigate the influence of thermodynamics and reduced mass changes on the reaction.

The Dynamics of Bond Formation

Most chemical reactions that are studied with femtosecond spectroscopy are unimolecular reactions (photodissociation); however, bimolecular reactions (photo-association) are interesting and chemically relevant in trying to understand how chemical reactions occur. In the photo-association process, a pair of atoms or molecules cooperatively absorbs the photon undergoing a free-to-bound transition and forming a bond. When the ultrashort pulse initiates the reaction in femtosecond photo-association spectroscopy, a clearly defined initial time for the reaction is given

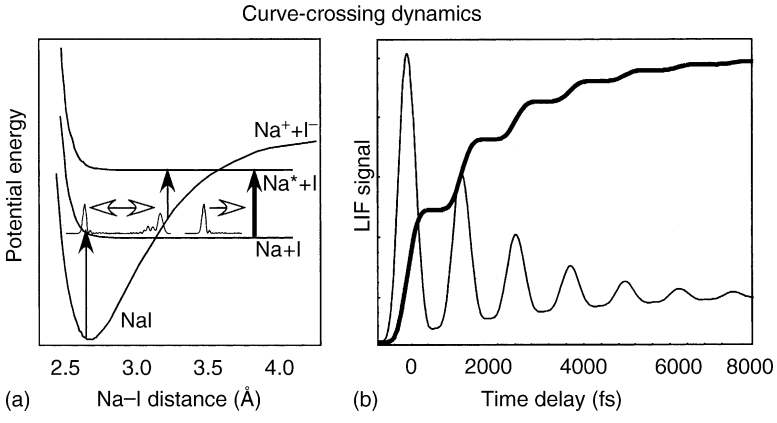


Figure 8 Photodissociation of sodium iodide. (a) Potential energy curve schematic showing the curve-crossing between the ground (ionic) and excited (covalent) states. As the wave packet oscillates on the excited curve, it splits into two wave packets at the curve crossing. (b) The thin line shows the vibrations of the excited complex as it oscillates on the excited surface. The signal intensity decreases because some of the molecules have dissociated to form Na and I atoms when the wave packet splits. The thick line shows the stepwise increase in free Na atom signal each time the wave packet oscillates and splits.

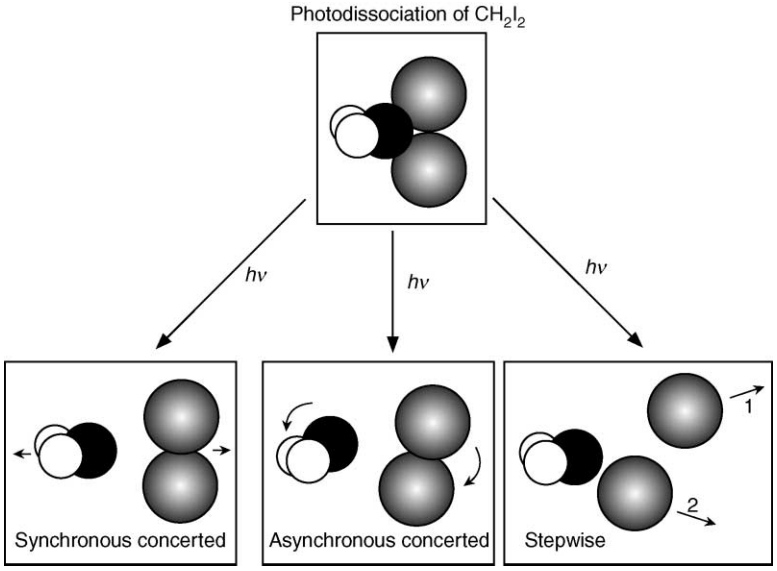


Figure 9 Possible mechanisms for the photodissociation of CH_2I_2 following multiphoton excitation. The molecule could break both C–I bonds and form an I–I bond all at the same time (synchronous concerted mechanism); this would result in very little rotational motion in the products. The molecule could begin to break one C–I bond slightly before breaking the second one and forming the I–I bond, but all three events still occur within the duration of the laser pulse (asynchronous concerted mechanism). This would result in significant rotational motion in the products. The molecule could also lose one I atom and at some later time lose the other I atom; these I atoms could collide later to form I_2 (stepwise mechanism).

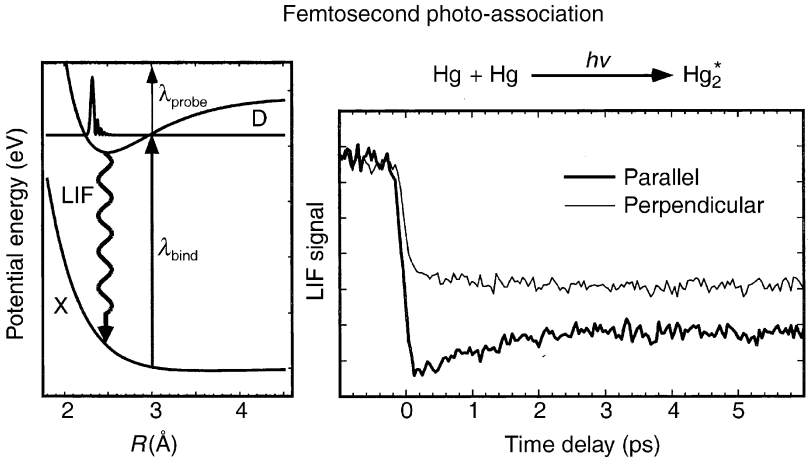


Figure 10 Photo-association of mercury dimers. The pump (bind) pulse is absorbed by two ground-state Hg atoms to form Hg_2^* on the excited state which is fluorescent. The probe pulse depletes this fluorescence causing the LIF signal to decrease at $t = 0$. Analysis of the transient reveals that the time of association is <60 fs. Parallel and perpendicular pump-probe data were examined to obtain the rotational distribution of this reaction.

and the alignment conditions of the reaction can be examined by taking advantage of the polarization of the lasers. The femtosecond photo-association spectroscopy technique was used by the Dantus group to study the formation of an excited mercury dimer. Their experiment is depicted in Figure 10. Two ground-state mercury atoms absorbed the pump (bind) pulse (at 312 nm) forming a wave packet in the excited D state. The probe pulse depleted the fluorescence from the newly formed molecules. The time of association was determined to be within the laser pulse duration (<60 fs). The dimers were found to have a narrow rotational distribution which together with the wavelength of the binding pulse reflects the impact parameter (geometry) of the collision. Calculations reveal that the 'optimal'

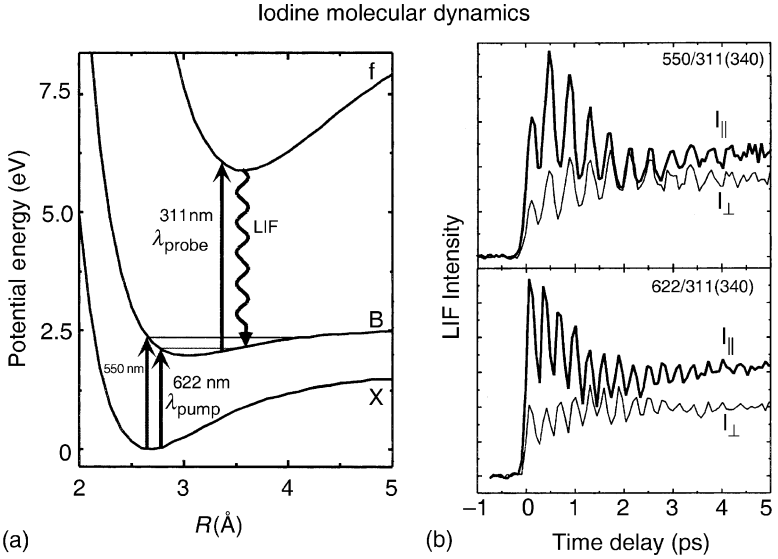


Figure 11 Molecular dynamics of iodine using two different pump laser wavelengths. (a) The pump laser, either at 550 or 622 nm, causes an excitation from the ground state to the B excited state. The choice of wavelength determines the vibrational levels which are excited on the B state. In either case, the probe at 311 nm causes an excitation to the f excited state which fluoresces around 340 nm. (b) Parallel and perpendicular pump-probe transients were obtained using each pump wavelength. At higher pump energies (550 nm), the molecule vibrates more slowly (higher vibrational level) and is reflected by the 410 fs vibrational period observed in the data. At lower pump energies (622 nm), the molecule vibrates faster and corresponds to the 300 fs vibrational period observed in the data.

binding pulse for this association reaction is 10 fs; such short pulses would overlap the entire absorption band from the repulsive ground state to the bound D state.

Bimolecular reactions can also be studied by preparing a low-temperature cluster that contains the reagents of interest. For example, a cluster containing one HI and one CO_2 molecule can be used to study the reaction $H + CO_2 \rightarrow OH + CO$. Initiation of the reaction takes place when the pump laser cleaves the H–I bond and the H atom attacks the CO_2 molecule present in the cluster. The reaction dynamics are probed as a function of time delay between pump and probe pulses as for other pump-probe experiments. The femtosecond dynamics of this reaction have been studied independently by Wittig's and Zewail's research groups.

Studies of Molecular Dynamics

I_2-Pump-Probe, Rotation and Vibration (Bound Surface)

Aside from observing chemical reactions, using ultrafast spectroscopy allows for studying the motion of isolated molecules to elucidate the internal motions of molecules themselves and to extract spectroscopic parameters. Iodine is one of the most frequently used molecules in femtosecond spectroscopy. Here we highlight only a small subset of the femtosecond studies on I_2 that have been conducted by Zewail's and other research groups. Using a pump beam at 550 nm, ground state I_2 is excited to the B state; probing with 310 nm excites the molecule to an even higher bound potential (f state) from which fluorescence at 340 nm can be obtained (as shown in Figure 11). By changing the pump-probe time delay, oscillations on the B excited state are clearly visible and show a vibrational period of approximately 410 fs. Excitation at 622 nm produces a wave packet lower in the potential well, with a vibrational period of approximately 300 fs. The differences observed for parallel and perpendicular probing correspond to the early rotational motion of the molecules. Fourier transfer analysis of the data yields the energy of each vibrational state that forms part of the vibrational wave packet. Different pump wavelengths can be used to access different vibrational levels such that vibrational frequencies and rotational constants can be obtained for a wide range of vibrational levels of the potential. This information can then be used to refine the B excited state potential energy curve of iodine.

Different Measurements Possible with Three-Pulse FWM

Ground and excited state molecular dynamics

Molecular dynamics can also be examined with four-wave mixing techniques. Once again we use iodine as a model system to illustrate that unlike the previous experimental transients of iodine obtained with pump-probe measurements, both excited and

ground state dynamics of I_2 are revealed with FWM. Materny, Kiefer, and co-workers have carried out coherent anti-Stokes Raman spectroscopy and degenerate four-wave mixing measurements on this molecule using two pulses overlapped in time followed by a third pulse after a time delay; both ground and excited state dynamic are observed in the Fourier transform of data. Dantus and co-workers explored all the pulse sequences possible with three-pulse FWM using molecular iodine as a model system (see below).

Double-sided Feynman diagrams and third-order response functions can be used to describe the interaction of the three pulses with the molecular system in three-pulse four-wave mixing. These mathematical tools are described in several books on nonlinear optical spectroscopy (see Further Reading). Figure 12 illustrates how different pulse sequences reveal different molecular dynamics. The transient grating method (Figure 12a) shows both ground (160 fs oscillations) and excited state (310 fs oscillations) wave packet dynamics. When data collection occurs over long time delays, the transient grating signal persists with a similar time constant to that of the lifetime of the excited state. The photon echo pulse sequence (shown in Figure 12b) reveals only excited state dynamics (310 fs oscillations). Note that at longer time delays, the signal decays exponentially with a time constant that reflects the rate of electronic coherence dephasing. This fundamental decay parameter indicates how long molecules remember the phase of the electric field that acted on them.

Off-resonance transient grating molecular dynamics

Off-resonance transient grating can be used to study the molecular motion of molecules in the ground state. The pulse sequence used is shown in Figure 6b. With this method it is possible to examine the rotational and vibrational dynamics of diatomic and polyatomic molecules. The vibrational dynamics belong to the Raman-active modes. From the rotational recurrences, one can obtain rotational constants with great accuracy. Dantus and co-workers have studied carbon dioxide at room temperature and in flames. At high temperatures, rotational recurrences narrow because more rotational levels are populated and the rotational recurrence moves towards longer times because of centrifugal distortion. Further studies on carbon dioxide and carbon disulfide have shown that strong fields can impart an increase in initial alignment (due to torque along the polarization vector of the laser) as well as molecular deformation such as bending (due to electronic state mixing). Heterodyne detected off-resonance transient grating data for CO_2 at room temperature are shown in Figure 13. The data closely resemble the schematic illustration in Figure 5 where rotational dynamics were introduced. The data in Figure 13 have been analyzed to yield the B and D rotational constants which give the structural parameters of the molecules and the centrifugal distortions, respectively. The type of analysis given here for CO_2 can be done on molecules

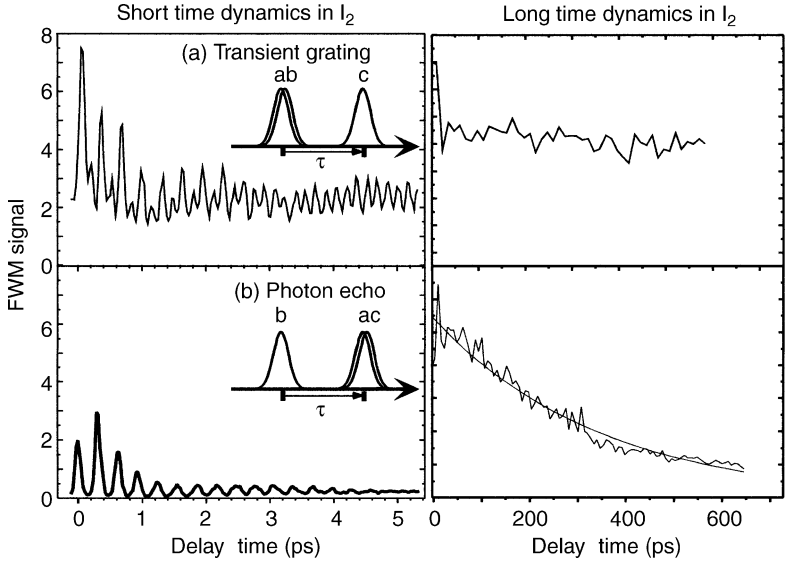


Figure 12 FWM signal in molecular iodine using two pulse sequences at short and long delay times. (a) Transient grating – E_a and E_b overlapped in time followed by E_c. Similar amounts of ground (208 cm^{-1}) and excited (105 cm^{-1}) state dynamics are observed with short delay times. In the transients, these correspond to oscillations of 160 fs and 307 fs respectively. (b) Photon echo – E_a and E_c overlapped in time but preceded by E_b. Excited state dynamics are observed in the transient with short delay times. At long time delays, the decrease in the signal reflects (a) the lifetime of the excited state or (b) the rate of the electronic coherence dephasing.

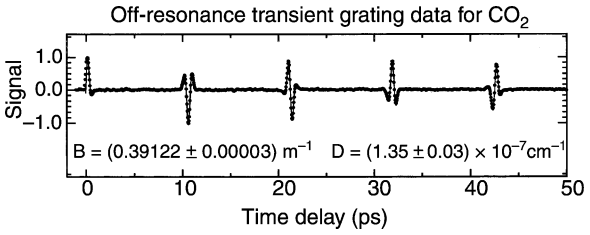


Figure 13 Heterodyne detected off-resonance transient grating on carbon dioxide at room temperature. The experimental data (dots) reflect the rotational motion of the CO_2 molecule. The data have been analyzed (the fit is shown as a line that goes through the experimental data points) to obtain the rotational constant (B) and the centrifugal distortion constant (D).

and molecular fragments resulting from chemical reactions. This method can also be used for molecular identification.

Perspectives for the Future

Most of this article has focused on the observation of ultrafast dynamics involved in chemical reactions. Presently, it is possible to generate pulses with durations as short as 5 fs. Such short pulses are capable of observing motion involving frequencies near 3000 cm^{-1}. This implies that one can observe the dynamics of most chemical bonds. The availability of shorter pulses will allow one to observe electronic dynamics. Advances in laser technology have also produced ultrashort pulses in the infrared region as well as in the vacuum ultraviolet and X-ray region. The wide spectrum available permits experiments that follow specific chemical bonds or specific transitions in elements. This in turn gives a wide range of possibilities for probing ultrafast chemical reaction dynamics.

Having pulses that are shorter than the duration of the dynamics they probe allows one to consider influencing the evolution of a chemical reaction as it is taking place. Controlling chemical reactions with lasers is not a new theme; however, the availability of ultrashort laser pulses has given new power to these efforts. Here we present a brief sample of some of the current efforts.

Control Using Three-Pulse Four-Wave Mixing

Coherent control of chemical reactions depends on the relative phase of two different laser pulses that interact with the sample in order to optimize the transfer of population between electronic states. The three-pulse four-wave mixi\ng technique can be utilized to attain coherent control; the specific timing between the pulses can be used to achieve near-unity or near-zero values of the transfer of population for molecules that

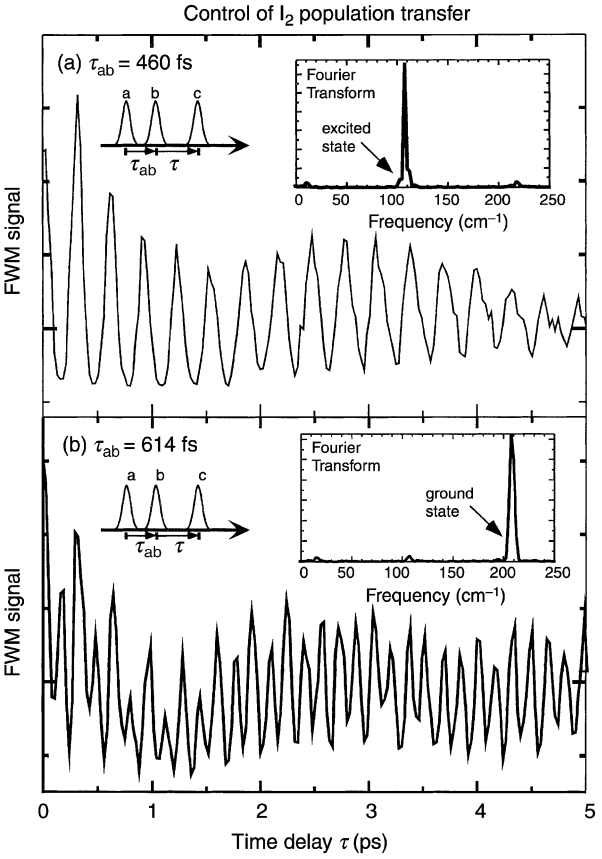


Figure 14 FWM signal in molecular iodine demonstrating control over observation of molecular dynamics. By using the pulse sequence shown, τ_{ab} was defined to be either (a) 460 fs or (b) 614 fs while τ was scanned in time. As is clearly evident in the data and the FFT, excited-state dynamics alone (307 fs oscillations) are observed with $\tau_{ab} = 460$ fs and ground-state dynamics (160 fs oscillations) are observed with $\tau_{ab} = 614$ fs.

interacted with the two electric fields. The Dantus group has also used three-pulse FWM to demonstrate control over optimizing the population transfer between two electronic states in iodine. The populations of the ground and excited states are controlled based on the time delay between the first two pulses τ_{ab} in the pulse sequence shown in Figure 6d. The transient is obtained as a function of the time delay between the second and third pulses, which probes the vibrational dynamics of the population; this experiment is illustrated in Figure 14. For Figure 14a using $\tau_{ab} = 460$ fs, only excited-state dynamics are observed; while for Figure 14b using $\tau_{ab} = 614$ fs, primarily ground-state dynamics are observed. This observation is confirmed by the Fourier transform of the time-resolved transient. The top transient has a vibrational frequency of 108 cm^{-1}, corresponding to the vibrational frequency of the excited B state. The bottom transient has a vibrational

frequency of 208 cm^{-1}, corresponding to the vibrational frequency of the ground state of iodine molecules.

Control Using Feedback and Evolutionary Algorithm

For complex molecules, one seldom knows the precise sequence of pulses or phase modulation of a given pulse that can optimize the yield of a chemical reaction. In these cases, a pulse-shaping device which manipulates the phase and amplitude of frequencies within an ultrashort pulse can be used. The device is part of a learning loop that includes a computer and an experimental setup. The learning algorithm running in the computer generates different pulses to be evaluated experimentally. Signal from the experiments is sent back to the computer for further refinement. The learning algorithm iteratively arrives at the optimum pulse. This method, first introduced by Rabitz, was first tested by Wilson. Gerber and co-workers have demonstrated automated control over the photodissociation reaction of organometallic compounds using an evolutionary algorithm with feedback. The organometallic compounds chosen were iron pentacarbonyl ($Fe(CO)_5$) and dicarbonylchloro(η^5-cyclopentadienyl)iron ($CpFe(CO)_2Cl$). For the first molecule, multiphoton absorption of femtosecond pulses at 800 nm caused ionization and fragmentation resulting in numerous ionic species – Fe^+, $Fe(CO)^+$, $Fe(CO)_2^+$, $Fe(CO)_3^+$, $Fe(CO)_4^+$, and $Fe(CO)_5^+$. These species were detected by mass spectrometry; the measured signal was used as feedback for the evolutionary algorithm that adjusts the phase of the femtosecond laser pulse. The $Fe^+$and $Fe(CO)_5^+$ exit channels were used to demonstrate the control available over branching ratios with this automated control method. Maximization (5.0) of the $Fe(CO)_5^+/Fe^+$ ratio was achieved with very short laser pulses while minimization (0.057) of this ratio was achieved with long (ps) pulses. These results were found to correspond well to changing the pulse duration of bandwidth-limited pulses. The optimization procedure did not attempt to control the other ionic species. This same automated optimization procedure was used to control the branching ratio of $CpFeCOCl^+/FeCl^+$ for the second organometallic compound. The maximum branching ratio was 4.9; the minimum was 1.2; and bandwidth-limited pulses yielded 2.4. Unlike the simpler carbonyl compound, simple changes to the pulse duration of a laser pulse could not optimize this branching ratio. The evolutionary algorithm with the feedback from the mass spectrometer allowed for these tailored pulses to be obtained and optimized the desired exit channel without any knowledge of the specific molecular system. The full potential of shaped femtosecond pulses has yet to be realized. It is possible that shaped femtosecond pulses will play an important role in future time-resolved experiments where they will be used to improve time resolution or to access potential energies that could not be reached with simpler pulses. Pulse shapers will also play a role in the preparation of ultrashort pulses with single femtosecond or attosecond duration.

List of Units and Nomenclature

Å	angstrom (10^{10} Å = 1 meter)
Attosecond	10^{18} as = 1 second
B	rotational constant
cm^{-1}	wavenumbers
D	centrifugal distortion constant
electron volts	1 eV = 8065 cm^{-1}
fs	femtosecond (10^{15} fs = 1 second)
FTS	femtosecond transition state spectroscopy
FWM	four-wave mixing
k	wavevector
LIF	laser-induced fluorescence
nm	nanometer (10^9 nm = 1 meter)
ps	picosecond (10^{12} ps = 1 second)
R(Å)	internuclear distance (given in angstroms)
$t = 0$	initiation of reaction (pump and probe are overlapped in time)
τ	time delay between pump and probe pulses being applied
TG	transient grating
ω_j	rotational frequency

See also

Chemical Applications of Lasers: Non-Linear Spectroscopies; Pump and Probe Studies of Femtosecond Kinetics. **Coherent Control:** Experimental; Theory. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** Four-Wave Mixing; Ultrafast and Intense-Field Nonlinear Optics; $\chi^{(3)}$ – Third-Harmonic Generation. **Spectroscopy:** Nonlinear Laser Spectroscopy; Raman Spectroscopy. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses. **Ultrafast Technology**: Femtosecond Condensed Phase Spectroscopy: Structural Dynamics.

Further Reading

Bandrauk AD, Fujimura Y and Gordon RJ (eds) (2002) *Laser Control and Manipulation of Molecules*, vol. 821 Washington, DC: ACS Press.

Chergui M (ed.) (1996) *Femtochemistry: Ultrafast Chemical and Physical Processes in Molecular Systems*. Singapore: World Scientific.

Dantus M (2001) Coherent nonlinear spectroscopy: from femtosecond dynamics to control. *Annual Reviews of Physical Chemistry* 52: 639–679.

De Schriber FC, De Feyter S and Schweitzer G (eds) (2001) *Femtochemistry: With the Nobel Lecture of A. Zewail*. Weinheim, Germany: Wiley-VCH.

Douhal A and Santamaria J (eds) (2002) *Femtochemistry and Femtobiology*. Singapore: World Scientific.

Goswami D (2003) Optical pulse shaping approaches to coherent control. *Physics Reports* 374: 385–481.

Grimberg I, Lozovoy VV, Dantus M and Mukamel S (2002) Femtosecond three pulse spectrocopies in the gas phase: density matrix representation. *Journal of Physical Chemistry* 106: 697–718.

Manz J and Woste L (eds) (1995) *Femtosecond Chemistry*, vols. I and II. Heidelberg, Germany: Verlag Chemie.

Mukamel S (1995) Principles of Nonlinear Optical Spectroscopy. New York: Oxford University Press.

Sundstrom V (ed.) (1997) *Femtochemistry and Femtobiology*. Singapore: World Scientific.

Zewail AH (ed.) (1994) *Femtochemistry – Ultrafast Dynamics of the Chemical Bond*, vols. I and II. Singapore: World Scientific.

Femtosecond Condensed Phase Spectroscopy: Structural Dynamics

E T J Nibbering, Max Born Institut, Berlin, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Ultrafast Chemistry

Chemistry means rearrangement of atoms in/between molecules. Chemical bonds are broken and/or formed, leading to a change in chemical composition, from reactants to products, during a reaction. Following the temporal behavior of the reaction all the way from reactant – via possible intermediates – towards the product state(s), amounts to deciphering the molecular processes that lead to these chemical reaction dynamics. Molecular rearrangements from reactant to product may encompass many orders of magnitude in time (see **Figure 1**). Typically the larger and more involved the rearrangements are, the more time it takes for a molecular species to reach the final state. Rearrangements of biomolecular structure occur in micro- to milliseconds (spontaneous folding of small proteins consisting of several hundreds of amino acid units), or longer (DNA multiplication, protein synthesis in ribosomes). In these conformational dynamics, hundreds to thousands of chemical bonds change their nature. Bimolecular reaction dynamics in liquid solution occur on time-scales of nanoseconds or longer, being controlled by the relatively slow diffusional motions of the reaction partners to each other. Many elementary processes in chemistry, however, occur on much faster time-scales. In fact, when one considers the dynamical event of a single-bond rearrangement, the relocation of the (relatively few) atoms appears to take place as an ultrafast event. Bond fission, hydrogen and proton transfer, electron transfer, and cis/trans-isomerizations have been found to occur on femto- to picosecond time-scales ('femtochemistry'). The dynamics of these elementary processes are not only determined by the energy landscapes of the reacting partners. The energy landscapes (potential energy surfaces) are determined by molecular parameters such as relative orientation (distances, angles) of the reaction partners, the energy levels of the reactant, intermediate and product species, and – very important – the energy barriers represented by transition states. When considering condensed phase reaction dynamics, the important role of the surrounding solvent has been recognized. The solvent shells modulate energy levels of the reactant, intermediate and product states, through electrostatic interactions. In the liquid phase the solvent fluctuations lead through these interactions to fluctuating energy levels of the reaction partners. In addition, the solvent may facilitate chemical reactions through energy exchange with the reaction partners, leading to efficient dissipation of excess energy, making chemical reactions often irreversible. The processes that are extremely dominated by these solvent interactions (such as molecular collisions, electronic and vibrational dephasing, vibrational relaxation) are also often found to occur on femto- to picosecond time-scales. The outcome of a bond fission may be a null effect, when the dissociating fragments are forced by the surrounding solvent to recombine and relax ('cage effect').

Spectroscopy has been a prime tool to obtain insight into these key molecular processes. Reactant, intermediate, and products contribute to the spectra with their respective molecular resonances. With spectral domain spectroscopy one may, in principle, obtain a precise determination of structural information. Fluctuations of energy levels of these species by the solvent interactions, and finite lifetimes, contribute to a broadening of these molecular resonances. Using steady-state spectroscopy one

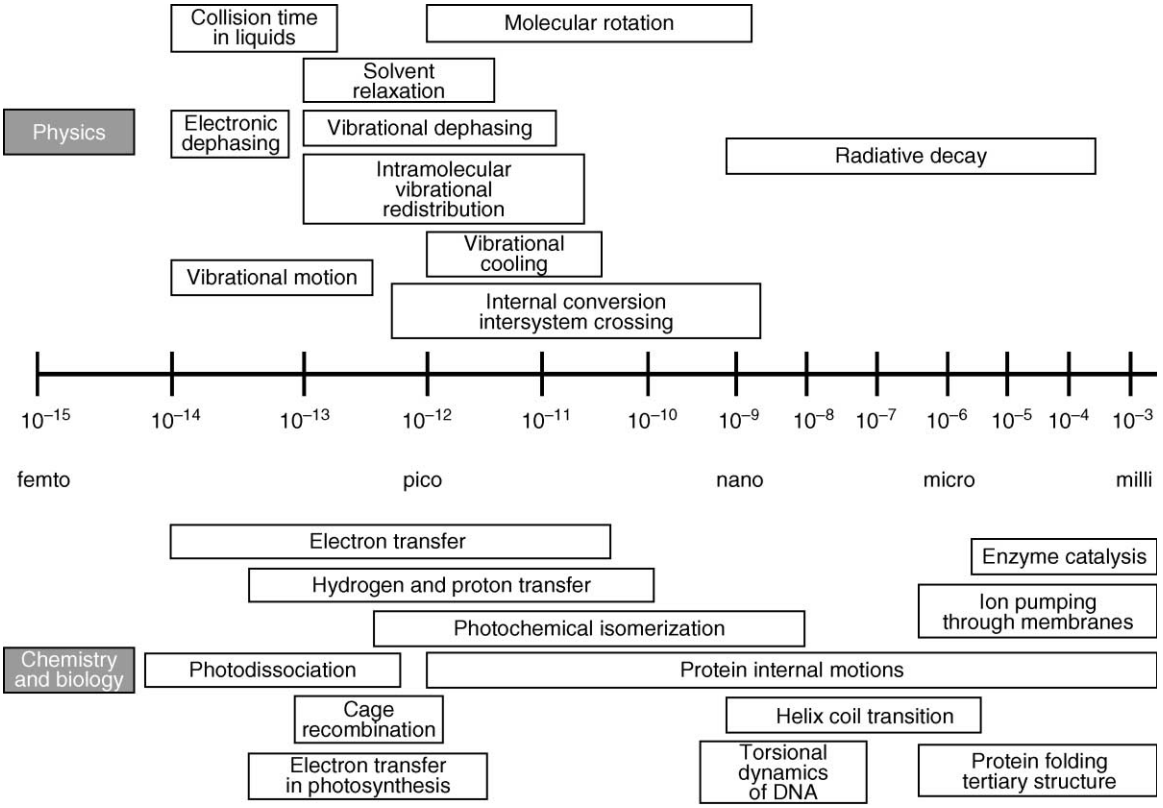


Figure 1 Chart showing the ultrafast time-scales of fundamental physical and chemical processes (inspired by charts of Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. Oxford, UK: Clarendon Press and Zewail AH (2000) Femtochemistry: atomic-scale dynamics of the chemical bond. Nobel lecture as appeared in *Journal of Physical Chemistry A* 104(24): 5660–5694).

could try to determine these effects through a lineshape analysis. Using this approach, the results have to be analyzed with an a priori defined model representing the dynamics. The disadvantage lies in the averaging over all time-scales inherent in spectral domain spectroscopy, where details of transient states may be lost in this averaging procedure. In contrast, in experiments designed to grasp the molecular dynamics in real time, the temporal resolution is the prerequisite quantity in observation of these processes. In time-resolved experiments one can trigger a chemical reaction at a well-defined point in time, after which one can follow the conformational changes from reactant – via possible intermediates – to the product, and identify any possible transition states along the reaction pathways.

During the last century the technological advances in time-resolved spectroscopy have prompted significant breakthroughs in the study of chemical reaction dynamics. Initially the temporal resolution was given by the duration of flashes from light bulbs ('flash' spectroscopy) or by rapid mixing of reaction partners (stopped-flow technique), as developed by Nobel Laureates Eigen, Norrish, and Porter, and many others. These techniques allow a temporal resolution of at best milliseconds. Already, up to the early 1960s, it was understood that elementary reaction dynamics occur on much faster time-scales. Fourier spectroscopy may improve time resolution down to microseconds or slightly less. With the advent of pulsed laser sources with ever-increasing temporal resolution, a wide arena of time-resolved spectroscopic techniques has emerged. Nanosecond laser systems are known since the early 1960s, picosecond pulses can be generated since the 1970s, the first femtosecond laser system was reported in 1981, and currently laser engineering groups are breaking into the attosecond domain. The majority of pulsed laser systems operate at visible and near-infrared wavelengths. This means that usually chemical events in ultrafast time-resolved spectroscopy are initiated by electronic excitation ('photochemistry'). Subsequent molecular rearrangements have then to be followed by probing the electronic states through electronic resonances (UV/VIS pump–UV/VIS probe or UV/VIS four-wave mixing

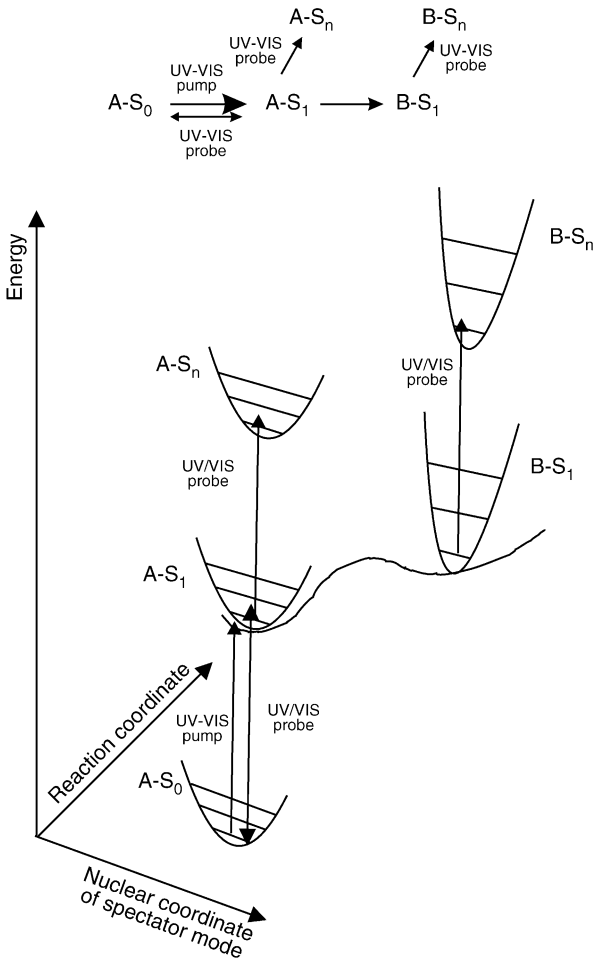


Figure 2 Ultrafast time-resolved electronic spectroscopy. A UV/VIS pump pulse excited an electronic resonance, and a probe pulse determines, depending on the tuning in the UV/VIS population changes in the reactant ground or excited states or in the product state.

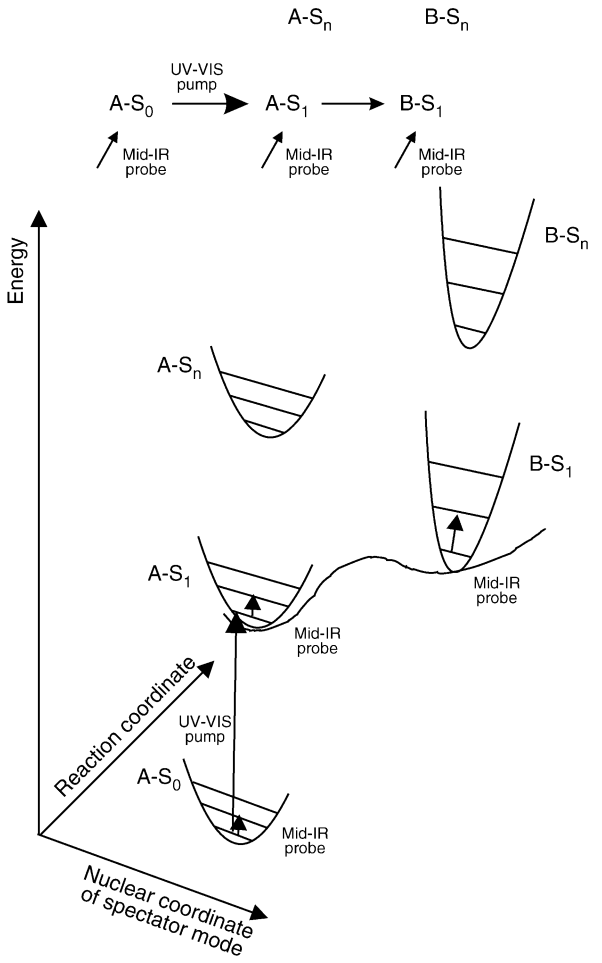


Figure 3 Ultrafast time-resolved infrared spectroscopy. A UV/VIS pump pulse excites an electronic resonance, after which the populations are followed with an infrared probe pulse by inspection of marker modes in the different electronic states.

spectroscopy) (Figure 2), or via vibrational resonances (UV/VIS pump–IR/Raman probe spectroscopy) (see Figures 3 and 4).

Laser Technology

Since the beginning of the 1990s, a major advancement in femtosecond laser technology has emerged by the discovery of Kerr lens mode locking in laser oscillators with Ti:sapphire as lasing material. Ultrastable and ultrashort laser pulses tunable between 700 and 1,000 nm can be generated with durations as short as 6 fs. The second advancement is due to the ability of Ti:sapphire as laser amplifying material, using the method of chirped pulse amplification. Pulses with output powers up to several W, with repetition rates in the kHz-regime, are now routinely generated. These developments facilitate the efficient generation of ultrashort pulses tunable from the UV to the mid-infrared, and even beyond to the THz-regime using schemes based on nonlinear optics. Frequency conversion of the fundamental output of an amplified Ti:sapphire laser system is then achieved with nonlinear processes such as self-phase modulation and supercontinuum generation, n^{th}-order harmonic generation ($n = 2, 3, 4,\ldots$), sum and difference frequency generation, or parametric generation and amplification. As a result, one can now almost arbitrarily tune the excitation and probe wavelengths of the applied laser pulses. Additional parameters, that one may alter, are pulse duration (from less than 10 fs up to several ps) and pulse energy (in the mJ range or less; typically one uses for

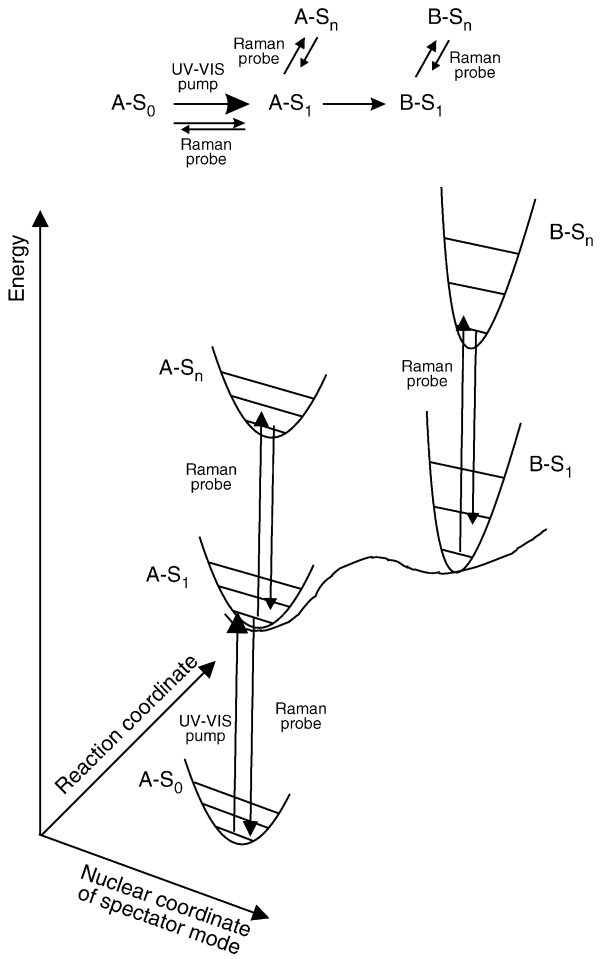


Figure 4 Ultrafast time-resolved Raman spectroscopy. A UV/VIS pump pulse excites an electronic resonance, after which the populations are followed with (resonance) Raman probe pulse by inspection of marker modes in the different electronic states.

condensed phase spectroscopy pulse energies in the range of nJ to μJ). More advanced pulse manipulation schemes use amplitude and phase masks. The latter aspect holds the promise of amplitude and phase control of molecular excitation, feeding the ultrafast chemist's dream of laser control of molecular reaction dynamics.

Ultrafast Electronic Spectroscopic Techniques

Pulsed laser sources enable a variety of experimental approaches to determine transient states. Already, since the development of nanosecond laser sources, time-resolved pump–probe and four-wave mixing spectroscopy have been widely used. In a first approach, often one uses these forms of spectroscopy to probe in real-time populations of reactant,

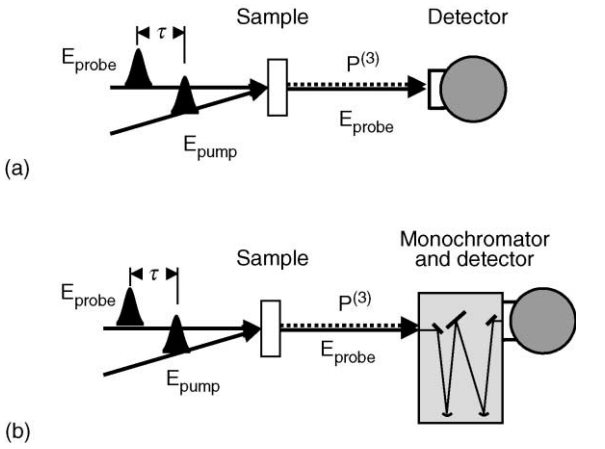


Figure 5 Pump–probe technique: A pump pulse induces an absorbance change in a sample, that is subsequently measured by a time delayed probe pulse, either spectrally integrated (a) or spectrally resolved (b).

transient, and product states as function of pulse delay. With the advent of pico- and femtosecond laser sources, coherence properties of the material response have been investigated as well.

The following techniques have been widely used in ultrafast electronic condensed phase spectroscopy.

UV/VIS Pump–UV/VIS Probe Spectroscopy

In this technique a 'pump' pulse resonant to a transition promotes the molecules to an electronic excited state, and the time-dependent populations are determined by measurement of the transmission of a time-delayed 'probe' pulse (Figure 5). The absorbance change $\Delta A(\tau) = -^{10}\log[T(\tau)/T(\tau = 0)]$ (with T for transmission change and τ the pulse delay) is a direct indication of populations of states. Here one can tune the probe pulse to the same transition as the pump pulse, and one measures an absorbance decrease due to a bleach of ground state population (i.e., fraction of molecules being excited), and due to molecules in the excited state stimulated back to the ground state. One could also tune the probe pulse to transitions between the transient state and higher lying states. For electronic spectroscopy, often transitions between the ground and first excited state ($S_0 \rightarrow S_1$) overlap with excited state absorptions ($S_1 \rightarrow S_n$). One has to rely then on numerical analysis procedures such as singular value decomposition and decay associated spectra, that correlate spectral features with temporal behavior.

The time resolution of the experiment is given by the cross-correlation between pump and probe pulses. In the case of extremely short pulses one has

to consider the effects of group velocity dispersion between different frequency components due to different travel times through the samples, leading to pulse temporal broadening. For the case of pump and probe tuned to different wavelengths, the time-resolution is usually dominated by group velocity mismatch between pump and probe pulses.

When ultrashort laser pulses are used, often the applied spectral bandwidth enables simultaneous excitation of several vibrational sublevels in vibronic transitions. Due to the phase-relationship of the applied laser pulses, these vibrational sublevels are prepared in states that are well-defined in temporal phase with respect to each other. Such a coherent superposition of vibrational eigenstates is nothing less than a vibrational wavepacket, that will evolve in time. The occurrence of vibrational wavepacket motions leads to oscillatory modulations of the pump–probe signals.

Time-Resolved Fluorescence Spectroscopy

Here again a 'pump' pulse excites an electronic resonance. The transient excited states are now probed through their spontaneous fluorescence emission, either by use of time-correlated single photon counting or by fluorescence upconversion with a gating pulse in a nonlinear medium (Figure 6). Single-photon counting has a time resolution of at best a few picoseconds. Fluorescence upconversion can be performed with a time resolution of 100–200 fs. The advantage lies in the fact that the measured signals are more likely due to the specific excited states only, and not corrupted by other contributions.

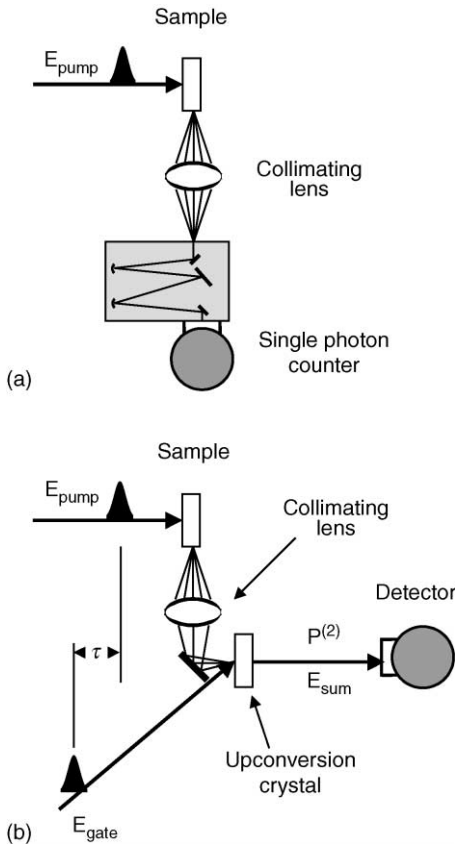


Figure 6 Time-resolved emission detection: The fluorescence emission is either temporally resolved by time-correlated single photon counting (a) or by up-conversion with a gating pulse (b).

Electronic Four-Wave Mixing Spectroscopy

In time-resolved degenerate electronic four-wave mixing spectroscopy, where all the three applied laser pulses (with wavevectors $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_3$) are tuned to the same electronic resonance, one can detect a nonlinear signal in the phase-matched direction $\mathbf{k}_S = \mathbf{k}_3 + \mathbf{k}_2 - \mathbf{k}_1$ (Figure 7). The first laser pulse generates a coherent superposition between the electronic ground and excited states. This electronic coherence evolves in time during the period (denoted as coherence time τ), until the second pulse converts this coherence in a population frequency grating in the electronic ground and excited states. After a second time period of free evolution (denoted as population or waiting time T), this frequency grating is again transformed into an electronic coherence, which after some time evolution generates a macroscopic polarization in the phase-matched directions. This macroscopic polarization is either directly measured by a time-integrating detector (in which case one measures the intensity signal proportional to the absolute square of the nonlinear polarization), or one uses another light pulse that mixes as local oscillator with the nonlinear polarization (and the signal is proportional to the amplitude of the polarization). The time resolution is determined by the third-order autocorrelation of the applied pulses.

Depending on the delay between the three input pulses, and which delay one varies, different dynamics can be measured. When one performs an experiment with τ scanned and T set at zero, one measures the electronic coherence decay (due to electronic dephasing) time in a two-pulse photon echo experiment. When, on the other hand, τ is set to zero, and T is scanned, one measures the evolution of a population grating, and electronic lifetimes can be deduced. In a three-pulse photon echo experiment both pulse delays are scanned, and the signals give insight into both electronic coherence and frequency grating decay (due to electronic dephasing and spectral diffusion, respectively). A novel extension of four-wave mixing technique is the three-pulse echo peak shift (3PEPS). In a 3PEPS measurement, the delay time τ_{max} is

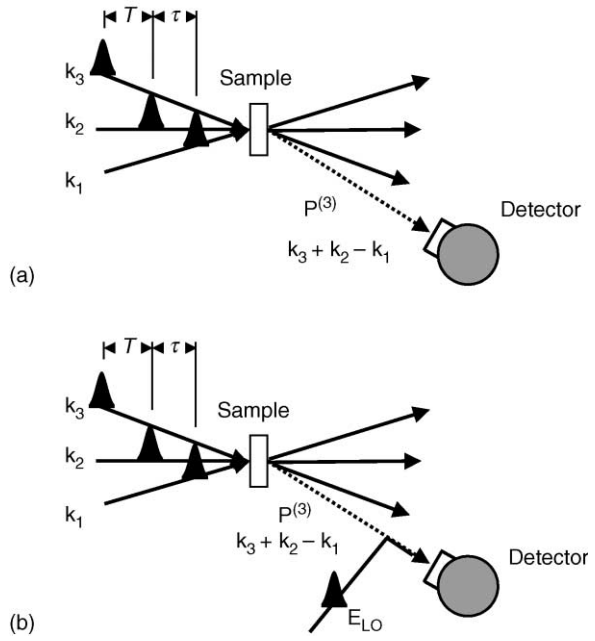


Figure 7 Beam geometry for time-resolved four-wave mixing (grating/photon echo) spectroscopy: The nonlinear signal is transmitted by the sample in a phase-matched direction and its intensity signal (a) or its amplitude by interference with a local oscillator field (b) is detected by a time-integrating detector.

determined, where the echo signal has a maximum along the coherence delay τ, as a function of the population delay T. This method provides direct insight into the frequency fluctuation correlation function that governs the linear and nonlinear signals. Extension of the method to multi-color four-wave mixing spectroscopy has been demonstrated.

Electronic Condensed Phase Spectroscopy: Femtochemistry and Solvation Dynamics

In the narrow sense of meaning, 'femtochemistry' deals with the real-time observation of nuclear motions during chemical reactions. Quantum mechanics teaches us that this is only possible when molecular systems are prepared in vibrational superpositions. Only in the case of quantum states describing vibrational wavepackets evolving in time the narrow-defined positions of nuclei allow for a real-time description of molecular rearrangements, or in other words: dynamics. This has been demonstrated in the gas-phase landmark experiments by Nobel Laureate Zewail (Figure 8).

In the condensed phase at room temperature, however, often such well-defined quantum states are hard to prepare, since the ensembles of molecular systems exist in mixed states. Coherent laser excitation on an ensemble will then prepare mixed states as well, and then information that one obtains from the experiments is less explicit. In addition, room temperature liquid solutions induce extremely fast fluctuations in the states of the molecules under study and as a consequence the electronic and vibrational coherences show ultrafast decaying behavior. This limits the temporal window with which one can follow the evolution of vibrational wavepackets.

Most of the ultrafast spectroscopic experimental work performed on photochemistry in the condensed phase, being described as 'femtochemistry', only deals with the time-scales of changes in electronic states (usually indicated with reaction rates typified with exponentially decaying or rising functions). In this regard it is better to speak of 'kinetics' rather than 'dynamics', although both terms are used in condensed phase spectroscopy. Due to the relative broad electronic spectra of condensed phase molecules masking any structural detail, it is hard to derive nuclear motions in real time, and as such it is difficult to make any statements about dynamics of molecular structures during chemical reactions.

Electronic resonances have also been used to probe the dynamical interactions between a nonreactive solute and the surrounding fluctuating solvent. The idea behind these experiments is to achieve information about the time-scales of solvent fluctuations on (and solvent motions due to a change in) the electronic charge distribution in a molecule (as induced by an electronic excitation). Usually the solvent motions are characterized with a frequency fluctuation correlation function (Figure 9). This information is highly relevant for the case of reacting molecules, where electronic motions accompany the relocations of nuclei, even in electron transfer reactions, where the nucleic rearrangements are modest. Solvent motions have been shown to often control reaction rates. Solvent rearrangement, as a response to a change in electronic charge distribution, is known as solvation dynamics. Experimental probes for solvent motional fluctuations are electronic coherence decay (electronic dephasing) and population frequency grating decay (spectral diffusion) that can be measured in two- and three-pulse photon echoes. Solvent adaptation to new charge distributions can be followed with the time-dependent Stokes-shift in fluorescence emission, as well as in three-pulse photon echoes (in so-called three-pulse echo peak shift measurements described above). Ultimately these experiments resolve the temporal characteristics of the frequency fluctuation correlation function.

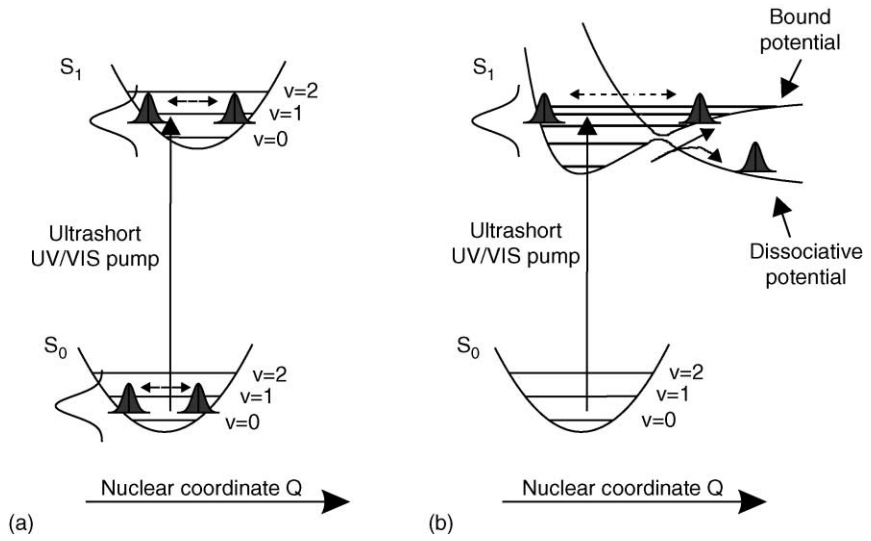


Figure 8 Ultrafast excitation with broadband pulses creates coherent superpositions of vibrational eigenstates, i.e., vibrational wavepackets. These wavepackets are generated in electronic excited and ground states (a), and their time-dependence can be determined in pump–probe and four-wave mixing experiments. When level crossing occurs to another electronic state, the vibrational wavepacket may be observed to evolve along the reaction coordinate (b).

Ultrafast Vibrational Spectroscopy in Photochemistry: Structural Dynamics

Vibrational spectroscopy has, despite a smaller signal sensitivity due to smaller transition cross-sections, several advantages over electronic spectroscopy. Vibrational transitions have typically smaller bandwidths than electronic transitions, due to longer dephasing times (the exception to the rule: O–H/O–D stretching bands of hydrogen bonded hydroxyl groups). Specific signal contributions are thus easier tractable in vibrational spectroscopy than in electronic spectroscopy. Vibrational bands can often be correlated to specific vibrational motions by inspection of the transition frequencies, (e.g., O–H, N–H, and C–H stretching bands can be found in the 3000 cm^{-1} range, C=O and C=N stretching modes are located near 1,600–1,750 cm^{-1}, O–H, N–H, and C–H bending vibrations between 1,500–1,650 cm^{-1}, C–O stretching mode around 1,200–1,300 cm^{-1} etc.), and, in particular, the frequency range between 1,000 and 1,800 cm^{-1} is called the fingerprint region because of this reason. Measurement of vibrational bands thus leads to identification of particular vibrational motions, and conclusions can be drawn on specific structural motifs in the molecules. Vibrational bands can be infrared (IR) or Raman active, sometimes both, and both IR and Raman techniques have been widely applied in structural determination studies ranging from small molecules to larger biomolecular systems. In ultrafast photochemistry one thus excites the molecules with a UV/VIS pump pulse, and an IR/Raman probe pulse follows the outcome of the chemical reaction by inspection of vibrational bands of reactant, transients, and products.

Vibrational spectroscopy has the potential of revealing site-specific information if the marker modes are due to nuclear motions of specific molecular side-groups. For instance, hydrogen bonding induces marked shifts of O–H, N–H, C=O, and C=N bands. Observation of changes in spectral shifts reveals important information on hydrogen bond interactions (weakening/strengthening or even hydrogen bond cleavage). Infrared spectroscopy is able to probe small molecular species in solution that typically have their electronic resonances in the far-UV (making electronic spectroscopy impossible, since normally the solvent would absorb this radiation) (Figure 10). For instance, in acid–base neutralization reactions, where the acidity of a so-called photoacid is switched on by use of a UV pulse, a VIS probe pulse would only be able to probe the photoacid (or its conjugate photobase), thus revealing only when a proton leaves the photoacid. An infrared pulse can, besides probing vibrational resonances of the photoacid, also probe bands of a small base (or its conjugated acid) indicating when a proton arrives at the base.

In the case where the vibrational normal modes do not allow such a structural insight into site-specific groups, one can make a comparison of the

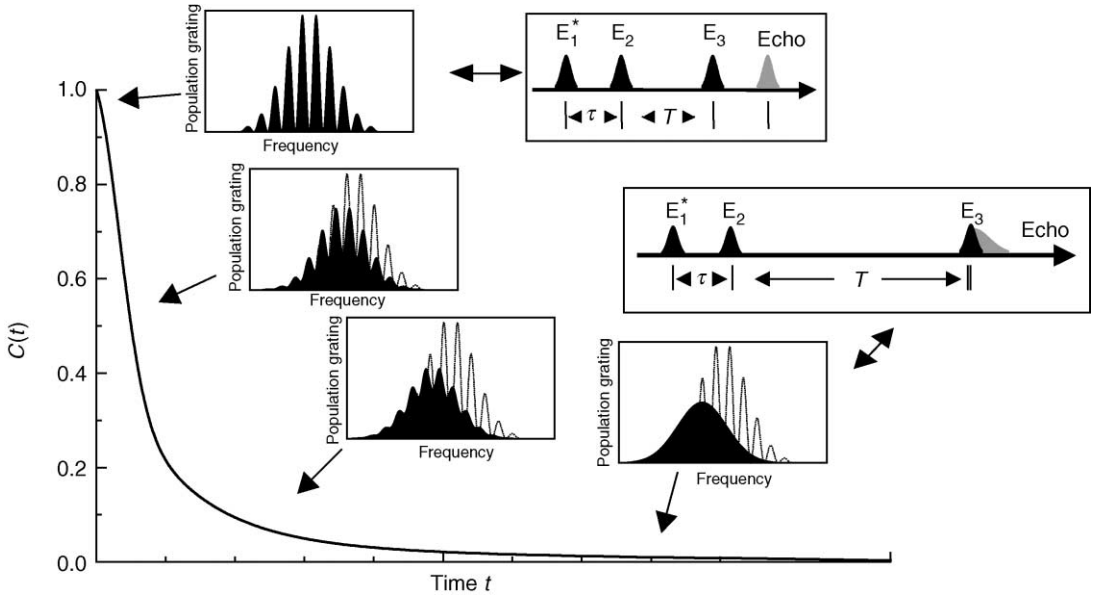


Figure 9 Solute–solvent interactions as characterized by the transition frequency fluctuation correlation function $C(t)$. This quantity can be measured in photon echo and time-resolved fluorescence experiments. In a three-pulse photon echo experiment phase information as given by the excited state frequency grating created with the first two interactions, is washed out due to solvent fluctuations and Stokes shifts to lower frequencies due to solvent rearrangement. A similar plot can be drawn for the ground state frequency grating.

experimentally observed vibrational bands with predictions made by quantum chemical calculations. When a full correspondence between experiment and theory is possible, one can make statements about the three-dimensional structure. With current quantum chemical calculational routines, such as density functional theory, medium-sized molecules are routinely calculated for electronic ground-state conformations. Reliable results of transient and product states in electronic excited states can be obtained with the routine *ab initio* complete active space self-consistent field (CASSCF), albeit for mid-size molecules of at most on the order of about 20 atoms. New developments in numerical procedures, such as time-dependent density functional theory (TD-DFT) may prove fruitful in the calculation of larger molecular systems in electronic excited states.

During recent years, femtosecond IR spectroscopy has been used in photo-induced chemical reactions ranging from excited state hydrogen and proton transfer, transformations of photochromic switches, excited state charge transfer, and cis–trans isomerizations. Here an ultrashort UV/VIS 'pump' pulse promotes the molecule to an electronic excited state, and the reaction is followed by a measurement of the absorbance change of a 'probe' pulse tuned in the mid-IR region where the vibrational marker modes sensitive to structural changes have their resonances. Experimentally, one performs spectrally resolved transient IR spectroscopy. Femtosecond IR parametric devices deliver pulses with bandwidths of 150 cm^{-1} or more. In order to be able to observe shifts as small as the linewidths of IR-active vibrations, one usually measures the IR absorbance change with a detector after spectral dispersion with a monochromator. As a side-effect of this spectral dispersion ground-state bleach signals often appear to grow in at negative pulse delay with the dephasing time of the transition. This effect, known as 'perturbed free induction decay' is a common feature of spectrally resolved nonlinear pump–probe spectroscopy of bleached transitions with dephasing times much longer than the time resolution of the experiment. The time resolution of the experiment is given by the cross-correlation between the UV/VIS-pump and IR-probe pulse (about 100–200 fs), and is typically dominated by group velocity mismatch in samples with thicknesses of about 100 μm.

In principle, the same approach can be followed by probing Raman-active vibrations. In this case the spectral resolution is not only determined by the monochromator through which the spontaneous Raman emission dispersed, but also by the bandwidth of the gating pulse by which the Raman effect is induced. As a result, UV/VIS pump–Raman probe

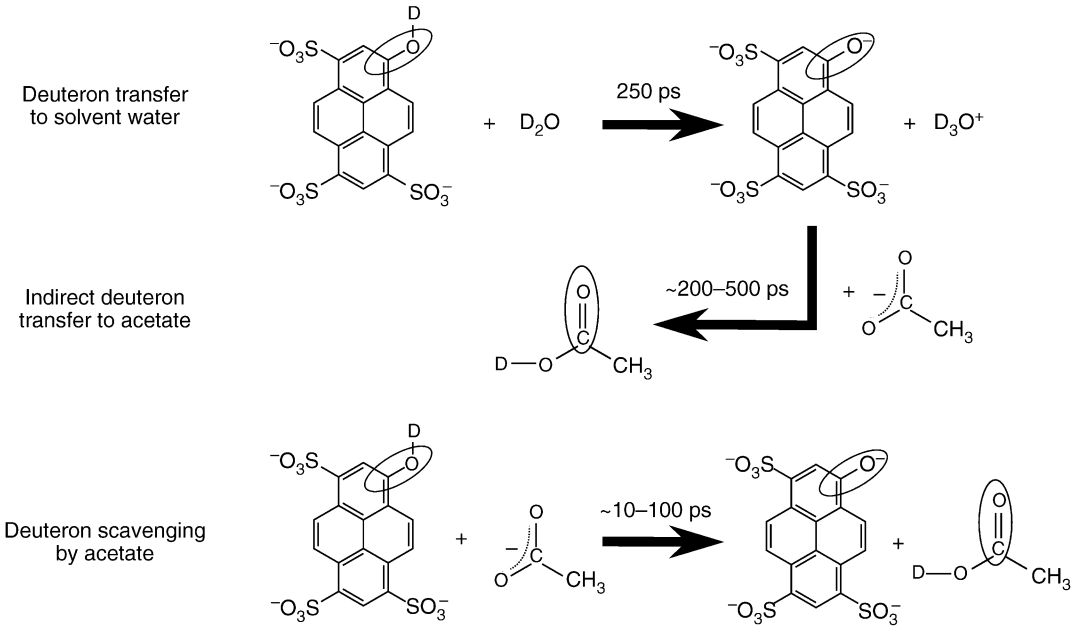


Figure 10 In excited state acid–base neutralization experiments, the proton dissociation can be followed on two sides with ultrafast infrared spectroscopy. Vibrational marker modes of photoacid and conjugated photobase indicate when the proton leaves the acid. The marker mode of acid formed by proton pick-up by the accepting base is a direct measure of the arrival time of the proton at the base. Depending on relative concentrations, either the indirect proton transfer via the solvent or direct scavenging of the proton by the base dominates the dynamics.

spectroscopy has a temporal resolution of around 1 ps (a compromise between spectral and time resolution). Due to the even weaker cross-sections of Raman transitions, one often uses resonance enhancement by tuning the gating pulse close to or resonant with an electronic transition of the state that is probed. The advantage is that one usually only observes Raman bands of the state under inspection. The drawback is that fluorescence resulting from the resonant electronic excitation often inhibits detection of Raman bands in extended spectral ranges.

Raman spectroscopy offers an additional insight into how chemical reactions evolve. By comparison of the intensities of anti-Stokes and Stokes lines of a particular vibration, it is possible to derive time-dependent changes in the excitation level of this vibration. One can draw conclusions on whether particular modes initially drive a chemical reaction ('state transition promoting modes'), or only get excited after the transition is made, by taking up the excess energy released by the reaction ('accepting modes'). In such a way, insight is obtained on the energy flow inside a molecular system (intramolecular vibrational redistribution, abbreviated as IVR) and vibrational energy dissipation to the surrounding solvent (vibrational cooling). Infrared spectroscopy is less powerful in revealing this, since red-shifted transient absorption of vibrational bands may either mean that the particular mode is highly excited (the red-shift is a consequence of the diagonal anharmonicity of the vibration), or it may mean that other modes are highly excited and these cause a red-shift of the marker mode under inspection (the red-shift is then due to off-diagonal anharmonic coupling with other modes).

Equilibrium Structural Dynamics in the Electronic Ground State

Existence of anharmonic coupling between vibrational modes means that the vibrational motions are not decoupled from each other. These couplings always exist, since otherwise no IVR and vibrational cooling (no relaxation) would occur. When the couplings are significantly large, excitation of one specific mode will induce a significant instantaneous shift of other vibrations anharmonically coupled to this particular mode. Estimation of the magnitude of these couplings should lead to a determination of the curvatures of potential energy surfaces along the respective coordinates. With current state-of-the-art femtosecond infrared technology (with time resolution of about 150 fs or less) it is possible to excite

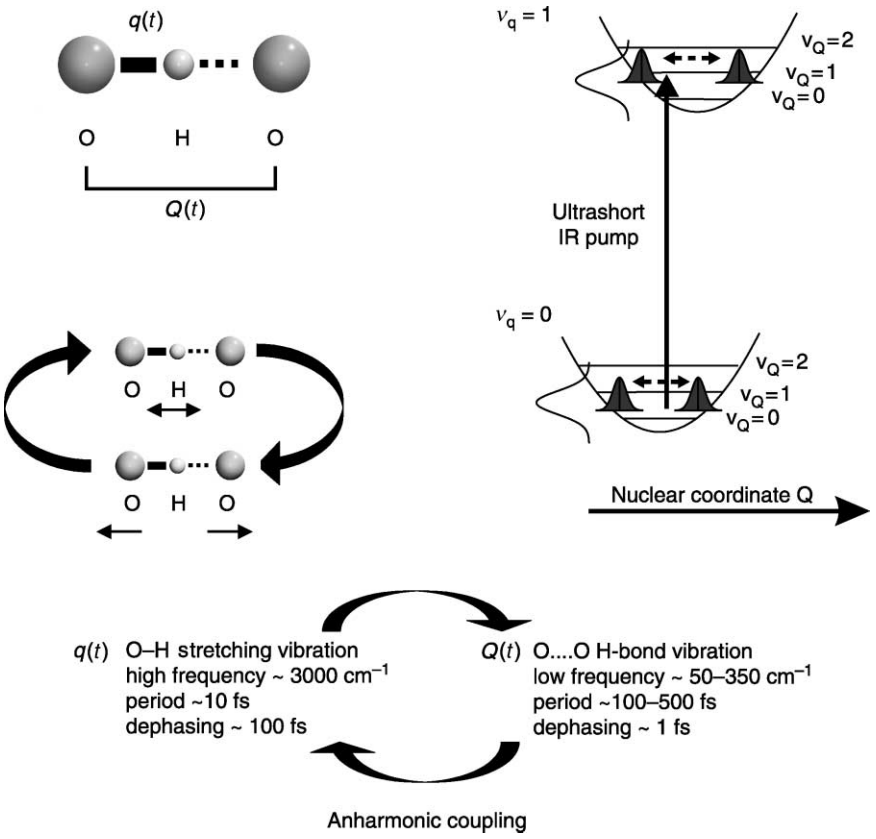


Figure 11 In hydrogen-bonded hydroxyl groups a strong anharmonic coupling exists between the hydroxyl stretching and modes modulating the hydrogen-bond distance. A Born–Oppenheimer-like separation of time-scales of vibrational motion between the high-frequency O–H/O–D stretching and the low-frequency hydrogen bond modes enables a description with potential energy surfaces for the low-frequency mode as a function of the quantum state of the hydroxyl stretching vibration. The formation of low-frequency vibrational wavepackets in ultrafast IR spectroscopy of O–H/O–D stretching modes is then fully analogous to wavepacket formation in ultrafast electronic spectroscopy.

infrared-active vibrations and to probe the same or other vibrations (either in the IR or by the Raman process), enabling insight into the anharmonic couplings between these vibrational modes.

Direct anharmonic couplings are probed in IR-pump–IR-probe and IR photon echo spectroscopy (Figure 11). In the case of peptides, the excitonic (Davydov) couplings between amide I vibrations of the amino acid units have been the subject of extensive study since these lead to spatial information of the relative orientations of the different amide I modes, and thus, of the spatial orientation of the amino acid units inside the peptide. Exploring the effects of the delay between the IR pulses on the observed signals, give insight into the fluctuations of these orientational features, or in other words, into structural dynamics of peptides. In the case of hydrogen bonded O–H/O–D stretching vibrations, the anharmonic couplings with under-damped low-frequency modes modulating the hydrogen bond distance, lead to marked modulations of observed pump–probe and echo signals as function of pulse delay. These experiments reveal which modes couple strongly to the hydrogen stretching oscillator.

Fluctuations in couplings between vibrational modes (in particular with those of the solvent) induce dephasing of the vibrational coherences. The dephasing and spectral diffusion dynamics of vibrational transitions, e.g., the O–H stretch vibration of HOD dissolved in D_2O, have been determined by IR photon echo spectroscopy. From these studies the time-scales of structural memory decay of hydrogen bond networks can be derived. Couplings between vibrational modes ultimately lead to population relaxation through intramolecular vibrational redistribution effects and vibrational cooling. Ultrafast two-color IR-pump–IR-probe reveals whether excitation of one IR-active vibration is followed by population transfer into another IR-active mode.

With IR-pump–Raman-probe, one can estimate whether Raman-active vibrations are transiently excited. In the case of hydrogen-bonded O–H/O–D stretching bands, an extremely short population relaxation time T_1 (below 1 ps) is observed, as well as an ultrarapid dephasing time T_2 (200 fs or less). These values are definitely longer for other vibrations, from which it has been concluded that the strong anharmonicities in hydrogen-bonded vibrational systems enable dephasing and relaxation processes to have such a rapid impact.

Future Developments

Future activities in the determination of dynamically evolving molecular structures in the condensed phase, may involve the aforementioned application of amplitude and phase controlled excitation pulses in optimal control of chemical reactions. Vibrational spectroscopy may be used as a tool in determination of evolving structures after electronic excitation by a shaped pulse tuned in the UV/VIS. Infrared pulses could also be shaped in amplitude and phase, and may be used in the exploration of steering of chemical reactions in the electronic ground state. Multidimensional vibrational spectroscopy is an object of extensive research. Extension of the method to determination of structure of transient states will lead to new results on molecular rearrangements, e.g., polypeptides reaching a new equilibrium after inducing a UV-excitation induced geometric change of a photochromic switch incorporated in the peptide structure. Much activity is currently also being put in the development of ultrafast structure resolving techniques such as electron diffraction, X-ray diffraction, and X-ray spectroscopy. While still extremely demanding at a technological level, final implementation of these latter techniques will open up new areas of ultrafast chemical dynamics.

See also

Chemical Applications of Lasers: Pump and Probe Studies of Femtosecond Kinetics.

Further Reading

Cerullo G and De Silvestri S (2003) Ultrafast optical parametric amplifiers. *Review of Scientific Instruments* 74(1): 1–18.

de Boeij WP, Pshenichnikov MS and Wiersma DA (1998) Ultrafast solvation dynamics explored by femtosecond photon echo spectroscopies. *Annual Review of Physical Chemistry* 49: 99–123.

Elsaesser T and Bakker HJ (eds) (2002) *Ultrafast Hydrogen Bonding Dynamics and Proton Transfer in the Condensed Phase*. Dordrecht, Germany: Kluwer.

Elsaesser T and Kaiser W (1991) Vibrational and vibronic relaxation of large polyatomic molecules. *Annual Review of Physical Chemistry* 42: 83–107.

Fayer MD (ed.) (2001) *Ultrafast Infrared and Raman Spectroscopy*. New York: Marcel Dekker.

Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. Oxford, UK: Clarendon Press.

Fleming GR and Cho M (1996) Chromophore-solvent dynamics. *Annual Review of Physical Chemistry* 47: 109–134.

Fleming GR and Hänggi P (eds) (1993) *Activated Barrier Crossing: Applications in Physics, Chemistry and Biology*. Singapore: World Scientific.

Kaiser W (ed.) (1988) *Ultrashort Laser Pulses and Applications*. Berlin: Springer.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. Oxford, UK: Oxford University Press.

Nibbering ETJ and Elsaesser T (2004) Ultrafast vibrational dynamics of hydrogen bonds in the condensed phase. *Chemical Reviews* 104: 1887–1914.

Owrutsky JC, Raftery D and Hochstrasser RM (1994) Vibrational relaxation dynamics in solutions. *Annual Review of Physical Chemistry* 45: 519–555.

Sundström V (ed.) (1996) *Nobel Symposium: Femtochemistry and Femtobiology: Ultrafast Reaction Dynamics at Atomic-Scale Resolution*. Singapore: World Scientific.

Wiersma DA (ed.) (1994) *Femtosecond Reaction Dynamics*. Amsterdam: North-Holland.

Zewail AH (2000) Femtochemistry: atomic-scale dynamics of the chemical bond. Nobel lecture as appeared in *Journal of Physical Chemistry A* 104(24): 5660–5694.

Web-information: http://www.nobel.se
Nobel web-site showing information on Nobel Laureates M Eigen, RGW Norrish and G Porter (The Nobel Prize in Chemistry 1967: 'for their studies of extremely fast chemical reactions, effected by disturbing the equilibrium by means of very short pulses of energy') and AH Zewail (The Nobel Prize in Chemistry 1999: 'for his studies of the transition states of chemical reactions using femtosecond spectroscopy').

Ultrafast Illumination and Processing

Y Fainman, University of California, San Diego, CA, USA

D M Marom, Bell Laboratories, Holmdel, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Mode-locked lasers emitting ultrashort pulses have expanded the range of available instruments and techniques to the scientific community. The unique properties of these ultrashort pulses – their high peak powers and short duration – enable such diverse applications as multiphoton microscopy, pump-probe experiments, and coherent control of molecular and atomic wavepackets, to name a few. Ultrashort pulses can also be used as a basis for ultrahigh data rate optical communications using time division multiplexing (TDM) in a fiber communication application. These ultrafast applications often require complementary real-time optical processors based on instantaneous nonlinear phenomena, since conventional electronic means are limited by the response time of electronic devices.

During the last few years, we have developed real-time optical processors utilizing nonlinear three- and four-wave mixing of signals originating from spatial or temporal channels for synthesizing, processing, and detecting ultrafast waveforms. We perform real-time optical signal processing that can be applied to data conversion between slow parallel channels in space and an ultrafast serial signal. A time-to-space mapping processor has been built for detection of ultrafast waveforms. It mixes two spatially inverted temporal frequency information waves in a three-wave mixing arrangement, generating a quasi-monochromatic spatial signal carrying the temporal image for detection by slower electronic means. A space-to-time mapping processor was constructed for generation of arbitrary ultrafast waveforms from an input ultrashort pulse. The waveform is a time-scaled replica of the employed input spatial image, converted by a four-wave mixing arrangement. By introducing only temporal information channels to our four-wave mixing processors, real-time optical processing of temporal waveforms was enabled. We have demonstrated time reversal experiments based on performing spectral phase conjugation and spectral inversion operations, achieving time reversal of the electrical field and of the complex amplitude waveform, respectively. The three- and four-wave mixing arrangements are performed within a spectral processing device – a free-space optical setup consisting of diffraction gratings and lenses – utilizing a $\chi^{(2)}$ nonlinear crystal. A cascaded second-order nonlinearity technique is used for four-wave mixing within the $\chi^{(2)}$ crystal (a frequency-sum generation process followed by a frequency-difference generation process satisfying the type-II noncollinear phase matching condition). However, a $\chi^{(3)}$ nonlinear crystal may be readily substituted without loss of generality (albeit with a likely smaller conversion efficiency).

This article reviews the spatial/temporal processing techniques for synthesizing, processing, and detecting ultrafast waveforms. The generation of spectrally decomposed waves within a spectral processing device is described in the following section. The subsequent section articulates on the employed nonlinear wave mixing processes, and the ensuing sections express the synthesis, processing, and detection of ultrafast waveforms.

Generation and Characteristics of Spectrally Decomposed Waves

Our wave mixing experiments are performed with spectrally decomposed waves, in which the spectral frequency content is spatially dispersed. When the pulse is decomposed with high resolution, the signal at every spatial location can be treated as quasi-monochromatic. This allows for increased interaction length within a nonlinear crystal, allowing us to use long crystals, whereas typically the interaction length is limited by temporal walkoff. In this section we analyze the creation of spectrally decomposed waves.

We introduce an ultrashort optical waveform, with a temporal envelope of $s(t/\tau)$, into the spectral processing device, where τ parameterized the waveform's duration. The waveform is propagating in free space towards a diffraction grating, at an incidence angle θ (see Figure 1). The input ultrashort waveform is characterized in its coordinate system (x_1, z_1) as:

$$E_{\text{pulse}}(x_1, z_1; t) = w\left(\frac{x_1}{L_1}\right) s\left(\frac{t - t_0 - z_1/c}{\tau}\right) \times \exp\left(\mathrm{i}\left(\omega_0 t - \frac{\omega_0}{c} z_1\right)\right) \qquad [1]$$

where $w(\cdot)$ defines the transversal field distribution with a scale size parameter L_1, t_0 is a time delay parameter, c is the speed of light in vacuum, and ω_0 is the center optical frequency. The waveform is

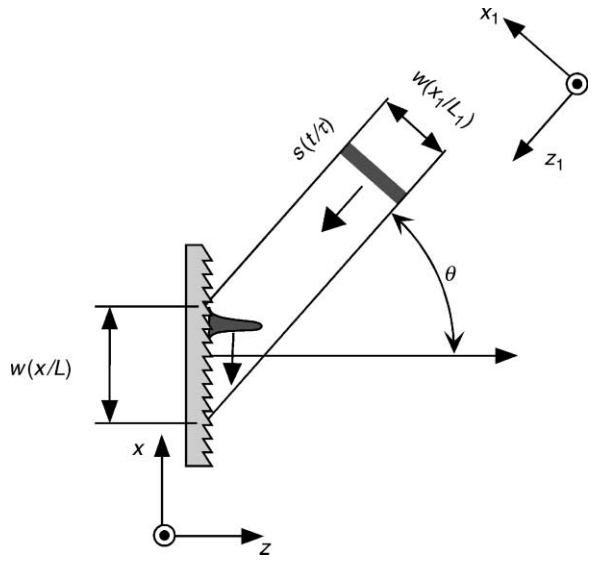


Figure 1 An ultrashort optical waveform, $s(t/\tau)$, propagating in free-space is incident upon a diffraction grating at angle θ. The transverse spatial mode of the pulse is stretched due to the projection onto the grating.

propagating in the z_1 direction at a group velocity and phase velocity of the speed of light in vacuum. The spatial confinement in the y_1 direction is omitted for brevity. We perform a time-domain Fourier transform (TFT) on eqn [1] to perform a frequency decomposition of the ultrafast waveform, yielding:

$$\tilde{E}_{\text{pulse}}(x_1, z_1; \omega) = \tau w\left(\frac{x_1}{L_1}\right) \times \exp\left(-\mathrm{i}\frac{\omega}{c} z_1\right) \tilde{s}(\tau(\omega - \omega_0)) \times \exp(-\mathrm{i}(\omega - \omega_0)t_0) \quad [2]$$

where the tilde over the script denotes a FT relationship, defined for the TFT as $\tilde{s}(\omega)A \int s(t)\exp(-\mathrm{i}\omega t)\mathrm{d}t$. Next we perform a rotation of the coordinate system from (x_1, z_1) to (x, z), for compatibility with the coordinate system of the spectral processing device, yielding:

$$\tilde{E}_{\text{pulse}}(x, z; \omega) = \tau w\left(\frac{x\cos\theta - z\sin\theta}{L_1}\right) \times \exp\left(-\mathrm{i}\frac{\omega}{c}(x\sin\theta + z\cos\theta)\right) \times \tilde{s}(\tau(\omega - \omega_0))\exp(-\mathrm{i}(\omega - \omega_0)t_0) \quad [3]$$

To find the incident field on the input grating of the spectral processing device, the field of eqn [3] is evaluated at $z = 0$. The effect of the grating diffraction can be modeled by adding the grating momentum, k_g, in the x direction to the k-vector that characterizes the propagation direction of the wave, resulting in:

$$\tilde{E}_{\text{input}}(x, 0; \omega) = \tau w\left(\frac{x\cos\theta}{L_1}\right) \times \exp\left(-\mathrm{i}\left(\frac{\omega}{c}\sin\theta - k_g\right)x\right) \times \tilde{s}(\tau(\omega - \omega_0))\exp(-\mathrm{i}(\omega - \omega_0)t_0) \quad [4]$$

The grating momentum, k_g, and the incidence angle, θ, are chosen such that the center frequency ω_0 will diffract in the direction of the optical axis of the system. Setting $\omega_0/c \cdot \sin\theta = k_g$ and substituting the grating's k-vector, $k_g = 2\pi/\Lambda$, where Λ is the grating period, yields $\sin\theta = \lambda_0/\Lambda \equiv \alpha$, where λ_0 is the optical wavelength corresponding to the center carrier frequency. To further simplify the notation, we define a new spatial width scaling parameter $L = L_1/\cos(\theta)$, accounting for the aperture size increase due to the projection onto the diffraction grating. With these substitutions, the expression takes on the more familiar form, given by

$$\tilde{E}_{\text{input}}(x; \omega) = \tau w\left(\frac{x}{L}\right)\exp\left(-\mathrm{i}\frac{(\omega - \omega_0)\alpha x}{c}\right) \times \tilde{s}(\tau(\omega - \omega_0))\exp(-\mathrm{i}(\omega - \omega_0)t_0) \quad [5]$$

We may perform an inverse TFT on eqn [5] to characterize the input signal in the time domain, yielding:

$$E_{\text{input}}(x; t) = w\left(\frac{x}{L}\right)s\left(\frac{t - t_0 - \alpha x/c}{\tau}\right)\exp(\mathrm{i}\omega_0 t) \quad [6]$$

Equation [6] describes the input waveform scanning across the fixed aperture at velocity of c/α in the x-direction.

The spectrally decomposed wave is generated by performing a spatial Fourier transform (SFT) employing a lens of focal length f on the input wave (either frequency domain representation, eqn [4], or time domain representation, eqn [6]). Using the frequency domain representation, the spectrally decomposed wave U_{SDW} of the input waveform is given by

$$U_{\text{SDW}}(x'; \omega) = \tau\tilde{s}(\tau(\omega - \omega_0)) \times \exp(-\mathrm{i}(\omega - \omega_0)t_0)\int_{-\infty}^{\infty} w\left(\frac{x}{L}\right) \times \exp\left(-\mathrm{i}(\omega - \omega_0)\frac{\alpha x}{c}\right) \times \exp\left(-\mathrm{i}2\pi\frac{xx'}{\lambda_0 f}\right)\mathrm{d}x \quad [7]$$

where we use the central wavelength in the Fourier transform kernel. Evaluating the integration of eqn [7], yields:

$$U_{\mathrm{SDW}}(x';\omega) = L\tau\tilde{s}\left(\tau(\omega-\omega_0)\right)\exp(-\mathrm{i}(\omega-\omega_0)t_0) \times \tilde{w}\left(\frac{\alpha L}{2\pi c}\left[\frac{\omega_0 x'}{\alpha f}+(\omega-\omega_0)\right]\right) \quad [8]$$

where the SFT of the input pupil function, $\tilde{w}(f_x)A\int w(x)\exp(-\mathrm{i}2\pi x f_x)\mathrm{d}x$, has mixed space-domain and time-domain frequency variables in its argument, characterizing the dispersion. A spectral component shifted away from the center frequency, $\Delta\omega = \omega - \omega_0$, will be centered at a displaced location $\Delta x' = -\Delta\omega\cdot\alpha f/\omega_0$, defining the spatial dispersion term $\partial x'/\partial\omega = -\alpha f/\omega_0$. One can see that in order to get fine localization of each spectral component, the spatial extent of the function $\tilde{w}(\cdot)$ must be minimized. This can be readily achieved by increasing the size of the input aperture, or the parameter L, as well as increasing α (grating's angular dispersion). Let us make the following observation on system resolution: The ultrafast pulse's spectral width is proportional to $1/\tau$, and its spatial extent is therefore proportional to $\alpha f/\omega_0\tau$ (multiplying by the spatial dispersion term). The spatial width of each spectral component is proportional to $cf/\omega_0\tau$. The dimensionless ratio of these two spatial widths, which is indicative of the number of resolvable spectral components, is $N \sim \alpha L/c\tau$, where N is defined as the system spectral resolution. When the spectral processing device is designed to satisfy the requirement of $N \gg 1$, then the spectrally decomposed wave can be considered quasi-monochromatic at every spatial position. In our experiments, we use $\lambda_0 = 800$ nm, gratings of 600 lines/mm frequency ($\Lambda = 1666$ nm), $L \approx 1$ cm, and $\tau \approx 100$ femtoseconds, yielding $\alpha \cong 0.5$ and $N \sim 160$, easily satisfying the high spectral resolution requirement. Let us make a second observation on system resolution using the time domain representation of eqn [6]: The spatial extent of the input aperture is proportional to L, whereas the instantaneous spatial extent of the scanning waveform is proportional to $\tau c/\alpha$ (see **Figure 2**). The dimensionless ratio of these two spatial widths is again $N \sim \alpha L/c\tau$, exemplifying the concept of resolvable positions and demonstrating the conservation of the space-bandwidth product.

To further facilitate the ensuing analysis, let us assume that we operate in the high-resolution regime (satisfied with the values given above), and we seek to simplify the expressions defining the spectrally decomposed wave. The spectrally decomposed wave

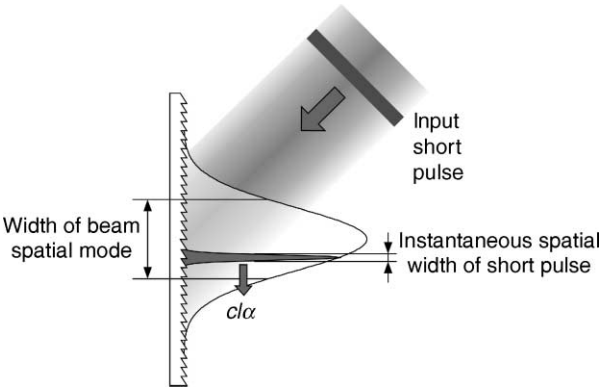


Figure 2 The input short pulse scans across the diffraction grating at velocity of c/α. The ratio of the width of the spatial mode of the beam to the instantaneous spatial width of the scanning waveform determines the system resolution. © 2001 Optical Society of America.

can be defined in the time domain by performing a SFT on eqn [6], yielding:

$$\begin{aligned} U_{\mathrm{SDW}}(x';t) &= \int w\left(\frac{x}{L}\right)s\left(\frac{t-t_0-\alpha x/c}{\tau}\right)\exp(\mathrm{i}\omega_0 t) \times \exp\left(-\mathrm{i}\frac{2\pi}{\lambda_0 f}xx'\right)\mathrm{d}x \\ &\approx w\left(\frac{c(t-t_0)}{\alpha L}\right)\exp\left(-\mathrm{i}\frac{\omega_0(t-t_0)x'}{\alpha f}\right) \times \tilde{s}\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\exp(\mathrm{i}\omega_0 t) \end{aligned} \quad [9]$$

where by account of the high-resolution approximation, the instantaneous location of the traveling signal waveform samples the value of the aperture function and removes the aperture dependence from the Fourier integral. This approximation decouples the aperture function from the temporal waveform, and illustrates that the duration of the spectrally decomposed wave is defined by the aperture function (time of flight through aperture of size L), and that the spatial dependence is a mapping of the incident waveform spectrum. A rotating wavevector in time is also present, on account of the traveling (moving) waveform on the input plane. Again, we may perform a TFT on eqn [9], defining the high-resolution limit of the spectrally decomposed wave, yielding:

$$U_{\mathrm{SDW}}(x';\omega) = \tilde{s}\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\exp(-\mathrm{i}(\omega-\omega_0)t_0) \times \tilde{w}\left(\frac{\alpha L}{2\pi c}\left[\frac{\omega_0 x'}{\alpha f}+(\omega-\omega_0)\right]\right) \quad [10]$$

One can compare eqns [8] and [10], as both characterize the spectrally decomposed wave in the

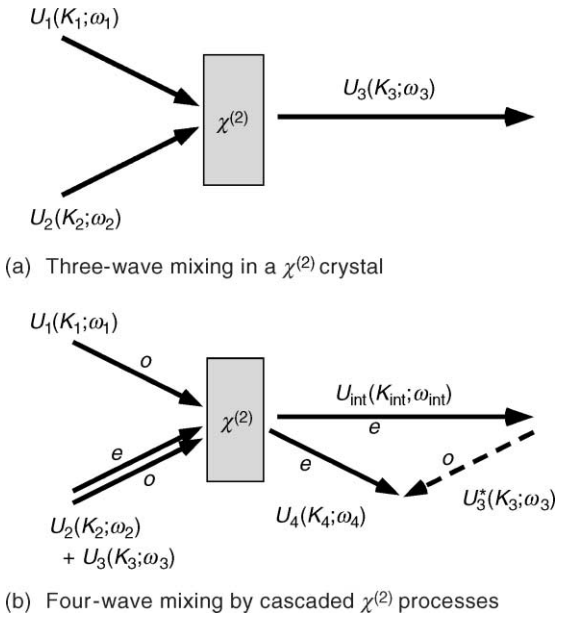


Figure 3 (a) Three-wave mixing interaction: two input waves U_1 and U_2, generate a third wave U_3. (b) Four-wave mixing by cascaded second order nonlinearities: three input waves, U_1, U_2, and U_3, generate a fourth wave U_4. The first two waves generate an intermediate wave U_{int}, by a frequency-sum process, followed by a frequency-difference process between the intermediate wave and the third input wave.

temporal frequency domain, and observe that the difference lies with the argument of the spectral term. If we apply the high-resolution approximation to eqn [8], then each spectral component would occupy a small region in space, as defined by the extent of the aperture function's SFT. Substituting for $\omega - \omega_0$ by $-\omega_0 x'/\alpha f$ for the spectral argument of eqn [8], corresponding to the high-resolution limit, equates the two expressions (essentially the same approximation as was performed in deriving eqn [9]).

Nonlinear Wave Mixing With Nondepleting Pumps

A nonlinear crystal exhibiting a large nonlinear susceptibility $\chi^{(2)}$ is placed at the Fourier plane of the spectral processing device. We employ noncollinear phase matching in our experiments. Either type-I or type-II phase matching conditions can be used in three-wave mixing experiments, with the resultant frequency-sum wave propagating along the bi-sector direction (see Figure 3a). Since both input waveforms are at ω_0 frequency, the generated field will be at a doubled carrier frequency, $2\omega_0$, due to the frequency-sum process. Four-wave mixing experiments use type-II phase matching in the $\chi^{(2)}$ crystal, involving a frequency-sum process generating an intermediate wave followed by a frequency-difference process between the intermediate wave and the third input wave. The resultant wave will copropagate with one of the input waves and at the same temporal carrier frequency (propagation direction and frequency ω_0), albeit at an orthogonal polarization state (see Figure 3b). The output signal can be extracted using a polarizer or a polarizing beam splitter. The interacting input waves propagate noncollinearly at the Fourier transform plane by offsetting the signals from each other at the input plane of the spectral processing device.

While it is possible, though often difficult, to obtain a precise analytic solution to the coupled mode equations governing the wave-mixing processes for the generated waves, it is beyond the scope of this article whose focus is on the signal processing functionality. We therefore limit the analysis to weak interactions with nondepleting input waves. Using the nondepleted input waves approximation, the nonlinear polarization term, which drives the generation of new waves, is constant and an ordinary differential equation characterizes the generation process.

The nonlinear polarization for the frequency-sum process is proportional to the product of the input waves, when expressed in the time domain: $P_{NL}^{sum}(x';t) = \chi_{eff}^{(2)} U_1(x';t) U_2(x';t)$. The nonlinear polarization for the frequency-difference process is proportional to the product of an input waves and a conjugated input wave: $P_{NL}^{diff}(x';t) = \chi_{eff}^{(2)} U_1(x';t) U_2^*(x';t)$.

The solution of the ordinary differential equation driving the generation of the new wave, takes into account the possible phase mismatch between the propagation of the nonlinear polarization and the generated wave within the nonlinear crystal medium. Since we are interested in the signal processing functionality only, we approximate the generated waves as being directly proportional to the driving nonlinear polarization. This approach is valid for obtaining the generated signal forms in the case of short nonlinear crystals (limited interaction length) satisfying the phase matching condition. In practice, the temporal and spatial bandwidths, as well as the crystal length, will introduce filtering effects which will slightly modify the output signals. However, it can be shown that when wave mixing with spectrally decomposed waves, especially mutually inverted waves, the phase matching requirement is well satisfied and the filtering effects are negligible for the employed crystal lengths. In our experiments, we typically used 2 mm long β-barium borate (BBO) crystals. Using the simplified method for obtaining

the form of the generated waves, arrangements for synthesizing, processing, and detecting ultrafast waveforms can be explored.

Detection of Ultrafast Waveforms

Ultrafast optical waveforms are too rapid for acquisition by opto-electronic means alone, due to the latter's finite response time. An optical processor is usually required to convert a finite time window (typically in the tens to hundreds of ps range) to a stationary signal in space for detection by slower means. We mix two mutually inverted spectrally decomposed waves in the Fourier plane of a spectral processing device, and spatially Fourier transform the resultant wave to the output plane. As will be shown below, when one of the spectrally decomposed waves is generated by a transform limited short pulse, the output spatial signal carries the temporal information of the input waveform. We call this process a time-to-space converter.

Let two ultrafast waveforms, $s((t-t_0)/\tau)$ and $r(t/\tau)$, enter the spectral processing device from opposite directions (one strikes the grating at angle θ and the other at $-\theta$, see **Figure 4**). Each waveform will generate a spectrally decomposed wave in the spatial Fourier transform plane. The spectrally decomposed wave of the signal $s(t)$, as defined by eqn [9], yields the first input field:

$$U_1(x';t) \approx w\left(\frac{c(t-t_0)}{\alpha L}\right)\exp\left(-\mathrm{i}\frac{\omega_0(t-t_0)x'}{\alpha f}\right) \times \tilde{s}\left(-\frac{\omega_0 \tau x'}{\alpha f}\right)\exp(\mathrm{i}\omega_0 t) \quad [11a]$$

whereas the inverted spectrally decomposed wave for $r(t/\tau)$ is given by

$$U_2(x';t) \approx w\left(-\frac{ct}{\alpha L}\right)\exp\left(\mathrm{i}\frac{\omega_0 t x'}{\alpha f}\right)\tilde{r}\left(\frac{\omega_0 \tau x'}{\alpha f}\right)\exp(\mathrm{i}\omega_0 t) \quad [11b]$$

Without loss of generality, we set the time delay parameter in eqn [11b] to zero, such that the parameter t_0 in the spectrally decomposed wave of the signal waveform is a measure of the relative time delay of the two waveforms. With full knowledge of the two input waves, U_1 and U_2, we can express the resultant waveform of the three-wave mixing process as:

$$U_3(x';t) \propto P_{\mathrm{NL}}^{\mathrm{sum}}(x';t) = w\left(\frac{c(t-t_0)}{\alpha L}\right)w\left(-\frac{ct}{\alpha L}\right) \times \exp\left(\mathrm{i}\frac{\omega_0 t_0 x'}{\alpha f}\right)\tilde{s}\left(-\frac{\omega_0 \tau x'}{\alpha f}\right)\tilde{r}\left(\frac{\omega_0 \tau x'}{\alpha f}\right) \times \exp(\mathrm{i}2\omega_0 t) \quad [12]$$

Next we apply a SFT to propagate the wave U_3 to the output plane, yielding:

$$U_{\mathrm{out}}(x'';t) = \int U_3(x';t)\exp\left(-\mathrm{i}\frac{2\pi}{\lambda_0 f/2}x'x''\right)\mathrm{d}x' = y\left(\frac{2\alpha x''}{c}-t_0\right)w\left(\frac{c(t-t_0)}{\alpha L}\right)w\left(-\frac{ct}{\alpha L}\right) \times \exp(\mathrm{i}2\omega_0 t_0) \quad [13]$$

where $y(t) = s(t/\tau) \otimes r(-t/\tau)$, as the Fourier transform of a product is the convolution of their Fourier transforms (the $\otimes$-symbol denotes the convolution operation), and we use the spatial Fourier transform kernel at the doubled carrier frequency. The stationary spatial signal, $y(\cdot)$, is proportional to the convolution of the two input waveforms, where

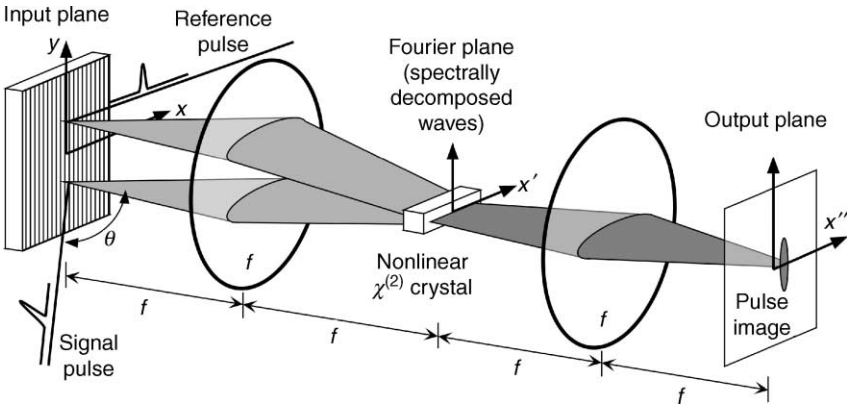


Figure 4 Arrangement for imaging ultrafast waveforms using a time-to-space conversion. The mutually inverted, spectrally decomposed waves of a signal waveform and that of a reference pulse are mixed in a nonlinear crystal placed at the Fourier plane. The generated wave is spatially Fourier transformed to the output plane, yielding a stationary image that carries the signal waveform information. © 2001 Optical Society of America.

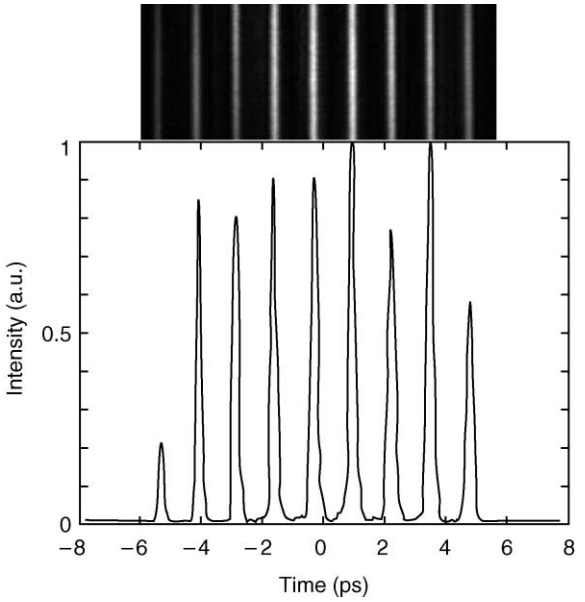


Figure 5 Output of a time-to-space converter. Top: image acquired by a CCD camera of the stationary image of a pulse sequence. Botttom: extracted time domain data from the image, showing inter-pulse time separation of 1.25 ps.

one of them is reversed in time. This mixing process was developed for time-to-space conversion by using a reference transform limited pulse for the reversed waveform. The resulting spatial image will be that of the signal waveform, albeit smoothed out due to the convolution operation with the reference pulse. The spatial signal is centered at $x'' = ct_0/2\alpha$, illustrating that the relative time delay between the two waveforms shifts the location of the output signal. The temporal characteristic is determined by the product of the time-scaled input apertures. The relative time delay, t_0, serves as an attenuation mechanism when the time windows are not perfectly aligned. When the time delay between the signal pulse and the reference pulse is too large, no output signal will emerge as the two signals do not overlap for the wave mixing process. Therefore, the time window of the spectral processing device is determined by the input aperture size, not the physical dimension of the nonlinear crystal. Consequently, we have been able to demonstrate high-resolution temporal imaging with time windows in the range of tens of picoseconds (see Figure 5). Additionally, as mentioned earlier, one of the significant advantages of this technique is the favorable phase matching condition with the spectrally decomposed wave and its inverted counterpart.

Processing of Ultrafast Waveforms

The simplest form of ultrafast waveform processing uses linear filtering, or placement of an element with transmittance characteristics that are frequency selective. However, a linear filter is a time invariant process and cannot perform more advanced processing operations of ultrafast waveforms such as interaction between ultrafast waveforms. For such processing applications, we mix three spectrally decomposed waves in the Fourier plane of a spectral processing device to generate a resultant output ultrafast waveform in real time. We illustrate advanced processing by performing a time reversal experiment using the spectral phase conjugation and spectral inversion techniques in four-wave mixing experiments.

Consider three ultrafast waveforms, $s_1(t - t_0/\tau)$, $s_2((t - t_1)/\tau)$ and $s_3(t/\tau)$, which enter a spectral processing device for generation of their respective spectrally decomposed waves. Let all three waves enter the spectral processing device from the same direction, such that the spatial dispersion is in the same direction for each signal (as defined by eqn [11a]), and generate the spectrally decomposed waves $U_1(x';t)$, $U_2(x';t)$, and $U_3(x';t)$, respectively (see Figure 6a). Since we utilize a cascaded second-order nonlinearity process, we let the first two waves interact first in a frequency-sum process, generating the intermediate field:

$$U_{\text{int}}(x';t) \propto P_{\text{NL}}^{\text{sum}}(x';t) = w\left(\frac{c(t-t_0)}{\alpha L}\right) w\left(\frac{c(t-t_1)}{\alpha L}\right) \times \exp\left(-\text{i}\frac{\omega_0(2t-t_0-t_1)x'}{\alpha f}\right) \tilde{s}_1\left(-\frac{\omega_0 \tau x'}{\alpha f}\right) \times \tilde{s}_2\left(-\frac{\omega_0 \tau x'}{\alpha f}\right) \exp(\text{i}2\omega_0 t) \quad [14]$$

It should be noted that upon a more rigorous analysis of the generated frequency-sum wave, phase matching is not well met across the temporal bandwidth of the waveform, as is the case when wave mixing mutually inverted spectrally decomposed waves. This will introduce time domain spectra filtering effects, but for simplicity these effects are ignored below. The generated frequency-sum wave interacts with the third input wave in a frequency-difference process, giving rise to the fourth wave:

$$U_4(x';t) \propto P_{\text{NL}}^{\text{diff}}(x';t) = w\left(\frac{c(t-t_0)}{\alpha L}\right) w\left(\frac{c(t-t_1)}{\alpha L}\right) \times w\left(\frac{ct}{\alpha L}\right) \exp\left(-\text{i}\frac{\omega_0(t-t_0-t_1)x'}{\alpha f}\right) \times \tilde{s}_1\left(-\frac{\omega_0 \tau x'}{\alpha f}\right) \tilde{s}_2\left(-\frac{\omega_0 \tau x'}{\alpha f}\right) \tilde{s}_3^*\left(-\frac{\omega_0 \tau x'}{\alpha f}\right) \exp(\text{i}\omega_0 t) \quad [15]$$

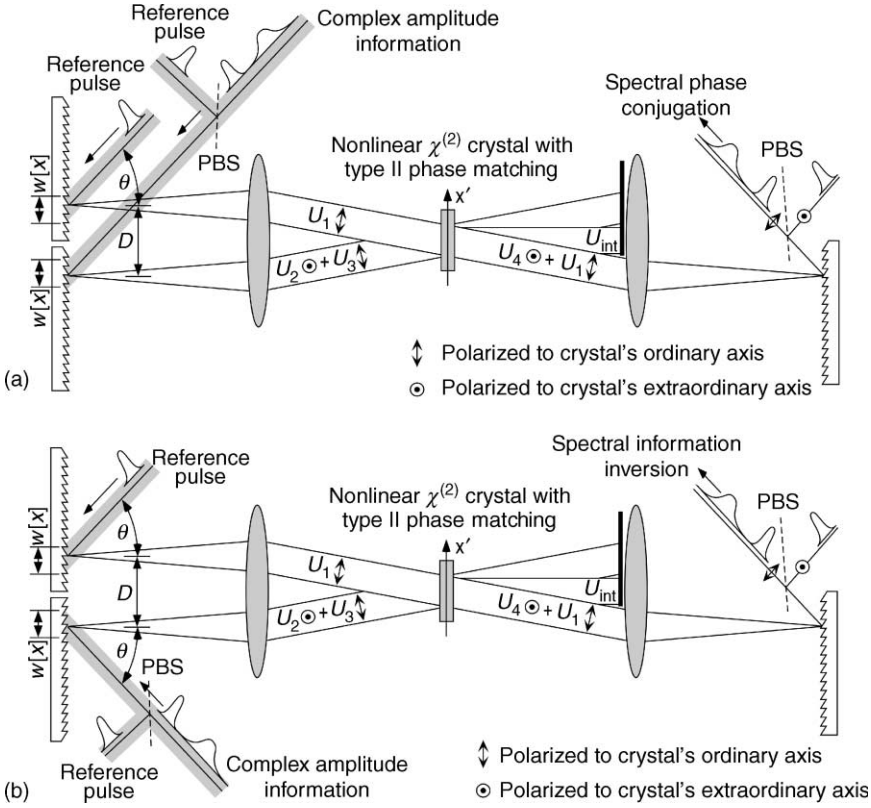


Figure 6 Arrangements of processing ultrafast waveforms within a spectral processing device. (a) Three input waveforms are introduced from the same direction, and their spectrally decomposed waves interact via four-wave mixing. (b) The three input waveforms are introduced from opposite directions, giving rise to mutually inverted spectrally decomposed waves. These two arrangements can perform time reversal, the first by spectral phase conjugation and the second by spectral information inversion. © 2001 IEEE.

where we assume that the window function $w(\cdot)$ is real. Next we apply a SFT to propagate the wave U_4 to the output plane, yielding:

$$\begin{aligned} U_{\text{out}}(x'';t) &= \int U_4(x';t)\exp\left(-\mathrm{i}\frac{2\pi}{\lambda_0 f}x'x''\right)\mathrm{d}x' \\ &= y\left(\frac{\alpha x''}{c}+t-t_0-t_1\right)w\left(\frac{c(t-t_0)}{\alpha L}\right) \\ &\quad\times w\left(\frac{c(t-t_1)}{\alpha L}\right)w\left(\frac{ct}{\alpha L}\right)\exp(\mathrm{i}\omega_0 t_0) \end{aligned} \quad [16]$$

where $y(t)=s_1(t/\tau)\otimes s_2(t/\tau)\otimes s_3^*(-t/\tau)$. The new ultrafast waveform $y(\cdot)$ scans along the output plane at a velocity of $-c/\alpha$ in the x-direction (compare to the input signals scanning in the opposite direction along the input plane, eqn [6]). By placing another diffraction grating at this plane, the new ultrafast waveform can be recomposed and made to propagate once more in free space. The waveform information consists of a convolution between the first and second signals, followed by a convolution with the third signal, which is both conjugated and reversed. If we set the first and second signals to be transform limited pulses, then the resultant waveform will carry the reversed and conjugated information of the third signal. Such capability can be very important for dispersion compensation in optical fiber communication links.

Let us now consider an alternate arrangement, where the second and third input signals are introduced to the spectral processing device from the opposite direction, and consequently have reversed spatial dispersion direction (see **Figure 6b**). Therefore, the first signal will generate a spectral decomposed wave $U_1(x';t)$, as defined by eqn [11a], whereas the second and third signals will give rise to the spectrally decomposed waves $U_2(x';t)$ and $U_3(x';t)$, with inverted spectra relative to that of U_1, as defined by eqn [11b]. The interaction of the first two waves gives rise to:

$$\begin{aligned} U_{\text{int}}(x';t) &\propto P_{\text{NL}}^{\text{sum}}(x';t) = w\left(\frac{c(t-t_0)}{\alpha L}\right) \\ &\quad\times w\left(-\frac{c(t-t_1)}{\alpha L}\right)\exp\left(\mathrm{i}\frac{\omega_0(t_0+t_1)x'}{\alpha f}\right) \\ &\quad\times \tilde{s}_1\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\tilde{s}_2\left(\frac{\omega_0\tau x'}{\alpha f}\right)\exp(\mathrm{i}2\omega_0 t) \end{aligned} \quad [17]$$

which is identical to the wave mixing process of the time-to-space converter (see eqn [12]). However, in this case, we follow the first wave mixing process by a frequency-difference process, generating:

$$U_4(x';t) \propto P_{\mathrm{NL}}^{\mathrm{diff}}(x';t) = w\left(\frac{c(t-t_0)}{\alpha L}\right) w\left(-\frac{c(t-t_1)}{\alpha L}\right) \times w\left(-\frac{ct}{\alpha L}\right) \exp\left(-\mathrm{i}\frac{\omega_0(t-t_0-t_1)x'}{\alpha f}\right) \tilde{s}_1\left(-\frac{\omega_0 \tau x'}{\alpha f}\right) \times \tilde{s}_2\left(\frac{\omega_0 \tau x'}{\alpha f}\right) \tilde{s}_3^*\left(\frac{\omega_0 \tau x'}{\alpha f}\right) \exp(\mathrm{i}\omega_0 t) \quad [18]$$

Next we apply a SFT to propagate the wave U_4 to the output plane, yielding the output waveform:

$$U_{\mathrm{out}}(x'';t) = \int U_4(x';t) \exp\left(-\mathrm{i}\frac{2\pi}{\lambda_0 f} x' x''\right) \mathrm{d}x' = y\left(\frac{\alpha x''}{c} + t - t_0 - t_1\right) w\left(\frac{c(t-t_0)}{\alpha L}\right) \times w\left(-\frac{c(t-t_1)}{\alpha L}\right) w\left(-\frac{ct}{\alpha L}\right) \exp(\mathrm{i}\omega_0 t_0) \quad [19]$$

where $y(t) = s_1(t/\tau) \otimes s_2(-t/\tau) \otimes s_3^*(t/\tau)$. As before, the new ultrafast waveform $y(\cdot)$ scans along the output plane at a velocity of $-c/\alpha$ in the x-direction, and will propagate in free-space upon another grating diffraction. This time, however, the waveform information consists of a convolution between the first signal and a time-reversed second signal, followed by a convolution with the conjugated third signal. If we set the first and second signals to be transform limited pulses, then the resultant waveform will carry the conjugated information of the third signal. If we set the first and third signals to be transform limited pulses, then the resultant waveform will carry the reversed information of the second signal.

We have experimentally demonstrated the two time-reversal techniques: the first technique whereby the resultant signal is both reversed and conjugated is labeled 'spectral phase conjugation', as the signal's spectrum is conjugated by the frequency difference process, and the second technique, where the signal is purely reversed in time, is named 'spectral information inversion', as the spectrum is flipped about the carrier frequency (see Figure 7). By proper selection of the input waveforms, one can also obtain higher-level operations and functionality such as convolutions and correlations among the different signals. All these operations are time-variant, and cannot be performed by passive filtering alone.

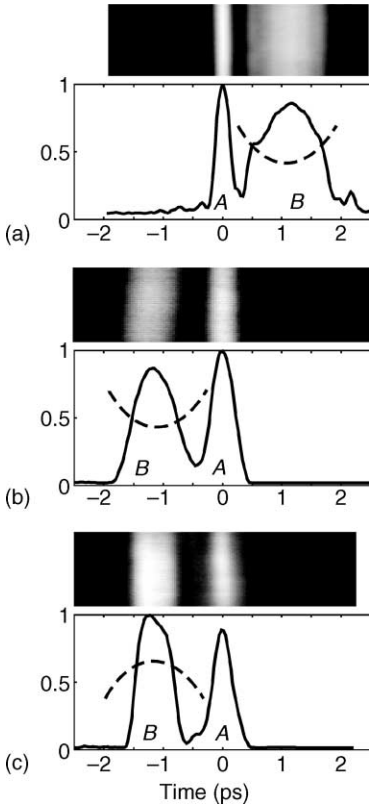


Figure 7 Time reversal experimental results. (a) Input pulse pair consisting of a transform limited pulse followed by a chirped pulse. (b) Time reversal by spectral information inversion technique. Chirped pulse leads transform limited pulse and maintains the sign of the chirp. (c) Time reversal by spectral phase conjugation technique. Chirped pulse leads transform limited pulse and flips the sign of the chirp. Data acquired with time-to-space imaging technique and sign of quadratic term (dashed line) determined by translating output plane.

Synthesis of Ultrafast Waveforms

Shaping of ultrafast waveforms is usually performed by filtering an input transform limited pulse by a fixed filter. Adaptive filtering can be performed by various means incorporating opto-electronic modulators such as liquid-crystal-based or acousto-optic-based spatial light moduloators, however the response time required for changing the waveform is typically limited to the ms and μs regime, respectively. For ultrafast reconfigurability, we utilize our spectral processing device and wave mixing, to achieve pulse shaping controlled by light, or a second optical signal.

Let the three input signals entering the spectral processing device be an ultrashort waveform, $s(t/\tau)$, and two additional spatial domain signals; one carrying spatial information $m(x/\chi)$, where χ is a scaling factor, and the second a point source (see Figure 8). Both spatial domain signals are illuminated

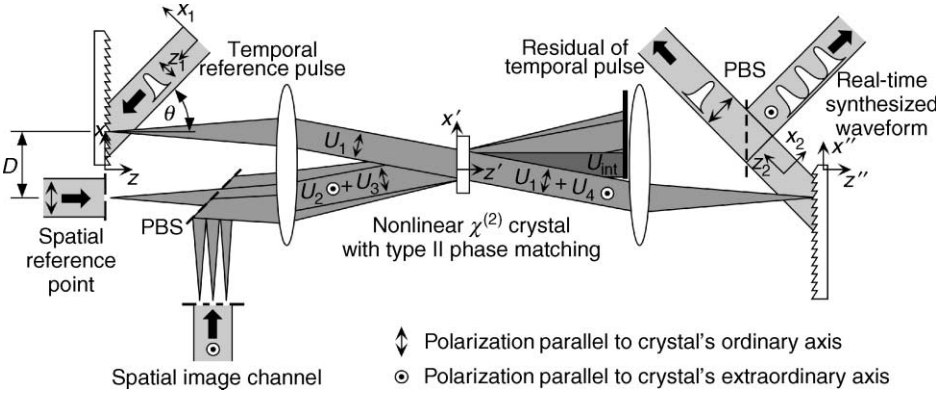


Figure 8 Arrangement for ultrafast waveform synthesis by wave mixing temporal and spatial information channels. Three input channels consist of a temporal channel, a spatial channel carrying information, and a second spatial channel consisting of a point source. The information from the spatial channel is transferred to the temporal channel via the wave mixing process. © 2001 IEEE.

by a monochromatic light source at frequency ω_0. The input ultrashort waveform gives rise to a spectrally decomposed wave $U_1(x';t)$, as described by eqn [11a]. The spatial information channel is SFT by a lens, generating the wave:

$$U_2(x';t) \propto \tilde{m}\left(\frac{\omega_0 \chi x'}{2\pi c f}\right) \exp(i\omega_0 t) \qquad [20]$$

The signal from the point source (or $\delta(x)$) gives rise to a wave $U_3(x';t)$ with no spatial information (plane wave). As before, we utilize cascaded second-order nonlinearities for the wave mixing process. The interaction of the first two waves gives rise to the intermediate frequency-sum wave:

$$U_{\text{int}}(x';t) \propto P_{\text{NL}}^{\text{sum}}(x';t) = w\left(\frac{ct}{\alpha L}\right)\tilde{s}\left(-\frac{\omega_0 \tau x'}{\alpha f}\right)\tilde{m}\left(\frac{\omega_0 \chi x'}{cf}\right)\exp(i2\omega_0 t) \qquad [21]$$

The frequency-difference process with an informationless waveform serves to reduce the carrier frequency of resultant waveform back to ω_0. The down-converted wave U_4 is SFT to the output plane, yielding:

$$U_{\text{out}}(x'';t) = y\left(\frac{\alpha x''}{c} + t\right)w\left(\frac{ct}{\alpha L}\right)\exp(i\omega_0 t_0) \qquad [22]$$

where $y(t) = s(t/\tau) \otimes m(-ct/\alpha\chi)$. As in the processing of ultrafast waveforms section above, the new ultrafast waveform $y(\cdot)$ scans along the output plane at a velocity of $-c/\alpha$ in the x-direction. The synthesized ultrafast waveform can propagate in freespace after an additional grating diffraction. The waveform information, in this case, consists of a convolution between the input temporal signal and a time-scaled and reversed version of the spatial information channel. Since the input temporal channel is typically a transform limited pulse, the output waveform carries the spatial information image, mapped to the time domain. In analogy to the imaging of ultrafast waveforms section, we call this technique space-to-time conversion. We have generated experimentally numerous waveforms that are controlled by a space domain image, or mask (flat-top waveforms, Figure 9, and pulse sequences of Figure 5 were generated by a cylindrical lens array in the spatial information channel).

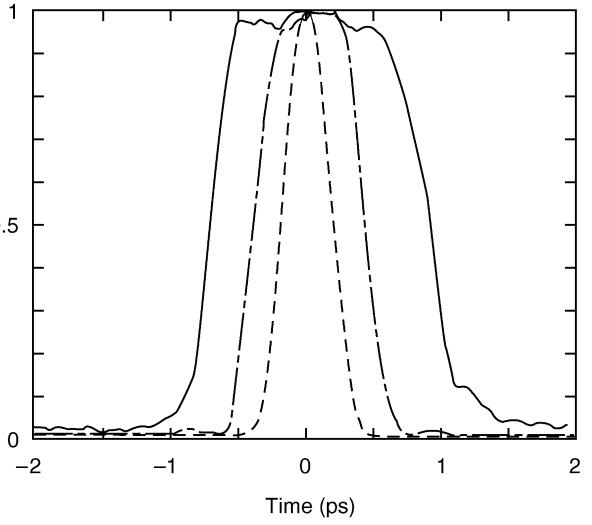


Figure 9 Synthesized flat-top waveforms via space-to-time conversion utilizing a variable slit in the spatial information channel. Increasing the width of the slit results in broader square-shaped pulses in real-time. Solid line: 2 ps pulse, dot-dash line: ~1 ps pulse, dashed line: ~300 fs pulse.

Conclusions

To properly utilize ultrafast waveforms in scientific and advanced technology applications requires instantaneous techniques from generation to processing and imaging of the waveforms. We have studied the nonlinear wave-mixing approach within a spectral processing device, and demonstrated the ability to perform all of these tasks. The wave mixing was performed in a second-order nonlinear crystal, using three-wave mixing for detection and four-wave mixing for synthesis and processing of ultrafast waveforms. The three-wave mixing process results in a second-harmonic output image, suitable for acquisition by silicon-based CCD imagers. The four-wave mixing processes result in an output waveform of an identical carrier frequency, as typically required in processing and synthesis applications. The high nonlinear coefficient of $\chi^{(2)}$ crystals enabled us to demonstrate these techniques with ultrashort pulse energies ranging from nJ to mJ levels. Further advantages afforded by the techniques are expanded time windows and favorable phase matching.

This review of the instantaneous optical signal processing capabilities with optical nonlinearities and interaction with ultrafast waveforms introduced and demonstrated the basic signal processing capabilities to manipulate optical signals in amplitude and phase. The same techniques can be used and expanded upon to achieve different signal processing abilities, which are beyond the scope of this article.

See also

Coherent Transients: Ultrafast Studies of Semiconductors. **Nonlinear Optics, Basics:** Ultrafast and Intense-Field Nonlinear Optics. **Ultrafast Laser Techniques:** Pulse Characterization Techniques. **Ultrafast Technology:** Femtosecond Chemical Dynamics: Gas-Phase; Femtosecond Condensed Phase Spectroscopy: Structural Dynamics; Ultrafast Illumination and Processing.

Further Reading

Diels J-C and Rudolph W (1996) *Ultrashort Laser Pulse Phenomena: Fundamentals, Techniques, and Applications on a Femtosecond Time Scale*. San Diego: Academic Press.

Marom DM, Panasenko D, Sun P-C and Fainman Y (2000) Femtosecond-rate space-to-time conversion. *Journal of the Optical Society of America B* 17: 1759–1773.

Marom DM, Panasenko D, Sun P-C, Mazurenko YY and Fainman Y (2001) Real-time spatial-temporal signal processing with optical nonlinearities. *IEEE Journal of Selected Topics in Quantum Electronics* 7: 683–693.

Miller DR, Murnane MM, Scherer NF, Weiner AM (eds) (2003) *Ultrafast Phenomena XIII, Proceeding of the 13th International Conference*. Berlin: Springer-Verlag.

Weiner AM (1995) Femtosecond optical pulse shaping and processing. *Progress in Quantum Electronics* 19: 161–238.

Weiner AM and Kanan M (1998) Femtosecond pulse shaping for synthesis, processing, and time-to-space conversion of ultrafast optical waveforms. *IEEE Journal of Selected Topics in Quantum Electronics* 4: 317–331.

EARLY HISTORY OF QUANTUM ELECTRONICS

C H Townes, University of California, Berkeley, CA, USA

© 2005, C H Townes. Published by Elsevier Ltd. All rights reserved.

All the ideas essential to making a laser were known before 1930, but there was no operating laser before 1960. So why didn't the laser come sooner? There are several reasons. One important impediment to the laser invention was that a combination of ideas from quantum mechanics and from electrical engineering was needed, and these two fields were not well mixed in the early days. Another is that, while some physicists recognized that amplification could occur if there was a population inversion of states, they did not consider the coherence of such amplifications and did not recognize its importance or usefulness. It was just something that in principle could happen, but was not very interesting. What was required was a combination of physics and electrical engineering thinking and recognition of the coherence of such amplification. In addition the importance of such a development had to be visualized and recognized in a way that it led to very devoted work towards its achievement.

It is striking that lasers grew out of the study of microwave spectra of molecules, for which engineering and quantum mechanics were both important and which helped orient thinking in the appropriate direction. That this origin was not accidental is convincingly demonstrated by the fact that three independent ideas for amplification by stimulated emission were generated about 1950. They were from Joe Weber, at the University of Maryland, from Nikolai Basov and Alexander Prokhovov at the Soviet Academy, and myself at Columbia University. Weber primarily wanted to point out the possibility, but didn't try to do it. In addition, his numbers were a bit off and no practical system was suggested. Basov and Prokhovov actually worked towards microwave amplification using a beam of molecules, as did I.

That stimulated amplification was recognized early, but not thought through, is illustrated by the fact that Prof. Richard Tolman, a theoretical physicist, wrote a discussion in 1924 of the net absorption of light by molecules, pointing out in particular that induced emission counteracted absorption and noting that if there were more molecules in the upper than in the lower state there could be 'negative absorption'. But, he wrote, 'This would usually be very small'. The Russian physicist Vitaly Ginsburg wrote me, after the maser and laser had appeared, that his professor, S.M. Levi, had been well aware of such effects back in the 1930s and had told him 'create an overpopulation at higher atomic levels and you will obtain an amplifier; the whole trouble is that it is difficult to create a substantial overpopulation of levels'.

The German physicist F.G. Houtermans said to me that in 1932, when told by a colleague of an unusual light intensity in a gaseous discharge, he had thought it might be a 'photon avalanche', i.e. multiplication of photons by stimulated emission.

Another Russian physicist, V.A. Fabricant, wrote a thesis in 1939 in which he discussed absorption and emission of light radiation in a gas and looked for 'negative absorption', or amplification. He did not discuss coherence or a resonant cavity, and was not able to achieve any amplification so his work was quickly forgotten. None of these early mentions of stimulated emission proposed how to actually get amplification, that it would be useful, nor clearly noted its coherence. Tolman did, however, write in 1927 that, 'We should expect radiation induced by an external field to be coherent with the radiation associated with that field'. I know of no proposal to actually make use of such amplification, until those made by microwave spectroscopists in the early 1950s.

Another clear indicator that even in the 1950s physicists and engineers did not think amplification by stimulated emission was particularly interesting nor useful is that during the $2\frac{1}{2}$ years when Jim Gordon, Herb Zeiger, and I were working on trying to obtain microwave amplification (the maser), a large number of scientists visited my laboratory, saw what we were doing, but no one bothered to also try to obtain such amplification. After the maser worked, it hit the newspapers and then very quickly became a popular and intense field of interest for a number of physicists.

My own drive to produce oscillators by stimulated emission came from my strong interest in obtaining sources of waves shorter than the few millimeters wavelength which could be produced by electronic devices, in order to extend the high resolution study of molecular spectra to wavelengths shorter than microwaves, down into the far infrared. My students and I worked on several possible schemes for producing waves shorter than those produced by klystrons or magnetrons – frequency multiplication by nonlinearities, electronic beams passing over surfaces of solid materials with resonances, and anything else I could think of. None worked very well.

In early 1950, I was asked by the Office of Naval Research to form and chair a committee which would examine possible research towards obtaining wavelengths down to one millimeter and shorter. I chose outstanding scientists and engineers in a variety of fields which might touch on this problem. We met together, and visited many pertinent laboratories and individuals interested in such research. Nothing very promising seemed to turn up. But we wanted, of course, to at least provide a report summarizing the situation as we saw it. Our last meeting was April 26, 1951, in Washington, D.C. Worrying over our lack of success, I woke up early before the meeting. Breakfast was not ready, so I walked over to nearby Franklin Park, sat on a bench in front of beautiful blooming azaleas, and puzzled over why neither I nor the Committee had found any promising solutions.

I went over all the ideas I had had previously. Molecules, of course, can produce high frequencies. But I had previously concluded, I thought wisely and rigorously, that one could not obtain intense radiation from them because radiation intensity was a function of temperature, and the temperature could not be very high without destroying the molecules. Suddenly I realized that they did not need to have a temperature in the usual sense – they need not be in temperature equilibrium. There could be more molecules in an upper than a lower excitation state, which could in principle produce indefinitely intense radiation. I pulled a piece of paper out of my pocket and wrote down the equations and numbers for such a case, using a molecule beam sent into a resonant cavity and ammonia molecules with which I was very familiar. My equations said one could get enough excited molecules, low enough loss in a resonant cavity, and it would work! Why hadn't I thought of it before?!

Back at Columbia, where I worked, and about 4 months later, the graduate student Jim Gordon agreed to work on trying to obtain such an oscillator using a beam of ammonia. I assured him that if it didn't work he could modify the experiment to do interesting spectroscopy and thus complete a thesis. But we both thought he had a good chance of making it work. And a young post doc working with me, Herb Zeiger, joined the effort.

Actually, I had first thought of obtaining stimulated emission from a molecular beam back in 1948, but simply as a demonstration of physical principles rather than as a useful amplification. Also, about a year later a young post doc, J.W. Trischka, working on molecular beams with Professors Rabi and Kusch, had also thought of demonstrating stimulated emission. We talked about it together, and he decided it wasn't worthwhile just demonstrating the effect because it wouldn't really prove any new physics. Neither of us at that time had recognized the real point and the possibility of useful amplification, which is one of the reasons mentioned above that the idea was delayed as long as it was.

It is perhaps important to emphasize again and to illustrate how out-of-the-way the use of stimulated emission was at that time for physicists, and how distant stimulated emission was from engineers. While we were working on the ammonia beam maser, Prof. L.H. Thomas, an outstanding physicist known for the 'Thomas Effect', frequently would run into me in the hallway at Columbia University and say that I didn't understand, and that my proposed ammonia oscillator could not work. However, I never got a clear explanation from him of what it was I didn't understand. And after we had been working on the ammonia maser system for about 2 years (a more-or-less normal time for a student thesis project), the Physics Department Chairman, Prof. Polycarp Kusch, and the previous chairman, Prof. I.I. Rabi, came into my office to object. They were excellent physicists, both received Nobel Prizes, and both were experts on molecular beams. They sat down in my office and said 'Look, Charlie, that is not going to work. We know it won't work and you know it won't work. You are wasting departmental money, and must stop!' Other people had also questioned what I was doing, frequently in particular whether the stimulated radiation would be coherent. So I had already thought over the situation many times. I had full notes from the quantum mechanics course I took as a student back in 1938, and could derive from them a proof of coherence. I felt also that I knew the quantitative numbers, such as the molecular beam intensity and the possible 'Q', or loss, in the cavity resonator, very well. I was also by then an Associate Professor, and the department chairman could not fire me simply because of stupidity. I replied to Rabi and Kusch 'No, I believe it has a reasonable chance of working and I'm going to continue!' Annoyed, they stomped out of the room.

About 2 months after the Rabi/Kusch incident, Jim Gordon dashed into the classroom where I was lecturing and said loudly 'It's working!' Most of the class then went up to the lab to see the new device. Rabi and Kusch were not against me; even though they were outstanding physicists, they probably just didn't quite comprehend the device. A couple of months after its successful operation, Kusch more or less apologized by saying 'Well, I should have realized that you probably know more about what you are doing than I do.'

After the maser successfully operated, there were other incidences showing the lack of appropriate

focus of physicists on stimulated emission. I was a friend of Aage Bohr, the son of Niels Bohr, and in that connection was visiting him in Denmark. Niels Bohr asked me what research I was presently doing, so I told him about our new oscillator, the maser, and its remarkably pure frequency. He looked at me and said 'Oh no, that's not possible. You must be misunderstanding something.' I emphasized again what it was really doing, but he still seemed not to believe it could function that way. I presumed he was thinking in terms of the uncertainty principle and the finite time of passage of a molecule through the cavity, though I never was quite sure just why he felt it impossible. A similar thing happened shortly after that at a cocktail party in Princeton, where John Von Neumann asked the same question – what was my research at the moment? After telling him about the maser oscillator and the purity of frequency he reacted very similarly 'Oh no,' he said, 'That can't be right. You must be misunderstanding something.' After arguing a little more, he left to get another drink. Fifteen minutes later he came back, saying 'Hey, you are right!' He had understood, and wanted to talk much more about the maser and of possibly using excited semiconductors. Only after his death did I learn from his notebooks that he had considered exciting electrons in semiconductors with neutrons from a reactor, and had written Edward Teller about whether some experiments might be done with this to obtain intense light. However, he had not considered coherence, and Teller was apparently not interested enough to respond, so the matter was dropped.

The above account reminds me a bit of the amusing comments of Arthur Clark on 'Change'. He writes:

'People go through four states before any revolutionary development:

1. It's nonsense, don't waste my time
2. It's interesting, but not important
3. I always said it was a good idea
4. I thought of it first'.

It's clear that the world of physics was thinking very little in the direction of masers or lasers, and that many preconceptions as well as lack of interest stood in the way. My own experience with engineering at Bell Telephone Labs during World War II, in designing radar and electronic systems, plus my intense interest in obtaining short-wave oscillators, were clearly important in bringing me to the right ideas.

As masers became very interesting to the physics community and the field grew rapidly, my engineering experiences continued to be of help. I was well acquainted with the theoretical examination of noise in vacuum tube amplifiers by various individuals at Bell Labs, and recognized that the maser could provide much more sensitive amplification than could common electronic amplifiers, where discrete electronic charges produce the basic noise. On sabbatical in France, I worked on electron spin masers with Jean Combrisson and Arnold Honig who had the appropriate equipment. And then in Japan, Koichi Shimoda, Hidetoshi Takahashi, and I wrote a theoretical paper on the basic quantum noise of maser (or laser) amplification. The theory showed that maser amplifiers of microwaves should be about 100 times more sensitive than the existing electronic ones.

After 2–3 years of maser experiments and development I felt I wanted to move on to the shorter wavelengths for which I had generated the maser idea. Although I had first tried the idea at microwave frequencies because that seemed the easiest way to test out the general idea and the result had been exciting, I still wanted those shorter wavelengths. I had not come up with any great ideas of just how to get to much shorter wavelengths, which is why I waited several years after the maser worked before moving on. However, in 1957, 3 years after the first successful operation of the maser, I decided it was high time to simply figure out what was the best way I could imagine to move on into the infrared region and do it. A number of physicists had concluded that of course one couldn't make masers work at much shorter wavelengths, certainly not in the visible region, because spontaneous emission becomes so much faster as the wavelength is shortened and adequate inversion of population was not practical. But that was intuition, not quantitative science. As I wrote down equations for what might be done to move towards shorter wavelengths using a model of atomic or molecular excitation by radiation and a reasonably high Q resonator, it became clear that it was quite practical to move on even into the visible region. That was exciting! Why hadn't I, or someone else, looked at it carefully and quantitatively before that time?

I was at the time consulting at Bell Labs, with the assignment to spend a day there every 2 weeks and just talk with the Lab's scientific personnel. Since my former post doc and now brother-in-law Arthur Schawlow was then at Bell Labs, I of course talked with him. On telling him of my ideas for an 'optical maser', using a resonant cavity and excitation of atoms with optical radiation, he said he too was interested in that, and we decided to work together to optimize a system. It was Art who then suggested use of a Fabry–Perot resonator rather than the cavity with large holes I had used for a model, and that was

an excellent addition. Why I did not think of that is a mystery, but Art had done his thesis at the University of Toronto in Fabry–Perot spectroscopy, and that might have been why the thought came to him.

Since Art Schawlow was participating, I decided the patent for the new 'optical maser' should belong to the Bell Labs (I already claimed ownership of the basic maser patent, which covered all wavelengths). Hence we recognized that the new idea must be kept confidential until Bell Labs lawyers had worked out and applied for an appropriate patent. This delay in public information sheds some additional light on how the scientific and technical world was thinking at the time. I had written down my original ideas for an 'optical maser' in my notebook and had it witnessed by my student Joe Giordmaine at Columbia University. I had also talked with a Columbia student Gordon Gould because he had been doing research with an intense light source and I wanted to know how much intensity he had in order to be sure I could get enough excited atoms and provide an oscillator at these short wavelengths. Except for these two persons, neither Art Schawlow nor I told anyone outside of Bell Labs about the 'optical maser', or laser idea until after Bell Labs had properly prepared its patent, which was about 11 months after my first notebook entry. For that entire time, no one has produced a record of any thoughts about extending the maser to optical wavelengths except Gordon Gould, who entered some ideas in his notebook about 1 month after I talked with him about the possibility of an optical maser, and 2 months after my original notebook entry. His notes were later to be the source of a long patent case.

The striking observation is that no one outside of Bell Labs except Gordon Gould, to whom I had explained my ideas, seems to have written or noted down anything about extending masers to these shorter wavelengths during the 11 months from my recognition that it could well be done until after the Schawlow and Townes paper on how to do it became available. After that there was considerable excitement and a number of ideas.

It is also noteworthy that, because of low general interest and competition, I did not publish a paper on how a maser might be made, but waited for publication until we demonstrated its operation, about 3 years after the idea had arrived. But after the maser worked the field became exciting and competitive. Hence, Schawlow and I thought we surely should publish a theoretical paper establishing the idea before taking the time to make a laser work. And indeed, there was much competition to build the first laser.

I myself helped a couple of my graduate students start work on trying to build an 'optical maser' or laser. But at about that time (1958), I was urged to undertake a job in Washington as Vice President and Director of Research of the Institute for Defense Analysis, an organization put together largely by the presidents of several universities to try to help advise the government on matters of science and technology. Sputnik had gone up the year before, and everyone was worried about the position of the U.S. with respect to Communist Russia, which seemed to be ahead particularly in some areas important to defense. I decided I should try to help, and accepted a two-year appointment in Washington. I recognized that this seriously distracted my attention from developing a laser quickly, but knew there were many others working towards lasers and so such devices would certainly be developed and our doing it just at Columbia University was neither critically important nor highly probable.

The field of masers and 'optical masers' or lasers was becoming so exciting and a bit scrambled that The Office of Naval Research asked me if I would organize a meeting on the subject. I did, with the help of a committee of many distinguished people in the field or closely related science and technology. And it was in a meeting of the Committee that we christened the field with the name 'Quantum Electronics'. This first international meeting on the subject was at Shawanga Lodge in New York State in September 1959. It made an occasion for the Russians Basov and Prokhorov to visit the United States (and my lab and home), and was about 6 months before Ted Maiman made the first working laser. There were many interesting discussions of masers and their coming operation at optical wavelengths.

It is significant to note that, while the maser grew out of basic research in universities (with Russian work at the Russian Academy) and industry had little to do with early masers, all the first lasers were made in industrial laboratories. This illustrates the sociology and the strengths and weaknesses of industrial and of academic laboratories. While masers (and from them lasers) originated from research on microwave spectroscopy of molecules, industrial laboratories believed the field of microwave spectroscopy had little to offer in the way of commercial results. Because of equipment available and the interest of physicists in industry, the field was initiated and pursued immediately after World War II in commercial laboratories – by myself at Bell Labs, my friends at the RCA Labs, at Westinghouse, and at General Electric. For lack of interest in industry, such work was soon shut down,

except at Bell Labs, and it moved to universities. Bell Labs generously allowed me to continue such work, although they wanted me to do some 'more useful' engineering. Once the field had obvious commercial possibilities, commercial laboratories began to support it well and the clear importance of masers and lasers made industry very interested. Really interested industry can more easily put strong new resources and push harder on a field than can academic laboratories, where money has to be granted and professors have a variety of other assignments. The first laser was made to work, of course, by Ted Maiman at Hughes Research Laboratories. Ted had been a student of Willis Lamb at Stanford, and there worked on radio spectroscopy. The second type of laser, rather similar to Maiman's but using a different material, was made to work at the General Electric Laboratories by Peter Sorokin and Mirek Stevenson. Sorokin had been a student of Bloembergen at Harvard in microwave or radio spectroscopy and Stevenson was one of my students in microwave spectroscopy at Columbia University. The next type of laser, and one I particularly admire, was made by Ali Javan, William Bennett, and Don Herriott. Javan had been a student with me in microwave spectroscopy at Columbia University, Bill Bennett a student in the molecular beam group at Columbia working on radio spectroscopy, and Don Harriott was an optics specialist. All of these originators, with the exception of Herriott, had been working at universities in the field which originated the idea and were recently hired by industry. The next important laser, involving semiconductors, was invented by Robert Hall and collaborators at General Electric Labs. Note that every one of these early lasers was created in industry.

After the first laser was operated, my own students at Columbia quickly turned from trying to make a laser to using lasers to explore more physics, the normal university function. And I am delighted that masers and lasers have provided such excellent tools for research, as well as for commercial and medical applications.

After a few years of exploring new physics with lasers, I decided that since there were many excellent scientists doing such work, I should move into fields which it seemed to me were being relatively neglected. I moved to the University of California at Berkeley to look for molecules in interstellar space by microwave spectroscopy, and to do infrared astronomy. Very soon after initiating work at Berkeley, one of my students, Albert Cheung, not only discovered the first polyatomic molecules in space, he found powerful water masers. A while before our discovery of water and identification of its radiation as due to maser action, it had been deduced that some microwave radiation of OH must be due to maser action. And since then, many, many masers due to a wide variety of molecules have been found in astronomical sources as well as a few powerful lasers. Since deviations from thermal equilibrium are common in the very thin gases excited by powerful sources in space, we should have expected this, but didn't. And these masers in space could have been easily detected with radio technology available back in the 1930s if anyone had searched the microwave spectrum carefully.

Clearly, masers and lasers could have been discovered and used much sooner than they actually were. We simply were neither thinking nor looking in the right directions. And this raises the natural question – what important and more-or-less obvious ideas are we missing now because of our lack of imaginative exploration?

Further Reading

Basov N and Prokhorov A (1954) Application of molecular beams to radio spectroscopic studies of rotation spectra of molecules. *J. Experimental. Theoretical Physics U.S.S.R.* 27: 431–438.

Bloembergen N (1956) Proposal for a new type of solid state maser. *Physical Review* 104: 324–327.

Cheung A, Rank D, Townes C, Thornton D and Welch J (1969) Detection of water in interstellar regions by its microwave radiation. *Nature* 121: 626–628.

Combrisson J, Honig A and Townes C (1956) Utilization de la resonance de spins electroniques pour realiser un oscillateur ou un amplificateur en hyperfrequencies. *Comptes Rendues* 242: 2451–2453.

Gordon J, Zeiger H and Townes C (1954) Molecular microwave oscillator and new hyperfine spectrum of NH_3. *The Physical Review* 95: 282–284.

Hall R, Fenner G, Kingsley G, Soltys J and Carlson R (1962) Coherent light emission from GaAs junction. *Physical Review Letters* 9: 366–368.

Javan A, Bennett W and Herriot D (1961) Population inversion and continuous optical maser oscillation in a gas discharge containing a He–Ne mixture. *Physical Review Letters* 6: 106–110.

Knowles S, Mayer C, Cheung A, Rank D and Townes C (1969) Spectra, variability, size, and polarization of H_2O microwave emission sources in the galaxy. *Science* 163: 1055–1057.

Maiman T (1960) Stimulated optical radiation in ruby. *Nature* 187: 493–494.

Perkins F, Gold T and Salpeter E (1966) Maser action in interstellar OH. *Ap. J.* 145: 361–366.

Schawlow A and Townes C (1958) Infrared and optical masers. *Physical Review* 112: 1940–1949.

Shimoda K, Takahashi H and Townes C (1957) Fluctuations in amplification of quanta with

application to maser amplifiers. *Journal of the Physical Society of Japan* 12: 686–700.

Sorokin P and Stevenson M (1960) Stimulated infrared emission from trivalent uranium. *Physical Review Letters* 5: 557–559.

Tolman R (1924) Weak quantization. *The Physical Review* 24: 287–295.

Tolman R (1927) Statistical mechanics with applications to physics and chemistry. The Chemical Catalog Co., New York, pp. 169.

Weber J (1953) Amplification of microwave radiation by substances not in thermal equilibrium. *Trans. Inst. Radio Engineers Professional Group on Electron Devices* 3: 1–4.

LET THERE BE COHERENT LIGHT: THE DISCOVERY OF THE LASER

T H Maiman, Vancouver, BC, Canada

© 2005, T H Maiman. Published by Elsevier Ltd. All rights reserved.

Editor's Note

The discovery of the first laser was a remarkable culmination of human creativity and ingenuity. However the achievement was finally made by one individual young scientist, Theodore Maiman, working independently of the mainstream and without the benefit of strong government or private financial support. With persuasive arguments, he convinced his management at the Hughes Research Laboratories to allow him to continue his work, albeit with limited resources. And then, working alone with his assistant he saw the light that now illuminates the world of laser science and engineering. This is the story of how thinking outside the box created a new field of science, whose developing technology has advanced the state of medicine and engineering and the quality of life in general.

Introduction

I am frequently asked how it was that I discovered the laser. Many assume that the concept evolved from some sudden, inspirational thought. It didn't happen that way.

It is dramatic and exciting to have a scientific discovery emerge from a dream or a vision that comes out of nowhere, but it seldom does. In reality, almost all scientific discoveries come from building on other, prior, scientific developments.

So it was with the laser. In 1916, Albert Einstein laid the foundation and conceived the basic underlying principles on which lasers are based. He formulated and explained the relations that govern the way that atoms and molecules absorb and emit radiation. He introduced the key concept of stimulated emission. Then, in the 1920s, physicists C.H. Füchtbauer and Rudolph Ladenburg added formulations that used the Einstein theory to go further, and tie the absorption of light in a material medium to the fundamental properties of its constituent atoms. But it was the Russian physicist A.V. Fabricant who first had the vision to propose the concept of a laser in 1940.

Unlike the Sun and other incoherent light sources, the laser produces light of one precise color. Its waves are emitted in phase with each other, and its radiation can be efficiently focused to a precise spot. In short, the laser produces coherent light, with the properties we were already used to with radio, television and microwave sources.

Fabricant specified the conditions needed for amplification of light via stimulated emission. He appreciated the concept of an inverted population and the concept of coupling such an inverted population medium to a resonant structure (resonator). He proposed using a gaseous electrical discharge to achieve laser action as one possibility. Later, he proposed the use of a helium discharge lamp to optically pump the gaseous form of cesium.

Scientists Purcell and Pound were the first to report evidence of net stimulated emission in 1951. Their observation, in the radio part of the spectrum, was a by-product of their pioneering developments in the techniques of nuclear induction, the foundation for magnetic resonance imaging.

The first achievement of a working device that utilized stimulated emission as its operating principle was a microwave oscillator/amplifier. That device used a beam of ammonia molecules as the working medium. The design details were worked out independently by groups at the Lebedev Institute in the Soviet Union and Columbia University in the USA. The device was dubbed a *maser*, which is an acronym for *microwave amplification by stimulated emission of radiation*.

In 1955, the year after I was awarded my Ph.D., I joined the Hughes Research Laboratory in Culver City, California. One of my major assignments was to head a project developing a miniature, liquid-nitrogen ruby maser for the US Army Signal Corps.

Race to the Light

Before starting work on the maser I had begun to germinate some ideas about the possibility of a laser. My concept so far was to try to use a solid material for the lasing medium, and to fabricate the potential laser material into a rod shape. In microwave thinking this configuration would be called a *dielectric wave-guide*. In optics it is known as a *light pipe*.

Abridged from "The Laser Odyssey" by Theodore H. Maiman. Copyright © 2000 Theodore H. Maiman. Reproduced with the permission of the author and Laser Press, Blaine, WA, USA.

I planned to put mirrors on each end of the rod to form a resonator.

Despite the advances in physics it represents, the maser does not in any way extend the coherent electromagnetic spectrum. Furthermore, its use as an amplifier turned out to be impractical because cryogenic temperatures were required for proper operation. There was strong motivation to push the frontiers of coherent electromagnetic radiation to higher frequencies, principally the promise of higher energies and greater concentrations. The reduction in wavelength from microwave to visible is a factor of about 10^4. This would lead to a consequent 10^4-fold reduction in the minimum achievable focal size, and a 10^8-fold increase in energy concentration delivered by the laser.

Work on the maser project therefore proved something of a distraction, and it was not until August 1959 that I was able to devote full effort to the consideration and analysis of my concepts for a laser.

Why would I be willing to enter such a race? There was rather formidable global competition already in play, well-funded and very competently manned. The answer lies in my knowledge of the proposals that had been floated about. Generally, the scientists who pursued these notions were not particularly secretive about what they were working on. They published and gave talks at conferences. On the whole, however, I found that the authors and conference presenters were only offering vague proposals. To be sure, this served as an exchange of information and stimulated ideas. But it didn't appear to me that anyone was close to the answer. In short, no viable laser concept was yet in existence.

As I look back, I was a little brash. I would be thrusting myself, in a sense, into a technological Olympics. The competition was of the best quality and of international scope. But my competitive spirit won out. The challenge of working in the top league of such an exciting project, that had so many questions and problems to resolve, was very compelling to me.

We know *now* that many kinds of lasers can be made. But back then, in 1959, we didn't know. We didn't even know with any confidence if it was really possible to make a laser at all. My strategy was to limit myself to potential solutions to the making of a laser that did not have appreciable distractions in the design. That way I could focus strictly on just the laser problem itself.

For various reasons I was reluctant to work on alkali-vapor or gas discharge systems. I chose to work with solid-state crystals. The main appeal that solids held for me was simplicity. By that, I mean simple in analysis and understanding, and simple in device conception.

In contrast to a gas discharge, the energy level diagram in an appropriate crystal is very limited. There are relatively few possibilities for the energy states, and by and large, the pertinent parameters for a potential laser candidate are amenable to a combination of calculations and relatively straightforward direct measurements.

Another advantage, in principle, to a solid crystal is its relatively high gain coefficient. By that, I mean the amplification in a given length of material is of reasonable proportions. This meant that the laser medium could be relatively small in size and short in extent, and I would not have the problem of developing or depending on the use of special mirrors. Indeed, my first laser used a crystal that was only 2 cm long.

I was also intrigued with the concept of a solid medium since I would not have to deal with vacuum pumps, impurity problems and gas handling apparatus, or complex mirror mechanisms. I could put simple silver mirror coatings directly on the crystal as I had done with my small ruby maser. In principle, a solid crystal laser could be designed to be very simple, compact and rugged.

My first choice to study and contemplate was that of a ruby crystal. Ruby is the result when a water-clear cast of aluminum oxide is doped with a small amount of chromium oxide impurity. It is the chromium that is responsible for the red color. The rubies used for devices are usually not natural gemstones, which have a chromium impurity level of about 0.5%. The chromium concentration in device rubies is around ten times less than this, and hence they display a lighter red color than gemstone ruby and are referred to as *pink ruby*.

Why choose ruby as a potential laser candidate? Most importantly, I was quite familiar with and fascinated by the interesting optical properties of the crystal. Ruby is a fluorescent mineral; if ultraviolet light is shone on a ruby, it will glow with deep red fluorescence. Furthermore, ruby also gives off a red glow when either blue or green light is shone upon it. It is these blue and green absorption bands that give ruby its red color. When a green photon impinges on and is absorbed by the ruby, a chromium impurity ion is raised from its ground state into a broad, excited band. Although the chromium ion has the possibility of radiating by spontaneous emission from that excited level, another process comes into play. The competing process uses the thermal vibrations of the crystal lattice to interact with the excited ion and deposit most of the excitation energy to another slightly lower excited level of the ruby chromium

where it stays for a while. This latter process is much more probable and dominates.

The energy from this third level is radiated as spontaneous incoherent emission. This spontaneous emission is made up of red photons and is the observed fluorescence. The level where the fluorescence emanates is sometimes called a metastable level, since the chromium ions linger in that energy state for a comparatively long time before they radiate red photons.

I developed a model that could be mathematically analyzed and I set up kinetic equations to take into account the various mechanisms taking place in the fluorescent process. I also set up simple intuitive criteria for establishing the condition for laser action. This model and these equations have subsequently become a standard way for others to analyze crystal lasers.

I was able to determine which material parameters were important and relevant to a laser by getting the solution to the equations describing the model. Using known and estimated values for the pertinent parameters in ruby, I found that ruby would require a very bright pump lamp to excite the crystal sufficiently to allow laser action to take place. The brightness of a lamp is an important parameter for lasers. It is a measure not of the total power radiated by the lamp, but rather the power per unit area.

Obstacles and Solutions

Ruby has many desirable qualities. It is a very stable, very rugged crystal to work with. It has some broad pumping bands that make it helpful to get a reasonable energy transfer from an incoherent pumping lamp, and compared to some other crystals I considered (e.g., gadolinium-doped crystals), a working ruby laser would produce visible light.

However, there was some uncertainty over the crystal's measured quantum efficiency. The quantum efficiency for a fluorescent crystal is the number of fluorescent photons emitted, divided by the number of pumping photons. In the case of ruby, this is the number of red photons radiated out compared to the number of green photons absorbed. A report in November 1959 had suggested that ruby's quantum efficiency was only 1%. This would effectively mean that it would be impossible for ruby to achieve laser action. However, I was able to make my own measurements of ruby's fluorescent efficiency, and these showed that it was nearer 75%.

I now had enough information that I could proceed to work out an actual laser design. But how would I pump the ruby? What would be the ruby's shape and size?

I knew that I needed a very bright lamp. Among the brightest around was a high-pressure mercury arc lamp manufactured by General Electric. In addition to being one of the brightest available laboratory lamps, it had the advantage of radiating most of its energy in the green and blue-violet parts of the spectrum. This was a good match to what I needed for the ruby.

I drew up a paper design with the mercury arc lamp at the focus of an elliptic cylinder. A small ruby rod was to be mounted on the other focus of the elliptic cylinder. It is a property of an elliptically shaped reflector that a point of light placed at one focus of the ellipse will be imaged at the other focus. The elliptic cylinder that I had in mind would be highly polished on its inner surface to a mirror finish.

But I had a problem. My paper analysis of the design showed that although it should work, it would only do so by a slim margin. I studied the design in more detail, looking for ways to optimize and improve it, but I couldn't convince myself that it was anything but marginal.

I stewed over the problem at hand. I started to think about other ways to look at the problem. I was frustrated: I felt that, on the one hand I was very close to an answer, but on the other hand it was eluding me.

I went back to my analytical model. I pondered my options and decided to put the pumping lamp requirement for the ruby in a different form. I calculated the equivalent black-body temperature of a suitable pumping lamp. The temperature I calculated to have enough brightness capable of driving a ruby into laser action was close to 5000 K. By contrast an ordinary tungsten light bulb has a brightness temperature of about 2800 K.

Once I put the pumping requirement in terms of brightness temperature, I began to think in a different way. I remembered reading an article about photographic strobe lamps, a camera's flash mechanism. The article said that strobe lamps could reach brightness temperatures of 8000 K or more.

I now had my 'aha'!

Most scientists had been thinking in terms of a continuous laser, and that was certainly my thinking to start with. But why should I place such restrictions on myself? At this juncture I was only trying to demonstrate that coherent light could be made at all. Besides, a pulsed source is not only acceptable in many applications, it may even be preferable.

I went back to my analytical model, modified the equations to account for a pulsed light source, and then analyzed the results to guide me in the actual laser design. I had already determined that the most important parameter of the light source would be its

brightness, that is, a lamp that maximized the power per unit area into the useful pump bands of the ruby.

My calculations also instructed me that this lamp brightness requirement is largely independent of the ruby crystal dimensions and its chromium concentration over a reasonably wide range. The model assumption was that of a three-level system (appropriate to ruby) where the lower laser-energy level is the ground state.

Next, I scoured every flashlamp catalogue that I could find. I calculated the luminous intensity for promising candidates and found that the highest intensity came from three General Electric spiral-shaped flashlamps. My calculations showed that I would have an adequate margin of safety, in terms of excess brightness intensity for the laser, of some two to three times, even with the smallest lamp.

Let there be 'Coherent' Light

Based on my most current calculations and measurements, I was becoming optimistic about the possibility of creating a laser. The tension was building and I started to have dreams that I actually did it. It was a wildly exciting thought!

Still, I was worried that I might be missing something. Instead of proceeding directly to the actual design, I decided to do an experiment that would check the validity of my model.

I made up a cube of pink ruby with each dimension equal to one centimeter in length, about the size and shape of a sugar cube. The crystal axis was perpendicular to one set of faces of the cube. I then placed the cube between two parallel plates, thus forming a microwave cavity. The cavity resonance was designed to be equal to the natural ground-state splitting of the ruby. These are the ground-state sublevels used in a ruby maser.

The purpose of this arrangement and the subsequent experiment was to monitor the ground state population of the ruby. My calculations had shown that it would be possible to substantially reduce the ruby ground state ion population necessary to make a laser, but I wanted tangible evidence to confirm and justify my optimism.

I was still harassed by the thought that no one had yet ever made a laser, although by now several teams of scientists had been diligently working away for nearly two years. Was there a fundamental insurmountable problem that I had missed?

I proceeded with the experiment using the ruby cube. As described above, one set of parallel faces of the cube served as the resonant microwave cavity. I connected one of the open (second set of) faces of the cube to one end of a polished quartz rod called a light pipe. The other end of the light pipe was connected to the flashlamp. The third set of faces of the cube was used to probe with selected wavelengths of light. I used the 'loading' on the parallel-plate microwave cavity to monitor the ground state population of the ruby, making use of the microwave properties of ruby that I had learned from my work on the ruby maser.

When I flashed the lamp, the cavity loading did indeed change. It decreased because, as expected, there were fewer chromium ions in the ground state. More important, the *magnitude* of the population change, 3%, was very close to the value I had predicted from my analysis of the experimental parameters. This was an extremely gratifying result since it gave further confirmation to my model and its analysis.

Now I was getting more excited. Since I couldn't think of anything else to check, it was time to proceed with the definitive design of a laser.

The Laser Design

My first thought was to use some modification of the elliptic-cylinder configuration that I considered for the mercury arc lamp as discussed in the previous section. The problem was, I would need a straight, rod-shaped flashlamp, but the straight lamps in the catalogues that I checked didn't have high enough intensity. Therefore, I resorted to a different design and stuck with what was available – the spiral-shaped lamps. I didn't want to get sidetracked at this time into the development of a special lamp.

It dawned on me that I would not have to devise a focusing arrangement, since the brightness of the radiation that is attainable at the focus of a mirror or lens can only approach, but not exceed, the brightness of the source. So, why not just place the crystal in close proximity to the lamp? That is, why not place the crystal *inside* the lamp helix?

To help gather the light, I placed a polished aluminum cylinder around the outside of the spiral. To fit the lamp, I used a pink ruby crystal in the shape of a right circular cylinder rod, about 1×2 cm. The ends of the ruby cylinder were polished flat, parallel to each other and perpendicular to the axis of the cylinder. For mirrors, I used evaporated silver.

I chose silver for the mirrors because it has the highest reflectivity of any metal at the deep red ruby laser wavelength. However, silver tarnishes quickly, and so with time, the thin output layer will change its transmission characteristics; it is not stable. To solve this problem, I had a thick layer of silver evaporated on both ends of the crystal and I scraped a tiny hole in the coating at one end. The laser beam would emanate from that coupling hole.

The Crucial Test

As I neared completion of my laser design, Bob Hellwarth, one of my colleagues in the Atomic Physics Department, asked, 'How will you know if it's working?'

At first, I was concerned that if I were stuck with a ruby crystal that deviated too much from optical perfection, these imperfections could prevent measurable laser behavior.

To get a better understanding of the processes, I elaborated on my existing analysis to account for an imperfect crystal. I found that if I were able to drive the ruby crystal reasonably far above the point of inverted population, or laser threshold, that I would see very substantial evidence of an inverted population and impressive laser behavior even with a poor crystal.

Just exactly what would I see?

I planned to monitor the red light emitted from the ruby through the hole in the output silver coating and expected to observe three kinds of laser behavior evidence: a reduction in the decay time of the fluorescent level; the bunching of the radiation into a beam; and a significant reduction of the spectral width of the red light.

The Laser Process

Laser action can be understood by following through the operational details in the above-mentioned ruby laser design. The process starts when the ruby crystal is excited by the flashlamp and chromium ions are excited into the metastable fluorescent level. The ions lose energy by radiating randomly red photons. This is the familiar red fluorescence (spontaneous radiation).

When the excitation level is great enough, that revered 'inverted population' condition is reached. In that case, more chromium ions are in the ruby's metastable upper fluorescent level than are in the ground state. Therefore, chromium ions can radiate by *stimulated* emission as well as by the normal spontaneous emission process.

Because the metastable level now empties much faster than it normally would, the fluorescent lifetime is reduced. That decreased lifetime can be observed by monitoring the red ruby glow with a photoelectric cell connected to an oscilloscope when the ruby is flashlamp-excited.

When the ruby is in the inverted population condition, as explained previously, it becomes an amplifier. The red photons are amplified as they progress through the crystal.

An important selection process starts to take place. The red photons are initially emitted in random directions. But the fluorescent photons that happen to be radiated at large angles to the mechanical axis of the ruby cylinder are lost through the sides of the crystal. On the other hand, photons radiated along the crystal axis, or at small angles with respect to the crystal axis, are in effect, trapped. They are reflected when they strike either one of the end mirrors and move back and forth through the crystal. As these axial photons move through the ruby they are amplified and consequently generate more photons in the same direction. They quickly become the dominant stimulated radiation from the excited metastable level.

Keep in mind that photons moving through the crystal are responsible for the stimulated emission. The axial photons pick up a following and march down that crystal axis. Consequently, the photons that emerge from the coupling hole in the output silver mirror are concentrated in a direction along, or nearly along, the crystal axis.

The red ruby fluorescent (spontaneous) emission extends over a distribution of frequencies in a curve that resembles a bell shape. The fluorescent photons are most concentrated at the center of that curve. When the inverted population condition exists, the top of the curve has the highest amplification. The consequence is (as with the discussion of beam angle) that photons near the center of the distribution are favored, since it is these photons that get amplified most in the back and forth transits, through the crystal from the multiple mirror reflections. It is this last process that explains why the frequency distribution, the linewidth of a laser, is so small.

I planned to vary the excitation to the flashlamp. In doing so, as explained above, I would expect to see a reduction of the fluorescent lifetime giving evidence of stimulated emission. As I varied the ruby excitation level, the fluorescent intensity should increase proportionately, as long as I was below threshold. But, when above threshold, small increases in excitation should make much bigger changes in the output since the detector and monochromator combination are more responsive to the smaller beam and narrower linewidth.

Do it!

It was the afternoon of May 16, 1960; it was time to confirm or deny all the fears of why 'ruby can't work' or why 'lasers can't be made to work.' No more new calculations, no more diversionary experiments. This was the moment of truth!

The laser head was mounted on a workbench. The flashlamp was connected to the power supply. The trigger electrode was connected to the spark

coil (the mechanism that initiates the flash from the strobe lamp). The light output from the coupling hole in the end of the ruby was directed through the monochromator to a photomultiplier tube. The electrical signal from the photomultiplier was connected to an oscilloscope.

Irnee D'Haenens, my laboratory assistant, and I were the only ones performing and observing the experiment. We first took a test shot so that we could adjust the monitoring equipment. We turned the power supply up to about 500 volts. We fired the flashtube. Indeed, we observed a trace on the oscilloscope!

The trace was a recording of the red ruby fluorescence. The decay in the trace was about three milliseconds, the lifetime of the upper possible laser level. We made the appropriate adjustments to optimize the monitor display.

We continued. We progressively increased the supply voltage, each time monitoring and recording the light output trace. As we did so, the peak output increased proportionately to the energy input, the decay time remained the same. So far, so good.

But, when we got past 950 volts on the power supply, everything changed! The output trace started to shoot up in peak intensity and the initial decay time rapidly decreased.

Voilà. This was it! The laser was born!

The Light Fantastic

When Irnee and I observed the first laser go into action, Irnee was smiling and jumping up and down with glee. I was numb and emotionally drained from all the tension and excitement.

Imagine: in the 10 years prior to the laser, the coherent electromagnetic spectrum had been extended by perhaps a factor of five. Now, with the advent of the laser, there was a quantum jump in that spectrum of ten thousand!

The significance of my historic accomplishment didn't sink in right away. I'm not sure that it has yet.

THE HELIUM–NEON LASER

A Javan, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005, A Javan. Published by Elsevier Ltd. All rights reserved.

The date, twelve of December nineteen sixty, and the time, 4:20 PM, are the date and the time of an event that triggered such an imortant segment of what has now evolved in modern optics. Here is how:

It had snowed heavily on that day. The laboratory at Murray Hill, New Jersey, the Bell Telephone Laboratory, BTL, had been closed at 4. What occurred was a triumph of our depth of understanding of the atomic spectra, the physics of the inter-atom collisions and electron impacts and exchange effects, the fascinating sequence of what takes place in a *gas discharge plasma* when the discharge is switched on. The triumph was also that of the art in spectroscopy, this time that of the colorful glow of a gas discharge plasma in a controlled mixture of helium and neon. Performing the spectroscopy, one could *effectively* watch and follow a set of fascinating physics effects in the discharge plasma, that would lead to an *inversion* to set-in in a specific transition between a high-lying pair of energy levels in neon, by the now well-known transition at 1.15 μm wavelength in neon. Helium atoms in a metastable state in the discharge plasma played the key role.

On that day: The laser had already been de-gassed for 5 days and nights at 600 °C under the high vacuum, to drive off the impurities prior to introducing the gas mixture in the laser-plasma quartz tubing and sealing it off. We had a way to optimize the mixture, at the exact mixture ratio for the optimum inversion and the laser gain at the 1.15 μm neon transition wavelength. The laser design had a *daring* interferometer length *one meter long*. Why *daring* at *one meter long*, I will have a chance to explain why below. The laser-mirrors at each end of the one-meter-long interferometer were situated internally in the laser, exposed to the gas mixture. Two fine micrometers at each end of the laser provided the fine control for the laser–mirror alignment control externally.

To convey the scene, we had known already, that the inversion and the laser gain at the 1.15 μm neon transition will be *higher* in the discharge '*afterglow*', the afterglow following *rapidly switching-off the discharge*. The inversion and the laser gain at the 1.15 μm wavelength, we knew, would last for as long as a millisecond in the afterglow. Certain that the inversion and the gain will be higher in the afterglow, we had set-up the electronics to switch on the discharge for about 10 milliseconds, and switching it off fast at a one microsecond switch-off time, and leaving it off for the same 10 milliseconds time interval, and repeating it at a low rep-rate, about 20 times a second on that day. Expecting that while searching for the laser signal, the laser will first break into oscillation in the afterglow where the laser-gain will be higher, knowing that pulling it into the glow afterward will be easy.

With me on that day I had two of the three good people who worked with me on the experiment. One of the two, Donald Herriott, was about to miss his car pool. He had his hand at the time on the laser–mirror tuning-alignment, the fine micrometer at one end of the laser interferometer on his side. I had my back towards the oscilloscope that would display the laser signal, aligning the detection system. My journey had begun over two years previously. I was *certain* on that day, and at that time, that possibly right then, or no later than hours later or by midnight, the moment would come for me and others that I had with me, to witness the onset of the event sought after so hard. And then … .

I had my back towards Don Herriott, urging him to stay a bit longer. I heard Don's voice, saying '*What is that*?'. I turned and looked. The laser signals in the discharge afterglow on the oscilloscope screen, synchronized to the gas discharge rep rate. I reached the laser interferometer fine-tuning micrometer on my end … a gentle tuning, in a mere few seconds time, optimized the afterglow signal, and then pulled the laser signal right into the glow. I loudly called on 'Ed' to switch on to CW. He, Edward Ballick, jumped over the equipment and the electronics on the floor in the lab, reached the switching electronics, switched it on to CW:

> The laser signal on the oscilloscope … switched to CW … a coherent light beam … CW … continuous … and at a degree of coherence … that as it proved later and I know then … to the limits that the laws of nature will permit. I looked at my watch, 4:20 PM.

The year 1960 was the beginning. Some months earlier, in the summer of that year, Theodore Maiman at Hughes Laboratories, in Malibu, CA, succeeded to operate an optically pumped laser, operating in short-duration laser pulses, the Ruby laser at 0.694 μm. That approach to the laser differed fundamentally

from the discharge plasma approach. The two were different inventions and served different purposes. One, the *optically pumped*, converts the *optical energy* in the pump, to the laser light output. The other, *electrically pumped*, converts the *electrical energy* in the *electrical current* that excites the laser plasma, to the laser light output. The rest on how things evolved is history.

The Scene Before Lasers

Before lasers, optics had long passed its peaks. The last words had been written by Born and Wolf in 1957, in an impressive volume on Physical Optics. When I went through my graduate studies at Columbia University in the early fifties, there were no graduate courses offered in optics. The subject was dealt with incidentally in courses in electricity and magnetism. Action was all on the microwaves at the time. Optical spectroscopy had reached its peaks much earlier yet, before World War II. Condon and Shortly had written the last words on Theory of Atomic Spectra in 1935, amazingly, as it seems, in *hardly a decade* after the discoveries in quantum mechanics. As events had it, optics gave way to the microwaves after the war, and optical spectroscopy to the RF and microwave spectroscopy – and all this, a result of the discoveries in Radar at the microwave frequencies in World War II.

We owed everything we knew of the atomic spectra to the *gas discharges* used as spectroscopic light sources at the time, from the high-resolution spectra of the hydrogen atoms at one end of the periodic table, to that of the trans-uranium elements on the other. We need not be reminded, everyone knows, how much we owed then and do now, to the high-resolution spectra of the hydrogen atoms, all because of the early hydrogen discharge light sources and then at the beginning, what led to the discovery of the quantum mechanics, and the wave nature of the atoms. The excitement that followed has remained unparalleled in physics since. Within hardly a decade after as a result, the cleverly designed spectroscopic gas-discharge sources, of different makes and designs, exhaustively revealed every feature predicted by the quantum mechanics in the high-resolution spectra of every element we know in nature. Gas discharge light sources set the course at the time, as did the lasers after the events in 1960.

Optics had given way to the microwaves in the fifties after the war. The physics of the gas discharges, however, continued as an active research field after the war, although of short duration. Bell Telephone Laboratory had a most distinguished research department in gas discharges in the early fifties. Towards the late fifties, as it seemed, the gas discharges as a field in research, also seemed to have been exhausted. The payoff at the time had evolved in the colorful neon light sources and fluorescence lamps, a thriving industry all over the country.

In the summer of 1958, Bell Telephone Laboratory had dissolved their research department in gas discharges. They had transferred every one of their distinguished research people to the other research areas in the laboratory. Some good people had left earlier, already.

As the fate in gas discharge physics had it, I joined the Laboratory in September 1958 – with ZERO backgrounds in gas discharge physics. I knew physics.

How Things Evolved

It has been said that when time is ripe for a new idea, if one person misses it, the next one will not. Time was ripe in the late fifties for the optically pumped idea for the laser, but far from it for the *electrically pumped*. Optical pumping was an active field in research at the time. The French had introduced the 'pompage optic', the optical pumping. With the French pompage optic, one would by optical pumping, resonantly excite an atomic species from its ground into its excited electronic state, and perform a double-resonance at an RF frequency on a fine or hyperfine structure of the excited electronic state. There were other versions of the French pompage optic. Each version used the optical pumping to induce resonantly, an atomic species into an excited electronic state.

Important, and more so, in the fifties, in the second half, the discoveries of the three-level masers at the microwave frequencies had attracted much attention. The three-level masers are the counterparts at the microwaves of our now optically pumped lasers. In three-level masers, one drives to saturation a pair of energy levels at microwave energy-spacing in a paramagnetic solid, and allows the inversion and the gain to set-in in a transition to a low-lying third energy level at the microwave frequencies. Nicholas Blumbergen at Harvard had proposed the three-level maser idea in an important publication in 1956. I had my own original three-level maser ideas, and had them published in 1957. In my work I had discovered a fascinating new effect, and for the first time, that in the transition to a third level in a three-level maser, it will be possible to drive the maser into its self-oscillation, *without* requiring the inversion. The effect had led me to the idea of the Raman masers, requiring a '*pumping*' microwave field, followed by a Raman transition across a pair of energy levels at microwave energy-spacing. In my publication of the

effect, I had also considered the pumping field at an optical frequency, instead of at the microwave.

The ideas were 'in the air' at the time, to extend the three-level maser approach into the optical frequencies. Yet, in the summer of 1958, on an occasion in June of that year, I think on the first Monday in June, *flashed* through my mind a far different approach, all because of the *helium metastables*. Ideas always come to mind in a *flash*, the rest follows.

Summer of 1958: A Bit of History

I think it was on the first Monday in June 1958. I had been on an occasion visiting the Bell Telephone Laboratory (BTL) on that day. They had approached me that I join their research staff at Bell. My work on masers – I had done as a part of my post-doctoral research position in physics at Columbia University in New York City – had in part overlapped the then ongoing work at Bell. They knew me, and I knew Bell.

At the visit meeting with Arthur Schawlow, whom I knew well, he gave me his news on that day, that he and Charles Townes, had submitted to the *Physical Review* a paper for publication proposing an optically pumped counterpart of a three-level maser at the optical frequencies – by-now, the well-known Schawlow–Townes optically pumped laser publication. Specially, I was attracted to Schawlow's idea on a two-mirror Perot–Fabry-type interferometer as a resonator at the optical frequencies for the laser. The use of it in that publication was Schawlow's idea.

My thoughts, it may seem paradoxical, right there and then moved far away from an optically pumped laser as a possibility. The helium atoms in a *metastable state* flashed through my mind. I will avoid explaining why or how, other than, of a three-level atom, two of the levels were now to belong to a *different* atomic species, the helium atom, and that there will be no optical pumping … and so much '*electrical energy*' could be converted to and stored in the energetic metastable helium. A new possibility presented itself, with no resemblance to the optical pumping – an energy-exchange effect taking place resonantly, between the highly energetic helium metastables, and – as it evolved – another inert gas, the neon atoms in a helium–neon gas mixture. The resonant nature of the effect embodied everything we know of the wave nature of the atoms. I will have a chance to explain the effect shortly below.

Driving home at the end of the day on that day, a two hour drive from Murray Hill, New Jersey to uptown New York City, the thought of the metastable helium had crystallized before me the task to examine the effect in a helium-plasma in a gas discharge, in a gas mixture with the other *four* well-behaved *inert gases*, from neon through xenon. There are four.

Within weeks in June, searching through the literature I learned, that over the years, and mostly in the early fifties, the helium plasmas in a gas discharge in mixtures with *every other* inert gases had been the subject of most extensive research, *except in a mixture with neon*. The mixture with neon, for the reason that the reader will find below, had been thought of as *uninteresting*. The matter was left for me.

I had accepted in June already, to join the Bell research staff at Murray-Hill, and I did on September. From my work that entire summer, unfolded the entire theory for the helium–neon gas laser. There was also another far-reaching offshoot.

The Physics Effect

Next to the hydrogen atom, the helium is the most fundamental of all atoms in physics. Hydrogen has one electron. Its spectrum, detached from the complexities of the multi-electron atoms, carries the entire signature of the wave nature of the atom, in ways that are subject to accurate predictions and the measurements. The helium atom has two electrons. Its spectrum carries the entire signature of the physics effect of the entire multi-electron atoms, also in ways that are subject to accurate predictions and the measurements. In its metastable states, it carries special *dynamic features*, deeply fundamental, and special *only* to the helium metastables.

In a gas discharge in pure helium at a moderate gas pressure, the discharge color takes on a grayish-white with a tint of blue. Over 500 excited electronic states in helium have been identified – their excitation by the energetic electrons in the discharge, followed by their spontaneous radiative decays, the emissions of photons, accounts for the high-resolution spectra of the discharge grayish-white with a tint of blue in color. Of these 500, all of them, high lying in energy, are short-lived radiatively, with the exception, however, of two amongst them. The two have the lowest excitation energy. Their radiative decays to the ground electronic state are deeply forbidden; they cannot decay radiatively. The two form the well-known helium metastables, identified by their spectroscopic designations: $^3S_{1/2}$, the triplet S one half, and $^1S_0$, the singlet S zero – generally referred to in the literature as, the triplet S and the singlet S metastables. Of the two, the triplet S metastable has the lowest excitation energy. It lies at 19.8 electron volts (eV), above the helium ground state. The singlet S metastable lies higher, at 20.2 eV.

Helium atoms in their ground electronic state are chemically inert. A helium atom's two electrons form a closed shell. It cannot enter in a chemical reaction, neither with its own species, nor with any other, as we all know. The situation is, however, far different with every one of the 500 excited electronic states in a helium gas discharge. They are energetically high-reactive, except, and with the exception of the two metastables, they are all *short-lived*, and cannot play an important role.

The effect manifests itself differently in the long-lived metastables, and accounts for their *special dynamic features,* special only to the helium metastables. What follows outlines the effect.

Dynamic Features: Why so Special

By far the majority of the atoms in the helium gas discharge are in their inert, ground electronic states. Once in its energetic metastable state, a helium metastable in the gas discharge encounters in an impact – in a collision – with a helium atom in its ground electronic state. The encounter in the impact is of a short duration in our scale of time, but long in the atomic time-scale, of about 2×10^{-13} second, a fifth of a picosecond. What occurs is strictly wave-mechanical, dictated by the wave nature of the electrons in the atoms. During the encounter, the electron wave functions in the encountering atoms overlap momentarily. A momentary binding occurs. In no ways can one tell during the encounter which electron is which: Electrons are the properties of their wave functions – once overlapped, individual electrons lose their identity, in no way can one tell which is which. As the atoms fly apart, caused by the kinematics of the impact, an *electron exchange* occurs, *wave-mechanically*. The helium atom initially in its ground electronic state, will fly away from the encounter with one of its two electrons now in the *highly excited* energetic *metastable* state. The other, the one initially in the excited metastable state, emerges from the encounter in the *unexcited* ground electronic state.

At a gas pressure in the discharge of about one torr, a millimeter, the mean free-time between two impacts is about one microsecond, a very long time in the scale of the atoms. The helium atom that flew away from the encounter in the metastable state remains in its now metastable state, during the entire mean-free collision time. Once it encounters the next helium atom in its ground electronic state, the act repeats again, a momentary binding occurs, an electron exchange takes place, and a new metastable flying away. The effect repeats over and over again, at about every microsecond. A *diffusion* of the metastability through the helium in the discharge sets in. Depending on the diameter of the discharge tubing, it takes about a millisecond until the helium that emerges from the encounter in the metastable state reaches the walls of the discharge tubing. The act will then *terminate*, in a *most violent* collision effect with the wall. The energetic metastable helium *knocks out* an electron off the wall, as itself emerges from the collision in its ground electronic state – all intact.

With one exception, the encounter between a helium metastable with any species other than of its own kind, a helium, is energetically *highly violent*. The exception is the neon atoms. In encounters with all atomic species other than neon, the metastable helium *knocks out* an electron off the species, ionizes the species, and itself emerges from the encounter in its ground electronic state, intact. The ionizing impacts had attracted so much attentions in the early fifties, the effect in the helium gas discharge in mixtures specially with every other inert gases, had been exhaustively investigated, except with neon. Neon has its ionization potential higher than the two metastable excitation energies. It can not undergo an ionizing impact with either of the two metastables. The mixture with neon had thus been considered uninteresting, and was discarded.

What emerged by the end of the summer of 1958 was the entire theory for the helium neon laser – and the recognition of a deeply wave-mechanical effect, the resonant energy exchange taking place between the energetic helium metastable and an inert neon atom in the discharge gas mixture. In the mixture, a momentary binding occurs between the helium metastable encountering an inert neon atom in the gas discharge. Wave-function overlaps occur for the electrons in the encountering atoms. A *resonant* energy exchange takes place during the encounter. Helium metastable will fly away from the impact in its now ground electronic state, while neon emerges in one of its high-lying excited energy states, at the near-resonant with the helium metastable.

A special effect sets in in the energy exchange, as well. The kinetics of the impact in the encounter enters in satisfying the exact resonance in the energy exchange. The singlet S metastable flying away from the encounter after the energy exchange, will be slightly 'cooler' kinetically, and the triplet S flying away from the encounter after the energy-exchange, will be slightly 'hotter'. The kinetic energy of the impact enters in the energy exchange.

Optical Gain and Amplification Effect

Within weeks, early in the summer of 1958, investigating the energy transfer effect led to the

prediction that there can be inversion and optical gain in two sets of transitions in neon, one as a result of the excitation transfer from the triplet S metastables, and the other a result of the transfer from the singlet S. By far the more challenging task that summer, however, was to estimate theoretically the *values* for the *gain and amplification factors* obtainable for the predicted inverted transitions, and more specifically, for the inverted transitions that will have the highest gain and amplification factors. There followed the theoretical estimates on which relied the entire experimental phase of the work, the planning of it commenced within weeks after my joining the Bell Telephone Laboratory on September of that year.

The set of transitions originating from the triplet S energy transfer to neon, occur in the near-red region of the spectrum, the 1 μm wavelength region. The strongest inverted transition at the highest predicted gain value, lies at 1.15 μm transition wavelength, the spectral laser line that first broke into self-oscillation (see the opening page in this article). The set of transitions originating from the singlet S energy transfer to neon occur in the red-yellow region of the spectrum. The strongest, although at less gain-value than 1.15 μm, occurs at the 0.633 μm in the red.

In the early phase of the investigation, it came to light also that the achievement of the inverted transitions with optical gains, was, interestingly, a *commune occurrence* in gaseous discharges in a variety of the atomic species, including in pure neon. However, the complexities in the excitations by the hot electrons of the high-lying energy levels, and in particular in the presence of the secondary excitation mechanism, and a list of other similar effects, in no way would allow a quantitative estimate of the gain and amplification factors in the inverted transitions obtainable in the gas discharge.

The reliance on the excitations of the high-lying energy levels in neon via the resonant energy transfer from the helium metastables, as per the physics effect I have outlined above, placed at my disposal the one gas discharge system that lent itself to the theoretical estimates of the predicted optical gains and the amplification factor for the inverted transitions.

Without a knowledge of the predicted values for the optical gain and the amplification factor, it would have been in no way possible to perform the experiment.

Experimental Phase

The Bell Telephone Laboratory, considered as a foremost research institution in the country at the time, operated at a $250 million yearly budget – in 1960s' dollars. Transistors had been invented at Bell in the forties. With the awareness at the time at Bell, that a light beam at the degree of frequency-purity and the coherence expected of the laser, will have far-reaching consequences on the telecommunication technology, without hesitation I received from Bell an open ticket to do what it takes – one key person in the administration understood in-depth my approach to the laser – via the gas discharge and the physics effects I had invoked. Within weeks following joining Bell in September, I had my planning underway for what came to be known in a short while later, as a most aggressive research in gas discharges – and at a time when Bell had already, in the summer of 1958, dismantled the entire research activities in gas discharges at Bell. That speaks a great deal for Bell at the time … at the time before the AT&T break up.

In lasers, the knowledge of the full length of the amplifying medium across which the optical gain obtainable at the laser transition will suffice to drive the laser into its self-oscillation, will dictate the full length of the laser interferometer resonator. The spacing between the two high-reflectance interferometer end-mirrors will have to accommodate the full length of the amplifying medium, thus dictating the interferometer full length. With respect to the mirror reflectivity obtainable at the time, one had to be content with the values at the highest close to 98%. The highest gain transition from my theoretical estimates, that of the neon transition at the 1.15 μm wavelength in the near-red, indicated an interferometer length of no less than 50 centimeters, or possibly as long as one meter – and there lay the challenge, achieving the required degree of parallelism needed for the interferometer end-mirrors at such spacing. One could show, from simple inspection, that the parallelism needed for the mirrors had to be at values better than one part in several-hundredths of a wavelength. This value applied to the use of the flat interferometer mirrors known to me at that time.

An alignment to such a degree of accuracy could well be obtainable at the time for Perot–Fabry-type interferometer mirrors at the mirror-spacing not exceeding about eight or ten centimeters at the most – where one would rely on the observation of the interference fringes using the available spectroscopic light sources to observe the fringes, a principle well known in physical optics. At longer mirror spacing, the spectroscopic light sources lacked the degree of coherence, the color purity, needed to observe fringes for the mirror alignment. The fringes will get wiped out.

These days, using a helium–neon laser light source available at every laboratory in optics, one can align a two-mirror interferometer at tens of meters mirror-spacing, or hundreds, or even as far apart as

several miles in some experiments. Ideas are floating around now of the possibility, using gas lasers, to align mirrors placed on satellites in orbit, with the mirror on two different satellites at controlled spacing as far apart as thousands of miles – the experiment thought of is that of the detection of the gravity waves. Without a laser operating and available at the time, in no ways would it have been possible to align an interferometer at a *daring* mirror-spacing as far apart as one meter, or even half as long. One had to rely on an auto-collimator, an optical instrument that at the best would provide an alignment to within one or two wavelengths, nowhere sufficient to drive the laser into its self oscillation.

The challenge mapped the entire course for the experimental phase, foremost aimed at showing every feature of the predicted effects experimentally, from the energy transfer effect, to observing and measuring the gain and amplification effect in the predicted inverted transitions in neon. The feat was that of the art in spectroscopy. The observation of gain and amplification effect was a highlight in the course of the experiment. Up to that point, in optical spectroscopy, one always measured the spectral lines in either emission or in absorption. The feat was the first in spectroscopy, observing the spectral line of an atomic species not in absorption, but in amplification – the amplification of an external light source transmitted through the gas discharge in the experiment. The feat, yet, lay in the art in spectroscopy, throughout the experiments.

The event at 4:20 PM in December two years later, speaks for itself.

> It may seem now, as if one needed an operating helium–neon laser at the time, to align the laser mirrors for the first operating helium–neon laser. It seemed evident from the beginning, that the presence in the very first helium–neon laser, of the optical-gain at the predicted value in the predicted inverted transition, will guide achieving the alignment of its own mirrors, to the exact optical degree of the alignment accuracy needed for the self-oscillation. Driving the laser into its self-oscillation signaled the feat. The feat, however, was that of the power in spectroscopy, making it possible to predetermine the presence in the first helium–neon laser, of the optical gain, at the predicted optimum value needed for self oscillation.
>
> There were doubters and doubters, in the technical staff and the administration at Bell. Donald Herriott, William Bennett and Edward Ballick collaborated with me in the experiment.

The Impact

Within months of the publication of the work in January 1961 in the *Physical Review Letters*, walking through the long hallways at BTL where the laboratories were situated, one could see through the open doors in many laboratories down the long halways, the pink colors of long helium–neon gas discharges plasma tubing, with people in the labs searching for new laser transitions. BTL had revived the gas discharge research activities they had dismantled in the summer of 1958. Not surprisingly, the second set of new laser transitions driven into self-oscillation took place at Bell in 1962, that of the set of transitions in neon originating from the energy transfer from the helium singlet S metastable – with the most intense in the series, that of the by now well-known 0.6328 μm neon transition in the red. There followed the discoveries of the argon–ion gas laser and shortly thereafter the CO_2 gas laser, the two amongst the foremost in the many new gas lasers driven to self-oscillation at Bell.

Every gas laser discovered since the December 12 event, has enjoyed the use of a '*no-longer*' a '*daring*' '*long*' interferometer resonator, merely because one would use a He–Ne laser to *pre-align* the laser mirrors to within a small fraction of the wavelength needed, prior to searching for a new laser line in the gas discharge chosen for the search. The already aligned resonator removed the need to *predetermine* the presence of the gain and the amplification effect at the values needed for self-oscillation. Within a span of about a decade, virtually hundreds of different gases were driven to self-oscillation with this method, operating literally at thousands of different wavelengths, ranging from the far infrared, the 100 100 ηn region, to the near-UV.

The helium–neon laser triggered so many of the key discoveries that followed. Within months, in the spring of 1961, one of the distinguished engineers at BTL, Rudolf Komphner, the discoverer of the backward tube oscillators at the microwave frequencies, began experimenting with the transmission through the improvised thin fiberoptic, of the output beam of a helium–neon laser at the 1.15 μm transition in the near-red, the near-IR, having his eyes on the potential someday in communication on the use of a continuously operating laser light beam at the near-IR frequencies. I recollect the very first ad that appeared on lasers nationally was a BTL ad in the spring of 1961, showing a cross-section of a beam of light, underneath it saying: '*Someday a Beam of Light This Size Will Carry Millions of Telephone Conversations Simultaneously*'. The ad speaks for itself.

In every gas laser, the active laser medium – being a gas – is highly homogeneous. As a result, with a reasonably good optics, one can obtain a nearly perfect Gaussian output beam. For this reason, the helium–neon lasers available commercially since the

mid-sixties, have served as the indispensable tools for the alignment of the complex optics in every research laboratory in lasers, or in every manufacturer of the lasers in the country, or the manufacturers of the precise optical components, or the medical uses and so on.

In the applied areas, the helium–neon laser triggered so many of the key experiments in the early days. Philips at Eindhoven can be considered as the first discoverer of a video disc recorder. The model used a helium–neon red laser, at a most impressive fidelity and color resolution, in recording and in reading for the display. I recollect, in an exhibit shown to me at a Philips Lab, one groove on the recorder was that of a horse in a prairie, and the next groove, one '*micrometer away*' on the recorder, was that of a crowd in a park – so impressive at the time, with no cross talks. They had it packaged in a compact unit already. In other areas, all bar-code readings were done using a helium–neon laser, and some still do.

In the scientific area, from high-resolution spectroscopy at unprecedented accuracy, to the accurate clocks operating at optical frequencies, or in the measurements of the absolute frequencies of the accurate laser clocks, or the early experiments in relativity and so on, all are owed to the precision gas lasers. Helium–neon laser has played the key role, in the early days throughout, and continues today.

A note to the reader

In the early publications on laser, the device was generally known and referred to at the time as the 'Optical Maser'. The word laser became an accepted designation for the device in the late 1960s and the early 1970s.

Suggested Readings

Javan A (1961) In: Singer J (ed.) *Advances in Quantum Electronics*, pp. 18. Columbia University Press.

Javan A, Bennett WR, Jr. and Herriott DR (1961) Publications reporting operation of the first helium–neon laser. *Physical Review Letters* 6: 106.